$\mathcal{S}^1$: A PRECISELY INTERPRETABLE STATE SPACE MODEL VIA AN END-TO-END CLOSED-FORM OPERATOR

Anonymous authors
Paper under double-blind review

ABSTRACT

The linear dynamics of state-space models are – in principle – mathematically solvable, yet the nonlinear readouts required for expressivity render the full computation inaccessible to mathematical analysis. We introduce $\mathcal{S}^1$, a state-space model derived from a system of complex nonlinear oscillator networks in which the entire computation – both dynamics and readout – admits an exact, end-to-end mathematical description. The key is a nonlinear coordinate transformation that simultaneously linearises the recurrent dynamics and uniquely determines the readout nonlinearity, which in turn enables an exact analytical expression for the end-to-end computation done by the network. A broad class of connectivity patterns gives rise to traveling wave modes in the recurrent layer, and classification emerges from nonlinear interactions between wave modes, which we illustrate in an exact manner. We validate $\mathcal{S}^1$ on EEG classification and sequential MNIST, providing a first-principles account of classification that no existing sequence model can offer, while also achieving competitive performance on benchmarks. The same mathematical framework also enables principled control: specific eigenmodes can be selectively amplified or suppressed to causally redirect model classifications. These results demonstrate a new SSM architecture that achieves competitive performance while adding precise interpretability and control.

1 INTRODUCTION

Sequence models are increasingly deployed in high-stakes domains – clinical timeseries, neural decoding, and scientific forecasting – yet the computations driving their predictions remain mathematically opaque. Transformers Vaswani et al. (2017) compute outputs through pairwise attention across all tokens, leading to a nonlinear operations on high-dimensional patterns that resist principled analysis. *Post-hoc* methods such as attention visualization or gradient-based attribution Elhage et al. (2021); Jain & Wallace (2019); Vig (2019) can identify *which* inputs drive the model strongly – similar in spirit to receptive fields in the brainHubel & Wiesel (1962) – but not *how* the system's internal computation produces that response. This is a critical distinction for fundamental mechanistic understanding of these models. While additional analysis techniques can provide further insight, a new architecture whose computation is mathematically transparent by construction could address many of these problems simultaneously.

State-space models (SSMs) Gu et al. (2021; 2022); Smith et al. (2022); Gu & Dao (2023); Orvieto et al. (2023) have emerged as a potential alternative to transformers. SSM systems are composed of two parts: a linear recurrence that receives sequence input and a nonlinear readout. The first part is quite tractable: one can identify eigenmodes and write the response to any input in terms of them. The second part makes the whole architecture intractable, however: SSM systems require a nonlinear readout to achieve high performance on sequence tasks, and this severs the connection between the eigenmodes of the linear recurrence and the system's output. The challenge is, then, not to eliminate the nonlinear readout – without it, the SSM architecture cannot approximate the range of input-output relationships encountered in practice Wang & Xue (2023); Muca Cirone et al. (2024) – but to derive an architecture that admits a tractable, closed-form end-to-end mathematical description grounded in first principles. In existing architectures, any such account can only be constructed post hoc, through approximations or attribution methods that are not derivable from the model's mathematical structure,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and therefore lack the precision and generality of a principled analytical treatment Lundberg & Lee (2017); Ali et al. (2025); Rudin (2019).

We address this challenge by grounding an SSM in the physics of a complex-valued nonlinear oscillator network. A specific nonlinear coordinate transformation renders this system exactly solvable by linearising it, with the inverse transformation uniquely determining the readout nonlinearity, leading to a system where the linear state and nonlinear readout admit an exact analytical description without approximation. The nonlinearity is thus not appended to the model as a heuristic activation function: it is derived from the model's own analytic structure.

The consequences for interpretability are immediate. The full computation of this new architecture – $\mathcal{S}^1$ – is expressible as an exact, closed-form operator that captures the full mapping from input to output. This operator can be completely parameterized in terms of the eigenvalues and eigenvectors of the connectivity matrix in the recurrent layer. With a broad class of connection patterns in the recurrent layer, these eigenmodes are waves that travel across the network, which can all be expressed precisely mathematically. The nonlinear readout induces specific interactions between wave modes, which can be precisely understood in terms of the end-to-end operator. Every stage of the computation admits a closed-form account – with no post-hoc attribution required.

We demonstrate $\mathcal{S}^1$ on EEG classification Birbaumer (1999) and sequential MNIST datasets LeCun et al. (1998), providing a mathematically grounded analysis of how the architecture performs classification in each case, and report the architecture achieves performance on the UEA multivariate timeseries benchmarks Bagnall et al. (2018); Walker et al. (2024) that is competitive with S4, Mamba, and LRU Gu et al. (2021); Gu & Dao (2023); Orvieto et al. (2023). These results provide initial evidence that complete mathematical explainability does not come at an empirical cost for sequence modeling problems.

Contributions. We summarise the contributions of this paper as follows:

- We derive $\mathcal{S}^1$ from a system of exactly solvable nonlinear oscillators, establishing a principled connection between nonlinear oscillator dynamics and linear state-space modeling in which the readout nonlinearity is uniquely determined by the model's mathematical structure.
- We show that $\mathcal{S}^1$ admits an exact closed-form operator description fully parametrised by the eigenvalues and eigenvectors of the connectivity matrix, with the nonlinearity arising from inter-modal interference – making $\mathcal{S}^1$ the first SSM in which the complete linear-nonlinear cascade is analytically tractable end-to-end.
- We develop a mechanistic interpretability framework grounded in the closed-form operator, connecting input-driven modal excitation to node-level phase coherence and class-specific spatial phase patterns, and demonstrate it on EEG and image classification tasks.
- We show competitive performance on UEA time series classification and sequential MNIST, establishing that analytical tractability and empirical competitiveness are mutually compatible design goals.

2 Mathematical background

We consider a system of N coupled oscillators on a one-dimensional ring, whose dynamics under external forcing are given by

$$\dot{\psi}_i = \sum_j \rho_{ij} \sin(\psi_j - \psi_i + \alpha_{ij}) - \mathrm{i} \left(\sum_j \rho_{ij} \cos(\psi_j - \psi_i + \alpha_{ij}) + [\mathbf{B}u(t)]_i e^{-\mathrm{i}\psi_i} \right), \quad (1)$$

where $\rho_{ij} = |[\mathbf{CKC}^{-1}]_{ij}|$, $\alpha_{ij} = \mathrm{Arg}([\mathbf{CKC}^{-1}]_{ij})$, $\mathbf{CKC}^{-1}$ is the adjacency matrix, $\mathbf{u}(t)$ is the external input, and $\mathbf{B}$ projects input channels to the N nodes. Using $\sin\theta - \mathrm{i}\cos\theta = -\mathrm{i}e^{\mathrm{i}\theta}$ and $[\mathbf{CKC}^{-1}]_{ij} = \rho_{ij} e^{\mathrm{i}\alpha_{ij}}$, Eq. equation 1 simplifies to

$$\mathrm{i}\dot{\psi}_i = \sum_j [\mathbf{CKC}^{-1}]_{ij} e^{\mathrm{i}(\psi_j - \psi_i)} + [\mathbf{B}u(t)]_i e^{-\mathrm{i}\psi_i}. \quad (2)$$

Exact solution and LTI system. Defining the lifted state $\mathbf{Cx} = e^{\mathrm{i}\boldsymbol{\psi}}$ (so that $\mathbf{y} := \mathbf{Cx}$ satisfies $y_i = e^{\mathrm{i}\psi_i}$), Eq. equation 2 transforms into the LTI system

$$\dot{\mathbf{x}} = \mathbf{K}\,\mathbf{x} + \mathbf{C}^{-1}B\,\mathbf{u}(t), \quad \mathbf{Cx}(0) = e^{\mathrm{i}\boldsymbol{\psi}(0)}, \tag{3}$$

whose continuous-time solution is $\mathbf{x}(t) = e^{\mathbf{K}t}\mathbf{x}(0) + \int_0^t e^{\mathbf{K}(t-s)}\mathbf{C}^{-1}B\,\mathbf{u}(s)\,\mathrm{d}s$. Discretising with timestep τ via zero-order hold and indexing so that u_k denotes the input driving the transition into step k (as in Gu et al. (2021); Gu & Dao (2023)), gives … gives

$$\mathbf{x}_k = e^{\mathbf{K}\tau}\mathbf{x}_{k-1} + \bar{\mathbf{B}}\,\mathbf{u}_k, \qquad \mathbf{x}_{-1} = \mathbf{x}_0, \qquad \bar{\mathbf{B}} := \int_0^\tau e^{\mathbf{K}r}\mathbf{C}^{-1}B\,\mathrm{d}r. \tag{4}$$

Nonlinear readout. SSMs such as S4 Gu et al. (2021) append a nonlinear readout to the LTI output, but this nonlinear readout is not fundamentally connected to the underlying dynamics. $\mathcal{S}^1$ produces mathematical explainability by deriving the nonlinear readout directly from the coordinate transformation that solves the network dynamics. Specifically, the inverse of the coordinate transformation between $\boldsymbol{\psi}$ and $\boldsymbol{x}$ satisfies

$$\boldsymbol{\psi}(t) = \mathrm{Arg}(\mathbf{Cx}(t)) - \mathrm{i}\log|\mathbf{Cx}(t)|. \tag{5}$$

Defining $\boldsymbol{y} = \boldsymbol{C}\boldsymbol{x}$ and using Eq.equation 5 we can identify a unique and canonical nonlinearity: the element-wise phase normalisation $y_{ik}/|y_{ik}| = [Cx]_{ik}/|[Cx]_{ik}| = e^{\mathrm{i}Re[\psi_i(k\tau)]}$ which exactly recovers the oscillator phase from the linear coordinate y_{ik}. The normalisation $y_{ik}/|y_{ik}|$ thus serves as the readout nonlinearity of the $\mathcal{S}^1$ model, uniquely determined by the map between the non-linear ψ and linear y systems rather than chosen heuristically. The output of the model, consisting of c class logits, is then

$$O_c = W_{ci}\left(\frac{1}{T}\sum_{k=1}^{T}\frac{y_{ik}}{|y_{ik}|}\right), \tag{6}$$

where $\mathbf{W} \in \mathbb{C}^{C\times N}$ is the learned readout matrix.

Spectral structure of the end-to-end computation

Diagonalization of the propagator. The operator equation 6 can be expressed entirely in terms of the eigenvalues and eigenvectors of the $\mathbf{y}$-system adjacency matrix $\mathbf{CKC}^{-1}$, which determines the propagator $\mathbf{D}_\tau = \mathbf{C}e^{\mathbf{K}\tau}\mathbf{C}^{-1}$. Using $\mathbf{K} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1}$, the propagator diagonalises as

$$\mathbf{D}_\tau = (\mathbf{CV})\,\boldsymbol{\Lambda}\,(\mathbf{CV})^{-1}, \qquad \Lambda_{\ell\ell} = e^{\lambda_\ell\tau}, \qquad \lambda_\ell = \alpha_\ell + \mathrm{i}\omega_\ell. \tag{7}$$

Defining modal coordinates $\tilde{\mathbf{y}}_k := (\mathbf{CV})^{-1}\mathbf{y}_k$ and modal input coupling $\tilde{\mathbf{B}} := \mathbf{V}^{-1}\bar{\mathbf{B}}$, $\tilde{b}_\ell := [\mathbf{V}^{-1}\bar{\mathbf{B}}]_{\ell,:}$, each mode evolves independently as

$$\tilde{\mathbf{y}}_k = \boldsymbol{\Lambda}^{(k+1)}\tilde{\mathbf{y}}_0 + \sum_{j=0}^{k}\boldsymbol{\Lambda}^{(k-j)}\tilde{\mathbf{B}}\,\mathbf{u}_j. \tag{8}$$

Modal decomposition. The state $\mathbf{y}_k$ can be written as a superposition of the eigenvectors $\boldsymbol{C}\boldsymbol{v}_\ell$ of the adjacency matrix $\boldsymbol{CKC^{-1}}$:

$$\mathbf{y}_k = \sum_\ell A_{\ell k}\,e^{\mathrm{i}\omega_\ell k\tau}\,\boldsymbol{C}\boldsymbol{v}_\ell, \tag{9}$$

where the complex modal amplitude $A_{\ell k}$ accumulates the input history through the filter $e^{-\lambda_\ell j\tau}$:

$$A_{\ell k} := e^{\alpha_\ell k\tau}\left(\tilde{y}_{\ell 0} + \sum_{j=0}^{k}e^{-\lambda_\ell j\tau}\,\tilde{b}_\ell\,\mathbf{u}_j\right). \tag{10}$$

Each mode ℓ therefore acts as a resonator: $\tilde{b}_\ell = [V^{-1}\bar{B}]_{\ell,:}$ defines the *input receptive field* of mode ℓ, determining which input patterns excite it, while the exponential kernel $e^{-\lambda_\ell(j)\tau}$ provides temporal filtering tuned to the mode's natural frequency ω_ℓ and decay rate α_ℓ. It is important to note that $\boldsymbol{C}$ acts on the node index of each eigenvector $\mathbf{v}_\ell$, not on the mode index. Each mode ℓ retains its eigenvalue λ_ℓ, natural frequency ω_ℓ, and decay rate α_ℓ — the effect of $\boldsymbol{C}$ is purely to reshape the spatial profile of mode ℓ across the N nodes, producing the transformed eigenvectors $\boldsymbol{C}\mathbf{v}_\ell$ without altering the temporal dynamics or spectral structure of the decomposition.

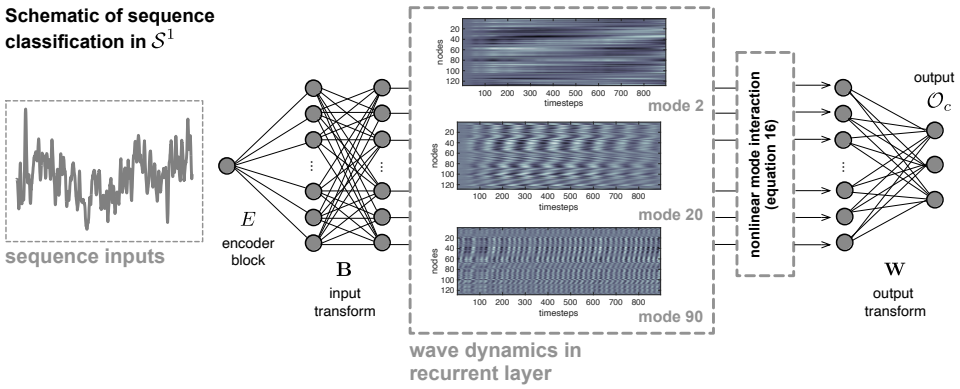


Figure 1: Schematic of sequence classification in $\mathcal{S}^1$. An input sequence is projected by encoder E and input transform $\mathbf{B}$ into a recurrent layer of N coupled oscillators, whose dynamics decompose into traveling wave eigenmodes of increasing spatial frequency. The nonlinear mode interaction (Eq. equation 12) extracts complex phase patterns across modes, and the readout $\mathbf{W}$ maps the resulting phasors to class logits $\mathcal{O}_c$.

Final operator. Projecting the modal decomposition onto the n-th node coordinate, the squared magnitude expands as

$$|y_{nk}|^2 = \sum_{\ell,m} A_{\ell k}\,\overline{A_{mk}}\,[\boldsymbol{C}\mathbf{v}_\ell]_n\,\overline{[\boldsymbol{C}\mathbf{v}_m]_n}\,e^{\mathrm{i}(\omega_\ell-\omega_m)k\tau}. \tag{11}$$

Substituting into Eq. equation 6 gives the fully explicit form of the readout:

$$O_c = \sum_n W_{cn}\left(\frac{1}{T}\sum_{k=1}^{T}\frac{\sum_\ell A_{\ell k}\,[\boldsymbol{C}\mathbf{v}_\ell]_n\,e^{\mathrm{i}\omega_\ell k\tau}}{\sqrt{\sum_{\ell,m} A_{\ell k}\,\overline{A_{mk}}\,[\boldsymbol{C}\mathbf{v}_\ell]_n\,\overline{[\boldsymbol{C}\mathbf{v}_m]_n}\,e^{\mathrm{i}(\omega_\ell-\omega_m)k\tau}}}\right). \tag{12}$$

Eq. equation 12 clarifies the rationale behind the choice of nonlinearity $[y/|y|]_{nk}$. Although the underlying dynamics are linear in the coordinates $\mathbf{x}$ or $\mathbf{y}$, the denominator explicitly couples all eigenmodes through inter-modal interference in the magnitude of the state. The nonlinearity therefore has a precise spectral interpretation: it is not an external activation but the mechanism through which the model accesses the nonlinear phase dynamics of the original oscillator system. This stands in contrast to existing SSMs, where the nonlinearity is applied after the LTI block and has no connection to the model's internal spectral structure.

The operator O_c is exact and fully parametrised by the eigenvalues and eigenvectors of the oscillatory system. Every stage of the computation – from the modal decomposition of $\mathbf{y}_k$, through the inter-modal interference encoded in the normalisation, to the final linear readout – admits a precise spectral interpretation. As a result, one can trace in closed form how any given input $\mathbf{u}$ influences the output O_c through identifiable eigenmodes, making the $\mathcal{S}^1$ model analytically tractable and interpretable by design rather than **post hoc** analysis, after the model's inferences are obtained.

3 The $\mathcal{S}^1$ Model

Architecture

The $\mathcal{S}^1$ model is a sequence-to-sequence architecture built around the linear time-invariant system derived in Section 2 and as illustrated in Fig. 1. An input sequence $\{u_j\}_{j=1}^T$ with d_{in} channels is first projected to N dimensions by a learned encoder $\boldsymbol{W}_{\text{enc}} \in \mathbb{R}^{N\times d_{\text{in}}}$. The encoded sequence is then processed by the LTI block, which computes the state trajectory $\{\mathbf{x}_k\}_{k=1}^T$ according to Eq. equation 4. The physical state $\boldsymbol{y}_k = \mathbf{C}\mathbf{x}_k$ is subsequently obtained by applying $\mathbf{C}$ to the output of the LTI block. The nonlinear readout $y_{kn}/|y_{kn}|$ is applied element-wise at every timestep, temporally averaged, and contracted with the complex readout matrix $\mathbf{W} \in \mathbb{C}^{c\times N}$ to produce c class logits via Eq. equation 6. The real part of the logits is used for classification.

Efficient computation via convolution

The full solution Eq. equation 4 separates into an initial condition term and an input-driven convolution:

$$\mathbf{x}_k = \underbrace{e^{\mathbf{K}\tau(k+1)}\,\mathbf{x}_0}_{\text{initial condition}} + \underbrace{\sum_{j=0}^{k} e^{\mathbf{K}\tau(k-j)}\,\bar{B}\,\mathbf{u}_j}_{\text{causal convolution}}, \tag{13}$$

where $\boldsymbol{x}_0$ is the initial state determined by the oscillator phases at the start of the input sequence, and the second term is a causal convolution of the input with the matrix-valued kernel $h_k = e^{\mathbf{K}\tau k}\bar{B}$. Following S4 Gu et al. (2021), this convolution is computed efficiently in the frequency domain. The convolution kernel is precomputed from the eigendecomposition of $\mathbf{K}$ and applied via FFT. This reduces the per-layer complexity from $\mathcal{O}(T^2)$ to $\mathcal{O}(T \log T)$, matching the computational profile of existing SSMs.

DFT Eigenbasis and Spectral Parametrisation

Fixing the connectivity within the recurrent layer to be a circulant graph renders the eigenvector matrix $\mathbf{V}$ to be the Discrete Fourier Transform basis:

$$v_{\ell n} = \frac{1}{\sqrt{N}}\, e^{-2\pi \mathrm{i}\ell n/N}, \qquad \ell,\, n = 0, \ldots, N-1, \tag{14}$$

while leaving the eigenvalues $\{\lambda_\ell\}$ as free learnable parameters. In practice the convolution is carried out in the modal coordinates $\tilde{\mathbf{x}} = \mathbf{V}^{-1}\mathbf{x}$ of the $\mathbf{x}$ system, where the dynamics are diagonal and each mode ℓ evolves independently under $e^{\lambda_\ell \tau}$. The physical state $\mathbf{x}_k$ is then recovered by mapping back via $\mathbf{x}_k = \mathbf{V}\tilde{\mathbf{x}}_k$, and the state $\mathbf{y}_k = \mathbf{C}\mathbf{x}_k$ follows. Crucially, regardless of what the eigenvalues $\{\lambda_\ell\}$ are or how they evolve during training, the full connectivity matrix of the $\mathbf{x}$ system can always be recovered as $\mathbf{K} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$. Similarly, the adjacency matrix of the $\mathbf{y}$ system is $\mathbf{C}\mathbf{K}\mathbf{C}^{-1} = (\mathbf{C}\mathbf{V})\mathbf{\Lambda}(\mathbf{C}\mathbf{V})^{-1}$, whose eigenvectors $\mathbf{C}\mathbf{v}_\ell$ are learned spatial transformations of the DFT basis. This has two consequences. First, since $\mathbf{V}$ is fixed throughout training, each eigenmode ℓ always corresponds to a traveling wave of spatial frequency $2\pi\ell/N$ in the $\mathbf{x}$ system, preserving a physically transparent spectral interpretation regardless of how the eigenvalues evolve. Second, the full eigensystem is available analytically without numerical diagonalisation at any point during training, keeping the closed-form operator Eq. equation 12 directly evaluable from the model parameters throughout.

Learnable parameters

The trainable parameters of the $\mathcal{S}^1$ model are: the basis change matrix $\boldsymbol{C} \in \mathbb{C}^{N\times N}$, the input projection matrix $\boldsymbol{B} \in \mathbb{R}^{N\times N}$, the encoder $W_{\text{enc}} \in \mathbb{R}^{N\times d_{\text{in}}}$ with bias $b_{\text{enc}} \in \mathbb{R}^N$, and the complex readout matrix $W \in \mathbb{C}^{C\times N}$ with bias $b \in \mathbb{C}^C$. The nonlinearity $\mathbf{y}/|\mathbf{y}|$ contains no learnable parameters – it is fixed by the coordinate transformation and introduces no additional degrees of freedom. The eigenvalues of the connectivity matrix $\mathbf{K} \in \mathbb{C}^{N\times N}$ are not trained in the experiments reported here, but they can be made learnable without any loss of analytical tractability

4 Experiments

4.1 UEA Multivariate Time-Series Classification

Although empirical performance is not the primary objective of $\mathcal{S}^1$, we evaluate it on six datasets from the UEA multivariate time-series classification archive Walker et al. (2024), covering sequence lengths $T = 400$–$18{,}000$ across diverse domains. Table 1 reports test accuracies averaged over five runs, comparing against S5 Smith et al. (2022), Mamba Gu & Dao (2023), LRU Orvieto et al. (2023), and LinOSS Rusch & Rus (2024), following the training protocol and 70/15/15 splits of Rusch & Rus (2024); Walker et al. (2024) exactly. All runs were conducted on a single NVIDIA RTX 4090 GPU. Beyond the UEA benchmark, $\mathcal{S}^1$ achieves 97% accuracy on sequential MNIST, and 63% on sCIFAR-10 with a single layer — comparable to what S4 achieves with a single layer and without stacking SSMs in parallel Gu et al. (2021) — further demonstrating that analytical tractability does not come at an empirical cost.

	Worms	SCP1	SCP2	Ethanol	Heartbeat	Motor
Seq. length	17,984	896	1,152	1,751	405	3,000
#Classes	5	2	2	4	2	2
NRDE	77.2 ± 7.1	76.7 ± 5.6	48.1 ± 11.4	31.4 ± 4.5	73.9 ± 2.6	54.0 ± 7.8
NCDE	62.2 ± 2.2	80.0 ± 2.0	49.1 ± 6.2	22.0 ± 1.0	68.1 ± 5.8	51.6 ± 6.2
Log-NCDE	82.8 ± 2.7	82.1 ± 1.4	54.0 ± 6.2	35.9 ± 6.1	74.2 ± 2.0	57.2 ± 6.5
LRU	85.0 ± 6.2	82.5 ± 1.4	53.6 ± 10.6	23.8 ± 8.2	78.1 ± 7.6	55.7 ± 8.4
LinOSS-IMEX	80.0 ± 2.7	87.5 ± 4.0	58.9 ± 8.1	29.9 ± 1.0	75.5 ± 4.3	57.9 ± 5.2
LinOSS-IM	95.0 ± 4.4	87.8 ± 2.6	58.2 ± 6.9	29.9 ± 0.6	75.8 ± 3.7	60.0 ± 7.5
S5	83.9 ± 4.1	87.8 ± 1.5	47.5 ± 4.6	25.6 ± 3.5	73.9 ± 6.1	51.4 ± 5.7
Mamba	70.9 ± 15.8	83.8 ± 7.0	43.9 ± 8.5	26.6 ± 4.1	76.2 ± 8.3	47.3 ± 6.4
S6	85.0 ± 16.1	82.7 ± 1.4	49.9 ± 4.6	25.4 ± 6.8	76.5 ± 8.3	53.7 ± 6.2
$\mathcal{S}^1$	72.8 ± 5.4	83.8 ± 2.3	49.1 ± 2.2	33.9 ± 3.8	68.4 ± 3.6	50.5 ± 6.1

Table 1: Test accuracy (averaged over five training runs) on a selection of datasets from the UEA Multivariate Time Series Classification Archive. Dataset names are abbreviated as follows: EigenWorms (Worms), SelfRegulationSCP1 (SCP1), SelfRegulationSCP2 (SCP2), EthanolConcentration (Ethanol), Heartbeat, and MotorImagery (Motor).

5 Results and analysis

We can recast Eq.equation 12 using Eq.equation 5 as

$$O_c = \mathbf{W}_c \sum_{k=1}^{T} e^{i\,\mathrm{Arg}(\mathbf{y}_k)} = \sum_n W_{cn} \sum_{k=1}^{T} e^{iRe[\psi_{kn}]} = \sum_n W_{cn} r_n e^{i\phi_n} \tag{15}$$

where r_n is the resultant phasor amplitude after adding the phasors over time and ϕ_n is the mean phase angle (resultant direction). Finally, $\mathbf{W}$ reads out the phase amplitude and phase direction in order to classify the input data. Throughout the following analyses, interpretability results are computed exclusively on correctly classified test samples, ensuring that the phase coherence and modal decomposition reflect the model's learned decision mechanism rather than failure modes.

5.1 Case study 1: EEG data (SelfRegulationSCP1)

Case Study: SCP1

Fig. 2 traces the full computational pipeline of $\mathcal{S}^1$ on the SCP1 EEG dataset((correctly classified samples only). The modal decomposition (top left) shows that energy concentrates in modes $\ell \leq 5$, consistent with the low-frequency nature of the signals, with class-specific amplitude and phase patterns in the dominant modes. The modal power spectrum (Fig. 5, Appendix) confirms a $1/f$-like decay with class 0 carrying consistently higher power than class 1 in modes $\ell \leq 5$, with the largest separation at the lowest frequencies. The canonical nonlinearity $y/|y| = e^{i\psi_n}$ (top right) extracts the instantaneous oscillator phase at each node, producing qualitatively different spatiotemporal phase fields for the two classes. Temporally averaging these unit phasors yields the mean resultant $r_n e^{i\phi_n}$ per node (bottom): class 0 sustains higher phase coherence and concentrates its mean phase direction in a distinct angular sector relative to class 1. The readout $O_c = \sum_n W_{cn} r_n e^{i\phi_n}$ discriminates between classes by weighting nodes according to both their coherence magnitude and mean direction, providing a complete closed-form account of the classification decision.

5.2 Case study 2: sMNIST

Case Study: sMNIST

Fig. 3 traces the pipeline of $\mathcal{S}^1$ on sequential MNIST for digits 1 and 7. The modal decomposition shows class-specific spectral resonance: digit 7 drives modes $\ell = 5$–8 strongly while digit 1 concentrates energy in lower modes, reflecting how the learned coupling $\tilde{b}_\ell$ selectively excites different frequency modes for different digit structures. The canonical nonlinearity $y/|y| = e^{i\psi_n}$ converts

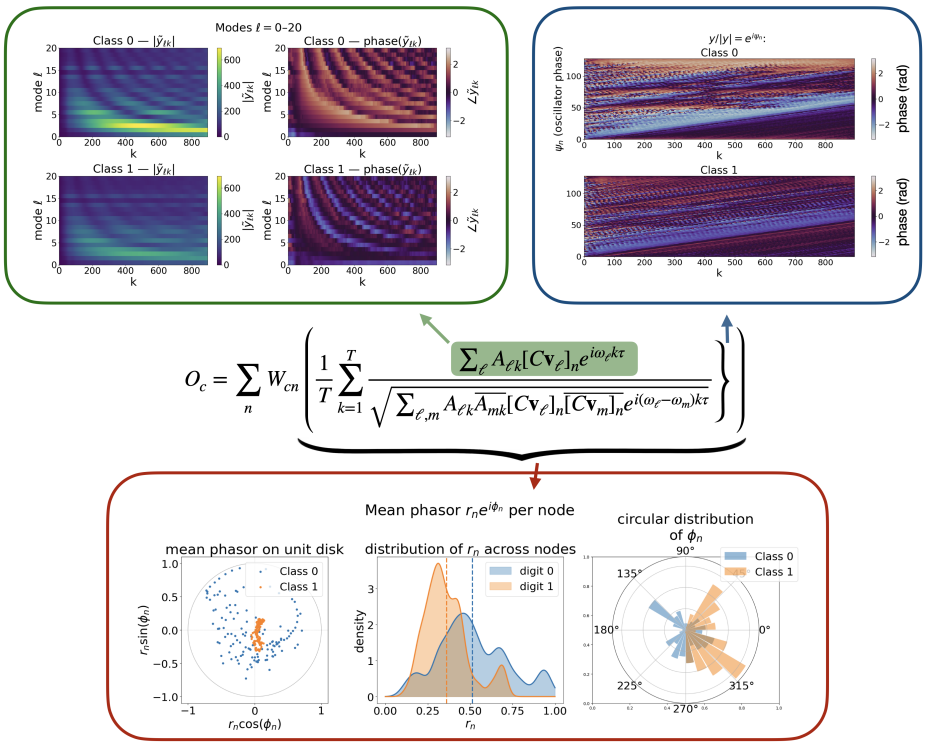


Figure 2: End-to-end interpretability of $\mathcal{S}^1$ on SCP1. **Top left:** Class-averaged modal amplitude $|\tilde{y}_{\ell k}|$ and phase $\angle \tilde{y}_{\ell k}$ for modes $\ell = 0$–20. Energy concentrates in low-frequency modes, with class-specific differences in both amplitude and phase. **Top right:** The canonical nonlinearity $y/|y| = e^{i\psi_n}$ recovers the oscillator phase ψ_{nk} at each node, revealing distinct spatiotemporal phase trajectories per class. **Bottom:** Temporal averaging yields the mean resultant phasor $r_n e^{i\phi_n}$ per node. Class 0 exhibits higher resultant phasor amplitude (KDE, centre) and a distinct preferred phase direction (rose diagram, right), both of which are read out by $\mathbf{W}$ via $O_c = \sum_n W_{cn} r_n e^{i\phi_n}$.

these modal differences into distinct spatiotemporal phase patterns. After temporal averaging, the mean phase directions ϕ_n concentrate in clearly separated angular sectors for the two classes, while coherence r_n remains similar — demonstrating that classification is driven by *phase direction* rather than coherence magnitude. The readout $\mathbf{W}$ implements $O_c = \sum_n W_{cn} r_n e^{i\phi_n}$, providing a complete closed-form account of the classification decision.

6 Modal Forcing-Gain Steering

The closed-form operator (q. equation 12 lets us go beyond the observational analyses of Section 5 and ask a causal question: do specific eigenmodes drive the classifier's decision? We intervene directly on identifiable modes while freezing the readout, and ask whether the prediction follows.

Intervention. Fix a switch index $k_\star$. For $k < k_\star$ the trajectory follows the original modal recursion; for $k \geq k_\star$ we replace it by

$$\tilde{y}^{\mathrm{st}}_{k+1,\ell} = \underbrace{\gamma_\ell}_{\text{prop. gain}} \underbrace{|\mu_\ell|}_{\text{decay}} \underbrace{e^{\mathrm{i}(\omega_\ell \tau + \delta\theta_\ell)}}_{\text{rotation}} \tilde{y}^{\mathrm{st}}_{k,\ell} + \underbrace{\eta_\ell}_{\text{forcing gain}} f_{k,\ell}, \tag{16}$$

where $\mu_\ell = e^{\lambda_\ell \tau}$ and $f_{k,\ell}$ is the modal forcing inferred from the unperturbed trajectory. Each control plays a distinct dynamical role: γ_ℓ amplifies or damps mode ℓ, $\delta\theta_\ell$ retunes its oscillation rate, and η_ℓ scales the inferred drive. Controls are unmasked only on the top-k modes of the discriminative score $s_{t,\ell} - s_{s,\ell}$ with $s_{c,\ell} := |(\boldsymbol{W}\boldsymbol{C}\boldsymbol{v}_\ell)_c|$, and are optimised against a hinge on the source–target

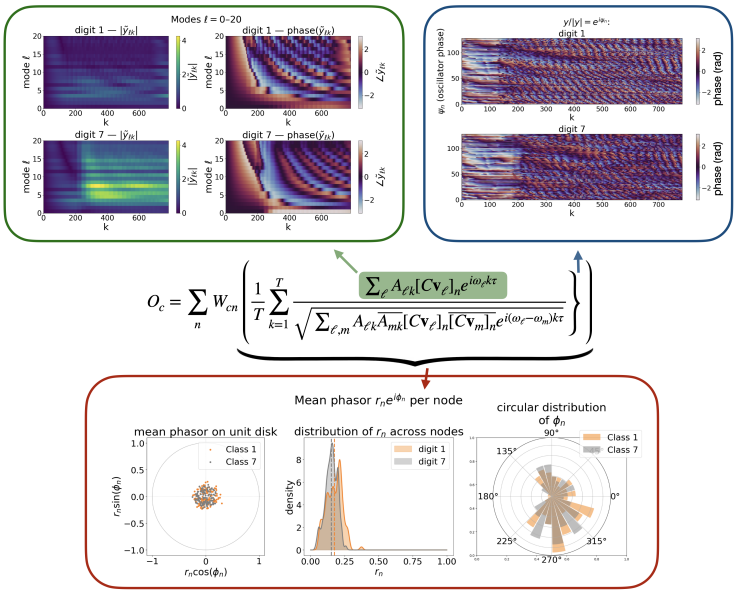


Figure 3: End-to-end interpretability of $\mathcal{S}^1$ on sMNIST for digits 1 and 7. **Top left:** Modal amplitude $|\tilde{y}_{\ell k}|$ and phase for modes $\ell = 0$–20. Digit 7 excites modes $\ell = 5$–8 more strongly while digit 1 concentrates energy in lower modes, reflecting class-specific spectral resonance. **Top right:** The nonlinearity $y/|y| = e^{i\psi_n}$ recovers the oscillator phase, producing distinct spatiotemporal patterns per class. **Bottom:** The mean resultant phasor $r_n e^{i\phi_n}$ shows that while per-node coherence is not very distinct across classes, the preferred phase directions are clearly separated — digit 1 around $315°$–$0°$ and digit 7 around $225°$–$270°$ — which $\mathbf{W}$ exploits via $O_c = \sum_n W_{cn} r_n e^{i\phi_n}$.

margin (Appendix B). Crucially, $\boldsymbol{W}$ and $\boldsymbol{C}$ are frozen, so any change in the predicted class is causally attributable to the modal perturbation alone.

Result. Figure 4 reports steering on SELFREGULATIONSCP1 for the $0 \to 1$ pair. Every source-class sample crosses $O_1 - O_0 = 0$ after $k_\star$, and the population mean of the margin moves from ≈ -5 to $\approx +5$ — a true $\arg\max$ flip, not a margin shift. Only five modes ($\ell \in \{1, 5, 22, 32, 37\}$) carry non-zero controls, and each perturbation satisfies $|\gamma_\ell - 1|, |\eta_\ell - 1|, |\delta\theta_\ell| 5 \times 10^{-3}$. A sparse, small intervention on readout-discriminative modes is therefore sufficient to causally redirect the classifier. The converged controls constitute a finite, mechanistic description — "mode ℓ had to be amplified by γ_ℓ, phase-shifted by $\delta\theta_\ell$, driven by η_ℓ" — which the analytic transparency of the $\mathcal{S}^1$ operator makes available by construction.

7 DISCUSSION

In this work, we have introduced $\mathcal{S}^1$, a novel SSM architecture that achieves competitive performance on sequence classification tasks while also being mathematically tractable. This combination of performance and mechanistic insight is possible because the nonlinear readout – which typically renders SSM systems intractable – is derived directly from the system's own dynamics, allowing an exact mathematical expression for the mapping from individual inputs to outputs in this architecture. This mathematical expression allows us to trace how the network processes inputs in real-world classification tasks, including EEG and sMNIST. This framework allows us to introduce a procedure for controlling the system's output – going beyond steering in a modern sequence problem for the first time.

7.1 LIMITATIONS

The current work focuses on sequence classification; we have not yet extended $\mathcal{S}^1$ to generative or regression settings. There is no fundamental barrier to doing this – the closed-form operator is not

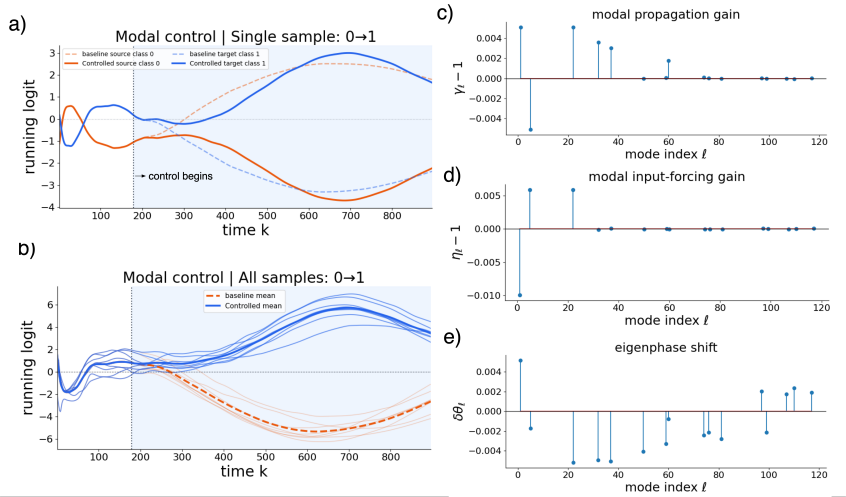


Figure 4: Modal forcing-gain control on SelfRegulationSCP1. Starting from $k_\star = 179$ (dotted line; shaded region), Eq. (16) is applied with controls $\{\gamma_\ell, \eta_\ell, \delta\theta_\ell\}$ optimised on the top-$k = 16$ discriminative modes; $\mathbf{W}$ and $\mathbf{C}$ are held fixed. **(a)** Running class logits for a representative sample. Baseline (dashed) settles into class 0; controlled (solid) crosses over within $\sim$150 steps and finishes firmly in class 1. **(b)** Running target$-$source margin per sample (thin) and class-mean (thick): all 8 samples flip to class 1, with the controlled mean reaching $\approx +6$ by sequence end (mean final margin $+3.47$, flip rate 1.00). **(c)** Propagation gain $\gamma_\ell - 1$: small bipolar adjustments ($\approx \pm 5 \times 10^{-3}$), amplifying modes $\ell \in \{1, 22, 32, 37, 60\}$ while damping $\ell = 5$. **(d)** Input-forcing gain $\eta_\ell - 1$: sparsely active, attenuating mode $\ell = 1$ ($\approx -10^{-2}$) and amplifying $\ell \in \{5, 22\}$. **(e)** Eigenphase shift $\delta\theta_\ell$: the most broadly active control, retuning low- and mid-index modes by a few milliradians per step. Controls of magnitude 10^{-2} on a sparse mode set suffice to flip every source sample, providing causal evidence that those modes carry the discriminative information identified in Section 5.

specific to classification – but validating this extension is left to future work. Performance on benchmarks such as Long Range Arena also remains to be established, and we expect that architectural refinements to the recurrent layer or the encoder-decoder structure may be needed to close the gap with state-of-the-art models at those scales. More broadly, $\mathcal{S}^1$ should be understood as an initial demonstration of the architecture, showing that a sequence model can achieve competitive performance while remaining fully interpretable. Future work may explore alternative connectivity structures, or hybrid architectures that preserve mathematical tractability while expanding expressivity.

7.2 Broader Impact

As sequence models are deployed in increasingly consequential settings, the absence of first-principles interpretability poses both a scientific and practical problem. $\mathcal{S}^1$ demonstrates that this need not be an inherent trade-off: mathematical tractability and strong empirical performance can be compatible components of sequence model architectures. The closed-form operator governing the end-to-end computation in this system opens new avenues for understanding how SSMs transform sequential inputs into decisions, with practical benefits for researchers seeking interpretable foundations and practitioners seeking to audit model behaviour. Beyond interpretability, the same mathematical framework enables a principled control framework in which undesired model behaviour can be corrected by intervening directly in the system's dynamics. This goes beyond standard methods of steering through fine-tuning or output filtering: by acting on identifiable eigenmodes with known roles in the system's dynamics, the control framework offers a transparent and mathematically grounded basis for safe deployment of future models that can be controlled online.

Code available at `https://anonymous.4open.science/r/S1-D9F3/`.

AI USE STATEMENT

In this work, we used generative AI tools for editing the manuscript. We have not used generative AI tools for developing the model or the mathematical analysis that the explainability claims are based on. We have used AI tools for formatting the code and the repository, but all the key functions and classes are written and verified manually. We have reviewed all AI-assisted work and rigorously checked all the suggested edits. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have shared the code along with the hyperparameters used for the benchmarks in the anonymous repo `https://anonymous.4open.science/r/S1-D9F3/`.

REFERENCES

Ameen Ali Ali, Itamar Zimerman, and Lior Wolf. The hidden attention of mamba models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1516–1534, 2025.

Anthony Bagnall, Hoang Anh Dau, Jason Levy, Germain Forestier, Che Hou, Germain Jehan, and Li Ye. The uea multivariate time series classification archive, 2018. *arXiv:1811.00075*, 2018.

Niels Birbaumer. Slow cortical potentials: Plasticity, operant control, and behavioral effects. *The Neuroscientist*, 5(2):74–78, 1999.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv:2312.00752*, 2023.

A. Gu, K. Goel, and C. Ré. Efficiently modeling long sequences with structured state spaces. *arXiv:2111.00396*, 2021.

A. Gu, K. Goel, A. Gupta, and C. Ré. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35, 2022.

David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Nicola Muca Cirone, Antonio Orvieto, Benjamin Walker, Cristopher Salvi, and Terry Lyons. Theoretical foundations of deep selective state-space models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

A. Orvieto, S. L. Smith, A. Gu, A. Fernando, C. Gulcehre, R. Pascanu, and S. De. Resurrecting recurrent neural networks for long sequences. In *International Conference on Machine Learning*. PMLR, 2023.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

T. K. Rusch and D. Rus. Oscillatory state-space models. *arXiv:2410.03943*, 2024.

J. T. H. Smith, A. Warrington, and S. W. Linderman. Simplified state space layers for sequence modeling. *arXiv:2208.04933*, 2022.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th annual meeting of the association for computational linguistics: system demonstrations*, pp. 37–42, 2019.

B. Walker, A. D. McLeod, T. Qin, Y. Cheng, H. Li, and T. Lyons. Log neural controlled differential equations: The lie brackets make a difference. *arXiv:2402.18512*, 2024.

Shida Wang and Beichen Xue. State-space models with layer-wise nonlinearity are universal approximators with exponential decaying memory. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A Appendix: Analysis

Modal Power Spectra

The time-averaged modal power spectra $|\tilde{y}_{\ell k}|^2$ for SCP1 and sequential MNIST, respectively, computed on correctly classified test samples are shown in Figs. 5 and 6.

For SCP1 (Fig. 5), the spectrum decays monotonically with mode index in a $1/f$-like fashion, consistent with the low-frequency nature of slow cortical potentials. Class 0 carries consistently higher power than class 1 across the dominant low-frequency modes $\ell \leq 5$, indicating that the two cortical states differ in the degree to which they drive the oscillator network, rather than in which modes they excite.

For sequential MNIST (Fig. 6), the spectral structure is qualitatively different: the power envelope is non-monotonic, with two distinct peaks around modes $\ell = 5$–15 and $\ell = 60$–80, reflecting the periodic spatial structure of rasterised digit images. Class 1 is sharply separated from classes 2 and 5 across almost the entire spectrum, with particularly large differences in the two peak regions. This selective suppression of modal power for digit 1 is consistent with the simple vertical stroke structure of that digit, which excites fewer frequency modes than the more spatially complex digits 2 and 5. Together, these spectra confirm that $\mathcal{S}^1$ learns input couplings $\tilde{b}_\ell$ that produce class-discriminative modal excitation patterns, providing direct spectral evidence for the resonance mechanism described in Section 2.

Fig. 7 shows the temporal evolution of the most discriminative modal amplitudes for sMNIST and SCP1, respectively. In both cases, the class separation emerges early in the sequence and persists, confirming that the input-driven modal excitation captured by $A_{\ell k}$ carries stable, class-specific information throughout the trial rather than only at specific timepoints.

B Appendix: Modal forcing-gain steering protocol

Inferred forcing. The modal recursion satisfies $\tilde{y}_{k+1,\ell} = \mu_\ell \tilde{y}_{k,\ell} + f_{k,\ell}$ with $\mu_\ell = e^{\lambda_\ell \tau}$. We infer the forcing directly from the unperturbed trajectory as the one-step residual $f_{k,\ell} := \tilde{y}_{k+1,\ell} - \mu_\ell \tilde{y}_{k,\ell}$, which isolates the dynamics from the input. The steered state is then $\boldsymbol{y}_k^{\text{st}} = \boldsymbol{V}\tilde{\boldsymbol{y}}_k^{\text{st}}$, with logits computed by the standard readout Eq. equation 12.

Bounded controls and mode selection. Each control is the bounded image of a free parameter,

$$\gamma_\ell = 1 + (\gamma_{\max} - 1) m_\ell \tanh \tilde{\gamma}_\ell, \quad \eta_\ell = 1 + (\eta_{\max} - 1) m_\ell \tanh \tilde{\eta}_\ell, \quad \delta\theta_\ell = \theta_{\max} m_\ell \tanh \tilde{\theta}_\ell, \tag{17}$$

so the steered eigenvalue $\mu_\ell^{\text{st}} = \gamma_\ell |\mu_\ell| e^{\mathrm{i}(\omega_\ell \tau + \delta\theta_\ell)}$ lies in an explicit annulus around μ_ℓ. The mask $m_\ell \in \{0, 1\}$ unmasks the top-k entries of $r_\ell = \max(0, s_{t,\ell} - s_{s,\ell})$ with $s_{c,\ell} = |(\boldsymbol{W}\boldsymbol{C}\boldsymbol{v}_\ell)_c|$.

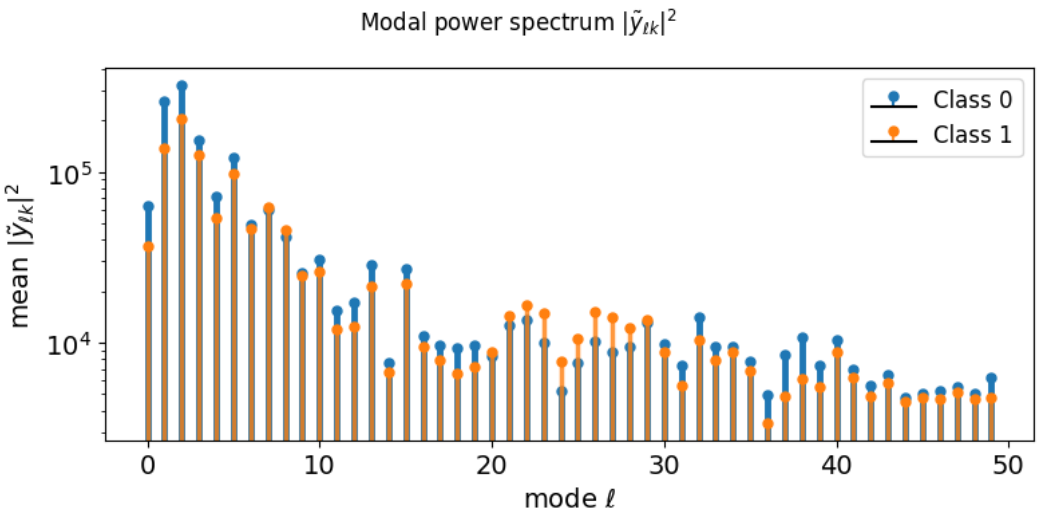


Figure 5: Modal power spectrum $|\tilde{y}_{\ell k}|^2$ for the first 50 modes on SCP1. Both classes exhibit a $1/f$-like decay with energy concentrated in modes $\ell \leq 5$. Class 0 consistently carries higher modal power than class 1 across the dominant low-frequency modes, with the separation most pronounced at $\ell = 0$–3.

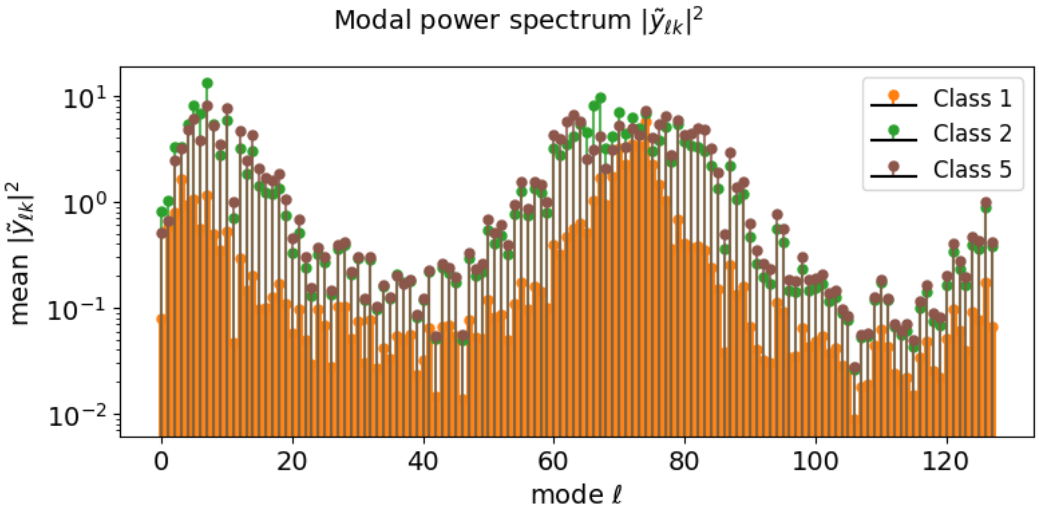


Figure 6: Modal power spectrum $|\tilde{y}_{\ell k}|^2$ for three classes of sequential MNIST. Class 1 carries substantially lower power than classes 2 and 5 across nearly all modes, with the largest separation occurring in modes $\ell = 0$–15 and $\ell = 55$–80. The two-peaked spectral envelope reflects the frequency content of the rasterised pixel sequences, with class-specific power differences indicating selective modal excitation by each digit's spatial structure.

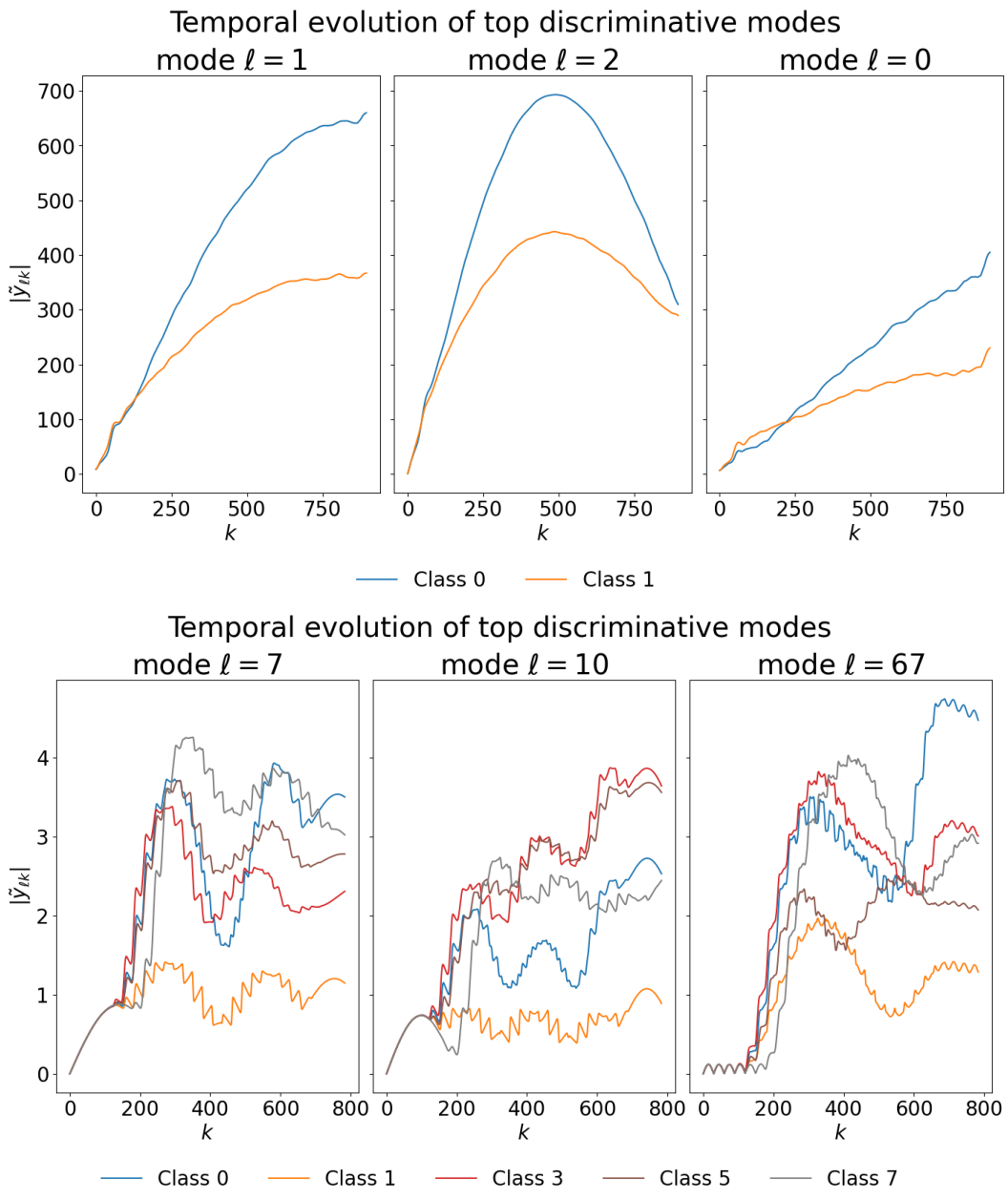


Figure 7: Temporal evolution of the most class-discriminative modal amplitudes $|\tilde{y}_{\ell k}|$. **Top (SCP1):** Modes $\ell = 0$–2 show clear sustained separation between classes, with class 0 accumulating consistently higher amplitude across the full trial. **Bottom (sMNIST):** Modes $\ell = 7$, 10, and 67 exhibit persistent class-specific trajectories, with class 1 consistently suppressed relative to all other classes.

Objective. The free parameters minimise

$$\mathcal{L} = \mathbb{E}_b\big[\max(0, \mathtt{m} - (O_t^{\mathrm{st}} - O_s^{\mathrm{st}}))\big] + \lambda_\gamma \|\tilde{\gamma}\|^2 + \lambda_\eta \|\tilde{\eta}\|^2 + \lambda_\theta \|\tilde{\theta}\|^2, \tag{18}$$

with the expectation over correctly-classified source samples and $\boldsymbol{W}, \boldsymbol{C}$ held fixed. For $|\mathcal{C}| \geq 3$ we replace O_s^{st} by $\max_{c \neq t} O_c^{\mathrm{st}}$ to demand a true $\arg\max$ flip. We optimise