DISSECTING CHRONOS: SPARSE AUTOENCODERS REVEAL CAUSAL FEATURE HIERARCHIES IN TIME SERIES FOUNDATION MODELS

Anurag Mishra
Rochester Institute of Technology
am2552@rit.edu

ABSTRACT

Time series foundation models (TSFMs) are increasingly deployed in high-stakes domains, yet their internal representations remain opaque. We present the first application of sparse autoencoders (SAEs) to a TSFM, training TopK SAEs on activations of Chronos-T5-Large (710M parameters) across six layers. Through 392 single-feature ablation experiments, we establish that every ablated feature produces a positive CRPS degradation, confirming causal relevance. Our analysis reveals a depth-dependent hierarchy: early encoder layers encode low-level frequency features, the mid-encoder concentrates causally critical change-detection features, and the final encoder compresses a rich but less causally important taxonomy of temporal concepts. The most critical features reside in the mid-encoder (max single-feature ΔCRPS $= 38.61$), not in the semantically richest final encoder layer, where progressive ablation paradoxically improves forecast quality. These findings demonstrate that mechanistic interpretability transfers effectively to TSFMs and that Chronos-T5 relies on abrupt-dynamics detection rather than periodic pattern recognition.

Track: Research

1 INTRODUCTION

Time series foundation models (TSFMs) such as Chronos-T5 (Ansari et al., 2024), TimesFM (Das et al., 2024), MOMENT (Goswami et al., 2024), and Moirai (Woo et al., 2024) achieve competitive or state-of-the-art performance across diverse forecasting benchmarks, often in zero-shot settings. Yet despite rapid adoption in production systems, the internal representations of these models remain entirely unexamined from a mechanistic perspective.

In natural language processing, mechanistic interpretability (MI) has become a productive research program. Sparse autoencoders (SAEs) decompose the dense, superposed activations of language models into interpretable features (Bricken et al., 2023; Cunningham et al., 2024; Templeton et al., 2024), and circuit-level analyses have identified interpretable computational subgraphs (Wang et al., 2022; Olsson et al., 2022; Conmy et al., 2023). For time series, interpretability has instead relied on post-hoc methods: saliency maps (Zhao et al., 2023; Kechris et al., 2025), perturbation-based explanations (Enguehard, 2023; Liu et al., 2024; Queen et al., 2023), counterfactual approaches (Yan & Wang, 2023), and concept-based frameworks (van Sprang et al., 2024; Boileau et al., 2025; Santamaria-Valenzuela et al., 2025). Only Kalnāre et al. (2025) have applied any form of mechanistic analysis to a time series transformer, targeting a small custom classifier rather than a foundation model.

To our knowledge, **no prior work has applied sparse autoencoders to a time series foundation model**. The T5 architecture underlying Chronos is well-understood (Raffel et al., 2020), SAE training protocols are mature, and Chronos's discrete tokenization (4,096 bins) provides a natural unit of analysis. We address this gap with three contributions:

1. We train TopK SAEs at six extraction points in Chronos-T5-Large and demonstrate via systematic ablation that learned features are causally relevant (392 ablations, 100% positive CRPdeltaS).

2. We show that temporal concepts follow a depth-dependent hierarchy: low-level frequency features in early layers, causally critical change-detection features in the mid-encoder, and rich semantic compression in the final encoder.
3. We find that causal importance is inversely related to semantic richness: the mid-encoder is most critical (max ΔCRPS $= 38.61$) while the final encoder paradoxically improves under progressive ablation.

2 METHOD

Chronos-T5. Chronos-T5 (Ansari et al., 2024) adapts the T5 encoder-decoder architecture (Raffel et al., 2020) for probabilistic time series forecasting via a quantization-based tokenization scheme: each value in a univariate time series is normalized and mapped to one of $V = 4{,}096$ discrete bins. Chronos-T5-Large has 710M parameters distributed across 24 encoder and 24 decoder layers with hidden dimension $d_{\text{model}} = 1{,}024$. We target six extraction points spanning the full processing pipeline: encoder blocks 5 (early), 11 (mid), and 23 (final), decoder blocks 11 and 23 (residual stream), and the cross-attention output at decoder block 11.

Sparse Autoencoders. We train TopK sparse autoencoders (Cunningham et al., 2024) on the residual stream activations at each extraction point. Given an activation vector $\mathbf{x} \in \mathbb{R}^{d_{\text{model}}}$, the SAE computes:

$$\mathbf{z} = \text{TopK}\big(\mathbf{W}_{\text{enc}}(\mathbf{x} - \mathbf{b}_{\text{dec}}) + \mathbf{b}_{\text{enc}},\ k\big), \qquad \hat{\mathbf{x}} = \mathbf{W}_{\text{dec}}\,\mathbf{z} + \mathbf{b}_{\text{dec}} \tag{1}$$

where TopK retains only the k largest activations and zeros the rest. We set $d_{\text{sae}} = 8 \times d_{\text{model}} = 8{,}192$ features per layer, $k = 64$, and train with MSE reconstruction loss for 50,000 steps using Adam (learning rate 3×10^{-4}, cosine decay). Dead features are periodically resampled following Templeton et al. (2024).

Activation Extraction. Activations are collected by registering forward hooks at each target layer during Chronos-T5 inference. We use two data sources: (i) a synthetic diagnostic suite providing ground-truth temporal properties (trends, seasonality, level shifts, frequency sweeps, heteroscedastic noise) for taxonomy validation, and (ii) the ETT benchmark (Zhou et al., 2021) for causal experiments.

Feature Taxonomy. Each SAE feature is classified into one of eleven temporal concept categories: `trend_up/down`, `seasonality`, `level_shift_up/down`, `frequency_high/low`, `high/low_volatility`, `noise`, and `unknown`. Classification uses Pearson correlations between each feature's activation pattern on synthetic data and the ground-truth properties of each diagnostic category. Features with maximum correlation below a threshold are assigned `unknown`.

Causal Validation. We validate feature importance through two ablation protocols. In *single-feature ablation*, we zero each feature's sparse code ($z_j \leftarrow 0$), decode to produce modified activations, patch them back into the forward pass, and measure the CRPS change (Gneiting & Raftery, 2007): $\Delta\text{CRPS}_j = \text{CRPS}_{\text{ablated}} - \text{CRPS}_{\text{original}}$. We define the *causal importance* of feature j as ΔCRPS_j, i.e., the degradation in forecast quality when feature j is zeroed. In *progressive ablation*, we cumulatively ablate features sorted by decoder-norm contribution ($1, 2, 4, \ldots, 64$ features) and measure CRPS at each checkpoint. Experiments use 256 context windows from ETT with prediction length 64 and 4 forecast samples. An extended configuration (1,024 windows, 8 samples, 200 features) was run at the final encoder for higher statistical power.

3 RESULTS

3.1 SAE FEATURES ARE UNIVERSALLY CAUSALLY RELEVANT

Across 392 single-feature ablation experiments spanning three encoder layers, **every ablation produced a strictly positive ΔCRPS** (Table 1). This establishes that each tested feature encodes information the model actively uses for forecasting, and that this information cannot be recovered from the remaining features.

The distribution of causal impact is markedly layer-dependent. At encoder block 11, the top feature (`4616`) produces a ΔCRPS of 38.61 with a max-to-median ratio of 30.5$\times$, revealing an extremely heavy-tailed importance distribution where a small number of features carry disproportionate causal

Table 1: Single-feature ablation summary across encoder layers. All features are causally relevant (ΔCRPS > 0). The mid-encoder (block 11) shows the highest mean and maximum impact with a heavily right-skewed distribution.

Layer	n	**Mean**	**Med.**	**Max**	**Std**	**+Frac**	**Max/Med**
Enc. block 5	64	3.05	0.95	26.32	5.12	1.00	27.7×
Enc. block 11	64	**5.15**	1.26	**38.61**	6.95	1.00	**30.5**×
Enc. block 23	64	3.73	2.98	11.65	2.07	1.00	3.9×
Enc. block 23†	200	2.37	2.37	2.44	0.03	1.00	1.03×

†Extended sweep (1,024 windows, 8 samples). All ΔCRPS values report $\text{CRPS}_{\text{ablated}} - \text{CRPS}_{\text{original}}$.

Table 2: Feature taxonomy across layers. The final encoder (block 23) has the richest semantic coverage; the mid-encoder (block 11) concentrates change-detection features. Counts below 10 omitted.

Concept	**Enc 5**	**Enc 11**	**Enc 23**	**Dec 11**	**Dec 11$_{ca}$**	**Dec 23**
Seasonality	12	45	**1,439**	33	91	64
Level shift ↑	66	**1,024**	1,097	28	37	28
Level shift ↓	46	–	210	14	26	60
Trend ↑	48	220	228	13	12	37
Trend ↓	11	16	89	–	13	32
Freq. high	97	91	668	16	–	24
Freq. low	10	–	456	83	82	93
High volatility	68	268	260	–	13	14
Low volatility	14	28	138	33	43	94
Noise	32	413	315	24	17	–
Labeled (%)	**4.9**	**25.8**	**59.8**	**3.1**	**4.2**	**5.5**

Dec 11$_{ca}$ = cross-attention output at decoder block 11. Labels are heuristic correlation-based assignments on synthetic diagnostics.

weight. A similarly skewed pattern appears at encoder block 5 (max/median $= 27.7\times$), while encoder block 23 shows a much more uniform distribution (max/median $= 3.9\times$ in the 64-feature sweep and $1.03\times$ in the 200-feature sweep). This power-law structure has practical implications: uniform pruning strategies would disproportionately affect the critical few features in middle layers.

3.2 A Depth-Dependent Hierarchy of Temporal Concepts

Table 2 presents the concept distribution across all six extraction points. Of 49,152 total SAE features (8,192 per layer $\times$ 6 layers), 8,462 (17.2%) receive non-unknown labels, with coverage ranging from 3.1% at decoder block 11 to 59.8% at encoder block 23.

Early encoder (block 5, 4.9% labeled). Only 404 features receive labels, dominated by `frequency_high` (97) and `high_volatility` (68), suggesting preliminary local feature extraction.

Mid-encoder (block 11, 25.8% labeled). This layer shows a distinctive profile dominated by `level_shift_up` (1,024 features, 12.5%), `noise` (413, 5.0%), and `high_volatility` (268, 3.3%), while seasonality is nearly absent (45, 0.5%). The mid-encoder functions as a change-detection hub.

Final encoder (block 23, 59.8% labeled). The richest layer: `seasonality` dominates (1,439, 17.6%), followed by `level_shift_up` (1,097, 13.4%), `frequency_high` (668, 8.2%), and `frequency_low` (456, 5.6%). All concept categories are represented, compressing a full temporal characterization for cross-attention.

Decoder layers (3.1–5.5% labeled). All decoder points show low labeling, with `low_volatility` and `frequency_low` relatively prominent, possibly reflecting the decoder's focus on smooth forecast generation.

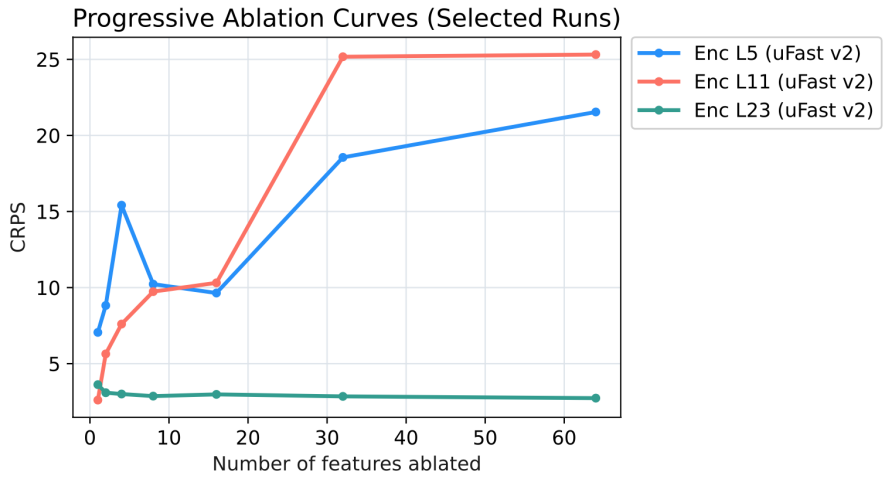

Figure 1: Progressive ablation curves across three encoder layers. CRPS is measured as features are cumulatively removed in order of decreasing decoder-norm contribution. Early and mid-encoder layers degrade sharply, while the final encoder remains flat or improves.

3.3 Causal Importance Is Inversely Related to Semantic Richness

Combining ablation (Table 1) with taxonomy (Table 2) reveals a counterintuitive pattern. The mid-encoder, with only 25.8% labeled features, is the most causally critical layer (mean ΔCRPS $= 5.15$, max $= 38.61$). The final encoder, with 59.8% labeled, has lower per-feature impact and a far more uniform distribution.

Progressive ablation (Figure 1) makes this divergence dramatic. At encoder block 11, CRPS rises from 2.61 to 25.32 (64 features ablated), indicating catastrophic dependence. At block 5, CRPS rises from 7.05 to 21.54. At encoder block 23, however, CRPS *decreases* from 3.62 to 2.73, a net improvement of 0.89. The extended 200-feature configuration confirms this: CRPS stays flat from 3.59 to 3.58 ($\Delta = -0.01$).

This paradoxical improvement may reflect the final encoder containing features that serve generalization across Chronos-T5's diverse pretraining domains but are suboptimal for ETT specifically; ablating them functions as implicit domain adaptation. The high feature utilization at block 23 (54.0% active vs. 23.3% at block 11 and 4.8% at block 5) is consistent with representational redundancy that mildly degrades the cross-attention signal.

4 Discussion and Conclusion

MI transfers to time series. Our central finding is that SAEs produce causally meaningful features when applied to a TSFM. The 100% positive ablation rate across 392 features is consistent with the utility of SAE features as causal handles.

The mid-encoder as computational bottleneck. The concentration of causal importance at encoder block 11, dominated by level shifts (1,024 features) and noise (413), suggests that detecting abrupt distributional changes rather than periodic patterns is central to Chronos-T5's forecasting on ETT data.

Limitations. The taxonomy classifier is heuristic, with 82.8% of features globally remaining unlabeled and decoder coverage below 6%. Ablation experiments use ETT data only, and we analyze only Chronos-T5-Large. The ultra-fast ablation configuration (256 windows, 4 samples) provides directional but statistically limited findings. Future work includes SAEs at larger expansion factors, supervised probes for improved taxonomy coverage, cross-architecture comparisons, and circuit-level analyses connecting features to forecast outputs.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2403.07815.

Matthieu Boileau, Philippe Helluy, Jérémy Pawlus, and Svitlana Vyetrenko. Towards interpretable time series foundation models. *arXiv preprint arXiv:2507.07439*, 2025.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/.

Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2024.

Joseph Enguehard. Learning perturbations to explain time series predictions. In *International Conference on Machine Learning (ICML)*, 2023.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *International Conference on Machine Learning (ICML)*, 2024.

Matīss Kalnāre, Sofoklis Kitharidis, Thomas Bäck, and Niki van Stein. Mechanistic interpretability for transformer-based time series classification. *arXiv preprint arXiv:2511.21514*, 2025.

Christodoulos Kechris, Jonathan Dan, and David Atienza. Time series saliency maps: Explaining models across multiple domains. *arXiv preprint arXiv:2505.13100*, 2025.

Zichuan Liu, Yingying Zhang, Tianchun Wang, Zefan Wang, Dongsheng Luo, Mengnan Du, Min Wu, Yi Wang, Chunlin Chen, Lunting Fan, and Qingsong Wen. Explaining time series via contrastive and locally sparse perturbations. In *International Conference on Learning Representations (ICLR)*, 2024.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Owen Queen, Thomas Hartvigsen, Teddy Koker, Huan He, Theodoros Tsiligkaridis, and Marinka Zitnik. TimeX: Encoding time-series explanations through self-supervised model behavior consistency. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Inmaculada Santamaria-Valenzuela, Victor Rodriguez-Fernandez, Javier Huertas-Tato, Jong Hyuk Park, and David Camacho. Decoding latent spaces: Assessing the interpretability of time series foundation models for visual analytics. *arXiv preprint arXiv:2504.20099*, 2025.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Angela van Sprang, Erman Acar, and Willem Zuidema. Interpretability for time series transformers using a concept bottleneck framework. *arXiv preprint arXiv:2410.06070*, 2024.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*, 2024.

Jingquan Yan and Hao Wang. Self-interpretable time series prediction with counterfactual explanations. In *International Conference on Machine Learning (ICML)*, 2023.

Ziqi Zhao, Yucheng Shi, Shushan Wu, Fan Yang, Wenzhan Song, and Ninghao Liu. Interpretation of time-series deep models: A survey. *arXiv preprint arXiv:2305.14582*, 2023.

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.

A LLM Usage Disclosure

In accordance with ICLR 2026 policies on LLM usage, we disclose the following: Claude (Anthropic) and OpenAI Prism was used to assist with drafting and editing portions of this manuscript and for generating experiment infrastructure code. All experimental results, numerical findings, and scientific claims were produced by the authors' own experiments and verified independently against raw output artifacts. The authors take full responsibility for all content.

B Additional Experimental Details

SAE Training. Each SAE was trained on residual stream activations extracted from approximately 100,000 time series windows with a batch size of 2,048 and MSE reconstruction loss. TopK sparsity $k = 64$ was selected for reconstruction quality while maintaining interpretable sparsity.

Ablation Configurations. The *ultra-fast* configuration (uFast v2) used 256 ETT windows, 4 forecast samples, batch size 32, features selected by decoder-norm ranking, and progressive checkpoints at $\{1, 2, 4, 8, 16, 32, 64\}$. The *extended* configuration (fast v1, encoder block 23 only) used 1,024 windows, 8 samples, batch size 16, and checkpoints at $\{1, 2, 4, 8, 16, 32, 64, 96, 128, 160, 200\}$.

Taxonomy Progression. The classifier was iteratively improved: Stage 0 had a wiring bug (0% labeled), Stage 1 fixed activation loading (15.3%), and Stage 2 (this paper) added correlation-aware directional labeling (17.2%).