TOPOPRIMER: THE MISSING TOPOLOGICAL CONTEXT IN FORECASTING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many leading time series foundation models (TSFMs) forecast individual series based predominantly on their own temporal histories, relying on implicit cross-attention to capture any broader context. It remains largely unexplored whether explicit, domain-wide population structure can improve these models. Specifically, it is unclear if a series' relative position within a population, or the global shape of the population itself, offers a recoverable signal for forecasting. To investigate this, we introduce TopoPrimer, a framework that injects the global topological structure of a series population as an explicit prior into forecasting backbones. TopoPrimer captures this structure through two frozen encodings, precomputed once per domain: a spectral relational coordinate that locates each series within the population, and a persistent-homology fingerprint that describes the population's global shape. These descriptors are injected either per-token into models trained from scratch, or via a lightweight adapter into frozen, pre-trained foundation models. We find a strict division of labor: the relational coordinate varies across series and drives accuracy on frozen backbones when history is abundant, whereas the topological fingerprint is constant within a population and relies entirely on cross-cohort structural diversity to provide predictive signal. TopoPrimer improves accuracy on three of four public benchmarks, reducing MSE by 7.4% on ECL. Notably, on the ECL dataset, the relational coordinate outperforms full fine-tuning of Chronos while training fewer than 1% of the parameters. On a large-scale internal corpus, the fingerprint reduces peak-demand error by 9 to 14% and improves cold-start MAE by 13% at launch. Because a replication on the public M5 dataset yielded no aggregate cold-start gain, we scope this zero-shot advantage to training distributions spanning structurally distinct populations.

1 INTRODUCTION

Leading time series foundation models (TSFMs), such as Chronos (Ansari et al., 2025) and TimesFM (Das et al., 2024), encode sequences independently and capture cross-series structure only implicitly via attention. This paradigm leaves a critical, domain-wide signal unexploited: the underlying geometry of the series population. In reality, individual trajectories exhibit complex behavioral groupings, form cyclical dependencies, and cluster into distinct topological regions. We formalize this recoverable cross-series signal as *topological context*: the global shape of the population and the relative position of each series within it (App. U.1).

TopoPrimer makes this context a frozen, precomputed input to any forecasting backbone. Its first component, a 256-dimensional *relational coordinate*, locates each series using the leading left singular vectors of the entity-time matrix (sheaf motivation (Curry, 2014; Hansen and Ghrist, 2021) in Appendix W). Unlike sheaf networks that train graph convolutions (Bodnar et al., 2022; Mostafa et al., 2026), it needs no training or graph. The second component is a 125-dimensional *persistent homology* fingerprint (H_0 clusters, H_1 loops, H_2 voids) applied to the cross-series correlation-distance matrix. While prior work applies homology to individual series (Zeng et al., 2021; Lin et al., 2025; 2026), we compute it globally across the population. Each descriptor enters through its own learned projection: the projections are summed and broadcast-added to every token of a fully-trained model, or concatenated inside a lightweight adapter (< 0.2M parameters) that corrects a frozen pre-trained forecast.

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

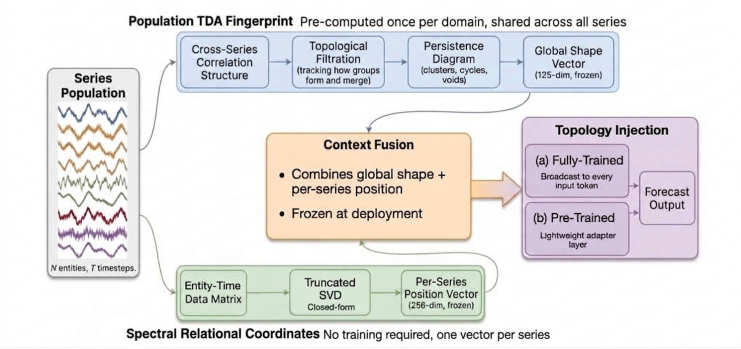

Figure 1: **TopoPrimer overview.** A 256-dim per-series spectral relational coordinate (truncated SVD) and a 125-dim population TDA fingerprint are precomputed and frozen, then injected by broadcast-addition to every input token *(a)* or via a lightweight adapter on a frozen backbone *(b)*.

Across four public datasets, TopoPrimer improves the performance of both models trained from scratch and frozen pre-trained backbones (e.g., yielding a 7.4% MSE gain on ECL with Chronos). Notably, on ECL, it outperforms both LoRA and full fine-tuning. On a large-scale internal corpus, these gains are especially pronounced during seasonal spikes and zero-shot cold starts.

Contributions. (i) **Graph-free relational coordinates.** Derived in closed form from the entity-time matrix, this sheaf-motivated prior requires no graph construction and drives the primary accuracy gains on frozen backbones. (ii) **Unified topological injection.** The same frozen descriptors condition both from-scratch Transformers (via per-token broadcast) and frozen TSFMs (via a lightweight residual adapter). On ECL, this frozen-backbone approach surpasses LoRA and full fine-tuning using a fraction of the parameters. This efficiency stems primarily from the adapter design, while the topological descriptors provide a consistent supplementary gain holds under two distinct adapter implementations, each applied to a different pair of datasets (App. K). (iii) **Evaluation of population-level TDA.** We introduce population-level persistent homology to time series forecasting. While its impact on standard single-population benchmarks is inherently marginal, it yields a substantial 13% reduction in launch-MAE for cold starts on our internal corpus. By contrast, a seeded public analog on the M5 Household yielded no aggregate gain and degraded performance on novel items. We trace this discrepancy directly to M5's lack of cross-cohort structural diversity. Rather than claiming blanket generalizability, we explicitly scope this cold-start advantage to training distributions with distinct cohort topologies—a requirement we characterize in App. M and pair with a screening heuristic (App. N).

2 Related Work

Relational structure is typically learned, not supplied. In modern forecasting, relational coordinates are almost exclusively learned during training or derived from predefined graphs. Global-factor models (Wang et al., 2019) learn low-rank series coordinates jointly with the forecasting objective. While conceptually similar to ours, these latent positions are intrinsically tied to a specific training run. Alternatively, spatio-temporal Transformers rely on spectral positional encodings extracted from a provided graph (Dwivedi and Bresson, 2020; Jiang et al., 2023). Graph neural networks (Li et al., 2018; Wu et al., 2019; 2020) and sheaf networks (Bodnar et al., 2022; Mostafa et al., 2026) go further, entirely replacing or heavily augmenting the forecasting backbone to learn relational structure (e.g., DCRNN and MTGNN learn adaptive adjacencies, while ST-Sheaf GNN acts as the end-to-end model). Crucially, these approaches all require a graph to exist (or be learned), inherently coupling the relational structure to the forecaster's architecture. By contrast, our relational coordinate is computed in closed form, requires no graph, and leaves the backbone completely untouched—functioning as a frozen prior rather than an architectural modification.

Frozen backbones are adapted, not informed. This distinction is critical: the backbones we adapt—standard Transformer forecasters (Zhou et al., 2021; Nie et al., 2023; Lim et al., 2021) and TSFMs (Ansari et al., 2025; Das et al., 2024)—process individual series independently and are

typically adapted via LoRA (Hu et al., 2022) or in-context fine-tuning (Das et al., 2025). While foundation models with native cross-variate attention have been proposed (Woo et al., 2024; Gao et al., 2024; Ekambaram et al., 2024), we contrast these architectures with our frozen prior approach in App. X. Standard adaptation methods expend parameters and gradient steps attempting to re-derive cross-series structure from individual histories—structure that the domain already exposes in closed form. We invert this paradigm: the backbone remains frozen, and the population's structural context is directly injected as a precomputed prior. Section 4.1 rigorously evaluates this approach against LoRA, full fine-tuning, and a parameter-matched partial fine-tune, strictly controlling for the backbone, data splits, and random seeds.

Topology is applied per series, not per population. Prior applications of TDA to time series apply persistent homology (Carlsson, 2009; Edelsbrunner and Harer, 2010) to individual series via sliding-window or zigzag constructions (Perea and Harer, 2015; Zeng et al., 2021; Lin et al., 2026; 2025; Kim et al., 2026; Chen et al., 2021; Tymochko et al., 2020; Zeng et al., 2024; Kim and Mémoli, 2021; Cohen-Steiner et al., 2006), yielding a separate descriptor for each window of *within*-series dynamics. In contrast, we apply TDA once to the entire population within a cross-series correlation-distance space, yielding a single global descriptor for the domain. Because these two approaches capture fundamentally different signals, we evaluate the per-series construction as a direct empirical competitor rather than a mere theoretical foil. Specifically, Section 4.2 evaluates a capacity- and injection-matched sliding-window (Takens) descriptor alongside our coordinate, where it outperforms our method in only one of six experimental settings (Table 10). Conceptually closest to our work is TADA (Chazal et al., 2024), which computes Rips persistence over channel correlations to detect distribution shifts in dependence structure. Rather than using this topology for anomaly detection, we extract and freeze it into a structural prior that conditions a foundation-model adapter (App. X).

3 METHOD

TopoPrimer extracts two population-level descriptors offline from the training window $\mathbf{X} \in \mathbb{R}^{N \times T_{\text{train}}}$ prior to validation and testing (Appendix I). These are fused and injected into forecasting backbones without modifying pre-trained weights (Figure 1; Appendices A and U).

Spectral relational coordinates. The coordinate of series i (denoted **"Rel"**) is row i of U, the left factor of a truncated singular value decomposition (SVD) $\mathbf{X} \approx U\Sigma V^\top$ of the entity-time matrix, whose rows are ℓ_2-normalized before the factorization. When series span unrelated categories (e.g., M5), we apply SVD within semantically coherent groups. We retain the top $k = \min(256,\, N - 1,\, T - 1)$ left singular vectors, zero-padding to a **256-dimensional coordinate** per series (Fig. 6). An entity with no row in $\mathbf{X}$ — one that launches after the descriptor cutoff — has no coordinate, and no out-of-sample projection recovers one: the usual formula $\mathbf{x}_{\text{new}} V\Sigma^{-1}$ requires an observed context window, which at launch is empty. Such an entity instead inherits the mean coordinate of a set $\mathcal{N}$ of established entities, chosen purely from static metadata known on the launch date (identity siblings in other locations where they exist, otherwise the $K{=}10$ nearest established entities in a static metadata-and-price feature space) and expressed in one shared basis so that the average is well defined. This is the only case in which a coordinate is not a row of U, and it arises only in our cold-start evaluation, which gives the construction in full (App. T). This closed-form approach matches or beats a learned neural sheaf encoder at a fraction of the cost on M5, where both were run (App. H).

Population TDA fingerprint. To capture the population's global shape, we compute the correlation-distance matrix $D_{ij} = 1 - \rho_{ij} \in [0, 2]$, where ρ_{ij} is the Pearson correlation between series i and j. This metric naturally places anti-correlated series far apart. We sparsify this matrix to each series' $k{=}50$ nearest neighbors across all datasets (clamping $k \leq N$; see App. G for metric robustness). We run a Vietoris-Rips filtration (Tralie et al., 2018) up to dimension 2, extracting the birth-death pairs for H_0 (clustering), H_1 (cyclic co-movement), and H_2 (enclosed voids). Each diagram is vectorized via the persistence landscape (Bubenik, 2015) by sampling λ_1 and λ_2 at 25 points for H_0 and H_1, and λ_1 at 25 points for H_2. All points are sampled on a single grid spanning the full filtration range (App. K). This yields a fixed **125-dimensional TDA fingerprint** that is computed once per population and broadcast to every series within it: one per domain on ECL, Monash Weather, and METR-LA; one per store×category on M5; and one per cohort on the internal corpus (Fig. 5). While the persistence landscape is Lipschitz-continuous with respect to the persistence diagram (App. V), the upstream k-

NN sparsification introduces discontinuities at neighbor-rank ties. Consequently, end-to-end stability from the raw panel is demonstrated empirically rather than theoretically guaranteed (App. G). Because the fingerprint is constant across a given population, it provides no series-discriminating variance within it. On a closed, single-population benchmark, it is simply absorbed into a layer bias, meaning the +TDA variant spans exactly the same function class as the vanilla model. It acts strictly as a cohort-level prior, becoming informative only when the *training* distribution spans multiple cohorts.

Fully-trained Transformers. Our primary from-scratch backbone is a standard Transformer encoder ($d_{\text{model}} = 256$, 6 layers, 8 heads, pre-norm) equipped with a learned linear input embedding and sinusoidal positional encodings. We map the contextual features into $\mathbb{R}^{d_{\text{model}}}$ using two distinct learned linear projections. The first, $\mathbf{W}_{\text{ctx}}$, processes static inputs (learned entity embeddings, temporal features, and the TDA fingerprint). These remain constant across the sequence because calendar features are sampled at a single index per window (App. J). The second, $\mathbf{W}_{\text{rel}}$, projects the relational coordinate. We define the global context vector $\mathbf{g}$ as their sum, which is broadcast-added to every temporal token: $\mathbf{z}_t \leftarrow \mathbf{z}_t + \mathbf{g}$ for $t = 1, \ldots, L$. Crucially, no gradients propagate back to the frozen inputs (Figure 2). We maintain separate projections as a design precaution against early optimization dynamics suppressing the relational signal (App. K). The model is trained to minimize a Huber quantile loss ($\delta = 1.0$) over 9 quantiles (though Monash Weather uses a deterministic point forecast, following the Chronos protocol).

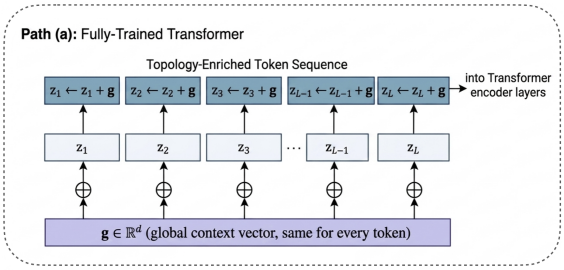

Figure 2: **Global context broadcast injection (Path (a)).** $\mathbf{g} \in \mathbb{R}^d$ is broadcast-added to every temporal token before the encoder. Gradients flow back through $\mathbf{g}$ into the learned projections $\mathbf{W}_{\text{ctx}}$ and $\mathbf{W}_{\text{rel}}$ (and the entity embeddings), but stop there: the fingerprint and the relational coordinate are precomputed and frozen, so no gradient reaches them.

Pre-trained foundation models. For pre-trained backbones, we freeze all network weights and train a lightweight *topology adapter* to correct the base forecast. We explore two adapter variants (Table 17 in App. K), both taking four shared inputs: **Rel** (the 256-dimensional coordinate), **TDA** (the 125-dimensional fingerprint), **Stat** (z-scored series statistics), and **Forecast** (the cached median base forecast). Both adapters predict a learned residual that is added to the base forecast, differing only in their input mixing strategies. **Fused adapter.** For ECL and M5, the four inputs are directly concatenated without per-input projections and passed into a two-layer MLP to predict the residual. **Four-branch adapter.** For Monash Weather and METR-LA, to prevent any single input from dominating by sheer dimensionality, each input is first projected into a common hidden space via independent branches (Figure 3). These representations are then concatenated and passed through an output MLP. **Evaluation variants.** Across both settings, we evaluate four variants: **Vanilla** (no topology), +**TDA**, +**Rel**, and +**TDA** + **Rel** (the full model). For the fused form, inactive inputs are dropped from the concatenation. For the four-branch form, inactive branches are retained but zero-filled, ensuring the Vanilla and +TDA variants are strictly parameter-matched. Full architectural specifications, including exact MLP widths and dataset-specific statistical features, are detailed in App. K.

4 RESULTS

Main results. As detailed in Table 1, the relational coordinate drives the primary accuracy gains. On ECL, adding +**Rel** reduces Chronos MAE from 0.3019 to 0.2899 (a 4.0% MAE improvement; 7.2% on MSE) and TimesFM MAE from 0.3015 to 0.2893, exhibiting remarkably high stability ($\sigma \leq 0.0005$, App. C). On Monash Weather, Chronos +**Rel** achieves the lowest MAE of any variant

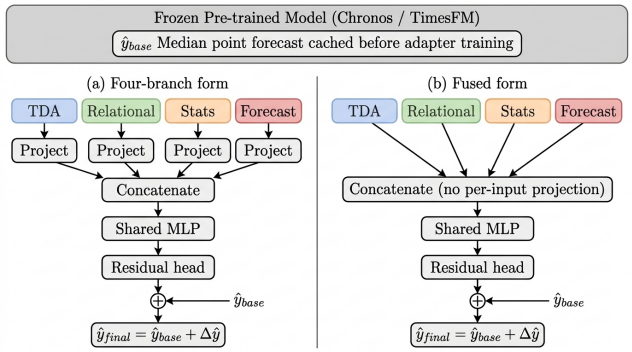

Figure 3: **Topology adapter for frozen pre-trained backbones.** Both forms take the same four inputs — the TDA fingerprint, the per-series spectral relational coordinate, z-scored series statistics, and the frozen backbone's cached median forecast — and predict a residual added to that base forecast. *(a)* The four-branch form, used on Monash Weather and METR-LA, first projects each input into a common hidden space so that no input dominates by sheer dimensionality. *(b)* The fused form, used on ECL and M5 and therefore supplying the headline results, concatenates the four inputs directly and uses a median-only head. Exact widths, parameter counts, and the per-dataset statistical features are given in App. K (Fig. 7, Table 17).

Table 1: **Forecasting performance** (mean over 5 seeds). All public-benchmark evaluations use a single forecast origin per series, predicting a single horizon block immediately following the split boundary, rather than the rolling test splits common in the LTSF literature (App. I). **Bold** indicates the best result per column within a backbone block (↓ lower is better). Zero-shot baselines are deterministic single runs and are excluded from bolding. Secondary metrics include MAPE (METR-LA, H=15 min), MSE (ECL, H=96 h), and WAPE (M5, H=4 wk). †Monash Weather is scored at a single lead time (day 30, the final step of the forecast), matching the protocol for METR-LA (scored at the 15-minute step). Conversely, ECL averages over its full prediction block to align with the LTSF literature baselines. Because Monash Weather lacks external baselines in this setting, no published comparisons are affected by this choice; full 30-day block means and per-day breakdowns are provided in App. Q. M5 reports the HOUSEHOLD category. *Comparisons should only be made within backbone blocks*, as M5 units vary across blocks, and adapter forms and training objectives differ across datasets (App. K). Standard deviations are reported in App. C. Adapter rows utilize the four-branch form for METR-LA and Monash Weather, and the fused form for ECL and M5 (App. K). Seed means alone characterize only initialization variance. We therefore add a second, orthogonal source of uncertainty — variability across test entities — scored on the seed-averaged forecast; seed variance is absorbed by that averaging (App. C.2). [a]Computed over 4 of 5 seeds.

	METR-LA		ECL		Monash Weather		M5 (Household)	
Model	MAE↓	MAPE↓	MAE↓	MSE↓	MAE† ↓	MSE† ↓	MAE↓	WAPE↓
Transformer	2.294	3.934	**0.1956**	0.0959	2.1765	25.912	1.8732	0.2653
Transformer + TDA	2.254	3.877	0.1969	0.0971	2.2071	26.396	**1.8674**	**0.2645**
Transformer + Rel	2.264	3.896	0.2006	0.1028	2.0494	25.285	1.8843	0.2669
Transformer + TDA + Rel	**2.240**	**3.856**	0.1964	**0.0942**	**2.0489**	**25.107**	1.8925	0.2681
Chronos Zero-Shot	3.348	5.615	0.586	0.610	2.3439	29.776	0.9132	0.9816
Chronos Vanilla Adapter	2.376	4.061	0.3019	0.2047	2.0156	28.283	**0.7965**	**0.8328**
Chronos + TDA	2.424	4.149	0.3018	0.2047	2.0135	28.319	0.7977	0.8340
Chronos + Rel	2.345	4.009	0.2899	0.1900	**1.9880**	**28.120**	0.7966	**0.8328**
Chronos + TDA + Rel	**2.337**	**4.003**	**0.2897**	**0.1895**	1.9960	28.155	0.7973	0.8335
TimesFM Zero-Shot	2.441	4.200	0.580	0.602	2.0316	28.032	0.9401	0.9828
TimesFM Vanilla Adapter	2.382	**4.097**	0.3015	0.2052	2.0363	28.126	0.7959	0.8321
TimesFM + TDA	2.387	4.104	0.3013	0.2050	**2.0355**	28.094	**0.7955**	**0.8317**
TimesFM + Rel	**2.381**	4.102	**0.2893**	0.1903	2.0582	28.088	0.7971[a]	0.8334[a]
TimesFM + TDA + Rel	2.385	4.105	**0.2893**	**0.1900**	2.0385	**28.000**	0.7957	0.8319

in Table 1 (1.9880), although weight-updating adaptation of the same backbone performs even better on this specific dataset (Section 4.1). This adapter ordering is specific to the day-30 lead: averaged over the full 30-day block, the vanilla adapters match or beat their topology variants, while the

from-scratch Transformer retains its gain (App. Q). For Transformers trained from scratch, +**Rel** reduces MAE by 5.8% on Weather and 1.3% on METR-LA (App. C.1). While +**Rel** alone slightly degrades performance on the compact ECL dataset, the combined +**TDA**+**Rel** model achieves the best overall MSE (0.0942). Because ECL consists of a single population (and thus a single fingerprint), this recovery reflects favorable optimization dynamics rather than the injection of new information (App. M). Conversely, M5 yields null results across all variants; its correlation structure appears heavily calendar-dominated, which likely explains this lack of topological signal (App. N, S). Finally, we note that these gains are horizon-dependent: the relational coordinate provides no advantage at Weather's shortest forecast steps (likely because local autocorrelation dominates), nor at ECL's longest horizons (where we hypothesize the backbone's learned long-range prior takes over) (App. Q, P).

In contrast to the relational coordinate, the TDA fingerprint alone has a marginal impact on standard benchmarks—a result entirely consistent with its single-population invariance (App. M). Its effect on MAE is small and inconsistent in sign (e.g., shifting METR-LA Transformer MAE by -0.040, but Weather by $+0.031$), and adding it to +**Rel** shifts performance by at most ± 0.024. The fingerprint's value instead emerges in the zero-shot cold-start regimes evaluated below. Separately, persistence statistics computed *over* cohorts serve as an offline pre-deployment screen; that screen is validated on internal cohorts rather than on these four domains (App. N).

Significance beyond seed noise. Seed means alone characterize only initialization variance. We therefore introduce a second, orthogonal source of uncertainty—variability across test entities. For the headline cells (Table 2), we compute paired cluster-bootstrap 95% BCa confidence intervals (B=5000, resampled by client or station) and Wilcoxon signed-rank tests. These are scored on the seed-averaged forecast, meaning seed variance is absorbed by the averaging (App. C.2). The relational coordinate demonstrates statistical significance on both metrics in the two settings we evaluate with paired intervals. On ECL with Chronos, it improves $\approx$81% of individual clients ($\widehat{\Delta\text{MAE}} = -0.0120$, 95% CI $[-0.0152, -0.0083]$, $p < 0.001$). On Monash Weather with the from-scratch Transformer, it achieves $\widehat{\Delta\text{MAE}} = -0.1265$ (95% CI $[-0.1461, -0.1071]$, $p < 0.001$)—the strongest statistically backed result in this paper. Conversely, the TDA fingerprint alone fails to achieve significance in either setting: on ECL its effect is negligible (-0.0001, p=0.097), and on Weather it performs significantly *worse* than the vanilla baseline ($+0.0348$, $p < 0.001$). We also report one further counter-example not covered in the table: on ECL with the from-scratch Transformer, +**Rel** alone is statistically worse than vanilla ($\widehat{\Delta\text{MAE}} = +0.0044$, 95% CI $[+0.0018, +0.0077]$), which is consistent with the $0.1956 \rightarrow 0.2006$ degradation in seed means noted above. Finally, on M5, every variant falls within the margin of seed noise (gain/$\sigma < 1.2$, $p > 0.05$ in every cell). This strict statistical bound confirms our M5 findings as a true null result rather than an artifact of initialization variance (App. C).

4.1 Competing Ways to Adapt the Same Backbone

Because the relational coordinate serves as an alternative to updating backbone weights, it must be evaluated against standard weight-updating methods. Table 7 in App. D isolates the adaptation method by strictly controlling for the Chronos-2 backbone, data splits, evaluation code, and random seeds. On ECL, the frozen adapter outperforms LoRA, full fine-tuning, and a partial fine-tune matched to its exact parameter count. It achieves this while training a parameter budget equivalent to just 0.64% of the backbone and requiring no backpropagation through the foundation model itself. Crucially, the majority of this performance margin stems from the adapter architecture rather than the topological prior. The topology-free Vanilla adapter achieves an MAE of 0.262 (already ahead of LoRA and full fine-tuning), while adding +**Rel** provides a further 0.005 improvement. Though numerically small, this refinement sits well outside the margin of seed noise (0.2620 ± 0.0002 vs. 0.2573 ± 0.0005). Thus, at this parameter budget, the frozen-backbone adapter fundamentally drives the advantage over weight-updating methods, while the relational coordinate acts as a statistically significant refinement.

On Monash Weather, our approach is highly competitive: among sub-1% methods, it trails LoRA by only 0.044 MAE, and it remains competitive with full fine-tuning despite the latter utilizing $157\times$ more trainable parameters. Note that the ECL column of Table 7 scores forecasts without a non-negativity clamp, differentiating it from Table 1 (App. D). We hypothesize that LoRA's slight edge on Weather stems from weight-level calibration to the dataset's specific seasonality. This suggests that weight-updating and closed-form relational priors offer complementary benefits

Table 2: **Test-set paired cluster-bootstrap CIs** for ΔMAE (topology $-$ vanilla), calculated from forecasts averaged across 5 training seeds (B=5000). ECL is resampled by client (N=321) using the Chronos backbone; Monash Weather is resampled by station (N=3,010) using the from-scratch Transformer at the day-30 lead time. Negative values indicate improvement; **bold** denotes the largest-magnitude significant gain per panel. Absolute columns report cluster-level means and are rounded independently of the paired Δ, meaning exact subtraction may differ slightly due to rounding. Notably, the Weather absolute columns are 0.025–0.029 lower than the mean-of-seeds values in Table 1. This occurs because averaging five independent forecasts inherently reduces absolute error; because the offset applies to all rows it largely cancels in the paired Δ, leaving residuals of at most 0.004. Full per-panel details are provided in App. C.2.

Variant	Vanilla	Topo	$\widehat{\Delta\text{MAE}}$	95% CI	Wilcoxon p
ECL, Chronos backbone (resampled by client, N=321)					
+TDA	0.3019	0.3018	-0.0001	$[-0.0003, -0.0000]$	$0.097^{\dagger}$
+Rel	0.3019	0.2899	-0.0120	$[-0.0152, -0.0083]$	$< 0.001^{***}$
+TDA+Rel	0.3019	0.2896	$\mathbf{-0.0123}$	$[-0.0155, -0.0089]$	$< 0.001^{***}$
Monash Weather, fully-trained Transformer, day-30 lead time (by station, N=3,010)					
+TDA	2.1475	2.1823	$+0.0348$	$[+0.0217, +0.0473]$	$< 0.001^{***}$
+Rel	2.1475	2.0210	$\mathbf{-0.1265}$	$[-0.1461, -0.1071]$	$< 0.001^{***}$
+TDA+Rel	2.1475	2.0231	-0.1243	$[-0.1439, -0.1037]$	$< 0.001^{***}$

$^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. On Weather the +Rel / +TDA+Rel ordering differs from Table 1, but both margins (0.0021 here, 0.0005 there) lie far inside the training-seed spread ($\sigma \approx 0.03$–0.07): the two variants are tied and the ordering should not be read.

rather than acting as strict substitutes. However, the practical advantage of our adapter lies in its computational efficiency. Because it trains exclusively on cached base forecasts, the foundation model requires only a single forward pass and is never differentiated through. Consequently, adapter training costs remain constant and do not scale with backbone size (App. L).

4.2 Controls: What Is the Topology Actually Buying?

A frozen prior might improve performance simply by exposing the model to generic population-level information. To test this directly, we inject unstructured statistical baselines into the identical architectural slot, matching the dimension and seed count (Table 8, App. E). Specifically, we evaluate `CovEig` (the top-125 eigenvalues of the train-window correlation matrix, bypassing the persistent-homology step) and `CorrRow` (each series' raw correlation row projected to 256 dimensions, replacing the relational coordinate).

A rigorous reading suggests that for frozen adapters, the majority of the gain stems from supplying *any* cross-series prior. Both +**Rel** and `CorrRow` reduce ECL MAE to $\approx$0.290; while +**Rel** statistically outperforms `CorrRow` on ECL (p=0.018), the two tie on Weather. The eigenvalue comparison, by contrast, is uninformative by construction: `CovEig` is broadcast to every series exactly as the fingerprint is, so on a single-population benchmark both are absorbed into a bias (App. M). We report that contrast for completeness but draw no conclusion from it either way. Nevertheless, we favor the spectral coordinate on structural grounds rather than sheer performance margins: its dimension (256) remains fixed, whereas a raw correlation row scales with the population size N and requires reprojection whenever the population composition changes.

Per-series topology as a competitor. The natural topological alternative is the per-series construction utilized in prior work. To evaluate this, we construct a capacity- and injection-matched sliding-window (Takens) persistence descriptor (utilizing a delay embedding $\rightarrow$ Vietoris–Rips $\rightarrow$ landscape pipeline, projected to 256 dimensions and injected identically). The empirical results are nuanced: across six dataset-by-backbone configurations, +**Rel** achieves the best performance in four, the sliding-window descriptor in one, and the vanilla baseline in one. Notably, the sliding-window descriptor outperforms the vanilla baseline in five of the six settings, confirming it is a strong, active baseline rather than an inert control. The single setting where the sliding-window descriptor wins is the ECL dataset with the from-scratch Transformer (0.1926 MAE, compared to +**Rel**'s 0.2006 and

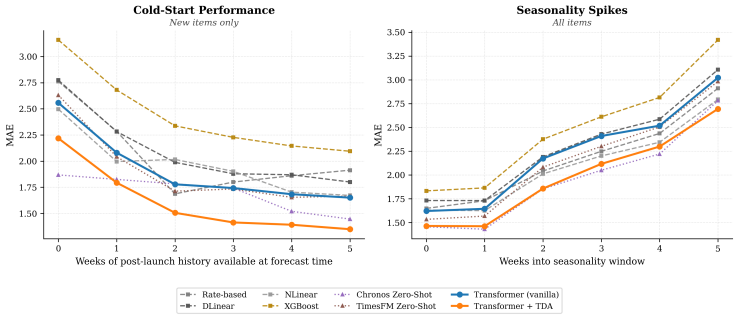

Figure 4: Performance in hard regimes on the internal corpus. (a) Cold-start MAE vs. weeks of post-launch history (evaluated on new items only, using an identical series set across models). The +**TDA** variant outperforms the vanilla Transformer at every origin; from week 2 onward, it establishes an error floor that the vanilla model never reaches within the evaluation window. (b) Seasonality window MAE over peak demand (evaluated on all items). While every method degrades as demand surges, the +**TDA** variant maintains the lowest error among all trained and classical models, eventually overtaking zero-shot Chronos at the week-5 peak. Per-week tabular results are provided in App. B. Single training seed (42).

vanilla's 0.1956). Crucially, this is the exact same from-scratch, smallest-population configuration where `CorrRow` also defeated +**Rel**. This reinforces our earlier scoping: the from-scratch model's trainable capacity absorbs the need for a precomputed relational prior. Conversely, on the four frozen-backbone configurations, the sliding-window descriptor never achieves the best performance. Furthermore, it is fundamentally undefined in cold-start scenarios, as there is no historical window to embed. In contrast, our precomputed relational coordinate is available from the very first time step (App. F). Per-horizon breakdowns, metric-robustness checks, and a comparison between spectral and learned sheaves are provided in Apps. G–O.

4.3 Two Hard Regimes

Because standard public benchmarks lack repeated seasonal spikes and true zero-shot cold starts, we evaluate these challenging regimes on a large-scale internal corpus (4,575 entities $\times$ 603 items) utilizing only the population fingerprint (+**TDA**). This is precisely the regime the fingerprint was designed to address, and it represents the only setting in this paper where it serves as the primary driver of performance. For this evaluation, each week index acts as an independent forecast origin, every model is scored on an identical set of series at that origin, and MAE is averaged exclusively over weeks exhibiting non-zero demand, at a single training seed (App. B).

Seasonal spikes. During a six-week peak-demand window, the from-scratch Transformer with +**TDA** achieves the lowest MAE of any trained or classical model across every week. It consistently reduces error by 9–15% relative to the vanilla Transformer and 2–10% relative to NLinear. Notably, zero-shot Chronos outperforms all models from week 0 through week 4 (e.g., 2.221 vs. 2.299 at week 4). However, Chronos degrades more rapidly as demand builds (a +91% vs. +84% error increase), allowing +**TDA** to surpass it at the week-5 peak (2.695 vs. 2.782 MAE; Fig. 4b, App. B).

Cold start. We define "new items" as those whose first non-zero sale falls within six weeks of the forecast origin. At launch, such an item's context window is entirely zero-filled, meaning no variant has access to item-level history. Consequently, the precomputed fingerprint (derived from a prior population and thus available before the item even exists) serves as the sole differentiating signal against the vanilla Transformer. Both variants are otherwise trained in a single batch with identical data, random seeds, learning schedules, and epoch budgets. Strikingly, injecting +**TDA** reduces launch MAE by 13.3% (2.219 vs. 2.558). This performance gap widens as history accumulates, peaking at an 18.8% improvement by week 3 (1.414 vs. 1.742). From week 1 onward, +**TDA** outperforms all evaluated models, including zero-shot TSFMs (App. B).

However, a seeded public analog on the M5 Household dataset yielded no aggregate gain and degraded performance on truly novel series (3% for $+$TDA, 8% for $+$Rel) (App. T). This discrepancy highlights a strict design condition: because the fingerprint is constant within a population, it provides predictive signal only when the training distribution spans structurally distinct cohorts (App. M). While our internal training set was deliberately widened to ensure this diversity, M5 fails this condition by construction. Its ten store$\times$category fingerprints are topologically near-identical (mean pairwise cosine 0.997), causing $+$**TDA** to collapse into an uninformative global bias. Repartitioning the panel until the training fingerprints measurably separate is the natural remedy, which we do not attempt here. This requirement for measurable topological variance across the training partition directly motivates the screening heuristic introduced in App. N.

5 Conclusion

TopoPrimer injects two precomputed, frozen topological priors into forecasting backbones: a spectral relational coordinate and a persistent-homology fingerprint. The relational coordinate improves accuracy on frozen backbones across our public benchmarks, with a documented counter-case on the smallest from-scratch setting. Notably, on ECL, our frozen-adapter approach outperforms both LoRA and full fine-tuning using only a fraction of the parameter budget. The coordinate refines this baseline further, and this adapter-driven efficiency proves robust across entirely different architectural implementations. Conversely, while the topological fingerprint yields only marginal gains on standard benchmarks, it excels in highly challenging regimes on a large-scale internal corpus. Specifically, it minimizes error during peak-demand windows and provides a critical zero-shot signal for new-item cold starts.

Perhaps our most surprising finding is the strict division of labor between these two descriptors. The relational coordinate dominates when historical data is abundant and the backbone lacks population-fitted parameters. The global fingerprint, conversely, acts strictly as a cohort-level prior, providing predictive signal only when the training distribution spans structurally distinct cohorts (App. M)—the condition behind our peak-demand and cold-start gains. Crucially, neither descriptor acts as a general-purpose performance booster. As demonstrated by our controls (Section 4.2), unstructured correlation statistics appear capable of doing much of the same work — most clearly once a model is trained on the target population, but also on the frozen adapter (Section 4.2). Ultimately, this suggests that the true value of a frozen topological prior lies in granting the model access to macro-structure it cannot otherwise observe, rather than simply increasing its raw expressiveness. The fingerprint's evidence base is bounded by cohort diversity, not by the method. Every setting evaluated here supplies few distinct cohorts, so the fingerprint contributes only a low-dimensional per-cohort offset—and still cuts launch MAE by 13.3%. Testing it properly needs corpora spanning many distinct populations: multi-retailer, multi-region, or multi-domain panels, where a per-cohort prior has room to vary.

Limitations. Our hard-regime evaluations rely on an unreleased internal corpus (App. B, J). Furthermore, the cold-start advantage appears corpus-specific: a five-seed public replication on the M5 Household dataset yielded no aggregate gain and degraded performance on truly novel series (App. T). We therefore explicitly scope our cold-start claims to training distributions spanning structurally distinct populations (Section 4.3). No public benchmark evaluated here meets that condition, so this scoping also concedes that our only positive fingerprint evidence rests on a corpus that cannot currently be replicated publicly. Additionally: (1) The relational coordinate's advantage is scoped strictly to frozen backbones lacking population-fitted parameters; it performs statistically worse than the vanilla baseline and both controls on our smallest benchmark when training a Transformer from scratch (Sections 4, 4.2). (2) The screening criterion (App. N) is retrospective — the median H_1 lifetime was selected as the strongest of sixteen candidates against the same nineteen cohorts used to measure the gains — so it is not a universal rule and does not perfectly rank the public domains. (3) Public results use a single forecast origin per series (App. I). While within-table comparisons are exact, they carry no estimate of temporal sampling variance and should not be directly compared to rolling-split baselines. (4) We offer no theoretical account for *why* the relational coordinate improves performance. Our sole formal proof is a fingerprint stability bound (App. V) guaranteeing robustness, not predictive accuracy.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance and for coding assistance. For writing, the tools were used to copy-edit, tighten, and reformat author-written prose and LaTeX tables, from author-supplied drafts, results and notes. Every claim, number, and citation in the resulting text was written or verified by the authors. For code, the tools were used to draft and refactor parts of the experimental pipeline (data loading, evaluation harnesses, plotting); all such code was read, executed, and verified by at least one author, and every reported result was produced by a pipeline run the authors inspected end to end.

We also used generative AI tools for retrieval and discovery, including finding and summarizing related work, and for research ideation and execution, including feedback on research methodology and on the design of the experiments and ablations. Every reference surfaced this way was retrieved, read, and checked against its primary source by an author before citation, and every suggested experiment was specified, run, and interpreted by the authors.

We used generative AI tools to assist in the writing of the theoretical result (Theorem V.1 and its proof in Appendix V); the statement, its assumptions, and the proof were checked by the authors. We did not use generative AI to generate synthetic data, and no reported number, table, or figure in this paper was produced by a generative AI tool rather than by an experimental run. Generative AI is not a component of the TopoPrimer method itself.

We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops a feature-construction method for time series forecasting and raises no concerns under the ICLR Code of Ethics. No human subjects are involved and no crowdsourcing was used. The experiments use observational time series only — electricity load, weather station readings, retail unit sales, and traffic sensor speeds — with no personal or sensitive attributes collected, inferred, or released. The electricity and retail data are aggregated at the client, store, and item level, not the individual level.

All four public benchmarks are used in accordance with their published terms and are credited with license details in Appendix Y: ECL (UCI, CC BY 4.0), Monash Weather (Godahewa et al., 2021; CC BY 4.0), M5 (Kaggle competition terms), and METR-LA (Li et al., 2018; MIT). Backbones and baselines — Chronos, TimesFM, XGBoost, PatchTST (Apache 2.0), DLinear and NLinear (MIT) — are likewise used under their licenses. The internal corpus is proprietary; results derived from it are reported in aggregate and the underlying data is not released.

The submission releases no datasets, pre-trained models, or other artifacts with misuse potential, so no release safeguards are needed. The method operates solely on observational time series and we identify no dual-use risk specific to it. We claim no societal benefit beyond forecast accuracy itself, and identify no negative societal impact specific to the method.

REPRODUCIBILITY STATEMENT

All architectural and pipeline details needed to reimplement TopoPrimer are in the paper. The persistent-homology fingerprint (Vietoris–Rips filtration to dimension 2, k=50 kNN sparsification of the correlation-distance matrix, 125-dimensional persistence landscape split H_0:50 / H_1:50 / H_2:25) is specified in Section 3 and Appendix K. The spectral relational coordinate (block-wise truncated SVD of the entity–time matrix, row-normalized, zero-padded to 256 dimensions, no training and no graph) in Section 3 and Appendix H. Backbone hyperparameters (d_{model}=256, 6 pre-norm layers, 8 heads, dropout 0.10, Huber quantile loss over 9 quantiles; AdamW with β_1=0.9, β_2=0.999, weight decay 10^{-3}, lr 10^{-4} for Transformer variants and 3×10^{-4} for adapters) and per-dataset context and horizon settings are tabulated in Appendix J. Adapter branch widths and measured trainable-parameter counts in Appendices K and D. Train/validation/test splits and descriptor cutoffs for the four public benchmarks are tabulated in Appendix I, which also states the descriptor cutoff for the internal corpus; the internal evaluation protocol, which uses rolling forecast origins rather than a single split boundary, is given in Appendix B. Every trained variant in Table 1 is run at 5 seeds and reported

as mean±standard deviation. Appendix C adds paired cluster-bootstrap 95% confidence intervals (B=5000, resampled by client or station) and Wilcoxon signed-rank tests. Compute requirements are reported in Appendix L.

Two limits on reproducibility are stated explicitly. First, code is not released with this submission; the four public datasets (ECL, Monash Weather, M5, METR-LA) are publicly available and cited with their original sources and licenses (Appendix Y), so the public results are reimplementable from the description above. Second, the hard-regime results — peak-demand and cold-start (Section 4.3, Appendix B) — use a proprietary internal corpus that cannot be released and therefore cannot be externally reproduced. The public M5 cold-start replication in Appendix T, which is reproducible and reports a null, is included for exactly this reason.

REFERENCES

A. F. Ansari et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

C. Bodnar, F. Di Giovanni, B. Chamberlain, P. Lio, and M. Bronstein. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in gnns. In *NeurIPS*, 2022.

P. Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16(1):77-102, 2015.

G. Carlsson. Topology and Data. *Bulletin of the American Mathematical Society*, 46(2):255-308, 2009.

F. Chazal, V. De Silva, and S. Oudot. Persistence stability for geometric complexes. *Geometriae Dedicata*, 173(1):193-214, 2014.

F. Chazal, C. Levrard, and M. Royer. Topological analysis for detecting anomalies in dependent sequences: Application to time series. *Journal of Machine Learning Research*, 25(365):1-49, 2024.

T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *KDD*, 2016.

Y. Chen et al. Z-GCNETs: Time zigzags at graph convolutional networks for time series forecasting. In *International Conference on Machine Learning (ICML)*, 2021.

D. Cohen-Steiner, H. Edelsbrunner, and D. Morozov. Vines and vineyards by updating persistence in linear time. In *Association for Computing Machinery*, 2006.

J. Curry. *Sheaves, cosheaves and applications*. PhD thesis, University of Pennsylvania, 2014.

A. Das et al. A decoder-only foundation model for time-series forecasting. In *ICML*, 2024.

A. Das et al. In-context fine-tuning for time-series foundation models. In *ICML*, 2025.

V. P. Dwivedi and X. Bresson. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*, 2020.

H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010.

V. Ekambaram, A. Jati, P. Dayama, S. Mukherjee, N. H. Nguyen, W. M. Gifford, C. Reddy, and J. Kalagnanam. Tiny time mixers (TTMs): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. In *NeurIPS*, volume 37, pages 74147–74181, 2024.

S. Gao, T. Koker, O. Queen, T. Hartvigsen, T. Tsiligkaridis, and M. Zitnik. UniTS: A unified multi-task time series model. In *NeurIPS*, volume 37, pages 140589–140631, 2024.

J. Hansen and R. Ghrist. Opinion dynamics on discourse sheaves. *SIAM Journal on Applied Mathematics*, 81(5):2033-2060, 2021.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

J. Jiang, C. Han, W. X. Zhao, and J. Wang. Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In *Proceedings of the AAAI conference on artificial intelligence*, 2023.

W. Kim and F. Mémoli. Spatiotemporal persistent homology for dynamic metric spaces. *Discrete & Computational Geometry*, 66:831-875, 2021.

N. Kim, H. Baik, and Y. Yoon. Topological contrastive learning for time series. *IEEE Transactions on Neural Networks and Learning Systems*, 2026.

Y. Li et al. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *ICLR*, 2018.

B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748-1764, 2021.

Z. Lin, N. F. S. Zulkepli, M. S. M. Kasihmuddin, and R. U. Gobithaasan. CrossTopoNet: A cross-attention framework on topological latent feature space for time-series forecasting. *Knowledge-Based Systems*, 2026.

Z. Lin et al. Time-series forecasting via topological information supervised framework with efficient topological feature learning. *Expert Systems with Applications*, 130032, 2025.

H. Liu et al. Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, p.4125-4129, 2023.

Y. Liu et al. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *International Conference on Learning Representations (ICLR)*, 2024.

A. Mostafa, R. Younis, and Z. Ahmadi. Dynamic Sheaf Diffusion Networks with Adaptive Local Structure for Heterogeneous Spatio-Temporal Graph Learning. *arXiv preprint arXiv:2604.11275*, 2026.

Y. Nie et al. A time series is worth 64 words: Long-term forecasting with transformers. In *ICLR*, 2023.

T. Papamarkou, T. Birdal, M. M. Bronstein, G. Carlsson, J. Curry, Y. Gao, M. Hajij, R. Kwitt, P. Lio, et al. Position: Topological deep learning is the new frontier for relational learning. In *Proceedings of machine learning research*, 235, p.39529, 2024.

J. A. Perea and J. Harer. Sliding windows and persistence: An application of topological methods to signal analysis. *Foundations of Computational Mathematics*, 15(3):799-838, 2015.

C. Tralie, N. Saul, and R. Bar-On. Ripser.py: A Lean Persistent Homology Library for Python. *Journal of Open Source Software*, 3(29):925, 2018.

S. Tymochko, E. Munch, and F. A. Khasawneh. Using zigzag persistent homology to detect Hopf bifurcations in dynamical systems. *Algorithms*, 13(11):278, 2020.

Y. Wang, A. Smola, D. Maddix, J. Gasthaus, D. Foster, and T. Januschowski. Deep factors for forecasting. In *International Conference on Machine Learning (ICML)*, 2019.

G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*, 2024.

Z. Wu et al. Graph wavenet for deep spatial-temporal graph modeling. In *IJCAI*, 2019.

Z. Wu et al. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020.

H. Wu, J. Xu, J. Wang, and M. Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in neural information processing systems*, volume 34, 2021.

S. Zeng, F. Graf, C. Hofer, and R. Kwitt. Topological attention for time series forecasting. In *Advances in neural information processing systems*, volume 34, 2021.

A. Zeng, M. Chen, L. Zhang, and Q. Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, 2023.

S. Zeng, F. Graf, M. Uray, S. Huber, and R. Kwitt. Neural persistence dynamics. In *Advances in Neural Information Processing Systems*, 2024.

H. Zhou et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, 2021.

A Architecture Diagrams

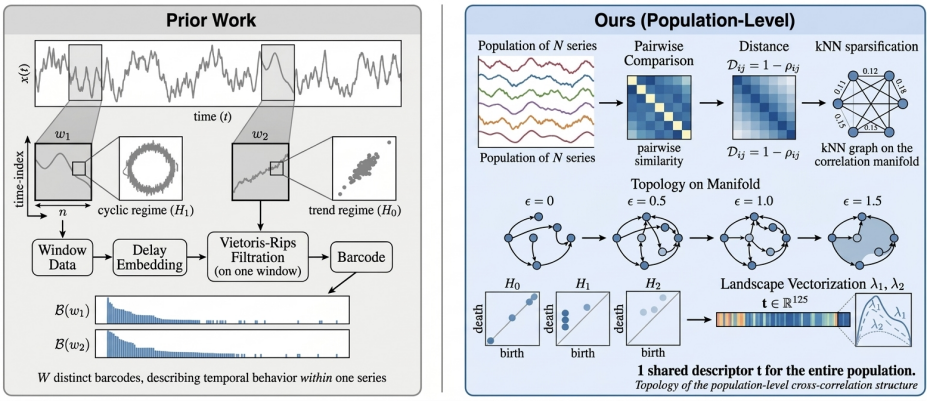

Figure 5: **Per-series vs. population-level TDA.** Prior work (left) computes a separate vector per sliding window, yielding independent descriptors of the temporal dynamics *within* a single series. TopoPrimer (right) treats the entire population as a single point cloud. It runs one Vietoris-Rips filtration on the cross-series correlation manifold to produce a global 125-dimensional persistence landscape shared by all series. This allows the model to condition on macroscopic structure that no individual trajectory contains.

Figure 5 illustrates our core methodological departure: rather than extracting isolated per-series features, TopoPrimer applies a single topological filtration to the full population manifold. Complementing this global descriptor, Figure 6 details the spectral relational pipeline, which derives a 256-dimensional per-series coordinate via a truncated SVD of the entity-time matrix.

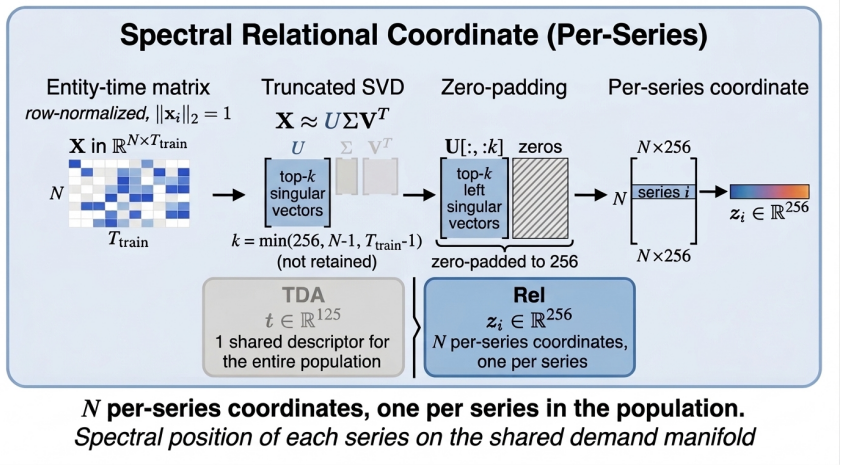

Figure 6: **Spectral relational coordinate computation (per-series).** The entity-time training matrix $\mathbf{X} \in \mathbb{R}^{N \times T}$ is formed directly from the observed series and row-normalized to unit ℓ_2 norm; no graph, correlation matrix or Laplacian is constructed on this path. A truncated SVD $\mathbf{X} \approx U\Sigma V^\top$ extracts the top-k left singular vectors; row i of U is the 256-dimensional spectral coordinate $\mathbf{z}_i \in \mathbb{R}^{256}$ for series i. This yields N per-series coordinates, one per series in the population, encoding each series' structural position in the shared demand manifold. In contrast to the TDA fingerprint $\mathbf{t} \in \mathbb{R}^{125}$ (one shared vector for the entire population), the relational coordinate is per-series and relational. The kNN correlation graph is used only by the trained neural variant of Appendix H, not by the spectral default reported here.

B Internal Corpus: Cold-Start and Seasonality MAE Tables

Section 4.3 introduces the cold-start and seasonality evaluations on our internal corpus. While the main text reports summary statistics and visualizes the trends (Figure 4), this appendix details the evaluation protocol and provides the complete per-week MAE results: Table 3 for the cold-start cohort and Table 4 for the seasonal-spike window. Note that these evaluations isolate the entity-manifold fingerprint (+TDA); the relational coordinate is inactive in this setting.

Evaluation protocol. We treat each week as an independent forecast origin: models observe history up to the cutoff and predict the subsequent four weeks. We report MAE averaged strictly over weeks with non-zero demand. Because the median weekly demand at the entity–item level is zero, including empty weeks artificially compresses the error of all methods toward a common floor, reflecting corpus sparsity rather than true forecast quality. To ensure fair comparisons, every model is scored on an identical set of series at each origin (the intersection of series all methods are capable of forecasting), preventing coverage discrepancies from confounding the results. The exact number of evaluated series, N, is reported in each table. Both internal-corpus evaluations use a single training seed (42); we did not run a seed sweep for this configuration. The reported percentages are therefore measured differences from one run rather than estimates with a seed band, and we note that the gap at launch is stable across three new-item cohort definitions (13.1–13.3%).

Cold-start cohort. We define "new items" as those whose first non-zero sale occurs within six weeks of the forecast origin (yielding 44 to 56 distinct items per window). At week w_0, these items enter with an entirely zero-filled context window, meaning no variant has access to historical item-level data. Consequently, the population-level topological fingerprint serves as the sole signal distinguishing the +TDA variant from the vanilla Transformer at launch. Because this fingerprint is computed offline from an earlier population snapshot, it is fully available prior to the new item's introduction. To strictly isolate the effect of the descriptor, both variants are trained under perfectly matched conditions, sharing identical data, random seeds, learning schedules, and epoch budgets; they differ only by the injection of the 125-dimensional fingerprint.

Table 3: Cold-start MAE vs. weeks of post-launch history (new items only). Identical series set across all models; non-zero-demand weeks. Lower is better; **bold** = best in column.

Model	Week 0	Week 1	Week 2	Week 3	Week 4	Week 5
Transformer variants (fully-trained)						
Transformer	2.558	2.081	1.778	1.742	1.684	1.651
Transformer + TDA	2.219	**1.794**	**1.506**	**1.414**	**1.392**	**1.349**
Classical baselines						
Rate-based	2.762	2.285	1.687	1.800	1.860	1.913
DLinear	2.774	2.282	1.988	1.879	1.869	1.800
NLinear	2.497	1.996	2.016	1.902	1.703	1.672
XGBoost	3.158	2.681	2.337	2.227	2.145	2.094
Zero-shot TSFMs						
Chronos	**1.868**	1.825	1.783	1.742	1.520	1.446
TimesFM	2.631	2.042	1.713	1.733	1.654	1.665
Series evaluated (N)	7,719	9,947	11,298	11,852	11,823	12,872

Table 4: Seasonality-spike MAE over the peak-demand window, evaluated on all items. To ensure fair comparison, metrics are computed over an identical set of series across all models, restricted strictly to weeks with non-zero demand, at a single training seed (42). Baseline implementations follow Zeng et al. (2023) for DLinear and NLinear, and Chen and Guestrin (2016) for XGBoost. Lower is better; **bold** indicates the best result per column.

Model	**Week 0**	**Week 1**	**Week 2**	**Week 3**	**Week 4**	**Week 5**
Transformer variants (fully-trained)						
Transformer	1.622	1.645	2.173	2.410	2.518	3.022
Transformer + TDA	1.465	1.462	1.857	2.117	2.299	**2.695**
Classical baselines						
Rate-based	1.649	1.730	2.042	2.249	2.437	2.911
DLinear	1.732	1.731	2.190	2.429	2.586	3.107
NLinear	1.631	1.625	2.012	2.200	2.344	2.794
XGBoost	1.833	1.865	2.376	2.611	2.815	3.419
Zero-shot TSFMs						
Chronos	**1.456**	**1.430**	**1.854**	**2.050**	**2.221**	2.782
TimesFM	1.534	1.569	2.083	2.301	2.505	2.983
Series evaluated (N)	21,154	20,123	21,350	20,602	19,823	14,362

C Statistical Significance and the Isolated Relational-Only Ablation

Table 1 reports 5-seed means; stds in Table 6. To provide a second, independent measure of uncertainty, this appendix details the test-set paired cluster-bootstrap 95% confidence intervals and Wilcoxon signed-rank tests for the headline ECL and Monash Weather comparisons (summarized in Table 2 of the main text). Section C.1 first unpacks the main-table MAE columns by dataset and backbone, allowing the individual contribution of each topological component to be read cell by cell.

C.1 Full Ablation Ladder at 5 Seeds

Table 5 restructures the MAE columns from Table 1 to explicitly isolate the standalone contribution of the relational coordinate across every experimental setting evaluated in this paper.

Table 5: **Full ablation results (5-seed mean±std MAE)** across all four public benchmarks and three backbone families. **Bold** indicates the best result per row. †Weather is scored at the day-30 lead time; the full 30-day block mean is provided in Table 24.

Dataset	Backbone	Vanilla	+TDA only	+Rel only	+TDA+Rel
ECL	Transformer	$\mathbf{0.1956 \pm .0044}$	$0.1969 \pm .0051$	$0.2006 \pm .0029$	$0.1964 \pm .0010$
ECL	Chronos	$0.3019 \pm .0002$	$0.3018 \pm .0001$	$0.2899 \pm .0003$	$\mathbf{0.2897 \pm .0002}$
ECL	TimesFM	$0.3015 \pm .0001$	$0.3013 \pm .0001$	$\mathbf{0.2893 \pm .0005}$	$\mathbf{0.2893 \pm .0001}$
Monash Weather†	Transformer	$2.1765 \pm .0710$	$2.2071 \pm .0442$	$2.0494 \pm .0737$	$\mathbf{2.0489 \pm .0283}$
Monash Weather†	Chronos	$2.0156 \pm .0098$	$2.0135 \pm .0319$	$\mathbf{1.9880 \pm .0279}$	$1.9960 \pm .0242$
Monash Weather†	TimesFM	$2.0363 \pm .0120$	$\mathbf{2.0355 \pm .0047}$	$2.0582 \pm .0163$	$2.0385 \pm .0374$
METR-LA	Transformer	$2.2938 \pm .0299$	$2.2540 \pm .0546$	$2.2641 \pm .0406$	$\mathbf{2.2402 \pm .0294}$
METR-LA	Chronos	$2.3755 \pm .0329$	$2.4241 \pm .1427$	$2.3454 \pm .0161$	$\mathbf{2.3371 \pm .0149}$
METR-LA	TimesFM	$2.3820 \pm .0221$	$2.3872 \pm .0235$	$\mathbf{2.3809 \pm .0102}$	$2.3849 \pm .0351$
M5 Household (neg. control)	Transformer	$1.8732 \pm .0269$	$\mathbf{1.8674 \pm .0112}$	$1.8843 \pm .0122$	$1.8925 \pm .0172$
M5 Household (neg. control)	Chronos	$\mathbf{0.7965 \pm .0014}$	$0.7977 \pm .0021$	$0.7966 \pm .0020$	$0.7973 \pm .0023$
M5 Household (neg. control)	TimesFM	$0.7959 \pm .0003$	$\mathbf{0.7955 \pm .0010}$	$0.7971 \pm .0042^a$	$0.7957 \pm .0005$

[a] 4 of 5 seeds; the fifth seed's evaluation output is missing.

The isolated +Rel variant accounts for nearly all of the combined +TDA+Rel performance gain in settings where topology is beneficial (e.g., ECL with Chronos $0.3019 \rightarrow 0.2899$; Monash Weather with the fully-trained Transformer $2.1765 \rightarrow 2.0494$). Conversely, +TDA alone yields negligible improvements and even mildly *degrades* performance on Monash Weather ($2.1765 \rightarrow 2.2071$). On M5 (our negative control, where the fingerprint is near-constant), the null result holds component-by-component across all three backbones. This aligns with our conclusion in Section 4: the relational coordinate drives within-dataset accuracy, whereas the TDA fingerprint is not a direct performance booster on single-population benchmarks. Its diagnostic use is as an offline screen over cohorts, validated on internal cohorts rather than across these public domains (Appendix N).

C.2 Test-Set Bootstrap Confidence Intervals

While training-seed variance (Table 5) captures one source of uncertainty, the finite size of the test sets (321 clients for ECL, 3,010 stations for Monash Weather) introduces another. To account for this, we additionally report paired cluster-bootstrap 95% BCa confidence intervals (B=5000, resampled by client/station, scored on the forecast averaged over the 5 training seeds) and Wilcoxon signed-rank tests for the two headline comparisons. These intervals are summarized in the main text (Table 2, Section 4). This subsection details how to interpret them and discusses the counter-cases.

Because these intervals resample entities at a fixed forecast origin using the seed-averaged forecast, they quantify cross-entity variability alone. Seed variance is absorbed by the averaging (reported separately in Table 5), and variability across forecast origins is not captured. Given the single-origin protocol detailed in Appendix I, temporal sampling variance remains unmeasured, meaning these intervals should be interpreted as conditional on that specific origin.

Table 6: **Detailed forecasting performance with per-seed standard deviations** (expanding Table 1). Metrics represent the mean±std over 5 seeds. **Bold** indicates the best result per column within a backbone block (↓ lower is better). Zero-shot baselines are deterministic single runs and are excluded from bolding. Secondary metrics include MAPE (METR-LA, $H{=}15$ min), MSE (ECL, $H{=}96$ h), and WAPE (M5, $H{=}4$ wk). †Weather is scored at a single lead time (day 30, the final step of the 30-day forecast); see Table 1 and App. Q. The M5 columns report the HOUSEHOLD category alone. Because M5 units differ between the Transformer and adapter blocks, comparisons should only be made within blocks. [a]Computed over 4 of 5 seeds.

	METR-LA		ECL		Monash Weather		M5 (Household)	
Model	MAE↓	MAPE↓	MAE↓	MSE↓	MAE† ↓	MSE† ↓	MAE↓	WAPE↓
Transformer variants								
Transformer	2.2938 ± 0.0299	3.9341 ± 0.0562	**0.1956 ± 0.0044**	0.0959 ± 0.0080	2.1765 ± 0.0710	25.912 ± 0.560	1.8732 ± 0.0269	0.2653 ± 0.0038
Transformer + TDA	2.2540 ± 0.0546	3.8769 ± 0.0919	0.1969 ± 0.0051	0.0971 ± 0.0074	2.2071 ± 0.0442	26.396 ± 0.287	**1.8674 ± 0.0112**	**0.2645 ± 0.0016**
Transformer + Rel	2.2641 ± 0.0406	3.8955 ± 0.0626	0.2006 ± 0.0029	0.1028 ± 0.0090	2.0494 ± 0.0737	25.285 ± 0.289	1.8843 ± 0.0122	0.2669 ± 0.0017
Transformer + TDA + Rel	**2.2402 ± 0.0294**	**3.8563 ± 0.0469**	0.1964 ± 0.0010	**0.0942 ± 0.0029**	**2.0489 ± 0.0283**	**25.107 ± 0.323**	1.8925 ± 0.0172	0.2681 ± 0.0024
Chronos 2.0 variants								
Chronos Zero-Shot	3.3482	5.6146	0.586	0.610	2.3439	29.776	0.9132	0.9816
Chronos Vanilla Adapter	2.3755 ± 0.0329	4.0606 ± 0.0455	0.3019 ± 0.0002	0.2047 ± 0.0002	2.0156 ± 0.0098	28.283 ± 0.085	**0.7965 ± 0.0014**	**0.8328 ± 0.0015**
Chronos + TDA	2.4241 ± 0.1427	4.1489 ± 0.2521	0.3018 ± 0.0001	0.2047 ± 0.0001	2.0135 ± 0.0319	28.319 ± 0.240	0.7977 ± 0.0021	0.8340 ± 0.0022
Chronos + Rel	2.3454 ± 0.0161	4.0085 ± 0.0235	0.2899 ± 0.0003	0.1900 ± 0.0004	**1.9880 ± 0.0279**	**28.120 ± 0.175**	0.7966 ± 0.0020	**0.8328 ± 0.0021**
Chronos + TDA + Rel	**2.3371 ± 0.0149**	**4.0028 ± 0.0297**	**0.2897 ± 0.0002**	**0.1895 ± 0.0002**	1.9960 ± 0.0242	28.155 ± 0.143	0.7973 ± 0.0023	0.8335 ± 0.0024
TimesFM 2.5 variants								
TimesFM Zero-Shot	2.4406	4.1996	0.580	0.602	2.0316	28.032	0.9401	0.9828
TimesFM Vanilla Adapter	2.3820 ± 0.0221	**4.0971 ± 0.0285**	0.3015 ± 0.0001	0.2052 ± 0.0001	2.0363 ± 0.0120	28.126 ± 0.098	0.7959 ± 0.0003	0.8321 ± 0.0004
TimesFM + TDA	2.3872 ± 0.0235	4.1039 ± 0.0328	0.3013 ± 0.0001	0.2050 ± 0.0001	**2.0355 ± 0.0047**	28.094 ± 0.051	**0.7955 ± 0.0010**	**0.8317 ± 0.0011**
TimesFM + Rel	**2.3809 ± 0.0102**	4.1021 ± 0.0144	**0.2893 ± 0.0005**	0.1903 ± 0.0006	2.0582 ± 0.0163	28.088 ± 0.045	0.7971 ± 0.0042[a]	0.8334 ± 0.0044[a]
TimesFM + TDA + Rel	2.3849 ± 0.0351	4.1046 ± 0.0459	**0.2893 ± 0.0001**	**0.1900 ± 0.0001**	2.0385 ± 0.0374	**28.000 ± 0.066**	0.7957 ± 0.0005	0.8319 ± 0.0005

How to read the intervals. We compute $\widehat{\Delta\text{MAE}}$ per cluster at full precision before averaging. Because the absolute Vanilla and Topo columns are rounded independently, their visible difference may not perfectly match the reported Δ. For ECL with Chronos, both relational-bearing variants place $\widehat{\Delta\text{MAE}}$ at 3.5 to 3.7 times the half-width of its own 95% CI, and improve predictions for $\approx$81% of clients. On Monash Weather (fully-trained Transformer, day-30 lead time), the absolute columns are scored on the seed-averaged forecast. Consequently, they sit 0.025 to 0.029 lower than the mean-of-seeds values in Table 1, as ensembling five forecasts naturally reduces absolute error. Because the offset applies to every row, the paired Δ is affected only by the spread between rows (0.0248 to 0.0290 here), which is why the Weather Δ values differ from the Table 1 seed-mean differences by up to 0.004. In this setting, +Rel reaches 6.5 times that half-width, marking the strongest statistically backed result in this paper. While the ordering of +Rel versus +TDA+Rel on Monash Weather differs between Table 2 and Table 1, the performance margins (0.0021 and 0.0005) fall well within the training-seed noise ($\sigma \approx 0.03$ to 0.07). The two variants are effectively tied, and their exact ranking should not be over-interpreted.

The counter-case. The fully-trained Transformer on ECL presents a counter-case: the +Rel-alone variant performs significantly *worse* than the vanilla baseline ($\widehat{\Delta\text{MAE}} = +0.0044$, 95% CI $[+0.0018, +0.0077]$). Although this specific comparison is not tabulated in Table 2 (which highlights ECL on the Chronos backbone), it aligns in both sign and magnitude with the $0.1956 \rightarrow 0.2006$ seed means from Table 1. As noted earlier, the paired estimate is computed per client at full precision, explaining the slight deviation from the 0.0050 difference between the independently rounded seed means. We explicitly report this counter-case alongside the successful results. Appendices E and F demonstrate that this is the exact same setting where the plain and per-series controls also outperform +Rel.

Ultimately, both training-seed variance and test-set sampling variance corroborate the core pattern reported in the main text. The relational coordinate produces a statistically significant gain in both settings we test with paired intervals (Chronos on ECL, Transformer on Monash Weather). Conversely, the TDA fingerprint alone fails to clear either uncertainty bar on its own. Finally, on the M5 negative control, every variant remains within the training-seed noise floor (gain/σ<1.2, p>0.05 for every cell in Table 5) across both deployment paradigms. This measured seed variance is precisely what allows us to confidently classify M5 as a true null result rather than an unlucky initialization.

D BACKBONE-ADAPTATION COMPETITORS

To ensure a rigorous comparison, all evaluated methods share the same Chronos-2 backbone, data splits, evaluation code, and five training seeds (Zero-Shot uses the frozen backbone directly). Unlike LoRA, partial fine-tuning, and full fine-tuning, our adapter never backpropagates through the foundation model; it trains exclusively on cached base forecasts. Table 7 summarizes this performance ladder. The remainder of this appendix details the evaluation protocol and the exact parameter accounting.

Table 7: **Parameter-matched adaptation ladder for Chronos-2.** Metrics represent MAE mean±std over 5 seeds (Zero-Shot is a deterministic single run). Standard deviations are rounded to three decimal places; the ECL Partial FT and Vanilla seed spreads are below 0.0005. The ECL column scores forecasts without a non-negativity clamp ($\hat{y} = f(x)$), whereas Table 1 applies one ($\hat{y} = \max(0, f(x))$). Both operate on the same z-normalized scale, making the clamp inappropriate here since roughly half the z-normalized ECL targets are negative (detailed below). Every ECL row is scored unclamped, while Weather values match the Table 1 basis. **Bold** indicates the best result among methods training under 1% of the backbone (full fine-tuning is a 100%-parameter reference and is excluded from bolding). The *Ours* row represents the fused ECL adapter from Appendix K, which concatenates the four inputs into a two-layer MLP of width 256, uses $[\text{mean}, \text{std}, \max, p_{75}]$ statistics, and employs a median-only head. The $182{,}880$ parameter count corresponds specifically to this variant. Parameter counts correspond to the ECL configuration; the Weather column uses the four-branch adapter with a different count.

Method	Trainable params	ECL MAE	Weather MAE
Ours (+Rel, frozen backbone)	182,880 (0.64%)	$\mathbf{0.257 \pm 0.001}$	1.988 ± 0.028
Vanilla adapter (frozen backbone)	117,344 (0.41%)	0.262 ± 0.000	2.016 ± 0.010
LoRA (r=2)	276,480 (0.96%)	0.268 ± 0.004	$\mathbf{1.944 \pm 0.027}$
Full fine-tune (reference)	28.7M (100%)	0.269 ± 0.003	1.923 ± 0.032
Partial FT (parameter-matched)	181,248 (0.63%)	0.292 ± 0.000	2.164 ± 0.031
Zero-Shot (in-context)	0 (0%)	0.302	2.344

ECL scoring: clamped vs. unclamped. Because ECL series are z-normalized individually before both training and evaluation, roughly half of all target values are negative. The backbone's default decoding applies a non-negativity clamp to its point forecast. While appropriate for raw-count data, this clamp fails on a z-normalized scale by truncating exactly the negative half of the prediction range, which disproportionately inflates the MAE for the least accurate forecasts. Letting $f(x)$ denote the backbone's raw median forecast, the two evaluation bases are:

$$\text{Table 1: } \hat{y} = \max\big(0,\, f(x)\big), \qquad \text{Table 7: } \hat{y} = f(x).$$

All other evaluation details remain identical, including the series, splits, horizon (H=96), z-normalization constants, and MAE definition. Every ECL row in Table 7 (adapter, Vanilla, LoRA, full fine-tune, partial fine-tune, and Zero-Shot) is scored unclamped. This explains why the identical +Rel configuration yields an MAE of 0.2899 in Table 1 but 0.2573 here, and why the Zero-Shot error drops from 0.5860 to 0.302: the weaker the raw forecast, the more probability mass the clamp incorrectly truncates.

Partial fine-tune parameter selection. To construct the parameter-matched partial fine-tuning baseline, we initially freeze every backbone parameter. We then unfreeze them in reverse registration order (starting from the output side of the network) until the trainable parameter count reaches the adapter's budget of $181{,}248$. This approach provides a standard last-layers partial fine-tuning baseline. Because no submodule is chosen by hand and the unfreezing rule is deterministic given the model architecture, the resulting comparison is a strict parameter-budget match rather than an optimized selection of weights. The resulting checkpoint is a standard backbone `state_dict`, evaluated through the exact same pipeline as the full fine-tune.

On ECL, the frozen adapter outperforms LoRA, full fine-tuning, and the parameter-matched partial fine-tune. Notably, the topology-free Vanilla adapter (0.262) already surpasses LoRA and full fine-tuning on its own, while adding +Rel refines this baseline by a further 0.005 MAE (0.2620 ± 0.0002 vs. 0.2573 ± 0.0005 over 5 seeds). On Monash Weather, among methods training under 1% of the

backbone parameters, our adapter trails only LoRA (by 0.044 MAE). We interpret LoRA's slight edge here as a successful calibration to Weather's specific seasonality via direct weight updates, a mechanism that is complementary to our closed-form relational prior rather than a direct substitute.

Vanilla adapter parameter count. The topology-free Vanilla adapter, used as the baseline throughout our main results, trains $117{,}344$ parameters (0.41%) on the Chronos-2 ECL configuration. This compares to $182{,}880$ for +Rel and $214{,}880$ for +TDA+Rel, with the increments arising from the 256×256 and 125×256 projections in the fused MLP's first layer. In the fused architecture, an inactive input is entirely removed from the concatenation rather than zero-padded. As a result, the trainable parameter counts differ by variant, even though the encoder and output-head widths remain fixed. In contrast, the four-branch architecture zero-fills inactive inputs, ensuring that the Vanilla and +TDA variants are strictly parameter-matched.

E Plain Covariance and Correlation Controls

To determine whether our structured topological features provide an advantage over unstructured population statistics, we inject plain controls into the identical architectural slots, matching both dimensionality and seed count. Specifically, `CovEig` extracts the top-125 eigenvalues from the same train-window correlation matrix used to construct the TDA fingerprint, bypassing the persistent-homology step entirely. `CorrRow` replaces the relational coordinate by projecting each series' own raw correlation row to 256 dimensions. Table 8 summarizes the frozen-adapter comparison. The remainder of this appendix discusses the fully-trained Transformer counterpart, where these unstructured controls match or outperform the topological features. Note that `CovEig`, like the fingerprint, is one vector broadcast to every series, so on these single-population benchmarks both are absorbed into a layer bias (App. M); the `CovEig`-vs-+TDA contrasts below therefore reflect optimization noise rather than a topological deficit. The `CorrRow`-vs-+Rel contrast is unaffected.

Table 8: **Ablation of structured vs. plain controls** using the frozen Chronos-2 adapter. Metrics are MAE (mean $\pm$ std over 5 training seeds). Parentheses report the gain over the Vanilla baseline divided by the seed standard deviation (gain/σ), computed at full precision and therefore not reproducible from the four-decimal values displayed; followed by a Wilcoxon signed-rank p-value versus Vanilla computed over test clusters (clients for ECL, stations for Weather) at a single training seed (42), not across seeds. Plain controls are injected into the identical architectural slot and dimension as the structured topological features they replace. **Bold** indicates the best result per column.

Variant	ECL MAE	Weather MAE
Vanilla	0.3019 ± 0.0002	2.0156 ± 0.0098
+TDA (real)	0.3018 ± 0.0001 (0.40, .42)	2.0135 ± 0.0319 (0.07, .88)
+CovEig (plain)	0.3023 ± 0.0002 (-1.34, .04)	2.0273 ± 0.0150 (-0.81, .15)
+Rel (real)	$\mathbf{0.2899 \pm 0.0003}$ (29.67, .00)	1.9880 ± 0.0279 (0.78, .16)
+CorrRow (plain)	0.2903 ± 0.0002 (39.87, .00)	$\mathbf{1.9859 \pm 0.0159}$ (1.68, .02)

In a direct comparison on the frozen adapter, +Rel outperforms +CorrRow on ECL (p=0.018; Wilcoxon signed-rank paired over test clusters, not over training seeds), while the two are statistically tied on Monash Weather. This pattern does not hold for the fully-trained Transformer: there, `CorrRow` matches or outperforms +Rel on both ECL ($+0.006$) and Weather ($+0.022$), pulling ahead in 4 of 5 training seeds (Table 9). The +TDA-vs-`CovEig` contrasts run in both directions (Weather $+0.086$ favoring `CovEig` on the Transformer; the ECL contrast is within seed noise), but per the scoping note above both compare two constant inputs and reflect optimization noise rather than a topological deficit. These reported deltas represent the 5-seed mean MAE differences based on the evaluation format of Table 1 (ECL over its 96-step block, and Weather at the day-30 lead time).

We offer the following interpretation for this discrepancy, noting that our experiments do not isolate the exact mechanism. We can confidently rule out an architectural explanation: the from-scratch encoder attends over the temporal tokens $t = 1, \ldots, L$ of a single series, exactly as the frozen backbones do, meaning it is not natively cross-variate in the manner of iTransformer or Crossformer. What distinguishes the fully-trained model from the frozen setting is its trainable capacity fitted directly to the population. Specifically, it possesses weights learned on the full panel alongside a 256-

Table 9: **Plain controls on the fully-trained Transformer** (MAE, mean±std over 5 training seeds; ECL over its 96-step block, Weather at the day-30 lead, matching Table 1). No p-values are reported: the available significance runs score a single training seed, and seed 42 carries the opposite sign for Weather `CorrRow`-vs-+Rel; per-seed sign counts are given in the text instead. Per the scoping note above, the `CovEig` rows are expressively null on both benchmarks. We therefore do not bold, as the leading ECL cell belongs to a `CovEig` row; the leading Weather cell (`CorrRow`) is informative.

Variant	ECL MAE	Weather MAE
Vanilla	0.1956 ± 0.0044	2.1765 ± 0.0710
+TDA (real)	0.1969 ± 0.0051	2.2071 ± 0.0442
+CovEig (plain)	0.1931 ± 0.0037	2.1216 ± 0.0380
+Rel (real)	0.2006 ± 0.0029	2.0494 ± 0.0737
+CorrRow (plain)	0.1950 ± 0.0044	2.0273 ± 0.0646

dimensional lookup embedding per series (Table 14) that updates during training. This combination may provide sufficient capacity to learn a population-relative coordinate and process an unstructured correlation row internally, leaving a structured prior with little to add. In contrast, the frozen adapter contains no per-series free parameters, relying entirely on fixed backbone weights, z-scored statistics, and a cached base forecast. Consequently, the injected descriptor acts as the sole channel through which population structure can influence the forecast, explaining why the structured topological features succeed in this regime. Directly validating this hypothesis would require an embedding-table ablation on the from-scratch model, which we leave to future work.

F Sliding-Window TDA Descriptor Comparison

As a strong baseline, we evaluate a capacity- and injection-matched per-series sliding-window (Takens) persistence descriptor (utilizing a delay embedding → Vietoris–Rips → landscape pipeline, projected to 256 dimensions and injected identically). The empirical results are nuanced: across the six dataset-by-backbone configurations in Table 10, +Rel achieves the best performance in four, the sliding-window descriptor in one, and the vanilla baseline in one. Notably, the sliding-window descriptor outperforms the vanilla baseline in five of the six settings, confirming it is an active, competitive baseline rather than an inert control.

The single setting where the sliding-window descriptor wins is the ECL dataset with the from-scratch Transformer (0.1926 MAE, compared to +Rel's 0.2006 and vanilla's 0.1956). Two observations contextualize this specific result without dismissing it. First, this configuration uses a fully-trained Transformer on our smallest population benchmark by series count after METR-LA (321 clients). Crucially, this is the exact same setting where the plain `CorrRow` control also matches or outperforms +Rel (Appendix E). This aligns with our scoping claim in Section 4: a precomputed relational prior provides the most value when the backbone lacks the capacity to fit the population structure internally, rather than when it can. Second, the sliding-window descriptor's margin over vanilla (0.0030) falls entirely within one seed standard deviation (± 0.0044). Therefore, this specific configuration cleanly separates the two descriptors from each other, but it does not statistically separate either descriptor from the vanilla baseline.

Conversely, on the four frozen-backbone configurations (where the adapter processes one series at a time and a population prior is highly beneficial), the sliding-window descriptor never achieves the best performance. Instead, +Rel wins three, and the vanilla baseline wins the Weather/TimesFM configuration where all topological augmentations fall flat. Furthermore, sliding-window descriptors are fundamentally undefined in cold-start scenarios, as there is no historical window to embed. In contrast, our precomputed relational coordinate acts as a manifold position available from the very first time step, inherited from established neighbors when a newly launched series lacks its own history (Section 3, Appendix T). Table 10 details this full comparison.

Table 10: **Sliding-window TDA vs. relational coordinate.** Metrics represent test MAE averaged over 5 seeds. Seed-level variability for the Vanilla and +Rel arms is reported in Table 5. **Bold** indicates the best result per row.

Dataset	Backbone	Vanilla	Sliding-window (matched)	+Rel (ours)
ECL	Transformer	0.1956	**0.1926**	0.2006
ECL	Chronos	0.3019	0.2923	**0.2899**
ECL	TimesFM	0.3015	0.2920	**0.2893**
Weather	Transformer	2.176	2.152	**2.049**
Weather	Chronos	2.016	1.997	**1.988**
Weather	TimesFM	**2.036**	2.037	2.058

G METRIC ROBUSTNESS OF THE TDA FINGERPRINT

To ensure our results are not sensitive to the choice of distance metric, we recompute the TDA fingerprint using Spearman, Euclidean, and distance-correlation (dCor) metrics in place of the default Pearson correlation distance. As shown in Table 11, altering the metric leaves the performance of the +TDA variant entirely within the seed noise of the original Pearson result on both ECL and Monash Weather, across both the Chronos and Transformer backbones. Because both benchmarks are single-population, this tests the descriptor's stability under metric choice, not its predictive value (App. M).

Table 11: **Robustness of the +TDA variant across distance metrics.** Metrics represent MAE mean±std over 5 seeds under four different distance functions. **Bold** indicates the best-performing distance metric per row (the Vanilla baseline is provided for reference).

Dataset (Backbone)	Vanilla	Correlation	Spearman	Euclidean	dCor
ECL (Chronos)	0.3019 ± 0.0002	$\mathbf{0.3018 \pm 0.0001}$	0.3019 ± 0.0001	0.3020 ± 0.0002	0.3019 ± 0.0001
ECL (Transformer)	0.1956 ± 0.0044	0.1969 ± 0.0051	$\mathbf{0.1957 \pm 0.0017}$	0.1988 ± 0.0049	0.1974 ± 0.0030
Weather (Chronos)	2.015 ± 0.010	2.014 ± 0.032	2.007 ± 0.020	$\mathbf{2.005 \pm 0.015}$	2.008 ± 0.026
Weather (Transformer)	2.177 ± 0.071	$\mathbf{2.207 \pm 0.044}$	2.232 ± 0.039	2.207 ± 0.080	2.213 ± 0.071

H SPECTRAL VS. NEURAL SHEAF ENCODER

As described in Section 3, we evaluate two implementations of the relational coordinate. The **Spectral (default)** approach computes per-series coordinates as the leading left singular vectors of the entity-time training matrix. This is done block-wise per store×category for M5 and globally for ECL, Monash Weather, and METR-LA. We retain $k = \min(256,\ N-1,\ T-1)$ left singular vectors and zero-pad to a fixed width of 256, meaning the coordinate is rank-limited whenever the population is smaller than this nominal width. In practice, only METR-LA is affected: with N=207 sensors, 50 of its 256 coordinates are identically zero, whereas ECL, Monash Weather, and M5 supply the full 256.

Alternatively, the **Sheaf (Neural)** approach trains per-series learnable embeddings $E_i \in \mathbb{R}^{256}$ on a k=10 nearest-neighbor Pearson-correlation graph with edge weights w_{ij}. The training objective is:

$$\mathcal{L}_{\text{sheaf}} = \lambda_c \sum_{(i,j)\in\mathcal{E}} w_{ij} \left\| RE_i - RE_j \right\|^2 \ + \ \lambda_r \left\| \mathrm{dec}(E) - \mathbf{X}_{\text{node}} \right\|^2,$$

where $\lambda_c = 1.0$ and $\lambda_r = 0.5$. Here, $R \in \mathbb{R}^{128\times 256}$ is a *single* linear map (without bias and orthogonally initialized) that is shared across every node and edge. The function dec is a two-layer MLP ($256 \to 128 \to 8$) with a ReLU activation, and $\mathbf{X}_{\text{node}} \in \mathbb{R}^{N\times 8}$ holds eight per-series training statistics (mean, standard deviation, p_{25}, p_{75}, lag-1 and lag-52 autocorrelation, a seasonality ratio, and a trend slope). The embeddings are initialized from the SVD of $\mathbf{X}_{\text{node}}$ and trained for 300 epochs using AdamW (learning rate 3×10^{-3}, weight decay 10^{-4}, cosine schedule).

Because this single map is shared across all incidences rather than being parameterized per incidence, the first term does not represent a general sheaf coboundary. Specifically, setting $\mathcal{F}_{i\trianglelefteq e} = \mathcal{F}_{j\trianglelefteq e} = R$ yields $\mathrm{tr}\big((ER^\top)^\top \mathbf{L}_W (ER^\top)\big)$, which is the ordinary weighted graph Dirichlet energy of the projected embeddings. Consequently, we characterize this variant as a learned graph-Laplacian regularizer on a projected embedding space, rather than a true incidence-parameterized cellular sheaf network.

Table 12 reports the comparison on M5, which is the only dataset where both variants were fully evaluated across all backbone families. The spectral encoder consistently matches or outperforms the neural variant. We specifically highlight Chronos 2.0 on the Household category due to its item heterogeneity rather than its performance margin. There, spectral coordinates achieve an MAE of 1.0251 versus 1.0343 for the neural encoder. This represents a 0.9% relative gap, which is narrower than the Foods gap (1.5%) and therefore serves as a conservative headline figure.

We do not expect the relational coordinate to improve upon the vanilla adapter on M5, and our primary results confirm this. M5 is the right domain for this comparison because its population fingerprint is uninformative by construction: the ten store×category vectors are near-identical (mean

pairwise cosine 0.997) and are a deterministic function of a store identity the model already receives via a learned embedding (App. M). An inert fingerprint leaves the relational encoder's performance unmasked (detailed below). Its high raw H_1 count reflects shared calendar periodicity rather than cross-series structure († footnote, Table 21). In the five-seed Household evaluation (Table 1), the vanilla baseline outperforms the relational coordinate (Chronos Vanilla Adapter 0.7965, +TDA+Rel 0.7973). While the single-seed rows of Table 12 place them in the opposite order (with every topology row outperforming vanilla), we resolve this discrepancy in favor of the more robust five-seed result. Crucially, the ordering this table is designed to establish (Spectral vs. Neural) holds regardless of the baseline comparison.

The neural encoder degrades relative to the spectral approach for three compounding reasons. First, it never starts from the spectral coordinate. Because initialization relies on the SVD of the eight-dimensional node-feature matrix, only eight of the 256 embedding dimensions are non-zero at initialization; subsequent gradient descent toward graph agreement only carries the embedding further from the optimal spectral position. Second, the coboundary loss and reconstruction loss work in opposition, as pushing two correlated items toward agreement in restriction-map space inherently reduces their individual reconstruction quality. Third, while the spectral encoder extracts the dominant structural axes of entity co-variation via a closed-form truncated SVD (requiring no training), the neural encoder attempts to approximate a more expressive but empirically inferior objective.

Given these factors, we adopt spectral coordinates as the default implementation wherever the relational coordinate is active. (Note that it remains inactive throughout the internal-corpus evaluations, which rely exclusively on the population fingerprint as detailed in Section 4.3 and Appendix J). The neural encoder is not used in any of our reported benchmark results.

Table 12: **Spectral vs. neural sheaf encoder on M5 (Walmart, 28,860 evaluated series; Appendix S).** Metrics represent MAE and WAPE per M5 category. **Bold** indicates the best result per section (↓ lower is better). The Household category exhibits the most heterogeneous item demand and serves as our primary comparison point, even though its performance margin is not the largest in the table. All rows use a 52-week context window and a 4-week forecast horizon. Note that these results are on the adapter scale of Table 27, rather than the M5 block of Table 1. They are strictly comparable to each other and to the adapter section of Table 27.

	Foods		Hobbies		Household	
Model	**MAE↓**	**WAPE↓**	**MAE↓**	**WAPE↓**	**MAE↓**	**WAPE↓**
Chronos 2.0 variants (adapter)						
Chronos Vanilla Adapter	0.984	1.408	1.054	1.788	1.040	1.643
Chronos + TDA	0.982	1.405	1.052	1.786	1.039	1.641
Chronos + TDA + Rel (Spectral, default)	**0.966**	**1.383**	**1.038**	**1.761**	**1.025**	**1.618**
Chronos + TDA + Rel (Neural sheaf)	0.981	1.403	1.043	1.770	1.034	1.633

Spectral coordinates outperform the neural encoder in every cell evaluated here, covering all three categories across both metrics. The performance gap is largest on Foods (MAE 0.966 vs. 0.981) and smallest on Hobbies (1.038 vs. 1.043). Ultimately, this experiment establishes a strict ordering between the two encoders; it does not present an accuracy case for the relational coordinate on M5, where the coordinate is null by construction. Combined with the massive cost asymmetry (a single closed-form SVD versus a full pre-training run requiring three hyperparameters), this empirical superiority justifies making the spectral coordinate the default throughout this paper. The accuracy case for the coordinate itself rests entirely on the topologically non-null domains presented in Table 1.

Why M5 is the right comparison domain. We select M5 as our primary comparison domain because it is topologically null. Its H_1 generators capture shared seasonal periodicity rather than underlying relational structure, yielding a fingerprint with no usable cross-series variance (Table **??**). This absence of topological signal provides a clean, isolated testbed for the relational encoder. On datasets like ECL and Monash Weather, the simultaneous injection of TDA and relational coordinates confounds the performance contributions of each. On M5, however, the TDA fingerprint is uninformative. This forces the relational coordinate to drive any performance gains, unmasking the true quality difference between the spectral (SVD) and neural encoders. We evaluate this comparison using Chronos 2.0, as it is the backbone used to test both encoder variants. Crucially, M5 is absent

from the pretraining corpora of both foundation models: TimesFM trains on M4 and Favorita (Das et al., 2024), while Chronos 2.0 reserves M5 strictly for evaluation (Ansari et al., 2025). Accordingly, neither zero-shot model exhibits an inherent advantage on M5 (Table 1: TimesFM 0.9401, Chronos 0.9132). This allows us to read the spectral encoder's advantage directly from the results (Household MAE 1.0251 vs. 1.0343). We omit the fully-trained Transformer from this specific ablation, as it relies exclusively on spectral coordinates by design.

Computational overhead. Beyond predictive accuracy, the spectral encoder offers a massive computational advantage. It requires only a single truncated SVD of the entity-time training matrix, computed block-wise per store×category group. For the 28,695 filtered series in M5, this operation completes in under one second on a single CPU core (comprising 30 block-wise SVDs of matrices no larger than $1{,}403 \times 269$) and requires zero hyperparameter tuning. In contrast, the neural encoder demands a full pre-training pipeline. It requires warm-start initialization from the spectral coordinates, followed by multiple epochs of gradient descent on a coboundary consistency loss, early stopping on a validation set, and the tuning of three separate hyperparameters (λ_c, λ_r, β). Because the spectral approach outperforms the neural encoder on Chronos 2.0, we do not evaluate the neural encoder on TimesFM; the steep pre-training cost is simply not justified. Given its superior performance and near-zero computational overhead, the spectral encoder is the clear practical choice.

I Train/Validation/Test Splits and Descriptor Cutoffs

To prevent data leakage, all topological and relational descriptors are computed strictly from the training split, chronologically preceding both validation and test sets. For ECL, we take the additional, over-conservative precaution of computing descriptors using only the first 40% of the training split to avoid any potential within-train weekday/weekend leakage. The dataset split conventions used follow their respective originating literature. Because the four public benchmarks use a chronological split over a fixed panel of entities, every evaluated series corresponds to a row in the matrix used for descriptor computation. Entities absent from this matrix appear only in the M5 cold-start study (Appendix T).

The internal corpus follows the same strict temporal separation. Its fingerprints are built from a demand-history snapshot ending eight months prior to the training cutoff, ensuring no validation data, test data, or items launched in the cold-start window influence the descriptor. We omit the internal corpus from Table 13 because it is not evaluated at a single split boundary. Instead, each week index serves as an independent forecast origin, following the protocol detailed in Appendix B.

Table 13: Splits and descriptor cutoffs. Descriptors never see validation or test data.

Dataset	**Split (train/val/test)**	**Descriptor cutoff**
ECL	70/10/20%, chronological	first 40% of train
Monash Weather	95.5/2.25/2.25%, H=30 end-windows	full train window
METR-LA	70/10/20%, chronological	full train window
M5	rest / 4 wk (1.4%) / 4 wk (1.4%)	$\leq$ last pre-val. week; novel items after

Evaluation origins. We evaluate all trained variants and adapters at a single forecast origin per series. Specifically, the context window ends at the final time step before the split boundary, and the model predicts a single horizon block. Evaluation thus comprises exactly one forecast per series (321 ECL customers, 3,010 Monash Weather stations, 207 METR-LA sensors, and the full M5 item set), with every variant scored on identical origins and series. Because no variant sees a different origin or series, comparisons *within* our tables are exact. However, absolute values are not directly comparable to published baselines that compute metrics over rolling test splits. This distinction is particularly important for METR-LA, which rests on a much smaller sample size than the other benchmarks; a per-horizon entry averages at most 207 absolute errors—one per sensor at a single timestamp, excluding sensors reading below 1 mph. Consequently, we treat METR-LA as a directional control rather than a primary benchmark result.

J TRANSFORMER ARCHITECTURE

We evaluate four variants on the public benchmarks: Vanilla, +TDA, +Rel, and +TDA+Rel. All four share identical encoder and head architectures, differing only in their active topology blocks. For the internal corpus, we evaluate only the Vanilla and +TDA models, leaving the relational block inactive (as detailed in Section 4.3).

Global context vector. For the public benchmarks, we form a context vector by projecting a learned per-series embedding, temporal features, and any active topology blocks to $d_{\text{model}}=256$ (Table 14). For the internal corpus, we concatenate learned node, item, and category embeddings with the temporal features prior to this projection (Table 15). The TDA fingerprint serves as the primary input to the main projection matrix, $\mathbf{W}_{\text{ctx}}$. When active, the relational block instead passes through a dedicated $\text{Linear}(256\to256)$ projection, and we sum its output with the main projection before broadcasting. We explain the rationale for this separate path, and its untested status, in Appendix K.

Table 14: **Context vector components: public benchmarks.** One topology block is active for +TDA and +Rel, both for +TDA+Rel, and none for the vanilla baseline. Series counts: 321 ECL customers, 3,010 Monash Weather stations, 207 METR-LA sensors, and the M5 item set.

Feature	Dim	Notes
Series embedding	256	Learned lookup, one per series
Group embedding	256	Learned lookup; single-valued except M5
Category embedding	64	Learned lookup; single-valued except M5
Temporal features	8	Calendar encodings at one index per window; content varies by sampling frequency
Topology (stackable):		
TDA block	125	$H_0+H_1+H_2$ persistence landscape
Rel block	256	Per-series spectral relational feature

Table 15: **Context vector components: internal corpus.** Learned entity embeddings are concatenated with temporal features before the main projection. The fingerprint is the only topology block active on this corpus; the relational block is not used.

Feature	Dim	Notes
Node embedding	256	Learned lookup, one per graph node
Item embedding	256	Learned lookup, one per item type
Category embedding	64	Learned lookup, all category levels
Temporal features	8	Calendar encodings
Topology:		
TDA block	125	Entity-manifold $H_0+H_1+H_2$ landscape

Temporal features. The 8 temporal dimensions consist of calendar encodings. We compute these at every time step but sample them at a single index per window, ensuring they remain constant across the sequence. For the internal corpus, ECL, METR-LA, and Monash Weather, this index corresponds to the forecast origin (the final observed step of the lookback window). For M5, we sample at the first forecast week. The specific encodings depend on the dataset's sampling frequency. Hourly ECL uses $(\sin, \cos)$ transformations of the hour-of-day and day-of-week, normalized hour and day-of-week values, and a weekend flag. METR-LA follows an identical setup but resolves the day into 288 five-minute steps. Daily Monash Weather replaces intra-day terms with $(\sin, \cos)$ of the day-of-year alongside normalized month and day-of-year. Weekly M5 uses $(\sin, \cos)$ of the fiscal week and month, a quarter indicator, and a normalized year. Finally, the weekly internal corpus incorporates ISO week, month, and season indices, the observed history length, and binary flags for holidays. On all four public benchmarks, the eighth dimension is a reserved placeholder set to zero. Crucially, every temporal feature is a deterministic function of the timestamp and leaks no information about the target series.

Encoder and head. We embed each demand scalar via a $\mathrm{Linear}(1{\to}256)$ layer, add it to the broadcast context, and augment it with sinusoidal positional encodings. The encoder comprises six identical pre-norm layers. Each layer features 8-head scaled dot-product attention (head dimension 32), a feed-forward network (hidden dimension 1024, ReLU), and a dropout rate of 0.10. Attention operates strictly over the L temporal tokens of a single series. Because the model never attends across series, population structure reaches the network exclusively through the per-series embedding table and the injected topological descriptors. The forecast head decodes the mean-pooled encoder representation via a sequential block ($\mathrm{Linear}(256{\to}256) \to \mathrm{ReLU} \to \mathrm{Dropout}(0.1) \to \mathrm{Linear}(256{\to}H{\times}Q)$). This produces Q outputs at each of the H forecast steps, where H represents the dataset's horizon (Table 16). We predict $Q{=}9$ quantiles ($\mathcal{Q}{=}\{0.02, 0.10, 0.20, 0.30, 0.50, 0.70, 0.80, 0.90, 0.98\}$) for the internal corpus, ECL, METR-LA, and M5. For Monash Weather, we output a point forecast ($Q{=}1$). The total parameter count is approximately 6.5M for the internal corpus and 5.5M for the public benchmarks, which have smaller entity vocabularies.

Optimizer and scheduler. We train all fully-trained Transformer variants using the AdamW optimizer ($\beta_1{=}0.9$, $\beta_2{=}0.999$, $\epsilon{=}10^{-8}$, weight decay 10^{-3}, learning rate 10^{-4}) alongside a OneCycleLR scheduler (maximum learning rate $3{\times}10^{-4}$). For topology adapters on frozen backbones, we use identical AdamW settings but increase the learning rate to $3{\times}10^{-4}$ and swap the scheduler for CosineAnnealingLR (period equal to the number of epochs, minimum learning rate set to $0.1\times$ the initial learning rate). The fully-trained Transformer optimizes a Huber quantile loss ($\delta{=}1.0$) over 9 output quantiles for the internal corpus, ECL, METR-LA, and M5. On Monash Weather, it optimizes a standard point Huber loss ($\delta{=}1.0$). The frozen-backbone adapters follow a different loss pattern determined by the adapter form rather than the dataset (detailed in Table 18).

Dataset-specific settings. Table 16 details the context window and forecast horizon for each dataset. While all variants share the core encoder architecture described above, the sequence length, horizon, temporal feature dimensionality, and output head vary by dataset. Following the Chronos evaluation protocol, the fully-trained Transformer uses a single-output point head trained with Huber loss for Monash Weather. Every other dataset utilizes the 9-quantile head, with reported MAE and MSE metrics computed from the median quantile. The frozen-backbone adapters deviate from this dataset-level split, as outlined in Table 18.

Table 16: **Context and forecast horizon per dataset.** Context windows follow established benchmark protocols. ECL uses 96 hr (4 days), matching standard long-term forecasting evaluations (Wu et al., 2021). METR-LA uses 12 steps (60 min) of context for the Transformer and 96 steps (8 hr) for PatchTST, following the standard protocol for each (Li et al., 2018; Nie et al., 2023), and forecasts a 12-step block scored at the 3-step (15 min) horizon (corresponding to the METR-LA column of Table 1). ECL trains one checkpoint per horizon, whereas all other datasets use a single horizon. All datasets use 8 temporal feature dimensions.

Dataset	Freq.	Context	Horizon	Temp. dim
Internal corpus	weekly	52 weeks	4 weeks	8
ECL	hourly	96 hr	96 / 192 / 336 / 720 hr	8
M5 (Walmart)	weekly	52 weeks	4 weeks	8
METR-LA	5-minute	12 / 96 steps	12 steps (scored at 3)	8
Monash Weather	daily	300 days	30 days	8

K Injection and Adapter Details

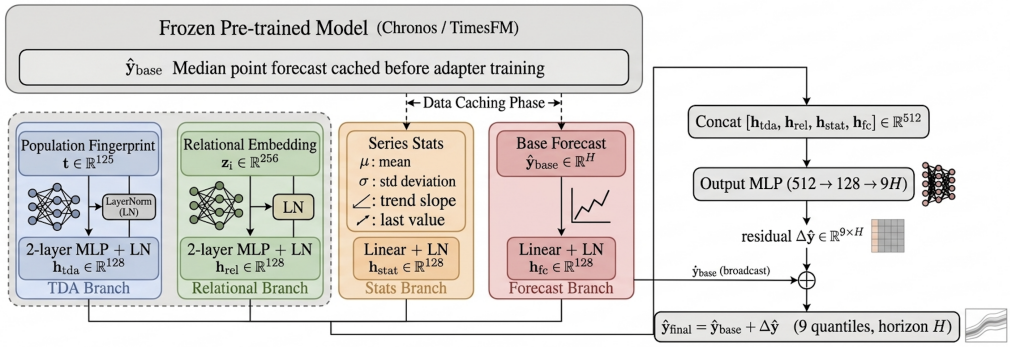

Figure 7: **Four-branch adapter, with dimensions.** The detailed form of Fig. 3a, used on Monash Weather and METR-LA. Four independent branches each project to a common hidden dimension d_h=128 — the TDA fingerprint (blue), the per-series spectral relational coordinate (green), four z-scored series statistics (orange), and the frozen backbone's cached median forecast (red). Separate projections ensure no branch dominates by sheer input dimensionality. The concatenated 512-dimensional representation — 384-dimensional when only one topology branch is active — passes through an output MLP $(\cdot \to d_h \to 9H)$ producing a residual correction $\Delta\hat{\mathbf{y}} \in \mathbb{R}^{9\times H}$, where H is the forecast horizon, added to the base forecast across all 9 quantiles and passed through a softplus to yield $\hat{\mathbf{y}}_{\text{final}}$. An inactive topology branch is retained and fed zeros here, so Vanilla and +TDA are parameter-matched in this form. The fused form used on ECL and M5 (Fig. 3b) instead concatenates the same four inputs without per-input projection (481 dimensions at ECL's H=96) into a two-layer MLP of width 256 with a zero-initialized, median-only head trained on MSE; there an inactive input is dropped from the concatenation rather than zeroed, so trainable parameters vary by variant (117,344 / 182,880 / 214,880 for Vanilla / +Rel / +TDA+Rel on the Chronos-2 ECL basis).

Filtration and vectorization details. We sparsify the correlation matrix to each series' k=50 nearest neighbors, reducing the memory footprint from $O(N^2)$ to $O(Nk)$. We apply the same k to every dataset (clamped to $N-1$). Consequently, the retained fraction varies based on the population over which a single fingerprint is computed, rather than the overall benchmark size (e.g., 24% for METR-LA at N=207, compared to 1.7% for Monash Weather at 3,010). M5 presents an exception because we compute its fingerprints per store×category rather than over the entire panel. After filtering, a block averages ≈1,350 (Foods), ≈990 (Household), or ≈530 (Hobbies) series, yielding a retained fraction of 4–9%. The pooled M5 populations are much larger ($50/28{,}695 \approx 0.2\%$ over the full filtered item×store panel, and $50/9{,}890 \approx 0.5\%$ over the Household subset used for the H_1 statistics in Appendix S). Because k bounds which cycles can appear in the filtration, readers should interpret the cross-dataset H_1 comparison (Appendix N, Table 21) with this spread in mind. The k-nearest-neighbor relation is directional, which we symmetrize by averaging the two directed weights. We retain an edge if it exists in either direction. Thus, a mutually-nearest pair enters the filtration at $1-\rho_{ij}$, while a pair retained in only one direction enters at $(1-\rho_{ij})/2$. All other pairs are discarded. The filtration parameter therefore acts as a rescaled correlation distance rather than a uniform $1-\rho_{ij}$. We run the Vietoris-Rips filtration up to dimension 2 to cover the three fundamental topological primitives. Higher dimensions are computationally expensive and empirically absent in correlation manifolds of this scale. Including λ_2 for H_0 and H_1 captures secondary structures (such as a two-cluster market split) that the top landscape alone misses. For H_2, where voids are sparse, we omit λ_2 because it contributes noise rather than signal.

Landscape sampling grid. We do *not* renormalize distances into $[0,1]$ before filtration. Instead, the filtration runs directly on the symmetrized, sparsified distances defined above. We evaluate the landscape on the interval $[0, t_{\max}]$, where $t_{\max}$ is the largest finite death time across the H_0, H_1, and H_2 diagrams. Sharing $t_{\max}$ across all three ensures the blocks of the 125-dimensional vector sit on a common scale. We do not truncate any features. Although D_{ij} can theoretically reach 2, the k=50 sparsification retains only the closest neighbors, keeping the largest observed death time below 1

on every benchmark ($t_{\max} = 0.94$ for ECL, 0.94 for Monash Weather, 0.84 for M5 `CA_1`/FOODS, and 0.72 for METR-LA). The $t \in [0, 1]$ axis in Figure 8 therefore represents the normalized index $t/t_{\max}$.

Global context injection. Although the spectral coordinates remain frozen, their contribution is gated through a learned projection, $\mathbf{W}_{\text{rel}}$. We use a dedicated path on the expectation that a shared joint layer would assign near-zero weights to the relational columns early in training, suppressing the signal before it can influence the model. This is a design precaution rather than a measured effect: we did not train a fused-projection variant, so we cannot quantify what the separate path contributes. Figure 2 (Section 3) diagrams this broadcast-injection mechanism.

Adapter architecture. Figure 7 diagrams the full four-branch adapter layout. In this configuration, the adapter processes four inputs through dedicated branches. Each branch projects to a common hidden dimension of 128, preventing any single input from dominating the representation by sheer size. The four branches are:

- **TDA branch:** 125-dim population fingerprint, two-layer MLP with LayerNorm.
- **Rel branch:** 256-dim spectral coordinate, two-layer MLP with LayerNorm.
- **Stats branch:** four z-scored series statistics (mean, standard deviation, linear trend slope, and last observed value), linear layer with LayerNorm.
- **Forecast branch:** cached median forecast $\hat{\mathbf{y}}_{\text{base}} \in \mathbb{R}^H$ from the frozen backbone, projected via linear layer with LayerNorm.

Z-scoring the series statistics removes cross-series scale variation, ensuring the adapter learns structural correction patterns rather than simple unit conversions. We concatenate the four projected representations, yielding 512 dimensions when all branches are active (4×128) or 384 dimensions when only a single topology branch is active. This concatenated vector passes through an output MLP ($\cdot \to d_h \to 9H$) to predict the residual $\Delta\hat{\mathbf{y}} = \text{OutputMLP}([\mathbf{h}_{\text{tda}}, \mathbf{h}_{\text{rel}}, \mathbf{h}_{\text{stat}}, \mathbf{h}_{\text{fc}}]) \in \mathbb{R}^{9\times H}$. The final forecast is computed as $\hat{\mathbf{y}}_{\text{final}} = \text{softplus}(\hat{\mathbf{y}}_{\text{base}} + \Delta\hat{\mathbf{y}})$. Here, we broadcast $\hat{\mathbf{y}}_{\text{base}}$ across the nine quantiles as a warm start, and the softplus activation ensures the final forecast remains strictly non-negative. We train the adapter using a Huber quantile loss ($\delta = 1.0$) and allow no gradients to flow back through the frozen backbone. The Rel branch is active only in the +TDA+Rel variant. For the Vanilla, +TDA, and +Rel configurations, the module runs three branches, with the single topology branch carrying a zero vector, the fingerprint, or the relational coordinate, respectively.

The fused form (used on ECL and M5) skips the per-input projections entirely. Instead, it concatenates the raw inputs (e.g., 481 dimensions at ECL's horizon H=96, derived from $96{+}4{+}125{+}256$) directly into a two-layer MLP of width 256. This variant uses $[\text{mean}, \text{std}, \max, p_{75}]$ as the series statistics and employs a zero-initialized, median-only head trained on mean squared error. Consequently, readers should interpret the ECL and M5 rows in Table 1 as using this fused variant rather than the four-branch module from Figure 7. Table 17 contrasts the two forms line by line. Crucially, the topology inputs, frozen backbone, and residual-correction design remain identical across both architectures.

Table 17: The two adapter forms. Both take the same four inputs, freeze the backbone, and train only on cached base forecasts; all reported metrics are computed from the median forecast in both cases.

	Fused	**Four-branch**
Datasets	ECL, M5	Monash Weather, METR-LA
Input handling	raw concatenation, no projection	per-input MLP to $\mathbb{R}^{128}$, LN
Concatenation	481 (96+4+125+256, ECL)	384–512 (3–4×128)
Inactive input	dropped from concatenation	branch retained, zero-filled
Series statistics	mean, std, max, p_{75}	mean, std, trend slope, last value
Trunk	2×[Linear 256, LN, GELU]	Linear 384–512→128, ReLU
Output head	Linear $256 \to H$, zero-init	Linear $128 \to 9H$
Output activation	none	softplus
Training objective	MSE on the point forecast	Huber quantile loss, 9 quantiles

The adapter employs the 9-quantile head for METR-LA and Monash Weather, and a median-only head for ECL and M5. Because we compute all reported metrics from the median forecast, these head differences do not affect comparisons. Table 18 summarizes the output head and training loss for every dataset and model family.

Table 18: **Output head and training loss by model family and dataset.** For the fully-trained Transformer, the configuration splits by dataset. For the frozen-backbone adapter, it splits by adapter form (Appendix K), explaining why ECL and M5 differ from METR-LA and Monash Weather. We compute all reported MAE, MSE, MAPE, and WAPE figures from the median (q=0.5) forecast in every row.

Model family	Dataset	Head	Training loss	Grid
Transformer	Internal corpus	9-quantile	Huber quantile (δ=1.0)	A
Transformer	ECL	9-quantile	Huber quantile (δ=1.0)	A
Transformer	METR-LA	9-quantile	Huber quantile (δ=1.0)	A
Transformer	M5	9-quantile	Huber quantile (δ=1.0)	A
Transformer	Monash Weather	point (1)	Huber (δ=1.0)	–
Adapter (fused)	ECL	median-only (1)	MSE on the point forecast	–
Adapter (fused)	M5	median-only (1)	MSE on the point forecast	–
Adapter (four-branch)	METR-LA	9-quantile	Huber quantile (δ=1.0)	B
Adapter (four-branch)	Monash Weather	9-quantile	Huber quantile (δ=1.0)	B

Grid A $= \{0.02, 0.10, 0.20, 0.30, 0.50, 0.70, 0.80, 0.90, 0.98\}$;
Grid B $= \{0.05, 0.10, 0.20, 0.30, 0.50, 0.70, 0.80, 0.90, 0.95\}$.

Training objective and quantile grid. The training objective varies between the two adapter forms. The four-branch adapter minimizes a Huber quantile loss over nine quantiles, whereas the fused adapter minimizes the mean squared error on the point forecast produced by its median-only head. For models utilizing the nine-quantile loss, the specific grids differ slightly. The fully-trained Transformer uses $\{0.02, 0.10, 0.20, 0.30, 0.50, 0.70, 0.80, 0.90, 0.98\}$, while the frozen-backbone adapters use $\{0.05, 0.10, 0.20, 0.30, 0.50, 0.70, 0.80, 0.90, 0.95\}$. We compute all reported MAE, MSE, MAPE, and WAPE figures from the median (q=0.5) forecast common to both grids. Consequently, the within-dataset comparisons in Table 1 (the only comparisons we draw) remain unaffected. Only the tail coverage depends on the specific grid. We note one direct caveat. The nine-quantile objective provides auxiliary multi-quantile supervision that the fused adapter's scalar MSE lacks, meaning the ECL and M5 adapter rows train without this regularization. Because this difference in adapter form remains constant across all variants within a given dataset, it does not affect our internal comparisons. It does, however, mean that the two adapter families are not interchangeable. Readers should not compare absolute values across dataset blocks. The ECL and M5 adapters train under a different objective than the METR-LA and Monash Weather adapters, in addition to the unit and normalization differences noted in the caption.

Vanilla adapter (precise definition). The "Vanilla" designation refers to the specific adapter form used for a given dataset, but with all topology inputs removed. For ECL and M5, this corresponds to the fused adapter described above, dropping both topology inputs from the concatenation. It observes only the base forecast $\hat{\mathbf{y}}_{\text{base}}$ and the four series statistics. This results in a 100-dimensional input at ECL's H=96, compared to 356 for +Rel and 481 for +TDA+Rel. The trainable parameter counts on the Chronos-2 ECL basis are therefore 117,344, 182,880, and 214,880. These increments stem from adding $256\times256 = 65{,}536$ and $125\times256 = 32{,}000$ weights to the fused MLP's first layer. The remaining layers (the $256\to256$ hidden layer, two LayerNorms, and the $256\to H$ head) remain identical across all three variants. Conversely, for Monash Weather and METR-LA, "Vanilla" utilizes the four-branch module with its topology branch retained but fed a zero vector. Consequently, the Vanilla and +TDA variants are exactly parameter-matched on those datasets.

Because the fused variants are not parameter-matched, the Vanilla row serves as a reference point rather than a strict capacity control. We control for capacity separately using topology-free descriptors injected at identical dimensions, slots, and seed counts. These include `CorrRow` (256-dim in the relational slot, matching the parameter count of +Rel) and `CovEig` (125-dim in the fingerprint slot) detailed in Appendix E, the capacity-matched sliding-window descriptor in Appendix F, and the

parameter-matched partial fine-tune ($181{,}248$ parameters) in Table 7. These controls establish that the gain is not a pure capacity effect, but they do not isolate it to topology: `CorrRow` recovers most of the ECL gain and edges +Rel on Weather (Section 4.2). When the main text states that the encoder and output head remain unchanged across variants, this refers strictly to their architectural *shape* (depth, width, output head), not to an identical parameter count across all adapter variants.

Parameter counts and per-benchmark variation. The total adapter size depends on the forecast horizon (which dictates the forecast branch and output head dimensions) and the number of active branches. Across the benchmarks reported here, every adapter contains fewer than 0.22M parameters. The largest is the $214{,}880$-parameter +TDA+Rel configuration on the Chronos-2 ECL basis (Appendix D). For reference, the frozen backbones are Chronos-2-small ($\approx$28M parameters) and TimesFM-2.5 ($\approx$200M parameters). The adapter thus represents less than one percent of the smaller backbone's parameter count (0.64% on the Chronos-2 ECL basis). In the fused form, we remove inactive inputs from the concatenation rather than zeroing them out. Consequently, the Vanilla and +TDA configurations have correspondingly fewer parameters, even though the encoder and output-head widths remain fixed across variants. The four-branch form instead retains the topology branch, feeding it a zero vector when inactive. As a result, Vanilla and +TDA are perfectly parameter-matched on Monash Weather and METR-LA; only the +TDA+Rel variant activates a fourth branch in those settings.

L COMPUTATIONAL EFFICIENCY

Hardware. We conducted all experiments on a cluster of 8×NVIDIA A100 (40 GB) GPUs. TDA precomputation (nearest-neighbor graph, Vietoris–Rips filtration to H_2, and persistence-landscape vectorization) is a CPU-only process that runs once offline per population, rather than at each training step or forecast. Its computational cost scales with the size of a single filtration rather than the total series count of the benchmark. Although M5 is the largest benchmark by series count (30,490 raw, 28,695 after filtering), it is never presented to Ripser as a single point cloud. Because we compute the fingerprint per store×category, the process consists of thirty independent filtrations of roughly 500 to 1,400 items each (with Foods 13,470, Household 9,890, and Hobbies 5,335 items spread across ten stores). These take only a few seconds apiece. Consequently, the largest single filtration among the public benchmarks is Monash Weather (a single point cloud of 3,010 stations), which takes 3 to 8 minutes. Sparsification keeps this computation tractable. We reduce the distance matrix to each series' k=50 nearest neighbors and pass it to Ripser in compressed sparse form. This bounds the 2-skeleton by the sparsified graph rather than the $\binom{N}{3}$ candidate triangles of a dense cloud. We advise against dense, unpartitioned filtrations for $N10^4$, as the $O(N^2)$ distance matrix and the H_2 computation become binding constraints. Computing the spectral relational coordinate takes less than one second (Appendix H).

FLOPs overhead. The encoder stack dominates the computational cost of the Transformer backbone. For the internal corpus (d_{model}=256, L=52, 6 pre-norm layers, feed-forward width 1024), self-attention requires $4Ld^2_{\text{model}} + 2L^2 d_{\text{model}} \approx 15.0$M MACs per layer (approximately 90.1M over the stack). The feed-forward sublayers add roughly 163.6M MACs, yielding a total of ≈253.7M MACs per forward pass. Because L, d_{model}, layer count, and head count remain identical across all Transformer variants, the encoder incurs the exact same FLOPs regardless of active topology features. TDA augmentation modifies only the shared context projection, which is evaluated exactly *once* per forecast. Its output $\mathbf{g}$ is broadcast-added to all L tokens (Appendix K, Figure 2), meaning this cost does not scale with sequence length. Adding the fingerprint expands the projection input dimension from 584 (Vanilla) to 709 (+TDA), which increases the projection cost from ≈0.15M to ≈0.18M MACs per forecast. This ≈32K MAC overhead represents less than 0.05% of the attention stack cost and 0.02% of the total encoder cost. Because topology features are precomputed once offline, they act as fixed-size vectors read from a lookup table during inference. They add zero runtime FLOPs beyond the initial projection layer.

Public benchmarks: cost of the relational projection. On the public benchmarks, the relational coordinate is active but bypasses $\mathbf{W}_{\text{ctx}}$. Instead, it utilizes the dedicated $\text{Linear}(256\to256)$ path detailed in Appendix J. Compared to the vanilla context projection ($584\to256$, costing 149,504 MACs), +TDA widens the input to 709 and adds 32,000 MACs. The +Rel variant adds 65,536 MACs for the dedicated projection, and the combined +TDA+Rel variant adds both. This totals 97,536 MACs and 98,304 parameters (including the bias and LayerNorm following each projection). Similar to the fingerprint path, the relational projection is evaluated once per forecast rather than once per token. Table 19 contextualizes this fixed cost against the encoder stack. This ratio varies across benchmarks solely due to differences in context length. METR-LA represents the least favorable case at 0.17%, precisely because its short 12-step context makes the baseline encoder exceptionally cheap.

Table 19: Encoder cost against topology overhead, measured in MACs per forward pass (d_{model}=256, 6 layers, feed-forward width 1024). Encoder counts are analytic (attention $4Ld^2 + 2L^2d$ and feed-forward $2Ldd_{\text{ff}}$ per layer, matching the internal-corpus convention above) and exclude normalization, softmax, and the output head. The overhead column represents the added context-projection cost of the fullest variant evaluated on that corpus, incurred once per forecast rather than once per token.

Corpus	Context L	Encoder MACs	Overhead	Share
Internal (+TDA)	52	253.7M	32,000	0.013%
ECL (+TDA+Rel)	96	481.3M	97,536	0.020%
M5 (+TDA+Rel)	52	253.7M	97,536	0.038%
METR-LA (+TDA+Rel)	12	57.1M	97,536	0.171%
Monash Weather (+TDA+Rel)	300	1,692.1M	97,536	0.006%

Frozen-backbone adapters. In the adapter path, the backbone remains frozen and its median forecast is cached. Consequently, the marginal inference cost of topology is strictly that of the adapter itself. On the Chronos-2 ECL basis (H=96), the fused adapter consists of three linear layers ($481 \to 256 \to 256 \to 96$ for +TDA+Rel), costing $213{,}248$ MACs per forecast. The vanilla adapter, which processes only the 100 dimensions of the cached forecast and series statistics, costs $115{,}712$ MACs. Topology therefore adds $97{,}536$ MACs per forecast. This matches the fully-trained path, as both cases add the 125-dimensional fingerprint and the 256-dimensional coordinate projected to a width of 256. For scale, a Chronos-2-small forward pass requires $\approx$560M FLOPs per sample at a sequence length of 512. The entire adapter uses $\approx$0.43M FLOPs, which is less than 0.1% of the backbone cost, and remains under 0.5% even if the backbone cost is scaled down linearly to ECL's 96-step context. We report arithmetic operations rather than wall-clock latency. Because base forecasts are precomputed and read from cache, end-to-end timing for the adapter variants primarily reflects lookup speeds and batch configurations rather than the adapter's intrinsic cost.

Parameter efficiency. The topology-augmented Transformer ($\approx$6.5M trainable parameters) adds only lightweight projections. It introduces $\approx$32K parameters for the 125-dimensional fingerprint on the internal corpus, and $\approx$98K on the public benchmarks when including the dedicated $\mathrm{Linear}(256 \to 256)$ relational path. In both cases, this represents less than 1.6% of the total model size. We provide the corresponding comparison against a parameter-matched adaptation of a pre-trained backbone in Appendix D.

M SCOPE OF THE POPULATION FINGERPRINT

As stated in Section 3 (*Population TDA fingerprint*), the topological fingerprint acts as a cohort-level prior rather than a per-series feature. This appendix formalizes this argument. We record the number of distinct fingerprints present in each experimental setting and outline the implications for interpreting Tables 1 and 2.

The invariance. Because the fingerprint $\mathbf{x}_{\text{tda}} \in \mathbb{R}^{125}$ is computed once per population and frozen, it remains shared across all constituent series. It varies neither across series nor across time steps within that population. Both architectural injection paths consume this fingerprint through an affine map. In the fully-trained Transformer, it forms one block of the input to the context projection ($\mathbf{g} = \mathbf{W}_{\text{ctx}}[\cdots \mid \mathbf{x}_{\text{tda}}] + \mathbf{W}_{\text{rel}}\mathbf{x}_{\text{rel}}$). Its contribution is therefore a fixed vector $\mathbf{c} = \mathbf{W}_{\text{ctx}}^{(\text{tda})}\mathbf{x}_{\text{tda}}$. Similarly, in the adapter, it passes through a dedicated branch whose output $\mathbf{h}_{\text{tda}}$ is concatenated with the other branches before the output MLP. In both cases, the contribution manifests as a fixed vector entering the subsequent linear layer. Because this contribution remains constant across the entire dataset, it is effectively absorbed into the bias of the consuming layer. Letting b denote that bias, the +TDA model computes exactly the same function as the vanilla model with a shifted bias of $b + \mathbf{c}$. The intervening `LayerNorm` does not alter this equivalence, as the shift occurs prior to normalization and is identical for every sample. Consequently, on a closed single-population benchmark, the +TDA and vanilla variants span the *exact same function class*. The fingerprint acts strictly as a macro-level intercept. It cannot supply intra-population, series-discriminating variance, and no amount of training can induce it to do so.

Two practical caveats prevent this theoretical equivalence from translating into identical measured metrics. First, expressive equivalence does not guarantee numerical equivalence. The +TDA model traverses a different optimization path because the induced bias is reparameterized through a $125 \times d$ weight block subject to its own effective step size and weight decay. Furthermore, the adapter's TDA branch applies dropout during training (though at evaluation, with dropout disabled, the contribution remains exactly constant). Second, the variants differ slightly in total parameter count (Appendix D). Together, these effects produce small, sign-inconsistent performance differences that carry no structural information, perfectly aligning with our empirical observations.

How many fingerprints each setting contains. Table 20 details the number of distinct fingerprints present in the training distribution for each setting. It also reports the mean pairwise cosine similarity among them to quantify their actual variance.

Table 20: **Distinct population fingerprints per setting.** Counts are of the vectors present in the *training* distribution; cosines are computed over the 125-dimensional fingerprints after ℓ_2 normalization. A count of one means the invariance argument above applies exactly. Beyond one, spread matters more than count: the internal cohorts are mutually distinct (mean cosine 0.940), whereas the ten M5 vectors are near-identical (0.997) and track a store ID the model already embeds.

Setting	Population unit	Count	Mean pairwise cosine
ECL	domain	1	— (single vector)
Monash Weather	domain	1	— (single vector)
METR-LA	domain	1	— (single vector)
M5 (Household)	store × category	10	0.997 (min 0.992)
M5 cold start	store × category	10	0.996 (min 0.990)
Internal corpus	cohort	*withheld*[‡]	0.940 (min 0.865)

[‡]The exact cohort count is withheld to preserve anonymity of the corpus. It exceeds one, so the exact invariance does not bind; whether the fingerprint is informative in practice depends on the diversity of the set, which the cosines quantify directly.

For ECL, Monash Weather, and METR-LA, a single vector is broadcast to all 321 clients, 3,010 stations, and 207 sensors, respectively. The invariance therefore holds exactly, rendering the +TDA addition expressively null on those benchmarks. M5 represents a near-miss rather than an exact invariance. Although ten store×category fingerprints are present (specifically for the evaluated Household category), their mean pairwise cosine similarity of 0.997 indicates they effectively

function as a single vector plus a small per-store perturbation. Crucially, this perturbation is a deterministic function of the store identity, which the model already receives via a learned embedding. The fingerprint is thus informationally redundant by construction. We apply this same argument to the M5 cold-start cohort in Appendix T, where the ten Household vectors exhibit a cosine similarity of 0.996 and the between-store spread of the TDA effect falls below the seed-to-seed noise floor.

Where the invariance does not bind. This invariance binding condition is a property of the *training* distribution, not the evaluation cohort. For the internal corpus, we train a single model across multiple cohorts possessing genuinely distinct fingerprints (mean pairwise cosine 0.940 with a minimum of 0.865, compared to 0.997 on M5). In this regime, the contribution $\mathbf{W}_{\text{ctx}}^{(\text{tda})}\mathbf{x}_{\text{tda}}$ is no longer constant. It becomes a learned map taking a different value per cohort. This acts as a per-cohort intercept, a feature that a vanilla model (which carries only a single bias shared across all cohorts) fundamentally cannot express. We conduct the reported cold-start and seasonality evaluations within a single cohort where the fingerprint is locally constant. What distinguishes the $+$TDA variant in these evaluations is not variance across the evaluated series, but rather the cohort-specific offset it learned from the wider training distribution. The vanilla model lacks the parameters to represent this offset.

This mechanism also explains why the performance effect concentrates at cold start rather than appearing uniformly across time. A newly launched item enters the model with an all-zero context window and an uninformed item embedding. Because the per-series channels the model normally relies on are empty, the cohort-level prior becomes the only informative signal remaining (Section 4.3). As post-launch history accumulates, the per-series channels recover and the prior's relative contribution shrinks. This aligns perfectly with Table 3, where the performance gap is largest in the early weeks.

We deliberately widened the training set for these runs beyond the evaluation cohort in advance for this exact reason. If trained on the evaluation cohort alone, every sample would receive the same fingerprint, causing the $+$TDA variant to collapse to a global bias and perform no better than the vanilla baseline. This is a design condition that follows from the invariance derived above, not an interpretation imposed after the fact. The fingerprint provides predictive value only when the model is trained across structurally distinct populations.

What follows for the reported tables. These structural properties dictate three key takeaways for interpreting the results. First, the minor $+$TDA performance shifts on ECL, Monash Weather, and METR-LA (Table 1) carry no actual information about topology. They should not be interpreted as weak evidence for or against the method. This caveat also applies to the $+$TDA$+$Rel MSE on ECL, which is expressively identical to the $+$Rel variant alone. Second, the fact that Monash Weather's $+$TDA variant performs *significantly worse* than vanilla in Table 2 ($+0.0348$, $p < 0.001$) does not represent a failure of the descriptor. Instead, it reflects the expected optimization cost of adding parameters to process a constant input, which we report transparently. Third, the topology diagnostic introduced in Appendix N remains completely unaffected by this invariance. The screen uses topology statistics purely as an offline diagnostic computed over cohorts rather than as a model input, meaning cross-cohort variation is precisely the signal it consumes.

N TOPOLOGY SCREENING DETAILS

Table 21 characterizes the four public domains by their H_1 count and loop density (H_1/N), alongside the five-seed Chronos +TDA+Rel ΔMAE. **METR-LA** and **ECL** share similar H_1/N densities (0.22 and 0.26, respectively). **Monash Weather** (H_1/N=0.61) exhibits a denser, genuine loop structure. While **M5 Household** shows a much higher density (H_1/N=4.12), we attribute this inflated count to artifacts. Specifically, shared weekly and annual seasonality likely produce calendar harmonics rather than genuine cross-series relational loops. We do not test this attribution directly.

Two facts limit what this table can establish. First, the realized gain does not rank with H_1/N: the largest improvement occurs on METR-LA, the *lowest*-density domain, and the ordering across the four domains is close to reversed. Second, and more fundamentally, the fingerprint is *expressively null by construction* on the three single-population benchmarks, because a single vector broadcast to every series is absorbed into the consuming layer's bias (Appendix M). The tabulated ΔMAE is therefore carried almost entirely by the relational coordinate rather than by topology: isolating the fingerprint's marginal contribution on top of +Rel gives -0.008 (METR-LA), $+0.008$ (Monash Weather), -0.0002 (ECL), and $+0.0007$ (M5), all within seed noise and inconsistent in sign. These four domains accordingly cannot test a fingerprint-based criterion at all, regardless of which delta we tabulate. We treat Table 21 as a descriptive characterization only, and validate the statistic on the internal cohorts below, where multiple genuinely distinct fingerprints are present.

Table 21: $\boldsymbol{H_1/N}$ **density across the public domains.** H_1 generators represent persistent loops in the correlation manifold, and H_1/N normalizes this count by the series population. The final column reports the five-seed Chronos +TDA+Rel ΔMAE against the vanilla adapter, read directly from Table 1 (Monash Weather at day 30). This delta is dominated by the relational coordinate; the fingerprint's isolated marginal effect is within ± 0.008 and sign-inconsistent (see text), as expected given that each of these domains carries a single fingerprint (Appendix M). The column therefore does not support raw H_1/N as a predictor of TDA amplification, and is not offered as evidence for or against the screening statistic; that statistic is validated on 19 internal cohorts in Table 22. We provide the corresponding H_0, H_1, and H_2 landscape curves in Appendix O.

Dataset	Domain	N	H_1	H_1/N	TDA+Rel ΔMAE
METR-LA	Traffic	207	46	0.22	−0.039
Monash Weather	Weather	3,010	1,847	0.61	−0.020
ECL	Electricity	321	83	0.26	−0.012
M5 Household	Retail	9,890	40,780†	4.12†	+0.001

†M5 H_1 inflated by shared weekly/annual calendar periodicity; genuine cross-series loop count unknown. N is the post-filter descriptor population (Appendix S).

UMAP projections (Figure 9) confirm this geometric basis. ECL and Monash Weather display clear arc and loop structures. METR-LA shows a filamentary topology. In contrast, M5 appears as a structureless, diffuse cloud, which is consistent with calendar-driven correlations and a lack of exploitable manifold geometry.

What this table is and is not. The H_1 and H_1/N columns are a descriptive summary of manifold structure, computed offline and independent of any model. The ΔMAE column records what the coordinates jointly deliver, and as noted above the relational coordinate accounts for essentially all of it while the fingerprint contributes nothing beyond noise. This is the expected outcome rather than a negative result: each of ECL, Monash Weather, and METR-LA carries exactly one fingerprint, so the +TDA variant spans the same function class as the vanilla model (Appendix M), and M5's ten store×category vectors have a mean pairwise cosine of 0.997. Consequently, we make no claim that these four domains support the screening criterion. They motivate it and bound the regime in which the fingerprint can matter; the evidence for it is the n=19 analysis that follows.

Beyond n=4: the criterion at scale. Because four public domains provide an insufficient sample size to validate a ranking criterion, we test the statistic on 19 internal cohorts spanning the full corpus; the cold-start and seasonality runs of Section 4.3 train on a subset of these (Table 20). Each cohort contributes a topology statistic and a realized accuracy gain derived from prior training runs

(cohort identities withheld, Section 4.3). The raw H_1 count is the weakest correlate, because it cannot distinguish a few long-lived structural loops from many short-lived noisy features near the diagonal. Both density-normalized and persistence-normalized statistics improve upon this baseline. The median H_1 lifetime is the strongest correlate ($\rho{=}-0.80$, $p{<}.001$ on Chronos), and is the statistic we report (Table 22). We deliberately do not extend this median lifetime column to the four public domains in Table 21. The statistic is calibrated within a single, homogeneous family of cohorts that share a common filtration scale. Raw lifetimes are not commensurable across public domains whose correlation-distance ranges and population sizes differ by orders of magnitude. Reporting those values would invite the exact cross-domain ranking that the $n{=}4$ evidence fails to support. Supplying a comparable column would require recomputing persistence across all four full populations under a unified normalization scheme, which we leave to future work.

Notably, this association does not hold for TimesFM. No statistic reaches significance for that backbone; the strongest correlate is raw H_1 per entity at $\rho{=}-0.35$ ($p{=}.14$), while median lifetime sits at $\rho{=}-0.21$ ($p{=}.40$). Because topology improves performance in 17 of the 19 TimesFM cohorts, there is very little negative-gain variance left to explain. We therefore present this screening method as a heuristic for prioritizing domains rather than a universally validated rule.

Table 22: Spearman ρ between the topology statistic and realized gain across $n{=}19$ internal cohorts for both adapter backbones. The gain target is ΔWAPE (the oracle best-topology arm against the vanilla adapter, evaluated on the bottom-25% long-tail slice) computed during the adapter sweep. This metric is deliberately distinct from the MAE used for the cold-start and seasonality tables in Appendix B, and the two are not interchangeable. The four statistics shown are the strongest of the sixteen candidates evaluated (see *Provenance of the statistic* below). The screening correlation is significant for Chronos but not for TimesFM.

	Chronos		**TimesFM**	
Statistic	ρ	p	ρ	p
Raw H_1 count	-0.53	.018	-0.03	.887
Raw H_1 / entity	-0.77	$< .001$	-0.35	.137
Persistence-weighted H_1	-0.73	$< .001$	-0.08	.743
Median H_1 lifetime (reported)	$\mathbf{-0.80}$	$< .001$	-0.21	.399

Provenance of the statistic. We state the provenance of this statistic plainly: the realized gains were measured first, and the median H_1 lifetime was selected as the strongest of sixteen candidate statistics evaluated against those same nineteen cohorts. It is therefore a *retrospective* diagnostic with no held-out validation, not a criterion committed to in advance. We report it because the association is strong and mechanistically motivated, but the reported ρ is optimistic by an unquantified selection margin, and prospective use requires validation on cohorts not used here.

O TDA ANALYSIS OF PUBLIC BENCHMARKS

Figure 5 (Appendix A) illustrates the population-level TDA pipeline. This section reports the two population-manifold diagnostics that bear directly on the screening argument of Appendix N. These diagnostics comprise the persistence landscape fingerprints and the UMAP projections of the correlation manifolds. Our multi-panel figures use a 2×2 layout. This arrangement displays ECL (top-left), Monash Weather (top-right), M5 Household (bottom-left), and METR-LA (bottom-right). We place the two topology-rich benchmarks alongside the two null or sparse cases for direct comparison.

Figure 8 overlays the persistence landscape vectors for all four public benchmarks. The H_0 signal (connected components) remains broadly similar across all datasets. Each population clusters into a small number of dominant groups at coarse scales, contributing little discriminating signal. The primary diagnostic information resides in the H_1 features. Monash Weather shows the broadest and highest-amplitude λ_1 (peak 0.27, with support across most of the filtration interval), which is consistent with cyclic co-movement at several scales. M5 reaches a comparable amplitude (0.21) but concentrates this signal in a single narrower band. It also carries an H_1 generator count an order of magnitude larger than Monash Weather ($40{,}780$ versus $1{,}847$). We attribute this spike to calendar repetition shared by every item (specifically, 7-day and 52-week periodicities) rather than to genuine relational geometry. This likely explains why the TDA fingerprint provides no useful grouping signal on M5, though we do not test this attribution directly. Conversely, ECL exhibits a low amplitude (0.069). This indicates that landscape shape alone cannot reliably rank these four domains. For this reason, the screening criterion in Appendix N relies on a persistence-weighted statistic rather than the raw H_1 count or the visual landscape shape. METR-LA falls between these extremes. While ring roads and interchanges produce a non-trivial H_1 count (46 generators), the physical network remains predominantly tree-like. Consequently, these cycles reflect static road geometry rather than dynamically correlated demand patterns. The H_2 panel (structural voids) shows meaningful activity only for Monash Weather and ECL. This indicates richer multi-dimensional structure in the descriptor itself; because both domains are single-population, it carries no implication for downstream accuracy (App. M). Taken together, these landscape signatures serve as a qualitative companion to the screening criterion in Appendix N rather than its quantitative basis. The criterion itself is a persistence-weighted statistic. Ultimately, neither this statistic nor the landscape shapes can perfectly rank the four public domains (Section 1).

Manifold structure (UMAP projection). Figure 9 presents UMAP 2D projections of each entity correlation manifold to visualize the global geometric structure quantified by persistent homology. The macroscopic shape of the projection directly reflects the underlying topological properties. Specifically, an arc or loop shape serves as the 2D signature of $\beta_1 > 0$. A diffuse cloud indicates low topological structure, whereas a filamentary shape suggests an approximately tree-like topology. Node color encodes the TDA-derived cluster assignments, demonstrating whether these geometric clusters map to interpretable, real-world entity groupings.

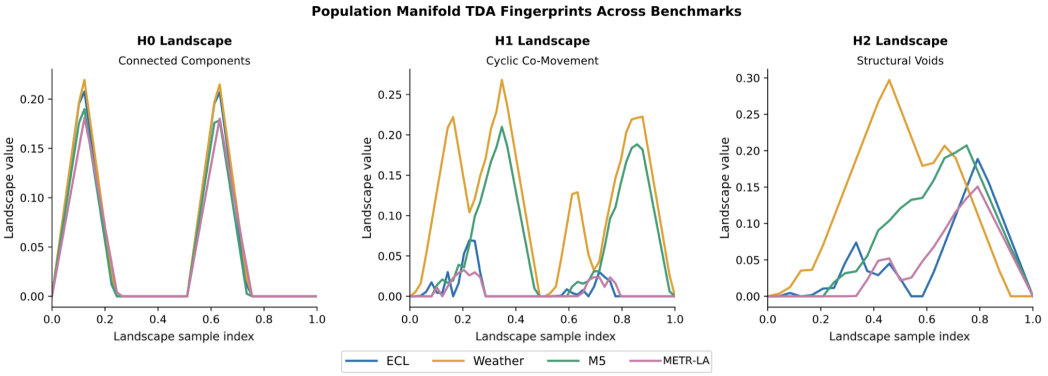

Figure 8: **Population manifold TDA fingerprints across benchmarks.** Each curve represents the 125-dimensional persistence landscape vector (50 dimensions for H_0, 50 for H_1, and 25 for H_2) for a single dataset. H_0 (connected components) remains similar across all datasets, as coarse cluster merging dominates the signal. H_1 (cyclic co-movement) serves as the primary diagnostic panel. Monash Weather has the broadest, highest-amplitude λ_1 (peak 0.27, supported across most of the interval). M5 reaches a comparable peak (0.21) but within a single narrower band. Both ECL (0.069) and METR-LA (0.032) display low amplitudes. H_2 (structural voids) shows activity only for Monash Weather and ECL. In the H_0 and H_1 panels, the first 25 points represent λ_1 and the subsequent 25 represent λ_2, with both sampled over the same interval $[0, t_{\max}]$. The horizontal axis therefore acts as a sample index rather than a single continuous filtration sweep. Consequently, the second rise in those panels indicates λ_2 restarting, rather than a new feature at larger t. The H_2 panel contains λ_1 alone and reads as a single sweep. Appendix K lists the per-dataset $t_{\max}$ values.

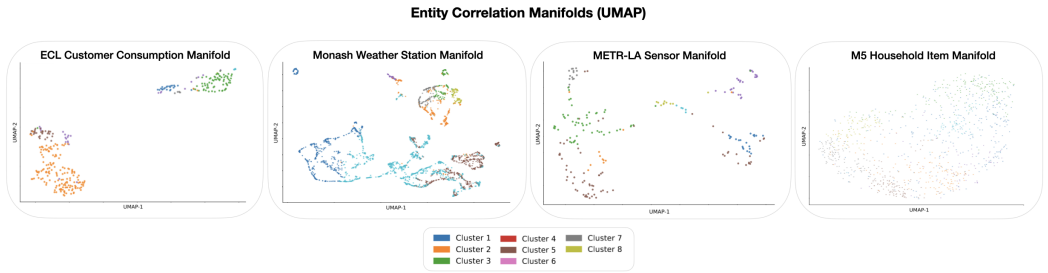

Figure 9: **Entity correlation manifolds (UMAP) across four benchmarks.** The macroscopic projection shape encodes the underlying topology. Arc and loop structures indicate $H_1 > 0$ (ECL and Monash Weather). A filamentary structure indicates a near-tree topology with sparse local loops (METR-LA). Finally, a diffuse cloud indicates low topological structure (M5 Household). Node color encodes the TDA-derived cluster assignments. M5 shows no visual clustering or geometric structure, appearing purely as a diffuse cloud with no arcs, loops, or filaments. This is consistent with a calendar-dominated rather than relational loop structure. In this domain, calendar-dominated correlations produce no exploitable manifold geometry, meaning no topological gain is expected or observed.

P ECL: FULL RESULTS

Note. The per-horizon numbers below derive from our earlier single-seed (seed 42) evaluation. Where they differ, Table 1 remains authoritative. Furthermore, we do not break out the isolated +Rel arm reported in that table here. Because ECL trains one checkpoint per prediction length, that specific arm exists only at $H{=}96$. We provide the complete four-way ladder (Vanilla, +TDA, +Rel, and +TDA+Rel) evaluated across five seeds in Table 5. At $H{=}96$, the seed spread is ± 0.0044 MAE for the fully-trained Transformer and ≤ 0.0005 for the adapters (Table 6). Consequently, readers should interpret per-horizon differences smaller than this spread as indicative rather than statistically resolved.

We benchmark on ECL (321 hourly electricity consumption series, UCI, 2012–2014). This dataset serves as the canonical long-term forecasting benchmark used by Autoformer, PatchTST, and iTransformer.

Protocol. All models use a 96-hour context window (4 days) to match the standard ECL evaluation context. We evaluate prediction horizons of $H \in \{96, 192, 336, 720\}$ hours. We normalize the data (zero-mean, unit-variance per series) prior to both training and evaluation, adhering to the canonical "normalized" protocol. We quote literature rows directly from their source papers under their respective protocols to provide orientation. Because our variants use different training and evaluation code, they are not directly comparable to these baselines (†). Therefore, readers should only draw comparisons within each section.

In Table 23, the TDA fingerprint alone provides no gain (or causes a mild regression) at every horizon for the two adapter families. For the fully-trained Transformer, its effect remains small and mixed in sign. Specifically, it regresses from 0.193 to 0.197 at H96, but improves from 0.234 to 0.227 at H192 and from 0.355 to 0.351 at H720. Unlike Monash Weather, each ECL horizon represents a separately trained checkpoint scored over its entire prediction block. The relational coordinate drives consistent improvements for both adapter families from H96 through H336. However, the two families diverge at H720. The Chronos +TDA+Rel variant still edges out its vanilla adapter (MAE 0.348 to 0.346, MSE 0.259 to 0.257), whereas the TimesFM +TDA+Rel variant reverses this trend (MAE 0.342 to 0.350, MSE 0.252 to 0.269). We have not isolated the exact cause of this divergence. One interpretation consistent with this pattern is that, at long ranges, the backbone's internal distributional prior already covers the information a static topological coordinate would otherwise contribute. This renders the coordinate marginal at best and, in the case of TimesFM, a net cost. We offer this as a plausible interpretation rather than a rigorously tested mechanism.

Table 23: ECL (321 electricity customers): MAE / MSE at $H \in \{96, 192, 336, 720\}$ hours. All trained variants use a 96-hour context window under the normalized protocol. Results reflect a single seed (42). **Bold** indicates the best result per section, and ↓ indicates lower is better. † Our variants are scored at a single forecast origin per series. In contrast, literature rows are quoted from their source papers over rolling test splits and are shown for orientation only (Appendix I). Readers should compare values only within a section, never across the rule.

	H96		H192		H336		H720	
Model	**MAE↓**	**MSE↓**	**MAE↓**	**MSE↓**	**MAE↓**	**MSE↓**	**MAE↓**	**MSE↓**
Literature: published LTSF benchmarks (rolling test split — different protocol)								
Autoformer (Wu et al., 2021)	0.317	0.201	0.334	0.222	0.338	0.231	0.361	0.254
PatchTST (Nie et al., 2023)	0.285	0.195	0.289	0.199	0.305	0.215	0.337	0.256
iTransformer (Liu et al., 2024)	**0.240**	**0.148**	**0.253**	**0.162**	**0.269**	**0.178**	**0.317**	**0.225**
Transformer variants† *(fully-trained)*								
Transformer	**0.193**	**0.091**	0.234	0.125	**0.243**	0.140	0.355	0.289
Transformer + TDA	0.197	0.102	**0.227**	**0.119**	**0.243**	0.141	**0.351**	**0.276**
Transformer + TDA + Rel	0.196	**0.091**	0.231	**0.119**	0.245	**0.136**	0.355	0.279
Chronos 2.0 variants† *(adapter)*								
Chronos Zero-Shot	0.586	0.610	0.594	0.616	0.612	0.638	0.642	0.688
Chronos Vanilla Adapter	0.302	0.205	0.305	0.207	0.319	0.221	0.348	0.259
Chronos + TDA	0.302	0.205	0.305	0.207	0.318	0.220	0.349	0.259
Chronos + TDA + Rel	**0.290**	**0.190**	**0.298**	**0.198**	**0.315**	**0.217**	**0.346**	**0.257**
TimesFM 2.5 variants† *(adapter)*								
TimesFM Zero-Shot	0.580	0.602	0.581	0.599	0.589	0.605	0.607	0.626
TimesFM Vanilla Adapter	0.300	0.204	0.303	0.206	0.315	0.219	**0.342**	**0.252**
TimesFM + TDA	0.300	0.204	0.303	0.206	0.315	0.219	0.343	0.254
TimesFM + TDA + Rel	**0.289**	**0.190**	**0.296**	**0.197**	**0.313**	**0.217**	0.350	0.269

Q MONASH WEATHER: FULL RESULTS

Note. The per-horizon numbers below derive from our earlier single-seed (seed 42) evaluation. Where they differ, Table 1 remains authoritative. We do not break out the isolated +Rel arm reported in that table here; instead, the ladder below evaluates Vanilla, +TDA, and +TDA+Rel at every horizon. We report the full four-way ladder (Vanilla, +TDA, +Rel, and +TDA+Rel) evaluated across five seeds at the terminal horizon in Table 5. The seed spread on Monash Weather reaches ± 0.074 MAE (Table 6). Consequently, readers should interpret the per-horizon differences below as indicative rather than statistically resolved.

Readout. Table 1 scores Monash Weather at a single lead time (day 30). The four horizons below represent step indices into a single 30-day forecast rather than separately trained models. Each row corresponds to a single checkpoint trained at $H{=}30$ and scored at days 1, 7, 14, and 30 of its own output block. We report the terminal step because our core hypothesis concerns slow-moving population structure, which we expect to matter primarily at multi-week lags. For completeness, Table 24 provides the 30-day block mean.

The performance gradient across the forecast block is strong and one-directional. At H1 and H7, the +TDA+Rel variant underperforms its own vanilla arm across all three backbones. For instance, at H1, MAE regresses from 1.555 to 1.617 for the fully-trained Transformer, 1.575 to 1.590 for Chronos, and 1.497 to 1.528 for TimesFM (with similar regressions at H7). Furthermore, zero-shot Chronos beats every Chronos adapter variant at both early steps, recording the lowest overall MAE in the table at H7 (1.468). This ordering reverses by H30, where all three backbones show improvement. The combined +TDA+Rel arm lifts Chronos from a vanilla-adapter MAE of 2.015 to 1.941 (the best MAE in the table), whereas TDA alone provides no improvement (or causes a mild regression) at the shorter steps. The isolated +Rel arm is not part of this specific ladder. However, where it is available (evaluated across five seeds at the terminal horizon), it accounts for essentially all of the observed gain (Table 5, showing 2.0156 for vanilla, 1.9880 for +Rel, and 1.9960 for +TDA+Rel).

When averaged over the full 30-day block, these early and late effects cancel out for the adapters. Table 24 shows the Chronos and TimesFM vanilla adapters matching or beating their respective topology variants. In contrast, the fully-trained Transformer retains a 3.6% MAE gain (2.0070 to 1.9355). We therefore claim a long-lead benefit on Monash Weather rather than a uniform improvement across all horizons. One interpretation consistent with this pattern is that climate-zone structure acts as a slow-moving prior. It carries valuable information about multi-week co-movement but provides little signal regarding a station's own short-lag autocorrelation. We state this strictly as an interpretation rather than a rigorously tested mechanism. We have not run an experiment designed to isolate this specific explanation. Table 25 provides the complete per-horizon breakdown.

Table 24: **Monash Weather scored over the full 30-day block** (mean±std over the same 5 seeds as Table 1, where n=90,300 derived from 3,010 stations × 30 days). This matches the standard Monash and Chronos aggregation convention, reported here alongside the day-30 lead time used in Table 1. The fully-trained Transformer gain persists across the block. However, for both adapter families, the vanilla arm matches or beats its topology variants. Zero-shot rows represent a single run and are excluded from bolding. **Bold** indicates the best result per column within a backbone block, and ↓ indicates lower is better.

Model	MAE↓	MSE↓
Transformer variants (fully-trained)		
Transformer	2.0070 ± 0.0287	15.218 ± 0.114
Transformer + TDA	2.0109 ± 0.0178	15.241 ± 0.111
Transformer + Rel	1.9423 ± 0.0244	14.978 ± 0.087
Transformer + TDA + Rel	**1.9355 ± 0.0165**	**14.916 ± 0.053**
Chronos 2.0 variants (adapter)		
Chronos Zero-Shot	1.9413	16.578
Chronos Vanilla Adapter	1.9131 ± 0.0096	**16.691 ± 0.089**
Chronos + TDA	**1.9123 ± 0.0105**	16.713 ± 0.112
Chronos + Rel	1.9385 ± 0.0064	16.885 ± 0.054
Chronos + TDA + Rel	1.9241 ± 0.0057	16.768 ± 0.072
TimesFM 2.5 variants (adapter)		
TimesFM Zero-Shot	1.8881	16.350
TimesFM Vanilla Adapter	**1.9064 ± 0.0044**	16.407 ± 0.011
TimesFM + TDA	1.9074 ± 0.0026	**16.406 ± 0.026**
TimesFM + Rel	1.9234 ± 0.0047	16.432 ± 0.041
TimesFM + TDA + Rel	1.9263 ± 0.0112	16.483 ± 0.033

Table 25: Monash Weather (3,010 daily series, BOM): MAE / MSE at $H \in \{1, 7, 14, 30\}$ days for all model families. No model has access to the geographic graph. All variants use `CTX_STEPS`= 300. Results reflect a single seed (42). **Bold** indicates the best entry per section per horizon, and ↓ indicates lower is better.

	H1		H7		H14		H30	
Model	**MAE↓**	**MSE↓**	**MAE↓**	**MSE↓**	**MAE↓**	**MSE↓**	**MAE↓**	**MSE↓**
Transformer variants (fully-trained)								
Transformer	**1.555**	**4.778**	1.729	5.586	**2.228**	**13.200**	2.175	25.935
Transformer + TDA	1.556	4.849	**1.667**	**5.250**	2.293	13.589	2.170	26.182
Transformer + TDA + Rel	1.617	5.059	1.743	5.727	2.270	13.468	**2.004**	**25.143**
Chronos 2.0 variants (adapter)								
Chronos Zero-Shot	**1.480**	**5.689**	**1.468**	**5.768**	**2.280**	15.494	2.344	29.776
Chronos Vanilla Adapter	1.575	6.165	1.494	6.067	2.334	15.608	2.015	28.487
Chronos + TDA	1.615	6.406	1.469	5.930	2.325	15.501	2.031	28.381
Chronos + TDA + Rel	1.590	6.228	1.520	6.107	2.290	**15.282**	**1.941**	**27.773**
TimesFM 2.5 variants (adapter)								
TimesFM Zero-Shot	1.478	5.662	**1.505**	5.911	2.350	15.744	2.032	28.032
TimesFM Vanilla Adapter	1.497	5.656	1.526	5.974	2.332	15.429	2.038	28.173
TimesFM + TDA	**1.468**	**5.544**	1.515	**5.781**	2.397	15.666	2.067	28.212
TimesFM + TDA + Rel	1.528	5.854	1.567	6.219	**2.280**	**15.231**	**1.974**	**27.875**

R METR-LA: FULL RESULTS

Note. The per-horizon numbers below derive from our earlier single-run evaluation. Where they differ, Table 1 remains authoritative. Furthermore, we do not break out the isolated +Rel arm reported in that table here. These are single-seed numbers. We provide the complete four-way ladder (Vanilla, +TDA, +Rel, and +TDA+Rel) evaluated across five seeds at the 15 min horizon in Table 5.

METR-LA's topological profile (46 H_1 generators, near-tree road hierarchy) is consistent with a null TDA gain, which the empirical results show. For the two fully-trained backbones, the fingerprint provides no predictive signal. +TDA improves the Transformer at every horizon (-0.007, -0.015, -0.031 MAE), matching the sign of the five-seed effect in Table 1, while its effect on PatchTST is mixed (-0.003, $+0.004$, -0.007). For the two frozen backbones, injecting the fingerprint degrades MAE by $+0.001$ to $+0.116$ across all evaluated settings. Both outcomes are consistent with that null profile. The relational coordinate retains a small signal at the 15 min horizon, where per-sensor behavioral types (e.g., mainline freeway, interchange, or ramp) provide marginal positional context. However, this advantage attenuates or reverses at the 30- and 60-min horizons as traffic dynamics become increasingly non-stationary. Crucially, our Transformer variants train with *no road graph access*. Their absolute MAE is not directly comparable to the literature baselines, which are computed over rolling test splits under the standard masking protocol. Therefore, we draw no conclusions from that contrast (Appendix I). Table 26 provides the complete per-horizon breakdown.

Table 26: METR-LA traffic speed forecasting (MAE / MAPE % at 15, 30, and 60 min, mph, lower is better). Literature models utilize a road-network adjacency matrix. In contrast, all trained variants use *no* graph access. **Bold** indicates the best trained entry per section per horizon. † Our variants are trained on per-sensor standardized speeds, with mean and standard deviation computed on the training period only and inverted at evaluation; we use no test-time instance normalization (in particular, no RevIN, which the reference PatchTST implementation includes). The Transformer variants minimize a Huber quantile loss (δ=1.0) over nine quantiles and are scored at the median; the PatchTST variants minimize MAE in raw mph. Splits are chronological (70%/10%/20%). The Transformer and adapter rows are scored at a single forecast origin per sensor, meaning each horizon cell averages at most 207 absolute errors—one per sensor at a single timestamp, excluding sensors reading below 1 mph; the PatchTST rows are scored over rolling windows ($n \approx 1.4$M per horizon) and are therefore not comparable to them cell-for-cell. The literature rows are quoted from their source papers over rolling test splits under a different masking protocol, and are shown for orientation only (Appendix I). Readers should compare values only within a section, never across the rule.

	15 min		30 min		60 min	
Model	**MAE↓**	**MAPE%↓**	**MAE↓**	**MAPE%↓**	**MAE↓**	**MAPE%↓**
Literature: road graph access (rolling test split — different protocol)						
DCRNN (Li et al., 2018)	2.770	7.300	3.150	8.800	3.600	10.500
Graph WaveNet (Wu et al., 2019)	**2.690**	**6.900**	**3.070**	**8.370**	**3.530**	**10.010**
Literature: no predefined graph (rolling test split — different protocol)						
STAEformer (Liu et al., 2023)	**2.650**	**6.850**	**2.970**	**8.130**	**3.340**	**9.700**
Classical baselines (no graph)						
Last Value	3.060	7.430	3.730	9.630	4.840	13.600
Hist. Average	3.090	7.580	3.690	9.410	4.750	**12.850**
NLinear (Zeng et al., 2023)	**2.900**	**7.070**	**3.550**	**9.250**	**4.640**	13.030
Transformer variants† (fully-trained)						
Transformer	2.213	3.821	1.555	2.683	1.527	2.488
Transformer + TDA	2.206	3.812	1.540	2.666	**1.496**	**2.456**
Transformer + TDA + Rel	**2.203**	**3.809**	**1.521**	**2.627**	1.525	2.507
PatchTST variants† (fully-trained)						
PatchTST	2.737	7.349	3.313	9.560	4.158	12.752
PatchTST + TDA	2.734	7.262	3.317	9.494	4.151	12.771
PatchTST + TDA + Rel	**2.695**	**7.178**	**3.241**	**9.272**	**3.966**	**12.054**
Chronos 2.0 variants† (adapter)						
Chronos Zero-Shot	3.348	5.615	5.502	9.091	8.645	14.025
Chronos Fine-Tuned	3.815	6.288	4.156	6.755	4.830	7.677
Chronos Vanilla Adapter	2.383	4.087	**2.020**	**3.395**	**2.718**	**4.350**
Chronos + TDA	2.392	4.091	2.136	3.569	2.799	4.462
Chronos + TDA + Rel	**2.378**	**4.063**	2.128	3.558	2.783	4.439
TimesFM 2.5 variants† (adapter)						
TimesFM Zero-Shot	2.441	4.200	2.235	3.736	2.980	4.783
TimesFM Vanilla Adapter	2.355	4.058	**2.046**	**3.459**	**2.173**	**3.543**
TimesFM + TDA	2.356	4.064	2.075	3.496	2.185	3.553
TimesFM + TDA + Rel	**2.336**	**4.033**	2.113	3.574	2.405	3.881

S M5: FULL RESULTS

Series counts. Four distinct M5 populations appear in this paper, and they are not interchangeable. The raw benchmark contains 30,490 item×store series (3,049 items × 10 stores). We compute topological descriptors exclusively on the training window. After dropping items active in fewer than the minimum required weeks, 28,695 rows remain (Foods 13,470, Hobbies 5,335, Household 9,890). This specific subset constitutes the population underlying all H_1 statistics. The per-category forecast metrics in Table 27 are reported on the 28,860 series that exhibit non-zero actuals over the test horizon (Foods 13,641, Hobbies 5,306, Household 9,913). **Crucially, the M5 block in Table 1 represents only the Household slice of that population** (9,913 series, evaluated across 5 seeds). Because we ran Foods and Hobbies at a single seed (appearing only in Table 27), the main table remains scoped strictly to the Household category. The cold-start study (Appendix T) utilizes a further-restricted Household universe of 9,258 *established* series (defined as those whose first non-zero sale falls on or before week 11404). The 98 items launching in week 11405 are held out as the evaluation cohort.

M5 presents a unique topological failure mode. Its H_1 structure is artificially inflated by shared weekly and annual seasonality. This creates spurious artifact loops rather than capturing genuine structural heterogeneity. We ran a full cross-category ablation across Foods ($\bar{H}_1 = 3.88$, 13,470 items), Household ($\bar{H}_1 = 4.12$, 9,890 items), and Hobbies ($\bar{H}_1 = 4.44$, 5,335 items). We report Household as the most reliable subcategory. The zero-shot and adapter rows in Table 27 do not share a common target normalization, so we draw no direct comparisons between them here. The definitive within-run comparison is the five-seed Household block in Table 1, where both adapters improve upon the zero-shot baseline. Consistent with M5's null topological profile, topology augmentation yields flat performance across all adapter variants. Household shows a nominal relational gain at this single seed (Chronos MAE $1.040 \rightarrow 1.025$; TimesFM $1.037 \rightarrow 1.025$). However, this gain does not survive replication. Over five seeds, the same comparison reverses (Table 1, $0.7965 \rightarrow 0.7973$). Furthermore, every M5 cell sits entirely within the seed noise floor (gain$/\sigma < 1.2$, $p > 0.05$; Appendix C). We therefore report this nominal gain as noise, which is entirely consistent with M5's status as a screened negative control. Table 27 provides the complete per-category breakdown.

Table 27: M5 benchmark results (28,860 evaluated series; see *Series counts* above) showing the per-category breakdown for a single seed (42) in raw per-category units. The M5 block in Table 1 reports the HOUSEHOLD category alone as a 5-seed mean. Therefore, it corresponds directly to the HOUSEHOLD columns here, rather than to any aggregate over the three categories. This correspondence holds for the fully-trained Transformer rows (1.866 vs. 1.8732 MAE, 0.264 vs. 0.2653 WAPE). However, it does not hold for the adapter rows, which utilize a different target normalization. **The adapter sections are likewise not on a common scale with the baseline and Transformer sections of this table.** This discrepancy explains why the adapter WAPE ranges from 1.2 to 1.8, while the baseline and Transformer WAPE ranges from 0.20 to 0.34. Readers should compare values only within a specific section. No cross-section ranking in this table is meaningful. All rows represent a single seed (42). The definitive five-seed Household figures are located in Table 1. Where the two tables disagree in direction, the five-seed result governs. No claim in this paper rests on a directional ordering taken from this specific table. **Bold** indicates the best entry per section per category, and ↓ indicates lower is better.

	FOODS		HOBBIES		HOUSEHOLD	
Model	**MAE↓**	**WAPE↓**	**MAE↓**	**WAPE↓**	**MAE↓**	**WAPE↓**
Baselines						
XGBoost	**2.937**	**0.201**	**1.443**	**0.280**	**1.734**	**0.246**
NLinear	3.633	0.249	1.602	0.311	1.975	0.280
Rate Based	4.045	0.277	1.494	0.290	1.988	0.282
DLinear	4.901	0.336	1.803	0.350	2.321	0.329
Transformer variants (fully-trained)						
Transformer	**3.768**	**0.258**	**1.417**	**0.275**	1.866	0.264
Transformer + TDA	3.769	**0.258**	1.421	0.276	1.865	0.264
Transformer + TDA + Rel	3.833	0.263	1.423	0.276	**1.827**	**0.259**
Chronos 2.0 variants (adapter)						
Chronos Zero-Shot	**0.854**	**1.222**	**0.941**	**1.597**	**0.918**	**1.450**
Chronos Vanilla Adapter	0.984	1.408	1.054	1.788	1.040	1.643
Chronos + TDA	0.982	1.405	1.052	1.786	1.039	1.641
Chronos + TDA + Rel	0.966	1.383	1.038	1.761	1.025	1.618
TimesFM 2.5 variants (adapter)						
TimesFM Zero-Shot	**0.843**	**1.207**	**0.947**	**1.607**	**0.914**	**1.443**
TimesFM Vanilla Adapter	0.983	1.406	1.057	1.793	1.037	1.636
TimesFM + TDA	0.980	1.402	1.055	1.790	1.034	1.632
TimesFM + TDA + Rel	0.982	1.405	1.054	1.789	1.025	1.618

T M5 COLD START: A PUBLIC REPLICATION

The cold-start result on the internal corpus (Appendix B) represents the only setting where we claim topology drives forecast performance independently. This claim requires a public, reproducible counterpart. We therefore constructed the closest available end-to-end analog on M5 and evaluated it across five training seeds. We emphasize that this is an analog rather than a direct replication, as the two experiments pose fundamentally different questions. The internal setting provides a launch item with an all-zero context to test whether a population fingerprint can substitute for absent history. In contrast, the M5 setting provides a neighbor-imputed context to test whether a per-series relational coordinate improves upon that imputation. We detail these divergences explicitly below (see *Relation to the internal setting*). This appendix reports the outcome, which proves negative in aggregate. However, we also detail the one robust effect it does contain, namely that *the relational coordinate's value scales inversely with series density*.

Protocol. We hold out every M5 Household series whose first sale occurs in week 11405. We train exclusively on series established at or before week 11404, using only data from weeks ≤ 11404. We construct the neighbor index, relational coordinates, and TDA fingerprints entirely from this same ≤ 11404 window. At forecast origin $w \in \{0, \ldots, 12\}$ weeks after launch, the model receives a blended history $h = \alpha\, h_{\text{own}} + (1 - \alpha)\, h_{\text{prior}}$, where $\alpha = \min(n_{\text{nonzero}}/\tau, 1)$ and $\tau = 8$. Here, h_{prior} denotes the mean normalized history of the item's established neighbors. Consequently, at $w = 0$, the input consists entirely of the neighbor prior. We apply this blend identically during both training and evaluation, ensuring that in-training launches exercise the exact same low-history regime. We report results over a four-week horizon, training five seeds $(42–46)$ per variant. The evaluation cohort comprises 98 series across 10 stores (64 representing items existing at other stores, and 34 representing truly novel items). After dropping blocks with zero realized demand, this yields 1,063 scored (series, origin) pairs.

Relation to the internal setting. Two key differences separate this experiment from Appendix B. Both stem from the benchmark's structure rather than arbitrary experimental choices. *Input.* Internally, a launch item enters with an all-zero context window. Here, it enters with h_{prior} (the neighbor-averaged history) because a launch series lacks its own coordinate, and the same neighbor set must supply both inputs. We apply this blend identically to every arm, including the vanilla baseline, preserving a fair comparison. However, this shifts the research question from *can a population descriptor stand in for missing history?* to *does a relational coordinate add value to an imputed history?*. *Descriptor.* Internally, the contrast isolates the 125-dimensional fingerprint while the relational coordinate remains inactive. Here, the fingerprint takes only ten distinct values across the entire cohort (with a mean pairwise cosine similarity of 0.9956). Because it is near-constant by construction (see *Why TDA is inert here*, below), the relational coordinate serves as the primary informative arm. Consequently, the two experiments test different descriptors under different input regimes. Neither confirms nor refutes the other. A closer analog would drop the blend entirely and score every arm using an all-zero context. However, we believe the M5 fingerprint is too degenerate for that test to be informative, so we reserve that evaluation for benchmarks where the fingerprint exhibits meaningful variance across the cohort.

Coordinates for entities absent from the SVD matrix. A truncated SVD assigns a coordinate to a *row* of the entity–time matrix. Therefore, a series with no observations prior to the cutoff lacks a coordinate, and no standard out-of-sample projection is defined for it. We resolve this by assigning cold-start coordinates via neighbor averaging in the established basis. We detail this construction in full below, as it exactly defines what the cold-start arm evaluates.

Basis. Let $\mathbf{X}_{\text{est}} \in \mathbb{R}^{N_e \times T}$ hold the established Household series (those with a first sale at or before week 11404) over the training weeks ≤ 11404, with rows ℓ_2-normalized. We compute a single thin SVD ($\mathbf{X}_{\text{est}} = U\Sigma V^\top$) over *all ten stores simultaneously*. Each established series j then receives $\mathbf{z}_j = U_{j,1:k}$, zero-padded to 256 dimensions, where $k = \min(256, \text{rank})$. This differs from the main M5 benchmark, which blocks the SVD per store×category (Appendix H). We require a single shared basis here to ensure a launch series and its established neighbors map to the same coordinate system for averaging.

Neighbor set. For a launch series s, we define the neighbor set $\mathcal{N}(s)$ as follows:

- **Tier (a):** Every established series with the same `item_id` at another store, if any exist (covering 64 of the 98 cohort series). If none exist, we fall back to Tier (b).
- **Tier (b):** The $K{=}10$ established series in the same store nearest to s in Euclidean distance, computed on a static feature vector $\varphi(s)$ (covering the remaining 34 truly novel items). If the store pool is empty, the search falls back to all established series.

The feature vector φ concatenates one-hot indicators for department, store, and state with price features derived from the item's pre-cutoff list price. These price features include the z-scored $\log$ price and its ratios to the department and category median prices. (In our implementation, the department ratio is entered twice, carrying double weight in the distance calculation). Every component of φ is available on the launch date and relies on zero demand history, making Tier (b) fully defined for items with no prior observations.

Assignment. The launch series then takes

$$\mathbf{z}_s \;=\; \frac{1}{|\mathcal{N}(s)|}\sum_{j\in\mathcal{N}(s)}\mathbf{z}_j, \qquad h_{\text{prior}}(s) \;=\; \frac{1}{|\mathcal{N}(s)|}\sum_{j\in\mathcal{N}(s)}h_j, \tag{1}$$

where h_j represents neighbor j's normalized last-52-week history. We average the denormalization mean and standard deviation over $\mathcal{N}(s)$ in the exact same manner. The history blend $h = \alpha h_{\text{own}} + (1-\alpha)h_{\text{prior}}$ then applies as described previously. Established series use their own $\mathbf{z}_j$ rather than a neighbor average. When exercising this blend during training, an established series excludes itself from its own neighbor set. Truly novel items lack a learned item embedding and fall back to the out-of-vocabulary slot. Consequently, at $w{=}0$, a Tier (b) series is represented to the model exclusively by a neighbor-averaged coordinate, a neighbor-averaged history, its store and category embeddings, and the store$\times$category fingerprint.

How to read the figures. Panel (a) of Figure 10 shares the same axes as Figure 4(a). It plots absolute MAE (lower is better) against the number of weeks of post-launch history accumulated at the forecast origin. Panels (b) and (c) switch to a *relative* quantity: the percentage MAE reduction compared to the vanilla Transformer. Therefore, the horizontal line at zero represents the vanilla arm. Points above the line indicate that topology lowers MAE, while points below indicate that topology raises MAE. Panel (b) aggregates this relative reduction into sextiles of the cohort ordered by demand, labeling each tick with the sextile's median weekly units. Panel (c) disaggregates this same quantity across the 98 individual series. The uncertainty encoding varies by panel and is specified in each subtitle. We show seed CIs as bands in (a), the five individual seeds as dots in (b), and a series-level bootstrap interval as a band in (c).

In Figure 11, each panel displays a histogram over *pairs* of vectors rather than individual vectors. Mass near zero indicates the descriptors point in largely unrelated directions, providing strong signal to separate one series from another. Conversely, mass near one indicates that all vectors are nearly identical, meaning the descriptor carries almost no separating information. We draw the two panels separately because their pair counts differ by five orders of magnitude. Readers should compare distributions along the x-axis, but never compare bar heights across panels.

No aggregate gain. Table 28 presents the headline results. The vanilla Transformer performs best in every column. On the truly novel cohort, the +Rel and +TDA+Rel seed intervals do not overlap the vanilla interval. Topology does not merely fail to help; it actively degrades MAE by 3 to 8%. This effect carries the opposite sign of the internal corpus results, and we measure it robustly over five seeds rather than one. Figure 10(a) shows this same null result resolved by forecast origin. The vanilla arm sits at or below every topology arm across all thirteen origins, and no variant exhibits the launch-week advantage reported internally in Figure 4(a).

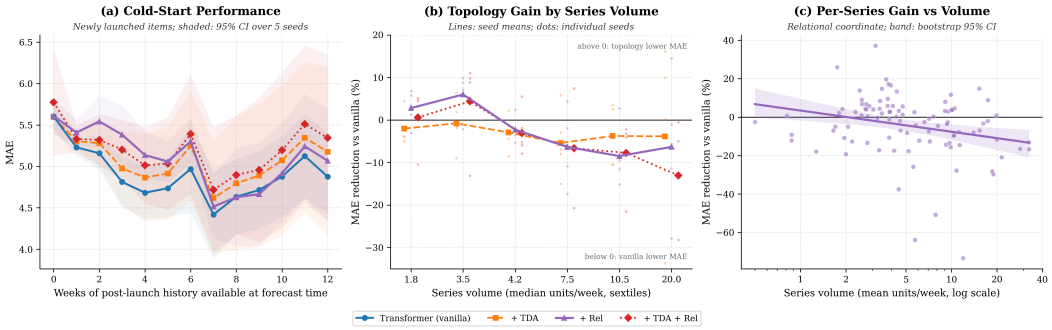

Figure 10: **M5 cold start: no aggregate effect, but a density gradient.** We evaluate five seeds throughout. **(a)** MAE against the number of weeks of post-launch history available at the forecast origin. Shaded bands represent 95% CIs over seeds. **(b)** MAE reduction relative to the vanilla Transformer, grouped by series-volume sextile. The x-axis is labeled by each sextile's median weekly demand. Solid lines represent seed means, while faint dots represent the five individual seeds. The y-axis is clipped to $[-35, 20]$. One seed point for +TDA+Rel at q6 falls below the axis at -59%. **(c)** The same relative quantity plotted per series for +Rel. Each point represents one of the 98 cohort series (averaged over seeds) plotted against its mean weekly demand. The line indicates a least-squares fit, and the shaded band represents a series-level bootstrap 95% interval. Its seed-wise slope is -0.048 ± 0.032 per log-unit of demand, negative in all five seeds. Reading the panels together reveals that topology performs at or worse than vanilla at every origin in (a). However, in (b) and (c), the relational coordinate shows positive gains on the two sparsest sextiles and negative results from the third onward. Meanwhile, TDA remains flat and slightly negative throughout.

Table 28: M5 cold-start MAE (mean ± 95% CI over five training seeds). MAE is averaged over the four-week horizon. *Known* indicates the item already sells at other stores. *Novel* indicates the item is completely new to the corpus. **Bold** indicates the best result per column, and ↓ indicates lower is better.

Model	All↓	Known↓	Novel↓
Transformer	**4.923 ± 0.323**	**5.321 ± 0.486**	**4.105 ± 0.046**
Transformer + TDA	5.092 ± 0.584	5.511 ± 0.820	4.229 ± 0.111
Transformer + Rel	5.118 ± 0.289	5.451 ± 0.418	4.434 ± 0.177
Transformer + TDA + Rel	5.215 ± 0.603	5.637 ± 0.845	4.348 ± 0.163

The effect is not uniform: a density gradient. Partitioning the cohort by realized weekly demand reveals a clear performance reversal. Table T shows that the relational coordinate improves MAE on the sparsest series but degrades it monotonically as demand rises. Aggregating the two extreme sextiles highlights this contrast. On the 39 sparsest series (q1–q2, median 2.5 units/week), +Rel attains an MAE of 2.213 compared to the vanilla 2.324, a 4.7% reduction holding in 4 of 5 seeds. Conversely, on the 27 densest series (q5–q6, median 13.0), it attains 8.769 against 8.258, a 6.8% *increase* in error. As in Table T, percentage changes are means of per-seed ratios and therefore do not equal the ratio of the rounded MAE means quoted here.

the ratio of the displayed MAE means. Positive values indicate that topology performs better. **Bold** indicates that topology improves upon the vanilla baseline. Note that the sign flips between q2 and q3 and does not revert.

the ratio of the displayed MAE means. Positive values indicate that topology performs better. **Bold** indicates that topology improves upon the vanilla baseline. Note that the sign flips between q2 and q3 and does not revert.

Table 29: M5 cold-start MAE grouped by series-volume sextile. Each cell reports the MAE (averaged over five seeds), computed as the mean of per-seed ratios, which differs from
the ratio of the displayed MAE means. Positive values indicate that topology performs better. **Bold** indicates that topology improves upon the vanilla baseline. Note that the sign flips between q2 and q3 and does not revert.

		Median	**Transformer**	**+ TDA**		**+ Rel**		**+ TDA + Rel**	
Sextile	n	**units/wk**	**MAE↓**	**MAE↓**	**Δ**	**MAE↓**	**Δ**	**MAE↓**	**Δ**
q1 (sparsest)	22	1.8	1.848	1.884	−2.0%	**1.797**	+2.8%	**1.837**	+0.6%
q2	17	3.5	2.807	2.825	−0.8%	**2.636**	+6.0%	**2.679**	+4.4%
q3	16	4.2	3.340	3.437	−2.9%	3.416	−2.3%	3.443	−3.1%
q4	16	7.5	5.433	5.724	−5.4%	5.772	−6.3%	5.781	−6.6%
q5	13	10.5	6.177	6.404	−3.8%	6.696	−8.5%	6.646	−7.7%
q6 (densest)	14	20.0	10.279	10.639	−3.8%	10.783	−6.3%	11.301	−13.0%

The gradient, not the cell, is the claim. Because individual cells in Table T are not separately significant across five seeds, we do not claim them as standalone results. Instead, the significance lies in the broader trend. We regress each series' relative MAE gain against the log of its mean demand (fitting one regression per seed). For +Rel, this yields a slope of -0.048 ± 0.032. This slope is negative across all five seeds, with a 95% confidence interval strictly excluding zero. The corresponding slopes for +TDA (-0.015 ± 0.072) and +TDA+Rel (-0.062 ± 0.102) share this negative sign but lack statistical significance. Testing a slope over an ordered covariate is far less vulnerable to selection bias than cherry-picking a subgroup. This distinction is crucial because we discovered this gradient through exploration rather than pre-registration.

We hypothesize that *topology substitutes for history rather than augmenting it*. For a sparse series selling only two units a week, its individual trajectory is virtually indistinguishable from noise. In this regime, a stable structural descriptor provides more value than the sparse history it displaces. Conversely, for a dense series selling twenty units a week, the historical trajectory already constitutes a rich, high-information signal. Here, injecting 256 additional input dimensions merely supplies extra capacity to overfit. This capacity-vs-signal tradeoff naturally predicts the observed sign change and remains testable on any benchmark featuring a wide density range.

Why TDA is inert here. The population fingerprint fails to improve performance in any subgroup we examined. This failure stems from a structural property of M5 rather than a flaw in the method itself. The cold-start cohort consists entirely of Household items across 10 stores. Consequently, the store×category fingerprint takes exactly *ten* distinct values over the entire evaluation. Furthermore, these ten vectors are near-identical, exhibiting a mean pairwise cosine similarity of 0.9956. The TDA fingerprint thus acts as a deterministic, almost-constant function of `store_id`, which the model already receives via a learned embedding. It is therefore informationally redundant by construction. Two empirical checks confirm this redundancy. First, across 39 evaluated subgroups (e.g., cohort, origin, volume, store, state, department), exactly one yielded a positive gain ($+0.1\% \pm 3.4\%$), and none reached statistical significance. Second, the between-store spread of the TDA effect (SD 0.163) remains strictly smaller than its seed-to-seed noise (SD 0.593), indicating that no individual store benefits systematically. M5 cold start is therefore *not* an informative test of the population fingerprint. It is, however, an informative test of the relational coordinate. The relational coordinates in this setting are genuinely rich, comprising 9,258 established Household series projected into 256 dimensions (with a mean pairwise cosine similarity of 0.008 and an effective rank of 160). Figure 11 contrasts these two descriptors directly. Ultimately, the relational arm's failure on dense series reflects a capacity effect rather than a degenerate input.

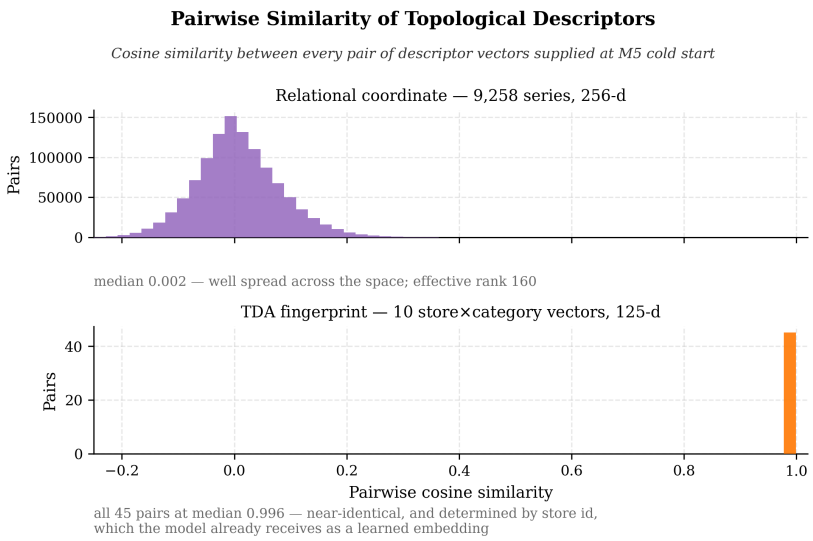

Figure 11: **Descriptor diversity at M5 cold start.** Distribution of pairwise cosine similarity between descriptor vectors. We show these as separate panels because their pair counts differ by five orders of magnitude. The relational coordinate spreads widely across the space (median 0.002). Conversely, the ten TDA vectors available to this cohort are mutually near-identical (all 45 pairs sit at a median of 0.996). These TDA vectors are entirely determined by `store_id`, which the model already receives as a learned embedding.

Comparison with the warm-start M5 results. A single-seed warm-start run (Table 27) shows the same qualitative split. Specifically, TDA remains flat, while a nominal relational gain concentrates in Household (the category with the highest demand heterogeneity). However, we do not treat this as corroboration. The rigorous five-seed evaluation of this exact comparison reverses its sign (Table 1). This discrepancy is visible on a common scale. The Transformer rows of Table 27 (which align with the scale of Table 1) improve from 1.866 to 1.827 at one seed, but degrade from 1.8732 to 1.8925 when averaged over five seeds. Ultimately, the only claim that survives replication on M5 is the weaker observation that the relational coordinate (rather than the fingerprint) is the sole component that induces any performance movement at all.

Scope and limitations. This analysis rests on a single benchmark, one category, 98 cold-start series, and five seeds. Therefore, the observed density gradient must be replicated on other datasets before it can be treated as a general phenomenon. Furthermore, this experiment does not test the entity-level fingerprint used on the internal corpus. It serves as a weak test of the population fingerprint for the structural reasons detailed above. It also poses a fundamentally different question from the internal cold-start study regarding both its input and its descriptor (see *Relation to the internal setting*). We report this experiment because it constitutes our only cold-start evaluation on public data, and because it does not support a direct cold-start reading of Appendix B. Consequently, we do not present M5 as corroboration of that internal result. Instead, we treat M5 strictly as the negative control indicated by its topological profile (Appendix N).

U Mathematical Preliminaries

U.1 On the Manifold Claim

The premise that a time series population exhibits a meaningful *geometry* (Section 1) is a modeling assumption standard to manifold learning (e.g., Isomap, UMAP, LLE) rather than a strict topological property of the raw data. A finite sample of discrete time series does not inherently form a continuous manifold. However, the pairwise correlation structure does induce a formal manifold: the correlation matrix $\mathbf{C}$ resides on the elliptope of positive-semidefinite matrices with unit diagonals. It is the population's induced point cloud in this correlation-distance space that persistent homology and UMAP (Appendix O) probe directly.

U.2 Simplicial Complexes and Filtrations

A *simplicial complex* $\mathcal{K}$ is a set of simplices (vertices, edges, triangles, and higher-dimensional analogs) that is closed under the inclusion of faces: if $\sigma \in \mathcal{K}$ is a simplex and $\tau \subseteq \sigma$, then $\tau \in \mathcal{K}$. A *filtration* is a nested sequence of subcomplexes $\emptyset \subseteq \mathcal{K}_{\varepsilon_1} \subseteq \mathcal{K}_{\varepsilon_2} \subseteq \cdots \subseteq \mathcal{K}$, parameterized by a monotonically increasing scale parameter $\varepsilon \geq 0$. The Vietoris-Rips construction provides a canonical method for generating such a filtration from pairwise point distances.

U.3 Vietoris-Rips Filtration

Given a finite point cloud $\mathcal{P}$ equipped with a distance metric d, the *Vietoris-Rips complex* at scale ε is defined as the set of all simplices whose pairwise vertex distances are at most ε:

$$\mathrm{VR}(\mathcal{P}, \varepsilon) = \big\{ \sigma \subseteq \mathcal{P} \;\big|\; \max_{p,q \in \sigma} d(p, q) \leq \varepsilon \big\}.$$

In our framework, $\mathcal{P} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ represents the population of time series, and $d_{ij} = 1 - \rho_{ij}$ is the Pearson correlation distance. This metric naturally places strongly anti-correlated series far apart. Applying this filtration yields topological summaries of the *entire series population*. These summaries capture macroscopic structures: series clusters (H_0), cyclic co-movements (H_1), and enclosed structural voids (H_2). Crucially, these are global descriptors of the population manifold, not properties of individual time series.

U.4 Persistent Homology

The k-th homology group $H_k(\mathcal{K})$ measures the number of independent k-dimensional holes in a complex. *Persistent homology* (Carlsson, 2009) tracks the evolution of these topological features across a filtration, recording the scale ε at which each homology class is born and dies. These events are represented as birth-death pairs (b, d), which collectively form a persistence diagram PD_k. The persistence of a feature, defined as $d - b$, quantifies its prominence: features with large persistence correspond to robust structural properties of the data, while short-lived features are typically attributed to topological noise.

U.5 Persistence Landscapes

To integrate topological features into neural architectures, the *persistence landscape* (Bubenik, 2015) maps the unstructured diagram PD_k into a functional space:

$$\lambda_k^{(\ell)}(t) = \ell\text{-th largest value of } \big\{ (\min(t - b,\, d - t))_+ \;:\; (b, d) \in \mathrm{PD}_k \big\},$$

where $(\cdot)_+ = \max(0, \cdot)$. Discretizing this function over a fixed grid yields an r-dimensional vector representation that is Lipschitz-stable with respect to the bottleneck distance d_B (see Appendix V).

U.6 Cellular Sheaves

A *cellular sheaf* $\mathcal{F}$ over a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ assigns a vector space (stalk) $\mathcal{F}_v \cong \mathbb{R}^d$ to each node and a linear *restriction map* $\mathcal{F}_{v \trianglelefteq e}$ to each node-edge incidence. The *sheaf Laplacian* $\mathbf{L}_{\mathcal{F}} = \delta^\top \delta$

measures the deviation from global consistency across the graph. For a signal $\mathbf{x}$ over the nodes, the Dirichlet energy is given by:

$$\mathbf{x}^\top \mathbf{L}_\mathcal{F} \mathbf{x} = \sum_{(u,v)\in\mathcal{E}} w_{uv} \left\| \mathcal{F}_{u\trianglelefteq e}(\mathbf{x}_u) - \mathcal{F}_{v\trianglelefteq e}(\mathbf{x}_v) \right\|^2.$$

This term penalizes the disagreement between adjacent nodes u and v after their features are projected into the shared edge space.

While this general formalism motivates our relational coordinates, our implementations operate in more restricted regimes. The trained *neural* encoder (Appendix H) learns a single global restriction map R, reducing the consistency term to the standard weighted graph Dirichlet energy of the projected embeddings. More importantly, the *spectral* encoder (which drives all main results in this paper) fixes these restriction maps to the identity. Under identity restrictions, the minimal-Dirichlet-energy subspace is spanned simply by the leading left singular vectors of the entity-time matrix. This yields a closed-form relational coordinate that requires no gradient-based training. Section 3 details both encoder variants.

Derivation: identity maps reduce to leading singular vectors. When all restriction maps are the identity ($\mathcal{F}_{v\trianglelefteq e} = I$), the sheaf Laplacian simplifies to the standard weighted graph Laplacian $\mathbf{L}_\mathcal{F} = \mathbf{D} - \mathbf{W}$, where $\mathbf{W}_{uv} = w_{uv}$ is the edge-weight matrix and $\mathbf{D}$ is the degree matrix.

We define the graph weights as $w_{uv} = |\rho_{uv}|$, the absolute Pearson correlation between time series. This affinity deliberately differs from the correlation *distance* $D_{ij} = 1 - \rho_{ij}$ used for the Vietoris-Rips filtration in Section 3. While the filtration places anti-correlated series far apart, the Laplacian treats strong negative coupling as equally informative to strong positive coupling. The two constructions serve distinct, complementary roles.

Let $\mathbf{M} \in \mathbb{R}^{N\times T}$ denote the entity-time matrix, where each row is centered across time ($\sum_t \mathbf{M}_{it} = 0$) and scaled to unit ℓ_2 norm. Under these operations, the Pearson correlation ρ_{uv} is exactly the inner product of rows u and v, yielding the weight matrix $\mathbf{W} = |\mathbf{M}\mathbf{M}^\top|$ (element-wise absolute value) with a unit diagonal. Row-centering is strictly required; otherwise, the inner product of non-negative data is dominated by the shared mean, falsely pushing uncorrelated series near 1.

The *minimal Dirichlet energy subspace* of dimension k minimizes the consistency loss:

$$\min_{\mathbf{X}\in\mathbb{R}^{N\times k},\, \mathbf{X}^\top\mathbf{X}=I} \operatorname{tr}\left(\mathbf{X}^\top \mathbf{L}_\mathcal{F} \mathbf{X}\right).$$

We deliberately avoid the term *harmonic*. A true harmonic section lies in $\ker(\mathbf{L}_\mathcal{F})$ and has zero Dirichlet energy. For a connected graph with identity restriction maps, this kernel is one-dimensional and spanned by $\mathbf{1}$. Because we use $k = 256$, the minimizer instead spans the k lowest-frequency Laplacian eigenvectors. All but one of these possess strictly positive energy and are therefore not harmonic.

Substituting $\mathbf{L}_\mathcal{F} = \mathbf{D} - \mathbf{W}$ reformulates the objective as $\operatorname{tr}(\mathbf{X}^\top\mathbf{D}\mathbf{X}) - \operatorname{tr}(\mathbf{X}^\top\mathbf{W}\mathbf{X})$. Assuming a d-regular graph, the degree term $\operatorname{tr}(\mathbf{X}^\top\mathbf{D}\mathbf{X}) = d\operatorname{tr}(\mathbf{X}^\top\mathbf{X}) = dk$ becomes constant under the orthonormality constraint. This reduces the problem to maximizing the weight term:

$$\max_{\mathbf{X}^\top\mathbf{X}=I} \operatorname{tr}\left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right).$$

By the Rayleigh-Ritz theorem, the solution $\mathbf{X} = \mathbf{U}_k$ is the matrix of leading eigenvectors of $\mathbf{W}$. Because $\mathbf{W} = |\mathbf{M}\mathbf{M}^\top|$, and time series domains often exhibit predominantly positive cross-correlations ($|\mathbf{M}\mathbf{M}^\top| \approx \mathbf{M}\mathbf{M}^\top$), the leading eigenvectors of $\mathbf{W}$ are well-approximated by the leading left singular vectors of $\mathbf{M}$ (via $\mathbf{M}\mathbf{M}^\top\mathbf{u}_i = \sigma_i^2\mathbf{u}_i$). Consequently, the minimal-Dirichlet-energy subspace of the identity-restriction sheaf is spanned by the truncated SVD of the entity-time matrix, entirely bypassing gradient-based training. We emphasize that this derivation is a theoretical motivation rather than a strict identity. It relies on the d-regular graph assumption, the row-centering hypothesis, and the approximation of $|\mathbf{M}\mathbf{M}^\top|$ by $\mathbf{M}\mathbf{M}^\top$. In practice, our implementation directly computes the SVD of the row-normalized entity-time matrix and never explicitly forms $\mathbf{W}$ (Appendix A).

V LIPSCHITZ STABILITY OF THE LANDSCAPE VECTORIZATION

Theorem V.1 (Lipschitz Stability)**.** *Let c and c' be any two entity populations. Let $\mathbf{h}_c, \mathbf{h}_{c'} \in \mathbb{R}^{125}$ be their respective persistence landscape fingerprints and $\mathcal{D}_c, \mathcal{D}_{c'}$ their persistence diagrams. Then:*

$$\|\mathbf{h}_c - \mathbf{h}_{c'}\|_2 \;\leq\; C \cdot d_B(\mathcal{D}_c, \mathcal{D}_{c'})$$

where $\|\cdot\|_2$ denotes the Euclidean norm, d_B is the bottleneck distance between persistence diagrams, and C depends only on the landscape sampling grid.

Proof sketch. The persistence landscape λ_1 is 1-Lipschitz with respect to the bottleneck distance (Bubenik, 2015). Concatenating the landscapes across H_0, H_1, and H_2 introduces at most an ℓ_2-to-supremum scaling factor of $\sqrt{125}$. □

Scope of this guarantee. Theorem V.1 bounds a single link of the topological pipeline: the vectorization map $\mathcal{D} \mapsto \mathbf{h}$ from persistence diagrams to $\mathbb{R}^{125}$. We emphasize that this is strictly a *robustness* guarantee, not a *performance* guarantee. It ensures the representation does not amplify noise, but the predictive gains reported in Section 4 are established purely empirically.

Furthermore, this stability does not trivially extend end-to-end from the raw time series panel $\mathbf{X}$ to the final fingerprint $\mathbf{h}$. The composite mapping passes through three prior stages, only two of which are continuous:

1. $\mathbf{X} \mapsto D$, the correlation-distance matrix. The Pearson correlation ρ is smooth in $\mathbf{X}$ assuming non-zero variance. This step is therefore locally Lipschitz, though the Lipschitz constant degrades as a series' variance approaches zero.
2. $D \mapsto \mathcal{D}$, the Vietoris-Rips filtration. Rips persistence is 1-Lipschitz in the ℓ_∞ norm on the dissimilarity matrix, yielding $d_B(\mathcal{D}(D), \mathcal{D}(D')) \leq \|D - D'\|_\infty$. Crucially, this bound holds for any arbitrary finite symmetric dissimilarity. Because the triangle inequality is not required, the fact that $d_{ij} = 1 - \rho_{ij}$ is not a strict metric does not invalidate the stability bound (Chazal et al., 2014).
3. The k=50 sparsification, which is **not** continuous. This operation acts as a hard threshold, retaining only the k nearest neighbors for each series. An arbitrarily small perturbation in ρ can alter the neighbor ordering, inserting or deleting an edge and shifting an entry in D between a finite distance and $+\infty$. At these rank boundaries, $\|D - D'\|_\infty$ diverges, allowing H_1 and H_2 features to jump discretely.

Consequently, the composite mapping is *piecewise* Lipschitz. It remains stable within a fixed neighbor ordering but permits jumps across rank boundaries, which is why Theorem V.1 is formally restricted to the vectorization step. The empirical insensitivity of the fingerprint to upstream perturbations is demonstrated in Appendix G, where substituting Pearson correlation with Spearman, Euclidean, or distance-correlation metrics yields downstream performance within the margin of random seed noise. We treat this end-to-end robustness as an empirical property of the data rather than a theoretical guarantee.

This stability result has three direct consequences for TopoPrimer.

1. **Noise robustness.** Once the persistence diagram shifts by d_B, the resulting fingerprint shifts by at most $C \cdot d_B$, ensuring the vectorization step introduces no signal amplification. Whether missing observations or measurement errors produce a small initial d_B depends on the discontinuous sparsification step, which we validate empirically (Appendix G).
2. **Cold-start coverage.** At launch, a novel item lacks individual history, yet the population fingerprint is already well-defined. Because it is computed offline from the prior population, it requires no per-item history and is available at the first forecast step. We do not recompute the fingerprint as the population absorbs the new item; the frozen descriptor carries it through its initial weeks (Section 4.3). While adding a series alters the k-nearest-neighbor graph (and thus falls outside the strict bounds of Theorem V.1), adding one series to a population of N changes at most k+1 rows of the sparsified matrix. The resulting perturbation is typically minimal, though we state this as a practical expectation rather than a formal corollary.
3. **Backbone agnostic.** The TDA fingerprint is precomputed once from the population and frozen. Because its ℓ_2 norm is fixed at computation time and never updated during training

or inference, it introduces no gradient-driven drift. This allows it to propagate as a stable conditioning signal regardless of the downstream model architecture.

W Cellular Sheaf Motivation for the Spectral Relational Coordinate

When restricted to identity maps, the sheaf Laplacian coincides with the ordinary weighted graph Laplacian. Its minimal-Dirichlet-energy subspace is spanned by the leading left singular vectors of the entity-time matrix. On a centered matrix, this operation is exactly principal component analysis (PCA). Because the resulting coordinate is a closed-form spectral embedding, we refer to it as "relational" rather than "sheaf" in the main text. We explicitly avoid the general sheaf formalism that requires learning distinct per-incidence restriction maps $R_e^u \neq R_e^v$. The learned neural encoder (Appendix H) shares a single map across all incidences, rendering it a projected graph-Laplacian regularizer rather than a true cellular sheaf network. Furthermore, the closed-form spectral coordinate matched or outperformed this neural encoder at a fraction of the computational cost on M5, the dataset where both were evaluated (Household, Chronos: 1.0251 vs. 1.0343 MAE).

Prior sheaf-based methods compute similar coordinates using learned sheaf Laplacians and graph convolutions (Bodnar et al., 2022; Mostafa et al., 2026). Similarly, relational-graph forecasters learn comparable structures using graph convolutions without the sheaf formalism (Li et al., 2018; Wu et al., 2019). Crucially, both approaches fundamentally replace or structurally augment the forecasting backbone. We instead adopt the closed-form spectral solution. The coordinate of series i is row i of U, the left factor of a truncated singular value decomposition (SVD) $\mathbf{X} \approx U\Sigma V^\top$ of the entity-time matrix $\mathbf{X} \in \mathbb{R}^{N\times T}$. This yields a 256-dimensional per-series vector computed once per domain (see motivating derivation in Appendix U.6). Because no parameters are trained and no graph convolutions are applied, the resulting signal remains entirely backbone-agnostic. This design choice also highlights the structural divergence between our two descriptors. The TDA fingerprint is *global* (one shared vector per population), whereas the relational coordinate is *per-series*, locating each individual series on the shared demand manifold.

Spectral coordinates serve as the default wherever a relational coordinate is reported. The learned neural encoder is evaluated exclusively on M5 (Appendix H), as it is the only dataset where both variants were tested across all backbone families. Finally, the internal corpus evaluations do not utilize a relational arm, relying solely on the Vanilla and +TDA configurations (Section 4.3, App. J). Therefore, neither encoder factors into the internal corpus results.

X Extended Related Work

Topological deep learning and TDA for time series. Topological deep learning (TDL) (Papamarkou et al., 2024) explicitly incorporates the topological structure of the underlying data space into neural architectures. Within time series forecasting, prior work applies persistent homology (Carlsson, 2009; Edelsbrunner and Harer, 2010) to sliding-window embeddings of individual series (Zeng et al., 2021; Lin et al., 2026; 2025; Kim et al., 2026). While these methods successfully capture within-series temporal dynamics (e.g., periodicity and local shape), they generate isolated descriptors for each window. Consequently, the overarching geometry of the series population remains unmodeled. TopoPrimer instead applies persistent homology directly to the cross-series correlation manifold, producing a single, shared fingerprint for the entire domain. This shift from per-series temporal topology to population-level relational topology constitutes our primary methodological departure from prior TDA forecasting literature.

Graph and relational forecasting. Graph-based forecasters, such as DCRNN (Li et al., 2018), Graph WaveNet (Wu et al., 2019), and MTGNN (Wu et al., 2020), learn directed or adaptive adjacencies over fixed entity graphs. These approaches fundamentally replace or heavily augment the forecasting backbone for each new domain. Conversely, standard Transformer-based models (Zhou et al., 2021; Nie et al., 2023; Lim et al., 2021) sidestep relational structure entirely by encoding each series independently. Global-factor models (Wang et al., 2019) learn a low-rank factorization jointly with the forecasting objective. While these latent coordinates are conceptually similar to ours, they are learned end-to-end rather than derived as closed-form, frozen priors. Unlike these existing

paradigms, TopoPrimer neither replaces nor modifies the underlying backbone. Instead, it injects population topology as a precomputed context that any standard forecaster can consume without architectural modification.

Cellular sheaf methods. Cellular sheaf theory (Curry, 2014; Hansen and Ghrist, 2021) generalizes graph convolutions by assigning restriction maps to node-edge incidences, capturing relational structures that standard shared-weight message-passing cannot represent. For instance, Bodnar et al. (Bodnar et al., 2022) learn distinct per-incidence restriction maps on heterophilic graphs, while ST-Sheaf GNN (Mostafa et al., 2026) applies diagonal maps for spatio-temporal forecasting. In both cases, the sheaf network serves as the end-to-end model and remains inherently local: each node's representation is shaped solely by its immediate neighbors, lacking a global perspective of its position within the broader population. TopoPrimer instead derives each series' coordinate in closed form from the leading left singular vectors of the entity-time matrix, requiring no gradient-based training. To our knowledge, deriving spectral sheaf coordinates as a frozen, backbone-agnostic prior for time series forecasting remains unexplored in prior literature.

Time series foundation models. Time series foundation models (TSFMs), such as Chronos 2.0 (Ansari et al., 2025) and TimesFM 2.5 (Das et al., 2024), are pre-trained on massive corpora to generalize across domains. When downstream adaptation is required, these models are typically updated via fine-tuning on individual series histories. This standard adaptation paradigm fails to introduce explicit population-topology signals. TopoPrimer addresses this gap by injecting precomputed topological features as a frozen context vector, enhancing the model without altering the pre-trained backbone weights.

Cross-variate foundation models. Not all foundation models are channel-independent. MOIRAI (Woo et al., 2024) introduces an any-variate attention mechanism designed to jointly model arbitrary cross-series and multivariate dependencies. Similarly, UniTS (Gao et al., 2024) utilizes multi-variable cross-attention and dynamic operators across channels, while Tiny Time Mixers (Ekambaram et al., 2024) apply multi-level cross-channel mixing in the decoder during transfer and adaptation. Because these models learn cross-series dependence end-to-end, they require series to be encoded together at inference. Consequently, their computational cost scales with the number of jointly attended series. TopoPrimer makes a fundamentally different trade-off. The relational coordinate is computed once offline in closed form. For example, on the M5 dataset, 30 block-wise SVDs complete in under one second on a single CPU core (Appendix H). This coordinate is then injected as a frozen 256-dimensional vector per series. As a result, the backbone performs no cross-series computation at inference and never directly observes the full population. These two paradigms are complementary rather than competing, and we make no comparative accuracy claims. Our reported gains are explicitly scoped to backbones lacking joint cross-series access, and we do not evaluate against natively multivariate architectures.

Y ASSET LICENSES

Datasets.

- **ECL** (Electricity Consuming Load, UCI repository): CC BY 4.0. https://creativecommons.org/licenses/by/4.0/
- **Monash Weather** (Godahewa et al., NeurIPS 2021; Zenodo DOI: 10.5281/zenodo.4654822): CC BY 4.0. https://creativecommons.org/licenses/by/4.0/
- **M5** (Kaggle M5 Accuracy competition): subject to Kaggle competition terms of use. https://www.kaggle.com/competitions/m5-forecasting-accuracy/rules
- **METR-LA** (Li et al., 2018): MIT License.

Models and libraries.

- **Chronos** (Amazon): Apache 2.0.
- **TimesFM** (Google): Apache 2.0.
- **XGBoost**: Apache 2.0.
- **PatchTST**: Apache 2.0.
- **DLinear**: MIT License.
- **NLinear**: MIT License.

All assets are used in accordance with their respective licenses.