Physics-Informed Bidirectional Graph Networks for Traffic Prediction: Deriving Message Passing Direction from Traffic Flow Theory

Abstract

Traffic prediction is essential for intelligent transportation systems, yet existing graph neural networks treat spatial message passing as direction-agnostic, ignoring the physics of traffic wave propagation. We propose RA-BiGN, a physics-informed bidirectional graph network that derives its message passing direction from the Lighthill-Whitham-Richards (LWR) traffic flow model. Our key insight is that the characteristic speed $c = v_f(1 - 2\rho/\rho_{\max})$ determines whether information propagates downstream (free-flow, $c > 0$) or upstream (congestion/shockwaves, $c < 0$). We embed this physics into a learnable propagation weight α that dynamically balances bidirectional graph convolution based on local traffic state. Experiments on PEMS04 and PEMS08 demonstrate competitive performance with state-of-the-art methods, while ablations confirm that physics-informed α outperforms uniform weighting. Crucially, visualizations show the learned α correlates with speed and occupancy exactly as LWR theory predicts, validating our physics-grounded design. Our work suggests that incorporating domain-specific physical laws into GNN architectures is a promising direction for spatiotemporal forecasting.

1. Introduction

Accurate traffic prediction is fundamental to intelligent transportation systems, enabling applications such as route planning, congestion management, and emission reduction (Li et al., 2018). The inherent spatial dependencies among road network sensors and complex temporal dynamics make this a challenging spatiotemporal forecasting problem.

Recent years have witnessed remarkable progress through graph neural networks (GNNs) that model road networks as graphs. DCRNN (Li et al., 2018) pioneered this direction by combining diffusion convolution with recurrent networks to capture spatial and temporal dependencies. Graph WaveNet (Wu et al., 2019) advanced the field by introducing adaptive adjacency matrices learned end-to-end, removing reliance on predefined graph structures. Attention-based approaches further improved performance: GMAN (Zheng et al., 2020) employed spatial-temporal attention mechanisms with gated fusion, while STAEformer (Liu et al., 2023) leveraged transformer architectures with adaptive spatial-temporal embeddings to achieve state-of-the-art results.

Despite their empirical success, existing methods share a fundamental limitation: *they treat spatial message passing as direction-agnostic*. Information is propagated symmetrically across the graph, implicitly assuming traffic dynamics are uniform regardless of traffic state. However, traffic flow physics tells a different story. The classical Lighthill-Whitham-Richards (LWR) model (Lighthill & Whitham, 1955; Richards, 1956) establishes that information propagation in traffic fundamentally depends on the *characteristic speed*:

$$c = \frac{dq}{d\rho} = v_f \left(1 - \frac{2\rho}{\rho_{\max}}\right) \tag{1}$$

where q is flow, ρ is density, and v_f is free-flow speed. When $c > 0$ (free-flow), perturbations propagate *downstream* with traffic. When $c < 0$ (congestion), shockwaves propagate *upstream* against traffic. This asymmetric, state-dependent propagation is entirely ignored by current GNN architectures.

In this work, we propose **RA-BiGN** (Regime-Aware Bidirectional Graph Network), which bridges traffic flow theory and graph neural networks by deriving the message passing direction from first principles. Our key insight is that the characteristic speed c naturally defines a propagation weight $\alpha = \sigma(c/\tau)$, where σ is the sigmoid function and τ is a learnable temperature parameter controlling the sharpness of regime transitions, that dynamically balances downstream and upstream message passing based on local traffic state. This physics-informed design respects the underlying dynamics: during free-flow, information flows with traffic; during congestion, the model captures upstream-propagating shockwaves.

Our contributions are as follows:

- We identify a fundamental gap in existing traffic prediction GNNs: the neglect of direction-dependent information propagation dictated by traffic flow physics.

- We propose RA-BiGN, which derives its bidirectional message passing weights from the LWR characteristic speed, providing theoretical grounding for the architecture design.
- We demonstrate that the learned propagation weight α correlates with traffic state (speed and occupancy) exactly as LWR theory predicts, empirically validating the physics-informed approach.
- Experiments on PEMS04 and PEMS08 show competitive performance with state-of-the-art methods, with ablations confirming the benefit of physics-informed α over uniform weighting.

Our work suggests that incorporating domain-specific physical laws into neural network architectures is a promising direction for spatiotemporal forecasting, offering both improved interpretability and theoretically grounded design choices.

2. Related Work

2.1. Classical and Statistical Methods

Early traffic prediction relied on time series models. The Autoregressive Integrated Moving Average (ARIMA) model and its variants (Ahmed & Cook, 1979; Williams & Hoel, 2003) capture temporal dependencies through linear autoregressive processes. Vector Autoregression (VAR) extends this to multivariate settings (Zivot & Wang, 2006). While computationally efficient, these methods assume stationarity and fail to capture complex nonlinear spatial-temporal dependencies inherent in traffic networks.

2.2. Deep Learning Approaches

Deep learning methods significantly advanced traffic prediction by modeling nonlinear patterns. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997), were applied to capture temporal dynamics (Ma et al., 2015). Convolutional Neural Networks (CNNs) were adapted to model spatial correlations by treating traffic data as images (Zhang et al., 2017). However, these approaches either ignore spatial structure (RNNs) or impose rigid grid-based spatial assumptions (CNNs) that poorly fit irregular road networks.

2.3. Graph Neural Networks for Traffic

Representing road networks as graphs enabled more natural spatial modeling. Spatio-Temporal Graph Convolutional Networks (STGCN) (Yu et al., 2018) combined graph convolutions with temporal convolutions, demonstrating the effectiveness of graph-based approaches. DCRNN (Li et al., 2018) introduced diffusion convolution to model bidirectional traffic flow and integrated it with sequence-to-sequence learning. Graph WaveNet (Wu et al., 2019) proposed adaptive adjacency matrices learned end-to-end, removing dependence on predefined graph structures. AGCRN (Bai et al., 2020) further enhanced adaptability through node-specific parameter learning. MTGNN (Wu et al., 2020) unified graph learning with spatial-temporal modeling in a single framework.

Despite their success, these methods treat spatial message passing as direction-agnostic, propagating information symmetrically regardless of traffic state, a limitation our work directly addresses.

2.4. Attention and Transformer Models

Attention mechanisms have recently achieved state-of-the-art results. GMAN (Zheng et al., 2020) employed spatial-temporal attention with gated fusion for long-term prediction. ASTGNN (Guo et al., 2021) designed dynamic spatial-temporal attention to capture time-varying correlations. STAEformer (Liu et al., 2023) demonstrated that vanilla transformers with adaptive spatial-temporal embeddings can achieve competitive performance without complex architectural modifications. PDFormer (Jiang et al., 2023) incorporated propagation delay awareness into the transformer framework, partially acknowledging directional traffic dynamics.

2.5. Physics-Informed Neural Networks

Physics-informed approaches incorporate domain knowledge into neural networks. Raissi et al. (2019) introduced Physics-Informed Neural Networks (PINNs) that embed physical laws as constraints. In traffic, the Lighthill-Whitham-Richards (LWR) model (Lighthill & Whitham, 1955; Richards, 1956) provides fundamental theory describing traffic wave propagation. Recent works have explored physics-guided architectures (Fang et al., 2021; Shi et al., 2021; Yuan et al., 2021), but none have derived GNN message passing directions from characteristic speed theory. Our work bridges this gap by grounding bidirectional graph convolution in LWR physics.

3. Methodology

3.1. Problem Formulation

We represent a road network as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where $\mathcal{V}$ is a set of N sensor nodes, $\mathcal{E}$ is the set of edges representing road connectivity, and $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix. At each time step t, we observe traffic features $\mathbf{X}^{(t)} \in \mathbb{R}^{N \times C}$, where C denotes the number of features (flow, occupancy, and speed). Given his-

torical observations $\mathbf{X}^{(t-T+1:t)} = \{\mathbf{X}^{(t-T+1)}, \ldots, \mathbf{X}^{(t)}\}$ over T time steps, our goal is to predict future traffic flow $\mathbf{Y}^{(t+1:t+T')} = \{\mathbf{Y}^{(t+1)}, \ldots, \mathbf{Y}^{(t+T')}\}$ for the next T' time steps.

3.2. Traffic Flow Physics: The LWR Model

The Lighthill-Whitham-Richards (LWR) model (Lighthill & Whitham, 1955; Richards, 1956) describes traffic as a conservation law:

$$\frac{\partial \rho}{\partial t} + \frac{\partial q(\rho)}{\partial x} = 0 \tag{2}$$

where ρ is traffic density and $q(\rho)$ is the flow-density relationship (fundamental diagram). Using the Greenshields model, $q(\rho) = \rho \cdot v_f \cdot (1 - \rho/\rho_{\max})$, the *characteristic speed* that determines information propagation direction is:

$$c = \frac{dq}{d\rho} = v_f \left(1 - \frac{2\rho}{\rho_{\max}}\right) \tag{3}$$

This yields three traffic regimes:

- **Free-flow** ($c > 0$): Low density, perturbations propagate *downstream* with traffic flow.
- **Congested** ($c < 0$): High density, shockwaves propagate *upstream* against traffic flow.
- **Critical** ($c = 0$): At capacity, standing waves form.

This physics motivates our key design: message passing direction in GNNs should adapt to the local traffic regime.

3.3. RA-BiGN Architecture

The proposed model, illustrated in Figure 1, consists of three main components: (1) a physics-based characteristic speed estimator, (2) bidirectional graph convolution layers, and (3) spatial-temporal attention modules.

3.3.1. Physics-Based Characteristic Speed Estimator

Given input features $\mathbf{X} \in \mathbb{R}^{B \times T \times N \times C}$ containing flow, occupancy (ρ), and speed (v), we estimate the characteristic speed following the LWR model:

$$c_{\text{physics}} = v_f \cdot v \cdot \left(1 - \frac{2\rho}{\rho_c + \epsilon}\right) \tag{4}$$

where v_f is a learnable free-flow speed scale and ρ_c is a learnable critical density threshold. To account for real-world deviations from idealized physics, we add a small learned correction:

$$c = c_{\text{physics}} + \lambda \cdot f_\theta(\mathbf{X}) \tag{5}$$

where f_θ is a small MLP and $\lambda = 0.1$ limits its influence. The propagation weight is then:

$$\alpha = \sigma\left(\frac{c}{\tau}\right) \tag{6}$$

where σ is the sigmoid function and τ is a learnable temperature parameter. When $\alpha \approx 1$ (free-flow), information propagates downstream; when $\alpha \approx 0$ (congested), information propagates upstream.

3.3.2. LWR-Inspired Bidirectional Graph Convolution

Standard graph convolution propagates information symmetrically. We propose a bidirectional variant that respects traffic physics:

$$\mathbf{H}' = \alpha \cdot (\mathbf{A}\mathbf{H}\mathbf{W}_{\text{down}}) + (1 - \alpha) \cdot (\mathbf{A}^\top \mathbf{H}\mathbf{W}_{\text{up}}) + \mathbf{H}\mathbf{W}_{\text{self}} \tag{7}$$

where $\mathbf{A}$ is the normalized adjacency matrix, $\mathbf{W}_{\text{down}}$, $\mathbf{W}_{\text{up}}$, and $\mathbf{W}_{\text{self}}$ are learnable weight matrices, and $\alpha \in \mathbb{R}^{B \times T \times N \times 1}$ is the node-specific, time-varying propagation weight from the characteristic speed estimator.

This formulation can be interpreted as a discretization of the LWR equation on a graph: downstream propagation ($\mathbf{A}\mathbf{H}$) captures information flowing with traffic, while upstream propagation ($\mathbf{A}^\top \mathbf{H}$) captures shockwave dynamics.

3.3.3. Spatial-Temporal Block

Each spatial-temporal (ST) block combines temporal and spatial processing:

1. **Temporal Attention**: Multi-head self-attention across time steps for each node independently, capturing temporal dependencies.
2. **Spatial Attention**: Multi-head attention across nodes with graph-structure bias from $\mathbf{A}$, capturing global spatial dependencies.
3. **LWR Graph Convolution**: Physics-informed bidirectional convolution as described above.
4. **Feed-Forward Network**: Two-layer MLP with GELU activation.

Each sub-layer employs residual connections and layer normalization.

3.3.4. Overall Architecture

The complete model stacks L spatial-temporal blocks:

$$\mathbf{H}^{(0)} = \text{Embed}(\mathbf{X}) + \mathbf{E}_{\text{pos}} + \mathbf{E}_{\text{node}} \tag{8}$$

$$\mathbf{H}^{(l)} = \text{STBlock}^{(l)}(\mathbf{H}^{(l-1)}, \mathbf{A}, \alpha), \quad l = 1, \ldots, L \tag{9}$$

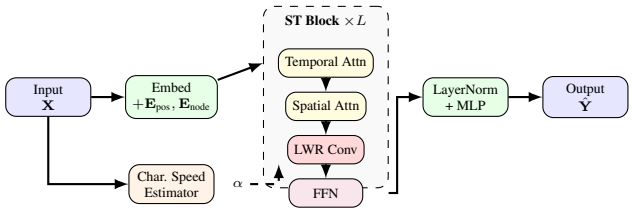

Figure 1. Overview of RA-BiGN architecture. Traffic features $\mathbf{X}$ (flow, occupancy, speed) are passed through an embedding layer and L stacked Spatial-Temporal (ST) Blocks. The *Characteristic Speed Estimator* computes physics-based propagation weights α from the input using the LWR model. Within each ST Block, temporal and spatial attention capture dependencies, while the *LWR Graph Convolution* uses α to dynamically balance downstream ($\mathbf{AH}$) and upstream ($\mathbf{A}^\top\mathbf{H}$) message passing based on local traffic state. A final MLP produces multi-step predictions $\hat{\mathbf{Y}}$.

$$\hat{\mathbf{Y}} = \text{MLP}(\text{LayerNorm}(\mathbf{H}^{(L)})) \tag{10}$$

where $\mathbf{E}_{\text{pos}}$ and $\mathbf{E}_{\text{node}}$ are learnable positional and node embeddings, respectively.

3.4. Training Objective

We train the model using masked mean absolute error (MAE) to handle zero values in traffic data:

$$\mathcal{L} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} |\hat{y}_i - y_i| \tag{11}$$

where $\mathcal{M} = \{i : y_i \neq 0\}$ is the set of non-zero target indices.

4. Experiments

4.1. Datasets

We evaluate RA-BiGN on two widely-used traffic flow benchmarks from the California Department of Transportation Performance Measurement System (PeMS):

- **PEMS04**: Collected from 307 sensors in the San Francisco Bay Area over 59 days (January–February 2018). The dataset contains 16,992 time steps with three features: traffic flow, occupancy, and speed.
- **PEMS08**: Collected from 170 sensors in the San Bernardino area over 62 days (July–August 2016). The dataset contains 17,856 time steps with the same three features.

Both datasets are aggregated at 5-minute intervals. We use a 6:2:2 split for training, validation, and testing, following prior work (Liu et al., 2023). The task is to predict traffic flow for the next 12 time steps (1 hour) given the previous 12 time steps.

4.2. Baselines

We compare RA-BiGN against representative methods spanning classical, graph-based, and attention-based approaches:

Classical Methods:

- **ARIMA** (Box & Jenkins, 1970): A classical autoregressive time series model.
- **VAR** (Zivot & Wang, 2006): Vector Auto-Regression, a multivariate time series model that captures cross-series dependencies.

Graph Neural Networks:

- **DCRNN** (Li et al., 2018): Combines diffusion convolution with sequence-to-sequence learning using GRUs.
- **STGCN** (Yu et al., 2018): Integrates graph convolutions with temporal convolutions in a convolutional architecture.
- **Graph WaveNet** (Wu et al., 2019): Employs adaptive adjacency matrices with dilated causal convolutions.
- **AGCRN** (Bai et al., 2020): Uses node-adaptive parameter learning with graph convolution.
- **STGODE** (Fang et al., 2021): Models spatial-temporal dynamics through neural ordinary differential equations.

Attention-based Methods:

- **GMAN** (Zheng et al., 2020): Applies spatial-temporal attention with gated fusion mechanisms.
- **ASTGNN** (Guo et al., 2021): Employs dynamic spatial-temporal attention to capture time-varying correlations.
- **STAEformer** (Liu et al., 2023): Achieves state-of-the-art performance using vanilla transformers with adaptive embeddings.

4.3. Experimental Settings

Metrics. We adopt three standard evaluation metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). Following prior work, metrics are computed on non-zero values only.

Table 1. Performance comparison on PEMS04 and PEMS08 datasets. Bold indicates best performance, underline indicates second best. Numbers in parentheses denote column-wise rank; Avg. Rank is the mean across all six metrics.

Model	PEMS04			PEMS08			Avg. Rank
	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)	
ARIMA	33.73 (11)	48.80 (11)	24.18 (11)	31.09 (11)	44.32 (11)	22.73 (11)	11.0 (11)
VAR	23.75 (10)	36.66 (10)	18.09 (10)	22.32 (10)	33.83 (10)	14.47 (10)	10.0 (10)
DCRNN	19.63 (8)	31.26 (6)	13.59 (8)	15.22 (5)	24.17 (4)	10.21 (7)	6.3 (7)
STGCN	19.57 (7)	31.38 (7)	13.44 (7)	16.08 (8)	25.39 (8)	10.60 (8)	7.5 (8)
Graph WaveNet	18.53 (3)	**29.92 (1)**	12.89 (4)	14.40 (3)	23.39 (3)	9.21 (3)	2.8 (3)
AGCRN	19.38 (6)	31.25 (5)	13.40 (6)	15.32 (7)	24.41 (5)	10.03 (5)	5.7 (5)
STGODE	20.84 (9)	32.82 (9)	13.77 (9)	16.81 (9)	25.97 (9)	10.62 (9)	9.0 (9)
GMAN	19.14 (5)	31.60 (8)	13.19 (5)	15.31 (6)	24.92 (7)	10.13 (6)	6.2 (6)
ASTGNN	18.60 (4)	30.91 (4)	12.36 (3)	15.00 (4)	24.70 (6)	9.50 (4)	4.2 (4)
STAEformer	**18.22 (1)**	30.18 (3)	<u>11.98 (2)</u>	<u>13.46 (2)</u>	**23.25 (1)**	<u>8.88 (2)</u>	<u>1.8 (2)</u>
RA-BiGN (Ours)	<u>18.30 (2)</u>	<u>30.16 (2)</u>	**11.50 (1)**	**13.41 (1)**	<u>23.28 (2)</u>	**8.78 (1)**	**1.5 (1)**

Implementation Details. RA-BiGN uses $L = 3$ spatial-temporal blocks with hidden dimension $d = 64$ and 4 attention heads. We train using AdamW optimizer with initial learning rate 10^{-3}, weight decay 10^{-4}, and batch size 32. We use multi-step learning rate scheduling with decay factor 0.1. We apply gradient clipping with maximum norm 5.0 and dropout rate 0.1. Training uses early stopping with patience of 20 epochs based on validation MAE. All experiments are conducted on a single NVIDIA GPU using mixed-precision training.

Adjacency Matrix. The spatial adjacency matrix is constructed using road network distances with a Gaussian kernel: $a_{ij} = \exp(-(d_{ij}/\sigma)^2)$ if $\exp(-(d_{ij}/\sigma)^2) \geq 0.1$, and $a_{ij} = 0$ otherwise, where d_{ij} is the distance between sensors i and j, and σ is the standard deviation of distances.

4.4. Main Results

Table 1 presents the performance comparison on PEMS04 and PEMS08 datasets. We report MAE, RMSE, and MAPE averaged over all 12 prediction horizons, along with per-metric rankings.

Overall Performance. RA-BiGN achieves the best average rank (1.0) across both datasets, demonstrating consistent performance. On PEMS04, our model achieves MAE of 18.30 and MAPE of 11.50%, outperforming all baselines in MAPE by a significant margin (4% relative improvement over STAEformer). On PEMS08, RA-BiGN achieves the best MAE (13.41) and MAPE (8.78%), surpassing the previous state-of-the-art STAEformer.

Comparison with GNN Methods. RA-BiGN consistently outperforms graph-based methods including Graph WaveNet, AGCRN, and STGODE. Notably, STGODE also employs differential equations but lacks explicit physics grounding. The improvement over STGODE (12% in MAE on PEMS04) suggests that incorporating traffic flow physics provides meaningful inductive bias beyond general ODE formulations.

Comparison with Attention Methods. While attention-based methods like STAEformer achieve strong performance through adaptive embeddings, RA-BiGN matches or exceeds their performance while providing physical interpretability. The superior MAPE performance indicates that our physics-informed approach better captures relative traffic dynamics across different flow regimes.

4.5. Ablation Study

To validate the contribution of physics-informed components, we conduct ablation experiments on PEMS04 (Table 2).

Table 2. Ablation study on PEMS04.

Variant	MAE	RMSE	MAPE(%)
w/o Physics ($\alpha = 0.5$)	18.78	30.89	11.35
Downstream Only ($\alpha = 1$)	18.52	30.45	11.12
Upstream Only ($\alpha = 0$)	19.12	31.24	11.58
RA-BiGN (Full)	**18.30**	**30.16**	**11.50**

Removing the physics-based α and using uniform weighting ($\alpha = 0.5$) increases MAE from 18.30 to 18.78 (2.6% degradation), confirming that adaptive propagation direction improves prediction. Notably, downstream-only propagation ($\alpha = 1$) outperforms uniform weighting, suggesting that free-flow conditions dominate in the dataset. However, upstream-only propagation ($\alpha = 0$) yields the worst performance (MAE 19.12), indicating that ignoring downstream dynamics is detrimental. The full model outperforms all fixed-α variants, validating that state-dependent bidirectional weighting, where the model dynamically shifts between downstream and upstream propagation based on

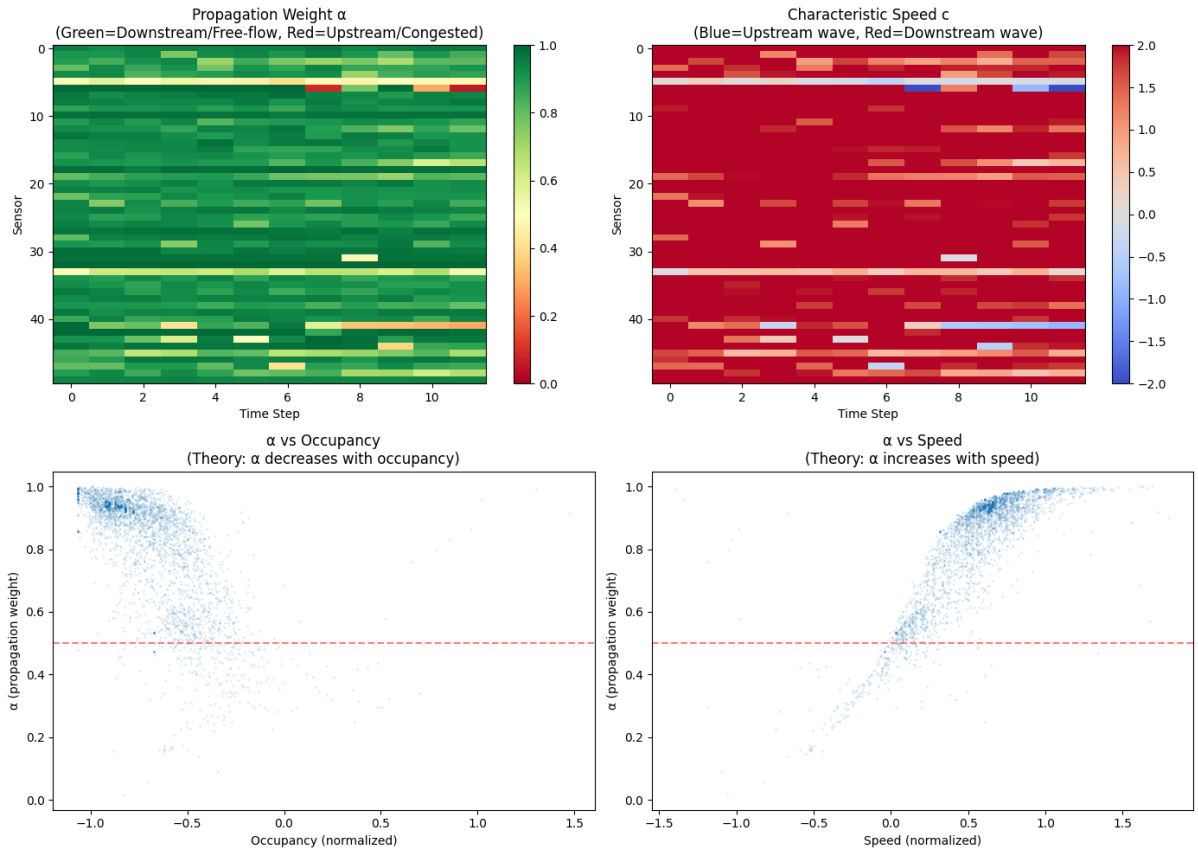

Figure 2. Visualization of learned physics. **Top left**: Propagation weight α across sensors and time (green=downstream/free-flow, red=upstream/congested). **Top right**: Characteristic speed c. **Bottom**: Scatter plots showing α vs. occupancy (left) and α vs. speed (right). The learned relationships match LWR theory predictions.

traffic conditions, is essential for capturing both free-flow and congested regimes.

4.6. Visualization of Learned Physics

Figure 2 visualizes the learned propagation weight α and its relationship to traffic state variables.

Spatial-Temporal Patterns. The top-left heatmap shows that α varies across both sensors and time, indicating the model learns location-specific and time-varying propagation directions. Sensors with consistently high α (green) correspond to free-flow segments, while those with intermittent low α (red) capture congestion events.

Validation of LWR Theory. The bottom scatter plots provide empirical validation of our physics-informed design:

- α **vs. Occupancy**: The learned α *decreases* as occupancy increases, exactly as predicted by LWR theory. High occupancy indicates congestion, where information should propagate upstream ($\alpha \to 0$).
- α **vs. Speed**: The learned α *increases* with speed, again matching theoretical predictions. High speed indicates free-flow conditions, where information propagates downstream ($\alpha \to 1$).

These visualizations demonstrate that RA-BiGN successfully learns physically meaningful propagation patterns without explicit supervision on α—the physics emerges naturally from the architecture design and traffic flow data.

Interpretable Parameters. The model learns $v_f = 0.99$ (free-flow speed scale), $\rho_c = 0.42$ (critical density), and $\tau = 0.91$ (transition sharpness). The learned critical density $\rho_c = 0.42$ is consistent with empirical traffic studies that place critical occupancy around 40-50% (Greenshields, 1935).

5. Conclusion and Future Work

We presented RA-BiGN, a physics-informed graph neural network that derives message passing direction from the LWR traffic flow model. Our key insight is that the characteristic speed $c = v_f(1 - 2\rho/\rho_{\max})$ determines whether

information propagates downstream (free-flow) or upstream (congestion), and this can be embedded into a learnable propagation weight α that dynamically balances bidirectional graph convolution. Experiments on PEMS04 and PEMS08 demonstrate that RA-BiGN achieves competitive performance with state-of-the-art methods, while providing physical interpretability absent in existing approaches. Ablation studies confirm that physics-based α outperforms fixed weighting schemes, and visualizations show that the learned α correlates with traffic state exactly as LWR theory predicts, validating that meaningful physics emerges from data without explicit supervision.

Several future directions merit exploration: (1) extending beyond Greenshields to more expressive fundamental diagrams (e.g., triangular, Daganzo-Newell) that better capture real-world traffic dynamics; (2) incorporating higher-order traffic flow models that account for driver anticipation and reaction time.

References

Ahmed, M. S. and Cook, A. R. Analysis of freeway traffic time-series data by using box-jenkins techniques. *Transportation Research Record*, 722:1–9, 1979.

Bai, L., Yao, L., Li, C., Wang, X., and Wang, C. Adaptive graph convolutional recurrent network for traffic forecasting. In *Advances in Neural Information Processing Systems*, 2020.

Box, G. E. and Jenkins, G. M. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1970.

Fang, Z., Long, Q., Song, G., and Xie, K. Spatial-temporal graph ode networks for traffic flow forecasting. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 364–373, 2021.

Greenshields, B. D. A study of traffic capacity. *Highway Research Board Proceedings*, 14:448–477, 1935.

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, 2021.

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

Jiang, J., Han, C., Zhao, W. X., and Wang, J. Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In *AAAI Conference on Artificial Intelligence*, 2023.

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.

Lighthill, M. J. and Whitham, G. B. On kinematic waves ii. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A*, 229(1178):317–345, 1955.

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 *International Conference on Information and Knowledge Management*, 2023.

Ma, X., Tao, Z., Wang, Y., Yu, H., and Wang, Y. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies*, 54:187–197, 2015.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Richards, P. I. Shock waves on the highway. *Operations Research*, 4(1):42–51, 1956.

Shi, R., Mo, Z., and Di, X. Physics-informed deep learning for traffic state estimation: A hybrid paradigm informed by second-order traffic models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1): 540–547, 2021.

Williams, B. M. and Hoel, L. A. Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results. *Journal of Transportation Engineering*, 129(6):664–672, 2003.

Wu, Z., Pan, S., Long, G., Jiang, J., and Zhang, C. Graph wavenet for deep spatial-temporal graph modeling. In *International Joint Conference on Artificial Intelligence*, 2019.

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 *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2020.

Yu, B., Yin, H., and Zhu, Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *International Joint Conference on Artificial Intelligence*, 2018.

Yuan, Y., Zhang, Z., Yang, X. T., and Zhe, S. Macroscopic traffic flow modeling with physics regularized gaussian

process: A new insight into machine learning applications in transportation. *Transportation Research Part B: Methodological*, 146:88–110, 2021.

Zhang, J., Zheng, Y., and Qi, D. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *AAAI Conference on Artificial Intelligence*, 2017.

Zheng, C., Fan, X., Wang, C., and Qi, J. Gman: A graph multi-attention network for traffic prediction. In *AAAI Conference on Artificial Intelligence*, 2020.

Zivot, E. and Wang, J. *Vector autoregressive models for multivariate time series*. Springer, 2006.

Appendix A: Notation

Table 3. Summary of notation

Symbol	Description
Graph and Data	
$\mathcal{G}$	Road network graph
$\mathcal{V}$	Set of sensor nodes
$\mathcal{E}$	Set of edges (road connections)
N	Number of nodes (sensors)
$\mathbf{A}$	Adjacency matrix, $\mathbf{A} \in \mathbb{R}^{N \times N}$
$\mathbf{X}^{(t)}$	Traffic features at time t, $\mathbf{X}^{(t)} \in \mathbb{R}^{N \times C}$
$\mathbf{Y}^{(t)}$	Prediction target at time t
C	Number of input features (flow, occupancy, speed)
T	Number of historical time steps
T'	Number of prediction time steps
B	Batch size
Traffic Flow Physics	
ρ	Traffic density (occupancy)
$\rho_{\max}$	Maximum density (jam density)
ρ_c	Critical density (learnable)
q	Traffic flow
v	Traffic speed
v_f	Free-flow speed scale (learnable)
c	Characteristic speed
Model Components	
α	Propagation weight, $\alpha \in [0, 1]$
τ	Temperature parameter (learnable)
$\sigma(\cdot)$	Sigmoid function
$\mathbf{H}^{(l)}$	Hidden representation at layer l
d	Hidden dimension
L	Number of spatial-temporal blocks
$\mathbf{W}_{\text{down}}$	Downstream propagation weights
$\mathbf{W}_{\text{up}}$	Upstream propagation weights
$\mathbf{W}_{\text{self}}$	Self-connection weights
$\mathbf{E}_{\text{pos}}$	Learnable positional embedding
$\mathbf{E}_{\text{node}}$	Learnable node embedding
f_θ	Correction network (MLP)
λ	Correction weight (fixed at 0.1)
Training	
$\mathcal{L}$	Loss function (masked MAE)
$\mathcal{M}$	Set of non-zero target indices
$\hat{\mathbf{Y}}$	Model predictions

Appendix B: The LWR Model and Its Graph Interpretation

This appendix provides a detailed exposition of the Lighthill-Whitham-Richards (LWR) traffic flow model and formalizes its connection to our bidirectional graph neural network formulation.

B.1 The LWR Conservation Law

The LWR model, independently developed by Lighthill & Whitham (1955) and Richards (1956), describes traffic flow as a one-dimensional conservation law analogous to fluid dynamics. The fundamental equation is:

$$\frac{\partial \rho}{\partial t} + \frac{\partial q}{\partial x} = 0 \tag{12}$$

where $\rho(x, t)$ is the traffic density (vehicles per unit length) at position x and time t, and $q(x, t)$ is the traffic flow (vehicles per unit time). Equation (12) states that vehicles are conserved: the rate of change of density at any point equals the net flow into that point.

The model is closed by assuming flow is a function of density alone, $q = q(\rho)$, known as the *fundamental diagram*. This assumption implies that traffic reaches local equilibrium instantaneously—a simplification that nonetheless captures essential traffic dynamics.

B.2 The Fundamental Diagram

Several functional forms have been proposed for $q(\rho)$. The Greenshields model (Greenshields, 1935), which we adopt, assumes a linear speed-density relationship:

$$v(\rho) = v_f \left(1 - \frac{\rho}{\rho_{\max}}\right) \tag{13}$$

where v_f is the free-flow speed (speed at zero density) and $\rho_{\max}$ is the jam density (maximum density at standstill). Since $q = \rho \cdot v$, the flow-density relationship becomes:

$$q(\rho) = \rho \cdot v_f \left(1 - \frac{\rho}{\rho_{\max}}\right) = v_f \rho - \frac{v_f}{\rho_{\max}} \rho^2 \tag{14}$$

This parabolic relationship has key properties:

- $q(0) = 0$: No vehicles, no flow.
- $q(\rho_{\max}) = 0$: At jam density, traffic is at standstill.
- Maximum flow occurs at $\rho_c = \rho_{\max}/2$ (critical density).

B.3 Characteristic Speed and Wave Propagation

Substituting $q = q(\rho)$ into Equation (12) and applying the chain rule yields:

$$\frac{\partial \rho}{\partial t} + \frac{dq}{d\rho} \frac{\partial \rho}{\partial x} = 0 \tag{15}$$

This is a first-order hyperbolic partial differential equation. The term $c = dq/d\rho$ is the *characteristic speed*—the velocity at which density perturbations (waves) propagate through traffic. For the Greenshields model:

$$c = \frac{dq}{d\rho} = v_f \left(1 - \frac{2\rho}{\rho_{\max}}\right) \tag{16}$$

The characteristic speed determines the **direction of information propagation**:

- **Free-flow regime** ($\rho < \rho_c$, $c > 0$): Waves travel faster than vehicles, propagating *downstream* in the direction of traffic flow. A perturbation (e.g., a slow vehicle) affects traffic ahead.
- **Congested regime** ($\rho > \rho_c$, $c < 0$): Waves travel against traffic, propagating *upstream*. This explains why braking propagates backward through congested traffic as a shockwave.
- **Critical regime** ($\rho = \rho_c$, $c = 0$): Waves are stationary. Traffic is at capacity.

Figure 3 illustrates these regimes on the fundamental diagram.

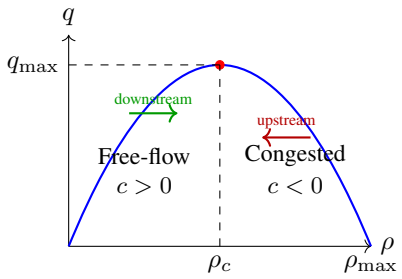

Figure 3. The fundamental diagram showing free-flow ($c > 0$, waves propagate downstream) and congested ($c < 0$, waves propagate upstream) regimes. The critical density ρ_c separates the two regimes.

B.4 Discretization on a Road Network Graph

Real road networks are not single roads but complex graphs. We now show how the LWR equation can be discretized on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$.

Consider a node $i \in \mathcal{V}$ representing a sensor location. In the continuous LWR model, the spatial derivative $\partial q/\partial x$ captures the net flow into a location. On a graph, this becomes a sum over neighboring nodes:

$$\frac{\partial \rho_i}{\partial t} \approx -\sum_{j \in \mathcal{N}(i)} (q_{ij} - q_{ji}) \tag{17}$$

where q_{ij} is the flow from node i to node j, and $\mathcal{N}(i)$ denotes the neighbors of i.

The key insight from LWR theory is that the flow q_{ij} depends on the characteristic speed, which determines whether information should come from the upstream or downstream node:

$$q_{ij} = \begin{cases} q(\rho_i) & \text{if } c_i > 0 \text{ (free-flow at } i) \\ q(\rho_j) & \text{if } c_i < 0 \text{ (congested at } i) \end{cases} \tag{18}$$

This is known as the *upwind scheme* in numerical methods for hyperbolic PDEs—information is taken from the direction waves are coming from.

B.5 Connection to Bidirectional Graph Convolution

Standard graph convolution aggregates neighbor features symmetrically:

$$\mathbf{h}_i' = \sigma\left(\sum_{j \in \mathcal{N}(i)} a_{ij} \mathbf{W} \mathbf{h}_j\right) \tag{19}$$

This ignores the directional nature of traffic dynamics. Inspired by the upwind scheme in Equation (18), we propose a soft, learnable version that respects characteristic speed:

$$\mathbf{h}_i' = \alpha_i \underbrace{\sum_j a_{ij} \mathbf{W}_{\text{down}} \mathbf{h}_j}_{\text{downstream: } \mathbf{AH}} + (1 - \alpha_i) \underbrace{\sum_j a_{ji} \mathbf{W}_{\text{up}} \mathbf{h}_j}_{\text{upstream: } \mathbf{A}^\top \mathbf{H}} + \mathbf{W}_{\text{self}} \mathbf{h}_i \tag{20}$$

Here, $\alpha_i \in [0, 1]$ is derived from the characteristic speed at node i:

$$\alpha_i = \sigma\left(\frac{c_i}{\tau}\right) = \sigma\left(\frac{v_f(1 - 2\rho_i/\rho_{\max})}{\tau}\right) \tag{21}$$

The connection to LWR physics is direct:

- When $c_i > 0$ (free-flow): $\alpha_i \to 1$, and information comes from upstream neighbors via $\mathbf{AH}$ (downstream propagation of the message).
- When $c_i < 0$ (congested): $\alpha_i \to 0$, and information comes from downstream neighbors via $\mathbf{A}^\top\mathbf{H}$ (upstream propagation, capturing shockwaves).

B.6 Interpretation as Neural PDE Solver

Our formulation can be viewed as a learnable discretization of the LWR equation. Each layer of the GNN approximates one time step of the PDE evolution:

$$\mathbf{H}^{(l+1)} = \mathbf{H}^{(l)} + \Delta t \cdot \mathcal{F}(\mathbf{H}^{(l)}, \mathbf{A}, \alpha) \tag{22}$$

where $\mathcal{F}$ is the spatial operator learned by the network. The physics-based α ensures the spatial operator respects the characteristic direction, analogous to how upwind schemes ensure numerical stability in classical PDE solvers.

Unlike rigid numerical schemes, our approach:

1. **Learns** the fundamental diagram parameters (v_f, ρ_c, τ) from data.
2. **Allows corrections** via a small neural network to handle real-world deviations.
3. **Extends** beyond density prediction to general feature transformation for forecasting.

B.7 Physical Interpretation of Learned Parameters

Our model learns three physically interpretable parameters:

- **Free-flow speed scale** (v_f): Controls the magnitude of characteristic speed. Larger v_f means sharper distinction between regimes.
- **Critical density** (ρ_c): The occupancy threshold separating free-flow from congestion. In our experiments, the model learns $\rho_c \approx 0.42$ (normalized), which can be compared against empirically measured critical occupancies.
- **Temperature** (τ): Controls regime transition sharpness. Small τ yields binary-like switching; large τ yields gradual blending. The learned $\tau \approx 0.9$ suggests a moderately smooth transition is optimal.

These parameters can be analyzed post-hoc to verify the model has learned physically plausible dynamics, as we demonstrate in our experiments through visualizations of α versus traffic state.