SPECOSS: SPECTRALLY GENERATED OSCILLATORY STATE-SPACE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A time-dependent signal can be represented in different bases, and the choice of basis determines how its temporal variation is decomposed across characteristic frequencies. Oscillatory state-space models (SSMs) provide a natural way to realise such frequency-based representations: their hidden state consists of second-order modes whose natural frequencies determine which temporal patterns are represented. In the common diagonal formulation, these squared frequencies appear directly as the diagonal entries governing the state dynamics. Existing oscillatory SSMs typically initialise the frequencies randomly and adapt them during training. We introduce the Spectrally Generated Oscillatory State-Space model (SpecOSS), which instead derives their initial values from the spectral structure of orthogonal bases. We obtain these frequencies in two ways. First, Sturm–Liouville (SL) theory associates orthogonal-function bases with differential operators whose eigenvalues give squared natural frequencies (Szegő, 1939). Second, online function approximation represents the past of a signal in an orthogonal basis whose coefficients evolve through a linear operator (Lee, 1932; Wiener, 1958); the spectrum of this operator provides a second construction of the oscillator frequencies. The resulting spectral initialisations may remain fixed or be adapted during training. Because an observed signal can superimpose dynamics from several underlying components, we represent several basis-specific spectra jointly in one model and learn how to weight their readouts, allowing distinct spectral components to contribute separately without increasing the cost of the recurrence. Furthermore, we derive error bounds for the finite spectral discretisation and analyse the expressivity of fixed spectral initialisations. Across five time-dependent signal classification benchmarks, heart-rate regression from wearable sensors at sequence length fifty thousand, real-world weather forecasting, and WikiText-103 and Pile language modelling, our spectral initialisations outperform randomly initialised oscillatory models and several first-order baselines.

1 INTRODUCTION

A measured time series is a sequence of observations whose temporal variation may result from several physical components at different characteristic frequencies. For classification, regression, or forecasting, a model therefore needs a representation that can model these physical components with their different characteristic frequencies. Oscillatory state-space models (SSMs) (Rusch & Mishra, 2021a;b; Rusch & Rus, 2025) provide such a representation using m harmonic modes. Given an input u, their state x evolves according to the second-order ODE $x'' = -Ax + Bu$, with a learned input map B and a learned linear output map. The state matrix $A = \mathrm{diag}(A_0, \ldots, A_{m-1})$ is non-negative and diagonal, with each entry $A_k = \omega_k^2$ parametrising the natural frequency of one mode. We refer to these frequencies collectively as the oscillator spectrum; they determine the oscillatory patterns which are weighted by the state representation. Existing oscillatory SSMs initialise these parameters at random and adapt them by gradient descent.

Instead of choosing the oscillator spectrum randomly, we determine it from structured orthogonal bases. This is useful because a basis specifies which modes are available to represent temporal variation and how those modes are distributed in frequency, giving the oscillator frequencies a principled structure before any training takes place. We use two constructions for this purpose; the first route comes from Sturm–Liouville (SL) theory (Zettl, 2005). For bases such as Hermite, Legendre,

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.

Chebyshev, Laguerre, and Fourier, the basis functions $\{p_k\}$ are eigenfunctions of a second-order differential operator $\mathcal{L}$, $\mathcal{L}p_k = \lambda_k p_k$. The eigenvalues λ_k define squared natural frequencies, where k is the mode index, so each basis, through its associated Sturm–Liouville operator, induces an oscillator spectrum. Different bases distribute these frequencies differently: for Hermite (Schrödinger, 1926) and Laguerre, the gaps between neighbouring frequencies decrease as k grows, whereas for Legendre (Legendre, 1785), Chebyshev (Chebyshev, 1854), and Fourier (Fourier, 1822), the gaps are approximately of equal size. (Table 1). Thus, each basis provides a different allocation of oscillatory modes across the frequency range. The second route is online function approximation (Lee, 1932; Wiener, 1958), which represents the past of an input signal f by its coefficients x in the chosen basis. The coefficients obey a linear law $x' = -A_{\mathrm{H}}x + B_{\mathrm{H}}f$ whose generator A_{H} and input vector B_{H} are fixed by the basis, and differentiating it once more turns it to second-order system of the same form as the oscillator layer. The spectrum follows from the eigenvalues of A_{H}, exactly for the bases whose generator is skew-symmetric and through a surrogate for the others (Section 4.1).

SpecOSS uses either the online function-approximation construction or the SL construction to initialise the state matrix A before training; A can subsequently remain fixed or be adapted. We hypothesise, and show empirically, that different components of an observed signal may be represented more effectively by different bases. We therefore assign several basis-specific spectra to separate state subspaces within a block-diagonal A and learn a gate over their readouts, without increasing the cost of the linear state update.

We then analyse the representational capacity of the resulting oscillator system. A continuous range of frequencies, weighted by a spectral density, defines an input-response kernel that specifies how past input values contribute to the current state; we refer to this as the memory kernel (Boyd & Chua, 1985). A finite set of m oscillators approximates this continuous kernel by numerical quadrature (Davis & Rabinowitz, 1984), and the error is controlled by the largest gap between neighbouring frequencies. Nearly evenly spaced spectra therefore converge at $O(m^{-1})$. Hermite and Laguerre converge at $O(m^{-1/2})$, but place progressively more modes toward the high-frequency end of the band, giving finer resolution of faster temporal variation. Having both types of spectra offers a trade-off between efficient frequency coverage and finer resolution where more modes are needed.

Our contributions are: (i) We introduce SpecOSS, an oscillatory SSM whose state dynamics are initialised from the spectral structure of a chosen basis rather than from random frequencies. This construction is performed once, after which the state matrix can either remain fixed or be optimised jointly with the rest of the model at the same computational cost. (ii) We derive two ways to obtain these oscillator frequencies, from Sturm–Liouville spectra and from the second-order form of online function approximation, and characterise when the latter is exact and when it provides a surrogate. (iii) We introduce a multi-basis representation in which different basis-specific spectra occupy separate state subspaces and their readouts are combined by a low-rank, input-dependent gate, allowing several frequency distributions to be represented within the same state update. (iv) We analyse the approximation error and expressiveness of the resulting oscillator representation. The error is controlled by the largest gap between neighbouring frequencies, and the representation converges to the kernel class induced by Lipschitz spectral densities as the number of oscillators increases, even when the state matrix is kept fixed. (v) We evaluate SpecOSS on nine benchmarks in total, spanning signal classification, heart-rate regression, real-world weather forecasting, and language modeling on WikiText-103 and the Pile, where basis-derived initialisations outperform randomly initialised oscillatory models and several first-order baselines, with ablations isolating the effects of the two spectral constructions and the multi-basis design.

2 Related work

Structured state-space models and their initialisation. S4 (Gu et al., 2022a) showed that the HiPPO-LegS matrix (Gu et al., 2020) is key to long-range memory in a linear SSM; S4D (Gu et al., 2022b) and DSS (Gupta et al., 2022) retain only its diagonal, and Gu et al. (2023) interpret the resulting S4D-Lin and S4D-Inv initialisations as generalised orthogonal-basis projections. S5 (Smith et al., 2023), LRU (Orvieto et al., 2023), and Mamba (Gu & Dao, 2024; Dao & Gu, 2024) learn diagonal first-order transitions, with Mamba making them input-dependent. These models are first-order, so their eigenvalues encode both decay rates and oscillation frequencies. We instead start from the second-order LinOSS (Rusch & Rus, 2025), whose diagonal contains squared natural frequencies, and generate these frequencies rather than learn them from scratch.

Oscillatory recurrences and their theory. coRNN and UnICORNN showed that second-order oscillatory dynamics avoid vanishing and exploding gradients; LinOSS made the oscillators linear and parallelisable, while D-LinOSS (Boyer et al., 2025) adds learned damping. Legendre Memory Units (Voelker et al., 2019), which maintain an online Legendre expansion of the past of a signal, are the closest precedent with a fixed spectrum. Universality of neural oscillators was proved by Lanthaler et al. (2023) through the time-varying sine transform and turned into quantitative bounds by Huang et al. (2025), in both cases for oscillators with free frequencies. We instead construct the oscillator frequencies from structured spectral priors and provide an error and expressivity analysis for the resulting oscillators. Approximation theory of first-order linear recurrences (Li et al., 2022; Orvieto et al., 2024; Wang & Xue, 2023; Muca Cirone et al., 2024) and multi-head SSMs with non-homogeneous initialisations (Fathullah et al., 2023) are discussed in Appendix E.

3 Background

Oscillatory state-space model. The model maps a length-T input sequence $\mathbf{I} = (i_1, \dots, i_T) \in \mathbb{R}^{T \times d_{\text{in}}}$ through L residual blocks of width H to a class label, a regression target, or a forecast of the last T_{fut} steps. An oscillatory SSM is built on a forced linear second-order ODE, $x''(t) = -A\,x(t) + B\,u(t)$, and a linear readout $y(t) = C\,x(t) + D\,u(t)$, with hidden state $x(t) \in \mathbb{R}^m$, layer input $u(t) \in \mathbb{R}^H$, and output $y(t) \in \mathbb{R}^H$. From here on we drop the time argument and write x, y, and u for these signals, and $x'' = -Ax + Bu$, $y = Cx + Du$ for the system, restoring (t) only where the value at one instant is meant. The weights are $A \in \mathbb{R}^{m \times m}$, $B \in \mathbb{R}^{m \times H}$, $C \in \mathbb{R}^{H \times m}$, and $D \in \mathbb{R}^{H \times H}$, with A nonnegative. The direct term Du bypasses the state and therefore does not contribute to the memory of the layer. We hence omit it from the subsequent exposition and write the readout as $y = Cx$. With $z = x'$, this is the first-order system $z' = -Ax + Bu$, $x' = z$. For $m = H = 1$ and $B = 0$, it reduces to $x'' + Ax = 0$, simple harmonic motion with natural frequency $\omega = \sqrt{A}$. A diagonal $A = \operatorname{diag}(A_0, \dots, A_{m-1})$ therefore gives m uncoupled forced oscillators, with mode k, $k = 0, \dots, m-1$, running at $\omega_k = \sqrt{A_k}$. A general A instead couples the modes; when a construction yields such a coupled matrix, we use a diagonal surrogate of it. Discretising the per-mode state $s_n = (z_n, x_n)^\top$ with a learned step $\Delta t_k \in (0, 1)$ gives a linear recurrence $s_n = M_k s_{n-1} + F_n$ under either the implicit (IM) or the implicit–explicit (IMEX) scheme (Ascher et al., 1997; Hairer et al., 2006), IM being dissipative and IMEX energy conserving inside the Courant–Friedrichs–Lewy (CFL) bound (Courant et al., 1928).

Two classical sources of spectra. The spectrum of A is its list of squared natural frequencies, and two theories produce such spectra from first principles. (i) Sturm–Liouville theory associates each basis of orthogonal polynomials $\{p_k\}$ with a self-adjoint second-order operator for which they are eigenfunctions, $\mathcal{L}p_k = \lambda_k p_k$. Its eigenvalues λ_k are real, nonnegative, and increase without bound, and correspond to the squared natural frequencies of the dynamics described by $\mathcal{L}$. (ii) Online function approximation keeps a state equal to the coefficients of the input's past in an orthonormal basis under a measure that slides or scales with time. Differentiating this definition gives a linear memory law $x' = -A_{\text{H}}x + B_{\text{H}}f$, whose generator A_{H} is fixed by the basis and the measure; differentiating once more yields a second-order system in the same coefficients, compatible with an oscillatory SSM. The squared frequencies are then obtained from the spectrum of A_{H}, directly when the resulting second-order operator is diagonalisable and through a diagonal surrogate when it is not.

4 The Spectrally Generated Oscillatory State-Space Model

4.1 Generation routes

We keep the notation of Section 3. A layer is the oscillator system $x'' = -Ax + Bu$ with $A = \operatorname{diag}(A_0, \dots, A_{m-1})$, where each $A_k = \omega_k^2$ is the squared natural frequency of mode k; the readout is generalised below to a gate over groups of modes drawn from different bases. SpecOSS constructs these frequencies from two sources rather than treating them as free parameters (Figure 1): route 1 uses the Sturm–Liouville spectrum λ_k^{SL} of a basis, while route 2 derives λ_k^{FA} from the spectrum of its function-approximation operator. In Figure 1, $\mathcal{E}$ denotes the input encoder that maps the raw sequence to the model width H, and $\mathcal{R}$ the task-specific readout that maps the final representation to the desired target. Since different bases induce different distributions of frequencies, we also

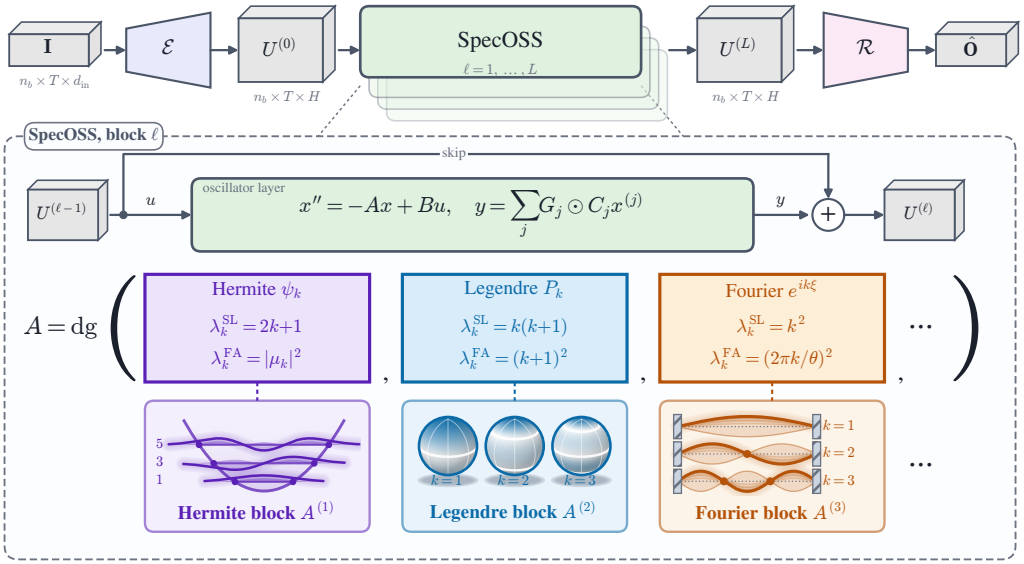


Figure 1: SpecOSS. The model stacks residual oscillator blocks whose diagonal state matrix A is assembled from basis-specific spectra generated by the Sturm–Liouville or function-approximation routes. The corresponding basis readouts are combined by a learned gate.

combine several such spectra within one A and learn how to weight their corresponding sub-states. All spectra are generated once, offline, before training, after which A may either remain fixed or be fine-tuned.

Route 1: Sturm–Liouville spectra. Each orthogonal basis $b \in \{\text{Hermite, Legendre, Chebyshev, Laguerre, Fourier}\}$ consists of functions $p_k(\xi)$ that satisfy a second-order eigenproblem $\mathcal{L}p_k = \lambda_k p_k$, where $\mathcal{L}$ is the differential operator associated with the basis and ξ is its spatial coordinate (Table 1). The eigenvalues λ_k are real and nonnegative (Olver, 2010), and correspond to squared natural frequencies: the k-th mode oscillates at $\omega_k = \sqrt{\lambda_k}$ (Appendix C.1). Route 1 therefore sets $\lambda_k^{\mathrm{SL}} := \lambda_k$ and $A = \mathrm{diag}(\lambda_0^{\mathrm{SL}}, \ldots, \lambda_{m-1}^{\mathrm{SL}})$, while B and C are learned. The bases differ mainly in how they distribute the frequencies ω_k. Hermite has $\lambda_k = 2k+1$, so $\omega_k = \sqrt{2k+1}$ and neighbouring frequencies become progressively closer as k increases. Legendre has $\lambda_k = k(k+1)$, giving nearly uniform spacing in ω_k, while Chebyshev and Fourier have $\lambda_k = k^2$ and hence exactly uniform frequencies $\omega_k = k$. Laguerre has $\lambda_k = k$, so $\omega_k = \sqrt{k}$, again giving denser spacing at higher frequencies. These different spectra therefore provide different distributions of temporal scales to the oscillator layer.

Route 2: function approximation. Online function approximation keeps a state equal to the coefficients of the input's past in an orthonormal basis $\{g_k^{(t)}\}$ under a measure that slides or scales with time, $x_k(t) = \langle u_{\le t}, g_k^{(t)} \rangle$. Differentiating this definition gives the first-order memory law $x' = -A_{\mathrm{H}} x + B_{\mathrm{H}} f$, where the generator A_{H} and input vector B_{H} are fixed by the chosen basis and measure. This is the form used by first-order memory SSMs; here we ask how the same state can be represented by a second-order oscillator. Differentiating once more gives

$$x'' = A_{\mathrm{H}}^2 x - A_{\mathrm{H}} B_{\mathrm{H}} f + B_{\mathrm{H}} f', \tag{1}$$

so matching the state term to $x'' = -Ax + Bu$ suggests $-A = A_{\mathrm{H}}^2$ and $B = -A_{\mathrm{H}} B_{\mathrm{H}}$. Two obstacles prevent this identification from always being exact. First, an oscillator requires $A \succeq 0$, hence $-A_{\mathrm{H}}^2 \succeq 0$. This holds when the eigenvalues of A_{H} are purely imaginary, as for skew-symmetric generators (the translating Hermite window and Fourier), because squaring an imaginary eigenvalue gives a negative real value. For a real spectrum such as Legendre, however, squaring gives positive growth rather than oscillation; and for defective generators such as Laguerre and Chebyshev, there is no eigenbasis in which to diagonalise the dynamics exactly. Second, Eq. (1) contains the additional forcing term $B_{\mathrm{H}} f'$, which depends on the derivative of the input and is unavailable to a standard oscillator layer.

Table 1: Squared frequencies generated by the two routes. Route 1 uses the Sturm–Liouville eigenvalues λ_k^{SL}, while route 2 uses $|\mu_k|^2$ from the function-approximation generator A_{H}, or $\varsigma_k(A_{\mathrm{H}})^2$ when A_{H} is defective. Indices start at $k = 0$, and θ denotes the Fourier-window length; derivations are given in Appendix C.

basis	Sturm–Liouville problem	route 1: λ_k^{SL}	route 2: λ_k^{FA}	route 2 generator
Hermite ψ_k	$(-\partial_\xi^2 + \xi^2)\psi_k = \lambda_k \psi_k$	$2k{+}1$	$\lvert\mu_k\rvert^2$	skew-symmetric, exact
Fourier $e^{ik\xi}$	$-\partial_\xi^2 e^{ik\xi} = \lambda_k e^{ik\xi}$	k^2	$(2\pi k/\theta)^2$	skew-symmetric, exact
Legendre P_k	$-\big[(1-\xi^2)P_k'\big]' = \lambda_k P_k$	$k(k{+}1)$	$(k{+}1)^2$	real spectrum
Chebyshev T_k	$-\sqrt{1-\xi^2}\big[\sqrt{1-\xi^2}\,T_k'\big]' = \lambda_k T_k$	k^2	$\varsigma_k(A_{\mathrm{H}})^2$	nilpotent, surrogate
Laguerre L_k	$-e^{\xi}\big[\xi e^{-\xi} L_k'\big]' = \lambda_k L_k$	k	$\varsigma_k(A_{\mathrm{H}})^2$	Jordan block, surrogate

We therefore construct a diagonal oscillator surrogate from the spectral scale of the generator:

$$\lambda_k^{\mathrm{FA}} := |\mu_k|^2, \qquad A = \mathrm{diag}(\lambda_0^{\mathrm{FA}}, \ldots, \lambda_{m-1}^{\mathrm{FA}}), \tag{2}$$

using the singular values $\varsigma_k(A_{\mathrm{H}})$ in place of $|\mu_k|$ when A_{H} is defective. This guarantees nonnegative squared frequencies while retaining the characteristic scales of A_{H}. The forcing in (1) cannot in general be transferred exactly to this diagonal representation: the map $-A_{\mathrm{H}}B_{\mathrm{H}}$ is expressed in the eigenvector coordinates of A_{H}, and the term $B_{\mathrm{H}}f'$ is unavailable to the oscillator. We therefore learn B and C, as in route 1, allowing them to adapt to the diagonal representation, similarly to the diagonal reduction used in S4D (Gu et al., 2022b). For skew-symmetric Hermite and Fourier generators, (2) exactly recovers the squared frequencies of the second-order state dynamics. For Legendre, it maps the real eigenvalue magnitudes to oscillator frequencies, preserving their characteristic scales but not the original dynamics. For the defective Laguerre and Chebyshev generators, the singular-value construction is a surrogate (Proposition 1, Appendix C.2). The two routes also need not produce the same spectrum: for Legendre, $\lambda_k^{\mathrm{SL}} = k(k{+}1) = 0, 2, 6, \ldots$ whereas $\lambda_k^{\mathrm{FA}} = (k{+}1)^2 = 1, 4, 9, \ldots$, with the same asymptotic growth, while for Fourier they coincide.

Remark 1 (Route 2 is a second-order HiPPO). Route 2 starts from the HiPPO memory operator (Gu et al., 2020): it uses the same first-order dynamics for the projection coefficients, differentiates them once more to obtain a second-order system, and retains only the spectral scales λ_k^{FA} needed by the oscillator model. For Hermite and Fourier, this reduction is exact. For the other bases, it preserves the characteristic frequencies of the HiPPO operator but not its exact dynamics, similar to the diagonal approximation of S4 by S4D.

Combining spectra from multiple bases. Because a block-diagonal matrix with diagonal blocks is still diagonal, several basis-specific spectra can be included in a single state matrix A without changing the recurrence. Choosing J bases $b_1, \ldots, b_J$ with block sizes $m_1 + \cdots + m_J = m$, and denoting by $\lambda_{b_j,k}$ the k-th squared frequency of basis b_j,

$$A = \mathrm{diag}\big(A^{(1)}, \ldots, A^{(J)}\big), \qquad A^{(j)} = \rho_j \ \mathrm{diag}\big(\lambda_{b_j,0}, \ldots, \lambda_{b_j,m_j-1}\big) \in \mathbb{R}^{m_j \times m_j}, \tag{3}$$

with no additional computational cost in the parallel scan. Here $\lambda_{b_j,k}$ can come from either route, $\lambda_{b_j,k}^{\mathrm{SL}}$ or $\lambda_{b_j,k}^{\mathrm{FA}}$. We set $\rho_j = 1/\max_k \lambda_{b_j,k}$ so that all blocks cover a comparable frequency range while retaining their different frequency spacings, e.g. Hermite with $\omega_k \propto \sqrt{k}$ and Fourier or Legendre with $\omega_k \propto k$ (Figure 5). The state is partitioned as $x = [x^{(1)}; \ldots; x^{(J)}]$, with one sub-state per basis. Each sub-state produces $y^{(j)} = C_j x^{(j)}$, and a learned gate combines these outputs as $y = \sum_j G_j \odot y^{(j)}$, where $G = \mathrm{softmax}_j(W_g[y^{(1)}; \ldots; y^{(J)}])$. This allows the model to weight different bases differently across channels and time steps. Simpler alternatives, including a shared readout and a rank-r factorisation of W_g with $2rJH$ instead of $(JH)^2$ parameters, are ablated in Appendix H.1. Although B and C are kept real above, both extend to the complex case, giving the readout $y = \sum_j G_j \odot \mathrm{Re}(C_j x^{(j)})$ in place of $y = \sum_j G_j \odot C_j x^{(j)}$; we ablate that extension in Appendix F.4 and report its sensitivity in Appendix J.1.

The oscillator layer is embedded in the same residual block for all spectral constructions, with $u^\ell = \mathrm{SpecOSS}(u^{\ell-1})$ for $\ell = 1, \ldots, L$: each block normalises $u^{\ell-1}$, applies the oscillator layer, and adds the result back to $u^{\ell-1}$ through a GELU and a gated linear unit (Figure 1 and Appendix B.5). Thus, changing the spectral construction changes only the oscillator dynamics, while the surrounding architecture and block width H remain fixed. Multiple bases can be combined either

within a single layer using the gated multi-basis construction above, or across depth by assigning different bases to successive blocks. The motivation is that different bases provide different distributions of frequencies and therefore cover different temporal scales; mixing them lets the model exploit these complementary representations, while the gate learns how much each basis should contribute. The across-depth construction introduces no additional parameter overhead.

Example 1 (A union of bases, a single basis, and a first-order core)**.** This example ablates two design choices: whether combining multiple bases improves class separation over a single basis, and whether the second-order oscillator representation improves over a first-order state-space core of the same size. All second-order models are initialised with λ^{SL}. Figure 2 compares three representations at the same state size, readout, and classifier: a union of bases, a single Hermite basis, and a first-order model with real poles. Row (i) shows predictions for a healthy twelve-lead ECG recording (Wagner et al., 2020), row (ii) shows the modes represented by each model, and row (iii) projects the learned representations into two dimensions to visualise class separation. The union of bases achieves 78.5% accuracy, compared with 75% for the single Hermite basis and 68% for the first-order model. The corresponding comparison using λ^{FA} is given in Appendix H.5.

4.2 Theory

We now ask how the chosen frequencies affect what an oscillatory layer can represent. For zero initial conditions, mode k stores a windowed sine transform of the input, $x_k(t) = \omega_k^{-1} \int_0^t \sin(\omega_k(t-\tau))\, b_k^\top u(\tau)\, d\tau$ (Lemma 1, Appendix D.1), so the memory of the layer is determined by the functions $\sin(\omega_k t)/\omega_k$ associated with its frequencies. To compare a finite set of m frequencies with a continuous range, let a spectral density ν weight all frequencies in $[\omega_{\min}, \omega_{\max}]$. This defines the continuous memory kernel

$$\mathcal{K}_\nu(t) = \int_{\omega_{\min}}^{\omega_{\max}} \nu(\omega)\, \frac{\sin(\omega t)}{\omega}\, d\omega, \qquad \mathcal{K}_\nu^m(t) = \sum_{k=0}^{m-1} \nu(\omega_k)\, |\mathcal{I}_k|\, \frac{\sin(\omega_k t)}{\omega_k}, \tag{4}$$

where $\mathcal{K}_\nu^m$ is the finite-frequency approximation obtained by assigning each ω_k a cell $\mathcal{I}_k$. The approximation quality is governed by the largest cell width, $h_m = \max_k |\mathcal{I}_k|$. The results below quantify both the resulting error and the expressivity obtained as the spectrum is refined.

Assumptions. **(A1)** (frequency band) All frequencies lie in $[\omega_{\min}, \omega_{\max}]$, with $\omega_{\min} = \sqrt{a_{\min}} > 0$ and $\omega_{\max} = 1$; this range is enforced by the model's frequency normalisation and is therefore a property of the construction, not of the input. **(A2)** (regular spectrum) The spectral density ν is Lipschitz continuous on the band with constant L_ν and bounded by $\bar{\nu} = \sup |\nu|$, so nearby frequencies receive similar weights and the continuous spectrum can be approximated by finitely many frequencies. **(A3)** (frequency discretisation) We use m distinct frequencies $\omega_0, \ldots, \omega_{m-1}$ and partition the band into cells $\mathcal{I}_k$ containing them, for example using midpoints between neighbouring frequencies; the largest cell width $h_m = \max_k |\mathcal{I}_k|$ measures the coarseness of this discretisation. The horizon T is finite throughout, and the results below concern the continuous-time kernel before time discretisation and training; implementation details are deferred to the appendix.

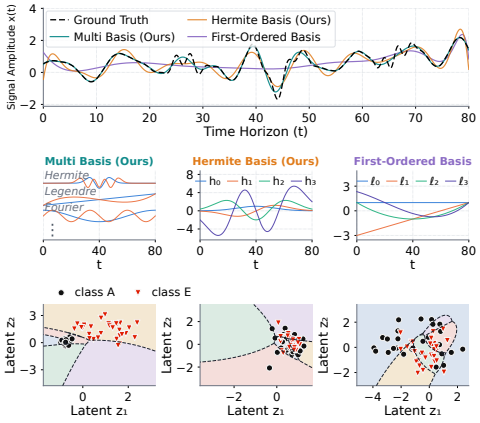


Figure 2: Comparison of state representations for five-class ECG classification. Rows show the predicted signal, modes, and two-dimensional class projection. Accuracy: 78.5 / 75 / 68%.

Theorem 1 (Spectral discretisation error; proof in Appendix D.2)**.** *Under (A1)–(A3), for every finite horizon $T > 0$,*

$$\sup_{t\in[0,T]} \left|\mathcal{K}_\nu(t) - \mathcal{K}_\nu^m(t)\right| \;\le\; (\omega_{\max} - \omega_{\min})\, h_m \left[L_\nu \min\{T, \tfrac{1}{\omega_{\min}}\} + \bar{\nu}\left(\tfrac{T}{\omega_{\min}} + \tfrac{1}{\omega_{\min}^2}\right)\right] = O(h_m). \tag{5}$$

For the normalised SL spectra of Table 1, $h_m = O(m^{-1/2})$ for Hermite and Laguerre, and $h_m = O(m^{-1})$ for Fourier, Legendre, and Chebyshev. The kernel approximation error therefore converges at the same respective rates (Corollary 1).

Interpretation. The approximation error is controlled by the largest spacing between neighbouring frequencies. The constants in the bound grow as $\omega_{\min} \to 0$, which motivates the lower frequency floor $a_{\min} = \omega_{\min}^2$, while the normalisation places all spectra on the same band. The same floor also bounds the forced response of every mode over any finite horizon, independently of m (Proposition 2). Bases with $\omega_k \propto \sqrt{k}$ therefore converge at rate $O(m^{-1/2})$, whereas the nearly uniformly spaced bases achieve $O(m^{-1})$. Combining several bases can further reduce the largest frequency gap at a fixed total number of modes.

Theorem 2 (Spectral expressivity; proof in Appendix D.3)**.** *Under (A1)–(A2), let $\mathfrak{K}$ be the class of kernels $\mathcal{K}_\nu$ induced by Lipschitz densities ν on the frequency band. For a spectrum satisfying (A3), define $\mathfrak{K}_m = \operatorname{span}\{\sin(\omega_k t)/\omega_k : k = 0, \dots, m-1\}$, the class of kernels representable by m oscillators. If the frequency discretisation is refined so that $h_m \to 0$ as $m \to \infty$, as for every normalised SL spectrum by Corollary 1, then every $\mathcal{K}_\nu \in \mathfrak{K}$ can be approximated uniformly on $[0, T]$ by elements of $\mathfrak{K}_m$, with the rate of Theorem 1. Hence $\mathfrak{K} \subseteq \overline{\bigcup_{m \geq 1} \mathfrak{K}_m}$ in the uniform on $[0, T]$.*

Interpretation. A frozen generated spectrum remains expressive as m increases: each SL basis covers the frequency band increasingly densely, allowing any kernel induced by a Lipschitz spectral density to be approximated. The choice of basis determines only the convergence rate and the distribution of frequencies across the band. This result concerns the memory kernel of the continuous-time layer, not arbitrary sequence maps; the learned Δt_k, B, C, and time discretisation introduce additional realisation error, treated separately in Appendix D.4.

Informal Theorem 3 (Frozen generated spectra). *For a causal target operator on a finite horizon, a layer with m fixed harmonic-type oscillator frequencies retains the free-frequency approximation-rate order $O((\log m)^2/m)$, up to the error of the nonlinear readout. Small departures from the harmonic grid incur a controlled additional error, while learned per-mode timesteps can absorb admissible spectral differences in the discrete dynamics.* The formal statement and proof are given in Theorem 3 in Appendix. Table 6 and Figures 5–6 in Appendix numerically illustrate the corresponding spectral geometry and the role of normalisation and flooring, regarding the theoretical findings.

Remark 2 (LinOSS as a random-frequency special case). LinOSS (Rusch & Rus, 2025) is recovered from SpecOSS by removing the spectral generation step. Replacing the generated state matrix A in (8) with randomly initialised diagonal entries $\hat{A} \sim \mathcal{U}[0, 1]^m$ leaves the rest of the layer unchanged and yields LinOSS-IM/IMEX. LinOSS therefore corresponds to the same architecture with unstructured random frequencies rather than generated ones.

5 Experiments

The experiments are organised into four blocks, ordered roughly by sequence length and scale: (5.1) classification of time-dependent signals on five long-range University of East Anglia (UEA) tasks; (5.2) heart-rate regression from wearable photoplethysmography (PPG) records of about fifty thousand steps; (5.3) real-world weather forecasting; and (5.4) language modelling on WikiText-103 and the Pile with one hundred million training tokens. Architectures, optimisation, hardware and the hyperparameter search behind every number and baseline's full tunning details are in Appendix F. The offline spectrum generation, the forward pass and the training loop are written out as Algorithms 1–3 and as an intuitive code sketch in Appendix K. Parameter counts, wall-clock and memory are analysed in Appendix I, the sensitivity of the results to the realisation, the discretisation and the tuning axes in Appendix J, and every complete result matrix in Appendix H.

5.1 Classification of time-dependent signals

The five datasets are SCP1, SCP2, Ethanol, Heartbeat, and Motor from the UEA archive (Bagnall et al., 2018); their preparation is described in Appendix F.1. For each dataset, we evaluate all bases under both generation routes, with frozen or trainable spectra, under IM and IMEX, together with the multi-basis construction. Table 2 focuses on two questions: whether multiple bases improve over a single Hermite basis, and how much is gained by training the initialised spectrum. Frozen variants remain close to their trainable counterparts across both routes, showing that much of the performance is already present at initialisation. Matched random-A controls use the same architecture and

Table 2: Time-series classification: test accuracy (%) on five long-range UEA tasks. **Bold**: best per column; underlined: second best. (SL) is the Sturm–Liouville spectrum of route 1 and (FA) the function-approximation spectrum of route 2.

Seq. length #Classes	**SCP1** 896 2	**SCP2** 1,152 2	**Ethanol** 1,751 4	**Heartbeat** 405 2	**Motor** 3,000 2	**Avg**
NRDE *ICML'21*	80.9 ± 2.5	53.7 ± 6.9	25.3 ± 1.8	72.9 ± 4.8	47.0 ± 5.7	56.0
NCDE *NeurIPS'20*	79.8 ± 5.6	53.0 ± 2.8	29.9 ± 6.5	73.9 ± 2.6	49.5 ± 2.8	57.2
Log-NCDE *ICML'24*	83.1 ± 2.8	53.7 ± 4.1	34.4 ± 6.4	75.2 ± 4.6	53.7 ± 5.3	60.0
LRU *ICML'23*	82.6 ± 3.4	51.2 ± 3.6	21.5 ± 2.1	78.4 ± 6.7	48.4 ± 5.0	56.4
S5 *ICLR'23*	89.9 ± 4.6	50.5 ± 2.6	24.1 ± 4.3	77.7 ± 5.5	47.7 ± 5.5	58.0
S6 *COLM'24*	82.8 ± 2.7	49.9 ± 9.4	26.4 ± 6.4	76.5 ± 8.3	51.3 ± 4.7	57.4
Mamba *COLM'24*	80.7 ± 1.4	48.2 ± 3.9	27.9 ± 4.5	76.2 ± 3.8	47.7 ± 4.5	56.1
LinOSS-IMEX *ICLR'25*	87.5 ± 4.0	58.9 ± 8.1	29.9 ± 1.0	75.5 ± 4.3	52.6 ± 5.3	60.9
LinOSS-IM *ICLR'25*	87.8 ± 2.6	58.2 ± 6.9	29.9 ± 0.6	75.8 ± 3.7	60.0 ± 7.5	62.3
D-LinOSS *2026*	88.9 ± 3.0	58.6 ± 2.3	29.9 ± 0.6	75.8 ± 4.9	61.1 ± 2.0	62.9
StableSSM *ICML'24*	84.7 ± 2.1	56.1 ± 2.1	25.3 ± 2.9	78.0 ± 2.4	$\underline{47.4 \pm 3.2}$	58.3
Neural Wave *ICML'23*	85.9 ± 3.2	50.9 ± 3.0	34.2 ± 1.7	72.6 ± 2.7	59.6 ± 3.4	60.6
Deep SL *ICML'25*	85.2 ± 1.9	49.1 ± 2.2	34.7 ± 1.8	77.4 ± 1.4	36.8 ± 1.6	56.6
Multi (SL, ours)	88.2 ± 2.4	61.4 ± 2.3	$\underline{40.5 \pm 3.1}$	$\mathbf{85.5 \pm 2.4}$	56.1 ± 3.8	$\underline{66.3}$
Multi (SL, frozen spectrum, ours)	87.2 ± 3.1	58.0 ± 2.2	40.0 ± 1.8	83.9 ± 2.4	55.9 ± 2.5	65.0
Multi (FA, ours)	$\mathbf{90.6 \pm 3.4}$	$\mathbf{68.4 \pm 2.2}$	40.2 ± 2.8	$\underline{84.2 \pm 3.3}$	54.4 ± 1.2	**67.6**
Multi (FA, frozen spectrum, ours)	90.1 ± 2.3	$\underline{67.7 \pm 2.4}$	36.7 ± 3.1	82.3 ± 2.4	53.8 ± 2.1	66.1
Hermite (SL, ours)	$\underline{90.3 \pm 3.1}$	56.1 ± 2.4	$\mathbf{44.3 \pm 3.9}$	77.4 ± 3.1	57.9 ± 2.6	65.2
Hermite (FA, ours)	84.7 ± 1.8	57.9 ± 2.8	31.6 ± 2.6	79.0 ± 3.3	$\mathbf{64.9 \pm 2.7}$	63.6
Δ best ours $-$ best baseline	$+0.7$	$+9.5$	$+9.6$	$+7.1$	$+3.8$	$+4.7$

parameter count, with the random spectrum either frozen or trainable is in Appendix H.1. Their substantially weaker frozen performance isolates the benefit of the orthogonal-basis spectral initialisation. Multi-basis models are strongest on average under both routes and achieve the best result on three of the five datasets; on Ethanol and Motor, however, Hermite alone performs better, indicating that a single basis can sometimes be sufficient. Overall, our best variant per dataset achieve the best result on all five datasets and the best average against the baselines. IM/IMEX comparisons, full per-basis and per-route results, and parameter counts are reported in Appendices J.2 and H.1, and Table 9. Animated data flow is available at: sites.google.com/view/specoss or our anonymous repo.

5.2 Heart-rate regression from wearable sensors

In this experiment we test SpecOSS in the case of very long-range interactions. We consider the PPG-DaLiA dataset, a multivariate time-series regression dataset designed for heart-rate prediction from data collected by a wrist-worn device (Reiss et al., 2019). It includes recordings from fifteen individuals, each with approximately 150 minutes of data sampled at a maximum rate of 128 Hz, over six channels: blood volume pulse, electrodermal activity, body temperature and three-axis acceleration.

We follow the standard preprocessing protocol of Walker et al. (2024), under which one input sequence spans almost fifty thousand steps. The architecture and the memory-bounded chunked scan that makes this length trainable are described in Appendices B.4 and F. The test MSE of our models and of the competing models is shown in Table 3. A trainable multi-basis spectrum reaches 2.39×10^{-2}, which improves on the best baseline, D-LinOSS at 6.16, by a factor of 2.6 and on Mamba and the LRU by a factor of more than four; the frozen union follows at 2.84, and the Hermite basis alone trails both whether its spectrum is frozen or trained. A union of bases therefore leads on this signal, and training the spectrum helps under either basis. Figure 3 draws every test window at once. The model places each recording at the right level, the per-window mean rate tracking the truth to about three and a half beats per minute, and it reproduces the varia-

Table 3: PPG-DaLiA, test MSE $\times 10^{-2}$. **Bold**: best; underlined: second.

Model	MSE
NRDE	9.90 ± 0.97
NCDE	13.54 ± 0.69
Log-NCDE	9.56 ± 0.59
LRU	12.17 ± 0.49
S5	12.63 ± 1.25
S6	12.88 ± 2.05
Mamba	10.65 ± 2.20
LinOSS-IMEX	7.5 ± 0.46
LinOSS-IM	6.4 ± 0.23
D-LinOSS	6.16 ± 0.73
Hermite (SL, frozen, ours)	3.46 ± 0.44
Hermite (SL, trainable, ours)	3.10 ± 0.51
Multi (SL, frozen, ours)	2.84 ± 0.77
Multi (SL, trainable, ours)	$\mathbf{2.39 \pm 0.27}$

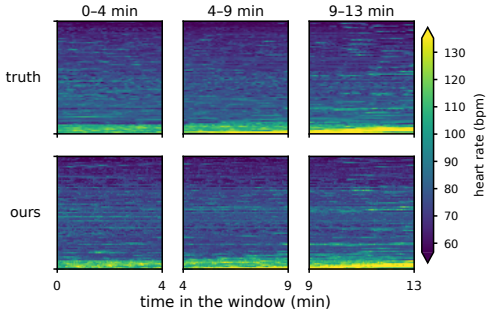


Figure 3: All 106 test windows: truth above ours, colour is heart rate.

tion inside a window rather than flattening it. The rows it fits least well are those at the top of the range, where exertion lifts the rate furthest above rest. The full results for every basis under both discretisations, and the extended results for the FA route, are in Appendix H.2.

5.3 Weather forecasting

Weather forecasting at a long horizon is a real-world test of whether a layer can carry information across thousands of steps of a signal it did not generate. The WTH dataset records twelve meteorological variables at hourly resolution, and the task is to predict the next 720 hours, a month ahead, from the preceding 720. We follow the protocol of Zhou et al. (2021); the data split and the architecture are described in Appendix F.1. The horizon is what makes the task demanding. The model is given no observations at all for the period it must predict, so everything it produces there has to come from what its state has retained of the month before. Table 4 reports the results. Results for every basis under both routes, and the MSE of the same rows, are in Appendix H.3, Table 14.

Table 4: WTH 720→720, test MAE. **Bold**: best; underlined: second.

Model	MAE
Informer	0.731 ± 0.010
LogTrans	0.773 ± 0.011
Reformer	1.575 ± 0.019
LSTMa	1.109 ± 0.017
LSTnet	0.757 ± 0.008
S4	0.578 ± 0.008
LinOSS-IMEX	0.508 ± 0.007
LinOSS-IM	0.528 ± 0.006
D-LinOSS	0.486 ± 0.005
Multi (SL, ours)	$\mathbf{0.380 \pm 0.005}$
Hermite (SL, ours)	0.430 ± 0.006
Multi (FA, ours)	0.400 ± 0.006
Hermite (FA, ours)	0.460 ± 0.006

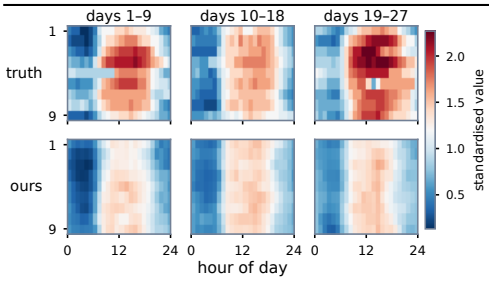


Figure 4: The forecast by day and hour of day: truth above ours.

What the layer retains over such a horizon is visible in Figure 4, which reshapes the forecast into days by hour of day. The daily temperature cycle is still there a month ahead and in phase, coldest before dawn and warmest in mid-afternoon to within an hour, at about three quarters of its true amplitude. What the forecast does not reproduce is the day-to-day weather, the warm and cold spells that make the rows of the upper panels differ from one another, so it repeats a similar day where the truth varies.

5.4 Language modelling

Finally, we evaluate the generated spectra in language modelling. This experiment is not intended to establish performance at large-scale training budgets, but to test whether the same idea remains effective in a causal language-model setting. With a budget of one hundred million training tokens, a single Hermite spectrum improves over every matched state-space baseline in Table 5. We evaluate on WikiText-103 (Merity et al., 2017) and the deduplicated Pile (Gao et al., 2020), using the same tokenizer (Black et al., 2022), token order, sequence packing, and optimisation setup for all variants. Full architectural and training details are given in Appendix F.6.

Table 5: Language modelling after 100M tokens: negative log-likelihood (NLL, nats per token) and perplexity (PPL). **Bold**: best per column.

		WikiText-103		Pile	
Model	Params	NLL ↓	PPL ↓	NLL ↓	PPL ↓
LRU	46.9M	3.977	53.37	4.837	126.07
S4D-Lin	46.9M	3.946	51.70	4.676	107.37
S5	46.9M	3.931	50.96	4.736	114.02
Mamba	46.7M	3.891	48.96	4.592	98.69
LinOSS	46.8M	3.901	49.45	4.660	105.65
D-LinOSS	46.8M	3.935	51.14	4.694	109.34
Hermite (SL, ours)	46.7M	**3.818**	**45.51**	**4.476**	**87.88**

6 Conclusion

We showed that the oscillator spectrum of a state-space model can be constructed from structured orthogonal bases rather than initialised randomly. Using Sturm–Liouville spectra and online function approximation, SpecOSS provides effective spectral initialisations that can remain fixed or be adapted during training, and can be extended to multiple bases within the same model. Across classification, regression, forecasting, and language modelling, these constructions improve over randomly initialised oscillatory models, and our theory characterises their approximation error and expressivity.

AI USE STATEMENT

Generative AI tools were used to edit and polish the text of this paper. The authors reviewed all AI-assisted text and take responsibility for the final content.

REFERENCES

Uri M. Ascher, Steven J. Ruuth, and Raymond J. Spiteri. Implicit-explicit Runge–Kutta methods for time-dependent partial differential equations. *Applied Numerical Mathematics*, 25(2-3):151–167, 1997.

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The UEA multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2015.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In *Workshop on Challenges & Perspectives in Creating Large Language Models*, 2022.

Guy E. Blelloch. Prefix sums and their applications. Technical Report CMU-CS-90-190, School of Computer Science, Carnegie Mellon University, 1990.

Stephen Boyd and Leon O. Chua. Fading memory and the problem of approximating nonlinear operators with Volterra series. *IEEE Transactions on Circuits and Systems*, 32(11):1150–1161, 1985.

Jared Boyer, T. Konstantin Rusch, and Daniela Rus. Learning to dissipate energy in oscillatory state-space models. *arXiv preprint arXiv:2505.12171*, 2025.

Pafnuty L. Chebyshev. Théorie des mécanismes connus sous le nom de parallélogrammes. *Mémoires des Savants Étrangers présentés à l'Académie Impériale des Sciences de Saint-Pétersbourg*, 7: 539–586, 1854.

Richard Courant, Kurt Friedrichs, and Hans Lewy. Über die partiellen Differenzengleichungen der mathematischen Physik. *Mathematische Annalen*, 100(1):32–74, 1928.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International Conference on Machine Learning (ICML)*, 2017.

Philip J. Davis and Philip Rabinowitz. *Methods of Numerical Integration*. Academic Press, 2nd edition, 1984.

Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, et al. Griffin: Mixing gated linear recurrences with local attention for efficient language models. In *arXiv preprint arXiv:2402.19427*, 2024.

Yassir Fathullah, Chunyang Wu, Yuan Shangguan, Junteng Jia, Wenhan Xiong, Jay Mahadeokar, Chunxi Liu, Yangyang Shi, Ozlem Kalinli, Mike Seltzer, and Mark J. F. Gales. Multi-head state space model for speech recognition. In *Interspeech*, 2023.

Joseph Fourier. *Théorie Analytique de la Chaleur*. Firmin Didot, 1822.

Daniel Y. Fu, Tri Dao, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. In *International Conference on Learning Representations (ICLR)*, 2023.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2024.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. HiPPO: Recurrent memory with optimal polynomial projections. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022a.

Albert Gu, Ankit Gupta, Karan Goel, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022b.

Albert Gu, Isys Johnson, Aman Timalsina, Atri Rudra, and Christopher Ré. How to train your HiPPO: State space models with generalized orthogonal basis projections. In *International Conference on Learning Representations (ICLR)*, 2023.

Ankit Gupta, Albert Gu, and Jonathan Berant. Diagonal state spaces are as effective as structured state spaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer, 2nd edition, 2006.

Boris Hanin. Universal function approximation by deep neural nets with bounded width and ReLU activations. *Mathematics*, 7(10):992, 2019.

Joshua Hanson and Maxim Raginsky. Universal simulation of stable dynamical systems by recurrent neural nets. *Learning for Dynamics and Control (L4DC)*, 2020.

Ramin Hasani, Mathias Lechner, Tsun-Hsuan Wang, Makram Chahine, Alexander Amini, and Daniela Rus. Liquid structural state-space models. In *International Conference on Learning Representations (ICLR)*, 2023.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Zifeng Huang, Konstantin M. Zuev, Yong Xia, and Michael Beer. Upper approximation bounds for neural oscillators. *arXiv preprint arXiv:2512.01015*, 2025.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015.

Taniya Kapoor, Abhishek Chandra, Daniel M. Tartakovsky, Hongrui Wang, Alfredo Nunez, and Rolf Dollevoet. Neural oscillators for generalization of physics-informed machine learning. In *AAAI Conference on Artificial Intelligence*, 2024.

T Anderson Keller and Max Welling. Neural wave machines: learning spatiotemporally structured representations with locally coupled oscillatory recurrent neural networks. In *International Conference on Machine Learning*, pp. 16168–16189. PMLR, 2023.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations (ICLR)*, 2020.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2018.

Samuel Lanthaler, T. Konstantin Rusch, and Siddhartha Mishra. Neural oscillators are universal. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yuk Wing Lee. Synthesis of electric networks by means of the Fourier transforms of Laguerre's functions. *Journal of Mathematics and Physics*, 11(1-4):83–113, 1932.

Adrien Marie Legendre. Recherches sur l attraction des spheroides homogenes. 1785.

Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Zhong Li, Jiequn Han, Weinan E, and Qianxiao Li. Approximation and optimization theory for linear continuous-time recurrent neural networks. *Journal of Machine Learning Research*, 23 (42):1–85, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Eric Martin and Chris Cundy. Parallelizing linear recurrent neural nets over sequence length. *International Conference on Learning Representations (ICLR)*, 2018.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017.

James Morrill, Cristopher Salvi, Patrick Kidger, and James Foster. Neural rough differential equations for long time series. In *International Conference on Machine Learning (ICML)*, 2021.

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 (NeurIPS)*, 2024.

Frank WJ Olver. *NIST handbook of mathematical functions hardback and CD-ROM*. Cambridge university press, 2010.

Antonio Orvieto, Samuel L. Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *International Conference on Machine Learning (ICML)*, 2023.

Antonio Orvieto, Soham De, Caglar Gulcehre, Razvan Pascanu, and Samuel L. Smith. Universality of linear recurrences followed by non-linear projections: Finite-width guarantees and benefits of complex eigenvalues. In *International Conference on Machine Learning (ICML)*, 2024.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning (ICML)*, 2023.

Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. Deep PPG: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14):3079, 2019.

T. Konstantin Rusch and Siddhartha Mishra. Coupled oscillatory recurrent neural network (coRNN): An accurate and (gradient) stable architecture for learning long time dependencies. In *International Conference on Learning Representations (ICLR)*, 2021a.

T. Konstantin Rusch and Siddhartha Mishra. UnICORNN: A recurrent model for learning very long time dependencies. In *International Conference on Machine Learning (ICML)*, 2021b.

T. Konstantin Rusch and Daniela Rus. Oscillatory state-space models. In *International Conference on Learning Representations (ICLR)*, 2025.

Erwin Schrödinger. Quantisierung als eigenwertproblem. *Annalen der Physik*, 384(4):361–376, 1926.

Jimmy T.H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Gábor Szegő. *Orthogonal Polynomials*. American Mathematical Society, 1939.

David Vigouroux, Joseba Dalmau, Louis Béthune, and Victor Boutin. Deep sturm–liouville: From sample-based to 1d regularization with learnable orthogonal basis functions. *arXiv preprint arXiv:2504.07151*, 2025.

Aaron Voelker, Ivana Kajić, and Chris Eliasmith. Legendre memory units: Continuous-time representation in recurrent neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020.

Benjamin Walker, Andrew D. McLeod, Tiexin Qin, Yichuan Cheng, Haoliang Li, and Terry Lyons. Log neural controlled differential equations: The lie brackets make a difference. In *International Conference on Machine Learning (ICML)*, 2024.

Shida Wang and Qianxiao Li. Stablessm: Alleviating the curse of memory in state-space models through stable reparameterization. *arXiv preprint arXiv:2311.14495*, 2023.

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 (NeurIPS)*, 2023.

Norbert Wiener. *Nonlinear Problems in Random Theory*. MIT Press, 1958.

Anton Zettl. *Sturm–Liouville Theory*, volume 121 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2005.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI Conference on Artificial Intelligence*, 2021.

TECHNICAL APPENDICES

A Notation Summary . . . 14
B Background details . . . 17
B.1 Sturm–Liouville spectra of the classical bases . . . 17
B.2 From a spectrum to a layer . . . 17
B.3 Discretisation . . . 17
B.4 Parallel scan and the exact chunked scan . . . 18
B.5 Residual block . . . 18
C Derivation of the generated spectra . . . 18
C.1 Route 1: why a Sturm–Liouville eigenvalue is a squared frequency . . . 19
C.2 Route 2: the twice-differentiated memory law . . . 21
D Theory and proofs . . . 24
D.1 Modal solution and forced-response bound . . . 24
D.2 Proof of Theorem 1 . . . 24
D.3 Proof of Theorem 2 . . . 25
D.4 The learned discrete kernel and the error decomposition . . . 25
D.5 Frozen spectra attain the free-frequency rate . . . 26
D.6 Sine-series approximation on harmonic-type spectra . . . 26
D.7 Frequency perturbation . . . 27
D.8 Proof of Theorem 3 . . . 27
D.9 Necessity of normalisation and floor . . . 28
D.10 Spectrum geometry . . . 28
E Extended related work . . . 29
F Experimental details: architectures, optimisation and hyperparameter tuning . . . 31
F.1 Datasets and protocols . . . 32
F.2 Architectures and hyperparameters . . . 32
F.3 Optimisation . . . 33
F.4 Hyperparameter tuning protocol . . . 33
F.5 Weather backbone . . . 33
F.6 Language-model setup . . . 34
G Data flow per experiment group . . . 34
G.1 Group 1: time-series classification and pooled regression . . . 35
G.2 Group 2: photoplethysmography regression . . . 35
G.3 Group 3: weather and masked forecasting . . . 35
G.4 Group 4: language modelling . . . 36
H Additional results . . . 37
H.1 Every basis, both routes, frozen and trainable . . . 37
H.2 PPG-DaLiA, all variants . . . 38
H.3 Weather forecasting, full core-swap matrix . . . 38
H.4 Language modelling, full result matrix . . . 40
H.5 Electrocardiogram classification . . . 40
I Parameter count, run time and memory . . . 40
J Sensitivity analysis . . . 43
J.1 Realisation of the layer around the spectrum . . . 43
J.2 IM versus IMEX . . . 43
J.3 Weather backbone and optimiser . . . 44
K Algorithms and code sketch . . . 44

A NOTATION SUMMARY

Every symbol used in the main text and in this appendix is listed below, and each carries one meaning throughout.

Symbol	Meaning
Sequences and data flow	
T	sequence length in steps; the continuous problem runs over $[0, T]$ at unit step spacing
T_{fut}, T_{out}	forecast horizon (masked steps); readout stride of the strided-regression head
T_{raw}, n_b	length of a raw series before windowing; batch size
T_c	chunk length of the memory-bounded scan
d_{in}, d_{out}	input / output dimension of a task
X, $\mathbf{I}$	raw record; the windowed input batch built from it by the pipeline of the group. Bold marks the batched task-level tensors, as against the per-layer signals u, x, y
$U^{(\ell)}$	output of block ℓ, at the model width ($U^{(0)}$ the encoded input, $U^{(L)}_{\text{str}}$ its strided rows)
$\mathbf{O}$, $\hat{\mathbf{O}}$	target and prediction, again bold and so distinct from the layer readout y
$\mathcal{E}$, $\mathcal{R}$	encoder and task head of the data flow (26)
$W_{\text{enc}}, W_{\text{head}}, W_{\text{emb}}$	encoder, task head and tied embedding matrices; biases $b_\bullet$
The layer and its stack	
H, m, L	model width (channels between blocks, and the input/output width of a layer); oscillators (modes) per layer; residual blocks
Blk_ℓ, $\text{SpecOSS}(\cdot)$	the residual block wrapping one oscillator layer; one application of the model, $u^\ell = \text{SpecOSS}(u^{\ell-1})$, (28)
W_1, W_2, v	the two maps of the gated linear unit $\text{GLU}(v) = W_1 v \odot \sigma(W_2 v)$ and its argument
u_n, $u(t)$	layer input at step n / continuous input signal in $\mathbb{R}^H$
y_n, $y(t)$	layer output (readout)
x, z, $s_n = (z_n, x_n)^\top$	oscillator position (the hidden state) in $\mathbb{R}^m$, velocity $z = x'$, and per-mode discrete state
$A = \text{diag}(A_0, \dots, A_{m-1})$	state matrix, nonnegative and diagonal as the layer runs it: the spectrum of squared frequencies
A_{grid}, $\hat{A}$	generated spectrum before training; free parameter of the trainable form (8)
$\omega_k = \sqrt{A_k}$, $a_{\min}$	natural frequency of mode k; floor of the spectrum (7)
$\Delta t_k = \sigma(\text{steps}_k)$	learned per-mode timestep, σ the logistic sigmoid; Δt a common fixed step
B, b_k	input map, $\mathbb{R}^{m \times H}$, and its k-th row
C, C_j	readout map, $\mathbb{R}^{H \times m}$; per-block readout of block j in multi-basis fusion
Re	real part
M_k^{IM}, M_k^{IMEX}	2×2 discrete transition of mode k under IM / IMEX
$S_k = (1 + \Delta t_k^2 A_k)^{-1}$	IM contraction factor
F_n, (c_z, c_x)	discrete forcing and its two scales
$\lambda(\cdot)$	eigenvalue of a matrix
Λ, γ	diagonal transition and forcing of the first-order (LRU, S5) scan
Language models	
d_{model}, d_{inner}, d_{state}	backbone width, inner width and per-channel state size
d_{vocab}	vocabulary size of the tokenizer
n_g, d_{head}, d_{conv}	channel groups of the grouped core, the width d_{inner}/n_g of one group, and the block's causal convolution width
W_{in}, W_{out}	the block's expansion and output maps
V, $\tilde{V}$, V_z	the block's expanded stream before and after its causal filter, and its gate branch
Generation routes	
$\mathcal{L}$, $\{p_k\}$, λ_k	SL operator of a basis, the basis it diagonalises (P_k Legendre, T_k Chebyshev, L_k Laguerre, H_k Hermite), and its eigenvalues
λ_k^{SL}, $\lambda_{b,k}^{\text{SL}}$, λ_k^{FA}	the spectrum of route 1 ($= \lambda_k$; with the basis b made explicit) and the spectrum of route 2 ($= \lvert\mu_k\rvert^2$)
$\lambda_{b,k}$	the spectrum of basis b read by either route, $\lambda_{b,k}^{\text{SL}}$ or $\lambda_{b,k}^{\text{FA}}$, as fused in (3)
b, θ, ξ	name of a basis (Hermite, Legendre, Chebyshev, Laguerre, Fourier); Fourier window length; coordinate on the interval of a basis, the argument of $p_k(\xi)$
w, π_2	weight and leading coefficient of the SL operator
A_{H}, B_{H}, f	generator and input vector of the memory law and its scalar input; the projection coefficients are the state x

continued on next page

Notation summary (continued from previous page)

Symbol	Meaning
$g_k^{(t)}$, $\mathcal{D}$	HiPPO orthonormal basis function at time t; the skew-symmetric tridiagonal generator of the translating Hermite window
μ_k, η_k, $\varsigma_k(A_{\mathrm{H}})$	eigenvalues of A_{H}; imaginary parts for skew A_{H}; singular values of A_{H}
Q, $\tilde{x}$, P	orthogonal eigenvector matrix of A_{H}^2 and rotated coefficients; eigenvector matrix of a diagonalisable matrix
ψ_k, $\tilde{p}_k$, $\langle\cdot,\cdot\rangle_w$	Hermite functions; orthonormalised function of a basis; inner product of the weight w
$\varphi(\xi,t)$, $\chi(\xi)$	latent field of the route-1 derivation, expanded in the basis, and the injection profile through which the input drives it
$\mathcal{X}$, a, $a^\dagger$	position matrix of the Hermite basis; annihilation and creation (ladder) operators
ϖ	warped coordinate $\xi = \cos\varpi$ of the Chebyshev problem
J, b_j, m_j, ρ_j	number of bases, basis, block size and scale of block j
$A^{(j)}$, $x^{(j)}$, $y^{(j)}$	spectrum, sub-state and per-block readout of block j (3)
G, W_g, $W_{g,1}$, $W_{g,2}$, r, b_g	gate, dense gate matrix, its low-rank factors and their rank, gate bias
β, $\mathrm{softmax}(\beta)$	static mixture logits over blocks (weighted fusion)
p_{drop}	mixture-dropout probability
Indices and generic symbols	
t, τ	time and integration variable
n, k, j, i, ℓ	index of a time step (and of the harmonics of Appendix D.5); of a mode, which is also the degree of its basis function, $k = 0,\dots,m-1$; of a block; of a lag; of a layer
$\mathcal{U}$, $\mathbb{E}$	uniform distribution; expectation
Theory: spectral error and expressivity (Section 4.2, Appendix D)	
$\omega_{\min}$, $\omega_{\max}$	ends of the normalised band, $\omega_{\min} = \sqrt{a_{\min}}$ and $\omega_{\max} = 1$
$\nu(\omega)$, $\bar{\nu}$, L_ν	spectral density over the band, its bound and its Lipschitz constant
$\mathcal{K}_\nu$, $\mathcal{K}_\nu^m$	continuous reference kernel of ν and its quadrature on the spectrum (4)
$\mathcal{I}_k$, h_m	cell of the band partition that contains ω_k; largest cell width
$\mathfrak{K}$, $\mathfrak{K}_m$	kernels induced by Lipschitz densities on the band; kernels realisable by m oscillators on the spectrum
Θ, $\mathcal{K}_{m,\ell}^\Theta$, $\mathcal{K}_{k,\ell}^\Theta$	trainable parameters $(\hat{A}, \mathrm{steps}, B, C)$; kernel of the learned discrete layer at lag ℓ and the contribution of mode k
τ_j, Δt_{ref}, e_x, $C_{:,k}$	reference lag grid $\tau_j = (j-1)\Delta t_{\mathrm{ref}}$; selector $(0,1)^\top$ of the position in s_n; k-th column of C
y_ℓ^ν, y_ℓ^Θ	responses of the quadrature kernel and of the learned layer to a discrete input
$\|\cdot\|_{L^\infty(0,T)}$, $\|\cdot\|_2$, $\|u\|_\infty$	sup norm in time; Euclidean or spectral norm; largest magnitude of a discrete input
Theory: frozen spectra and necessity (Appendix D.5–D.9)	
K, B_K, L_K	compact input set, its sup-norm bound and Lipschitz constant
Φ, ϕ_Φ, q	target causal operator, its modulus of continuity, its output dimension
$\mathcal{L}_t f(\omega)$	windowed sine transform $\int_0^t f(t-\tau)\sin(\omega\tau)\,d\tau$
$\omega_n^\natural$, δ	harmonic-type spectrum $(n+\delta)\pi/(2T)$ and its offset $\delta \in \{0, \frac{1}{2}\}$
$\tilde{\omega}_n$, Δ_n	an arbitrary spectrum and its deviation from $\omega_n^\natural$
Π, H_Π, ε_Π	readout MLP, its depth parameter, its approximation term
c_K, c_ρ, c_F, ε_K, B_Ω	constants of the free-frequency bound ($c_K = c_\rho c_F$); its truncation error $32 c_K H L_K T(\ln m)^2/m$; its domain constant
υ, ρ_υ, $\tilde{f}_t$, $\tilde{f}_\upsilon$, $\bar{f}_\upsilon$	mollifier width, mollifier, odd extension of $u(t-\tau)-u(0)$, its mollification and its even reflection
$\varkappa_n$, ϑ_n, p, $\mathcal{H}$	weight and phase of harmonic n in the sine-series approximant; exponent of the Jackson-type bound; Fourier transform
Ψ, L_{ψ_j}, e_h	continuous map from $(x, u(0), t)$ to $\Phi[u](t)$; Lipschitz constants of its components; h-th standard basis vector
κ_k, r_k, c_κ, $\lambda_\pm$	stiffness $\Delta t_k^2 A_k$ of mode k; IM spectral radius $(1+\kappa_k)^{-1/2}$; power-bound constant; eigenvalues of a 2×2 transition
$\ell_\epsilon(k)$, ϵ, σ_f^2, f_j	memory horizon of mode k at tolerance ϵ; variance and samples of a scalar forcing sequence

B Background details

B.1 Sturm–Liouville spectra of the classical bases

Each classical basis p_k is orthogonal on an interval with respect to a weight w and satisfies $\mathcal{L}p_k = \lambda_k p_k$ for a self-adjoint Sturm–Liouville operator

$$\mathcal{L}p = -\frac{1}{w}\big(w\,\pi_2\,p'\big)' + V(\xi)p,$$

with the corresponding operators listed in Table 1. For Legendre, Chebyshev, Laguerre, and Fourier we have $V = 0$, with $(w, \pi_2) = (1, 1-\xi^2)$, $((1-\xi^2)^{-1/2}, 1-\xi^2)$, $(e^{-\xi}, \xi)$, and $(1, 1)$, respectively, on their standard domains. For Fourier on $[0, \theta)$, $\lambda_k = (2\pi k/\theta)^2$; since the common scale is removed by the normalisation below, we use the equivalent spectrum k^2. For Hermite on $\mathbb{R}$, we use the Schrodinger form with eigenfunctions $\psi_k = H_k e^{-\xi^2/2}$ and eigenvalues $\lambda_k = 2k + 1$, giving a strictly positive spectrum without a zero mode. The resulting spectra are constructed in double precision, then normalised by (6) and floored by (7) in the layer.

B.2 From a spectrum to a layer

Before entering the layer, a generated spectrum is normalised and floored. That both steps are needed follows from the oscillator interpretation rather than from tuning; the value of $a_{\min}$ is selected on validation (Appendix F.4). Since $A_k = \omega_k^2$, the raw spectra can contain increasingly large frequencies as the number of modes m grows. At $\Delta t = 1$, frequencies above the Nyquist rate cannot be resolved by the sampled sequence and, under IM, sufficiently large frequencies rapidly suppress the input response (Proposition 4a). We therefore rescale

$$A \leftarrow A/\max_k A_k \qquad (\omega_{\max} = 1) \tag{6}$$

so that every generated spectrum occupies the same frequency range, independently of its original growth rate. Since $\Delta t_k = \sigma(\cdot) < 1$, this also guarantees $\Delta t_k^2 A_k \le 1 < 4$, and hence the CFL bound. Fourier, Legendre, Chebyshev, and Laguerre contain a zero mode, $A_0 = 0$, which corresponds to a double integrator; under forcing its response grows like $T^{3/2}$ (Proposition 4c). We therefore apply the floor

$$A \leftarrow \max(A, a_{\min}), \qquad a_{\min} > 0. \tag{7}$$

The resulting spectrum is used either frozen, with A stored as a non-trainable buffer, or trainable, with

$$A = a_{\min} + \hat{A}^2, \qquad \hat{A}_{\text{init}} = \sqrt{\max(A_{\text{grid}} - a_{\min}, 0)}, \tag{8}$$

so that optimisation starts exactly from the generated spectrum while keeping A strictly positive. The layer itself is agnostic to how A is generated: it only requires a nonnegative diagonal A, and all variants use the same A-independent initialisation of B (Appendix F.2).

B.3 Discretisation

Write the oscillator system in first-order form as $z' = -Ax + Bu$ and $x' = z$. The IM scheme treats both updates implicitly,

$$z_n = z_{n-1} + \Delta t(-Ax_n + (Bu)_n), \qquad x_n = x_{n-1} + \Delta t z_n.$$

Eliminating x_n gives

$$z_n = S\big(z_{n-1} - \Delta t A x_{n-1} + \Delta t(Bu)_n\big), \qquad x_n = S\big(\Delta t z_{n-1} + x_{n-1} + \Delta t^2 (Bu)_n\big),$$

where $S = (I + \Delta t^2 A)^{-1}$. This yields M^{IM} with $(c_z, c_x) = (\Delta t S, \Delta t^2 S)$. IMEX instead evaluates the restoring force explicitly,

$$z_n = z_{n-1} + \Delta t(-Ax_{n-1} + (Bu)_n), \qquad x_n = x_{n-1} + \Delta t z_n,$$

giving M^{IMEX} and $(c_z, c_x) = (\Delta t, \Delta t^2)$. For each diagonal mode of IM, the characteristic polynomial is $\lambda^2 - 2S\lambda + S = 0$, with determinant S; hence the complex pair satisfies $|\lambda| = \sqrt{S} = (1 + \Delta t^2 A)^{-1/2}$ and is therefore dissipative. For IMEX, the trace is $2 - \Delta t^2 A$ and the determinant is 1. Thus, for $0 < \Delta t^2 A < 4$, the eigenvalues form a complex-conjugate pair on the unit circle, while outside this Courant–Friedrichs–Lewy range they are real and one has modulus larger than one. These are the properties used in Proposition 4(a,b) and shown in the left panel of Figure 6.

B.4 Parallel scan and the exact chunked scan

The recurrence $s_n = M_n s_{n-1} + F_n$ is a first-order linear scan. Two consecutive updates can be combined with the associative operator

$$(M_a, F_a) \circ (M_b, F_b) = (M_b M_a, M_b F_a + F_b)$$

(Blelloch, 1990; Martin & Cundy, 2018), allowing the full sequence to be evaluated in $O(\log T)$ parallel depth. Since A is diagonal, each mode is represented only by the four entries of its 2×2 transition M_n and the two entries of its forcing F_n, and a Hillis–Steele scan requires $\lceil \log_2 T \rceil$ levels. Storing all of these levels for backpropagation becomes expensive at long sequence lengths; at $T \approx 5 \times 10^4$, it requires about 13 GB per layer for the LinOSS configuration used here. We therefore split the sequence into chunks of length T_c. Each chunk is scanned from zero to obtain $\hat{s}_n$, while the incoming state is propagated through its cumulative transition $M_{1:n}$, giving exactly

$$s_n = M_{1:n} s_0 + \hat{s}_n.$$

Thus chunking changes only how the recurrence is evaluated, not its result. With gradient checkpointing applied per chunk, the peak storage scales as $6 \log_2(T_c), n_b T_c m$ rather than $6 \log_2(T), n_b T m$ numbers, where n_b is the batch size. This makes PPG-DaLiA at length 49,920 trainable with batch size 4 on an A100. The first-order baselines LRU and S5 use the same chunking, with the analogous diagonal composition $(\Lambda_a, \gamma_a) \circ (\Lambda_b, \gamma_b) = (\Lambda_b \Lambda_a, \Lambda_b \gamma_a + \gamma_b)$ for transition Λ and forcing γ.

B.5 Residual block

The model consists of L residual blocks, $u^\ell = \mathrm{SpecOSS}(u^{\ell-1})$ for $\ell = 1, \ldots, L$, with

$$\mathrm{SpecOSS}(u) := u + \mathrm{Drop}\big(\mathrm{GLU}(\mathrm{Drop}(\mathrm{GELU}(\mathrm{SSM}(\mathrm{BN}(u)))))\big).$$

Here BN is channel-wise, affine-free batch normalisation over batch and time (Ioffe & Szegedy, 2015), followed by the oscillator SSM, a Gaussian error linear unit (GELU) (Hendrycks & Gimpel, 2016), dropout, and a gated linear unit $\mathrm{GLU}(v) = W_1 v \odot \sigma(W_2 v)$ (Dauphin et al., 2017), followed by a second dropout. This is the same residual block used by LinOSS. The oscillator layer uses $B \in \mathbb{R}^{m \times H}$ and $C \in \mathbb{R}^{H \times m}$, with readout $y_n = C x_n$. Before the residual stack, an affine encoder maps $\mathbb{R}^{d_{\mathrm{in}}} \to \mathbb{R}^H$, optionally including a time channel. A task-specific head is applied after the final block: mean pooling followed by a linear map for sequence-level prediction, or a per-step linear map for sequence-to-sequence prediction, with the required output steps selected by the task.

C Derivation of the generated spectra

This section derives the two spectra of Table 1 from first principles and shows in what sense each is the spectrum of a collection of oscillators. Route 1 explains why a Sturm–Liouville eigenvalue λ_k^{SL} is a squared natural frequency, by writing the basis's own second-order equation as a forced collection of oscillators, exactly and for every basis; the Hermite case is carried out in full because there the operator itself has the oscillator form. Route 2 starts from the first-order memory law of

online function approximation and shows why its second-order form can be matched to an oscillator only up to a reduction, what that reduction keeps (λ_k^{FA}) and what it gives up (the derived input map).

Table 1, entry by entry. Route-1 column: the Hermite entry $2k{+}1$ is (13), derived from the two recurrences of the Hermite functions in the paragraph "Hermite in full"; the Legendre $k(k{+}1)$, Chebyshev k^2, Laguerre k and Fourier k^2 entries are the eigenvalue equations written out in the paragraph "Legendre, Chebyshev, Laguerre and Fourier", each with the coordinate change or caveat that goes with it. Route-2 column: the generator A_{H} of every basis is given in closed form in the paragraph "Closed forms of the generators", where the Fourier entry $(2\pi k/\theta)^2$ is immediate; the Hermite entry $|\mu_k|^2$ is Proposition 1(a), whose A_{H} is the skew tridiagonal derivative matrix $\mathcal{D}$ with eigenvalues $\pm i\eta_k$; the Legendre entry $(k{+}1)^2$ is Proposition 1(b); the Chebyshev and Laguerre entries $\varsigma_k(A_{\mathrm{H}})^2$ are Proposition 1(c). The last column of the table names the structure of A_{H} that decides which case applies.

C.1 Route 1: why a Sturm–Liouville eigenvalue is a squared frequency

Let p_k be an orthogonal basis on an interval in the internal coordinate ξ, with weight $w(\xi)$, and let

$$\mathcal{L}p_k = \lambda_k p_k, \qquad \lambda_k \geq 0,$$

for the self-adjoint Sturm–Liouville operator

$$\mathcal{L}p = -\frac{1}{w}\big(w\pi_2 p'\big)' + V(\xi)p$$

of Appendix B.1. Together with the domain and boundary conditions, the functions w, π_2, and V specify the corresponding eigenvalue problem; Table 1 gives the operator for each basis. We now show why the eigenvalues λ_k naturally act as squared temporal frequencies. Orthonormalise the basis as $\tilde{p}_k = p_k/|p_k| * w$, so that $\langle \tilde{p}_m, \tilde{p}_k\rangle_w = \delta * mk$. A time-dependent field over the internal coordinate can then be expanded as

$$\varphi(\xi,t) = \sum_k x_k(t)\tilde{p}_k(\xi), \qquad x_k(t) = \langle \varphi(\cdot,t), \tilde{p}_k\rangle_w,$$

so that the ξ-dependence is carried by the fixed basis functions and the time dependence by the coefficients $x_k(t)$. Let this field satisfy

$$\partial_t^2\varphi(\xi,t) = -\mathcal{L}\varphi(\xi,t) + \chi(\xi), u(t), \tag{9}$$

where $\mathcal{L}$ acts as the restoring operator and the input $u(t)$ is applied through a fixed profile $\chi(\xi)$. Since the basis functions are independent of time,

$$\partial_t^2\varphi = \sum_k \ddot{x}_k(t)\tilde{p}_k(\xi),$$

while the eigenvalue relation gives

$$\mathcal{L}\varphi = \sum_k \lambda_k x_k(t)\tilde{p}_k(\xi).$$

Projecting (9) onto $\tilde{p}_k$ and using orthonormality therefore yields

$$\ddot{x}_k = -\lambda_k x_k + \langle \chi, \tilde{p}_k\rangle_w, u, \qquad k = 0, 1, \ldots. \tag{10}$$

Thus, after retaining the first m modes, the field equation reduces to an oscillator system with

$$A = \mathrm{diag}(\lambda_k^{\mathrm{SL}}), \qquad \lambda_k^{\mathrm{SL}} = \lambda_k.$$

The field construction would also induce the input coefficients $B_k = \langle \chi, \tilde{p}_k \rangle_w$; in the actual layer, however, B and C are learned (Appendix B.5), so only the squared-frequency matrix A is taken from this derivation.

This is the normal-mode decomposition of a linear second-order field equation: each eigenfunction of $\mathcal{L}$ defines an independent temporal mode whose squared frequency is the corresponding eigenvalue, $\omega_k^2 = \lambda_k$. It is the same principle by which the wave equation decomposes into oscillators, with the spatial operator $-\partial_\xi^2$ replaced here by the operator $\mathcal{L}$ associated with the chosen basis. The modal decomposition is exact for the complete basis; retaining only the first m modes gives the finite-dimensional system used by the layer. The forcing is chosen as $\chi(\xi)u(t)$, so the projected dynamics depend directly on the input and not on its derivatives. The main modelling choice is therefore the second-order field equation (9): it is second order in time to produce oscillator dynamics, while $\mathcal{L}$ is chosen so that its eigenfunctions coincide with the desired basis. More generally, replacing $\mathcal{L}$ by a spectral function $g(\mathcal{L})$ would give squared frequencies $g(\lambda_k)$; Route 1 uses the canonical choice $g(\lambda) = \lambda$. Finally, $\lambda_k > 0$ gives an oscillatory mode with frequency $\sqrt{\lambda_k}$, $\lambda_k = 0$ gives a zero-frequency mode, and negative eigenvalues would instead produce exponentially growing and decaying solutions. Thus, $\mathcal{L} \succeq 0$ excludes such unstable restoring modes, while strictly positive eigenvalues correspond to genuine oscillations.

Hermite in full. The normalized Hermite functions are $\psi_k(\xi) = (2^k k!\sqrt{\pi})^{-1/2} H_k(\xi) e^{-\xi^2/2}$, where H_k denotes the physicists' Hermite polynomial. They form a complete orthonormal basis of $L^2(\mathbb{R})$. We seek a second-order operator $\mathcal{L}$ for which $\mathcal{L}\psi_k = \lambda_k \psi_k$, because Route 1 uses the resulting eigenvalues λ_k as squared oscillator frequencies.

The standard Hermite identities $H_{k+1} = 2\xi H_k - 2kH_{k-1}$ and $H_k' = 2kH_{k-1}$, together with the Gaussian factor $e^{-\xi^2/2}$, give

$$\xi\,\psi_k = \sqrt{\frac{k}{2}}\,\psi_{k-1} + \sqrt{\frac{k+1}{2}}\,\psi_{k+1}, \qquad \psi_k' = \sqrt{\frac{k}{2}}\,\psi_{k-1} - \sqrt{\frac{k+1}{2}}\,\psi_{k+1}. \tag{11}$$

Thus multiplication by ξ and differentiation act almost identically on the basis, differing only in the sign of the ψ_{k+1} term. In matrix form, multiplication by ξ is represented by a symmetric tridiagonal matrix $\mathcal{X}$, while differentiation is represented by a skew-symmetric tridiagonal matrix $\mathcal{D}$.

Applying the two relations in (11) a second time gives

$$\begin{aligned} \psi_k'' &= \frac{\sqrt{k(k-1)}}{2}\psi_{k-2} - \frac{2k+1}{2}\psi_k + \frac{\sqrt{(k+1)(k+2)}}{2}\psi_{k+2}, \\ \xi^2\psi_k &= \frac{\sqrt{k(k-1)}}{2}\psi_{k-2} + \frac{2k+1}{2}\psi_k + \frac{\sqrt{(k+1)(k+2)}}{2}\psi_{k+2}. \end{aligned} \tag{12}$$

The two expressions have the same ψ_{k-2} and ψ_{k+2} terms, while the coefficient of ψ_k has opposite sign. Therefore, in the combination $-\psi_k'' + \xi^2\psi_k$, the neighbouring modes cancel exactly, leaving

$$\left(-\partial_\xi^2 + \xi^2\right)\psi_k = (2k+1)\psi_k. \tag{13}$$

Hence the Hermite functions diagonalise $\mathcal{L} = -\partial_\xi^2 + \xi^2$ with eigenvalues $\lambda_k = 2k+1$. Equivalently, in the Hermite basis, $-\mathcal{D}^2 + \mathcal{X}^2 = \mathrm{diag}(1, 3, 5, \ldots)$. Neither $-\partial_\xi^2$ nor multiplication by ξ^2 is diagonal by itself: each couples mode k to modes $k \pm 2$, while their combination cancels these couplings exactly. Among linear combinations of these two second-order terms, this cancellation fixes their relative coefficient up to an overall scale.

The same result can also be obtained from the ladder operators $a = (\xi + \partial_\xi)/\sqrt{2}$ and $a^\dagger = (\xi - \partial_\xi)/\sqrt{2}$. Equation equation 11 gives $a\psi_k = \sqrt{k}, \psi_{k-1}$ and $a^\dagger\psi_k = \sqrt{k+1}, \psi_{k+1}$. Since $-\partial_\xi^2 + \xi^2 = 2a^\dagger a + 1$ and $a^\dagger a, \psi_k = k\psi_k$, it follows again that $(-\partial_\xi^2 + \xi^2)\psi_k = (2k+1)\psi_k$. This is the harmonic-oscillator operator, but here the reasoning is used in the opposite direction: the Hermite basis is chosen first, and its recurrence relations determine the corresponding second-order operator and spectrum.

For the infinite Hermite basis, the cancellation above is exact. If $\mathcal{D}$ and $\mathcal{X}$ are first truncated to finite $m \times m$ matrices and then squared, the missing coupling from the last retained mode to ψ_m changes

the final diagonal entry. We therefore do not construct A by truncating and combining these matrices, but instead use the exact eigenvalues directly, with $A = \mathrm{diag}(1, 3, 5, \ldots, 2m-1)$. Consequently, the kth OSS mode satisfies $\ddot{x}_k = -(2k+1)x_k + B_k u$ and has frequency $\omega_k = \sqrt{2k+1}$. Unlike the Fourier, Legendre, Chebyshev, and Laguerre Route 1 spectra, the Hermite spectrum contains no zero mode.

If the continuous field construction of Appendix C.1 were used literally, the input coefficient would be $B_k = \langle \chi, \psi_k \rangle$. For example, a point injection $\chi(\xi) = \delta(\xi - \xi_*)$ gives $B_k = \psi_k(\xi_*)$. At $\xi_* = 0$, the odd Hermite functions vanish, so only even modes are excited, with $B_{2k} = (-1)^k \pi^{-1/4} \sqrt{(2k)!}/(2^k k!)$ and $B_{2k+1} = 0$. In SpecOSS, however, B is learned directly and no injection profile χ is fixed. The Hermite derivation is therefore used only to determine the squared-frequency matrix A.

Legendre, Chebyshev, Laguerre and Fourier. The same Sturm–Liouville argument applies to the remaining bases: once a basis function p_k satisfies $\mathcal{L}p_k = \lambda_k p_k$, Route 1 sets the corresponding squared frequency to λ_k. For the Legendre polynomials on $[-1, 1]$, the standard differential equation can be written as $-[(1-\xi^2)P_k']' = k(k+1)P_k$. Hence the eigenvalues are $\lambda_k = k(k+1)$, giving $A = \mathrm{diag}(k(k+1))$ and frequencies $\omega_k = \sqrt{k(k+1)}$. The $k = 0$ mode has eigenvalue zero, while for larger k the frequencies grow approximately linearly with k. For the Chebyshev polynomials, the differential equation $(1-\xi^2)T_k'' - \xi T_k' + k^2 T_k = 0$ gives the eigenvalue $\lambda_k = k^2$. This can also be seen directly from the change of variable $\xi = \cos\varpi$, for which $T_k(\cos\varpi) = \cos(k\varpi)$. In the coordinate ϖ, the equation becomes $\partial_\varpi^2 \cos(k\varpi) + k^2 \cos(k\varpi) = 0$, so the corresponding operator has eigenvalue k^2. Therefore $A = \mathrm{diag}(k^2)$ and $\omega_k = k$. For the Laguerre polynomials on $[0, \infty)$ with weight $e^{-\xi}$, the Sturm–Liouville equation is $-e^{\xi}[\xi e^{-\xi} L_k']' = kL_k$. Thus $\lambda_k = k$, so Route 1 gives $A = \mathrm{diag}(k)$ and frequencies $\omega_k = \sqrt{k}$. Finally, the Fourier modes satisfy $-\partial_\xi^2 e^{ik\xi} = k^2 e^{ik\xi}$ directly. Therefore their eigenvalues are $\lambda_k = k^2$, giving $A = \mathrm{diag}(k^2)$ and $\omega_k = k$ under the normalized coordinate convention. Legendre, Chebyshev, Laguerre, and Fourier all contain a zero mode at $k = 0$. In the implemented layer, this zero eigenvalue is replaced by the small positive floor introduced in (7), so that every retained mode has a strictly positive frequency.

Consequences for the discretisation. Because A is diagonal, each mode can be discretised independently using the scalar results of Appendix B.3. For a mode with squared frequency A_k, the IM scheme gives $|\lambda_k|^2 = (1 + \Delta t^2 A_k)^{-1}$, while IMEX preserves unit magnitude, $|\lambda_k| = 1$, whenever $0 < \Delta t^2 A_k < 4$. Thus the most restrictive IMEX condition comes from the largest entry of A, namely $\Delta t < 2/\sqrt{A_{m-1}}$. Before spectrum normalisation, this restriction depends on how fast the eigenvalues grow with the mode index. For Legendre, Chebyshev, and Fourier, A_k grows approximately as k^2, so the largest mode requires $\Delta t = O(1/m)$. For Laguerre and Hermite, A_k grows approximately as k, giving the weaker restriction $\Delta t = O(1/\sqrt{m})$. The normalisation in (6) rescales the spectra so that the same timestep condition can be used across all basis families. Legendre, Chebyshev, Laguerre, and Fourier also contain a zero mode with $A_0 = 0$. This mode does not oscillate and lies at the boundary of the IMEX condition above, which motivates replacing it by the small positive floor introduced in (7). Hermite is the exception: its smallest eigenvalue is already $A_0 = 1$, so it has no zero-frequency mode.

C.2 Route 2: the twice-differentiated memory law

The memory law and its second-order form. Online function approximation represents the input history through coefficients $x_k(t)$ in a time-dependent basis. For many such constructions, differentiating these coefficients gives a first-order state equation of the form $x' = -A_{\mathrm{H}} x + B_{\mathrm{H}} f$, where A_{H} is the basis-dependent memory generator and B_{H} determines how the current input $f(t)$ enters the state. Differentiating this equation once more gives

$$\begin{aligned} x' &= -A_{\mathrm{H}} x + B_{\mathrm{H}} f \\ &\Downarrow \\ x'' &= A_{\mathrm{H}}^2 x - A_{\mathrm{H}} B_{\mathrm{H}} f + B_{\mathrm{H}} f'. \end{aligned} \tag{14}$$

The homogeneous part can be compared with the OSS dynamics $x'' = -Ax + Bu$. If μ_k is an eigenvalue of A_{H}, then the corresponding eigencoordinate satisfies $z_k'' = \mu_k^2 z_k$. An oscillator instead

satisfies $z_k'' = -\lambda_k z_k$, so the exact relation would be $\lambda_k = -\mu_k^2$. In particular,

$$\mu_k = i\omega_k \quad \Longrightarrow \quad \lambda_k = -\mu_k^2 = |\mu_k|^2 = \omega_k^2. \tag{15}$$

This gives three cases.

Exact case: purely imaginary spectrum. If A_{H} is skew-symmetric, its eigenvalues occur on the imaginary axis, $\mu_k = i\omega_k$. Then $-\mu_k^2 = \omega_k^2 \geq 0$, so the twice-differentiated memory dynamics produce genuine oscillators and the squared frequencies are exactly $\lambda_k^{\mathrm{FA}} = |\mu_k|^2$. This is the exact Route 2 construction used for families whose first-order generator is skew-symmetric.

Surrogate 1: diagonalizable generator with non-imaginary eigenvalues. If $A_{\mathrm{H}} = P\,\mathrm{diag}(\mu_k)P^{-1}$ is diagonalizable but some μ_k are not purely imaginary, the exact values $-\mu_k^2$ are not generally nonnegative real numbers and therefore cannot be used directly as squared oscillator frequencies. In this case we retain only the spectral scale and define the positive surrogate

$$\lambda_k^{\mathrm{FA}} = |\mu_k|^2. \tag{16}$$

The eigenvector matrix P is not retained: its effect on the input and output maps can instead be absorbed into the learned B and C. This construction preserves the magnitude associated with each mode, but it is no longer dynamically equivalent to the original first-order memory system.

Surrogate 2: defective generator. If A_{H} is defective, it cannot be diagonalized and its eigenvalues alone may contain very little information about the dynamics. A nilpotent generator, for example, may have only zero eigenvalues while still having nontrivial state evolution. For such cases we therefore use the singular values of A_{H} and set $\lambda_k^{\mathrm{FA}} = \varsigma_k(A_{\mathrm{H}})^2$. These values are always nonnegative and provide a positive measure of the strength of each generated mode, but, as in the previous case, they are a surrogate rather than an exact second-order equivalence.

The input terms also prevent an exact match to the OSS layer. The twice-differentiated equation contains both $-A_{\mathrm{H}}B_{\mathrm{H}}f$ and $B_{\mathrm{H}}f'$, whereas the OSS receives only the input itself and not its derivative. Thus the coefficient $-A_{\mathrm{H}}B_{\mathrm{H}}$ alone cannot be identified with the OSS input map. In SpecOSS, B and C are therefore learned directly, and Route 2 is used only to generate the squared-frequency spectrum A.

Closed forms of the generators. The first-order generator A_{H} depends on the basis used for the memory representation. Indexed from $k = 0$, the generators considered here are as follows.

For LegS (Legendre),

$$A_{\mathrm{H}}[k,l] = \sqrt{(2k+1)(2l+1)} \quad \text{for } k > l, \tag{17}$$

with $A_{\mathrm{H}}[k,k] = k+1$ and $A_{\mathrm{H}}[k,l] = 0$ for $k < l$.

For LagT (Laguerre),

$$A_{\mathrm{H}}[k,l] = 1 \quad \text{for } k \geq l, \tag{18}$$

and $A_{\mathrm{H}}[k,l] = 0$ otherwise.

For Chebyshev,

$$A_{\mathrm{H}}[k,l] = 4k \quad \text{for } l < k \text{ and } k-l \text{ odd}, \tag{19}$$

with all other entries equal to zero. The column $l = 0$ is additionally scaled by $2^{-1/2}$.

For the translating Hermite basis, $A_{\mathrm{H}} = \mathcal{D}$ from (11), where $\mathcal{D}^\top = -\mathcal{D}$. For a Fourier window of length θ, the generator is diagonal in the Fourier basis with eigenvalues $\mu_k = 2\pi ik/\theta$.

These generators correspond to the three Route 2 cases described above. Hermite and Fourier have purely imaginary spectra, so their squared frequencies follow exactly from $\lambda_k^{\mathrm{FA}} = |\mu_k|^2$. For Fourier, this gives $\lambda_k^{\mathrm{FA}} = (2\pi k/\theta)^2$. For Hermite, the eigenvalues of the skew-symmetric matrix $\mathcal{D}$ occur in pairs $\pm i\eta_j$. For even m, this gives $m/2$ distinct squared frequencies η_j^2, each appearing twice.

LegS is lower triangular, so its eigenvalues are given by its diagonal entries, $\mu_k = k+1$. Since these eigenvalues are real, the exact values $-\mu_k^2$ would not produce oscillations. Route 2 therefore uses the positive surrogate $\lambda_k^{\mathrm{FA}} = (k+1)^2$.

LagT and Chebyshev are defective, so their eigenvalues do not adequately represent their dynamics. In particular, the Chebyshev generator is strictly lower triangular and therefore has only zero eigenvalues. For these two families, Route 2 instead uses the squared singular values, $\lambda_k^{\mathrm{FA}} = \varsigma_k(A_{\mathrm{H}})^2$, as the positive surrogate spectrum.

Proposition 1 (Exactness ladder). ***(a)*** *If A_{H} is real skew-symmetric, as for Hermite and for Fourier in a real sine–cosine representation, then A_{H}^2 is symmetric negative semidefinite. Hence $A_{\mathrm{H}}^2 = -Q\,\mathrm{diag}(\eta_k^2)Q^\top$ for some orthogonal Q, where the eigenvalues of A_{H} occur in pairs $\pm i\eta_k$. In the coordinates $\tilde{x} = Q^\top x$, the differentiated law becomes $\tilde{x}'' = -\,\mathrm{diag}(\eta_k^2)\tilde{x} + (\textit{forcing})$, so the squared frequencies are exactly $\lambda_k^{\mathrm{FA}} = \eta_k^2 = |\mu_k|^2$.*

(b) *If A_{H} is diagonalisable with real positive spectrum, as for LegS with $\mu_k = k+1$, then in its eigenbasis the differentiated law is $\tilde{x}'' = +\,\mathrm{diag}(\mu_k^2)\tilde{x} + (\textit{forcing})$. These modes are exponential rather than oscillatory. Route 2 therefore replaces the exact values $-\mu_k^2$ by the positive surrogate $\lambda_k^{\mathrm{FA}} = |\mu_k|^2$, preserving the modal scale but not the original dynamics. For LegS, this gives $(k+1)^2 = 1, 4, 9, \ldots$, compared with the Route 1 spectrum $k(k+1) = 0, 2, 6, \ldots$.*

(c) *If A_{H} is defective, as for Laguerre and Chebyshev, no complete eigenbasis exists and the eigenvalues alone do not capture the modal structure. Route 2 therefore uses the singular-value surrogate $\lambda_k^{\mathrm{FA}} = \varsigma_k(A_{\mathrm{H}})^2$, i.e. the eigenvalues of $A_{\mathrm{H}}^\top A_{\mathrm{H}}$.*

Proof. **(a)** If $A_{\mathrm{H}}^\top = -A_{\mathrm{H}}$, then $(A_{\mathrm{H}}^2)^\top = (A_{\mathrm{H}}^\top)^2 = A_{\mathrm{H}}^2$, and $v^\top A_{\mathrm{H}}^2 v = -(A_{\mathrm{H}}v)^\top(A_{\mathrm{H}}v) \le 0$. Thus A_{H}^2 is symmetric negative semidefinite and can be written as $A_{\mathrm{H}}^2 = -Q\,\mathrm{diag}(\eta_k^2)Q^\top$ with Q real orthogonal. The eigenvalues of a real skew-symmetric matrix occur in pairs $\pm i\eta_k$, so $\mu_k^2 = -\eta_k^2$ and therefore $\eta_k^2 = |\mu_k|^2$. With $\tilde{x} = Q^\top x$, the differentiated law becomes $\tilde{x}'' = -\,\mathrm{diag}(\eta_k^2)\tilde{x} + Q^\top(-A_{\mathrm{H}}B_{\mathrm{H}}f + B_{\mathrm{H}}f')$. Hence the homogeneous part is exactly a diagonal oscillator with $\lambda_k^{\mathrm{FA}} = |\mu_k|^2$, while $\tilde{x}$ is a fixed orthogonal transformation of the original projection coefficients.

(b) If $A_{\mathrm{H}} = P\,\mathrm{diag}(\mu_k)P^{-1}$ with real $\mu_k > 0$, then $A_{\mathrm{H}}^2 = P\,\mathrm{diag}(\mu_k^2)P^{-1}$. Therefore $\tilde{x} = P^{-1}x$ satisfies $\tilde{x}'' = +\,\mathrm{diag}(\mu_k^2)\tilde{x} + (\text{forcing})$, whose homogeneous solutions contain $e^{\pm\mu_k t}$ and are therefore non-oscillatory. No nonnegative diagonal A in the oscillator system reproduces these dynamics exactly. Route 2 instead uses $\lambda_k^{\mathrm{FA}} = \mu_k^2$ as a positive surrogate that preserves the characteristic modal scale. For LegS, A_{H} is lower triangular with distinct diagonal entries $k+1$, hence diagonalisable with $\mu_k = k+1$, giving $\lambda_k^{\mathrm{FA}} = (k+1)^2$, whereas Route 1 gives $\lambda_k^{\mathrm{SL}} = k(k+1)$.

(c) For a defective generator, the eigenvalues do not provide a complete modal description. In particular, the Chebyshev generator is nilpotent and therefore has only zero eigenvalues despite having nontrivial dynamics, while the Laguerre generator is defective with repeated eigenvalue 1. Since $A_{\mathrm{H}}^\top A_{\mathrm{H}}$ is always symmetric positive semidefinite, its eigenvalues are the squared singular values $\varsigma_k(A_{\mathrm{H}})^2$. These values are therefore used as the positive surrogate spectrum. □

Basis by basis, and three senses of a diagonal reduction. The four bases illustrate the three cases described above. For Legendre, the LegS generator is lower triangular with distinct diagonal entries $k+1$, so its eigenvalues are $\mu_k = k+1$. The generator is therefore diagonalisable, but its eigenvalues are real rather than imaginary. After differentiating the memory law, the homogeneous modes satisfy $\tilde{x}_k'' = \mu_k^2\tilde{x}_k$ and are exponential rather than oscillatory. Route 2 therefore replaces the exact values $-\mu_k^2$ by the positive surrogate $\lambda_k^{\mathrm{FA}} = (k+1)^2$. This preserves the modal scale but changes the dynamics, and differs from the Route 1 spectrum $\lambda_k^{\mathrm{SL}} = k(k+1)$, although the two become close for large k.

For Chebyshev, the generator is strictly lower triangular and hence nilpotent. All of its eigenvalues are therefore zero, even though the generator itself produces nontrivial dynamics. The eigenvalues consequently provide no useful frequency spread, so Route 2 uses the squared singular values $\varsigma_k(A_{\mathrm{H}})^2$ instead. Laguerre is similar in this respect: its all-ones lower-triangular generator is defective, with eigenvalue 1 of full algebraic multiplicity, and therefore has no complete eigenbasis. Its Route 2 spectrum is likewise obtained from $\varsigma_k(A_{\mathrm{H}})^2$ rather than from an eigendecomposition. The Laguerre memory law is intrinsically dissipative, so the numerically damped IM discretisation is more naturally aligned with its behaviour than the conservative IMEX scheme.

For the translating Hermite basis, $A_{\mathrm{H}} = \mathcal{D}$ is skew-symmetric, so $-\mathcal{D}^2$ is symmetric positive semidefinite and the homogeneous second-order reduction is exact. Its spectrum nevertheless differs

from the Route 1 values $2k+1$: Route 2 squares only the derivative generator, whereas Route 1 uses the full Hermite operator $-\partial_\xi^2 + \xi^2$. The missing potential term ξ^2 is therefore the reason the two Hermite constructions produce different spectra.

These examples give three distinct relations between the dense generator A_{H} and the diagonal oscillator matrix A. First, when A_{H} has a purely imaginary spectrum, as in the skew-symmetric case, the homogeneous second-order dynamics reduce exactly to oscillators with $\lambda_k^{\mathrm{FA}} = |\mu_k|^2$. Second, when A_{H} is diagonalisable but its spectrum is not purely imaginary, as for LegS, the change to eigencoordinates is exact but the replacement $-\mu_k^2 \mapsto |\mu_k|^2$ is a surrogate that changes the dynamics. Third, when A_{H} is defective, as for Laguerre and Chebyshev, no exact diagonal eigenbasis exists and the singular-value spectrum is used as a positive replacement. In all cases the diagonal form is computationally attractive because the recurrence and the contraction $S = (1 + \Delta t^2 A)^{-1}$ become elementwise; the learned B and C then absorb the input and output mixing that is not represented inside the diagonal recurrence.

D Theory and proofs

D.1 Modal solution and forced-response bound

Lemma 1 (Exact modal solution). *Mode k of the oscillator system, $x_k'' + A_k x_k = b_k^\top u$ with $A_k > 0$, $\omega_k = \sqrt{A_k}$ and initial position and velocity $x_k(0)$, $z_k(0)$, has the solution*

$$x_k(t) = x_k(0)\cos(\omega_k t) + \frac{z_k(0)}{\omega_k}\sin(\omega_k t) + \frac{1}{\omega_k}\int_0^t \sin\big(\omega_k(t-\tau)\big)\, b_k^\top u(\tau)\, d\tau. \tag{20}$$

For zero initial conditions and every $t \in [0,T]$, $|x_k(t)| \le \|b_k\|_2 \, \|u\|_{L^\infty(0,T)} \min\{t^2/2,\, t/\omega_k\}$ and $|z_k(t)| \le \|b_k\|_2 \, t \, \|u\|_{L^\infty(0,T)}$.

Proof. The homogeneous equation has the fundamental solutions $\cos(\omega_k t)$ and $\sin(\omega_k t)$, and variation of constants gives (20). With zero initial conditions only the integral remains, and $|\sin r| \le \min\{|r|, 1\}$ gives $|x_k(t)| \le \omega_k^{-1}\|b_k\|_2\|u\|_{L^\infty}\int_0^t \min\{\omega_k(t-\tau), 1\}\, d\tau \le \|b_k\|_2\|u\|_{L^\infty}\min\{t^2/2,\, t/\omega_k\}$. Differentiating (20), $z_k(t) = \int_0^t \cos(\omega_k(t-\tau))\, b_k^\top u(\tau)\, d\tau$, and $|\cos r| \le 1$ gives the second bound. □

The integral in (20) is $\omega_k^{-1}\mathcal{L}_t[b_k^\top u](\omega_k)$ with $\mathcal{L}_t f(\omega) = \int_0^t f(t-\tau)\sin(\omega\tau)\, d\tau$ the windowed sine transform, the form used in Appendix D.5.

Proposition 2 (Finite-horizon forced-response bound). *Let $x'' = -Ax + Bu$ with $A = \mathrm{diag}(A_0, \dots, A_{m-1})$, $A_k \ge a_{\min} > 0$, zero initial conditions and $\omega_{\min} = \sqrt{a_{\min}}$. For every $u \in L^\infty([0,T]; \mathbb{R}^H)$ and every $t \in [0,T]$,*

$$\|x(t)\|_2 \le \|B\|_2 \, \|u\|_{L^\infty(0,T)} \min\Big\{\frac{t^2}{2}, \frac{t}{\omega_{\min}}\Big\}, \qquad \|z(t)\|_2 \le t\, \|B\|_2 \, \|u\|_{L^\infty(0,T)}, \tag{21}$$

so $\sup_{t \le T}\|x(t)\|_2 \le \|B\|_2\|u\|_{L^\infty(0,T)} \min\{T^2/2,\, T/\sqrt{a_{\min}}\}$: the forced response is bounded on every finite horizon, depends on the spectrum only through its smallest entry, and does not depend on m.

Proof. A is diagonal, so by Lemma 1 $x(t) = \int_0^t \mathrm{diag}\big(\sin(\omega_k(t-\tau))/\omega_k\big)\, Bu(\tau)\, d\tau$. At $s = t - \tau$ the diagonal matrix has spectral norm $\max_k |\sin(\omega_k s)|/\omega_k \le \min\{s, 1/\omega_{\min}\}$, since $|\sin(\omega_k s)| \le \min\{\omega_k s, 1\}$ and $\omega_k \ge \omega_{\min}$; integrating $\min\{t-\tau, 1/\omega_{\min}\}$ over $[0,t]$ gives the first bound. For the velocity, $z(t) = \int_0^t \mathrm{diag}(\cos(\omega_k(t-\tau)))\, Bu(\tau)\, d\tau$ and the diagonal matrix has norm at most one. □

D.2 Proof of Theorem 1

For $\omega \ge \omega_{\min} > 0$ and $t \in [0,T]$ the integrand $\sin(\omega t)/\omega$ of (4) is bounded by $\min\{t, 1/\omega_{\min}\} \le \min\{T, 1/\omega_{\min}\}$, and its ω-derivative $(\omega t\cos(\omega t) - \sin(\omega t))/\omega^2$ by $t/\omega_{\min} + 1/\omega_{\min}^2 \le T/\omega_{\min} +$

$1/\omega_{\min}^2$. Hence for $\omega, \omega_k \in \mathcal{I}_k$,

$$\Big|\nu(\omega)\frac{\sin\omega t}{\omega} - \nu(\omega_k)\frac{\sin\omega_k t}{\omega_k}\Big| \le |\nu(\omega)-\nu(\omega_k)|\,\Big|\frac{\sin\omega t}{\omega}\Big| + |\nu(\omega_k)|\,\Big|\frac{\sin\omega t}{\omega} - \frac{\sin\omega_k t}{\omega_k}\Big|$$
$$\le \Big[L_\nu \min\{T, \tfrac{1}{\omega_{\min}}\} + \bar\nu\big(\tfrac{T}{\omega_{\min}} + \tfrac{1}{\omega_{\min}^2}\big)\Big]\,|\omega-\omega_k|,$$

and $|\omega - \omega_k| \le h_m$ on the cell. Integrating over $\mathcal{I}_k$ bounds the error of cell k by $h_m|\mathcal{I}_k|$ times the bracket; summing over k with $\sum_k |\mathcal{I}_k| = \omega_{\max} - \omega_{\min}$ and taking the supremum over $t \in [0,T]$ gives (5). □

Corollary 1 (Spacing of the normalised SL spectra)**.** *After (6) and (7) the spectrum $\omega_k = \sqrt{\lambda_k/\lambda_{m-1}}$, $k = 0,\dots,m-1$, of a classical basis has largest cell $h_m = O(m^{-1/2})$ for Hermite and Laguerre and $h_m = O(m^{-1})$ for Fourier, Legendre and Chebyshev, so by Theorem 1 the spectral discretisation error is $O(m^{-1/2})$ for the former and $O(m^{-1})$ for the latter.*

Proof. Take the cells as the midpoint intervals between consecutive frequencies, the end cells closed at $\omega_{\min}$ and $\omega_{\max}$; then h_m is at most the largest consecutive gap plus the width of the lowest cell. For Fourier and Chebyshev, $\omega_k = k/(m-1)$ and every gap is $1/(m-1)$. For Legendre, $\omega_k = \sqrt{k(k+1)/(m(m-1))}$ and $\sqrt{k(k+1)} = (k+\frac{1}{2})\sqrt{1-(2k+1)^{-2}}$ lies within $1/(8k)$ of $k+\frac{1}{2}$, so every gap is $1/\sqrt{m(m-1)}$ up to $O(k^{-2}m^{-1})$, all $O(m^{-1})$. For Hermite, $\omega_k = \sqrt{(2k+1)/(2m-1)}$ and the largest gap is the first, $(\sqrt{3}-1)/\sqrt{2m-1} = O(m^{-1/2})$; for Laguerre, $\omega_k = \sqrt{k/(m-1)}$ and the gap from the floored zero mode to $\omega_1 = 1/\sqrt{m-1}$ is $O(m^{-1/2})$. In every case the lowest cell has width at most ω_1, of the same order, and modes that the floor collapses onto $\omega_{\min}$ enlarge no cell. □

D.3 Proof of Theorem 2

The quadrature kernel $\mathcal{K}_\nu^m$ of (4) is a linear combination of the functions $\sin(\omega_k t)/\omega_k$, hence a member of $\mathfrak{K}_m$ by construction, and Theorem 1 bounds $\sup_{[0,T]}|\mathcal{K}_\nu - \mathcal{K}_\nu^m|$ by a constant times h_m. If $h_m \to 0$ the bound tends to zero, so $\mathcal{K}_\nu$ is the uniform limit of $(\mathcal{K}_\nu^m)_{m\ge1}$ and lies in the closure of $\bigcup_m \mathfrak{K}_m$; since ν was an arbitrary Lipschitz density, $\mathfrak{K} \subseteq \overline{\bigcup_m \mathfrak{K}_m}$. □

The theorem characterises the spectral expressivity of the finite collection of oscillators independently of the learned realisation: more modes enrich $\mathfrak{K}_m$, and refinement of the spectrum guarantees convergence to $\mathfrak{K}$ on every finite horizon. It does not assert universal approximation of arbitrary sequence maps, only of the memory kernels induced by Lipschitz densities on the band; the next subsection separates this capability from the learned parameters and the discretisation.

D.4 The learned discrete kernel and the error decomposition

The trained layer is a finite discrete-time collection of oscillators whose mode k has $A_k = a_{\min} + \hat A_k^2$, $\omega_k = \sqrt{A_k} \ge \omega_{\min}$, a learned step $\Delta t_k = \sigma(\mathrm{steps}_k) \in (0,1)$ and complex B, C with the real readout; let $\Theta = (\hat A, \mathrm{steps}, B, C)$ collect the trainable parameters. Under IM or IMEX the per-mode state $s_{k,n} = (z_{k,n}, x_{k,n})^\top$ evolves as $s_{k,n} = M_k s_{k,n-1} + (c_{z,k}, c_{x,k})^\top b_k^\top u_n$ with the transition and forcing scales of (28). With $e_x = (0,1)^\top$ selecting the position, the contribution of mode k to the output at lag $\ell \ge 1$ and the kernel of the layer are

$$\mathcal{K}^\Theta_{k,\ell} = C_{:,k}\, e_x^\top M_k^{\ell-1}(c_{z,k}, c_{x,k})^\top b_k^\top, \qquad \mathcal{K}^\Theta_{m,\ell} = \mathrm{Re}\sum_{k=0}^{m-1} \mathcal{K}^\Theta_{k,\ell}, \tag{22}$$

with $C_{:,k}$ the k-th column of C. Fix a reference lag grid $\tau_j = (j-1)\Delta t_{\mathrm{ref}}$, $j = 1,\dots,T$, with $\Delta t_{\mathrm{ref}} > 0$.

Proposition 3 (Decomposition of spectral and realisation error)**.** *For the continuous kernel $\mathcal{K}_\nu$, its quadrature $\mathcal{K}_\nu^m$ on the generated spectrum and the learned kernel (22),*

$$\max_{j\le T}\big|\mathcal{K}_\nu(\tau_j) - \mathcal{K}^\Theta_{m,j}\big| \le \underbrace{\max_{j\le T}\big|\mathcal{K}_\nu(\tau_j) - \mathcal{K}_\nu^m(\tau_j)\big|}_{\textit{spectral discretisation}} + \underbrace{\max_{j\le T}\big|\mathcal{K}_\nu^m(\tau_j) - \mathcal{K}^\Theta_{m,j}\big|}_{\textit{realisation mismatch}}, \tag{23}$$

and the first term is at most the right-hand side of (5).

Proof. The triangle inequality at each τ_j, then the maximum over j; since $\tau_j \in [0, T]$, Theorem 1 bounds the first term. □

Corollary 2 (Sequence-response error)**.** *For a bounded input $u_1, \dots, u_T$ with $\|u\|_\infty = \max_n |u_n|$, the responses $y^\nu_\ell = \sum_{n\le\ell} \mathcal{K}^m_\nu(\tau_{\ell-n+1})u_n$ of the quadrature kernel and $y^\Theta_\ell = \sum_{n\le\ell} \mathcal{K}^\Theta_{m,\ell-n+1}u_n$ of the learned layer satisfy* $\max_{\ell\le T} |y^\nu_\ell - y^\Theta_\ell| \le T \|u\|_\infty \max_{j\le T} |\mathcal{K}^m_\nu(\tau_j) - \mathcal{K}^\Theta_{m,j}|$.

Proof. Subtract the two sums, bound each of the at most T terms by $|u_n|$ times the largest kernel difference, and take the maximum over ℓ. □

The $O(h_m)$ rate is a property of the generated spectrum and of the continuous reference, not a convergence rate of the trained model: the learned $\hat{A}$, Δt_k, B, C and the choice of IM or IMEX enter only the second term of (23), which measures how far the trained discrete realisation sits from the quadrature kernel at the reference lags, and by Corollary 2 that term controls the discrepancy of the two responses.

D.5 Frozen spectra attain the free-frequency rate

The following result, in the setting of the neural-oscillator approximation bounds, shows that fixing the frequencies at a harmonic-type generated spectrum costs nothing in approximation rate against a layer whose frequencies are free, and that any positive spectrum is reachable through the learned per-mode timestep. Throughout, $u \in K \subset C^0([0,T];\mathbb{R}^H)$ is the input of a layer and Φ the target causal operator. **Assumptions. (F1)** K is compact, $\|u\|_{L^\infty} \le B_K$ and $|u(t_1) - u(t_2)| \le L_K|t_1 - t_2|$ for all $u \in K$. **(F2)** $\Phi : C^0([0,T];\mathbb{R}^H) \to C^0([0,T];\mathbb{R}^q)$ is causal and uniformly continuous with modulus ϕ_Φ, and satisfies the regularity conditions of Huang et al. (2025, Assumptions 3–6). **(F3)** (harmonic-type generated spectrum) the layer instantiates each of the m frequencies $\omega^\natural_n = (n+\delta)\pi/(2T)$, $n = 1, \dots, m$, $\delta \in \{0, \frac{1}{2}\}$, once per input channel, so its state has dimension mH. The normalised Fourier and Chebyshev SL spectra and the HiPPO-Legendre spectrum $(n+1)^2$ are of this type with $\delta = 0$ (the floored zero mode is dropped); the Legendre SL spectrum $\sqrt{n(n+1)}$ deviates from the $\delta = \frac{1}{2}$ spectrum by at most $\pi/(16nT)$.

Theorem 3 (Frozen generated spectra; proof in Appendix D.8)**.** *Under (F1)–(F3), for every admissible $m \ge 3$ and every depth parameter $H_\Pi \ge 1$ there exist an input map B and a rectified-linear (ReLU) readout multilayer perceptron (MLP) Π of depth $H_\Pi + 1$ and width $q(H(m+1)+4)$ acting on $(x(t), u(0), t)$ such that for all $u \in K$ and $t \in [0,T]$*

$$\big|\Phi[u](t) - \Pi\big(x(t), u(0), t\big)\big| \;\le\; \phi_\Phi\Big(\tfrac{32\, c_K H L_K T(\ln m)^2}{m}\Big) + \varepsilon_\Pi(m, H_\Pi), \tag{24}$$

with the frequencies fixed at the generated spectrum; $c_K > 1$ is the constant and ε_Π readout term, which decays as $H_\Pi^{-1/(H(m+1)+1)}$. Moreover:

1. *(spectrum deviation) if the frequencies are $\tilde{\omega}_n$ with $|\tilde{\omega}_n - \omega^\natural_n| \le \Delta_n$, the argument of ϕ_Φ in (24) increases by at most $4HB_KT\sum_{n=1}^m \Delta_n$;*
2. *(learned timesteps absorb the spectrum) for any spectrum $\tilde{\omega}_n$ with distinct positive entries and $\omega^\natural_n \le \tilde{\omega}_n$, the per-mode steps $\Delta t_n = \omega^\natural_n/\tilde{\omega}_n$ make the discrete transitions of the two spectra identical up to a rescaling of B; hence with learned Δt_k the spectrum shape constrains initialisation, not expressivity.*

D.6 Sine-series approximation on harmonic-type spectra

We restate the key lemma of Huang et al. (2025, Lemma 1) and note that its frequencies are generated, then extend it to the half-integer spectrum.

Lemma 2 (after Lemma 1 of Huang et al. (2025))**.** *Let K satisfy (F1) and let $\delta \in \{0, \frac{1}{2}\}$, $\omega^\natural_n = (n+\delta)\pi/(2T)$. For every $0 < \upsilon < T$ and every integer $m > 2$ there exist weights $\varkappa_n$ with*

$|\varkappa_n| \le 2/T$ *and phases* ϑ_n*, independent of* u*, such that for all* $t \in [0,T]$*,* $\tau \in [0,t]$ *and* $u \in K$*,*

$$\Big|u(t-\tau) - u(0) - \sum_{n=1}^{m} \varkappa_n \Big\{ \mathcal{L}_t u(\omega_n^\natural) + \tfrac{u(0)}{\omega_n^\natural}\big[\cos(\omega_n^\natural t) - 1\big] \Big\} \sin(\omega_n^\natural \tau - \vartheta_n) \Big| \le H L_K \upsilon + \frac{c_K^{p+1} H L_K T^{p+1} (p!)^2 \ln m}{\upsilon^p m^p} \tag{25}$$

for every positive integer p*; optimising* υ *and choosing* $p = \lfloor \ln m \rfloor$ *gives the right-hand side* $32 c_K H L_K T (\ln m)^2/m$ *for all admissible* m *(their eq. 15).*

Proof. Case $\delta = 0$ is their Lemma 1 with H channels and $\phi_K(t) = L_K t$: the odd extension $\tilde{f}_t(\tau)$ of $u(t-\tau) - u(0)$ to $[-t,t]$ is mollified with a Gevrey-class bump ρ_υ of width υ, the mollification $\tilde{f}_\upsilon$ is supported in $(-2T, 2T)$ and expanded in its Fourier series on $[-2T, 2T]$, whose frequencies are exactly $\omega_n = n\pi/(2T)$ (their eq. A.7); the Jackson-type bound of the truncated series gives the second term and the mollification error the first; finally each Fourier term is rewritten through $\mathcal{H}\tilde{f}_t(\omega) = -2i\{\mathcal{L}_t u(\omega) + u(0)[\cos \omega t - 1]/\omega\}$ with $|\varkappa_n| \le 1/T$ (their eq. A.13). For $\delta = \frac{1}{2}$ extend $\tilde{f}_\upsilon$, which vanishes on a neighbourhood of $\pm 2T$, evenly about $\tau = 2T$ (and about $\tau = -2T$) to an $8T$-periodic function $\bar{f}_\upsilon$. Being odd about 0 and even about $2T$, $\bar{f}_\upsilon$ has a Fourier series on $[-4T, 4T]$ containing only the odd harmonics $\sin((2n+1)\pi\tau/(4T)) = \sin((n+\frac{1}{2})\pi\tau/(2T))$; the reflection preserves the Gevrey bounds of the derivatives because $\tilde{f}_\upsilon$ is zero near the reflection points, so the Jackson-type bound applies with the period doubled, which at most doubles the constants ($|\varkappa_n| \le 2/T$). On $[-2T, 2T]$ the extension coincides with $\tilde{f}_\upsilon$, and the rewriting through $\mathcal{L}_t u$ is unchanged since it only uses that $\tilde{f}_t$ is odd and supported in $[-t, t]$. □

D.7 Frequency perturbation

Lemma 3. *Let* $|\tilde{\omega} - \omega| \le \Delta$ *and* $\|u\|_{L^\infty} \le B_K$ *with* H *channels. Then for* $t \le T$ *and* $\tau \le t$*: (i)* $|\mathcal{L}_t u(\tilde{\omega}) - \mathcal{L}_t u(\omega)| \le H B_K T^2 \Delta/2$*; (ii)* $\big|\frac{\cos \tilde{\omega} t - 1}{\tilde{\omega}} - \frac{\cos \omega t - 1}{\omega}\big| \le \frac{3}{2} T^2 \Delta$*; (iii)* $|\sin(\tilde{\omega}\tau - \vartheta) - \sin(\omega\tau - \vartheta)| \le T\Delta$*; (iv)* $|\mathcal{L}_t u(\omega)| \le H B_K T$ *and* $|u(0)[\cos \omega t - 1]/\omega| \le H B_K T$.

Proof. (i) $|\sin \tilde{\omega}\tau - \sin \omega\tau| \le \tau\Delta$ and $\int_0^t \tau \, d\tau = t^2/2$. (ii) With $g(\omega) = (\cos \omega t - 1)/\omega$, $g'(\omega) = -t \sin(\omega t)/\omega - (\cos \omega t - 1)/\omega^2$, and $|\sin \xi| \le |\xi|$, $|1 - \cos \xi| \le \xi^2/2$ give $|g'| \le t^2 + t^2/2$. (iii) is the Lipschitz bound of the sine. (iv) $|\sin| \le 1$ and $|1 - \cos x| \le |x|$. □

D.8 Proof of Theorem 3

Main statement. By Lemma 1, an oscillator layer whose mH modes carry the frequencies $\omega_n^\natural$, one per input channel, with B selecting the channel ($b_{(n,h)} = \omega_n^\natural e_h$), has state $x_{(n,h)}(t) = \mathcal{L}_t u_h(\omega_n^\natural)$, i.e. the vector of windowed sine transforms of the input at the generated spectrum. Lemma 2 shows that $u(t-\tau)$ on $[0,t]$ is determined up to $\varepsilon_K := 32 c_K H L_K T (\ln m)^2/m$ by $(x(t), u(0), t)$ through a fixed continuous map, so by causality and uniform continuity of Φ there is a continuous Ψ with $|\Phi[u](t) - \Psi(x(t), u(0), t)| \le \phi_\Phi(\varepsilon_K)$, which is Lemma 3 of Huang et al. (2025) with the frequencies fixed. Their Lemma 4 bounds the Lipschitz constants and the compact domain of Ψ (the arguments live in $[-TB_K, TB_K]^{mH} \times [-B_K, B_K]^H \times [0,T]$), and their Theorem 1 approximates Ψ by a ReLU MLP Π of depth $H_\Pi + 1$ and width $q(H(m+1)+4)$ using Hanin (2019), with error $\varepsilon_\Pi(m, H_\Pi) = B_\Omega \big(H(m+1)+1\big)^2 (0.5 H_\Pi)^{-1/(H(m+1)+1)} \sum_{j=1}^{q} L_{\psi_j}(m)$, $B_\Omega = 2\max(TB_K, B_K, T)$. The triangle inequality gives (24). The only change to their argument is that the frequencies are not chosen in the proof but supplied by the spectrum, which Lemma 2 allows for $\delta \in \{0, \frac{1}{2}\}$; the normalised Fourier and Chebyshev SL spectra, $\omega_n \propto n$, and the HiPPO-Legendre spectrum, $\omega_n = n + 1$, are of this form after the frequency scale $\pi/(2T)$ is fixed by (6) and Δt, with the floored zero mode unused.

Part 1 (spectrum deviation). Replace $\omega_n^\natural$ by $\tilde{\omega}_n$ in every term of the approximant of Lemma 2. By Lemma 3 the n-th term changes by at most $|\varkappa_n|\big[H B_K T^2 \Delta_n/2 + \frac{3}{2} H B_K T^2 \Delta_n + 2 H B_K T \cdot T\Delta_n\big] = 4|\varkappa_n| H B_K T^2 \Delta_n \le 4 H B_K T \Delta_n$ using $|\varkappa_n| \le 1/T$ (for $\delta = \frac{1}{2}$ the constant is 8). Summing over n adds $4 H B_K T \sum_n \Delta_n$ to ε_K, and the rest of the proof is unchanged. For the

Legendre SL spectrum at the harmonic frequency scale, $\sqrt{n(n+1)} = (n+\frac{1}{2})\sqrt{1-(2n+1)^{-2}}$, so $0 \le (n+\frac{1}{2}) - \sqrt{n(n+1)} \le \frac{1}{8n}$ and $\Delta_n \le \pi/(16nT)$, giving a total deviation penalty of at most $(\pi/4)HB_K(1+\ln m)$ (with $\delta = \frac{1}{2}$ constants, $(\pi/2)HB_K(1+\ln m)$). This worst-case bound does not vanish with m; it is loose because it sums absolute perturbations, and Part 2 shows the deviation can be removed exactly.

Part 2 (learned timesteps). The IM and IMEX transitions of mode n depend on A_n and Δt_n only through $\Delta t_n^2 A_n = (\Delta t_n \tilde{\omega}_n)^2$ in S_n and the entries, and through Δt_n in the off-diagonal entries and the forcing scales $(c_z, c_x) = (\Delta t_n S_n, \Delta t_n^2 S_n)$ or $(\Delta t_n, \Delta t_n^2)$. With $\Delta t_n = \omega_n^\natural/\tilde{\omega}_n \in (0,1]$ we get $\Delta t_n \tilde{\omega}_n = \omega_n^\natural$, so S_n and the $(1,1),(2,2),(1,2)$ entries agree with those of the harmonic spectrum at unit step, while the $(2,1)$ entry and the forcing scales differ by the factor Δt_n; rescaling the n-th row of B by $1/\Delta t_n$ and the velocity coordinate by Δt_n (a similarity transform of the 2×2 state that leaves the readout of x unchanged) makes the two recurrences identical. Hence any function realised by the harmonic spectrum is realised by the spectrum $\tilde{\omega}_n$ with these timesteps and a rescaled B. □

D.9 Necessity of normalisation and floor

Proposition 4 (Necessity of normalisation and floor)**.** *Let $\kappa_k := \Delta t_k^2 A_k$.* ***(a)*** *Under IM, $\|(M_k^{\mathrm{IM}})^\ell\|_2 \le c_{\kappa_k} r_k^\ell$ with $r_k = (1+\kappa_k)^{-1/2}$, so the state at step n depends on the input at lag ℓ through a factor at most $c_{\kappa_k}\|B_k\| r_k^\ell$: the memory horizon at tolerance ϵ is $\ell_\epsilon(k) = 2\ln(1/\epsilon)/\ln(1+\kappa_k)$. For the raw Hermite spectrum at $\Delta t = 1$, $\ell_\epsilon(k) \le 2\ln(1/\epsilon)/\ln 2$ for every k; after the normalisation the slowest mode has $\ell_\epsilon \ge 2\ln(1/\epsilon)\,(2m-1)\,(1-\frac{1}{2(2m-1)})$.* ***(b)*** *Under IMEX, $|\lambda(M_k^{\mathrm{IMEX}})| = 1$ iff $\kappa_k < 4$, otherwise $\max|\lambda| = |1-\kappa_k/2| + \sqrt{\kappa_k^2/4-\kappa_k} > 1$; the raw Hermite spectrum violates this for $m \ge 3$ at $\Delta t = 1$ and every normalised spectrum satisfies it for every $\Delta t_k \in (0,1)$.* ***(c)*** *If $A_k = 0$, both schemes reduce to a double integrator and, for i.i.d. zero-mean forcing of variance σ_f^2, $\mathbb{E}\,x_T^2 \ge \sigma_f^2 \Delta t^4 T^3/3$, whereas for $A_k \ge a_{\min} > 0$ under IM $\sup_T \mathbb{E}\,x_T^2 < \infty$.*

Proof. **(a)** From Appendix B.3, M_k^{IM} has eigenvalues $\lambda_\pm = S_k(1 \pm i\Delta t_k\sqrt{A_k})$ with $|\lambda_\pm| = \sqrt{S_k} = r_k$. For $\kappa_k > 0$ the eigenvalues are distinct, so $M_k^{\mathrm{IM}} = P\,\mathrm{diag}(\lambda_+,\lambda_-)P^{-1}$ and $\|(M_k^{\mathrm{IM}})^\ell\|_2 \le \|P\|_2\|P^{-1}\|_2\, r_k^\ell =: c_{\kappa_k} r_k^\ell$, where c_{κ_k} depends continuously on κ_k and on Δt_k (it is the condition number of the eigenvector matrix, bounded on compact subsets of $\kappa_k > 0$). The state at step n is $s_n = \sum_{\ell=0}^{n-1}(M_k^{\mathrm{IM}})^\ell F_{n-\ell}$, so the input at lag ℓ enters through $(M_k^{\mathrm{IM}})^\ell (c_z, c_x)^\top b_k^\top u_{n-\ell}$, bounded by $c_{\kappa_k} r_k^\ell \,\|(c_z,c_x)\|\,\|b_k\|\,|u_{n-\ell}|$. Setting $r_k^\ell = \epsilon$ gives $\ell_\epsilon(k) = \ln(1/\epsilon)/\ln(1/r_k) = 2\ln(1/\epsilon)/\ln(1+\kappa_k)$. For the raw Hermite spectrum at $\Delta t = 1$, $\kappa_k = 2k+1 \ge 1$ and $\ln(1+\kappa_k) \ge \ln 2$. After (6), the slowest mode has $\kappa = 1/(2m-1)$ and $\ln(1+\kappa) \le \kappa(1-\kappa/2)^{-1}$, hence $\ell_\epsilon \ge 2\ln(1/\epsilon)(2m-1)(1-\frac{1}{2(2m-1)})$. **(b)** is the IMEX eigenvalue computation of Appendix B.3: for $\kappa \ge 4$ the eigenvalues are $1-\kappa/2 \pm \sqrt{\kappa^2/4-\kappa}$, real, with product one, so the larger modulus is $|1-\kappa/2| + \sqrt{\kappa^2/4-\kappa} \ge 1$ with equality only at $\kappa = 4$. For the raw Hermite spectrum $\kappa_{\max} = 2m-1 \ge 4$ iff $m \ge 3$; after normalisation $\kappa_k \le \Delta t_k^2 \le 1$. **(c)** With $A_k = 0$ both schemes have $S = 1$ and reduce to $z_n = z_{n-1} + \Delta t f_n$, $x_n = x_{n-1} + \Delta t z_{n-1} + \Delta t^2 f_n$, so $z_n = \Delta t \sum_{j\le n} f_j$ and $x_T = \Delta t^2 \sum_{n=1}^T \sum_{j\le n} f_j = \Delta t^2 \sum_{j=1}^T (T-j+1) f_j$. For i.i.d. zero-mean f_j of variance σ_f^2, $\mathbb{E}\,x_T^2 = \Delta t^4 \sigma_f^2 \sum_{j=1}^T j^2 = \Delta t^4 \sigma_f^2 T(T+1)(2T+1)/6 \ge \Delta t^4 \sigma_f^2 T^3/3$. For $A_k \ge a_{\min} > 0$ under IM, part (a) gives $\mathbb{E}\,x_T^2 \le \big(c_\kappa \|(c_z,c_x)\|\big)^2 \sigma_f^2 \sum_{\ell\ge 0} r_k^{2\ell} < \infty$ uniformly in T. The right panel of Figure 6 verifies both halves numerically. □

D.10 Spectrum geometry

Figure 5 shows the natural frequencies $\omega_k = \sqrt{A_k}$ for $m = 16$. The generated spectra are shown after the common normalisation (6) and floor (7), before training; the LinOSS random initialisation is included as a reference in its native scale. For the trainable generated variants, (8) starts exactly from this grid because $\hat{A}_{\mathrm{init}}$ is chosen so that $a_{\min} + \hat{A}_{\mathrm{init}}^2 = A_{\mathrm{grid}}$. Table 6 reports the smallest post-floor frequency $\omega_{\min}$, the largest midpoint-cell width h_m over the distinct post-floor frequencies on $[\omega_{\min}, \omega_{\max}]$, and the number of distinct frequencies; the interval from 0 to $\omega_{\min}$ is not part of this

partition. By Theorem 1, h_m controls the spectral discretisation error, so smaller cells correspond to a finer coverage of the represented frequency band. SL Fourier/Chebyshev and FA Legendre are harmonic, while SL Legendre is close to the half-harmonic grid; accordingly, these spectra have nearly uniform coverage at this resolution ($h_m = 0.062$–0.067). SL Hermite and SL Laguerre are more nonuniform near the low-frequency end, with $h_m = 0.111$ and 0.133, consistent with Corollary 1. FA Hermite contains only eight distinct frequencies because its conjugate eigenvalue pairs produce repeated magnitudes. For FA Laguerre and FA Chebyshev, many modes are collapsed by the floor, leaving only six and nine distinct frequencies and substantially larger cells ($h_m = 0.399$ and 0.374); Corollary 1 concerns the SL spectra and does not apply to these FA surrogates. The multi-spectrum and LinOSS initialisation are shown only as reference geometries. Harmonicity is reported separately because it is used by Theorem 3, whereas Theorem 1 depends on the band coverage through h_m; it is assessed on the normalised generated spectrum before flooring, since the floor can merge several low-frequency modes and therefore need not preserve m distinct frequencies.

The quantity h_m characterises the worst-case discretisation of a continuous kernel with a Lipschitz spectral density; it is not a ranking criterion for trained models. In particular, the learned B, C, and Δt_k, the discretisation, nonlinear residual stack, and task-dependent input spectrum are not represented in Theorem 1. Consequently, a basis with a larger h_m, such as Hermite, may still perform better on a particular task than a more uniformly spaced spectrum. Likewise, Hermite$\oplus$Fourier does not reduce h_m relative to Fourier alone at fixed state size, so its empirical benefit should be interpreted as access to complementary spectral allocations rather than as a consequence of a smaller discretisation bound.

Theorem 3 is therefore an expressivity statement rather than an optimisation result: learned per-mode timesteps can compensate for differences between positive spectra, so the empirical advantage of a generated spectrum should be interpreted as an initialisation and inductive-bias effect rather than as a restriction of the attainable function class.

Table 6: Geometry and theoretical scope of the generated spectra at $m = 16$ after the normalisation (6) and floor (7), in units of $\omega_{\max}$. $\omega_{\min}$ is the smallest post-floor frequency and h_m is the largest midpoint-cell width on $[\omega_{\min}, \omega_{\max}]$. A smaller h_m gives a tighter worst-case spectral-discretisation bound in Theorem 1, but is not intended as a predictor of downstream task performance. The asymptotic rates shown for the SL spectra are those of Corollary 2; for the FA, multi-basis, and random spectra, h_m is reported only as a finite-m geometric diagnostic unless stated otherwise.

spectrum	$\omega_{\min}$	h_m	# distinct	Thm. 1 / spacing	Thm. 3
SL Hermite $2k{+}1$	0.180	0.111	16	$O(m^{-1/2})$	not harmonic
SL Laguerre k	0.100	0.133	16	$O(m^{-1/2})$	not harmonic
SL Legendre $k(k{+}1)$	0.100	0.065	15	$O(m^{-1})$	near $\delta = \frac{1}{2}$
SL Fourier / Chebyshev k^2	0.100	0.067	15	$O(m^{-1})$	exact, $\delta = 0$
FA Hermite $\lvert\mathrm{eig}(\mathcal{D})\rvert^2$	0.100	0.161	8	finite-m	not harmonic
FA Legendre $(k{+}1)^2$	0.100	0.062	16	finite-m	exact, $\delta = 0$
FA Laguerre $\varsigma_k(A_{\mathrm{H}})^2$	0.100	0.399	6	finite-m	not harmonic
FA Chebyshev $\varsigma_k(A_{\mathrm{H}})^2$	0.100	0.374	9	finite-m	not harmonic
Multi Hermite$\oplus$Fourier	0.100	0.085	15	finite-m	mixed
LinOSS $a_{\min} + \hat{A}^2$, $\hat{A} \sim \mathcal{U}[0,1]$	0.102	0.130	16	reference only	not covered

E Extended related work

HiPPO, S4 and diagonal SSMs. HiPPO derived the LegS, LagT and Fourier memory operators as the dynamics that keep the state equal to a projection of the input history onto an orthogonal basis; S4 (Gu et al., 2022a) made the LegS SSM trainable at scale with a normal-plus-low-rank parameterisation and a convolutional evaluation; DSS (Gupta et al., 2022) and S4D showed that the diagonal part suffices, with the S4D-Lin initialisation $\lambda_n = -\frac{1}{2} + i\pi n$ and S4D-Inv; Gu et al. (2023) interpreted these through generalised orthogonal-basis projections, including a truncated Fourier basis. S5 (Smith et al., 2023) moved to a single multi-input SSM with a parallel scan; LRU (Orvieto et al., 2023) showed that a complex diagonal linear RNN with a ring initialisation suffices; H3, Hyena and Liquid-S4 (Fu et al., 2023; Poli et al., 2023; Hasani et al., 2023) extended SSMs towards language; Mamba and Mamba-2 (Gu & Dao, 2024; Dao & Gu, 2024) made the transition input-

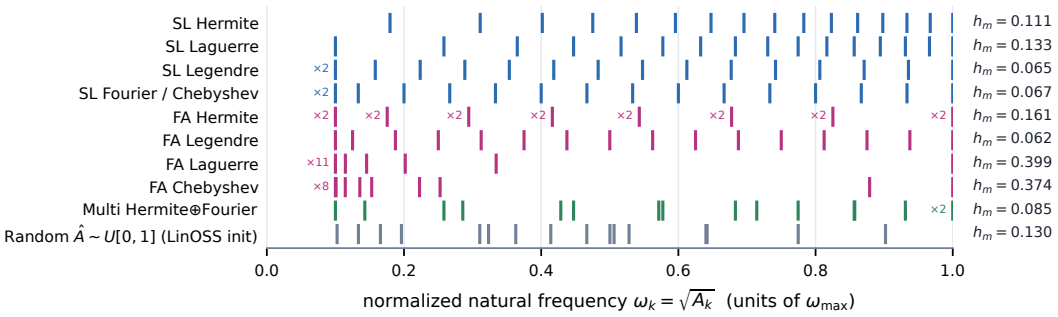


Figure 5: Normalised frequency geometry at $m = 16$. Each marker is a generated natural frequency $\omega_k = \sqrt{A_k}$ after the common normalisation (6) and floor (7), before training; $\times c$ marks a frequency carried by c modes, either modes collapsed onto the floor or the conjugate pairs of FA Hermite. The right margin reports the largest midpoint-cell width h_m between distinct frequencies; the interval from 0 to $\omega_{\min}$ is excluded. SL Fourier/Chebyshev and FA Legendre are harmonic and SL Legendre is nearly half-harmonic; SL Hermite and SL Laguerre widen towards low frequencies; FA Laguerre and FA Chebyshev leave one wide empty interval; the random LinOSS spectrum is not normalised, so its largest frequency lies below 1.

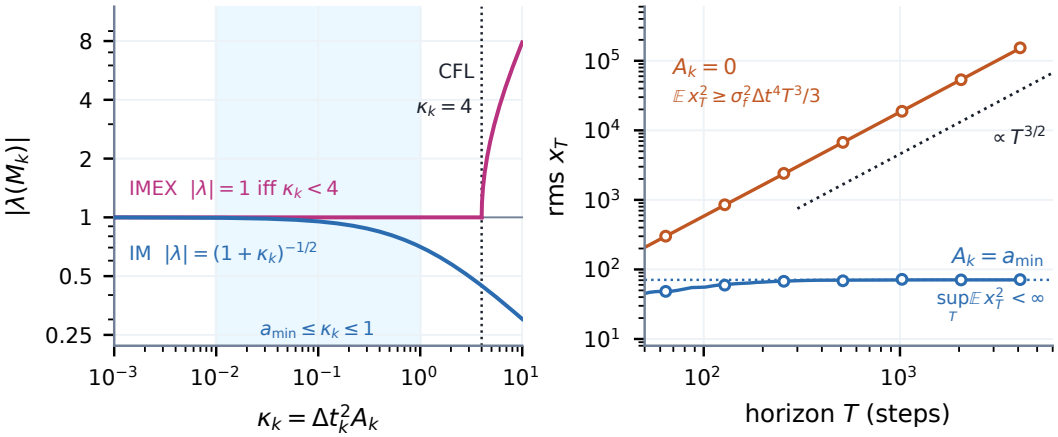


Figure 6: **Left:** spectral radius $|\lambda(M_k)|$ of the per-mode transition against $\kappa_k = \Delta t_k^2 A_k$, for the two schemes of Appendix B.3. IM contracts at $(1+\kappa_k)^{-1/2}$, which fixes its memory horizon (Proposition 4a); IMEX is marginal, $|\lambda| = 1$, up to the CFL threshold $\kappa_k = 4$ (dotted) and grows beyond it (Proposition 4b). The band is what a normalised spectrum spans at $\Delta t_k = 1$: from the floor $a_{\min}$ of (7) to 1 after (6). A smaller Δt_k only moves it left, so $\kappa_k \le \Delta t_k^2 \le 1$ and the CFL bound is never approached. **Right:** root-mean-square response x_T to unit-variance white noise under IM ($\Delta t = 1$), for a zero mode $A_k = 0$ and for a mode held at the floor $A_k = a_{\min}$ (Proposition 4c). The zero mode is a double integrator, so it grows like $T^{3/2}$ (dotted guide), whereas the floored mode saturates at the level $\sup_T \mathbb{E}\, x_T^2 < \infty$ (horizontal dotted); the same $a_{\min}$ bounds the continuous-time response in Proposition 2. Markers are 2000 simulated trajectories, lines the exact second moment of the recurrence.

dependent, and Griffin (De et al., 2024) mixed gated recurrences with local attention. All of these are first-order; the imaginary parts of their eigenvalues are frequencies but their real parts are decay rates that must be learned or constrained. LinOSS (Rusch & Rus, 2025) is second-order, so the entire state matrix is a spectrum of squared frequencies, which is what makes spectral generation natural.

Oscillatory networks and their theory. coRNN and UnICORNN (Rusch & Mishra, 2021a;b) introduced nonlinear coupled oscillators with provable gradient bounds, and neural wave ma-

chines (Keller & Welling, 2023) placed such oscillators on a grid in which each is coupled only to its neighbours through one shared convolution, so that travelling waves in the hidden state carry memory at few parameters; there the structure is imposed on the coupling between oscillators, whereas ours is imposed on the frequencies of an uncoupled diagonal. Lanthaler et al. (2023) proved that a linear second-order ODE followed by an MLP is a universal approximator of causal continuous operators via the time-varying sine transform, and Huang et al. (2025) gave quantitative rates, decomposing the error into a sine-transform truncation term and a readout-MLP term, and a separate polynomial rate for incrementally stable second-order systems. Neural oscillators have also been used to extend physics-informed models in time (Kapoor et al., 2024). D-LinOSS learns per-mode damping; a damped generated spectrum would take the frequency from the spectral theory and the damping from data, and is a natural extension of our work.

Approximation theory of linear recurrences. Li et al. (2022) characterised the approximation and optimisation of linear continuous-time RNNs and the curse of memory, and StableSSM (Wang & Li, 2023) showed that a freely trained recurrence learns stably only memory that decays exponentially, since longer memory drives its eigenvalues to the stability boundary where the gradients blow up, and removed that obstruction by computing each eigenvalue from its trainable weight through a map whose gradient bound grows only linearly in that weight; our generated spectra meet the same difficulty from the other side, by fixing where the frequencies sit rather than reparameterising how they move. Orvieto et al. (2024) proved finite-width universality of linear recurrences with nonlinear projections and the benefit of complex eigenvalues; Wang & Xue (2023) showed universality with exponentially decaying memory for layer-wise nonlinear SSMs; Muca Cirone et al. (2024) did so for selective SSMs. Fading-memory operators and Volterra series, and the simulation of stable dynamical systems by RNNs (Hanson & Raginsky, 2020), underlie these results.

Heterogeneous and multi-head SSMs. Multi-head SSMs with non-homogeneous initialisations (Fathullah et al., 2023) and the generalised bases of Gu et al. (2023) motivate unions of spectra; our multi-basis fusion differs in that the union is realised inside one diagonal matrix, so the scan cost is unchanged, and the heterogeneity is between spacings of the same band rather than between heads.

Sturm–Liouville structure in learned models. Deep Sturm–Liouville (Vigouroux et al., 2025) is a general function approximator rather than a sequence model: it learns a vector field, follows the field line through each input out to the boundary of the domain, solves a Sturm–Liouville eigenproblem with learned coefficients along that line, and predicts from a linear combination of the first few eigenfunctions evaluated at the input, which regularises the model along whole curves rather than at the training points alone. It is the closest use of Sturm–Liouville theory in deep learning we are aware of, and the contrast is the point: there the operator is learned and its eigenfunctions are the features, whereas here the operator is a classical one whose eigenvalues are known in closed form and enter a sequence model as the squared frequencies of an oscillator layer.

Benchmarks. The five long-range UEA tasks and the PPG-DaLiA protocol are those of Walker et al. (2024), who also introduced Log-NCDE and re-ran NRDE (Morrill et al., 2021), NCDE (Kidger et al., 2020), LRU, S5, Mamba and S6 under one tuning protocol; Rusch & Rus (2025) adopted the same protocol and reported the LinOSS rows we quote. The weather task, and with it the Informer, LogTrans (Li et al., 2019), Reformer (Kitaev et al., 2020), LSTMa (Bahdanau et al., 2015) and LSTnet (Lai et al., 2018) rows of Table 4, is from Zhou et al. (2021); WikiText-103 from Merity et al. (2017).

F Experimental details: architectures, optimisation and hyperparameter tuning

This section records, per experiment group, the data protocol, the architecture and its dimensions, the optimiser, and the hyperparameter search that produced every number of the experiments, so that each table can be reproduced from the configuration files of the supplementary code alone.

F.1 Datasets and protocols

UEA long-range classification. The five datasets, each given as (sequence length, channels, classes) – SCP1 (896, 6, 2), SCP2 (1,152, 7, 2), Ethanol (1,751, 3, 4), Heartbeat (405, 61, 2) and Motor (3,000, 64, 2) – come from the UEA archive (Bagnall et al., 2018). Labels are encoded, exact duplicates are removed, and a fixed random 70/15/15 split is drawn (boundaries $\lfloor 0.7n \rfloor$, $\lfloor 0.85n \rfloor$). No normalisation is applied. Test accuracy on the held-out split is reported.

PPG-DaLiA. Fifteen subjects' wrist recordings (blood volume pulse, electrodermal activity, temperature, three-axis acceleration; six channels min–max scaled to $[-1, 1]$) with the heart-rate target; per-subject temporal 70/15/15 split, then windows of length 49,920 with stride 4,992 against 390-step targets (stride 39), one readout every 128 input steps, following the protocol of Walker et al. (2024). Test MSE is reported on the scaled targets.

Weather (WTH, 720→720). The pipeline of Zhou et al. (2021) (12 variables, chronological 70/10/20 split, scaler fit on training rows) with 720 observed hourly steps followed by 720 zero-masked steps as input and the 720 future steps as target, in normalised space; the model input is $1{,}440 \times 12$ (Appendix F.5).

WikiText-103 and the Pile. Official WikiText-103 splits, GPT-NeoX tokenizer, end-of-text tokens between articles, 1,024-token packed blocks, one shared cache and data order for every variant; validation perplexity is token-weighted over the full validation split every 5M tokens.

PTB-XL electrocardiograms. The PhysioNet release of $21{,}799$ clinical twelve-lead recordings of ten seconds, taken at 100 Hz so that one record is $1{,}000$ steps on twelve channels. Each record carries the diagnostic statements of up to two cardiologists, which the standard aggregation maps onto five superclasses (normal, myocardial infarction, ST/T change, conduction disturbance, hypertrophy). We keep the $16{,}244$ records that carry exactly one superclass and predict it with five-way cross-entropy, so the $5{,}144$ records with several superclasses and the 411 with none are dropped. The dataset ships a patient-disjoint ten-fold assignment; we train on folds one to eight, select on fold nine and report fold ten ($12{,}957$ / $1{,}637$ / $1{,}650$ records).

F.2 Architectures and hyperparameters

Table 7 lists the configuration shared by every SpecOSS variant on a dataset: H is the residual block width, m the number of oscillator modes, L the depth in blocks, and B, C its input and output projections (Section 3). The complex B, C realisation (with the A-independent LinOSS initialisation, learned Δt_k, batch normalisation, dropout 0.05) is the default for every dataset; with real B the rows of B are whitened by $\sqrt{A_k}$ so that the per-mode impulse response, of magnitude $\|B_k\|/\omega_k$, is of order one for any spectrum.

Table 7: Per-dataset configuration of SpecOSS (identical for every generation route). H, F, Le, La: Hermite, Fourier, Legendre, Laguerre; Multi: the multi-basis union.

dataset	H	m	L	B, C	bases
Heartbeat	16	16	4	complex	H,F,Le,La,Multi
SCP1	32	32	3	complex	H,F,Le,La,Multi
SCP2	32	32	4	complex	H,F,Le,La,Multi
Ethanol	16	16	3	complex	H,F,Le,La,Multi
Motor	64	16	2	complex	H,F,Le,La,Multi
PPG-DaLiA	16	64	6	complex, chunk 12,480	H,F,Le,La,Multi
WTH 720→720	128	64	4	complex	H,F,Le,La,Multi
WikiText-103 / Pile	512	see Appendix F.6	12	complex	H,F,Le,La,Multi

F.3 Optimisation

The UEA classification runs (five datasets) and the PPG-DaLiA regression run minimise cross-entropy or mean-squared error respectively with Adam or AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) at learning rate 10^{-3} and no weight decay, a cosine schedule with 5% linear warm-up decaying to zero, gradient-norm clipping at 1, and batch size 32 (4 for PPG-DaLiA, 16 for Motor). Validation is run once per epoch (every 500 optimiser steps on PPG-DaLiA), the checkpoint with the best validation metric is kept, and test metrics are computed once on it. The electrocardiogram classification behind Example 1 and Table 17 uses the same optimiser at batch 64 for 40 epochs under a cosine schedule (Appendix H.5). The weather forecasting runs use Adam at a constant 10^{-3}, at most 20 epochs and early stopping with patience 5 on validation MSE. The language models use AdamW with learning rate 6×10^{-4}, moment coefficients $(0.9, 0.95)$, weight decay 0.1 (none on the recurrence parameters), 5% warm-up and cosine decay to 10% of the peak over a 100M-token schedule horizon on both corpora, 65,536 tokens per optimiser step, gradient clipping at 1 and bf16 autocast; the recipe is identical for every variant. Frozen spectra register A as a buffer; trainable spectra optimise $\hat{A}$ of (8) with the same optimiser and learning rate as every other parameter. Every reported mean, and every $\pm$ in this paper, is taken over five seeds under one protocol.

F.4 Hyperparameter tuning protocol

Classification and PPG-DaLiA. Two stages were run per dataset on the validation split at the architecture of Table 7. Stage A trains the full realisation grid, the three toggles {real, complex B, C} $\times$ {fixed, learned Δt_k} $\times$ {no, batch normalisation}, for every basis, both spectrum modes (frozen, trainable) and both discretisations, together with the random-A control. Stage B is a greedy coordinate search from the Stage A winner over $a_{\min} \in \{10^{-3}, 10^{-2}, 10^{-1}\}$, the input gain of B in $\{0.5, 1, 2\}$, dropout $\{0, 0.05, 0.1\}$, learning rate $\{10^{-3}, 3 \times 10^{-4}\}$ and depth $L \in \{2, 3, 4\}$, for at most two rounds. The winners were frozen into the per-dataset configurations of Table 7; the complex-B, learned-Δt, batch-normalised realisation with $a_{\min} = 10^{-2}$, unit input gain, dropout 0.05 and learning rate 10^{-3} won on every dataset, and Appendix J.1 reports the sensitivity to each toggle. The generation route, the basis, frozen against trainable and the discretisation are not tuned but enumerated, and the tables report the enumeration.

Weather. The core-swap backbone was fixed before any core was trained ($H = 128$, $m = 64$, $L = 4$). A search over width $\{128, 256\}$, state size $\{64, 128\}$, depth $\{4, 5, 8\}$, learning rate $\{10^{-3}, 7 \times 10^{-4}, 3 \times 10^{-4}\}$, dropout $\{0.05, 0.15\}$, schedule (constant, cosine) and an appended time channel was then run; its configurations are the rows of Table 20. The baselines are tuned under the same search.

Language models. The backbone architecture is shared across every model and every baseline; only the core differs between the state-space models.

Selection and evaluation. Every selection decision above – both tuning stages, the weather backbone search and the per-epoch checkpoint – is made on the validation split alone. The test split is evaluated once, on the selected configuration, and never enters tuning or model selection.

F.5 Weather backbone

Every core of Table 4 sits in the same residual block, a linear encoder from the 12 variables to $H = 128$, then $L = 4$ blocks of affine-free batch normalisation, the core, GELU, dropout 0.05, a gated linear unit, dropout and the skip, and a per-step linear head that reads the last 720 positions (Appendix G.3). The oscillator cores use $m = 64$ modes, $a_{\min} = 10^{-2}$, learned Δt_k and complex B, C; the LinOSS core is the faithful port of the official layer ($A = \mathrm{relu}(\hat{A})$ with $\hat{A} \sim \mathcal{U}[0, 1]$, learned Δt_k, complex B, C); Mamba is the official mixer with $d_{\mathrm{state}} = 64$ and its fused scan, and S5 has four blocks of shared state. The input is 720 standardised observed vectors followed by 720 zero vectors, without a mask channel or calendar covariates, and every core sees the identical optimiser and stopping rule of Appendix F.3.

F.6 LANGUAGE-MODEL SETUP

The backbone is the Mamba backbone (Gu & Dao, 2024) (12 layers, $d_{\text{model}} = 512$, expansion 2, convolution width 4, RMSNorm, tied embeddings), and every variant carries the same rank-16 gated linear unit on the mixer output. Our SpecOSS core takes the place of Mamba's selective scan as the sequence-mixing layer: the inner width $d_{\text{inner}} = 1{,}024$ is split into 128 groups, each with its own 36-mode state (18,432 state reals per layer), IMEX integration, log-spaced initial timesteps $\Delta t \in [10^{-3}, 10^{-1}]$, whitened B, $a_{\min} = 10^{-4}$, and $A = a_{\min} + \hat{A}^2$ (trainable) or a frozen buffer. Because the transition is input-independent the recurrence is a causal convolution and is evaluated with two real FFTs. Every baseline (LRU, S4D-Lin, S5 and Mamba) sits in the same backbone at its own reference hyperparameters. The optimiser, schedule and token budget are the same for every variant, tuned once for the shared backbone rather than per model, which is what keeps the comparison in Table 5 fair.

G DATA FLOW PER EXPERIMENT GROUP

The experiments fall into four groups, time-series classification, photoplethysmography regression, weather forecasting and language modelling. All four realise the same skeleton, and only the input $\mathbf{I}$ and the head change with the group,

$$\mathbf{I} \xrightarrow{\mathcal{E}} U^{(0)} \xrightarrow{\text{SpecOSS}} U^{(L)} \xrightarrow{\mathcal{R}} \hat{\mathbf{O}}, \tag{26}$$

with $\mathbf{I}$ the windowed input built from a raw record X by the data pipeline of the group, $\mathcal{E}$ the trainable lift to the model width and $\mathcal{R}$ the task head. Written with dimensions, the two stages that never change are

$$\begin{aligned}
\mathbf{I} &\in \mathbb{R}^{n_b \times T \times d_{\text{in}}}, && \text{windowed input},\\
U^{(0)} &= \big[\,\mathbf{I} \,\big|\, t\,\big] W_{\text{enc}}^{\top} + b_{\text{enc}} \in \mathbb{R}^{n_b \times T \times H}, && (\mathcal{E}:\ \text{linear lift}),\\
U^{(\ell)} &= \text{SpecOSS}\big(U^{(\ell-1)}\big) \in \mathbb{R}^{n_b \times T \times H}, \quad \ell = 1, \dots, L, && (\text{one block per step, (28)}),\\
\hat{\mathbf{O}} &= \mathcal{R}\big(U^{(L)}\big), && (\mathcal{R}:\ \text{task head}).
\end{aligned} \tag{27}$$

Here $t \in \mathbb{R}^T$ is a channel of normalised time, appended by the model on the datasets marked in Table 7 and absent otherwise, so W_{enc} has $d_{\text{in}} + 1$ or d_{in} columns. Opening the middle stage, one application of SpecOSS is one residual block Blk_ℓ at the constant width H, so that $U^{(L)} = \text{Blk}_L \circ \cdots \circ \text{Blk}_1(U^{(0)})$, and block ℓ normalises $U^{(\ell-1)}$, runs one oscillator layer on it, and adds the skip through a GELU and a gated linear unit (Appendix B.5). The layer inside is the second-order system this paper studies, in its own matrices,

$$\begin{aligned}
x'' &= -Ax + Bu, && y = \sum_j G_j \odot C_j x^{(j)},\\
x &\in \mathbb{R}^m, \quad u, y \in \mathbb{R}^H, && A = \text{diag}(A_0, \dots, A_{m-1}) \in \mathbb{R}^{m \times m},\\
B &\in \mathbb{R}^{m \times H}, \quad C \in \mathbb{R}^{H \times m}, && s_{n,k} = M_k s_{n-1,k} + F_{n,k},\\
s_{n,k} &= (z_{n,k}, x_{n,k})^{\top}, && F_{n,k} = (c_{z,k}, c_{x,k})^{\top} (B u_n)_k,\\
M_k^{\text{IM}} &= \begin{pmatrix} S_k & -\Delta t_k A_k S_k \\ \Delta t_k S_k & S_k \end{pmatrix}, && (c_{z,k}, c_{x,k}) = (\Delta t_k S_k,\ \Delta t_k^2 S_k),\\
M_k^{\text{IMEX}} &= \begin{pmatrix} 1 & -\Delta t_k A_k \\ \Delta t_k & 1 - \Delta t_k^2 A_k \end{pmatrix}, && (c_{z,k}, c_{x,k}) = (\Delta t_k,\ \Delta t_k^2).
\end{aligned} \tag{28}$$

the dimensions being those of Section 3: the state x carries the m modes while the input and the output carry the H channels of the residual stream, so B lifts $\mathbb{R}^H$ into the modes and C reads them back. Here $S_k = (1 + \Delta t_k^2 A_k)^{-1}$, $k = 0, \dots, m-1$ indexes the modes, $u_n \in \mathbb{R}^H$ is row n of the normalised block input, y_n is the layer output at that position, the readout above evaluated there and equal to Cx_n on a single basis, and $s_0 = 0$. B and C stay real above and no dimension changes (Appendix B.5); both extend to the complex case, with the readout $y_n = \sum_j G_j \odot \text{Re}(C_j x_n^{(j)})$ in place of $y_n = \sum_j G_j \odot C_j x_n^{(j)}$, an extension ablated in Appendix F.4 with its sensitivity in

Appendix J.1. Only A being a nonnegative diagonal is used anywhere else. Because A is diagonal the m modes decouple into m independent 2×2 systems, which is what lets the scan of Appendix B.4 evaluate the whole sequence exactly in $O(\log T)$ depth, and one layer carries $2m$ state reals per position. The normalisation of a block is batch normalisation on every time-series run and is the block's causal per-token normalisation in the language models. A multi-basis layer partitions the m modes into J contiguous blocks of size m_j and keeps either the single C of concat fusion or the per-block C_j that the weighted and gated fusions combine. The generation route changes A and nothing else, so the four subsections below differ only in how $\mathbf{I}$ is built and in $\mathcal{R}$. Numerical values of T, d_{in}, H, m, L, T_c and the head constants are in Appendix F.1 and Table 7.

G.1 Group 1: time-series classification and pooled regression

Figure 7 draws the flow of this group. This group holds the five long-range classification tasks. A raw record X is one multivariate recording stored as channels by steps, either read from the archive or generated. Windowing only transposes it, no scaling is applied, and the optional time channel is added by $\mathcal{E}$. $\mathcal{R}$ destroys the time axis by averaging and then maps to the task output,

$$\begin{aligned} X &\in \mathbb{R}^{d_{\text{in}}\times T}, && \text{one recording,}\\ \mathbf{I} &= \text{batch of } X^\top \in \mathbb{R}^{n_b\times T\times d_{\text{in}}}, && \text{(transpose and stack, no scaling),}\\ U^{(L)} &= \text{SpecOSS}(U^{(0)}) \in \mathbb{R}^{n_b\times T\times H}, && (\mathcal{E}:\ \text{linear lift to } U^{(0)}),\\ \hat{\mathbf{O}} &= \Big(\tfrac{1}{T}\textstyle\sum_{n=1}^{T} U^{(L)}_{:,n,:}\Big) W_{\text{head}}^\top + b_{\text{head}} \in \mathbb{R}^{n_b\times d_{\text{out}}}, && (\mathcal{R}:\ \text{mean-pool, linear}). \end{aligned} \tag{29}$$

d_{out} is the number of classes and the loss is cross-entropy, so the pooled vector has to carry a quantity accumulated over the whole sequence. Because the pooling is uniform over T, everything the head can use has to survive in $U^{(L)}$ across the full sequence, which is what makes the memory horizon of the spectrum the quantity under test.

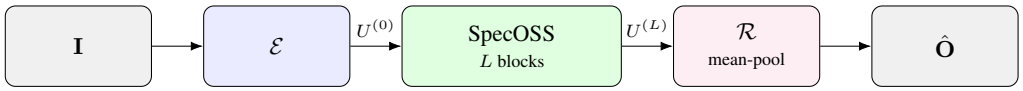


Figure 7: Group 1, (29). Blue is the shared encoder, green the SpecOSS stack, magenta the head; $\mathbf{I}$ is the windowed input and $\hat{\mathbf{O}}$ the prediction.

G.2 Group 2: photoplethysmography regression

Figure 8 draws the flow of this group. The raw record is one subject's wrist recording, min–max scaled to $[-1, 1]$, paired with a heart rate series sampled far more coarsely than the sensors. Each subject is split temporally and then windowed, so one input of T steps faces T/T_{out} targets. $\mathcal{R}$ neither pools nor runs at every step, it reads the stack out once every T_{out} positions, which is exactly the target's sampling rate,

$$\begin{aligned} X &\in \mathbb{R}^{T_{\text{raw}}\times d_{\text{in}}}, && \text{one subject's scaled channels,}\\ \mathbf{I} &\in \mathbb{R}^{n_b\times T\times d_{\text{in}}}, && \text{(split, window),}\\ U^{(L)} &= \text{SpecOSS}(U^{(0)}) \in \mathbb{R}^{n_b\times T\times H}, && (\mathcal{E}:\ \text{linear lift to } U^{(0)});\ (\text{chunk } T_c),\\ \hat{\mathbf{O}} &= \tanh\big(U^{(L)}_{\text{str}} W_{\text{head}}^\top + b_{\text{head}}\big) \in \mathbb{R}^{n_b\times (T/T_{\text{out}})\times d_{\text{out}}}, && (\mathcal{R}:\ \text{strided, } \tanh), \end{aligned} \tag{30}$$

where $U^{(L)}_{\text{str}} \in \mathbb{R}^{n_b\times(T/T_{\text{out}})\times H}$ collects the rows of $U^{(L)}$ at the positions $T_{\text{out}}, 2T_{\text{out}}, \dots, T$, and $d_{\text{out}} = 1$ matches the $[-1, 1]$ scaling of the target, so the trailing axis is dropped and $\hat{\mathbf{O}}$ is compared with the target series directly. This is the group with the longest T that is actually scanned in pieces, and the chunked scan of Appendix B.4 evaluates the stack in segments of T_c steps, which bounds peak memory without changing the result.

G.3 Group 3: weather and masked forecasting

Figure 9 draws the flow of this group. This group's experiment is the $720 \to 720$ weather task. The raw record X is one long multivariate series, split chronologically and standardised with statistics

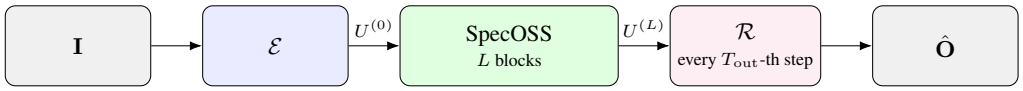


Figure 8: Group 2, (30). The head samples the SpecOSS stack on the target's grid instead of pooling, so the stack has to hold information at every readout position.

of the training portion only. Windows of T steps are cut in which the last T_{fut} rows are replaced by zeros, so the horizon enters the model only as a mask and the forecast is a sequence-to-sequence map rather than an autoregressive rollout. $\mathcal{R}$ runs the head at every position and keeps that horizon,

$$\begin{aligned} X &\in \mathbb{R}^{T_{\text{raw}} \times d_{\text{in}}}, && \text{one series,} \\ \mathbf{I} &= \left[\, X_{i:i+T-T_{\text{fut}}} \;;\; 0_{T_{\text{fut}} \times d_{\text{in}}} \,\right] \in \mathbb{R}^{n_b \times T \times d_{\text{in}}}, && \text{(window, zero the future),} \\ U^{(L)} &= \text{SpecOSS}(U^{(0)}) \in \mathbb{R}^{n_b \times T \times H}, && (\mathcal{E}: \text{ linear lift to } U^{(0)}), \\ \hat{\mathbf{O}} &= \big(U^{(L)} W_{\text{head}}^{\top} + b_{\text{head}}\big)_{:,\, T-T_{\text{fut}}+1:T,\,:} \in \mathbb{R}^{n_b \times T_{\text{fut}} \times d_{\text{out}}}, && (\mathcal{R}: \text{ per-step, tail}), \end{aligned} \tag{31}$$

with $d_{\text{out}} = d_{\text{in}}$ under the value-only protocol this run uses, and the target the true values of the masked steps, scored in the standardised units the windowing produced. The scan is chunked as in group 2. Because the masked half carries no information, the whole forecast is produced by what the SpecOSS stack propagates out of the observed half, which is the property this group tests.

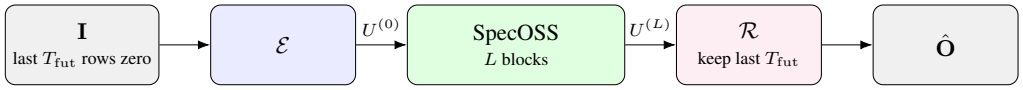


Figure 9: Group 3, (31). The last T_{fut} rows of $\mathbf{I}$ are literal zeros and $\hat{\mathbf{O}}$ predicts exactly those steps, so the forecast is whatever the SpecOSS stack carries across the mask.

G.4 Group 4: language modelling

Figure 10 draws the flow of this group. This group keeps the skeleton and replaces both ends of it. The raw record is text, so a tokenizer and packing produce $\mathbf{I}$, which holds identifiers rather than a real-valued signal, and $\mathcal{E}$ is a table lookup whose transpose is reused as the head,

$$\begin{aligned} X &\in \{0, \dots, d_{\text{vocab}} - 1\}^{*}, && \text{token stream, one EOS per document,} \\ \mathbf{I} &\in \{0, \dots, d_{\text{vocab}} - 1\}^{n_b \times T}, && \text{(pack into length-}T\text{ blocks),} \\ U^{(0)} &= W_{\text{emb}}[\mathbf{I}] \in \mathbb{R}^{n_b \times T \times d_{\text{model}}}, && (\mathcal{E}: \; W_{\text{emb}} \in \mathbb{R}^{d_{\text{vocab}} \times d_{\text{model}}}), \\ U^{(L)} &= \text{SpecOSS}(U^{(0)}) \in \mathbb{R}^{n_b \times T \times d_{\text{model}}}, && (L \text{ blocks around the core}), \\ \hat{\mathbf{O}} &= \text{RMSNorm}\big(U^{(L)}\big) W_{\text{emb}}^{\top} \in \mathbb{R}^{n_b \times T \times d_{\text{vocab}}}, && (\mathcal{R}: \text{ tied head}), \end{aligned} \tag{32}$$

scored by next-token cross-entropy and reported as token-weighted perplexity. Two blocks wrap the SpecOSS core. The Mamba block widens the residual stream, filters it locally, and hands the result to the core,

$$\begin{aligned} [\, V \mid V_z \,] &= U^{(\ell-1)} W_{\text{in}}^{\top} \in \mathbb{R}^{n_b \times T \times 2 d_{\text{inner}}}, && \text{expansion, stream and gate,} \\ \tilde{V} &= \text{SiLU}\big(\text{Conv1d}_{d_{\text{conv}}}(V)\big) \in \mathbb{R}^{n_b \times T \times d_{\text{inner}}}, && \text{depthwise causal filter,} \\ \tilde{V} &\mapsto \mathbb{R}^{n_b \times T \times n_g \times d_{\text{head}}}, \qquad d_{\text{head}} = d_{\text{inner}}/n_g, && \text{split into groups,} \\ U^{(\ell)} &= \big(\text{SpecOSS}(\tilde{V}) \odot \text{SiLU}(V_z)\big) W_{\text{out}}^{\top} \in \mathbb{R}^{n_b \times T \times d_{\text{model}}}, && \text{gate and project back,} \end{aligned} \tag{33}$$

where each group runs (28) with its own $A \in \mathbb{R}^{m \times m}$, $B \in \mathbb{C}^{m \times d_{\text{head}}}$ and $C \in \mathbb{C}^{d_{\text{head}} \times m}$, and n_g is fixed once as the smallest divisor of d_{inner} whose $4 n_g m$ state reals per layer reach the selective baseline's $d_{\text{inner}} d_{\text{state}}$. Because the transition does not depend on the input, the recurrence is a causal convolution and is evaluated by real fast Fourier transform. The plain block removes the expansion, the convolution and the input-dependent gate, and is the residual block Blk_{ℓ} of the other three groups with a causal normalisation, so the core runs at the model width with $n_g = 1$ and $d_{\text{head}} = d_{\text{model}}$ and the only parameter difference between two variants sits inside the core. Both corpora, WikiText-103 and the deduplicated Pile, are packed the same way and differ only in how the stream is built, the Pile blocks being cut so that the smaller token budgets are prefixes of the largest.

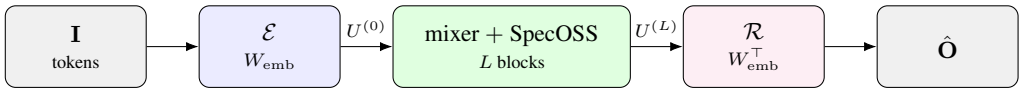


Figure 10: Group 4, (32). $\mathbf{I}$ holds packed tokens and the embedding table W_{emb} is reused transposed as the head. The SpecOSS core of (28) sits inside the block, Mamba's (33) or the plain one. WikiText-103 and the Pile share the whole diagram.

H Additional results

H.1 Every basis, both routes, frozen and trainable

Table 8 is the spectrum behind the summary rows of Table 2: every polynomial basis read once with the Sturm–Liouville spectrum of route 1 and once with the function-approximation spectrum of route 2, each with the spectrum frozen and with it trained from that initialisation and best over the two discretisations. Fourier and Laguerre were not run on Motor, whose sequence length puts the full matrix out of budget, so those rows carry no average.

Table 8: UEA test accuracy (%) by basis, generation route and whether the spectrum is frozen, best over IM and IMEX. The multi-basis row of each block is the best union of that route's spectra. **Bold** marks the cells that feed the summary rows of Table 2.

	SCP1	SCP2	Ethanol	Heartbeat	Motor	Avg
Route 1 (SL), spectrum frozen						
Hermite (SL, frozen)	87.1 ± 1.8	$\mathbf{56.1 \pm 2.4}$	40.5 ± 1.7	75.8 ± 2.1	$\mathbf{57.9 \pm 2.6}$	63.5 ± 1.6
Legendre (SL, frozen)	84.7 ± 2.1	61.4 ± 1.8	39.2 ± 2.4	79.0 ± 1.6	56.1 ± 2.2	64.1 ± 1.7
Fourier (SL, frozen)	84.7 ± 1.9	59.6 ± 2.5	39.2 ± 1.8	74.2 ± 2.3	–	–
Laguerre (SL, frozen)	83.5 ± 2.2	56.1 ± 1.7	39.2 ± 2.1	83.9 ± 1.9	–	–
Multi-basis (SL, frozen)	$\mathbf{87.2 \pm 3.1}$	$\mathbf{58.0 \pm 2.2}$	$\mathbf{40.0 \pm 1.8}$	$\mathbf{83.9 \pm 2.4}$	$\mathbf{55.9 \pm 2.5}$	$\mathbf{65.0 \pm 1.9}$
Route 1 (SL), spectrum trained from the generated initialisation						
Hermite (SL, trainable)	$\mathbf{90.3 \pm 3.1}$	54.4 ± 2.4	$\mathbf{44.3 \pm 3.9}$	$\mathbf{77.4 \pm 3.1}$	56.1 ± 1.7	64.5 ± 1.5
Legendre (SL, trainable)	85.9 ± 2.0	64.9 ± 1.9	36.7 ± 2.6	75.8 ± 1.8	52.6 ± 2.3	63.2 ± 1.8
Fourier (SL, trainable)	85.9 ± 1.7	56.1 ± 2.2	34.2 ± 2.5	79.0 ± 1.9	–	–
Laguerre (SL, trainable)	87.1 ± 1.9	57.9 ± 2.1	34.2 ± 2.3	83.9 ± 1.6	–	–
Multi-basis (SL, trainable)	$\mathbf{88.2 \pm 2.4}$	$\mathbf{61.4 \pm 2.3}$	$\mathbf{40.5 \pm 3.1}$	$\mathbf{85.5 \pm 2.4}$	$\mathbf{56.1 \pm 3.8}$	$\mathbf{66.3 \pm 2.1}$
Route 2 (FA), spectrum frozen						
Hermite (FA, frozen)	83.5 ± 2.1	$\mathbf{57.9 \pm 2.8}$	29.1 ± 2.7	75.8 ± 1.9	42.1 ± 2.4	57.7 ± 1.8
Legendre (FA, frozen)	84.7 ± 1.8	54.4 ± 2.3	32.9 ± 2.5	75.8 ± 2.0	36.8 ± 2.6	56.9 ± 1.9
Fourier (FA, frozen)	83.5 ± 2.0	63.2 ± 1.7	38.0 ± 2.4	72.6 ± 2.2	–	–
Laguerre (FA, frozen)	87.1 ± 1.6	57.9 ± 2.1	32.9 ± 2.6	77.4 ± 1.8	–	–
Multi-basis (FA, frozen)	$\mathbf{90.1 \pm 2.3}$	$\mathbf{67.7 \pm 2.4}$	$\mathbf{36.7 \pm 3.1}$	$\mathbf{82.3 \pm 2.4}$	$\mathbf{53.8 \pm 2.1}$	$\mathbf{66.1 \pm 2.0}$
Route 2 (FA), spectrum trained from the generated initialisation						
Hermite (FA, trainable)	$\mathbf{84.7 \pm 1.8}$	54.4 ± 2.2	$\mathbf{31.6 \pm 2.6}$	$\mathbf{79.0 \pm 3.3}$	$\mathbf{64.9 \pm 2.7}$	62.9 ± 1.6
Legendre (FA, trainable)	85.9 ± 1.8	49.1 ± 2.6	35.4 ± 2.3	74.2 ± 2.1	52.6 ± 2.4	59.4 ± 2.0
Fourier (FA, trainable)	84.7 ± 2.0	45.6 ± 2.7	32.9 ± 2.2	77.4 ± 1.9	–	–
Laguerre (FA, trainable)	87.1 ± 1.7	64.9 ± 1.8	34.2 ± 2.5	77.4 ± 2.0	–	–
Multi-basis (FA, trainable)	$\mathbf{90.6 \pm 3.4}$	$\mathbf{68.4 \pm 2.2}$	$\mathbf{40.2 \pm 2.8}$	$\mathbf{84.2 \pm 3.3}$	$\mathbf{54.4 \pm 1.2}$	$\mathbf{67.6 \pm 2.0}$

Table 9: Trainable parameters of our rows of Table 2, for the frozen SL spectrum at the architecture of Table 7. Every baseline is tuned under the same protocol and runs in the same backbone at a comparable budget (Appendix F.4).

	SCP1	SCP2	Ethanol	Heartbeat	Motor
ours	15.9k	21.1k	4.1k	7.4k	33.2k

Table 11 names the winning variant behind every entry of Table 8, read once by basis and once by route, together with our ports of the baseline architectures trained under the same budget.

Table 12 ablates how two bases are combined on Heartbeat, over the dense gate on the per-block readouts, the rank-r factorisation of that gate, mixture dropout on it, the single shared readout of concat fusion, a single basis, and one basis per block up the stack.

Table 10: Parameter efficiency against the published LinOSS architecture. Every comparison reported in this paper is run in a fixed backbone at a matched budget under one tuning protocol (Appendix F.4, Table 9). This table is a separate comparison, added to show the effectiveness of a generated spectrum: it places our counts next to the parameter counts reported with LinOSS by Rusch & Rus (2025) at that model's own reference architecture, at the accuracies of Table 2.

	SCP1	SCP2	Ethanol	Heartbeat	Motor
ours	15.9k	21.1k	4.1k	7.4k	33.2k
LinOSS (published)	991.2k	399.1k	6.7k	10.9k	91.8k
ratio	62×	18.9×	1.6×	1.5×	2.8×

Table 11: The best variant behind each basis and each route of Table 8 (test accuracy %, discretisation, trainable parameters), together with our ports of the baseline architectures trained under the same budget. Column headers give the epoch budget of the full training run. For the SL frozen, SL trainable, Multi and baseline-port rows the parameters are followed by the wall-clock of that run in minutes; – marks a wall-clock not measured for that architecture. **Bold** marks the accuracy value in a cell that also appears as an entry of Table 2. The LinOSS, LRU and S5 rows are the Table 2 baselines, LinOSS at the better discretisation per dataset; every baseline is tuned under the same protocol in the same backbone (Appendix F.4), and these rows repeat no parameter count or wall-clock.

	SCP1 [epoch:250]	SCP2 [epoch:200]	Ethanol [epoch:250]	Heartbeat [epoch:250]	Motor [epoch:150]
Best route and discretisation, by basis					
Hermite	SL trainable IM **90.3** (16.0k)	FA frozen IMEX **57.9** (25.4k)	SL trainable IMEX **44.3** (4.1k)	FA trainable IM **79.0** (7.5k)	FA trainable IMEX **64.9** (29.3k)
Legendre	SL trainable IM 85.9 (19.2k)	SL trainable IM 64.9 (25.6k)	SL frozen IMEX 39.2 (4.1k)	SL frozen IM 79.0 (6.3k)	SL frozen IM 56.1 (29.3k)
Fourier	SL trainable IM 85.9 (19.2k)	FA frozen IMEX 63.2 (25.4k)	SL frozen IMEX 39.2 (4.1k)	SL trainable IM 79.0 (6.4k)	–
Laguerre	SL trainable IM 87.1 (16.0k)	FA trainable IM 64.9 (25.6k)	SL frozen IM 39.2 (4.1k)	SL trainable IMEX 83.9 (7.5k)	–
Best basis and discretisation, by route					
SL frozen	Hermite IM 87.1 (15.9k, 1.3)	Legendre IMEX 61.4 (21.1k, 0.9)	Hermite IMEX 40.5 (4.1k, 1.3)	Laguerre IMEX 83.9 (7.4k, 2.0)	Hermite IM **57.9** (33.2k, 1.1)
SL trainable	Hermite IM **90.3** (16.0k, 2.2)	Legendre IM 64.9 (25.6k, 1.9)	Hermite IMEX **44.3** (4.1k, 2.0)	Laguerre IMEX 83.9 (7.5k, 2.2)	Hermite IMEX 56.1 (29.3k, 2.3)
FA frozen	Laguerre IMEX 87.1 (19.1k)	Fourier IMEX 63.2 (25.4k)	Fourier IM 38.0 (4.1k)	Laguerre IM 77.4 (7.4k)	Hermite IM 42.1 (29.3k)
FA trainable	Laguerre IMEX 87.1 (19.2k)	Laguerre IM 64.9 (25.6k)	Legendre IM 35.4 (4.1k)	Hermite IM **79.0** (7.5k)	Hermite IMEX **64.9** (29.3k)
Reference variants					
Multi	FA **90.6** (19.1k, 2.8)	FA **68.4** (27.9k, 4.3)	(SL, Hermite +Legendre) gate IMEX **40.5** (7.2k, 1.3)	(SL, Hermite +Laguerre) low-rank gate IMEX **85.5** (8.6k, 2.3)	(SL, Hermite +Fourier+Legendre) gate IMEX **56.1** (103k, 2.5)
LinOSS	IM **87.8** (–)	IMEX **58.9** (–)	IM **29.9** (–)	IM **75.8** (–)	IM **60.0** (–)
LinOSS frozen	IM 56.3 (–)	IMEX 47.9 (–)	IM 24.9 (–)	IM 51.2 (–)	IM 51.4 (–)
LRU	**82.6** (–)	**51.2** (–)	**21.5** (–)	**78.4** (–)	**48.4** (–)
S5	**89.9** (–)	**50.5** (–)	**24.1** (–)	**77.7** (–)	**47.7** (–)
Baseline ports, same budget					
GRU	70.6 (4.0k, 0.2)	49.1 (4.1k, 0.1)	35.4 (1.1k, 0.2)	69.4 (3.9k, 0.2)	50.9 (25.1k, 0.7)
LSTM	69.4 (5.3k, 0.2)	52.6 (5.4k, 0.1)	40.5 (1.4k, 0.2)	75.8 (5.2k, 0.2)	57.9 (33.4k, 0.6)
linear RNN	69.4 (–)	50.9 (–)	30.4 (–)	62.9 (–)	50.9 (–)

H.2 PPG-DaLiA, all variants

Table 13 lists every variant under both generation routes and both discretisations. The Legendre runs predict a near-constant heart rate (validation MSE flat from the first evaluation); under the strided readout this configuration did not train.

H.3 Weather forecasting, full core-swap matrix

Table 14 gives the test MSE and MAE of the baselines of Table 4 and of every basis under both generation routes and both spectrum modes. Table 15 lists our best combinations in the same fixed residual backbone ($H = 128$, $m = 64$, $L = 4$, Adam, 20 epochs with early stopping; Appendix F.5),

Table 12: Heartbeat fusion ablation: test accuracy (%) and trainable parameters, selected as in Table 2. **Bold**: best, the variant behind that table's Heartbeat column.

variant	test	params
Multi (SL, Hermite+Fourier), dense gate, IM	80.6	11,666
Multi (SL, Hermite+Fourier), low-rank gate, IM	82.3	8,594
Multi (SL, Hermite+Fourier), low-rank gate and dropout, IM	82.3	8,594
Multi (SL, Hermite+Fourier), low-rank gate and dropout, IMEX	79.0	8,594
Multi (SL, Hermite+Laguerre), low-rank gate, IM	82.3	8,594
Multi (SL, Hermite+Laguerre), low-rank gate and dropout, IMEX	**85.5**	8,594
Multi (SL, Hermite+Fourier), concat, IM	82.3	7,442
Laguerre (SL), single basis, IM	80.6	7,442
Hermite/Laguerre by depth, trainable, IM	82.3	7,506

Table 13: PPG-DaLiA, all variants: test MSE $\times 10^{-2}$ under IM and IMEX. **Bold**: best.

variant	IM	IMEX
Hermite (SL), frozen	3.46 ± 0.44	3.73 ± 0.55
Hermite (SL), trainable	3.10 ± 0.51	4.96 ± 0.88
Legendre (SL), frozen	29.97 ± 2.13	30.03 ± 3.14
Legendre (SL), trainable	30.45 ± 0.88	29.97 ± 0.88
Multi (SL, Hermite+Legendre), frozen concat	2.84 ± 0.77	4.08 ± 0.56
Multi (SL, Hermite+Legendre), frozen gated	3.72 ± 0.18	3.49 ± 0.29
Multi (SL, Hermite+Legendre), trainable concat	$\mathbf{2.39 \pm 0.27}$	3.22 ± 0.38
Multi (SL, Hermite+Legendre), trainable gated	3.68 ± 0.22	3.63 ± 0.31
Hermite (FA), frozen	4.07 ± 0.49	29 ± 4.12
Hermite (FA), trainable	3.10 ± 0.38	25 ± 1.41
Multi (FA, Hermite+Legendre), frozen	3.89 ± 0.19	5.54 ± 0.25
Multi (FA, Hermite+Legendre), trainable	4.53 ± 0.74	6.44 ± 0.31

with both discretisations for every oscillator core. The untouched Informer of Table 4 is an external architecture baseline and is not re-run.

Table 14: WTH 720→720, test MSE and MAE for the baselines of Table 4 and for every basis, generation route and spectrum mode. **Bold**: best; underlined: second, as in Table 4.

core	MSE	MAE
Informer	0.839 ± 0.022	0.731 ± 0.010
LogTrans	0.939 ± 0.026	0.773 ± 0.011
Reformer	3.897 ± 0.094	1.575 ± 0.019
LSTMa	1.932 ± 0.058	1.109 ± 0.017
LSTnet	0.900 ± 0.020	0.757 ± 0.008
S4	0.525 ± 0.014	0.578 ± 0.008
LinOSS-IMEX	0.405 ± 0.011	0.508 ± 0.007
LinOSS-IM	0.438 ± 0.011	0.528 ± 0.006
Hermite (SL, frozen)	0.377 ± 0.009	0.490 ± 0.006
Hermite (SL, trainable)	0.290 ± 0.008	0.430 ± 0.006
Legendre (SL, frozen)	0.792 ± 0.021	0.710 ± 0.009
Legendre (SL, trainable)	0.684 ± 0.021	0.660 ± 0.010
Fourier (SL, frozen)	0.748 ± 0.018	0.690 ± 0.008
Fourier (SL, trainable)	0.643 ± 0.018	0.640 ± 0.009
Laguerre (SL, frozen)	1.970 ± 0.043	1.120 ± 0.012
Laguerre (SL, trainable)	1.798 ± 0.047	1.070 ± 0.014
Multi-basis (SL, frozen)	0.304 ± 0.009	0.440 ± 0.007
Multi-basis (SL, trainable)	$\mathbf{0.227 \pm 0.006}$	$\mathbf{0.380 \pm 0.005}$
Hermite (FA, frozen)	0.393 ± 0.009	0.500 ± 0.006
Hermite (FA, trainable)	0.332 ± 0.009	0.460 ± 0.006
Legendre (FA, frozen)	0.931 ± 0.024	0.770 ± 0.010
Legendre (FA, trainable)	0.664 ± 0.020	0.650 ± 0.010
Fourier (FA, frozen)	0.664 ± 0.016	0.650 ± 0.008
Fourier (FA, trainable)	0.623 ± 0.017	0.630 ± 0.009
Laguerre (FA, frozen)	2.534 ± 0.056	1.270 ± 0.014
Laguerre (FA, trainable)	2.041 ± 0.053	1.140 ± 0.015
Multi-basis (FA, frozen)	0.347 ± 0.010	0.470 ± 0.007
Multi-basis (FA, trainable)	$\underline{0.251 \pm 0.007}$	$\underline{0.400 \pm 0.006}$

Table 15: WTH 720→720, controlled core swap in the backbone of Appendix F.5: test MSE and MAE in normalised units and trainable parameters, for our combinations under both discretisations. The IMEX rows are the trainable-spectrum numbers of Table 14. **Bold**: best per column.

core	MSE	MAE	params
SpecOSS SL Hermite, trainable, IM	0.392 ± 0.011	0.610 ± 0.009	267k
SpecOSS SL Hermite, trainable, IMEX	0.290 ± 0.008	0.430 ± 0.006	267k
SpecOSS FA Hermite, trainable, IM	0.367 ± 0.009	0.581 ± 0.005	267k
SpecOSS FA Hermite, trainable, IMEX	0.332 ± 0.009	0.460 ± 0.006	267k
SpecOSS Multi (Hermite+Legendre), trainable concat, IM	0.311 ± 0.004	0.491 ± 0.008	267k
SpecOSS Multi (Hermite+Legendre), trainable concat, IMEX	$\mathbf{0.227 \pm 0.006}$	$\mathbf{0.380 \pm 0.005}$	267k

H.4 Language modelling, full result matrix

Table 16 extends Table 5 to every generation route, basis and spectrum mode, holding fixed the placement that gives the best result: the generated spectrum in place of Mamba's selective scan (Appendix F.6), after 100M training tokens on both corpora. Mamba is the tuned baseline, the reference row every other variant is measured against.

Table 16: Every route, basis and spectrum mode in place of Mamba's selective scan, after 100M training tokens: negative log-likelihood (NLL, nats per token) and perplexity (PPL). **Bold**: best per column.

		WikiText-103		Pile	
variant	params	NLL ↓	PPL ↓	NLL ↓	PPL ↓
Mamba	46.7M	3.891	48.96	4.592	98.69
SpecOSS Hermite (SL), frozen	46.7M	3.895	49.17	4.616	101.12
SpecOSS Hermite (SL), trainable	46.7M	**3.818**	**45.51**	**4.476**	**87.88**
SpecOSS Legendre (SL), frozen	46.7M	4.053	57.55	4.684	108.16
SpecOSS Legendre (SL), trainable	46.7M	3.983	53.66	4.653	104.92
SpecOSS Hermite (FA), frozen	46.7M	3.900	49.39	4.613	100.77
SpecOSS Hermite (FA), trainable	46.7M	3.822	45.69	4.479	88.14
SpecOSS Legendre (FA), frozen	46.7M	4.064	58.19	4.699	109.88
SpecOSS Legendre (FA), trainable	46.7M	3.973	53.14	4.641	103.62
SpecOSS Multi (Hermite+Legendre), trainable	46.7M	3.822	45.71	4.477	87.97

H.5 Electrocardiogram classification

Table 17 is the experiment referred to in Section 4.1. One backbone ($H = 64$, $m = 64$, $L = 4$, dropout 0.1, batch 64, 40 epochs of Adam at 10^{-3} under a cosine schedule) is held fixed and only the sequence core changes, as in the weather core swap of Appendix H.3. Results are over the five seeds of Appendix F.3.

Two things are visible in the table. Within the generated spectra the implicit discretisation is ahead of the energy-conserving one on every single variant, which is the behaviour the dissipative scheme is expected to show on recordings that are noisy rather than conservative. Against the structure-free control the frozen Hermite spectrum is ahead both as first generated and once its band is placed, so the ordering between a generated spectrum and a learned random one of the same size is the same here as on the benchmarks of the main text.

Figure 11 is the companion to Figure 2 of Example 1, with every spectrum read with λ^{FA} in place of λ^{SL}.

I Parameter count, run time and memory

This section gives the asymptotic cost of one SpecOSS layer against the layers it is compared with, and then the measured parameter counts, wall-clock times and peak GPU memory of the runs behind the experiments, read from the run logs. All measurements were taken on shared cluster nodes with an NVIDIA A100 80 GB, and the wall-clock of a training run includes its per-epoch validation passes.

Table 17: PTB-XL twelve-lead electrocardiograms, five diagnostic superclasses: test accuracy (%) and trainable parameters on the official test fold. **Bold** marks the values also reported in Example 1.

core	accuracy	params
Reference cores in the same backbone		
First-order SSMs (S5, LRU)	**68.0 ± 0.2** S5	67.9k
LinOSS, IM	73.7 ± 0.9	100.8k
LinOSS, IMEX	71.8 ± 1.6	100.8k
Route 1 (SL)		
SpecOSS SL Hermite, frozen, IM	74.9 ± 1.4	100.5k
SpecOSS SL Hermite, trainable, IM	**75.0 ± 0.6**	100.8k
SpecOSS Multi (SL, Hermite+Laguerre), frozen concat, IM	74.8 ± 0.1	100.5k
SpecOSS Multi (SL, Hermite+Laguerre), frozen gated low-rank, IM	75.2 ± 1.2	105.2k
SpecOSS Multi (SL, Hermite+Laguerre), trainable concat, IM	**78.5± 0.7**	100.8k
SpecOSS Multi (SL, Hermite+Laguerre), trainable gated low-rank, IM	74.4 ± 0.5	105.4k
SpecOSS Multi (SL, Hermite+Legendre), frozen concat, IM	74.1 ± 0.9	100.5k
SpecOSS Multi (SL, Hermite+Legendre), frozen gated low-rank, IM	73.6 ± 0.1	105.2k
SpecOSS Multi (SL, Hermite+Legendre), trainable concat, IM	73.8 ± 1.1	100.8k
SpecOSS Multi (SL, Hermite+Legendre), trainable gated low-rank, IM	73.9 ± 0.7	105.4k
SpecOSS SL Hermite, frozen, IMEX	70.6 ± 0.2	100.5k
SpecOSS SL Hermite, trainable, IMEX	73.0 ± 0.5	100.8k
SpecOSS Multi (SL, Hermite+Laguerre), frozen concat, IMEX	71.3 ± 0.8	100.5k
SpecOSS Multi (SL, Hermite+Laguerre), frozen gated low-rank, IMEX	70.8 ± 0.5	105.2k
SpecOSS Multi (SL, Hermite+Laguerre), trainable concat, IMEX	73.0 ± 0.5	100.8k
SpecOSS Multi (SL, Hermite+Laguerre), trainable gated low-rank, IMEX	72.3 ± 0.6	105.4k
SpecOSS Multi (SL, Hermite+Legendre), frozen concat, IMEX	70.4 ± 0.3	100.5k
SpecOSS Multi (SL, Hermite+Legendre), frozen gated low-rank, IMEX	70.6 ± 0.1	105.2k
SpecOSS Multi (SL, Hermite+Legendre), trainable concat, IMEX	72.5 ± 0.4	100.8k
SpecOSS Multi (SL, Hermite+Legendre), trainable gated low-rank, IMEX	71.9 ± 1.8	105.4k
Route 2 (FA)		
SpecOSS FA Hermite, frozen, IM	73.8 ± 1.1	100.5k
SpecOSS FA Hermite, trainable, IM	75.6 ± 0.8	100.8k
SpecOSS Multi (FA, Hermite+Laguerre), frozen concat, IM	74.2 ± 0.6	100.5k
SpecOSS Multi (FA, Hermite+Laguerre), frozen gated low-rank, IM	76.0 ± 1.0	105.2k
SpecOSS Multi (FA, Hermite+Laguerre), trainable concat, IM	77.8 ± 0.7	100.8k
SpecOSS Multi (FA, Hermite+Laguerre), trainable gated low-rank, IM	75.1 ± 0.9	105.4k
SpecOSS Multi (FA, Hermite+Legendre), frozen concat, IM	72.9 ± 0.8	100.5k
SpecOSS Multi (FA, Hermite+Legendre), frozen gated low-rank, IM	74.5 ± 0.6	105.2k
SpecOSS Multi (FA, Hermite+Legendre), trainable concat, IM	74.6 ± 1.0	100.8k
SpecOSS Multi (FA, Hermite+Legendre), trainable gated low-rank, IM	73.1 ± 0.7	105.4k
SpecOSS FA Hermite, frozen, IMEX	71.4 ± 0.5	100.5k
SpecOSS FA Hermite, trainable, IMEX	72.2 ± 0.7	100.8k
SpecOSS Multi (FA, Hermite+Laguerre), frozen concat, IMEX	72.0 ± 0.6	100.5k
SpecOSS Multi (FA, Hermite+Laguerre), frozen gated low-rank, IMEX	69.9 ± 0.4	105.2k
SpecOSS Multi (FA, Hermite+Laguerre), trainable concat, IMEX	73.7 ± 0.6	100.8k
SpecOSS Multi (FA, Hermite+Laguerre), trainable gated low-rank, IMEX	71.6 ± 0.8	105.4k
SpecOSS Multi (FA, Hermite+Legendre), frozen concat, IMEX	69.8 ± 0.5	100.5k
SpecOSS Multi (FA, Hermite+Legendre), frozen gated low-rank, IMEX	71.2 ± 0.3	105.2k
SpecOSS Multi (FA, Hermite+Legendre), trainable concat, IMEX	73.3 ± 0.5	100.8k
SpecOSS Multi (FA, Hermite+Legendre), trainable gated low-rank, IMEX	72.6 ± 1.1	105.4k

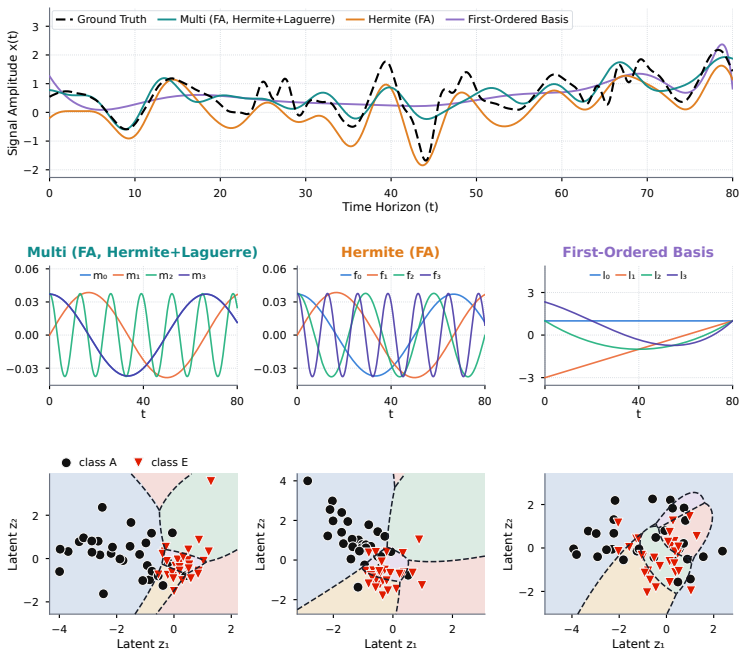


Figure 11: As Figure 2, with every spectrum read with λ^{FA}. Rows: predicted signal, modes, and the two-dimensional projection with the class boundaries. Five-class accuracy 77.8 / 75.6 / 68.0%.

Asymptotic cost. Table 18 counts, per layer, the trainable parameters and the work of one forward pass over a length-T sequence. A SpecOSS layer with a frozen spectrum stores $B \in \mathbb{R}^{m\times H}$ and $C \in \mathbb{R}^{H\times m}$ (real, $2mH$ parameters) or $B \in \mathbb{C}^{m\times H}$ and $C \in \mathbb{C}^{H\times m}$ (complex, $4mH$ parameters; the default, Appendix F.2) and the m timesteps, so the layer has $2mH + m$ or $4mH + m$ trainable parameters; a learned spectrum adds m more either way, and the state is the same, $2m$ reals per position. The work splits into the two dense products with B and C, $O(TmH)$, and the scan of the m decoupled 2×2 recurrences, $O(Tm)$ work in $O(\log T)$ depth (Appendix B.4), which is the cost of a diagonal first-order SSM with $2m$ real states. The multi-basis fusions keep the products and add the gate, $2rJH + JH$ parameters and $O(TrJH)$ work for the low-rank gate against $(JH)^2 + JH$ and $O(TJ^2H^2)$ for the dense one. Peak memory is dominated by the scan levels that autograd keeps alive, $6\lceil \log_2 T \rceil\, n_b T m$ numbers per layer, which the exact chunked scan with chunk length T_c lowers to $6\lceil \log_2 T_c \rceil\, n_b T_c m$; at the PPG-DaLiA setting ($n_b = 4$, $T_c = 12{,}480$, $m = 64$) this is 2.7×10^8 numbers, about 1.1 GB per layer in single precision, against 10^{10} numbers for the unchunked scan over $T = 49{,}920$. In the language models the time-invariant realisations evaluate the recurrence as a causal convolution by FFT, $O(n_g m T \log T)$ per layer, and the FFT buffers, not the parameters, set the peak memory reported below.

Classification tasks. Table 11 also gives, for the SL frozen, SL trainable, multi-basis and baseline-port rows, the wall-clock of the complete training run (validated after every epoch) alongside the parameters, with the epoch budget of each dataset in the column header. Frozen and trainable spectra cost the same per step; where the frozen entry is faster the winning frozen variant came from the earlier realisation without batch normalisation and dropout (SCP1, Motor), which is a cheaper block, not a cheaper spectrum. For our Heartbeat models the scan bound above gives $6 \cdot 9 \cdot 32 \cdot 405 \cdot 16 \approx 1.1 \times 10^7$ numbers, 45 MB per layer.

PPG-DaLiA and weather. The PPG-DaLiA models have 28,465 parameters with a frozen spectrum, 28,849 with a trainable one and 34,801–35,185 with a gated union. On an A100 with the chunked scan ($T_c = 12{,}480$, batch size 4), the 12,000 optimiser steps with a validation pass every 500 steps took 5.4 hours for the frozen Hermite spectrum under IM, 5.4–6.2 hours for the single-basis variants and 5.4–10.1 hours for the unions and the random control. On WTH the core-swap backbone has 267,404 parameters for every oscillator core, 267,660 for the LRU, 201,740 for S5 and 748,684 for the Mamba mixer; one epoch of the 1,440-step sequences takes 5.6 minutes for

Table 18: Asymptotic cost per layer for an input of length T and width H. Every sequence core is written with m states per channel group; for the first-order cores m counts complex states, for ours oscillators. J bases and gate rank r refer to the multi-basis fusion and n_g to the channel groups of the language-model core. The B, C column gives the realisation the parameter count assumes for every SpecOSS row; the baselines (LinOSS, S5, LRU) use their own fixed, complex convention.

layer	B, C	trainable parameters	state per position	forward work	scan / conv. depth
SpecOSS, frozen spectrum	real	$2mH + m$	$2m$	$O(TmH) + O(Tm)$	$O(\log T)$
SpecOSS, frozen spectrum	complex	$4mH + m$	$2m$	$O(TmH) + O(Tm)$	$O(\log T)$
SpecOSS, trainable spectrum	real	$2mH + 2m$	$2m$	$O(TmH) + O(Tm)$	$O(\log T)$
SpecOSS, trainable spectrum	complex	$4mH + 2m$	$2m$	$O(TmH) + O(Tm)$	$O(\log T)$
LinOSS	complex	$4mH + 2m$	$2m$	$O(TmH) + O(Tm)$	$O(\log T)$
SpecOSS multi-basis, concat	real	$2mH + m$	$2m$	$O(TmH) + O(Tm)$	$O(\log T)$
SpecOSS multi-basis, concat	complex	$4mH + m$	$2m$	$O(TmH) + O(Tm)$	$O(\log T)$
SpecOSS multi-basis, low-rank gate	real	$2mH + m + 2rJH + JH$	$2m$	$O(TmH) + O(TrJH)$	$O(\log T)$
SpecOSS multi-basis, low-rank gate	complex	$4mH + m + 2rJH + JH$	$2m$	$O(TmH) + O(TrJH)$	$O(\log T)$
SpecOSS multi-basis, dense gate	real	$2mH + m + (JH)^2 + JH$	$2m$	$O(TmH) + O(TJ^2H^2)$	$O(\log T)$
SpecOSS multi-basis, dense gate	complex	$4mH + m + (JH)^2 + JH$	$2m$	$O(TmH) + O(TJ^2H^2)$	$O(\log T)$
S5 / LRU (diagonal first-order, dense B, C)	complex	$4mH + 2m$	$2m$	$O(TmH) + O(Tm)$	$O(\log T)$
SpecOSS in the language model, FFT path	real	$2mH + 2m$ per group	$2n_gm$	$O(n_gmT \log T)$	$O(\log T)$
SpecOSS in the language model, FFT path	complex	$4mH + 2m$ per group	$2n_gm$	$O(n_gmT \log T)$	$O(\log T)$

the oscillator cores, 3.7 for Mamba, 2.2 for the LRU and 1.5 for S5 on an A100. Peak memory, logged for the tuning runs of Appendix J.3, is 1.8 GB at the backbone size and 4.2 GB at $H = 256$, $m = 128$, $L = 5$ (1.32M parameters).

J Sensitivity analysis

This section is a sensitivity analysis: it reports how much the results move when one setting is changed at a time.

J.1 Realisation of the layer around the spectrum

Stage A of the tuning protocol (Appendix F.4) trained every spectrum under three independent toggles of the layer around it: real against complex B, C, a fixed common step $\Delta t = 1$ against a learned per-mode Δt_k, and no normalisation against batch normalisation (real B is whitened by $\sqrt{A_k}$). Every number reported in this paper comes from the last of the eight combinations, complex B, C with a learned Δt_k and batch normalisation, which is the default of Table 7. Table 19 holds one configuration fixed, the multi-basis Sturm–Liouville spectrum with a trainable A, and turns those toggles off one at a time, so its last row is the Multi (SL, ours) row of Table 2. Every row reads route 1, whereas the Multi row of Table 11 takes the better of the two routes per dataset and lands on route 2 for SCP1 and SCP2, which is why those two columns stand higher there.

Table 19: Sensitivity to the realisation: test accuracy (%) of the multi-basis SL spectrum with a trainable A under each of the eight Stage A realisations, per dataset. The last row, complex B, C with a learned Δt_k and batch normalisation, is the realisation behind every number reported in the paper; its entries are the Multi (SL, ours) row of Table 2. Route 1 throughout, so SCP1 and SCP2 sit below the route-2 entries of Table 11's Multi row.

B, C	Δt	norm.	SCP1	SCP2	Ethanol	Heartbeat	Motor
real	fixed	no	84.0	58.2	36.9	83.1	51.7
real	fixed	yes	85.6	59.0	38.7	83.2	53.5
real	learned	no	86.1	59.6	38.9	83.0	53.8
real	learned	yes	87.1	60.7	39.9	84.3	55.7
complex	fixed	no	85.3	59.0	37.6	82.8	54.5
complex	fixed	yes	86.2	60.7	39.5	84.4	55.0
complex	learned	no	87.2	60.0	39.6	84.7	55.5
complex	learned	yes	88.2	61.4	40.5	85.5	56.1

J.2 IM versus IMEX

Across the five UEA datasets the best frozen SL variant used IMEX on SCP2, Ethanol and Heartbeat (Laguerre) and IM on SCP1 and Motor; trainable spectra used IMEX on Ethanol, Heartbeat and Motor. Which discretisation wins depends on the union as much as on the dataset, since on

Heartbeat the Hermite+Laguerre gate is best under IMEX at 85.5 while the Hermite+Fourier gate prefers IM, 82.3 against 79.0 (Table 12). On PPG-DaLiA IM is better for every single-basis and concat variant (Hermite frozen 3.46 against 3.73; Hermite trainable 3.10 against 4.96; Multi (SL, Hermite+Legendre) frozen concat 2.84 against 4.08), the gated unions and the trainable Legendre spectrum come out marginally the other way, and the route-2 Hermite spectra diverge under IMEX (Table 13). On the electrocardiograms IM is ahead under both routes, by 5.5 points at the best route-1 setting and 4.1 at the best route-2 one (Table 17). On WTH the picture reverses and IMEX is better for every oscillator core, by 0.035–0.102 MSE (Table 15). The raw-spectrum IMEX configurations are excluded by the CFL guard (Proposition 4b).

J.3 Weather backbone and optimiser

Table 20 lists the tuning runs of the SL Hermite trainable IMEX variant as changes from the backbone; the base widens and deepens the core, and every row below the base changes one setting from it. Moving from the backbone to the base buys 0.011 MSE, and on top of that dropout 0.15 is the single largest gain at a further 0.020, followed by a cosine schedule, the appended time channel and a lower learning rate. Depth beyond five blocks does not help, a longer patience adds nothing over the learning rate it accompanies, and narrowing the core while keeping the state size is the worst change tried. The spread of the whole search is 0.043 MSE and 0.025 MAE.

Table 20: WTH 720→720, SL Hermite trainable IMEX variant: change in test MSE and MAE from the backbone row, with trainable parameters. The backbone is the variant of Table 15, 0.290 MSE and 0.430 MAE; **bold** is the largest improvement.

configuration	ΔMSE	ΔMAE	params
backbone: $H = 128$, $m = 64$, $L = 4$, lr 10^{-3}	reference	reference	0.27M
base: $H = 256$, $m = 128$, $L = 5$, lr 7×10^{-4}	-0.011	-0.007	1.32M
$L = 8$	-0.009	$+0.005$	2.11M
lr 3×10^{-4}	-0.017	-0.009	1.32M
dropout 0.15	$\mathbf{-0.031}$	$\mathbf{-0.016}$	1.32M
cosine schedule	-0.021	-0.012	1.32M
appended time channel	-0.019	-0.007	1.32M
lr 3×10^{-4}, patience 12	-0.016	-0.003	1.32M
$H = 128$, $m = 128$, lr 3×10^{-4}	$+0.012$	$+0.009$	0.50M

K Algorithms and code sketch

Three algorithms cover the method end to end: the offline spectrum generation that runs once before training (Algorithm 1), one forward pass of a layer (Algorithm 2), and the training loop (Algorithm 3). All three are written at the level of the main text and refer only to its equations. A code sketch follows.

Algorithm 1 Offline spectrum generation (either route), run once before training

Input: bases $b_1, \dots, b_J$ with block sizes $m_1 + \cdots + m_J = m$ ($J = 1$ for a single basis); route per block (SL or FA); floor $a_{\min}$
for $j = 1, \dots, J$ **do**
  if route is SL **then**
    $\lambda_k \leftarrow \lambda^{\mathrm{SL}}_{b_j,k}$ for $k = 0, \dots, m_j - 1$ from the closed forms of Table 1 (route 1)
  else
    with A_{H} the generator of the memory law for basis b_j, $\lambda_k \leftarrow |\mu_k(A_{\mathrm{H}})|^2$ sorted, as in (2) (route 2)
  end if
  $A^{(j)} \leftarrow \rho_j \operatorname{diag}(\lambda_0, \dots, \lambda_{m_j-1})$ with $\rho_j = 1/\max_k \lambda_k$ (3)
end for
$A_{\mathrm{grid}} \leftarrow \operatorname{diag}(A^{(1)}, \dots, A^{(J)})$, floored at $a_{\min}$ (union (3))
return A_{grid} frozen as a buffer, or $\hat{A} = \sqrt{\max(A_{\mathrm{grid}} - a_{\min}, 0)}$ if the spectrum is to be trained, so that $A = a_{\min} + \hat{A}^2$ starts on the generated spectrum

Algorithm 2 One SpecOSS layer (frozen or trainable generated spectrum)

Input: $u \in \mathbb{R}^{T\times H}$; spectrum A_{grid} from Algorithm 1; parameters $\hat{A}$ (if trainable), steps, B, C
$A \leftarrow$ frozen: A_{grid}; trainable: $a_{\min} + \hat{A}^2$
$\Delta t \leftarrow \sigma(\mathrm{steps})$; $S \leftarrow 1$ for IMEX, $S \leftarrow (1 + \Delta t^2 A)^{-1}$ for IM
$f \leftarrow uB^\top$ (the forcing Bu of the oscillator system $x'' = -Ax + Bu$, $y = Cx$)
advance that system over the sequence, one mode at a time and independently of the others,
$z \leftarrow S \odot (z - \Delta t A \odot x + \Delta t f_n)$, $x \leftarrow x + \Delta t z$
evaluated as an associative scan, $O(\log T)$ depth, chunked if T is large
return $\mathrm{Re}(xC^\top)$ (or the gated fusion of the per-block readouts, (3))

Algorithm 3 Training SpecOSS

Input: training and validation sets; bases and route; L, H, m; number of epochs; learning rate and schedule
$A_{\mathrm{grid}} \leftarrow$ Algorithm 1 (once, offline; shared by all L layers)
initialise the encoder, L residual blocks (each with B, C, steps and, if trainable, $\hat{A}$) and the head
for each epoch **do**
  for each mini-batch $(\mathbf{I}, \mathbf{O})$ **do**
    $\hat{\mathbf{O}} \leftarrow$ head $\circ$ the L residual blocks $\circ$ encoder, applied to $\mathbf{I}$, with Algorithm 2 inside every block
    loss $\leftarrow$ cross-entropy or mean-squared error of $\hat{\mathbf{O}}$ against $\mathbf{O}$
    clip the gradient norm at 1; Adam(W) step at the current rate of the warm-up–cosine schedule
  end for
  evaluate the validation metric; keep the parameters if it improved
end for
return the parameters of the best validation epoch (test metrics are computed once on them)

An intuitive sketch. The pseudocode below is the whole method in a page, in Python-like notation rather than any real framework: the spectrum generators of Algorithm 1, one layer written as a loop over time, the residual block and a training loop. Comments name the equation of the main text each line implements. The frozen spectrum is a plain buffer, so the only line that distinguishes a frozen from a trainable SpecOSS layer, or either from LinOSS, is where the diagonal comes from.

```
# ---------- the grid, built once before training ----------

def sl_grid(family, m):
    """Route 1 ladder."""
    k = range(m)
    return {"hermite":   [2*i + 1   for i in k],      # Table 1
            "legendre":  [i*(i + 1) for i in k],
            "chebyshev": [i*i       for i in k],
            "fourier":   [i*i       for i in k],
            "laguerre":  [i         for i in k]}[family]

def fa_grid(family, m):
    """Route 2 ladder."""
    A_H = generator(family, m)                        # the generator of (1)
    return sorted(abs(mu)**2 for mu in eigenvalues(A_H))          # (2)

def to_band(a, a_min):
    """Scale onto the shared band, then floor the zero mode."""
    return [max(ai / max(a), a_min) for ai in a]      # rho_j of (3), then the floor

def grid(families, sizes, route, a_min=0.01):
    """One diagonal carrying every family's ladder."""
    ladder = sl_grid if route == "SL" else fa_grid
    return concat(to_band(ladder(f, mj), a_min)       # (3)
                  for f, mj in zip(families, sizes))

# ---------- one layer:  x'' = -A x + B u,  y = C x ----------

class SpecOSSLayer:
    def __init__(self, grid, H, trainable=False, disc="imex", a_min=0.01):
        m = len(grid)
        if not trainable:
            self.A = buffer(grid)                     # A is never a parameter
        else:
            self.A_hat = parameter([sqrt(g - a_min) for g in grid])   # A = a_min + A_hat^2
        self.trainable, self.a_min, self.disc = trainable, a_min, disc
        self.steps = parameter(random(m))             # dt_k = sigmoid(steps_k)
        self.B = parameter(random(m, H))              # learned under either route
        self.C = parameter(random(H, m))

    def diagonal(self):
        if not self.trainable:
            return self.A
        return [self.a_min + a*a for a in self.A_hat]

    def forward(self, u):                             # u: (T, H)
        A, dt = self.diagonal(), sigmoid(self.steps)
        f = u @ self.B.T                              # the forcing B u
        S = 1 if self.disc == "imex" else 1 / (1 + dt*dt*A)   # IMEX conserves, IM damps
        z = x = zeros(len(A))                         # velocity and position
        out = []
        for n in range(len(u)):                       # a scan in practice
            z = S * (z - dt*A*x + dt*f[n])
            x = x + dt*z
            out.append(x)
        return real(stack(out) @ self.C.T)            # y = C x

# ---------- the block around it, held fixed across routes ----------

class Block:
    def __init__(self, layer, H, p_drop=0.05):
        self.layer, self.norm, self.drop = layer, batch_norm(H), dropout(p_drop)
        self.w1, self.w2 = linear(H, H), linear(H, H)

    def forward(self, u):
        y = self.drop(gelu(self.layer(self.norm(u))))
        return u + self.drop(self.w1(y) * sigmoid(self.w2(y)))    # GLU, then the skip

# ---------- build the grid once, then train ----------

A_grid = grid(["hermite", "legendre"], [8, 8], route="SL")   # or one family, J = 1
model = sequential(linear(d_in, H),
                   *[Block(SpecOSSLayer(A_grid, H, trainable=False), H) for _ in range(L)],
                   mean_over_time, linear(H, n_classes))

for u, target in batches:
    loss = cross_entropy(model(u), target)
    loss.backward()
    clip_gradient_norm(model, 1.0)
    optimiser.step()
```

Every variant passes through the same normalisation, floor, discretisation and block, and only the line that produces the diagonal changes between them, which is what makes the comparison in this paper controlled.