Learnable Koopman-Enhanced Transformer-Based Time Series Forecasting with Spectral Control

Anonymous authors
Paper under double-blind review

Abstract

This paper proposes a unified family of learnable Koopman operator parameterizations that integrate linear dynamical systems theory with modern deep learning forecasting architectures. We introduce four learnable Koopman variants—`scalar-gated`, `per-mode gated`, `MLP-shaped` spectral mapping, and `low-rank` Koopman operators—which generalize and interpolate between strictly stable Koopman operators and unconstrained linear latent dynamics. Our formulation enables explicit control over the spectrum, stability, and rank of the linear transition operator while retaining compatibility with expressive nonlinear backbones such as `PatchTST`, `Autoformer`, and `Informer`. We evaluate the proposed operators in a large-scale benchmark that also includes LSTM, DLinear, and simple diagonal State-Space Models (SSMs), as well as lightweight transformer variants. Experiments across multiple horizons and patch lengths show that learnable Koopman models provide a favorable bias–variance trade-off, improved conditioning, and more interpretable latent dynamics. We provide a full spectral analysis, including eigenvalue trajectories, stability envelopes, and learned spectral distributions. Our results demonstrate that learnable Koopman operators are effective, stable, and theoretically principled components for deep forecasting.

1 Introduction

Time series forecasting underpins applications in retail demand Wen et al. (2017); Salinas et al. (2020), traffic management Lv et al. (2014); Li et al. (2018), energy systems Dimoulkas et al. (2019); Saxena et al. (2019); Forootani et al. (2025), and finance Callot et al. (2017); Yan et al. (2018). The growing scale, dimensionality, and non-stationarity of modern data challenge classical statistical methods Hyndman & Athanasopoulos (2018), motivating deep learning approaches with greater expressive power and scalability. Early neural models—including RNNs/LSTMs Hochreiter & Schmidhuber (1997); Salinas et al. (2020) and CNN/TCN architectures LeCun & Bengio (1995); Bai et al. (2018)—learned temporal dependencies directly from data. Transformer-based models Wu et al. (2021b); Zhou et al. (2021a); Nie et al. (2022); Vaswani et al. (2017) later became dominant due to self-attention's ability to capture long-range dependencies, achieving strong results in benchmarks such as M4 and M5 Smyl (2020); Makridakis et al. (2021).

Numerous variants improve efficiency and structure: `Informer` Zhou et al. (2021b;a) proposes ProbSparse attention; `Autoformer` Wu et al. (2021a;b) incorporates trend–seasonal decomposition; `FEDformer` Zhou et al. (2022a;b) leverages Fourier mixing; and `Pyraformer` Liu et al. (2022a; 2021) models multi-scale hierarchies. In contrast, `PatchTST` Nie et al. (2022) demonstrates that simple design choices can outperform deeper Transformers Zheng et al. (2014); Zeng et al. (2022). Moreover, `DLinear` Zeng et al. (2022; 2023) reveals that even channel-wise linear models can rival complex architectures, raising a key question: *how much architectural complexity is truly necessary for accurate forecasting?* Despite their success, deep models often lack interpretability, struggle under distribution shifts Kuznetsov & Mohri (2020); Liu et al. (2022b); Kim et al. (2022), and fail to encode explicit dynamical structure, limiting robustness. Recent work incorporates domain priors through hybrid statistical–deep models Wang et al. (2019); Sen et al. (2019); Smyl (2020), neural ODE frameworks Chen et al. (2018); Vialard et al. (2020), and physics-inspired designs, though such approaches frequently rely on fixed dynamics assumptions and face scalability challenges in complex multivariate settings.

1.1 Koopman Operators and Deep Learning

The *Koopman operator* provides a linear perspective on nonlinear dynamical systems by evolving observables instead of states. As an infinite-dimensional linear operator on state functions, it enables spectral analysis of nonlinear, nonstationary, or chaotic systems Mezić (2017); Khosravi (2023). Data-driven methods such as Dynamic Mode Decomposition (DMD) approximate the Koopman spectrum for prediction and modal analysis Williams et al. (2015); Kutz et al. (2016); Drgona et al. (2022); Skomski et al. (2021); Drgona et al. (2021), but often degrade in high-dimensional, noisy, or multivariate settings. Deep learning has renewed interest in Koopman theory by learning nonlinear embeddings with linear latent evolution. Neural Koopman models typically use an encoder $\mathcal{E}\theta$ to map an input window X_t to latent coordinates $z_t = \mathcal{E}\theta(X_t)$, followed by a linear propagator $z_{t+1} = \mathcal{K}z_t$, often combined with nonlinear decoders and spectral or stability constraints Lusch et al. (2018a); Yeung et al. (2019); Takeishi et al. (2017a); Lusch et al. (2018b); Azencot et al. (2020); Morton et al. (2018); Takeishi et al. (2017b). These approaches improve interpretability and rollout stability, with extensions incorporating low-rank structure, symmetry priors, and multiresolution representations Brunton et al. (2021). However, most remain architecturally shallow and struggle with long-range, non-local dependencies effectively handled by Transformer models.

Recent work integrates Koopman operators into Transformer architectures to achieve stable and interpretable latent dynamics, notably `DeepKoopFormer` Forootani et al. (2026); Forootani & Khosravi (2026). Unlike prior Koopman-based forecasting models Lusch et al. (2018b); Azencot et al. (2020), `DeepKoopFormer` provides provable spectral decay, Lyapunov-stable propagation, normality of the transition matrix, and closed-form error bounds within a flexible encoder–propagator–decoder framework aligned with physics-informed forecasting Karniadakis et al. (2021); Nghiem et al. (2023) and interpretable dynamical systems modeling Brunton et al. (2016).

1.2 Contributions: Learnable Koopman–Transformer Framework

Although promising, most Koopman–neural models rely on fixed or heavily constrained operators, leaving much of the learnable operator space unexplored. Existing approaches typically pair nonlinear encoders with a latent linear operator Takeishi et al. (2017b); Lusch et al. (2018b), but the operator is usually learned *without explicit spectral parameterization*—implemented as an unconstrained linear layer or least-squares fit. This offers no direct control over contraction, oscillation, or multi-scale temporal dynamics. Furthermore, such designs are rarely integrated with modern Transformer architectures, despite their strength in modeling long-range dependencies. The resulting combination of *spectral non-control* and *architectural underuse* limits both interpretability and dynamical expressiveness.

To address these limitations, we introduce the `Learnable-DeepKoopFormer`, a unified framework that equips Transformer backbones with *parameterized Koopman operators* whose spectra are learned directly from data. Our formulation includes four operator families: (i) `scalar-gated`: global trainable parameters (α, β) control spectral shifting and scaling, enabling tunable damping or persistent dynamics; (ii) `per-mode gated`: dimension-wise parameters (α_i, β_i) model anisotropic temporal responses and multi-frequency evolution; (iii) `MLP-shaped` spectral mapping: neural spectral mappings are projected onto stable linear propagators with controlled spectral radius; (iv) `low-rank` Koopman: relatively low rank factorizations of the Koopman operator to capture high dimensional latent dynamics efficiently through structured compression. These operator parameterizations are integrated into both lightweight and full Transformer forecasters—`PatchTST` Nie et al. (2022), `Informer` Zhou et al. (2021a), and `Autoformer` Wu et al. (2021b)—forming a unified family of `DeepKoopFormer` models. Spectral stability constraints and operator regularization enable joint optimization of latent dynamics and representations while retaining interpretability, as dominant modes, stability margins, and spectral envelopes remain directly analyzable.

We conduct an extensive simulation benchmark comparing the proposed learnable Koopman variants against two Koopman baselines (spectrally constrained and unconstrained), as well as standard forecasting models including LSTM, `DLinear` Zeng et al. (2022; 2023), and a diagonal SSM Gu et al. (2021b). Beyond forecasting accuracy, we present the first systematic *spectral evaluation* of Koopman–Transformers, analyzing eigenvalue densities, training trajectories, stability envelopes, and horizon sensitivity. In summary, this work provides: (i) a unified parameterization of *learnable Koopman operators* (scalar, per-mode, neural, and low-rank); (ii)

the `Learnable-DeepKoopFormer` framework, embedding these operators into three Transformer backbones; (iii) a benchmark of *21 models* spanning constrained, learnable, and unconstrained variants, plus LSTM, `DLinear`, and diagonal SSM baselines; (iv) the first comprehensive *spectral analysis of Koopman–Transformer models*, revealing stability, multi-scale expressiveness, and interpretable latent dynamics.

For real-world evaluation, we utilize high-dimensional datasets from multiple domains, including: the CMIP6 climate projections [1], focusing on wind speed and surface pressure forecasting over Germany; a financial time series dataset [2] for cryptocurrency market analysis; and an electricity generation dataset [3] for modeling energy supply dynamics. These datasets collectively span chaotic, periodic, and stochastic regimes, allowing for a comprehensive assessment of model accuracy, stability, and generalization across domains.

Compared to existing Koopman-based forecasting models Lusch et al. (2018b); Takeishi et al. (2017b); Azencot et al. (2020), `Learnable-DeepKoopFormer` is the first to offer provable spectral decay, Lyapunov-stable propagation, normality of the transition matrix, and closed-form error control—all within a general, extensible encoder–propagator–decoder pipeline aligned with the goals of physics-informed forecasting Karniadakis et al. (2021); Nghiem et al. (2023) and interpretable dynamical systems modeling Brunton et al. (2016); Korda & Mezić (2016).

This paper is organized as follows. Section 2 provides the preliminaries required for time series forecasting. In Section 3, we present `Learnable-DeepKoopFormer` architectures for multivariate time series forecasting. Numerical simulations are presented in Section 4 to evaluate the performance of `Learnable-DeepKoopFormer`versus other benchmarks. Finally, we conclude the article in Section 5.

2 `Learnable-DeepKoopFormer` Architecture

We begin by formalizing the multivariate time series forecasting task and introducing the key components of the proposed architecture. Given a multivariate time series $\{x_t\}_{t=1}^{T}$, with $x_t \in \mathbb{R}^{d_s}$, `Learnable-DeepKoopFormer` operates on context windows of length P and predicts the subsequent H samples. For each valid index t, an input segment $X_t = [x_t, \ldots, x_{t+P-1}] \in \mathbb{R}^{P \times d_s}$ is paired with its future sequence $Y_t = [x_{t+P}, \ldots, x_{t+P+H-1}] \in \mathbb{R}^{H \times d_s}$. The input window is passed through a Transformer encoder with positional encoding $\mathcal{E}_\theta$ (e.g., `PatchTST`, `Autoformer`, or `Informer`), producing a latent representation $z_t = \mathcal{E}_\theta(X_t) \in \mathbb{R}^{d_{\text{lat}}}$, where d_{lat} is the latent dimension. Typically $d_{\text{lat}} = d_{\text{model}}$, and not necessarily equal to the physical dimension d_s, where d_{model} denotes the hidden (embedding) dimension of the Transformer backbone, i.e., the dimensionality of the token representations and of the latent state produced by the encoder $\mathcal{E}_\theta$. With the slight abuse of notation and for the sake of simplicity, hereinafter we assume $d_{\text{model}} = d_{\text{lat}} = d_s = d$.

Temporal evolution in the latent space is modeled through a learned Koopman operator $\mathcal{K}_\phi$, which advances the latent state linearly according to $z_{t+1} = \mathcal{K}_\phi z_t$. The propagated latent state is decoded into a direct H-step forecast using a linear mapping $\mathcal{D}_\varphi$, yielding a vector $\hat{y}_t = \mathcal{D}_\varphi(z_{t+1}) \in \mathbb{R}^{H \cdot d}$, which is then reshaped into the output sequence $\hat{Y}_t \in \mathbb{R}^{H \times d}$.

Here, the subscript θ denotes the parameters of the Transformer encoder $\mathcal{E}_\theta$, ϕ collects the parameters of the Koopman propagator $\mathcal{K}_\phi$ (including its spectral coefficients and orthogonal factors), and φ denotes the parameters of the linear decoder $\mathcal{D}_\varphi$. During inference, a context window X_t is encoded once, propagated through the Koopman operator $\mathcal{K}_\phi$, and decoded to produce the entire prediction horizon H in a single forward pass, without autoregressive rollout.

All components are trained jointly in an end-to-end fashion using the Adam optimizer Kingma & Ba (2015), a first-order stochastic gradient method with adaptive moment estimation. In particular, training is performed end-to-end by minimizing a forecasting loss (e.g., mean squared error) augmented with a Lyapunov-inspired

[1] https://cds.climate.copernicus.eu/datasets/projections-cordex-domains-single-levels?tab=overview
[2] https://github.com/Chisomnwa/Cryptocurrency-Data-Analysis
[3] https://github.com/afshinfaramarzi/Energy-Demand-electricity-price-Forecasting/tree/main

penalty as follows

$$\begin{aligned}\mathcal{L} &= \mathbb{E}\Big[\|\widehat{Y}_t - Y_t\|^2\Big] \;+\; \lambda_{\text{Lyap}}\,\mathbb{E}\Big[\big\|\mathcal{K}_\phi \boldsymbol{z}_t\big\|_P^2 - \big\|z_t\big\|_P^2\Big]_+ \\ &= \mathcal{L}_{\text{MSE}} \;+\; \lambda_{\text{Lyap}}\mathcal{L}_{\text{Lyap}},\end{aligned} \tag{1}$$

where $\|z\|_P^2 = z^\top P z$, $P \succ 0$ is a fixed positive-definite matrix, $[x]_+ = \max\{x, 0\}$, and $\lambda_{\text{Lyap}} > 0$ is a scalar regularization weight controlling the strength of the Lyapunov stability penalty.

3 Learnable Koopman Variants

In `Learnable-DeepKoopFormer` we assume the latent evolution to be modeled by a linear dynamics through a learned Koopman propagator $\mathcal{K}_\phi \in \mathbb{R}^{d\times d}$, $z_{t+1} = \mathcal{K}_\phi z_t$. As shown later in the paper, to ensure identifiability and stable training, $\mathcal{K}_\phi$ is parameterized through an orthogonal–diagonal–orthogonal (ODO) factorization

$$\mathcal{K}_\phi \;=\; U_\phi \,\mathrm{diag}(\Sigma_\phi)\, V_\phi^\top, \tag{2}$$

where U_ϕ, V_ϕ are learned orthonormal matrices obtained via QR retraction, and $\Sigma_\phi = (\Sigma_{\phi,1}, \ldots, \Sigma_{\phi,d})^\top \in \mathbb{R}^d$ (or $\mathbb{R}^r$ in the low-rank case, see later in the paper) the vector containing the learnable spectral coefficients. Equation (2) ensures that all nonlinearity and expressiveness are concentrated in the spectrum Σ_ϕ, while the left/right singular directions evolve on the Stiefel manifold, i.e., the set of matrices with orthonormal columns Absil et al. (2008).

The key design freedom of our *learnable Koopman architecture* lies in how the diagonal spectral coefficients Σ_ϕ are generated. We consider Koopman variants that share the principle $\Sigma_{\phi,i} \;=\; \rho_{\max}\,\sigma_\phi\big(S_i\big), \qquad i = 1, \ldots, d,$ where $S_i \in \mathbb{R}$ are raw trainable parameters, $\rho_{\max} \in (0, 1)$ is a prescribed spectral bound, and $\sigma_\phi(\cdot)$ is a smooth squashing map, i.e. a nonlinear "spectral shaping" function, which preserves differentiability and enables end-to-end learning while providing explicit control over contraction rates in the latent Koopman dynamics. This map is chosen to enforce

$$|\Sigma_{\phi,i}| < \rho_{\max} \quad \text{for all } i, \qquad \rho_{\max} \in (0, 1), \tag{3}$$

whenever spectral stability is desired, while allowing different degrees of expressiveness. In particular, it is the composition of the logistic sigmoid $\sigma(x) = \frac{1}{1+e^{-x}}$ and a function chosen from a list of spectral shaping maps. More specifically, we introduce four learnable Koopman families that act on the vector of raw spectral parameters (i) `scalar-gated`. A shared affine gate controls the entire spectrum:

$$\sigma_\phi(S_i) \;=\; \sigma(\alpha S_i + \beta), \tag{4}$$

with $\alpha, \beta \in \mathbb{R}$ learned scalars.

(ii) `per-mode gated`. Each latent direction receives its own gate,

$$\sigma_\phi(S_i) \;=\; \sigma(\alpha_i S_i + \beta_i), \tag{5}$$

allowing anisotropic amplification/decay across modes.

(iii) `MLP-shaped` spectral mapping. A small neural operator $g_\phi : \mathbb{R} \to \mathbb{R}$ transforms raw spectral parameters before squashing:

$$\sigma_\phi(S_i) \;=\; \sigma(g_\phi(S_i))\,, \tag{6}$$

enabling flexible nonlinear reshaping of the spectrum.

(iv) `low-rank` Koopman. The operator is restricted to rank $r \ll d$ via

$$\mathcal{K}_\phi \;=\; U_{r,\phi}\,\mathrm{diag}(\Sigma_{r,\phi})\,V_{r,\phi}^\top, \quad U_{r,\phi}, V_{r,\phi} \in \mathbb{R}^{d\times r}, \tag{7}$$

with orthonormal columns and a spectral vector $\Sigma_{r,\phi} \in \mathbb{R}^r$ satisfying again $|\Sigma_{r,\phi,i}| < \rho_{\max}$. It encourages low-dimensional latent dynamics and reduces complexity.

In addition, for comparison purpose, we also consider a fully free operator (unconstrained baseline) obtained when $\mathcal{K}_\phi$ is a free dense matrix, without the ODO structure (2) and without the spectral bound (3). This variant maximizes expressiveness but lacks the stability and dynamical bias of the structured families above. Overall, the learnable Koopman architecture provides a continuum between rigorously constrained operator-theoretic propagators and fully flexible data-driven linear maps, allowing us to isolate and study the role of spectral structure in Transformer-based forecasting models.

3.1 Theoretical Properties of the Learnable Koopman Propagator

We consider latent states $\boldsymbol{z}_t \in \mathbb{R}^d$ evolving under a linear propagator

$$z_{t+1} = \mathcal{K}_\phi z_t, \qquad \mathcal{K}_\phi \in \mathbb{R}^{d\times d}, \tag{8}$$

where $\mathcal{K}_\phi$ is parameterized by the learnable weights ϕ of the Koopman module in `DeepKoopFormer`. This subsection collects several structural properties of the proposed parameterizations, focusing on stability, expressiveness, low-rank structure, and Lyapunov regularization.

3.1.1 Spectral stability, contraction, and invertibility

The spectral bound (3) directly implies stability and contractivity of the associated linear dynamics.

Proposition 1 (Spectral stability)**.** *Let $\mathcal{K}_\phi$ be parameterized as in (2), with orthonormal U_ϕ, V_ϕ and diagonal $\mathrm{diag}(\Sigma_\phi)$ satisfying (3). Then $\rho(\mathcal{K}_\phi) \le \|\mathcal{K}_\phi\|_2 = \max_i|\Sigma_{\phi,i}| < \rho_{\max}$, where $\rho(\cdot)$ denotes the spectral radius and $\|\cdot\|_2$ the spectral norm.*

Proof. Because U_ϕ and V_ϕ are orthonormal, left and right multiplication do not change the spectral norm:

$$\|\mathcal{K}_\phi\|_2 = \left\|U_\phi \,\mathrm{diag}(\Sigma_\phi)\, V_\phi^\top\right\|_2 = \left\|\mathrm{diag}(\Sigma_\phi)\right\|_2. \tag{9}$$

The spectral norm of a diagonal matrix is the maximum absolute diagonal entry, hence $\|\mathcal{K}_\phi\|_2 = \max_i|\Sigma_{\phi,i}|$. For any matrix, the spectral radius is bounded above by the spectral norm, $\rho(\mathcal{K}_\phi) \le \|\mathcal{K}_\phi\|_2$. Combining this with $|\Sigma_{\phi,i}| < \rho_{\max}$ for all i yields the claim. □

Corollary 1 (Exponential contraction in latent space)**.** *Under the assumptions of Proposition 1, for any $\boldsymbol{z}_0 \in \mathbb{R}^d$ and any $n \in \mathbb{N}$, $\|\mathcal{K}_\phi^n \boldsymbol{z}_0\|_2 \le \|\mathcal{K}_\phi\|_2^n \|z_0\|_2 \le \rho_{\max}^n \|z_0\|_2$, and hence $\lim_{n\to\infty} \mathcal{K}_\phi^n z_0 = 0$.*

Proof. Submultiplicativity of the spectral norm gives $\|\mathcal{K}_\phi^n\|_2 \le \|\mathcal{K}_\phi\|_2^n$, so $\|\mathcal{K}_\phi^n \boldsymbol{z}_0\|_2 \le \|\mathcal{K}_\phi^n\|_2 \|\boldsymbol{z}_0\|_2 \le \|\mathcal{K}_\phi\|_2^n \|\boldsymbol{z}_0\|_2$. Proposition 1 implies $\|\mathcal{K}_\phi\|_2 < \rho_{\max} < 1$, hence the right-hand side converges to zero as $n \to \infty$, i.e. the latent dynamics (8) is exponentially stable:

$$\|z_t\|_2 \le \rho_{\max}^t \|z_0\|_2, \quad t \ge 0. \tag{10}$$

□

In many forecasting settings, it is also natural to consider the inverse dynamics whenever the spectrum is bounded away from zero. This leads to the following simple consequence of the ODO structure.

Proposition 2 (Invertibility and stability of the inverse)**.** *Assume the ODO parameterization (2) satisfies the two-sided spectral bound*

$$0 < \rho_{\min} \le |\Sigma_{\phi,i}| \le \rho_{\max} < 1 \quad \textit{for all } i. \tag{11}$$

Then $\mathcal{K}_\phi$ is invertible with $\mathcal{K}_\phi^{-1} = V_\phi \,\mathrm{diag}\big(\Sigma_{\phi,1}^{-1}, \ldots, \Sigma_{\phi,d}^{-1}\big)\, U_\phi^\top$, and $\rho(\mathcal{K}_\phi^{-1}) \le \|\mathcal{K}_\phi^{-1}\|_2 = \max_i|\Sigma_{\phi,i}^{-1}| \le \rho_{\min}^{-1}$.

Proof. Under (11), all diagonal entries of $\mathrm{diag}(\Sigma_\phi)$ are nonzero, so it is invertible. Since U_ϕ and V_ϕ are orthonormal, $\mathcal{K}_\phi$ is invertible and its inverse has the stated form. The spectral norm calculation proceeds exactly as in Proposition 1, yielding $\|\mathcal{K}_\phi^{-1}\|_2 = \max_i |\Sigma_{\phi,i}^{-1}|$, and $\rho(\mathcal{K}_\phi^{-1}) \le \|\mathcal{K}_\phi^{-1}\|_2$. □

This shows that the same spectral factors that enforce forward-time contraction can also guarantee that the inverse dynamics are well-conditioned, provided the spectrum is kept away from zero.

3.1.2 Expressiveness of spectrally bounded and low-rank operators

The ODO factorization parameterizes exactly the class of linearly stable operators with bounded spectral norm.

Proposition 3 (Representation of all spectrally bounded operators)**.** *Fix $\rho_{\max} > 0$ and define $\mathcal{K}(\rho_{\max}) := \left\{\mathcal{K} \in \mathbb{R}^{d\times d} \,:\, \|\mathcal{K}\|_2 \le \rho_{\max}\right\}$. Then the set of matrices that admit an ODO factorization with $|\Sigma_{\phi,i}| \le \rho_{\max}$ is precisely $\mathcal{K}(\rho_{\max})$. In particular, any $\mathcal{K}$ with $\|\mathcal{K}\|_2 < \rho_{\max}$ can be represented exactly by some choice of $(U_\phi, V_\phi, \Sigma_\phi)$.*

Proof. (*Surjectivity onto $\mathcal{K}(\rho_{\max})$.*) Let $\mathcal{K} \in \mathcal{K}(\rho_{\max})$. By the singular value decomposition, there exist orthonormal matrices $U, V \in \mathbb{R}^{d\times d}$ and nonnegative singular values $\Sigma_1, \dots, \Sigma_d$ such that $\mathcal{K} = U\,\mathrm{diag}(\Sigma_1, \dots, \Sigma_d)\,V^\top$. The spectral norm is $\|\mathcal{K}\|_2 = \max_i \Sigma_i$, so $\|\mathcal{K}\|_2 \le \rho_{\max}$ implies $\Sigma_i \le \rho_{\max}$ for all i. Thus $\mathcal{K}$ admits an ODO factorization with $U_\phi = U$, $V_\phi = V$ and $\Sigma_{\phi,i} = \Sigma_i$ satisfying $|\Sigma_{\phi,i}| \le \rho_{\max}$.

(*Inclusion in $\mathcal{K}(\rho_{\max})$.*) Conversely, let $\mathcal{K}_\phi = U_\phi \mathrm{diag}(\Sigma_\phi) V_\phi^\top$ with $\max_i |\Sigma_{\phi,i}| \le \rho_{\max}$. As in the proof of Proposition 1,

$$\|\mathcal{K}_\phi\|_2 = \left\|\mathrm{diag}(\Sigma_\phi)\right\|_2 = \max_i |\Sigma_{\phi,i}| \le \rho_{\max}, \tag{12}$$

hence $\mathcal{K}_\phi \in \mathcal{K}(\rho_{\max})$. □

Thus the strictly stable and learnable Koopman variants with $|\Sigma_{\phi,i}| < \rho_{\max}$ retain the full expressiveness of all linear operators in the spectrally bounded class $\mathcal{K}(\rho_{\max})$, while providing an explicit stability margin.

3.1.3 Low-rank structure and approximation

For the low-rank family (7), the Koopman operator is restricted to a rank–r subspace of $\mathbb{R}^d$.

Proposition 4 (Low-rank structure and norm bound)**.** *For the low-rank Koopman parameterization (7): (i) $\mathrm{rank}(\mathcal{K}_\phi) \le r$; (ii) $\|\mathcal{K}_\phi\|_2 = \max_{i\le r} |\Sigma_{r,\phi,i}|$; (iii) if $|\Sigma_{r,\phi,i}| < \rho_{\max}$ for all i, then $\rho(\mathcal{K}_\phi) \le \|\mathcal{K}_\phi\|_2 < \rho_{\max}$.*

Proof. (i) The product of a $d \times r$ matrix, an $r \times r$ diagonal matrix and an $r \times d$ matrix has rank at most r. (ii) Since $U_{r,\phi}$ and $V_{r,\phi}$ have orthonormal columns, they describe partial isometries from $\mathbb{R}^r$ into $\mathbb{R}^d$, and the nonzero singular values of $\mathcal{K}_\phi$ coincide with the absolute values of the entries of $\Sigma_{r,\phi}$. Hence $\|\mathcal{K}_\phi\|_2 = \max_{i\le r} |\Sigma_{r,\phi,i}|$. (iii) The last claim follows from (ii) together with $\rho(\mathcal{K}_\phi) \le \|\mathcal{K}_\phi\|_2$ and the bound $|\Sigma_{r,\phi,i}| < \rho_{\max}$. □

Let $\mathcal{K}_\phi$ have singular values $\Sigma_{\phi,1} \ge \cdots \ge \Sigma_{\phi,d} \ge 0$, and let $\mathcal{K}_\phi^{(r)}$ denote the truncation of its singular value decomposition to rank r, i.e., $\mathcal{K}_\phi^{(r)}$ retains only the first r singular values and vectors. The Eckart–Young theorem (see Lin (2011) for more details) states that $\mathcal{K}_\phi^{(r)}$ minimizes $\|\mathcal{K}_\phi - \tilde{\mathcal{K}}_\phi\|_F$ over all rank–r matrices $\tilde{\mathcal{K}}_\phi$. Any such $\mathcal{K}_\phi^{(r)}$ admits a representation of the form (7), and thus belongs to the low-rank family. Consequently, whenever the latent Koopman dynamics are approximately low-dimensional, the low-rank variant can capture the principal modes while discarding high-rank noise.

3.1.4 Lyapunov regularization and energy decay

Besides spectral constraints, the training objective incorporates in (1) a Lyapunov-inspired penalty that encourages contractive behavior with respect to a fixed quadratic form. For latent pairs $(\boldsymbol{z}_t, \boldsymbol{z}_{t+1})$ the addition of the term $\mathcal{L}_{\mathrm{Lyap}} = \mathbb{E}\Big[\big(\|\boldsymbol{z}_{t+1}\|_P^2 - \|\boldsymbol{z}_t\|_P^2\big)_+\Big]$, $\|\boldsymbol{z}\|_P^2 = \boldsymbol{z}^\top P \boldsymbol{z}$, with $(x)_+ = \max\{x, 0\}$ and $P \succ 0$, provides a common regularizer across learnable, Koopman variants, which further biases optimisation toward strict contraction of latent energy by penalizing unstable latent growth and promoting Koopman-consistent dynamics. The following results formalize these considerations.

Proposition 5 (Lyapunov consistency)**.** *Suppose $P \succ 0$ and the support of $\boldsymbol{z}_t$ spans $\mathbb{R}^d$. If $\mathcal{L}_{\mathrm{Lyap}} = 0$, then $\mathcal{K}_\phi$ satisfies the discrete Lyapunov inequality $\mathcal{K}_\phi^\top P \mathcal{K}_\phi - P \preceq 0$. In particular, the quadratic energy $\|\boldsymbol{z}_t\|_P^2$ is non-increasing under the dynamics $\boldsymbol{z}_{t+1} = \mathcal{K}_\phi \boldsymbol{z}_t$.*

Proof. The equality $\mathcal{L}_{\text{Lyap}} = 0$ implies $\left(\|\boldsymbol{z}_{t+1}\|_P^2 - \|\boldsymbol{z}_t\|_P^2\right)_+ = 0$, almost surely, and thus $\|\boldsymbol{z}_{t+1}\|_P^2 \leq \|\boldsymbol{z}_t\|_P^2$, for all $\boldsymbol{z}_t$ in the support. Using $\boldsymbol{z}_{t+1} = \mathcal{K}_\phi \boldsymbol{z}_t$, $\boldsymbol{z}_t^\top \mathcal{K}_\phi^\top P \mathcal{K}_\phi \boldsymbol{z}_t \leq \boldsymbol{z}_t^\top P \boldsymbol{z}_t$, or equivalently $\boldsymbol{z}_t^\top (\mathcal{K}_\phi^\top P \mathcal{K}_\phi - P)\boldsymbol{z}_t \leq 0$, for all $\boldsymbol{z}_t$ in a set that spans $\mathbb{R}^d$. A symmetric matrix M satisfies $M \preceq 0$ if and only if $\boldsymbol{z}^\top M \boldsymbol{z} \leq 0$ for all $\boldsymbol{z}$, hence $\mathcal{K}_\phi^\top P \mathcal{K}_\phi - P \preceq 0$. □

Conversely, classical results in linear systems theory relate such a Lyapunov inequality to spectral stability.

Proposition 6 (Spectral stability from a Lyapunov certificate)**.** *Let $\mathcal{K} \in \mathbb{R}^{d\times d}$ and suppose there exists $P \succ 0$ such that $\mathcal{K}^\top P \mathcal{K} - P \prec 0$. Then $\rho(\mathcal{K}) < 1$.*

Proof. Since $P \succ 0$, write $P = R^\top R$ for some invertible $R \in \mathbb{R}^{d\times d}$ and consider the similarity transform $\tilde{\mathcal{K}} = R\mathcal{K}R^{-1}$. The Lyapunov inequality becomes

$$\mathcal{K}^\top P \mathcal{K} - P = \mathcal{K}^\top R^\top R \mathcal{K} - R^\top R = R^\top (\tilde{\mathcal{K}}^\top \tilde{\mathcal{K}} - I) R \prec 0. \tag{13}$$

Premultiplying and postmultiplying by $(R^{-1})^\top$ and R^{-1} yields $\tilde{\mathcal{K}}^\top \tilde{\mathcal{K}} - I \prec 0$, which implies $\|\tilde{\mathcal{K}}\boldsymbol{x}\|_2^2 < \|\boldsymbol{x}\|_2^2$ for all $\boldsymbol{x} \neq 0$, hence $\|\tilde{\mathcal{K}}\|_2 < 1$. Therefore $\rho(\tilde{\mathcal{K}}) < 1$, and since $\mathcal{K}$ and $\tilde{\mathcal{K}}$ are similar, $\rho(\mathcal{K}) = \rho(\tilde{\mathcal{K}}) < 1$. □

For finite Lyapunov regularization weight, the penalty $\mathcal{L}_{\text{Lyap}}$ need not vanish exactly; nevertheless, Propositions 5–6 show that it biases the learned propagator toward the set of operators admitting a Lyapunov certificate of stability, in a way that complements the direct spectral constraints imposed by the ODO parameterization.

Taken together, Propositions 1–6 show that the learnable Koopman family used in `Learnable-DeepKoopFormer`: (i) provides a clear, tunable stability margin via $\rho_{\max}$ and the Lyapunov weight; (ii) retains the full expressiveness of spectrally bounded linear dynamics; (iii) admits low-rank specializations capturing dominant latent modes; and (iv) supports well-conditioned inverse propagation when the spectrum is bounded away from zero. These guarantees hold uniformly across all instantiations of `Learnable-DeepKoopFormer` built on top of the `PatchTST`, `Autoformer`, and `Informer` backbones.

The following result shows that spectral stability of the Koopman propagator directly controls the sensitivity of forecasts to perturbations in the input history through the encoder–Koopman–decoder pipeline.

Proposition 7 (Finite-horizon forecast error bound)**.** *Let the forecasting model in `Learnable-DeepKoopFormer` be defined as $\widehat{Y}_t^{(h)}(X_t) = \mathcal{D}_\varphi\left(\mathcal{K}_\phi^h \mathcal{E}_\theta(X_t)\right)$, where $\mathcal{E}_\theta$ denotes the encoder, $\mathcal{K}_\phi$ is the Koopman propagator, and $\mathcal{D}_\varphi$ is the decoder.*

Assume that: (i) the encoder $\mathcal{E}_\theta$ is Lipschitz continuous with constant $L_\mathcal{E} > 0$, i.e. $\|\mathcal{E}_\theta(X) - \mathcal{E}_\theta(\tilde{X})\|_2 \leq L_\mathcal{E}\|X - \tilde{X}\|_2$, (ii) the decoder $\mathcal{D}_\varphi : \mathbb{R}^d \to \mathbb{R}^{H\cdot d}$ is Lipschitz continuous with constant $L_\mathcal{D} > 0$, (iii) the Koopman propagator satisfies $\|\mathcal{K}_\phi\|_2 \leq \rho_{\max} < 1$.

Then for any two input histories $X_t, \tilde{X}_t$ and any horizon $h \geq 0$,

$$\begin{aligned}\left\|\widehat{Y}_t^{(h)}(X_t) - \widehat{Y}_t^{(h)}(\tilde{X}_t)\right\|_2 &\leq L_\mathcal{D}\|\mathcal{K}_\phi^h\|_2 \|\mathcal{E}_\theta(X_t) - \mathcal{E}_\theta(\tilde{X}_t)\|_2 \\ &\leq L_\mathcal{D} L_\mathcal{E} \|\mathcal{K}_\phi^h\|_2 \|X_t - \tilde{X}_t\|_2 \\ &\leq L_\mathcal{D} L_\mathcal{E} \rho_{\max}^h \|X_t - \tilde{X}_t\|_2.\end{aligned}$$

In particular, the influence of perturbations in the input history on the forecast decays at least geometrically with the prediction horizon.

Proof. Starting from the model definition, $\widehat{Y}_t^{(h)}(X_t) = \mathcal{D}_\varphi\left(\mathcal{K}_\phi^h \mathcal{E}_\theta(X_t)\right)$.

Hence

$$\left\|\widehat{Y}_t^{(h)}(X_t) - \widehat{Y}_t^{(h)}(\tilde{X}_t)\right\|_2 = \left\|\mathcal{D}_\varphi(\mathcal{K}_\phi^h \mathcal{E}_\theta(X_t)) - \mathcal{D}_\varphi(\mathcal{K}_\phi^h \mathcal{E}_\theta(\tilde{X}_t))\right\|_2.$$

By Lipschitz continuity of the decoder $\mathcal{D}_\varphi$,

$$\big\|\widehat{Y}_t^{(h)}(X_t) - \widehat{Y}_t^{(h)}(\tilde{X}_t)\big\|_2 \le L_{\mathcal{D}} \big\|\mathcal{K}_\phi^h \mathcal{E}_\theta(X_t) - \mathcal{K}_\phi^h \mathcal{E}_\theta(\tilde{X}_t)\big\|_2.$$

Using linearity of the Koopman propagator,

$$\mathcal{K}_\phi^h \mathcal{E}_\theta(X_t) - \mathcal{K}_\phi^h \mathcal{E}_\theta(\tilde{X}_t) = \mathcal{K}_\phi^h\big(\mathcal{E}_\theta(X_t) - \mathcal{E}_\theta(\tilde{X}_t)\big),$$

hence

$$\big\|\widehat{Y}_t^{(h)}(X_t) - \widehat{Y}_t^{(h)}(\tilde{X}_t)\big\|_2 \le L_{\mathcal{D}} \big\|\mathcal{K}_\phi^h\big(\mathcal{E}_\theta(X_t) - \mathcal{E}_\theta(\tilde{X}_t)\big)\big\|_2.$$

By the induced spectral norm inequality,

$$\big\|\widehat{Y}_t^{(h)}(X_t) - \widehat{Y}_t^{(h)}(\tilde{X}_t)\big\|_2 \le L_{\mathcal{D}} \|\mathcal{K}_\phi^h\|_2 \, \|\mathcal{E}_\theta(X_t) - \mathcal{E}_\theta(\tilde{X}_t)\|_2.$$

Applying the Lipschitz property of the encoder,

$$\big\|\widehat{Y}_t^{(h)}(X_t) - \widehat{Y}_t^{(h)}(\tilde{X}_t)\big\|_2 \le L_{\mathcal{D}} L_{\mathcal{E}} \|\mathcal{K}_\phi^h\|_2 \|X_t - \tilde{X}_t\|_2.$$

Finally, by submultiplicativity of the spectral norm, $\|\mathcal{K}_\phi^h\|_2 \le \|\mathcal{K}_\phi\|_2^h$, and since $\|\mathcal{K}_\phi\|_2 \le \rho_{\max} < 1$, $\|\mathcal{K}_\phi^h\|_2 \le \rho_{\max}^h$. Substituting yields

$$\big\|\widehat{Y}_t^{(h)}(X_t) - \widehat{Y}_t^{(h)}(\tilde{X}_t)\big\|_2 \le L_{\mathcal{D}} L_{\mathcal{E}} \rho_{\max}^h \|X_t - \tilde{X}_t\|_2,$$

which proves the claim. □

Proposition 7 shows that spectral control of the Koopman propagator yields direct control over the sensitivity of forecasts to perturbations in the input history. In particular, whenever $\|\mathcal{K}_\phi\|_2 < 1$, perturbations propagate through the encoder–Koopman–decoder pipeline with a bounded amplification factor. Under the Lipschitz assumptions on the encoder and decoder, discrepancies in the input sequence are attenuated geometrically under repeated Koopman propagation, and their influence on the decoded forecast decreases at least at rate $\rho_{\max}^h$, up to the multiplicative constants $L_{\mathcal{E}}$ and $L_{\mathcal{D}}$. This provides a theoretical explanation for the improved robustness and tighter error distributions observed empirically for the constrained and learnable Koopman variants.

3.1.5 Theoretical stability advantage of the Koopman layer

We compare the latent dynamics induced by the proposed Koopman layer with those of an unconstrained linear state–space model (SSM). We have shown how all `Learnable-DeepKoopFormer` variants propagate a latent state $z_t \in \mathbb{R}^d$ according to $z_{t+1} = \mathcal{K}_\phi \, z_t$, where $\mathcal{K}_\phi$ is parameterised via an orthogonal–diagonal–orthogonal (ODO) factorisation, $\mathcal{K}_\phi = U_\phi \operatorname{diag}\big(\Sigma_\phi\big) V_\phi^\top$. Here $U_\phi, V_\phi \in \mathbb{R}^{d\times r}$ have orthonormal columns, obtained through QR retraction, and the spectral coefficients $\Sigma_\phi \in \mathbb{R}^r$ are generated from unconstrained raw parameters S via a componentwise squashing map $\Sigma_{i,\phi}(\theta) = \rho_{\max}\, \sigma\big(S_i\big), \quad \sigma : \mathbb{R} \to (0,1), \quad 0 < \rho_{\max} < 1$. In contrast, the SSM baseline evolves a hidden state $h_t \in \mathbb{R}^{d_h}$ as

$$h_{t+1} = Ah_t + Bx_t, \qquad \hat{y}_t = Ch_t, \tag{14}$$

where A, B, C are completely unconstrained. In particular, no analogous spectral and contraction constraints to the ones for $\mathcal{K}_\phi$ are imposed on the SSM transition matrix A in (14). In the following Theorem we formally address core stability advantage of the Koopman parameterisation.

Theorem 1. *Let $\mathcal{K}_\phi(\rho_{\max})$ denote the class of Koopman operators realised by the ODO-based spectral–squashing parameterisation, such that $\|\mathcal{K}_\phi\|_2 < \rho_{\max} < 1$ for all $\mathcal{K}_\phi \in \mathcal{K}_\phi(\rho_{\max})$. Let $\mathcal{A}$ denote the class of unconstrained state–space transition matrices. Then: (i) For every $\mathcal{K}_\phi \in \mathcal{K}_\phi(\rho_{\max})$, the latent dynamics $z_{t+1} = \mathcal{K}_\phi z_t$ are uniformly exponentially stable at all training iterates. (ii) The class $\mathcal{A}$ contains matrices with arbitrarily large spectral radius; moreover, for any finite prediction horizon, there exist unstable $A \in \mathcal{A}$ that achieve the same finite-horizon training loss as a stable model.*

Proof. (i) Stability of the Koopman parameterisation

By Proposition 1 and Corollary 1, for all $\mathcal{K}_\phi \in \mathcal{K}_\phi(\rho_{\max})$ the latent dynamics satisfy (see (10)): $\|z_t\|_2 \leq \rho_{\max}^t \|z_0\|_2$, which proves uniform exponential stability.

(ii) Unstable but loss-equivalent SSM solutions The unconstrained class $\mathcal{A}$ contains matrices with arbitrarily large spectral radius: for any $\gamma > 0$, $A = \gamma I$ satisfies $\rho(A) = \gamma$. Now consider a stable SSM $h_{t+1} = Ah_t + Bx_t$, $\hat{y}_t = Ch_t$ that achieves a given finite-horizon loss over $t = 0, \ldots, H$. Define an augmented state $\tilde{h}_t = [h_t^\top, w_t^\top]^\top$ with dynamics

$$\tilde{h}_{t+1} = \begin{bmatrix} A & 0 \\ 0 & \gamma I \end{bmatrix} \tilde{h}_t + \begin{bmatrix} B \\ 0 \end{bmatrix} x_t, \qquad \hat{y}_t = \begin{bmatrix} C & 0 \end{bmatrix} \tilde{h}_t, \tag{15}$$

where $\gamma > 1$ and $\tilde{h}_0 = [h_0^\top, 0^\top]^\top$. Then $w_t \equiv 0$ for all $t \leq H$, so the augmented model produces identical outputs and achieves the same finite-horizon loss, while its transition matrix has spectral radius $\gamma > 1$. Choosing γ arbitrarily large completes the proof. □

3.1.6 Invertibility in finite dimensions and Banach space extension

Let $(\mathcal{X}, \|\cdot\|)$ be a Banach space and let $T : \mathcal{X} \to \mathcal{X}$ denote the (possibly nonlinear) time-1 evolution map of an underlying dynamical system. The Koopman operator acting on observables $J : \mathcal{X} \to \mathbb{R}$ is $(\mathcal{K}_\phi J)(x) = J\big(T(x)\big)$, which is linear even if T is nonlinear. If T is bijective, then $\mathcal{K}_\phi$ is invertible with $\mathcal{K}_\phi^{-1} J = J \circ T^{-1}$. Thus invertibility of the latent propagator corresponds to bijectivity of the underlying state update. In `DeepKoopFormer` we learn a *finite-dimensional approximation* of $\mathcal{K}_\phi$ using an ODO-structured linear operator $\mathcal{K}_\phi = U_\phi \operatorname{diag}(\Sigma_\phi) V_\phi^\top, \quad U_\phi, V_\phi \in O(d), \;\; \Sigma_i \in (0, \rho_{\max})$.

Because $0 < \Sigma_{i,\phi} < \rho_{\max} < 1$, the operator $\mathcal{K}_\phi$ is invertible and belongs to the general linear group $GL(d) := \{M \in \mathbb{R}^{d\times d} \mid \det(M) \neq 0\}$, that is, the group of all invertible $d \times d$ real matrices. Its inverse is given by $\mathcal{K}_\phi^{-1} = V_\phi \operatorname{diag}(\Sigma_\phi^{-1}) U_\phi^\top$, with operator norm bounds $\|K_\phi\|_2 \leq \rho_{\max}$ and $\|\mathcal{K}_\phi^{-1}\|_2 \leq 1/\sigma_{\min}(\mathcal{K}_\phi) < 1/\rho_{\max}$.

This construction yields a finite-dimensional surrogate of an invertible discrete-time Koopman operator, with $\mathcal{K}_\phi^n \to 0$ as $n \to \infty$ while backward iterates $\mathcal{K}_\phi^{-n}$ remain well-defined and bounded. In contrast, standard SSMs with unconstrained transition matrices A—common in classical and deep state-space architectures—provide no explicit spectral guarantees: A may be singular, ill-conditioned, or unstable, and even when invertible its spectrum is uncontrolled Gu et al. (2021a); Smith et al. (2022); Hamilton (2020). By comparison, the Koopman propagator in `Learnable-DeepKoopFormer` ensures stable, invertible latent dynamics with explicit spectral bounds and well-conditioned evolution.

4 Numerical Simulations

To evaluate `Learnable-DeepKoopFormer` in a controlled and scalable setting, we use a unified Koopformer benchmark on multivariate real-valued time series. From each dataset, sliding windows of length P are paired with forecast targets of length H, over a grid $P \in \{80, 100, 120, 140\}$ and $H \in \{4, 8, 12, 16\}$. Inputs are feature-wise normalised and split chronologically into 80%/20% train/test sets.

Model hyperparameters To ensure fair comparison, model capacity is fixed across experiments. Both `PatchTST` and `Informer` use $L = 3$ encoder layers with $d_{\text{model}} = 96$, $h = 4$ attention heads, and $d_{\text{ff}} = 96$. `Autoformer` uses the same depth and heads with $d_{\text{model}} = 96$ and $d_{\text{ff}} = 64$. Patch size (for `PatchTST`, `Informer`) and moving-average window (for `Autoformer`) are tied to the input length P, and token representations are pooled into a single latent vector. For constrained and learnable Koopman regimes, the latent dimension is $d_{\text{lat}} = d_{\text{model}}$ for `PatchTST` and `Informer`, and $d_{\text{lat}} = H$ for `Autoformer`. Stable and learnable variants enforce $\rho(\mathcal{K}_\phi) < \rho_{\max}$ with $\rho_{\max} = 0.99$, while the unconstrained case uses a free matrix in $\mathbb{R}^{d_{\text{lat}} \times d_{\text{lat}}}$. Within `Learnable-DeepKoopFormer`, `scalar-gated` and `per-mode gated` use global and per-mode (α, β) gates; `MLP-shaped` spectral mapping employs a one-hidden-layer MLP for spectral shaping; and `low-rank` Koopman uses rank $r = 16$, $\mathcal{K}_\phi = U_{r,\phi} \operatorname{diag}(\Sigma_{r,\phi}) V_{r,\phi}^\top$, with $U_r, V_r \in \mathbb{R}^{d_{\text{lat}} \times 16}$. Among baselines, LSTM uses $L = 2$ layers with hidden size 96; DLinear applies a shared linear map from the length-P history to the

H-step forecast; and the SSM employs a 96-dimensional latent state with linear readout to $\mathbb{R}^{H \cdot d}$. All models are trained with Adam (learning rate 3×10^{-4}) for 4,000 steps per (P, H) configuration; the Lyapunov weight is $\lambda_{\text{Lyap}} = 0.1$ for Koopman variants unless stated otherwise.

Spectral logging and diagnostics For all models with an explicit linear propagator, we log spectral quantities to characterise stability and latent energy dynamics. In Koopman-based architectures, latent evolution $z_{t+1} = \mathcal{K}_\phi z_t$ is governed by the spectrum of $\mathcal{K}_\phi$, which controls contraction or amplification over time. In constrained and learnable ODO variants, $\mathcal{K}_\phi = U_\phi \Sigma_\phi V_\phi^\top$, with orthonormal U_ϕ, V_ϕ and diagonal $\Sigma_\phi = \text{diag}(\sigma_{i,\phi})$ obtained via a squashing map enforcing $\sigma_{i,\phi} \in (0, \rho_{\max})$, $\rho_{\max} < 1$. We log the singular values $\{\sigma_i(\mathcal{K}_\phi)\}$ and in particular $\max_i \sigma_i(\mathcal{K}_\phi)$ as a proxy for the spectral radius, allowing us to visualise how different parametrisations (`scalar`, `per-mode`, `MLP`, `low-rank`) populate the disc of radius $\rho_{\max}$.

We compare three linear-transition regimes: (i) *constrained Koopman* (`constr`), enforcing $\max_i |\Sigma_{\phi,i}| < \rho_{\max} < 1$ for uniform contraction; (ii) *unconstrained Koopman* (`unconstr`), replacing the ODO structure with a free matrix $\mathcal{K}_\phi \in \mathbb{R}^{d \times d}$ and monitoring its eigenvalues $\{\lambda_i(\mathcal{K}_\phi)\}$ and spectral radius $\rho(\mathcal{K}_\phi) = \max_i |\lambda_i(\mathcal{K}_\phi)|$; and (iii) a linear state-space model (SSM), $h_{t+1} = Ah_t + Bx_t$, $\hat{y}_t = Ch_t$, where A is unconstrained. Since A is generally non-normal, we log its singular values—particularly $\|A\|_2$—as a stability and conditioning diagnostic. This logging procedure maps each trained model to a spectral data point, enabling distributional comparisons (e.g., violin plots) that reveal stability–accuracy trade-offs across Koopman parametrisations and Transformer backbones.

Benchmark Datasets To assess the forecasting performance of `Learnable-DeepKoopFormer`, we consider three heterogeneous and dynamically rich domains: *climate systems*, *financial markets*, and *electricity generation*. These settings exhibit nonstationarity, multiscale dependencies, and nonlinear interactions, challenging standard deep forecasting models. For climate analysis, we use CMIP6 atmospheric data[4] covering regional wind speed and surface pressure over Germany Makula & Zhou (2022); Hersbach et al. (2020); Forootani et al. (2025). As a nonlinear financial benchmark, we analyse a public multivariate cryptocurrency dataset[5] including volatility, price, and trading indicators. Finally, we model Spanish national electricity generation[6] across fossil, wind, solar, hydro, and biomass sources, where physical and weather-driven couplings shape temporal dynamics. These datasets span chaotic, stochastic, and structured regimes, providing a rigorous testbed for evaluating Koopman-enhanced latent representations in nonlinear time-series forecasting.

Hardware Configuration All experiments were executed on a high-performance computing (`HPC`) system with dual `AMD EPYC 9554` processors (128 cores per node, multithreaded) and 1.5 TB shared memory, suitable for large context windows and attention-based models. Training was accelerated using `NVIDIA L40S GPUs` supporting mixed-precision tensor operations.

Code availability statement The `Learnable-DeepKoopFormer` framework—including source code, datasets, and representative figures—is accessible at `Zenodo` [7]. The entire implementation is written in `Python` and builds upon standard scientific computing libraries, including `PyTorch`, `NumPy`, and `SciPy`.

4.1 Discussion on Wind Speed dataset

For the CMIP6 wind–speed task, Fig. 1 shows train/test MSE and MAE across all (P, H) configurations. Koopman–enhanced Transformers (`PatchTST`, `Autoformer`, `Informer`) exhibit tightly concentrated, low-centred violins, indicating consistently low errors and strong robustness across forecasting settings. In contrast, LSTM, DLinear, and SSM display higher-centred and broader distributions, reflecting greater sensitivity to patch length and horizon. Within the Koopman family, the `constr` variant yields the most compact and stable error profiles—especially for `PatchTST`—highlighting the stabilising effect of explicit spectral constraints. Learnable variants (`scalar-gated`, `per-mode gated`, `MLP-shaped` spectral mapping, `low-rank` Koopman) achieve comparable medians with slightly wider spreads, reflecting an expressiveness–robustness

[4] https://cds.climate.copernicus.eu/datasets/projections-cordex-domains-single-levels?tab=overview
[5] https://github.com/Chisomnwa/Cryptocurrency-Data-Analysis
[6] https://github.com/afshinfaramarzi/Energy-Demand-electricity-price-Forecasting/tree/main
[7] https://doi.org/10.5281/zenodo.17988424, https://doi.org/10.5281/zenodo.18115612

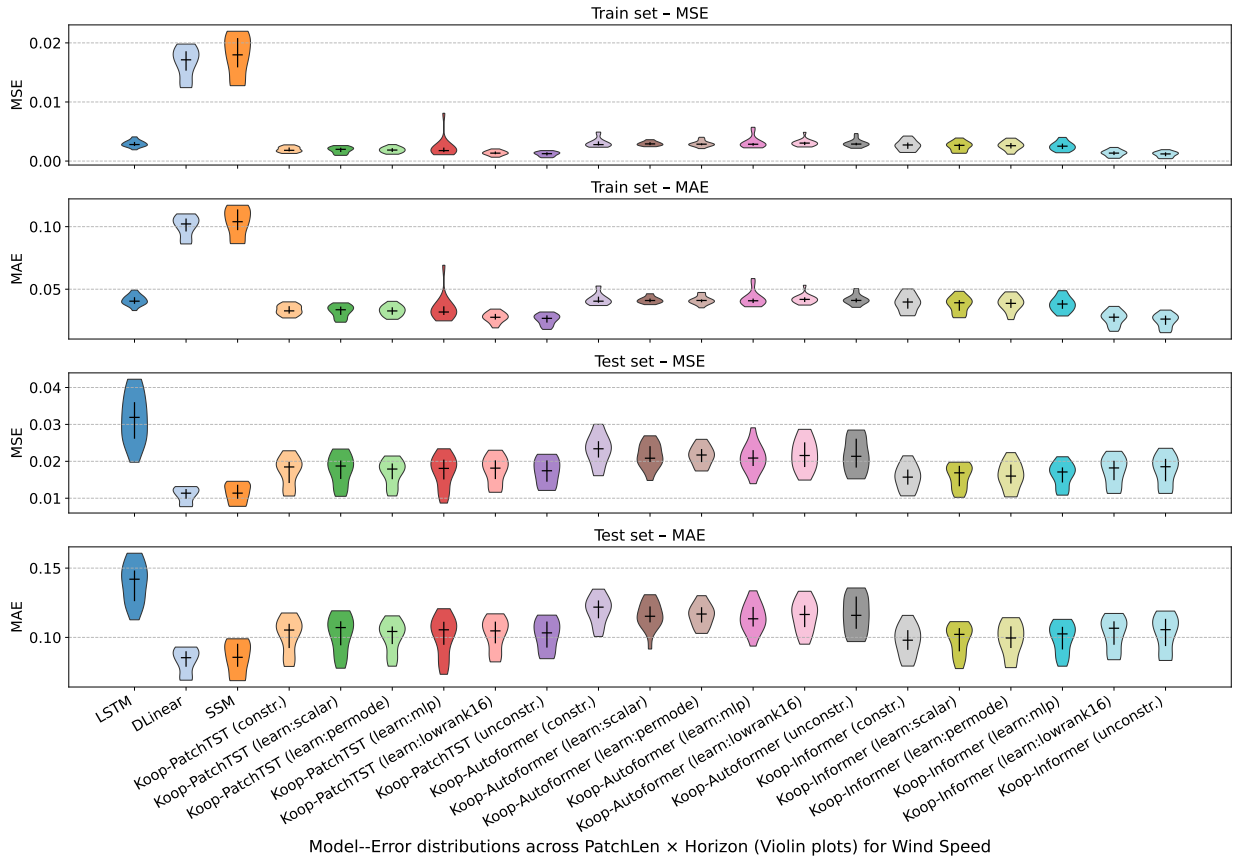

Figure 1: Violin plots of train/test MSE and MAE for all architectures on the wind speed forecasting task, aggregated over all patch lengths and horizons.

trade-off. The `unconstr` operator shows heavier tails, particularly under `PatchTST`, indicating reduced reliability despite competitive median accuracy. Overall, Koopman–Transformer models are both more accurate and more stable than recurrent and linear baselines.

Koopman spectra across backbones — Wind Speed For each configuration, we pool latent spectral magnitudes $\rho = |\lambda|$ across backbones and variants. Figure 2 reports these distributions alongside the SSM singular-value spectrum. Constrained Koopman models (`constr`) exhibit compact spectra concentrated in $\rho \approx 0.3$–0.8, with virtually no mass above 1, confirming enforcement of the spectral bound. Learnable variants (`scalar-gated`, `per-mode gated`, `MLP-shaped` spectral mapping, `low-rank` Koopman) occupy a similar range but with broader support, exploiting more of the admissible disc while remaining predominantly contractive.

In contrast, `unconstr` Koopman operators collapse toward small radii ($\rho \lesssim 0.2$), indicating strongly damped dynamics influenced by data fitting and Lyapunov regularisation. The SSM baseline concentrates at small magnitudes ($\rho \approx 0.05$–0.2) but exhibits a thin tail extending beyond the unit circle, reflecting occasional unstable or poorly conditioned transitions. Overall, spectral behaviour is primarily determined by the Koopman parametrisation rather than the backbone, with constrained and learnable variants achieving expressive yet stable latent dynamics.

4.2 Discussion on Pressure Surface dataset

For the CMIP6 pressure–surface task, Fig. 3 reports train/test MSE and MAE across all (P, H) configurations. Koopman–Transformer models (`PatchTST`, `Autoformer`, `Informer`) exhibit low-centred and compact

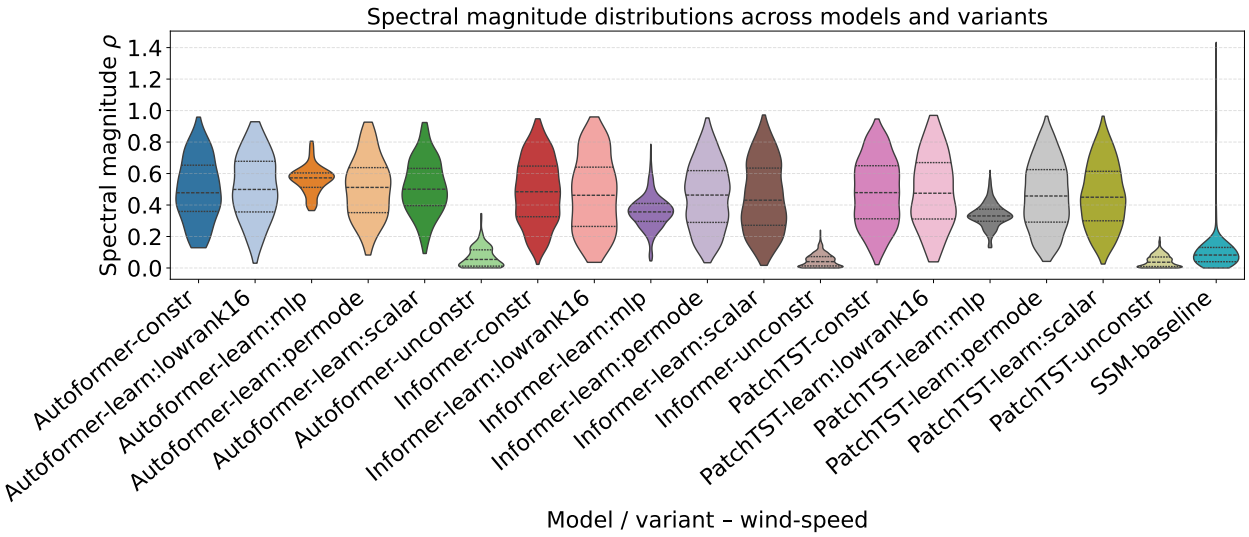

Figure 2: Spectral distributions of latent Koopman operators for CIMIP6 wind speed forecasting task.

violins, indicating stable performance across forecasting settings. In contrast, LSTM, DLinear, and SSM often show higher-centred and broader distributions, reflecting greater variability. Among Koopman variants, the constrained operator (`constr`) consistently yields the most compact and low-variance errors across backbones. Learnable families (`scalar-gated`, `per-mode gated`, `MLP-shaped` spectral mapping, `low-rank` Koopman) achieve comparable medians with slightly wider spreads, reflecting increased flexibility with limited loss of robustness. The unconstrained operator (`unconstr`) displays heavier tails—particularly for `PatchTST` and `Informer`—indicating reduced reliability without spectral regularisation. Although certain baseline configurations (notably DLinear and SSM) attain slightly lower *mean* test errors in some cases, these improvements are accompanied by substantially wider spreads, suggesting less consistent generalisation. In contrast, Koopman models maintain competitive means with markedly reduced variability, highlighting the stabilising effect of structured latent dynamics.

Koopman spectra across backbones — Pressure Surface For each configuration, we pool latent spectral magnitudes $\rho = |\lambda|$ by backbone and variant (Fig. 4). Constrained Koopman models (`constr`) exhibit compact spectra concentrated in $\rho \approx 0.3$–0.8, with no mass above 1, confirming effective enforcement of the spectral bound. Learnable variants (`scalar-gated`, `per-mode gated`, `MLP-shaped` spectral mapping, `low-rank` Koopman) occupy a similar range with broader support, remaining largely contractive while allowing greater flexibility. In contrast, `unconstr` variants collapse toward small radii ($\rho \lesssim 0.2$), indicating strongly damped dynamics. The SSM baseline also concentrates at small magnitudes but shows a thin tail extending occasionally beyond the unit circle, reflecting potential instability. Overall, spectral behaviour is primarily shaped by the Koopman parametrisation rather than the Transformer backbone, with constrained and learnable variants achieving stable latent dynamics for pressure–surface forecasting.

4.3 Discussion on Cryptocurrency dataset

For the cryptocurrency task, Fig. 5 shows train/test MSE and MAE across all (P, H) configurations. Koopman–Transformer models (`PatchTST`, `Autoformer`, `Informer`) exhibit compact, low-centred violins, indicating stable performance across hyperparameter choices. In contrast, LSTM, DLinear, and SSM display broader spreads and higher variability, reflecting greater sensitivity to architecture and window settings. Within the Koopman family, the constrained operator (`constr`) achieves the most concentrated error distributions on both train and test sets, underscoring the benefit of spectral stabilisation in volatile crypto dynamics. Learnable variants (`scalar-gated`, `per-mode gated`, `MLP-shaped` spectral mapping, `low-rank` Koopman) attain competitive medians with slightly wider spreads, reflecting a flexibility–stability trade-off.

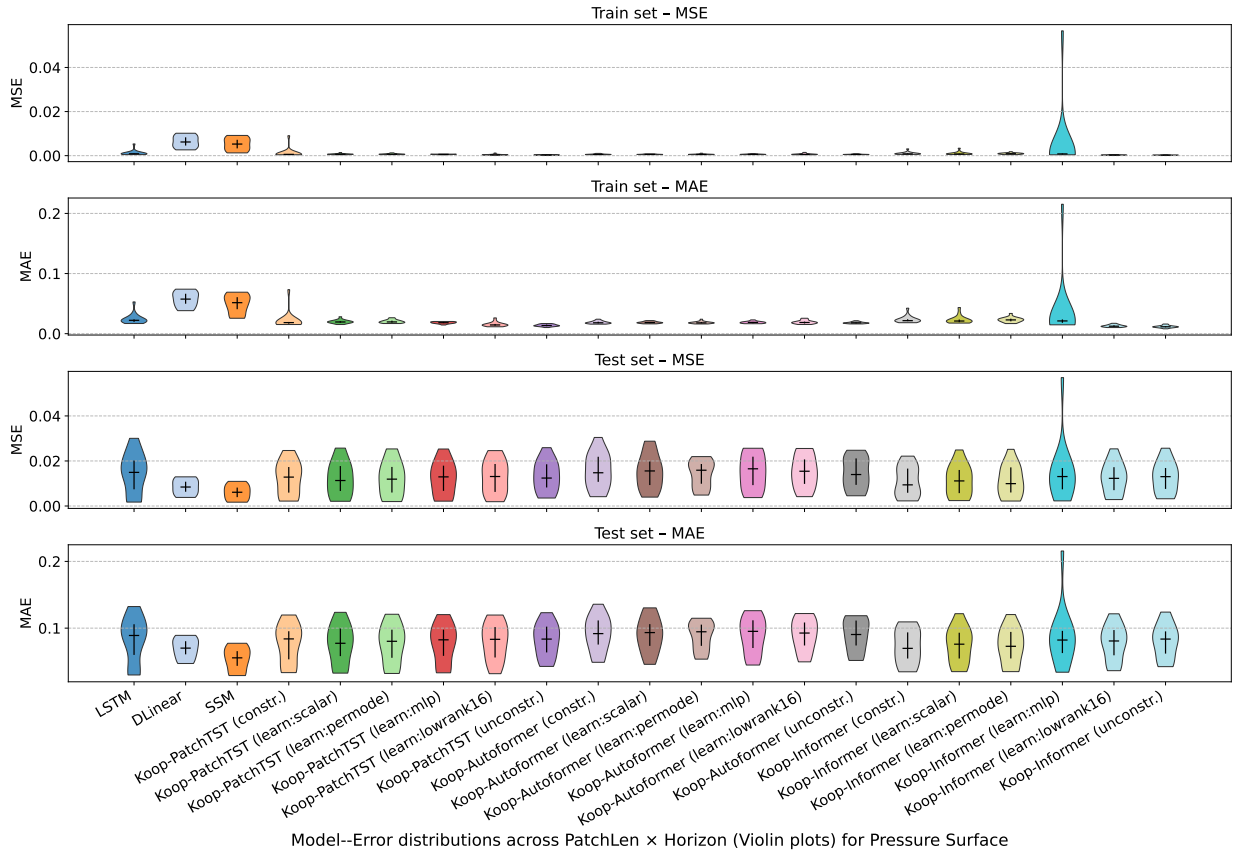

Figure 3: Violin plots of train/test MSE and MAE for all architectures on the pressure surface forecasting task, aggregated over all patch lengths and horizons.

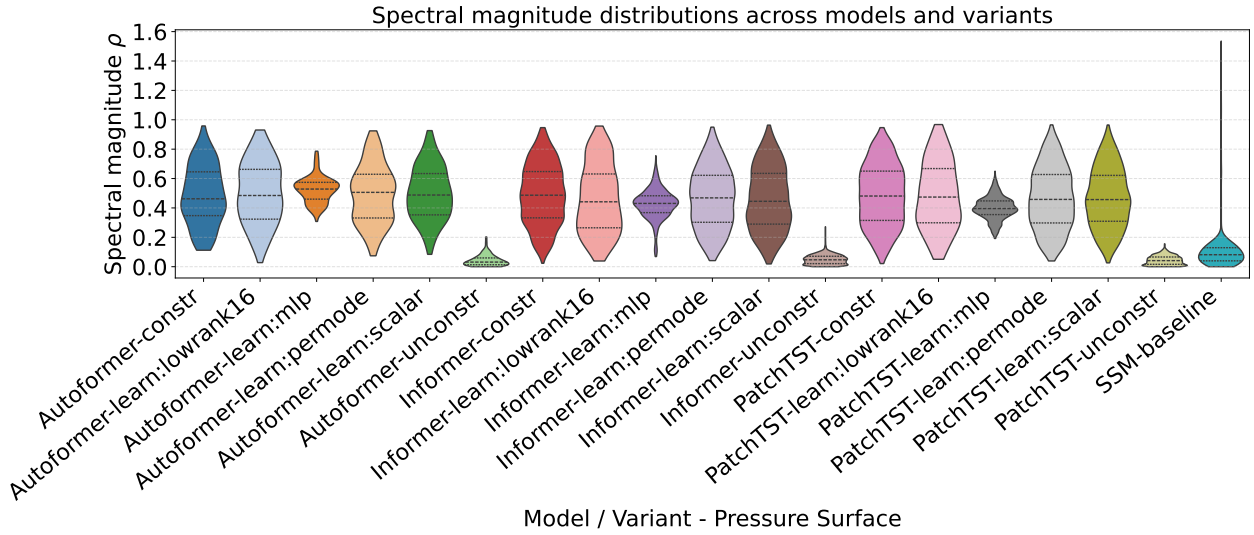

Figure 4: Spectral distributions of latent Koopman operators for CIMIP6 pressure surface forecasting task.

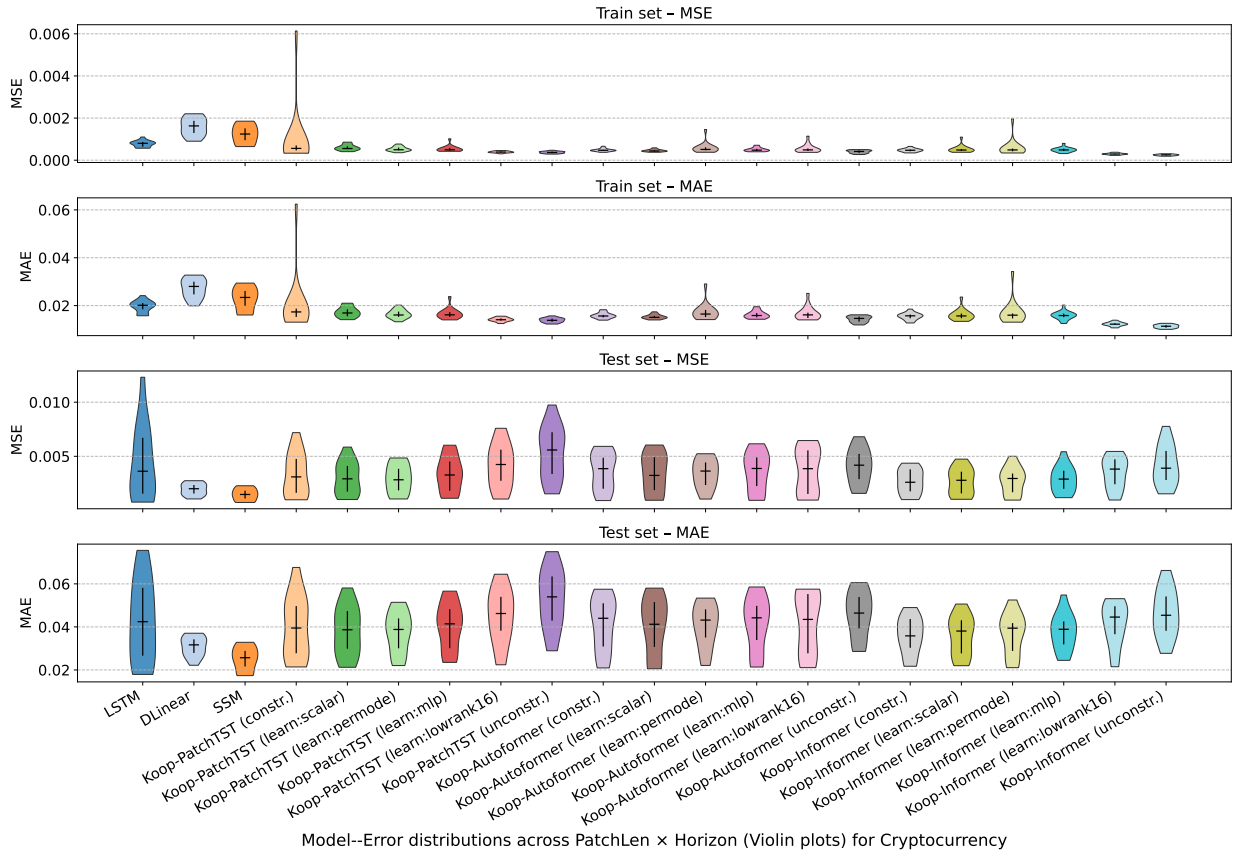

Figure 5: Violin plots of train/test MSE and MAE for all architectures on the Cryptocurrency forecasting task, aggregated over all patch lengths and horizons.

The `unconstr` operator shows heavier tails—particularly for `PatchTST` and `Informer`—indicating reduced robustness without spectral control.

Koopman spectra across backbones — Cryptocurrency Pooling latent spectral magnitudes $\rho = |\lambda|$ across configurations (Fig. 6), constrained Koopman models (`constr`) exhibit compact spectra centred around $\rho \approx 0.3$–0.6, with no values near the unit circle, confirming effective spectral enforcement. Learnable variants (`scalar-gated`, `per-mode gated`, `MLP-shaped` spectral mapping, `low-rank` Koopman) span a slightly broader contractive range ($\rho \approx 0.3$–0.8), reflecting increased expressive capacity while remaining largely stable. In contrast, `unconstr` variants collapse toward small radii ($\rho < 0.2$), indicating overly damped latent dynamics. The SSM baseline likewise concentrates at small magnitudes but exhibits a thin tail extending toward and beyond $\rho \approx 1$, revealing occasional unstable modes. Overall, spectral behaviour is primarily governed by the Koopman parametrisation rather than the Transformer backbone, with constrained and learnable variants achieving expressive yet stable latent evolution in highly volatile settings.

4.4 Discussion on Energy Systems Dataset

Figure 7 reports train/test MSE and MAE across all (P, H) settings. Unlike previous domains, `Autoformer` performs noticeably worse: both baseline and Koopman variants show high-centred violins with broad tails, indicating sensitivity to hyperparameters. LSTM achieves very low training errors but exhibits widely spread test errors, suggesting overfitting. DLinear and SSM occasionally reach competitive medians yet display inconsistent spreads across configurations, reflecting limited robustness. Across backbones, Koopman-enhanced models provide more reliable behaviour. The constrained operator (`constr`) yields consistently tight, mid-centred violins with low variability. While not always achieving the lowest median error, it is the most stable

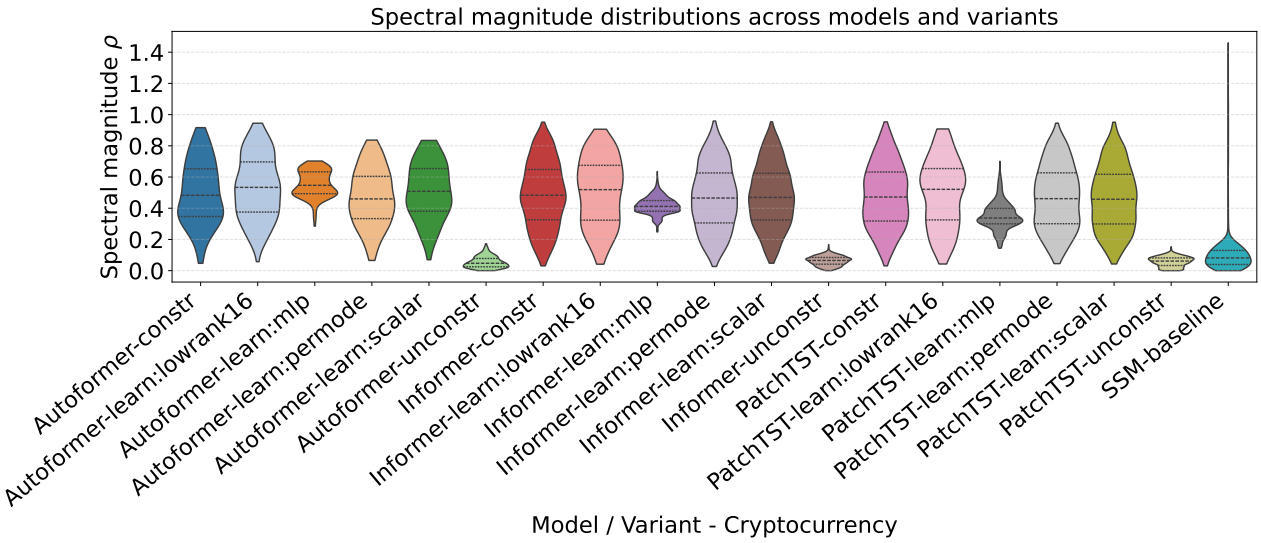

Figure 6: Spectral distributions of latent Koopman operators for the Cryptocurrency forecasting task.

across hyperparameters, highlighting the benefit of spectral bounds for slow-varying energy dynamics. Learnable variants (`scalar-gated`, `per-mode gated`, `MLP-shaped` spectral mapping, `low-rank` Koopman) attain competitive medians with slightly broader spreads, reflecting an expressiveness–robustness trade-off. Overall, results show that minimising training error alone (e.g., LSTM or `Autoformer`) does not ensure reliable forecasting, whereas spectrally controlled Koopman models favour stable generalisation under long-range temporal structure.

Koopman spectra across backbones — Energy Systems Figure 8 displays pooled latent spectral magnitudes $\rho = |\lambda|$. The constrained Koopman operator (`constr`) forms compact spectra around $\rho \approx 0.3$–0.6, well below $\rho = 1$, confirming stable latent evolution. Learnable families (`scalar-gated`, `per-mode gated`, `MLP-shaped` spectral mapping, `low-rank` Koopman) span a broader yet contractive range ($\rho \approx 0.3$–0.8), enabling greater flexibility. In contrast, `unconstr` variants collapse toward small radii ($\rho < 0.2$), indicating overly damped dynamics, while the SSM baseline shows mostly small values with a thin tail beyond $\rho \geq 1$, revealing occasional instability. Overall, explicit spectral control is key to stable temporal modelling in energy systems data.

4.5 Advantages of `Learnable-DeepKoopFormer` Modules.

The spectral analysis highlights a key benefit of the proposed learnable Koopman modules: they adapt the latent time scales without sacrificing stable and reversible dynamics. Whereas purely constrained variants tightly restrict the spectrum, and unconstrained models tend to collapse into strongly damped states ($|\lambda| \to 0$), the learnable families populate a non-collapsed but contractive range ($0.3 \lesssim |\lambda| \lesssim 0.8$). This region provides enough variability to capture both fast and slow temporal modes, while keeping the spectral radius below unity so that the latent dynamics remain stable. Crucially, because eigenvalues do not decay toward zero, the resulting operators are *invertible*, enabling well-posed backward propagation and preventing information loss over long-range iterative forecasts. Thus, learnability does not merely increase flexibility; it yields *invertible and non-degenerate latent dynamics*, offering richer temporal structure than constrained models and more reliable long-horizon behaviour than unstable or over-damped baselines.

5 Conclusion

This paper introduced the Learnable-DeepKoopFormer framework, which augments Transformer-based forecasters with spectrally controlled, learnable Koopman operators. Using an orthogonal–diagonal–

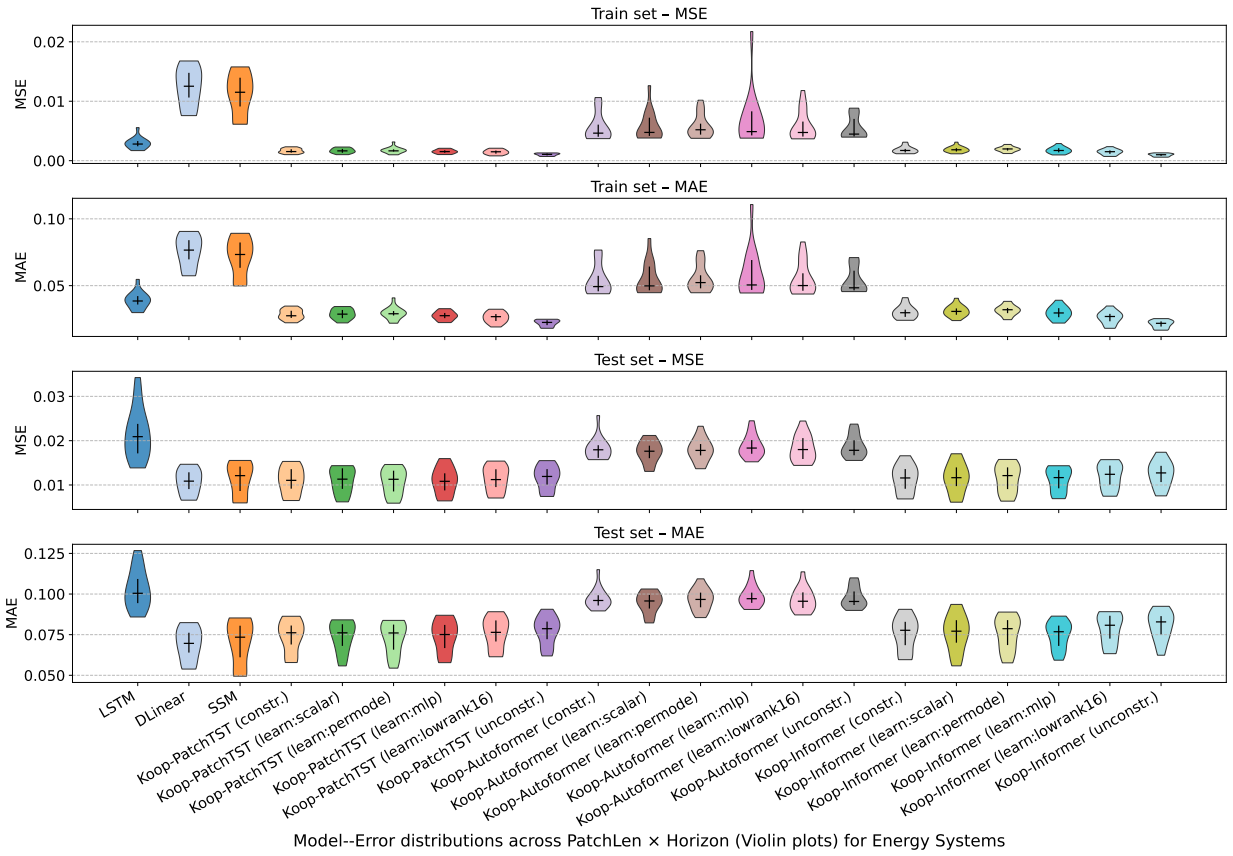

Figure 7: Violin plots of train/test MSE and MAE for all architectures on the energy systems forecasting task, aggregated over all patch lengths and horizons.

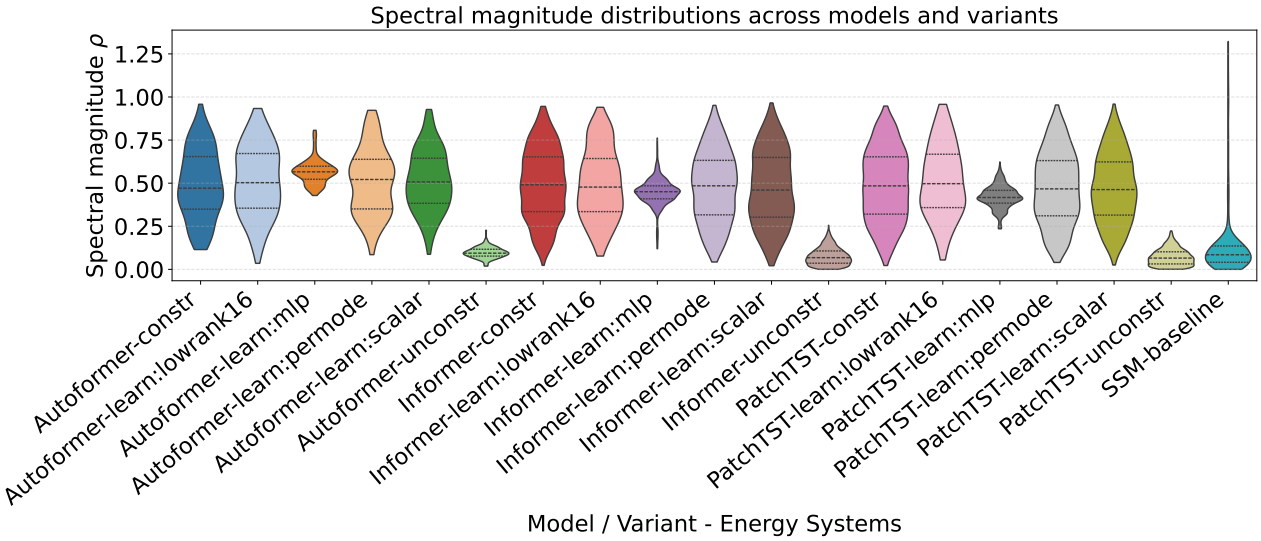

Figure 8: Spectral distributions of latent Koopman operators for the energy systems forecasting task.

orthogonal (ODO) parameterisation, we proposed scalar-gated, per-mode gated, MLP-shaped, and low-rank variants that interpolate between strictly constrained and fully unconstrained linear latent dynamics while preserving explicit control over spectral radius, stability, and rank. A unified training objective combining forecasting loss with Lyapunov regularisation provides additional stability bias. We established theoretical guarantees on spectral stability, exponential contraction, low-rank structure, invertibility, and Lyapunov consistency, showing that the proposed operators retain the full expressiveness of spectrally bounded linear dynamics while ensuring well-conditioned propagation.

Extensive experiments on heterogeneous real-world datasets—CMIP6 wind speed and surface pressure, cryptocurrency markets, and national-scale electricity generation—demonstrate that Koopman-enhanced Transformers achieve robust and accurate forecasting across diverse input–output configurations. Compared with LSTM, DLinear, and SSM baselines, Koopman variants embedded in PatchTST, Autoformer, and Informer consistently produce lower and more concentrated error distributions, indicating improved stability with respect to patch length and prediction horizon. Constrained operators yield the most compact and reliable behaviour, while learnable variants attain competitive median accuracy with a favorable expressiveness–robustness trade-off. Unconstrained linear propagators, although sometimes matching median errors, exhibit heavier tails and reduced reliability.

Spectral diagnostics confirm that the Koopman parameterisation primarily shapes latent dynamics. Constrained and learnable variants populate a non-collapsed contractive band, supporting stable yet expressive temporal evolution, whereas unconstrained models tend toward over-damped or occasionally unstable regimes. Overall, Learnable-DeepKoopFormer provides a principled synthesis of operator-theoretic structure and modern attention architectures, yielding stable, interpretable, and invertible latent dynamics suitable for large-scale time-series forecasting.

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

Omri Azencot, N Benjamin Erichson, Vanessa Lin, and Michael Mahoney. Forecasting sequential data using consistent koopman autoencoders. In *International Conference on Machine Learning*, pp. 475–485. PMLR, 2020.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

S. L. Brunton, B. W. Brunton, J. L. Proctor, and J.N. Kutz. Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control. *PLoS ONE*, 11(2):e0150171, 2016.

Steven L Brunton, Marko Budišić, Eurika Kaiser, and J Nathan Kutz. Modern koopman theory for dynamical systems. *arXiv preprint arXiv:2102.12086*, 2021.

Laurent AF Callot, Anders B Kock, and Marcelo C Medeiros. Modeling and forecasting large realized covariance matrices and portfolio choice. *Journal of Applied Econometrics*, 32(1):140–158, 2017.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 2018.

I. Dimoulkas, P. Mazidi, and L. Herre. Neural networks for GEFCom2017 probabilistic load forecasting. *International Journal of Forecasting*, 35(4):1409 – 1423, 2019.

Jan Drgona, Sayak Mukherjee, Jiaxin Zhang, Frank Liu, and Mahantesh Halappanavar. On the stochastic stability of deep markov models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24033–24047. Curran Associates, Inc., 2021.

Jan Drgona, Aaron Tuor, Soumya Vasisht, and Draguna Vrabie. Dissipative deep neural dynamical systems. *IEEE Open Journal of Control Systems*, 1:100–112, 2022. doi: 10.1109/OJCSYS.2022.3186838.

Ali Forootani and Mohammad Khosravi. Compact transformer variants for synthetic time series forecasting. *Neurocomputing*, 677:133140, 2026. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2026.133140. URL https://www.sciencedirect.com/science/article/pii/S0925231226005370.

Ali Forootani, Danial Esmaeili Aliabadi, and Daniela Thrä n. Climate aware deep neural networks (cadnn) for wind power simulation. *Array*, 28:100534, 2025. ISSN 2590-0056. doi: https://doi.org/10.1016/j.array.2025.100534. URL https://www.sciencedirect.com/science/article/pii/S2590005625001614.

Ali Forootani, Mohammad Khosravi, and Masoud Barati. Deepkoopformer: a koopman enhanced transformer based architecture for time series forecasting. *npj Artificial Intelligence*, 2:35, 2026. doi: 10.1038/s44387-026-00085-3.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021a.

Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state-space layers. *NeurIPS*, 2021b.

James D Hamilton. *Time series analysis*. Princeton university press, 2020.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, June 2021. ISSN 2522-5820. doi: 10.1038/s42254-021-00314-5. Publisher Copyright: © 2021, Springer Nature Limited.

Mohammad Khosravi. Representer theorem for learning koopman operators. *IEEE Transactions on Automatic Control*, 68(5):2995–3010, 2023.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *ICLR*, 2022.

Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2015.

M. Korda and I. Mezić. Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control. *arXiv:1611.03537*, 2016.

J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor. *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*. SIAM, 2016.

Vitaly Kuznetsov and Mehryar Mohri. Discrepancy-based theory and algorithms for forecasting non-stationary time series. *Annals of Mathematics and Artificial Intelligence*, 88(4):367–399, 2020.

Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10), 1995.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

Matthew M Lin. Discrete eckart–young theorem for integer matrices. *SIAM journal on matrix analysis and applications*, 32(4):1367–1382, 2011.

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *ICLR*, 2021.

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International Conference on Learning Representations*, 2022a.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Rethinking the stationarity in time series forecasting. *arXiv preprint arXiv:2205.14415*, 2022b.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1):4950, November 2018a. ISSN 2041-1723.

Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature communications*, 9(1):4950, 2018b.

Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang. Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873, 2014.

Spyros Makridakis, Evangelos Spiliotis, Vassilios Assimakopoulos, Zhi Chen, Anil Gaba, Ilia Tsetlin, and Robert L Winkler. The m5 uncertainty competition: Results, findings and conclusions. *International Journal of Forecasting*, 2021.

Exavery Kisesa Makula and Botao Zhou. Coupled model intercomparison project phase 6 evaluation and projection of east african precipitation. *International Journal of Climatology*, 42(4):2398–2412, 2022.

I. Mezić. *Spectral operator methods in dynamical systems: Theory and applications*. Springer, 2017.

Jeremy Morton, Antony Jameson, Mykel J Kochenderfer, and Freddie Witherden. Deep dynamical modeling and control of unsteady fluid flows. *Advances in Neural Information Processing Systems*, 31, 2018.

Truong X. Nghiem, Ján Drgoňa, Colin Jones, Zoltan Nagy, Roland Schwan, Biswadip Dey, Ankush Chakrabarty, Stefano Di Cairano, Joel A. Paulson, Andrea Carron, Melanie N. Zeilinger, Wenceslao Shaw Cortez, and Draguna L. Vrabie. Physics-informed machine learning for modeling and control of dynamical systems. In *2023 American Control Conference (ACC)*, pp. 3735–3750, 2023. doi: 10.23919/ACC55779.2023.10155901.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

Harshit Saxena, Omar Aponte, and Katie T. McConky. A hybrid machine learning model for forecasting a billing period's peak electric load days. *International Journal of Forecasting*, 35(4):1288 – 1303, 2019.

Rajat Sen, Hsiang-Fu Yu, and Inderjit S Dhillon. Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. *Advances in Neural Information Processing Systems*, 32, 2019.

Elliott Skomski, Soumya Vasisht, Colby Wight, Aaron Tuor, Ján Drgoňa, and Draguna Vrabie. Constrained block nonlinear neural dynamical models. In *2021 American Control Conference (ACC)*, pp. 3993–4000, 2021. doi: 10.23919/ACC50511.2021.9482930.

Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022.

Slawek Smyl. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, 36(1):75–85, 2020.

Naoya Takeishi, Yoshinobu Kawahara, and Takehisa Yairi. Learning Koopman invariant subspaces for dynamic mode decomposition. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017a.

Naoya Takeishi, Yoshinobu Kawahara, and Takehisa Yairi. Learning koopman invariant subspaces for dynamic mode decomposition. *Advances in neural information processing systems*, 30, 2017b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

François-Xavier Vialard, Roland Kwitt, Suan Wei, and Marc Niethammer. A shooting formulation of deep learning. In *Advances in Neural Information Processing Systems*, 2020.

Yuyang Wang, Alex Smola, Danielle Maddix, Jan Gasthaus, Dean Foster, and Tim Januschowski. Deep factors for forecasting. In *International Conference on Machine Learning*, pp. 6607–6617, 2019.

Ruofeng Wen, Kari Torkkola, Balakrishnan Narayanaswamy, and Dhruv Madeka. A multi-horizon quantile recurrent forecaster. *arXiv preprint arXiv:1711.11053*, 2017.

Matthew O Williams, Ioannis G Kevrekidis, and Clarence W Rowley. A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25:1307–1346, 2015.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with Auto-Correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, 2021a.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with Auto-Correlation for long-term series forecasting. In *NeurIPS*, 2021b.

Xing Yan, Weizhong Zhang, Lin Ma, Wei Liu, and Qi Wu. Parsimonious quantile regression of financial asset tail dynamics via sequential learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Enoch Yeung, Soumya Kundu, and Nathan Hodas. Learning deep neural network representations for Koopman operators of nonlinear dynamical systems. In *2019 American Control Conference (ACC)*, pp. 4832–4839, 2019. doi: 10.23919/ACC.2019.8815339.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *arXiv preprint arXiv:2205.13504*, 2022.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Yi Zheng, Qi Liu, Enhong Chen, Yong Ge, and J Leon Zhao. Time series classification using multi-channels deep convolutional neural networks. In *International conference on web-age information management*, pp. 298–310. Springer, 2014.

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*, 2021a.

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 *The Thirty-Fifth AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021b.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proc. 39th International Conference on Machine Learning*, 2022a.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *ICML*, 2022b.