Disentangling Dynamics: A Compositional Framework for Temporal Graph Foundation Models

Fabrizio De Castelli [1] **Alessio Gravina** [1] **Davide Bacciu** [1] **Moshe Eliasof** [2]

Abstract

Learning general representations for temporal graphs is challenging because temporal dynamics and structural changes are often intertwined, causing models to capture domain-specific patterns over generalizable ones. This acts as a barrier to the development of foundation models for temporal graphs. Current methods use sequential models alongside Graph Neural Networks (GNNs), producing entangled spatio-temporal representations that struggle to transfer to new graphs. In this work, we hypothesize that this limitation is in part structural. To address this, we propose TIDES, a compositional framework that systematically decouples time from space during learning, serving as a stepping stone toward foundation models for temporal graphs. TIDES relies on a one-time, parallel extraction process to create three domain-agnostic priors: (i) temporal embeddings from a frozen time-series foundation model; (ii) deterministic summaries of local temporal contexts; and (iii) supra-Laplacian positional encodings of the changing graph structure. These features are computed once and used as node inputs for a GNN. TIDES achieves strong performance on multiple benchmarks, reduces negative transfer, and improves cross-graph transferability, outperforming state-of-the-art temporal graph baselines on 20 unseen networks in the MiNT zero-shot benchmark. Overall, TIDES offers a systematic approach to Temporal Graph Foundation Models by diagnosing current issues and providing scalable, transferable representation strategies.

[1]Department of Computer Science, University of Pisa, Pisa, Italy [2]Faculty of Computer and Information Science, Ben-Gurion University of the Negev, Israel. Correspondence to: Fabrizio De Castelli <fabrizio.decastelli@phd.unipi.it>.

The Workshop on Graph Foundation Models at the 43rd International Conference on Machine Learning (ICML 2026), Seoul, South Korea.

1. Introduction

Many real-world systems are inherently dynamic, characterized by data that evolve in both structure and attributes over time (Kazemi et al., 2020; Longa et al., 2023; Gravina & Bacciu, 2024; Jin et al., 2024). Modeling these systems requires extending static graph learning to dynamic settings. The unprecedented success of Foundation Models (FMs) in natural language (Brown et al., 2020; Ouyang et al., 2022) and computer vision (He et al., 2021; Kirillov et al., 2023) has naturally inspired a growing interest in developing Graph Foundation Models (GFMs) (Wang et al., 2025a), primarily focused on static graphs. In the context of Temporal GFMs, existing methods broadly focus on coupling spatial and temporal processing. This is typically achieved by interleaving sequence models with Graph Neural Networks (GNNs) (Seo et al., 2018; Li et al., 2018; Zhao et al., 2020; Bai et al., 2021; Cini et al., 2023), or by integrating a GNN inside a Recurrent Neural Network (RNN) (Chen et al., 2022; Li et al., 2019; Micheli & Tortorella, 2022). While effective on single, specific datasets, scaling these coupled to foundation levels from scratch poses strong generalization challenges and limits their practical deployment.

Fundamental issues extensively studied in the pursuit of foundation models, such as cross-domain generalization (Wang et al., 2024) and dataset heterogeneity (Wang et al., 2025b; Liu et al., 2026; Eliasof et al., 2026), are further enhanced in the context of temporal graph learning: the sheer diversity of topological structures combined with significant distributional shifts in temporal dynamics, have been shown to lead to catastrophic negative transfer (Yuan et al., 2025a; Zi et al., 2025). That is, when standard models attempt to jointly learn across diverse domains, conflicting spatio-temporal patterns can lead to negative transfer and degraded performance under heterogeneous domain shifts (Li et al., 2022). We argue that this failure is structural: when sequence models and GNNs are trained jointly, every shared parameter is exposed to conflicting spatio-temporal patterns and is forced to encode domain identity to fit them, producing a leakage channel that scales poorly with corpus diversity. Recent transfer-learning analyses on dynamic interaction networks corroborate this view, showing that memory-based architectures fail catastrophically on entirely unseen graphs because their recurrent components mem-

orize train-time node identities (Chatterjee et al., 2025). Therefore, addressing this gap is essential; the inability to effectively share parameters without interference likely limits the progress toward scalable, multi-domain temporal GFMs.

To move towards mitigating negative transfer, we introduce **TIDES** (**T**emporal **I**njection of **D**ecoupled **E**ncodings into **S**tatic GNNs), a novel, memory-free modular framework that serves as a highly scalable stepping stone towards TGFMs. Instead of relying on a monolithic architecture trained from scratch (Li et al., 2025; Shamsi et al., 2026; Yuan et al., 2025b), our TIDES systematically decouples temporal and spatial dynamics. It leverages the robust priors of pre-trained Time Series FMs (Goswami et al., 2024; Ansari et al., 2024), augmented with local temporal statistics, and structural and positional encodings (PE) (Galron et al., 2026a) of the topology. These features are treated as fixed inputs to a GNN, which is trained to model spatial dependencies. In this way, we prevent shared parameters from encoding domain-specific patterns.

Our main contributions are as follows:

- **Decoupled Time-then-Space:** We propose TIDES, a novel modular framework that fully decouples temporal and spatial graph dynamics. By synergistically combining pre-computed temporal statistics, positional encodings, and a simple static GNN, our architecture achieves powerful temporal representation learning without the computationally prohibitive need to train temporal sequence models from scratch.
- **Zero-Shot Transfer with Reduced Negative Transfer:** On the MiNT zero-shot benchmark (Shamsi et al., 2026), we substantially reduce the incidence of negative transfer relative to memory-based pretraining baselines, while reaching competitive performance with as few as four pretraining networks.

2. Related Work

Spatio-Temporal Foundation Models. Recent advancements in spatio-temporal forecasting have sought to generalize predictions across unseen geographic regions. OpenCity (Li et al., 2025) achieves zero-shot transfer for urban traffic forecasting by leveraging Reversible Instance Normalization (Kim et al., 2022) over traffic regions and temporal patch embeddings. Similarly, frameworks such as ST-LLM (Liu et al., 2024) and UniFlow (Yuan et al., 2025b) align raster and graph-based spatio-temporal data by leveraging the cognitive routing of Large Language Models (LLMs) (Brown et al., 2020) or shared memory banks for urban patterns. However, these architectures are intrinsically tied to specific geographic grids and regression tasks, and they assume a time-invariant underlying topology. In contrast, TIDES operates on general temporal graphs with evolving topology and is not restricted to regression-based urban forecasting tasks.

Transfer Learning on Dynamic Interaction Networks. In contrast to spatio-temporal forecasting, temporal graph learning on pure interaction networks has historically struggled with transferability due to a fundamental memory bottleneck. Standard dynamic graph architectures, (Gravina & Bacciu, 2024), rely on recurrent modules (Rumelhart et al., 1986; Cho et al., 2014; Hochreiter & Schmidhuber, 1997) that memorize train-time node identities, leading to catastrophic failure on entirely unseen graphs (Chatterjee et al., 2025). Recent efforts have attempted to bypass this limitation through multi-network pre-training. MiNT (Shamsi et al., 2026) demonstrates zero-shot transfer across temporal transaction networks by shuffling input orders and context-switching across 64 datasets to prevent identity leakage. In the domain of Temporal Knowledge Graphs (TKGs), POSTRA (Pan et al., 2025) achieves inductive transfer by abandoning absolute timestamps in favor of relative temporal signals and purely topological relation embeddings. These models mitigate the node-identity problem within their respective domains, but remain bounded by a shared task interface and, as we show in Section 4.3, still exhibit negative transfer on a substantial fraction of unseen networks. Our framework, instead, relies on the computation of strong temporal and structural priors that provide a domain-agnostic representation of the input, without explicitly encoding node identities or dataset-specific memory.

Time-Series Foundation Models as Temporal Priors. Recent time-series research has rapidly consolidated around universal foundation models. Architectures such as MomentFM (Goswami et al., 2024), Chronos (Ansari et al., 2024), and PatchTST (Nie et al., 2023) demonstrate that massive pre-training on diverse sequential data yields robust, domain-agnostic representations of temporal dynamics. By treating time windows as continuous patches or quantizing them into discrete tokens, these models remove the need for dataset-specific temporal resolutions. However, these approaches operate on temporal sequences and do not explicitly model relational structure or interactions between entities. In contrast, TIDES combines their domain-agnostic temporal representations with positional and structural encodings, and leverages a GNN to model information propagation over an evolving graph topology.

3. A Compositional Framework for Temporal Graph Representation Learning

In this section, we introduce **TIDES** (**T**emporal **I**njection of **D**ecoupled **E**ncodings into **S**tatic GNNs), a unified framework for encoding temporal graph data as a composition of structured operators. Notably, rather than training a mono-

lithic spatio-temporal architecture from scratch, our approach decomposes representation learning into complementary components, each introducing a distinct inductive bias. Our approach follows the time-then-graph paradigm (Gao & Ribeiro, 2022), which has proved advantageous for expressivity compared to entangled paradigms such as time-and-graph. In Figure 1, we provide a visual overview of the components of our TIDES.

3.1. Notation and Preliminaries

We focus on Discrete-Time Dynamic Graph (D-TDG) (Kazemi et al., 2020; Gravina & Bacciu, 2024), also known as Snapshot-based Temporal Graphs (Longa et al., 2023), where the graph evolves over time through a sequence of snapshots with evolving node features and, potentially, time-varying topology. We define a D-TDG as a sequence of graph snapshots $\mathcal{G}_{1:T} = \{\mathcal{G}_1, \ldots, \mathcal{G}_T\}$. Each snapshot $\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t, \mathbf{X}_t, \mathbf{E}_t)$ consists of a non-empty set of n nodes $\mathcal{V}$, a time-varying edge set $\mathcal{E}_t \subseteq \mathcal{V} \times \mathcal{V}$, a node feature matrix $\mathbf{X}_t \in \mathbb{R}^{n \times d_n}$, and an edge feature matrix $\mathbf{E}_t \in \mathbb{R}^{|\mathcal{E}_t| \times d_e}$. The structural information expressed by $\mathcal{E}_t$ is encoded in the adjacency matrix $\mathbf{A}_t \in \{0,1\}^{n \times n}$, where $[\mathbf{A}_t]_{uv} = 1$ if $(u, v) \in \mathcal{E}_t$, and 0 otherwise. We note that when $\mathbf{A}_t = \mathbf{A}_1 \, \forall t$ the D-TDG is often called spatio-temporal graph. Finally, we define the degree matrix as $\mathbf{D}_t = \mathbf{A}_t \mathbf{1}$, and the ordered sequence of input signals as $\mathcal{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T) \in \mathbb{R}^{T \times n \times d_n}$.

3.2. Time-Signal-Topology Decoupling

We formulate our framework as a composition of three components, in which temporal signals, structural topology, and positional information are encoded independently and then integrated into a unified representation space. This design is formalized through a representation map $\Phi : \mathbb{R}^{T \times n \times d_n} \times \mathfrak{G}^T \to \mathbb{R}^{T' \times n \times d}$ that produces node-wise features suitable for downstream prediction tasks, where $\mathfrak{G}$ denotes the space of graphs over $\mathcal{V}$ and $T' \leq T$ accounts for context-window boundary effects. We define Φ as the concatenation of three complementary operators:

$$\Phi(\mathcal{X}, \mathcal{G}_{1:T}) \;=\; \big[\, \Phi_{\mathrm{f}}(\mathcal{X}) \;\big\|\; \Phi_{\mathrm{s}}(\mathcal{X}) \;\big\|\; \Phi_{\mathrm{p}}(\mathcal{G}_{1:T}) \,\big], \qquad (1)$$

where Φ_{f} extracts transferable temporal representations from a frozen pre-trained sequence model, Φ_{s} captures low-order statistical structure from local temporal contexts, and Φ_{p} injects positional and structural information from the evolving graph topology. Together, these three components provide complementary temporal, statistical, and structural priors for downstream prediction tasks. This decomposition induces a partition of the representation space

$$\mathbb{R}^d \;=\; \mathcal{H}_{\mathrm{f}} \oplus \mathcal{H}_{\mathrm{s}} \oplus \mathcal{H}_{\mathrm{p}}, \qquad (2)$$

in which each subspace captures a distinct source of variation: $\mathcal{H}_{\mathrm{f}}$ encodes data-driven temporal dependencies, $\mathcal{H}_{\mathrm{s}}$ captures low-order statistical structure of the signal, and $\mathcal{H}_{\mathrm{p}}$ encodes structural priors induced by the graph topology and temporal indexing. Crucially, the parameters of Φ_{f}, Φ_{s}, and Φ_{p} are frozen: they are precomputed once and never updated during downstream training. This design reduces the extent to which shared parameters can encode domain-specific temporal patterns during multi-domain training, which can lead to negative transfer in coupled spatio-temporal architectures.

Temporal signal encoding. The operator Φ_{f} maps the temporal signal to a sequence of node-wise embeddings via a pre-trained sequence model f_θ, with frozen set of parameters θ, applied over a sliding context of length w_f. Formally,

$$\Phi_{\mathrm{f}}(\mathcal{X})_t \;=\; f_\theta\big(\mathcal{X}_{t-w_f+1:t}\big) \;\in\; \mathbb{R}^{n \times d_f}. \qquad (3)$$

Here, f_θ acts independently on each node trajectory, modeling temporal dependencies while ignoring the underlying graph structure. This design lets Φ_{f} leverage transferable, sequence-level representations learned from large-scale pre-training, while remaining agnostic to spatial interactions.

Local statistical summaries. The operator Φ_{s} complements Φ_{f} by extracting a localized representation of node features over a possibly different temporal context w_s. By either directly preserving or aggregating the recent signal, this component introduces a controlled inductive bias that grounds the deep temporal representation of Φ_{f} in the immediate state of the network.

Supra-Laplacian positional encoding. We define Φ_{p} to inject structural and temporal priors via the supra-Laplacian (Karmim et al., 2024) of a sliding topology window. While standard Laplacian positional encodings (Dwivedi & Bresson, 2021) are highly effective for static graphs, computing them independently per snapshot suffers from the temporal misalignment and worse generalization capabilities (Galron et al., 2026a). Therefore, given a window of length w_p ending at time t, we lift the snapshots $\mathcal{G}_{t-w_p+1:t}$ into a multilayer graph (Kivelä et al., 2014) and assemble the supra-Laplacian matrix

$$\mathbf{L}_{\mathrm{supra}} \;=\; \mathbf{D}_{\mathrm{supra}} - \mathbf{A}_{\mathrm{supra}} \;\in\; \mathbb{R}^{nw_p \times nw_p}, \qquad (4)$$

where the supra-adjacency $\mathbf{A}_{\mathrm{supra}}$ is block tridiagonal with diagonal blocks $[\mathbf{A}_{\mathrm{supra}}]_{\tau\tau} = \mathbf{A}_\tau$ for $\tau \in [t - w_p + 1, t]$ and off-diagonal blocks $[\mathbf{A}_{\mathrm{supra}}]_{\tau,\tau\pm1} = \mu \mathbf{I}_n$, with $\mu > 0$ controlling the inter-layer coupling and all other blocks zero. $\mathbf{D}_{\mathrm{supra}}$ is the supra-degree matrix associated with $\mathbf{A}_{\mathrm{supra}}$. The k smallest eigenpairs $\{(\lambda_j, \mathbf{u}_j)\}_{j=1}^{k}$ of $\mathbf{L}_{\mathrm{supra}}$

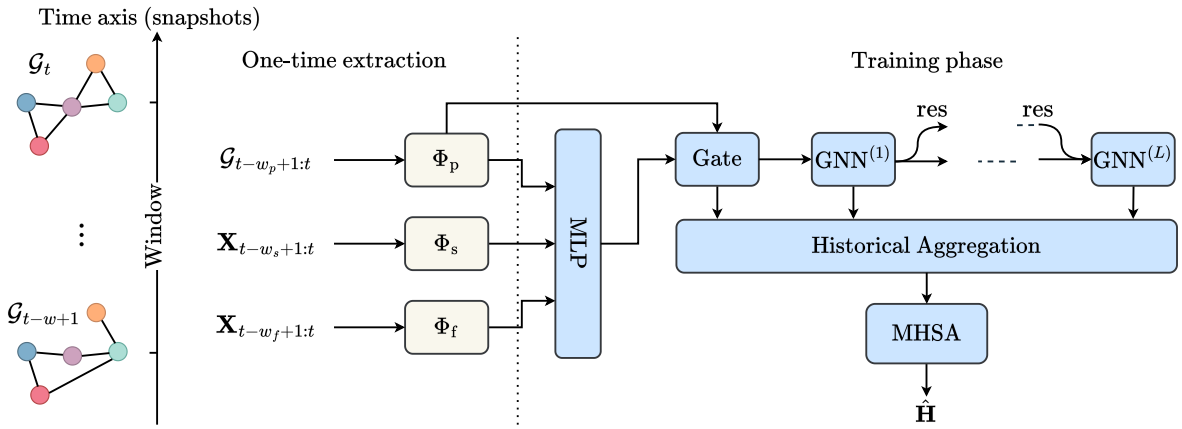

Figure 1. Overall architecture of TIDES. On the left, a high-level block architecture showing components across space and time. On the right, the trainable architecture, which consists of the projections, L GNN layers, and historical aggregations of latent states.

minimize the variational objective

$$\begin{aligned} \min_{\{\mathbf{U}_\tau\}_\tau} \quad & \sum_{\tau=t-w_p+1}^{t} \operatorname{tr}\big(\mathbf{U}_\tau^\top \mathbf{L}_\tau \mathbf{U}_\tau\big) \\ & + \mu \sum_{\tau=t-w_p+2}^{t} \|\mathbf{U}_\tau - \mathbf{U}_{\tau-1}\|_F^2 \\ \text{s.t.} \quad & \mathbf{U}^\top \mathbf{U} = \mathbf{I}, \end{aligned} \tag{5}$$

where $\mathbf{U} = [\mathbf{U}_{t-w_p+1}^\top, \dots, \mathbf{U}_t^\top]^\top \in \mathbb{R}^{nw_p \times k}$ stacks the per-snapshot blocks and $\mathbf{L}_\tau$ denotes the Laplacian of $\mathcal{G}_\tau$. The first term in (5) enforces intra-layer structural smoothness through the Laplacian quadratic form, while the second term enforces inter-layer consistency, jointly producing eigenmodes that span the spatio-temporal window smoothly and admit consistent sign assignments across snapshots (Galron et al., 2026a). We define the positional encoding at target time t by extracting the time-t slice of the eigenvectors and concatenating the corresponding eigenvalues:

$$\Phi_{\mathrm{p}}(\mathcal{G}_{t-w_p+1:t}) = \big[\, \mathbf{U}_t \,\big\|\, \mathbf{1}_n \boldsymbol{\lambda}^\top \,\big] \in \mathbb{R}^{n \times 2k}, \tag{6}$$

where $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_k)^\top \in \mathbb{R}^k$ collects the corresponding eigenvalues and $\mathbf{1}_n \in \mathbb{R}^n$ broadcasts them across nodes. Augmenting the eigenvectors with their associated eigenvalues provides the model with explicit access to the spectrum of $\mathbf{L}_{\mathrm{supra}}$, allowing it to weight modes by their frequency content rather than treating all eigenvectors uniformly. To control the cubic cost of dense eigendecomposition, we approximate the k smallest eigenpairs via the Locally Optimal Block Preconditioned Conjugate Gradient (LOBPCG) method (Knyazev, 2001). By providing the model with a strong inductive bias on the local geometry of the evolving graph, Φ_{p} explicitly grounds the data-driven temporal summaries Φ_{f} and Φ_{s} within a structural context.

Trajectory-Aware Layer Aggregation. To instantiate our decoupled blueprint, we design a static architecture that resolves spatial dependencies while explicitly leveraging the precomputed temporal priors. The three operator outputs $\{\Phi_{\mathrm{f}}, \Phi_{\mathrm{s}}, \Phi_{\mathrm{p}}\}$ are projected into a shared d-dimensional latent space, additively combined, and passed through a non-linear activation gated by the positional encodings, adaptively modulating the structural bias to the topology of the temporal window. Because sequence modeling is handled in a one-time preprocessing stage, the GNN that models spatial dependencies can be scaled to L layers with residual connections at a reduced training cost compared to recurrent spatio-temporal architectures. This expands the spatial receptive field without the memory overhead of recurrent backpropagation.

The interaction between residual connections and depth-wise readout is central to our design. A residual GNN propagates messages as $\mathbf{H}_l = \mathbf{H}_{l-1} + f_l(\mathbf{H}_{l-1})$, so that the trajectory $\{\mathbf{H}_0, \dots, \mathbf{H}_L\}$ accumulates contributions monotonically: each $\mathbf{H}_l$ encodes the receptive field of an l-hop neighborhood while preserving the information available at every shallower scale. The standard approach reading out the final hidden state $\mathbf{H}_L$ alone collapses this multi-scale signal into a single embedding and discards information that residuals were specifically designed to preserve. We instead treat the residual trajectory as the model's primary representation and aggregate across it, following the historical activation pooling of (Galron et al., 2026b).

The full residual trajectory $\mathbf{H} = (\mathbf{H}_0, \mathbf{H}_1, \dots, \mathbf{H}_L) \in \mathbb{R}^{n \times L+1 \times d}$ enters the historical aggregation module of (Galron et al., 2026b) as a single multi-scale tensor: rather than reading out the final hidden state alone, all L activations are jointly attended to. Sinusoidal positional encodings $\mathbf{P}$ are added along the layer axis to make depth ordering explicit, $\tilde{\mathbf{H}} = \mathbf{H} + \mathbf{P}$, and a single attention head treats the

final-layer state as the query while keys and values span the entire trajectory:

$$\begin{aligned}\mathbf{Q} &= \tilde{\mathbf{H}}_{L-1}\mathbf{W}^Q \in \mathbb{R}^{n\times 1\times d},\\ \mathbf{K} &= \tilde{\mathbf{H}}\mathbf{W}^K \in \mathbb{R}^{n\times L\times d},\\ \mathbf{V} &= \tilde{\mathbf{H}} \in \mathbb{R}^{n\times L\times d}.\end{aligned} \tag{7}$$

Scaled dot-product scores are then averaged across nodes, producing a single weighting vector

$$c_l \;=\; \frac{1}{n}\sum_{u\in\mathcal{V}}^{n} \frac{\mathbf{Q}_u^\top \mathbf{K}_{u,l}}{\sqrt{d}}, \qquad l \in \{0,\dots,L-1\}, \tag{8}$$

that summarizes how much each depth scale contributes to the final-layer view of the graph. Following (Galron et al., 2026b), the weights are normalized as $\alpha_l = c_l / \sum_{l'} c_{l'}$ rather than via softmax, since the latter forces the aggregation to be a convex combination of layer activations, while signed normalization permits negative weights and recovers the ability to take differences between layers. The pairing with residual connections highlights its usefulness: since $\mathbf{H}_l - \mathbf{H}_{l-1} = f_l(\mathbf{H}_{l-1})$ isolates the contribution of layer l, signed weights can subtract layers and surface intermediate-scale structure that no individual depth expresses on its own. The cross-layer pooled representation

$$\mathbf{Z} \;=\; \sum_{l=0}^{L} \alpha_l\, \tilde{\mathbf{H}}_l \;\in\; \mathbb{R}^{n\times d} \tag{9}$$

is then refined by node-wise Multi-Head Self-Attention (MHSA) (Vaswani et al., 2017) to restore inter-node mixing discarded by the layer-wise pooling, $\mathbf{H} = \mathrm{MHSA}(\mathbf{Z},\mathbf{Z},\mathbf{Z})$. Coupling skip connections with trajectory-aware aggregation thus converts a deep GNN into a multi-scale spatial encoder.

Implementation Details. In our experiments in Section 4, we instantiate the components of our framework as follows. *(i)* We set Φ_{f} to the pre-trained model `Moment-1-small` from the family of open time series foundation models proposed in (Goswami et al., 2024). The context window length w_f is selected according to the task. The model contains approximately 40M parameters and is used without fine-tuning. *(ii)* We define Φ_{s} to be the identity mapping (i.e., $\Phi_{\mathrm{s}}(\mathcal{X})_t = \mathcal{X}_{t-w_s+1:t}$), or a fixed-dimensional representation obtained via adaptive average pooling over the temporal and channel dimensions. The choice depends on the task and enforces a controlled inductive bias toward low-variance components of the signal. *(iii)* For spatio-temporal data, we simulate mild topology variations by randomly perturbing the graph structure at each training step, inserting and removing up to 5% of edges. This stochastic augmentation is intended to improve robustness to temporal variability in the underlying graph. *(iv)* As the trainable backbone, we employ a Graph Convolutional Network (GCN) (Kipf & Welling, 2017) with L layers, residual (skip) connections, and the trajectory-aware aggregation mechanism described in the previous paragraph.

Complexity Analysis. The computational cost of TIDES is dominated by two distinct phases that fundamentally separate the heavy temporal sequence modeling from the spatial optimization.

Preprocessing. This one-time preprocessing step handles the temporal and structural priors independently. Extracting the temporal embeddings requires passing n independent sequences of length w_f through the time-series FM, thereby inheriting its computational complexity. In the case of MomentFM (as used in our implementation), this corresponds to the standard Transformer inference complexity (Goswami et al., 2024). Extracting the supra-Laplacian PE requires the k smallest eigenpairs of the block matrix $\mathbf{L}_{\text{supra}}$; with LOBPCG, the cost is $\mathcal{O}(knw_p)$ per temporal window (Galron et al., 2026a). Crucially, both operations are gradient-free and can be parallelized across temporal windows, with the temporal embeddings parallelizable over nodes as well.

Training phase. During training, the temporal dimension is entirely factored out, circumventing the memory-intensive Backpropagation Through Time (BPTT) of standard sequence models (Rumelhart et al., 1986). The L-layer GNN with residual connections propagates messages at the standard cost $\mathcal{O}(L(|\mathcal{E}|+n))$ (when message-passing based GNNs (Gilmer et al., 2017) are considered), and the historical activation pooling adds $\mathcal{O}(n^2)$ (Galron et al., 2026a). The total per-step training complexity is therefore $\mathcal{O}(L(|\mathcal{E}|+n)+n^2)$. In the typical regime $L \ll n$, this is dominated by the spatial attention term $\mathcal{O}(n^2)$. In contrast, end-to-end temporal models incur an additional multiplicative factor of the sequence length T.

4. Experiments

We empirically assess TIDES across two settings: *(i)* **Supervised Forecasting**, where we perform full training of the proposed architecture (i.e., the training only of the GNN) to validate its performance under isolated data distributions; and *(ii)* **Zero-Shot Transfer**, pre-training TIDES on 64 token networks of the MiNT benchmark (Shamsi et al., 2026) and evaluating it on 20 unseen networks. Implementation details, baselines, and dataset statistics are reported in Appendix A. All experiments were performed on NVIDIA A100-SXM4-40GB GPUs.

4.1. Supervised Node-Level Forecasting

We evaluate TIDES on 7 popular benchmarks spanning web activity, public health, social interactions, and traffic forecasting: Wikipedia Math, ChickenPox Hun-

Table 1. Component ablation on Wikipedia Math, Twitter Tennis, ChickenPox Hungary (10 seeds, MSE) and Pems07 (5 seeds, MaskedMAE). TIDES combines all three priors $\{\Phi_f, \Phi_s, \Phi_p\}$. `no_X` variants drop one prior; `only_X` variants retain only one. When Φ_p is removed, the PE-conditioned gate is bypassed. Lower is better; **best** and **second-best** per column are highlighted.

Configuration	Wiki Math MSE ↓	Twitter Tennis MSE ↓	ChickenPox MSE ↓	Pems07 MaskedMAE ↓
TIDES (full)	**0.440 ± 0.002**	**0.262 ± 0.002**	**0.745 ± 0.003**	**19.767 ± 0.055**
Leave-one-out				
w/o Φ_f (`no_temporal_fm`)	**0.439 ± 0.004**	0.293 ± 0.007	**0.839 ± 0.029**	21.043 ± 0.087
w/o Φ_s (`no_stats`)	0.669 ± 0.019	0.300 ± 0.003	1.006 ± 0.029	21.202 ± 0.144
w/o Φ_p (`no_pe`)	0.452 ± 0.002	**0.269 ± 0.003**	0.845 ± 0.026	**19.962 ± 0.036**
Single-prior				
only Φ_f (`only_temporal_fm`)	0.674 ± 0.016	0.287 ± 0.006	0.971 ± 0.014	21.243 ± 0.028
only Φ_s (`only_stats`)	0.471 ± 0.007	0.286 ± 0.001	0.927 ± 0.012	22.483 ± 0.075
only Φ_p (`only_pe`)	0.775 ± 0.007	0.332 ± 0.010	1.098 ± 0.001	45.098 ± 0.002

Table 2. Test MSE ↓ on Wikipedia Math, ChickenPox Hungary, and Twitter Tennis, averaged over 10 runs. **First**, **second**, and **third** best results are color-coded.

Model	Wiki Math	Twitter Tennis	ChickenPox
DCRNN	$0.679_{\pm 0.007}$	$0.478_{\pm 0.004}$	$1.097_{\pm 0.006}$
GCRN-GRU	$0.680_{\pm 0.021}$	$0.477_{\pm 0.007}$	$1.103_{\pm 0.004}$
GCRN-LSTM	$0.678_{\pm 0.008}$	$0.477_{\pm 0.006}$	$1.097_{\pm 0.006}$
GC-LSTM	$0.677_{\pm 0.009}$	$0.475_{\pm 0.010}$	$1.095_{\pm 0.005}$
DyGrAE	$0.621_{\pm 0.012}$	$0.480_{\pm 0.005}$	$1.102_{\pm 0.013}$
EvolveGCN-H	$0.779_{\pm 0.031}$	$0.481_{\pm 0.003}$	$1.137_{\pm 0.026}$
EvolveGCN-O	$0.807_{\pm 0.047}$	$0.484_{\pm 0.002}$	$1.135_{\pm 0.011}$
A3T-GCN	$0.618_{\pm 0.008}$	$0.477_{\pm 0.005}$	$1.078_{\pm 0.009}$
T-GCN	$0.616_{\pm 0.011}$	$0.478_{\pm 0.004}$	$1.083_{\pm 0.011}$
MPNN LSTM	$0.856_{\pm 0.021}$	$0.482_{\pm 0.001}$	$1.125_{\pm 0.005}$
DynGESN	$\mathbf{0.610}_{\pm \mathbf{0.003}}$	$\mathbf{0.300}_{\pm \mathbf{0.003}}$	$\mathbf{0.907}_{\pm \mathbf{0.007}}$
HMM4G	$\mathbf{0.542}_{\pm \mathbf{0.008}}$	$\mathbf{0.333}_{\pm \mathbf{0.004}}$	$\mathbf{0.939}_{\pm \mathbf{0.013}}$
TIDES	$\mathbf{0.440}_{\pm \mathbf{0.002}}$	$\mathbf{0.262}_{\pm \mathbf{0.002}}$	$\mathbf{0.745}_{\pm \mathbf{0.003}}$

gary, Twitter Tennis (Rozemberczki et al., 2021), and PeMS03/04/07/08 (Song et al., 2020). We adhere to the data splits and protocols of the original works.

Results. Tables 2 and 3 report forecasting performance. TIDES outperforms all baselines on ChickenPox Hungary, Wikipedia Math, and Twitter Tennis, and remains competitive on the four Pems03/04/07/08 traffic forecasting benchmarks. These results show that the trainable part benefits from the pre-computed temporal and structural priors in these real-world scenarios.

4.2. Component Ablation

We evaluate six ablated variants of TIDES on the datasets of Table 2 and on PeMS07. For each variant, we fix the optimal hyperparameters and drop one or two of the three priors $\{\Phi_f, \Phi_s, \Phi_p\}$, to isolate their contributions. Table 1 shows that *(i)* All three priors are complementary. Removing any single component degrades performance on nearly every fixed-topology benchmark, the sole exception being Φ_f on Wikipedia Math, where removal yields a statistically indistinguishable result (0.439 vs. 0.440). Apart from this case, the full TIDES achieves the lowest error. The `only_pe` variant collapses everywhere, confirming the importance of these encodings in node-level tasks. *(ii)* Local statistics carry the most signal. Removing Φ_s causes the largest performance drop among the leave-one-out variants across all four datasets. We attribute this to the need of providing real inputs to make predictions aware of the local input data scale and variance. *(iii)* Positional encodings are needed. The ablation demonstrates that removing Φ_p degrades performance across all datasets, confirming its necessity regardless of whether the underlying graph is static or time-varying. Interestingly, the most substantial performance degradation in the leave-one-out setting for Φ_p occurs on the static ChickenPox dataset, rather than the dynamically evolving Twitter Tennis network (D-TDG).

4.3. Zero-Shot Transfer on MiNT

We evaluate TIDES on the 84 Ethereum transaction networks proposed in the MiNT benchmark (Shamsi et al., 2026). The task is binary graph property prediction at weekly resolution: forecasting whether overall network growth (transaction volume) will increase or decrease over the next week.

We consider two variants. **TIDES-sup** is trained in a fully supervised fashion on each test network independently using a shallow GCN without employing the layer aggregation of Equation (9), serving as our single-network reference (per network performance is discussed in Appendix B.1). **TIDES-64** shares the same backbone but is pre-trained on the maximum data package of 64 token networks, mirroring the MiNT-64 protocol. During pre-training we uniformly sample a token ID and feed the model a batch of that token's graphs with shuffled snapshots (i.e., with no strict sequentiality), then evaluate directly on the 20 unseen test networks without target-specific training.

Table 3. Performance Comparison with Baseline Models on PEMS03, 04, 07, 08. **First**, **second**, and **third** best results are color-coded. Baselines results are reported from (Gao et al., 2024).

Model	PEMS03			PEMS04			PEMS07			PEMS08		
	MAE ↓	RMSE ↓	MAPE ↓	MAE ↓	RMSE ↓	MAPE ↓	MAE ↓	RMSE ↓	MAPE ↓	MAE ↓	RMSE ↓	MAPE ↓
TCN	19.31	33.24	19.86%	31.11	37.25	15.48%	32.68	42.23	14.22%	22.69	35.79	14.04%
Transformer	17.50	30.24	16.80%	23.83	37.19	15.57%	26.80	42.95	12.11%	18.52	28.68	13.66%
DCRNN	18.18	30.31	18.91%	24.70	38.12	17.12%	25.30	38.58	11.66%	17.86	27.83	11.45%
STGCN	17.49	30.12	17.15%	22.70	35.55	14.59%	25.38	38.78	11.08%	18.02	27.83	11.40%
ASTGCN	17.69	29.66	19.40%	22.93	35.22	16.56%	28.05	42.57	13.92%	18.61	28.16	13.08%
GWNet	19.85	32.94	19.31%	25.45	39.70	17.29%	26.85	42.78	12.12%	19.13	31.05	12.68%
STSGCN	17.48	29.21	16.78%	21.19	33.65	13.90%	24.26	39.03	10.21%	17.13	26.80	10.96%
STFGNN	16.77	28.34	16.30%	19.83	31.88	13.02%	22.07	35.80	9.21%	16.64	26.22	10.60%
STGODE	16.50	27.84	16.69%	20.84	32.82	13.77%	22.99	37.54	10.14%	16.81	25.97	10.62%
DSTAGNN	15.57	27.21	14.68%	19.30	31.46	12.70%	21.42	34.51	9.01%	15.67	24.77	9.94%
ST-WA	15.17	26.63	15.83%	**19.06**	31.02	**12.52%**	20.74	34.05	8.77%	15.41	**24.62**	9.94%
ASTGNN	15.07	26.88	15.80%	19.26	31.16	12.65%	22.23	35.95	9.25%	15.98	25.67	9.97%
EnhanceNet	16.05	28.33	15.83%	20.44	32.37	13.58%	21.87	35.57	9.13%	16.33	25.46	10.39%
AGCRN	16.06	28.49	15.85%	19.83	32.26	12.97%	21.29	35.12	8.97%	15.95	25.22	10.09%
Z-GCNETs	16.64	28.15	16.39%	19.50	31.61	12.78%	21.77	35.17	9.25%	15.76	25.11	10.01%
STNorm	15.32	25.93	**14.37%**	19.21	32.30	13.05%	20.59	34.86	8.61%	15.39	24.80	**9.91%**
STEP	**14.22**	**24.55**	**14.42%**	**18.20**	**29.71**	**12.48%**	**19.32**	**32.19**	**8.12%**	**14.00**	**23.41**	**9.50%**
PDFormer	**14.94**	**25.39**	15.82%	**18.32**	29.97	**12.10%**	**19.83**	32.87	**8.53%**	**13.58**	**23.51**	**9.05%**
TIDES	**14.68**	**22.87**	**14.57%**	19.08	**29.55**	13.00%	**19.68**	**34.45**	**8.30%**	**14.09**	27.51	10.31%

Results. Table 5 reports per-network AUC and average rank, computed as the mean position (1 = best, 4 = worst) of each model across the 20 test networks. TIDES-64 achieves the best average rank (1.45), ranks first on 12 networks and second on 7, and outperforms MiNT-64 despite using identical training data. This is evidence that the decoupled priors provide a robust latent space that the architecture can easily separate. However, the zero-shot transfer setting also exposes distinct dataset-specific sensitivities and trade-offs. The most notable anomaly occurs on the SDEX network, where both TIDES-64 and MiNT-64 suffer a severe negative transfer. This divergence can be attributed to distribution shifts between the source training distribution and the specific structural or temporal dynamics of this network, indicating that the source training data inadvertently biased both models against its diverse structure and evolution.

Quantifying negative transfer. We measure the impact of multi-network pretraining via the per-network transfer gain $\Delta_i = \text{AUC}_i^{\text{multi}} - \text{AUC}_i^{\text{single}}$, where $\Delta_i < 0$ indicates negative transfer. For both TIDES and MiNT we use their respective backbones (i.e., TIDES-sup and HTGN) as baselines, ensuring each model is compared against its respective single-network counterpart. As can be visualized in Figure 2, TIDES-64 achieves a mean gain of 0.107, more than four times that of MiNT-64 (+0.024), and exhibits negative transfer on only 2/20 networks compared to 7 for MiNT (Table 4).

Scaling Behavior with Pretraining Corpus Size. We train TIDES-k on $k \in \{2, 4, 8, 16, 32, 64\}$ token networks and evaluate on 20 unseen test networks, comparing with MiNT-k for $k \in \{4, 8, 16, 32, 64\}$. Figure 3 reports mean test AUC across networks. First, it emerges that TIDES is highly data-efficient. At $k = 4$, it achieves 0.732 mean AUC, comparable to MiNT-64 (0.692) trained on 16× more data, while $k = 2$ already reaches 0.706, indicating strong transfer from minimal pretraining. Second, the performance saturates early: TIDES peaks at $k = 32$ (0.749) and slightly degrades at $k = 64$ (0.729). This suggests that performance is likely governed more by matching distributions between pretraining tokens and target graphs than by corpus size. In contrast, MiNT scales more monotonically (Avg. Rank 4.40 → 2.20 from $k = 4$ to $k = 64$).

Table 4. Aggregate transfer statistics on the 20 MiNT test networks.

Statistic	MiNT	TIDES
Mean $\Delta \uparrow$	+0.024	**+0.107**
Median $\Delta \uparrow$	+0.043	**+0.075**
# networks with $\Delta < 0 \downarrow$	7/20	**2/20**
# networks with $\Delta > +0.1 \uparrow$	5/20	**9/20**

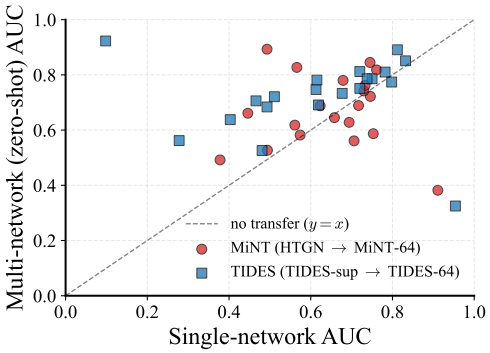

Figure 2. Per-network transfer on 20 MiNT test networks. Each point shows single-network AUC (x-axis) vs. zero-shot AUC (y-axis); the diagonal indicates no transfer.

Table 5. Zero-Shot Transfer Performance (AUC) on MiNT Test Networks, averaged over 3 seeds. HTGN and MiNT-64 are baseline models. TIDES-sup and TIDES-64 are proposed variants. Only **first** and **second** best performance per dataset are highlighted.

Network	HTGN	MiNT-64	TIDES-sup	TIDES-64
MIR	0.745 ± 0.023	**0.845 ± 0.035**	0.720 ± 0.027	**0.812 ± 0.025**
DOGE2.0	0.446 ± 0.164	**0.661 ± 0.047**	0.481 ± 0.051	**0.526 ± 0.011**
MUTE	0.574 ± 0.022	0.582 ± 0.078	**0.613 ± 0.107**	**0.747 ± 0.004**
EVERMOON	0.494 ± 0.127	**0.527 ± 0.118**	0.511 ± 0.141	**0.721 ± 0.063**
DERC	**0.717 ± 0.035**	0.689 ± 0.096	0.615 ± 0.104	**0.781 ± 0.119**
ADX	**0.753 ± 0.013**	0.587 ± 0.014	0.677 ± 0.018	**0.733 ± 0.029**
HOICHI	**0.746 ± 0.010**	0.722 ± 0.034	0.720 ± 0.002	**0.751 ± 0.053**
SDEX	**0.911 ± 0.104**	0.382 ± 0.280	**0.954 ± 0.032**	0.325 ± 0.158
BAG	0.493 ± 0.043	0.893 ± 0.074	0.098 ± 0.036	**0.923 ± 0.079**
XCN	0.566 ± 0.199	**0.827 ± 0.025**	0.466 ± 0.074	**0.706 ± 0.095**
ETH2x-FLI	0.561 ± 0.037	0.618 ± 0.025	**0.832 ± 0.033**	**0.851 ± 0.018**
stkAAVE	0.623 ± 0.077	**0.688 ± 0.019**	0.493 ± 0.104	**0.684 ± 0.021**
GLM	0.761 ± 0.031	**0.818 ± 0.074**	0.812 ± 0.149	**0.891 ± 0.004**
QOM	**0.658 ± 0.150**	0.645 ± 0.109	0.619 ± 0.026	**0.691 ± 0.044**
WOJAK	0.378 ± 0.028	**0.492 ± 0.107**	0.278 ± 0.179	**0.562 ± 0.013**
DINO	**0.706 ± 0.120**	0.561 ± 0.006	0.403 ± 0.021	**0.638 ± 0.203**
Metis	0.679 ± 0.039	0.780 ± 0.041	**0.782 ± 0.014**	**0.810 ± 0.008**
REPv2	0.730 ± 0.007	0.742 ± 0.041	**0.750 ± 0.006**	**0.786 ± 0.021**
TRAC	0.733 ± 0.009	**0.762 ± 0.028**	0.738 ± 0.004	**0.786 ± 0.045**
BEPRO	0.694 ± 0.009	0.628 ± 0.017	**0.798 ± 0.037**	**0.774 ± 0.065**
Avg. Rank ↓	2.45	2.05	2.60	**1.45**

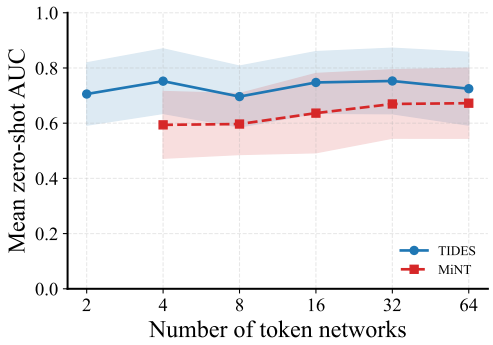

Figure 3. Scaling behavior. Zero-shot mean AUC vs. pretraining size k. TIDES outperforms the largest competitor (i.e., MiNT-64) using only 4 token networks. Error bands show standard deviation across networks.

5. Conclusions

In this work, we introduce TIDES, a modular framework for temporal graph learning that fundamentally disentangles temporal sequence modeling from spatial message passing. TIDES extracts complex temporal dynamics and supra-Laplacian positional encodings in a one-time, parallelizable phase. By injecting these pre-computed priors into a trainable GNN, the model optimizes only spatial parameters. This memory-free inductive bias sidesteps the historical bottlenecks of recurrent models, enabling temporal parallelization during training. Empirically, TIDES achieves state-of-the-art or competitive performance in diverse supervised settings and demonstrates robust zero-shot transfer, significantly outperforming baselines on unseen networks. Future work can focus on extending the framework to Continuous-Time Dynamic Graphs (C-TDGs) (Gravina et al., 2024), relaxing the channel-independence assumption during temporal feature extraction, and scaling the pre-training corpus across heterogeneous domains to realize a universal foundation model.

Acknowledgements

FDC, AG, and DB acknowledge funding from EU-EIC EMERGE (Grant No. 101070918)

References

Ansari, A. F., Stella, L., Turkmen, A. C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., and Wang, B. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=gerNCVqqtR. Expert Certification.

Bai, J., Zhu, J., Song, Y., Zhao, L., Hou, Z., Du, R., and Li, H. A3T-GCN: Attention Temporal Graph Convolutional Network for Traffic Forecasting. *ISPRS International Journal of Geo-Information*, 10(7), 2021. ISSN 2220-9964. doi: 10.3390/ijgi10070485.

Bai, L., Yao, L., Li, C., Wang, X., and Wang, C. Adaptive graph convolutional recurrent network for traffic forecasting. *arXiv preprint arXiv:2007.02842*, 2020.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chatterjee, A., Ikica, B., Ravandi, B., and Palowitch, J. Transfer learning for temporal link prediction. *arXiv preprint arXiv:2504.10925*, 2025.

Chen, J., Wang, X., and Xu, X. GC-LSTM: Graph convolution embedded LSTM for dynamic network link prediction. *Applied Intelligence*, pp. 1–16, 2022.

Chen, Y., Segovia-Dominguez, I., and Gel, Y. R. Z-gcnets: Time zigzags at graph convolutional networks for time series forecasting. *arXiv preprint arXiv:2105.04100*, 2021.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734. Association for Computational Linguistics, October 2014. doi: 10.3115/v1/D14-1179.

Cini, A., Marisca, I., Bianchi, F. M., and Alippi, C. Scalable Spatiotemporal Graph Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6): 7218–7226, Jun. 2023. doi: 10.1609/aaai.v37i6.25880.

Cirstea, R.-G., Kieu, T., Guo, C., Yang, B., and Pan, S. J. Enhancenet: Plugin neural networks for enhancing correlated time series forecasting. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pp. 1739–1750, 2021. doi: 10.1109/ICDE51399.2021.00153.

Cirstea, R.-G., Yang, B., Guo, C., Kieu, T., and Pan, S. Towards spatio-temporal aware traffic time series forecasting–full version. *arXiv preprint arXiv:2203.15737*, 2022. URL https://arxiv.org/abs/2203.15737.

Deng, J., Chen, X., Jiang, R., Song, X., and Tsang, I. W. St-norm: Spatial and temporal normalization for multi-variate time series forecasting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 269–278, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467330. URL https://doi.org/10.1145/3447548.3467330.

Dwivedi, V. P. and Bresson, X. A Generalization of Transformer Networks to Graphs. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*, 2021.

Eliasof, M., Mantri, K. S. I., Bevilacqua, B., Ribeiro, B., and Schönlieb, C.-B. Bridging input feature spaces towards graph foundation models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Dt4XAIKYbf.

Errica, F., Gravina, A., Bacciu, D., and Micheli, A. Hidden Markov Models for Temporal Graph Representation Learning. In *Proceedings of the 31st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2023.

Fan, J., Weng, W., Tian, H., Wu, H., Zhu, F., and Wu, J. Rgdan: A random graph diffusion attention network for traffic prediction. *Neural Networks*, 172:106093, 2024. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2023.106093.

Fang, Z., Long, Q., Song, G., and Xie, K. Spatial-temporal graph ODE networks for traffic flow forecasting. *CoRR*, abs/2106.12931, 2021. URL https://arxiv.org/abs/2106.12931.

Galron, Y., Frasca, F., Maron, H., Treister, E., and Eliasof, M. Understanding and improving laplacian positional encodings for temporal gnns. In Ribeiro, R. P., Pfahringer, B., Japkowicz, N., Larrañaga, P., Jorge, A. M., Soares, C., Abreu, P. H., and Gama, J. (eds.), *Machine Learning and Knowledge Discovery in Databases. Research Track*, pp. 420–437, Cham, 2026a. Springer Nature Switzerland. ISBN 978-3-032-05981-9.

Galron, Y., Sinai, H., Maron, H., and Eliasof, M. Learning from historical activations in graph neural networks. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=8SnAGYf2wM.

Gao, H., Jiang, R., Dong, Z., Deng, J., Ma, Y., and Song, X. Spatial-temporal-decoupled masked pre-training for

spatiotemporal forecasting. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/442.

Gao, J. and Ribeiro, B. On the equivalence between temporal and static equivariant graph representations. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7052–7076. PMLR, 17–23 Jul 2022.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for Quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *ICML'17*, pp. 1263–1272. JMLR.org, 2017.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., and Dubrawski, A. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

Gravina, A. and Bacciu, D. Deep Learning for Dynamic Graphs: Models and Benchmarks. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):11788–11801, 2024. doi: 10.1109/TNNLS.2024.3379735.

Gravina, A., Lovisotto, G., Gallicchio, C., Bacciu, D., and Grohnfeldt, C. Long Range Propagation on Continuous-Time Dynamic Graphs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16206–16225. PMLR, 21–27 Jul 2024.

Guo, S., Lin, Y., Feng, N., Song, C., and Wan, H. Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):922–929, Jul. 2019. doi: 10.1609/aaai.v33i01.3301922.

Guo, S., Lin, Y., Wan, H., Li, X., and Cong, G. Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 34(11):5415–5428, 2022. doi: 10.1109/TKDE.2021.3056502.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9:1735–1780, 11 1997. doi: 10.1162/neco.1997.9.8.1735.

Jin, M., Koh, H. Y., Wen, Q., Zambon, D., Alippi, C., Webb, G. I., King, I., and Pan, S. A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection. *arXiv preprint arXiv:2307.03759*, 2024.

Karmim, Y., Lafon, M., S' niehotta, R. F., and Thome, N. Supra-laplacian encoding for transformer on dynamic graphs. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 17215–17246. Curran Associates, Inc., 2024. doi: 10.52202/079017-0547.

Kazemi, S. M., Goel, R., Jain, K., Kobyzev, I., Sethi, A., Forsyth, P., and Poupart, P. Representation Learning for Dynamic Graphs: a Survey. *J. Mach. Learn. Res.*, 21(1), jan 2020. ISSN 1532-4435.

Kim, T., Kim, J., Tae, Y., Park, C., Choi, J.-H., and Choo, J. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=cGDAkQo1C0p.

Kipf, T. N. and Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 09 2014. ISSN 2051-1310. doi: 10.1093/comnet/cnu016. URL https://doi.org/10.1093/comnet/cnu016.

Knyazev, A. V. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM Journal on Scientific Computing*, 23(2):517–541, 2001. doi: 10.1137/S1064827500366124. URL https://doi.org/10.1137/S1064827500366124.

Lan, S., Ma, Y., Huang, W., Wang, W., Yang, H., and Li, P. DSTAGNN: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11906–11917. PMLR,

17–23 Jul 2022. URL https://proceedings.mlr.press/v162/lan22a.html.

Li, H., Wang, X., Zhang, Z., and Zhu, W. Out-of-distribution generalization on graphs: A survey. *arXiv preprint arXiv:2202.07987*, 2022.

Li, J., Han, Z., Cheng, H., Su, J., Wang, P., Zhang, J., and Pan, L. Predicting Path Failure In Time-Evolving Graphs. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 1279–1289, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330847.

Li, J., Wu, R., Jin, X., Ma, B., Chen, L., and Zheng, Z. State space models on temporal graphs: A first-principles study. *Advances in Neural Information Processing Systems*, 37: 127030–127058, 2024.

Li, M. and Zhu, Z. Spatial-temporal fusion graph neural networks for traffic flow forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35 (5):4189–4196, May 2021. doi: 10.1609/aaai.v35i5.16542. URL https://ojs.aaai.org/index.php/AAAI/article/view/16542.

Li, Y., Yu, R., Shahabi, C., and Liu, Y. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJiHXGWAZ.

Li, Z., Xia, L., Shi, L., Xu, Y., Yin, D., and Huang, C. Open spatio-temporal foundation models for traffic prediction. *ACM Trans. Intell. Syst. Technol.*, 2025. ISSN 2157-6904. doi: 10.1145/3773912.

Liu, C., Yang, S., Xu, Q., Li, Z., Long, C., Li, Z., and Zhao, R. Spatial-temporal large language model for traffic prediction. *arXiv preprint arXiv:2401.10134*, 2024.

Liu, H., Dong, Z., Jiang, R., Deng, J., Deng, J., Chen, Q., and Song, X. Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4125–4129, 2023.

Liu, L., Xia, X., Xie, Q., Liu, B., Xu, W., and Peng, M. Enhanced expert merging for mixture-of-experts in graph foundation models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=fyqqd1lHDb.

Longa, A., Lachi, V., Santin, G., Bianchini, M., Lepri, B., Lio, P., Scarselli, F., and Passerini, A. Graph neural networks for temporal graphs: State of the art, open challenges, and opportunities. *arXiv preprint arXiv:2302.01018*, 2023.

Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Micheli, A. and Tortorella, D. Discrete-time dynamic graph echo state networks. *Neurocomputing*, 496(C):85–95, jul 2022. ISSN 0925-2312. doi: 10.1016/j.neucom.2022.05.001.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Pan, J., Nayyeri, M., Mohammed, O., Hernandez, D., Zhang, R., Cheng, C., and Staab, S. Towards foundation model on temporal knowledge graph reasoning. *arXiv preprint arXiv:2506.06367*, 2025.

Pareja, A., Domeniconi, G., Chen, J., Ma, T., Suzumura, T., Kanezashi, H., Kaler, T., Schardl, T. B., and Leiserson, C. E. EvolveGCN: Evolving Graph Convolutional Networks for Dynamic Graphs. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. Film: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Rozemberczki, B., Scherer, P., He, Y., Panagopoulos, G., Riedel, A., Astefanoaei, M., Kiss, O., Beres, F., López, G., Collignon, N., and Sarkar, R. PyTorch Geometric Temporal: Spatiotemporal Signal Processing with Neural Machine Learning Models. In *Proceedings of the 30th ACM International Conference on Information & Knowledge*

Management, CIKM '21, pp. 4564–4573. Association for Computing Machinery, 2021. ISBN 9781450384469. doi: 10.1145/3459637.3482014.

Ruiz, L., Gama, F., and Ribeiro, A. Gated graph recurrent neural networks. *IEEE Transactions on Signal Processing*, 68:6303–6318, 2020.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. ISSN 1476-4687. doi: 10.1038/323533a0.

Seo, Y., Defferrard, M., Vandergheynst, P., and Bresson, X. Structured sequence modeling with graph convolutional recurrent networks. In *25th International Conference on Neural Information Processing*, ICONIP, pp. 362–373. Springer, 2018.

Shamsi, K., Poursafaei, F., Huang, S., Ngo, B. T. G., Coskunuzer, B., and Akcora, C. G. Graphpulse: Topological representations for temporal graph property prediction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DZqic2sPTY.

Shamsi, K., Ngo, T. G. B., Shirzadkhani, R., Huang, S., Poursafaei, F., Azad, P., Rabbany, R., Coskunuzer, B., Rabusseau, G., and Akcora, C. G. MiNT: Multi-network transfer benchmark for temporal graph learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=Za7IcsXIRV.

Shang, C., Chen, J., and Bi, J. Discrete graph structure learning for forecasting multiple time series. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=WEHSlH5mOk.

Shao, Z., Zhang, Z., Wang, F., and Xu, Y. Pre-training enhanced spatial-temporal graph neural network for multivariate time series forecasting. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pp. 1567–1577. ACM, 2022.

Sofianos, T., Sampieri, A., Franco, L., and Galasso, F. Space-time-separable graph convolutional network for pose forecasting, 2021.

Song, C., Lin, Y., Guo, S., and Wan, H. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):914–921, Apr. 2020. doi: 10.1609/aaai.v34i01.5438.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, volume 2 of *NIPS'14*, pp. 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

Taheri, A. and Berger-Wolf, T. Predictive temporal embedding of dynamic graphs. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, pp. 57–64. Association for Computing Machinery, 2020. ISBN 9781450368681. doi: 10.1145/3341161.3342872.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Wang, Z., Zhang, Z., Chawla, N. V., Zhang, C., and Ye, Y. GFT: Graph foundation model with transferable tree vocabulary. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0MXzbAv8xy.

Wang, Z., Liu, Z., Ma, T., Li, J., Zhang, Z., Fu, X., Li, Y., Yuan, Z., Song, W., Ma, Y., Zeng, Q., Chen, X., Zhao, J., Li, J., Jiang, M., Lio, P., Chawla, N., Zhang, C., and Ye, Y. Graph foundation models: A comprehensive survey. *arXiv preprint arXiv:2505.15116*, 2025a.

Wang, Z., Zhang, Z., Ma, T., Chawla, N. V., Zhang, C., and Ye, Y. Towards graph foundation models: Learning generalities across graphs via task-trees. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=BSqf2k01ag.

Wu, Z., Pan, S., Long, G., Jiang, J., and Zhang, C. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, pp. 1907–1913. AAAI Press, 2019. ISBN 9780999241141.

Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., and Zhang, C. 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*, KDD '20, pp. 753–763. Association for Computing Machinery, 2020. ISBN 9781450379984. doi: 10.1145/3394486.3403118.

Yang, M., Zhou, M., Kalander, M., Huang, Z., and King, I. Discrete-time temporal network embedding via implicit hierarchical learning in hyperbolic space. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21,

pp. 1975–1985, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467422. URL https://doi.org/10.1145/3447548.3467422.

You, J., Du, T., and Leskovec, J. ROLAND: graph learning framework for dynamic graphs. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2358–2366, 2022.

Yu, B., Yin, H., and Zhu, Z. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pp. 3634–3640. AAAI Press, 2018. ISBN 9780999241127.

Yuan, H., Sun, Q., Shi, J., Fu, X., Hooi, B., Li, J., and Yu, P. S. How much can transfer? BRIDGE: Bounded multi-domain graph foundation model with generalization guarantees. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=bjDKZ3Roax.

Yuan, Y., Ding, J., Han, C., Sheng, Z., Jin, D., and Li, Y. Uniflow: A foundation model for unified urban spatio-temporal flow prediction. *arXiv preprint arXiv:2411.12972*, 2025b.

Zhang, X., Chen, X., Tang, H., Wu, Y., Shen, H., and Li, J. Adpstgcn: Adaptive spatial–temporal graph convolutional network for traffic forecasting. *Knowledge-Based Systems*, 301:112295, 2024. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2024.112295.

Zhao, L., Song, Y., Zhang, C., Liu, Y., Wang, P., Lin, T., Deng, M., and Li, H. T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 21(9):3848–3858, 2020. doi: 10.1109/TITS.2019.2935152.

Zheng, C., Fan, X., Wang, C., and Qi, J. Gman: A graph multi-attention network for traffic prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 (01):1234–1241, Apr. 2020. doi: 10.1609/aaai.v34i01.5477.

Zhu, Y., Cong, F., Zhang, D., Gong, W., Lin, Q., Feng, W., Dong, Y., and Tang, J. Wingnn: Dynamic graph neural networks with random gradient aggregation window. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 3650–3662. Association for Computing Machinery, 2023. ISBN 9798400701030. doi: 10.1145/3580305.3599551.

Zi, C., LIU, B., Sun, X., Cheng, H., and Li, J. Rethinking graph prompts: Unraveling the power of data manipulation in graph neural networks. In *The Fourth Blogpost Track at ICLR 2025*, 2025. URL https://openreview.net/forum?id=fQtOTcZhXI.

A. Experimental Details

In this section, we provide details regarding our experiments.

A.1. Implementation Details

To ensure the full reproducibility of our framework, we detail the specific configurations and hyperparameters used to instantiate TIDES across our experiments.

Time Series Foundation Model (Φ_{f}). For the extraction of temporal features, we utilized `Moment-1-small` from the family of open time series foundation models proposed in (Goswami et al., 2024) as our temporal encoder Φ_{f}. The temporal embeddings were generated without fine-tuning the base model to fully exploit its pre-learned, domain-agnostic temporal priors. For statistical summaries we choose $\Phi_{\mathrm{s}} = \mathbf{I}$ for all spatio-temporal traffic forecasting benchmarks and we use adaptive average pooling to project the input window to 16 output dimensions for all other tasks.

Supra-Laplacian Hyperparameters. The multi-layer eigendecomposition requires setting the inter-layer temporal coupling weight μ and the number of lowest-frequency eigenmodes k. Across our evaluations, we set $\mu = 1$ to avoid unbalancing inter-layer information flow, and extracted the $k = 8$ smallest eigenpairs for all datasets. Since in spatio-temporal data the topology is not evolving, we perform random edge drops and insertions, to strengthen the input-positional encodings associations with controlled perturbations of them.

Optimization. Models were trained using the Adam optimizer with an initial learning rate of 10^{-3}, a weight decay of 10^{-4}, and a dropout probability of 0.1 shared among layers. We also apply cosine annealing (Loshchilov & Hutter, 2017) to the learning rate for a maximum of 200 or 300 epochs, depending on the task.

Data normalization. For spatio-temporal tasks we adopt a two-stage normalization procedure inspired by Reversible Instance Normalization (Kim et al., 2022). In particular, we first perform a global z-score normalization computing the mean $\mu^{(g)}$ and standard deviation $\sigma^{(g)}$ of each of the d_{in} channels independently, aggregating over the entire temporal and spatial dimensions of the training set. The raw inputs are transformed into globally normalized sequences $\tilde{\mathbf{X}} = (\mathbf{X} - \mu^{(g)})/\sigma^{(g)}$.

Then, at training time, we extract the local statistical context of each input window of length w_s. For a given node $v \in \mathcal{V}$ and its temporal window $[\tilde{\mathbf{X}}_{t-w_s+1:t}]_v$, we compute the local temporal mean $\mu_v^{(l)}$ and standard deviation $\sigma_v^{(l)}$ strictly along the time dimension. Real-world spatio-temporal datasets, such as those used for traffic forecasting, are prone to sensor failures that result in missing or constant flatline readings over a given window. Under these conditions, the local variance collapses to zero, which can cause numerical instability and severe distribution shifts during denormalization. To ensure robustness against such anomalies, we bypass scaling for flatlined sensors by replacing near-zero standard deviations with ones. Specifically, we introduce a threshold τ (e.g., $\tau = 0.05$) and apply the following fallback mechanism:

$$\hat{\sigma}_v^{(l)} = \begin{cases} \sigma_v^{(l)} & \text{if } \sigma_v^{(l)} > \tau \\ 1 & \text{otherwise} \end{cases} \tag{10}$$

The final predictions $\hat{\mathbf{Y}}$ in the original physical magnitude are reconstructed by directly applying the sequential reversal of both the local window statistics and the global dataset statistics to the model's raw predictions $\hat{\mathbf{Y}}_{\text{raw}}$:

$$\hat{\mathbf{Y}} = \underbrace{\left(\hat{\mathbf{Y}}_{\text{raw}} \odot \hat{\sigma}^{(l)} + \mu^{(l)}\right)}_{\text{Local Denormalization}} \odot \sigma^{(g)} + \mu^{(g)} \tag{11}$$

where $\odot$ denotes element-wise multiplication. This adaptive target scaling ensures the model inherently adjusts to localized temporal spikes while maintaining stability against anomalous, flat-lining sensor data.

Gate. To integrate the temporal positional encodings Φ_{p} into the joint latent space we employ an adaptive gating mechanism evaluated for each node. Let $\mathbf{z}_v \in \mathbb{R}^d$ represent the projected features of node $v \in \mathcal{V}$ in the joint latent space. First, the positional encoding vector for node v, denoted as $[\Phi_{\mathrm{p}}]_v$, is mapped to the hidden dimension d through a projection network, followed by a learnable affine transformation with a strictly positive scale to ensure stability:

$$[\tilde{\Phi}_{\mathrm{p}}]_v = \text{Softplus}(\mathbf{s}) \odot \text{MLP}_{\text{proj}}([\Phi_{\mathrm{p}}]_v) + \mathbf{b} \tag{12}$$

where $\mathbf{s}, \mathbf{b} \in \mathbb{R}^d$ are learnable scale and shift parameters shared across all nodes, $\odot$ denotes element-wise multiplication, and MLP_{proj} is a shared projection module comprising layer normalization, a linear mapping, and a ReLU activation.

Next, we compute a node-specific, feature-conditioned gate that dynamically controls the amount of structural information injected into node v. The gate is derived from the concatenation of the node features and its processed positional encoding:

$$\mathbf{g}_v = \sigma\Big(\text{Clamp}\big(\text{MLP}_{\text{gate}}([\mathbf{z}_v \,\|\, [\tilde{\Phi}_{\text{p}}]_v]), -10, 10\big)\Big) \tag{13}$$

where $\|$ denotes concatenation along the feature dimension, MLP_{gate} is a two-layer network utilizing a GELU activation, Clamp bounds the pre-activation logits to $[-10, 10]$ to prevent numerical instability and gradient saturation, and σ is the sigmoid activation function.

Finally, the gated positional encodings are added to the original node features via a residual connection to produce the updated representation for node v:

$$\mathbf{z}'_v = \mathbf{z}_v + \mathbf{g}_v \odot [\tilde{\Phi}_{\text{p}}]_v \tag{14}$$

This formulation ensures that the model can selectively attend to the structural and temporal priors provided by Φ_{p} based on the current context of each individual node's features, seamlessly fusing them into a final joint latent representation $\mathbf{z}'_v$ per node.

Handling exogenous variables. To incorporate exogenous variables, such as datetime features, into the predictive process, we condition the node representations using Feature-wise Linear Modulation (FiLM) (Perez et al., 2018). Let $\mathbf{C} \in \mathbb{R}^{w_s \times F_c}$ denote the sequence of exogenous features of dimension F_c over the window of length w_s. We first compress this sequence into a single graph-level temporal context vector $\mathbf{c} \in \mathbb{R}^d$ by applying a 1D convolution with kernel size w_s along the temporal dimension, followed by a LeakyReLU activation:

$$\mathbf{c} = \text{LeakyReLU}\big(\text{Conv1D}(\mathbf{C})\big) \tag{15}$$

Next, we map this context vector to modulation parameters $\gamma, \beta \in \mathbb{R}^d$ via a two-layer Multi-Layer Perceptron (MLP_{FiLM}):

$$[\gamma \,\|\, \beta] = \text{MLP}_{\text{FiLM}}(\mathbf{c}) \tag{16}$$

where MLP_{FiLM} projects the context to 16 dimensions with a LeakyReLU activation and dropout, before expanding to $2d$ dimensions. To guarantee numerical stability during training and prevent extreme scaling, the multiplicative factor γ is tightly bounded around 1:

$$\tilde{\gamma} = \mathbf{1} + 0.1 \tanh(\gamma) \tag{17}$$

Finally, we apply this global modulation to the gated features of each individual node. For every node $v \in \mathcal{V}$, the feature vector $\mathbf{z}'_v$ is shifted and scaled via:

$$\mathbf{z}''_v = \tilde{\gamma} \odot \mathbf{z}'_v + \beta \tag{18}$$

This mechanism is the last applied step before the historical aggregation defined in Section 3.

A.2. Employed Baselines

To evaluate the effectiveness of TIDES, we compare it against a comprehensive set of baselines. Specifically:

A3T-GCN (Bai et al., 2021), **AdpSTGCN** (Zhang et al., 2024), **AGCRN** (Bai et al., 2020), **ASTGCN** (Guo et al., 2019), **ASTGNN** (Guo et al., 2022), **DCRNN** (Li et al., 2018), **DSTAGNN** (Lan et al., 2022), **DynGESN** (Micheli & Tortorella, 2022), **DyGrAE** (Taheri & Berger-Wolf, 2020), **EnhanceNet** (Cirstea et al., 2021), **EvolveGCN-H** and **EvolveGCN-O** (Pareja et al., 2020), **FC-LSTM** (Sutskever et al., 2014), **GC-LSTM** (Chen et al., 2022), **GCRN-GRU** and **GCRN-LSTM** (Seo et al., 2018), **GGRNN** (Ruiz et al., 2020), **GMAN** (Zheng et al., 2020), **GraphPulse** (Shamsi et al., 2024), **GraphSSM** (Li et al., 2024), Graph WaveNet (**GWNet**) (Wu et al., 2019), **GTS** (Shang et al., 2021), **HMM4G** (Errica et al., 2023), **HTGN** (Yang et al., 2021), **LRGCN** (Li et al., 2019), **MiNT** (Shamsi et al., 2026), **MPNN-LSTM** (Gilmer et al., 2017), **MTGNN** (Wu et al., 2020), Persistence Forecast (**P.F.**) (Shamsi et al., 2026), **RGDAN** (Fan et al., 2024), **ROLAND** (You et al., 2022), **STAEformer** (Liu et al., 2023), **STD-MAE** (Gao et al., 2024), **STEP** (Shao et al., 2022), **STFGNN** (Li & Zhu, 2021), **STGCN** (Yu et al., 2018), **STGODE** (Fang et al., 2021), **STNorm** (Deng et al., 2021), **STSGCN** (Sofianos et al., 2021), **ST-WA** (Cirstea et al., 2022), **T-GCN** (Zhao et al., 2020), **WinGNN** (Zhu et al., 2023), **Z-GCNETs** (Chen et al., 2021).

A.3. Employed Datasets

In our experiments, we evaluate TIDES on a variety of dynamic graph benchmarks. For node-level forecasting, we employ **Wikipedia Math**, which tracks daily page views; **ChickenPox Hungary**, representing epidemiological spread across counties; **Twitter Tennis**, modeling social interaction intensity; the large-scale traffic speed datasets **Metr-LA** and **PeMS-Bay**, which require capturing fine-grained correlations in sensor networks to predict the speed of passing vehicles; and the traffic flow benchmarks **PEMS03**, **PEMS04**, **PEMS07**, and **PEMS08**, which measure vehicle counts across California highway sensor networks at different geographic scales. In Table 6, we report the statistics of the employed datasets.

Table 6. Employed datasets statistics

Dataset	#Graphs	#Nodes	#Edges	Sequence Length	Evolving Topology
Twitter Tennis	1	1000	40839	120	✓
ChickenPox	1	20	102	517	✗
Wikipedia Math	1	1068	27079	723	✗
Metr-LA	1	207	1515	34272	✗
PeMS-BAY	1	325	2369	52116	✗
PEMS03	1	358	1093	26208	✗
PEMS04	1	307	680	16992	✗
PEMS07	1	883	1732	28224	✗
PEMS08	1	170	548	17856	✗

Instead, the statistics of the 84 employed token networks are shown in Table 7.

Table 7. All token networks' statistics.

Token	#Node	#Transaction	#Snapshots (days)	Growth rate	Novelty	Surprise	Token	#Node	#Transaction	#Snapshots (days)	Growth rate	Novelty	Surprise
ARC	11325	70968	606	0.43	0.32	0.88	Metis	52586	343141	907	0.44	0.48	0.89
CELR	65350	235807	1691	0.49	0.56	0.96	cDAI	52753	358050	1437	0.45	0.46	0.9
CMT	86895	205961	309	0.45	0.72	0.92	BITCOIN	34051	347054	178	0.48	0.39	0.63
DRGN	113453	341849	2164	0.44	0.57	0.97	INJ	60472	312822	1113	0.46	0.52	0.98
GHST	35156	180955	1146	0.43	0.51	0.93	MIM	23038	269366	885	0.44	0.4	0.89
INU	8556	66315	154	0.27	0.41	0.59	GLM	53385	234912	1080	0.5	0.53	0.96
IOTX	63079	288469	1993	0.45	0.56	0.99	Mog	14590	240680	107	0.37	0.38	0.55
QSP	117977	299671	2178	0.45	0.67	0.99	DPI	40627	234246	1150	0.49	0.5	0.86
REP	83282	224843	346	0.46	0.69	0.96	LINA	45342	227147	1144	0.45	0.46	0.95
RFD	23208	173695	169	0.3	0.39	0.6	Yf-DAI	22466	226875	1158	0.42	0.31	0.87
TNT	88247	316352	1216	0.43	0.55	0.93	BOB	42806	212099	199	0.35	0.48	0.73
TRAC	71667	299181	2110	0.46	0.54	0.97	RGT	35277	211932	1110	0.44	0.46	0.98
RLB	28033	240291	129	0.43	0.49	0.76	TVK	42539	208082	1062	0.41	0.48	0.93
steCRV	19079	211538	1033	0.45	0.53	0.9	RSR	50645	205906	659	0.47	0.62	0.91
ALBT	63042	434881	1152	0.43	0.44	0.89	WOJAK	34341	198653	201	0.37	0.48	0.73
POLS	128159	554705	1132	0.45	0.61	0.94	ANT	36517	200262	1107	0.47	0.46	0.93
SWAP	69230	509769	1213	0.46	0.45	0.79	LADYS	37486	192176	181	0.37	0.52	0.79
SUPER	83299	502030	986	0.47	0.46	0.85	ETH2x-FLI	11008	199088	965	0.47	0.28	0.84
RARI	87186	502960	1207	0.43	0.47	0.91	TURBO	38638	189048	189	0.33	0.48	0.72
KP3R	39323	493258	1102	0.43	0.33	0.88	REPv2	39061	191367	1194	0.48	0.5	0.97
MIR	79984	444998	1066	0.45	0.43	0.92	NOIA	29798	185528	1133	0.46	0.37	0.7
aUSDC	23742	475680	1067	0.46	0.4	0.73	0x0	21531	182430	283	0.51	0.46	0.81
LUSD	25852	430473	943	0.48	0.36	0.87	PSYOP	25450	168896	169	0.32	0.39	0.59
PICKLE	28498	430262	1149	0.48	0.34	0.69	ShibDoge	40023	134697	680	0.43	0.53	0.8
DODO	47046	390443	1131	0.47	0.45	0.91	ADX	14567	123755	1188	0.44	0.4	0.91
YFII	43964	391984	1196	0.44	0.44	0.96	BAG	11860	122634	298	0.31	0.44	0.87
STARL	71590	369913	856	0.46	0.48	0.86	QOM	21757	118292	598	0.46	0.41	0.81
LQTY	34687	374230	943	0.45	0.34	0.91	BEPRO	26521	120261	1132	0.46	0.48	0.87
FEG	118294	367584	1007	0.4	0.62	0.92	AIOZ	29231	119926	947	0.43	0.49	0.89
AUDIO	91218	362685	1108	0.45	0.58	0.95	PRE	40476	118625	1113	0.5	0.55	0.86
OHM	45728	377068	690	0.43	0.46	0.88	CRU	19990	117712	1144	0.5	0.43	0.95
WOOL	16874	351178	716	0.41	0.18	0.41	POOH	27245	111641	193	0.26	0.49	0.69
DERC	24277	111205	824	0.45	0.49	0.83	aDAI	13648	187050	1068	0.45	0.46	0.82
stkAAVE	37355	110924	1128	0.42	0.57	0.71	ORN	44010	239451	1134	0.46	0.47	0.87
BTRFLY	8450	108371	453	0.48	0.34	0.44	DOGE2.0	7664	79047	123	0.45	0.38	0.66
SDEX	9127	104869	240	0.41	0.44	0.75	HOICHI	5075	77361	436	0.36	0.32	0.71
XCN	20085	104185	607	0.46	0.42	0.84	EVERMOON	7552	79868	163	0.24	0.35	0.52
HOP	37004	102650	514	0.41	0.6	0.88	MUTE	12426	82345	977	0.43	0.46	0.95
MAHA	18401	96180	749	0.43	0.47	0.91	crvUSD	2950	88647	174	0.61	0.37	0.73
DINO	15837	94140	358	0.44	0.44	0.74	SLP	6675	95368	1151	0.43	0.36	0.91
bendWETH	1454	96898	593	0.51	0.21	0.51	sILV2	12838	92905	611	0.4	0.34	0.48
PUSH	14501	93103	936	0.46	0.38	0.83	SPONGE	25852	90468	184	0.31	0.66	0.81

Table 8. The grid of hyperparameters employed by TIDES during model selection for all evaluated tasks. Datasets include: Traffic Forecasting (**Pems03**, **Pems04**, **Pems07**, **Pems08**), Node Prediction (**ChickenPox**, **Wikipedia Math**, **Twitter Tennis**), and the **MiNT** benchmark.

Hyperparameters	**Values**		
	Traffic Forecasting	**Node Prediction**	**MiNT**
Optimizer	Adam	Adam	Adam
Learning rate	0.003, 0.001	0.001	0.001
Weight decay	10^{-4}	10^{-4}	10^{-4}
d	512	256, 512	32, 64
L	10	5, 10	2
Aggregate layers	✓	✓	✗
w_f	72, 144, 288	16, 32	8
w_s	12	$w_s = w_f$	$w_s = w_f$
w_p	8	8	8
k	8	8	8
MHSA #heads	2, 4	2, 4	2, 4

A.4. Hyperparameter Space

The hyperparameter space employed by TIDES across all tasks is reported in Table 8.

B. Additional Results and Comparisons

In this section we report additional results and comparisons with other datasets and baselines.

B.1. Supervised Evaluation on MiNT Networks

Table 9 reports the per-network supervised AUC of TIDES-sup against seven baseline architectures across the 20 MiNT test networks. TIDES-sup is a 2-layer GCN with skip connections and no historical aggregation, trained in a fully supervised fashion on each of the 20 test networks independently. It exhibits a specialized performance profile: it achieves state-of-the-art results on SDEX (0.954 vs. 0.762), ETH2x-FLI (0.832 vs. 0.688), and BEPRO (0.798 vs. 0.783), but shows high variance across the remaining tokens. Among the baselines, GraphPulse is the strongest, ranking first on 10 of 20 networks, while GC-LSTM and HTGN frequently appear in the top three. EvolveGCN and the Persistence Forecast (P.F.) baseline of (Shamsi et al., 2026) struggle on this benchmark, with P.F. yielding the weakest predictive performance overall. These results show the effectiveness of our pre-training, in which we bring substantial positive transfer across the majority of test networks, outperforming the MiNT counterparts (see Table 5).

Table 9. Supervised Evaluation Performance (AUC) on the 20 MiNT Test Networks. This table compares our fully trained TIDES against the single supervised baselines. Best results in **green**, second best in **blue**, and third best in **bold**.

	Single Model on Individual Networks							
Network	**HTGN**	**GC-LSTM**	**EvolveGCN**	**GraphPulse**	**ROLAND**	**WinGNN**	**P.F.**	**TIDES-sup**
WOJAK	$0.479_{\pm 0.005}$	$0.484_{\pm 0.000}$	$\mathbf{0.505_{\pm 0.023}}$	$0.467_{\pm 0.030}$	$\mathbf{0.529_{\pm 0.005}}$	$\mathbf{0.511_{\pm 0.026}}$	0.378	$0.278_{\pm 0.179}$
DOGE2.0	$\mathbf{0.590_{\pm 0.059}}$	$0.538_{\pm 0.000}$	$\mathbf{0.551_{\pm 0.022}}$	$0.384_{\pm 0.180}$	$0.513_{\pm 0.022}$	$\mathbf{0.577_{\pm 0.038}}$	0.250	$0.481_{\pm 0.051}$
EVERMOON	$0.512_{\pm 0.023}$	$\mathbf{0.562_{\pm 0.179}}$	$0.451_{\pm 0.046}$	$\mathbf{0.519_{\pm 0.130}}$	$0.349_{\pm 0.119}$	$\mathbf{0.525_{\pm 0.114}}$	0.241	$0.511_{\pm 0.141}$
QOM	$0.633_{\pm 0.017}$	$0.612_{\pm 0.001}$	$0.618_{\pm 0.002}$	$\mathbf{0.775_{\pm 0.011}}$	$\mathbf{0.641_{\pm 0.003}}$	$\mathbf{0.645_{\pm 0.099}}$	0.334	$0.619_{\pm 0.026}$
SDEX	$\mathbf{0.762_{\pm 0.034}}$	$0.720_{\pm 0.002}$	$\mathbf{0.733_{\pm 0.028}}$	$0.436_{\pm 0.030}$	$0.483_{\pm 0.254}$	$0.726_{\pm 0.000}$	0.423	$\mathbf{0.954_{\pm 0.032}}$
ETH2x-FLI	$0.610_{\pm 0.059}$	$\mathbf{0.670_{\pm 0.009}}$	$\mathbf{0.688_{\pm 0.010}}$	$0.666_{\pm 0.047}$	$0.621_{\pm 0.023}$	$0.617_{\pm 0.056}$	0.355	$\mathbf{0.832_{\pm 0.033}}$
BEPRO	$0.655_{\pm 0.038}$	$0.632_{\pm 0.019}$	$0.610_{\pm 0.012}$	$\mathbf{0.783_{\pm 0.003}}$	$0.439_{\pm 0.125}$	$\mathbf{0.736_{\pm 0.018}}$	0.393	$\mathbf{0.798_{\pm 0.037}}$
XCN	$\mathbf{0.668_{\pm 0.099}}$	$0.306_{\pm 0.092}$	$0.512_{\pm 0.067}$	$\mathbf{0.821_{\pm 0.004}}$	$\mathbf{0.765_{\pm 0.015}}$	$0.586_{\pm 0.029}$	0.592	$0.466_{\pm 0.074}$
BAG	$\mathbf{0.673_{\pm 0.227}}$	$0.196_{\pm 0.179}$	$0.329_{\pm 0.040}$	$\mathbf{0.934_{\pm 0.020}}$	$0.418_{\pm 0.016}$	$0.485_{\pm 0.105}$	**0.792**	$0.098_{\pm 0.036}$
TRAC	$0.712_{\pm 0.071}$	$\mathbf{0.748_{\pm 0.000}}$	$\mathbf{0.748_{\pm 0.000}}$	$\mathbf{0.767_{\pm 0.001}}$	$0.495_{\pm 0.223}$	$\mathbf{0.752_{\pm 0.007}}$	0.400	$0.738_{\pm 0.004}$
DERC	$\mathbf{0.683_{\pm 0.013}}$	$\mathbf{0.703_{\pm 0.022}}$	$0.669_{\pm 0.009}$	$\mathbf{0.769_{\pm 0.040}}$	$0.405_{\pm 0.357}$	$0.674_{\pm 0.044}$	0.353	$0.615_{\pm 0.104}$
Metis	$\mathbf{0.715_{\pm 0.122}}$	$0.646_{\pm 0.023}$	$0.688_{\pm 0.027}$	$\mathbf{0.812_{\pm 0.011}}$	$0.696_{\pm 0.108}$	$0.690_{\pm 0.039}$	0.423	$\mathbf{0.782_{\pm 0.013}}$
REPv2	$\mathbf{0.760_{\pm 0.012}}$	$0.725_{\pm 0.014}$	$0.709_{\pm 0.002}$	$\mathbf{0.830_{\pm 0.001}}$	$\mathbf{0.751_{\pm 0.003}}$	$0.744_{\pm 0.026}$	0.321	$0.750_{\pm 0.006}$
DINO	$0.730_{\pm 0.195}$	$\mathbf{0.874_{\pm 0.028}}$	$\mathbf{0.868_{\pm 0.029}}$	$\mathbf{0.801_{\pm 0.020}}$	$0.497_{\pm 0.092}$	$0.628_{\pm 0.251}$	0.431	$0.403_{\pm 0.021}$
HOICHI	$0.807_{\pm 0.047}$	$\mathbf{0.857_{\pm 0.000}}$	$\mathbf{0.856_{\pm 0.001}}$	$0.714_{\pm 0.010}$	$\mathbf{0.815_{\pm 0.036}}$	$0.769_{\pm 0.101}$	0.374	$0.720_{\pm 0.002}$
MUTE	$\mathbf{0.649_{\pm 0.015}}$	$0.593_{\pm 0.030}$	$\mathbf{0.617_{\pm 0.010}}$	$\mathbf{0.779_{\pm 0.004}}$	$0.289_{\pm 0.042}$	$0.593_{\pm 0.054}$	0.536	$0.613_{\pm 0.107}$
GLM	$\mathbf{0.830_{\pm 0.029}}$	$0.451_{\pm 0.003}$	$0.501_{\pm 0.033}$	$\mathbf{0.769_{\pm 0.018}}$	$0.559_{\pm 0.357}$	$0.530_{\pm 0.004}$	0.427	$\mathbf{0.812_{\pm 0.149}}$
MIR	$\mathbf{0.750_{\pm 0.005}}$	$\mathbf{0.768_{\pm 0.026}}$	$\mathbf{0.745_{\pm 0.015}}$	$0.689_{\pm 0.097}$	$0.228_{\pm 0.060}$	$0.742_{\pm 0.015}$	0.327	$0.720_{\pm 0.027}$
stkAAVE	$\mathbf{0.702_{\pm 0.042}}$	$0.368_{\pm 0.011}$	$0.397_{\pm 0.022}$	$\mathbf{0.743_{\pm 0.006}}$	$\mathbf{0.591_{\pm 0.122}}$	$0.572_{\pm 0.018}$	0.426	$0.493_{\pm 0.104}$
ADX	$\mathbf{0.769_{\pm 0.018}}$	$0.723_{\pm 0.002}$	$0.718_{\pm 0.004}$	$\mathbf{0.784_{\pm 0.002}}$	$\mathbf{0.761_{\pm 0.011}}$	$0.733_{\pm 0.023}$	0.362	$0.677_{\pm 0.018}$
Top rank ↑	2	4	0	10	1	0	0	3

B.2. Traffic Speed Forecasting Benchmarks

In Table 10 we report the MasekdMAE, MaskedRMSE and MaskeMAPE of TIDES on the Metr-LA and PeMS-Bay benchmarks. Overall, TIDES demonstrates competitive MaskedRMSE, which penalizes large prediction errors. This specific result highlights the model's reliability in handling extreme variance and anomalous traffic conditions. While TIDES does not strictly dominate the top-performing models across all remaining MaskedMAE and MaskedMAPE metrics, this reflects a deliberate structural trade-off. TIDES operates with strong inductive biases, providing a highly reliable and theoretically grounded alternative.

Table 10. Time series forecasting errors (MAE, RMSE, MAPE) on the Metr-LA and PeMS-Bay datasets for Horizon 12 and a Window of 12. **First**, **second**, and **third** best results are color-coded.

Model	**Metr-LA**			**PeMS-Bay**		
	MAE ↓	RMSE ↓	MAPE ↓	MAE ↓	RMSE ↓	MAPE ↓
AdpSTGCN	**3.40**	**7.21**	**9.45%**	**1.92**	4.49	4.62%
ASTGCN	6.51	12.52	11.64%	2.61	5.42	6.00%
DCRNN	3.60	7.60	10.50%	2.07	4.74	4.90%
GMAN	3.44	7.35	10.07%	1.86	**4.32**	**4.37%**
Graph WaveNet	3.53	7.37	10.01%	1.95	4.52	4.63%
GTS	3.46	7.31	9.98%	1.95	**4.43**	4.58%
MTGNN	3.49	7.23	**9.87%**	1.94	4.49	4.53%
RGDAN	**3.26**	**7.02**	9.73%	**1.82**	**4.20**	**4.28%**
STEP	**3.37**	**6.99**	**9.61%**	**1.79**	**4.20**	**4.18%**
STGCN	4.59	9.40	12.70%	2.49	5.69	5.79%
STSGCN	5.06	11.66	12.91%	2.26	5.21	5.40%
GGRNN	3.88	8.14	10.59%	2.34	5.14	5.21%
GraphSSM-S4	3.74	7.90	10.37%	1.98	4.45	4.77%
TIDES	3.85	**7.21**	11.17%	2.10	**4.20**	5.02%