EpiRDN: A Learnable Anisotropic Reaction-Diffusion Network for Epidemic Time Series Prediction

Asela Hevapathige[1]

[1]School of Computing, Australian National University, Canberra, Australia

Abstract

Epidemic spread on graphs involves local infection dynamics and directional propagation through population movement. Current epidemic graph neural networks combine these aspects using a single symmetric aggregation operator, which limits the learning of their distinct timescales and fails to capture the asymmetry of real epidemic spread. This leads to node representations collapsing to a common state, losing diversity necessary for distinguishing infection stages. To address these issues, we propose EpiRDN, which discretizes the reaction-diffusion equation on graphs and learns its components end-to-end. EpiRDN features a local reaction network for region-level transitions and an anisotropic diffusion operator that utilizes asymmetric attention for directional infection flow. A feature-conditioned damping coefficient balances preserving local identity and neighborhood aggregation. We demonstrate that feature diversity in EpiRDN decays at most geometrically with depth, providing a significant advantage over existing methods. Experiments on four real-world datasets related to influenza and COVID-19 show consistent improvements, especially at longer forecasting horizons.

1 INTRODUCTION

Predicting epidemic time series on graphs is an important challenge, significantly impacting public health [Zeger et al., 2006, Zhang et al., 2014, Jin et al., 2025]. This task involves forecasting changes in infection counts across a network of regions over time [Liu et al., 2025, Babanejaddehaki et al., 2025]. Unlike standard graph learning, this process follows a directional flow, with the relationship between local infection dynamics and cross-regional spread evolving as the epidemic progresses. Accurate multi-step predictions require maintaining distinct representations for different regions at various infection stages.

Early epidemic forecasting methods relied on compartmental models [Brauer, 2008] and statistical approaches [Bharambe and Kalbande, 2016], which are based on fixed mechanistic assumptions about transmission dynamics. However, these methods often struggle with accurate parameter estimation and adapting to diverse, data-driven patterns. Deep learning offers a solution by learning complex non-linear dynamics directly from data [Maaliw et al., 2021, Ahmed et al., 2022, Bamana et al., 2024]. Yet, deep learning models focused solely on temporal aspects often overlook the relational structures that influence epidemic spread. Graph neural networks (GNNs) have been developed to incorporate these structures, allowing for topology-aware propagation of epidemic states [Wu et al., 2020, Liu et al., 2024, Xu et al., 2025]. Many recent approaches combine graph convolutional networks and graph attention models with temporal sequence models and incorporate domain-specific signals, such as mobility patterns and historical incidence curves, to enhance epidemic forecasting [Deng et al., 2020, Xie et al., 2022, Liu et al., 2023, Zheng et al., 2024, Zhu et al., 2024, Fujita and Akutsu, 2025]. Despite their successes, these methods face several limitations, especially when it comes to longer forecasting horizons:

- **Isotropic Propagation.** Current message-passing and epidemic GNNs [Gilmer et al., 2017, Liu et al., 2025] symmetrically aggregate information from neighboring nodes. However, infection pressure flows asymmetrically along contact and mobility networks—from high-prevalence to susceptible regions. This symmetry fails to capture the directionality inherent in the spread of disease.
- **Conflation of Local Dynamics and Spatial Spread.** Epidemic processes consist of two distinct mechanisms: local state transitions governed by intrinsic transmission and recovery rates, and the spatial propa-

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:2137–2151.

gation of infection pressure influenced by population movement. Most existing epidemic GNNs [Yan et al., 2018, Deng et al., 2020, Xie et al., 2022] combine these dynamics into a single graph convolution operation, losing the crucial distinction between local behavior and spatial interactions.

- **Feature Collapse under Temporal Depth.** When predicting multi-step epidemics, it's essential to have enough steps for the propagation. However, if aggregation is applied repeatedly, the node representations tend to converge into a common steady state [Li et al., 2018a, Jin and Zhu, 2025]. This results in a loss of the diversity necessary to differentiate between areas at various stages of infection.

All the above limitations arise from the same architectural decision: collapsing the distinct mechanisms of epidemic spread into a single symmetric aggregation operator.

We introduce **EpiRDN** (Epidemic Reaction-Diffusion Network), which discretises the classical reaction-diffusion equation [Murray, 2006] on graphs and learns both components end-to-end. Each propagation step decomposes into a local reaction network capturing region-level state transitions and an anisotropic diffusion operator that encodes directional infection flow via asymmetric attention. A feature-conditioned damping coefficient balances local identity preservation against neighbourhood aggregation at each node and time step. EpiRDN breaks the information propagation symmetry at the level of the diffusion operator, maintains an explicit separation between local dynamics and spatial spread, and provides a formal guarantee against feature collapse.

Our main contributions are as follows:

- **Anisotropic Reaction-Diffusion Network.** We formulate epidemic propagation as a learnable reaction-diffusion process on graphs, decomposing each propagation step into a local reaction and an anisotropic diffusion operator.
- **Feature-Conditioned Damping.** We introduce a per-node, per-step scalar coefficient conditioned on the current node state. The damping floor is learned jointly with all other parameters via a lightweight regularisation term, making the anti-collapse guarantee data-adaptive rather than hand-tuned.
- **Formal Guarantee against Feature Collapse.** We prove that feature diversity decays at most geometrically with propagation depth, avoiding feature collapse at any finite depth. This guarantee holds under learned parameters and is directly relevant for accurate long-horizon prediction.

We evaluate EpiRDN on four real-world datasets spanning influenza and COVID-19. EpiRDN consistently outperforms or provides competitive performance over baselines across all settings, with particularly pronounced gains at longer forecasting horizons where feature collapse most severely degrades competing approaches.

2 RELATED WORK

Message-passing GNNs. Message-passing GNNs learn node representations by iteratively aggregating neighbourhood information via symmetric message passing [Kipf and Welling, 2017, Hamilton et al., 2017]. Traditional message-passing methods treat all edges identically, making them incapable of encoding the directed, asymmetric flow inherent to epidemic spread. Attention-based methods such as GAT [Veličković et al., 2018] and GATv2 [Brody et al., 2022] partially address this by weighting edges by feature similarity, but project source and target features through a shared weight matrix, preventing the model from learning structurally different representations for sender and receiver nodes. A natural extension is to separate the treatment of incoming and outgoing edges altogether, and DirGNN [Rossi et al., 2024] takes this step by maintaining distinct aggregation channels for each direction. However, this separation is a purely architectural correction with no grounding in underlying propagation dynamics, and offers no protection against feature collapse under depth. In EpiRDN, separate projection matrices for source and target nodes are embedded within the diffusion operator itself, coupling directionality to the propagation mechanism rather than imposing it as a post-hoc structural choice.

Diffusion GNNs. A growing line of work frames message passing as a continuous dynamical system. GRAND [Chamberlain et al., 2021a] and GRAND++ [Thorpe et al., 2022] model feature evolution via graph Laplacian ODEs, BLEND [Chamberlain et al., 2021b] extends this to Beltrami flow, and GraphCON [Rusch et al., 2022] introduces oscillator dynamics to slow energy decay. ACMP [Wang et al., 2023] and GREAD [Choi et al., 2023] augment diffusion with reaction terms, but both fix the functional form of the reaction to physics-inspired constraints rather than learning it from data. ADR-GNN [Eliasof et al., 2024] introduces directed edge velocities as a separate advection term combined with symmetric diffusion and a generic reaction.

Despite their appeal, these methods struggle with three key challenges in epidemic time series prediction. First, their symmetric diffusion operators fail to capture the directed infection flow from high-prevalence to susceptible areas. Second, they are optimized for static graphs and don't effectively manage historical data for future predictions. Third, they lack a strategy for maintaining diversity in representation across multiple forecasting steps, often relying on heuristic methods. EpiRDN addresses these issues by learning an anisotropic diffusion operator, using a feature gener-

ation stage for initial hidden states, and ensuring Dirichlet energy retention with learned parameters.

Epidemic GNNs. Spatio-temporal GNNs [Yan et al., 2018, Li et al., 2018b, Corradini et al., 2025] combine graph convolution with sequence models for spatial flow prediction. Epidemic forecasting methods [Deng et al., 2020, Xie et al., 2022, Liu et al., 2023, Zheng et al., 2024] specialise this paradigm for epidemic spread by incorporating domain-specific information injection and multi-scale temporal signal modelling. Despite their performance improvements, these methods share several limitations. Their propagation operators are isotropic and cannot capture the directed, asymmetric flow that characterizes real epidemic processes. They also combine reaction and diffusion dynamics within a single aggregation operator, preventing the model from learning these two processes independently. In addition, they lack a mechanism to preserve representational diversity as propagation depth increases. EpiRDN addresses all these limitations.

3 BACKGROUND

3.1 PRELIMINARIES

Let $G = (V, E, A)$ be a graph with nodes V, $|V| = n$, edges E, $|E| = m$, and adjacency matrix $A \in \{0,1\}^{n\times n}$. Each node i has a feature vector $\mathbf{h}_i \in \mathbb{R}^f$, collected in $\mathbf{H} \in \mathbb{R}^{n\times f}$, and $\mathcal{N}(i) = \{j \in V : (i,j) \in E\}$ denotes its neighbourhood. In the epidemic forecasting setting, each node corresponds to a geographic region and edges encode population movement or contacts between regions.

3.2 EPIDEMIC TIME SERIES PREDICTION

Each region $i \in V$ is associated with a time series of observed infection counts $y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(T)} \in \mathbb{R}_{\geq 0}$. Let $\mathbf{Y}^{(t)} = [y_1^{(t)}, \ldots, y_n^{(t)}]^\top \in \mathbb{R}^n$ denote the system-wide incidence vector at time t.

Definition 1 (Epidemic Time Series Prediction)**.** *Given the infection counts of all n regions over a historical window of w time steps, the epidemic time series prediction problem is to learn a function $\mathcal{F}_\theta$ such that*

$$\hat{\mathbf{Y}}^{(T+1:T+\tau)} = \mathcal{F}_\theta\Big(\mathbf{Y}^{(T-w+1:T)},\, G\Big), \tag{1}$$

where $\mathbf{Y}^{(T-w+1:T)} \in \mathbb{R}^{n\times w}$ collects the incidence windows of all regions over the most recent w steps, $\hat{\mathbf{Y}}^{(T+1:T+\tau)} \in \mathbb{R}^{n\times\tau}$ are the predicted counts over the next τ time steps, and θ denotes learnable parameters.

The model is optimised by minimising a reconstruction loss between $\hat{\mathbf{Y}}$ and the ground truth future counts $\mathbf{Y}^{(T+1:T+\tau)}$ over a set of training windows.

3.3 GRAPH NEURAL NETWORKS

Graph Neural Networks (GNNs) learn node representations by iteratively combining feature information with graph topology [Wu et al., 2020]:

$$\mathbf{h}_i^{(\ell+1)} = \mathrm{UPD}\Big(\mathbf{h}_i^{(\ell)},\ \mathrm{AGG}\Big(\{\mathbf{h}_j^{(\ell)} : j \in \mathcal{N}(i)\}\Big)\Big), \tag{2}$$

where AGG is a permutation-invariant aggregation function and UPD is a learned transformation. Standard GNNs treat all edges symmetrically and suffer from feature collapse under repeated aggregation [Li et al., 2018a, Jin and Zhu, 2025], limiting their utility for multi-step temporal prediction where distinct node representations must be maintained across many propagation steps.

3.4 DIFFUSION-BASED GNNS

A growing line of work frames message passing as a diffusion process [Chamberlain et al., 2021a, Gasteiger et al., 2018, 2019], evolving node features as:

$$\mathbf{H}^{(t+1)} = \mathbf{H}^{(t)} + \Phi\Big(\mathbf{H}^{(0)},\, \mathbf{H}^{(t)},\, G;\, \Theta\Big), \tag{3}$$

or in continuous time as the Ordinary Differential Equation (ODE):

$$\tfrac{d\mathbf{H}}{dt} = \Phi(\mathbf{H}^{(0)}, \mathbf{H}^{(t)}, G; \Theta) \tag{4}$$

These formulations connect GNNs to dynamical systems theory and yield strong empirical results on static tasks, yet existing methods retain isotropic propagation, treating all edges symmetrically regardless of the directed, asymmetric nature of underlying epidemic dynamics.

3.5 REACTION-DIFFUSION SYSTEMS

Directed propagation can be naturally motivated by the reaction-diffusion equation [Murray, 2006]:

$$\frac{\partial u}{\partial t} = R(u) + \nabla \cdot \big(D(x)\,\nabla u\big), \tag{5}$$

where $R(u)$ models local state changes such as infection and recovery dynamics at each region, and $D(x)$ is a spatially varying diffusivity tensor governing the rate and orientation of spatial spread. Isotropic diffusion, where $D(x) = cI$ for some scalar c and identity matrix I, corresponds to the symmetric graph Laplacian in standard GNN message passing. Anisotropic D enables heterogeneous, direction-dependent propagation, motivating graph propagation mechanisms that better capture the asymmetric contact structure of real epidemic networks. EpiRDN instantiates this distinction on graphs by learning both R and D end-to-end from epidemic data.

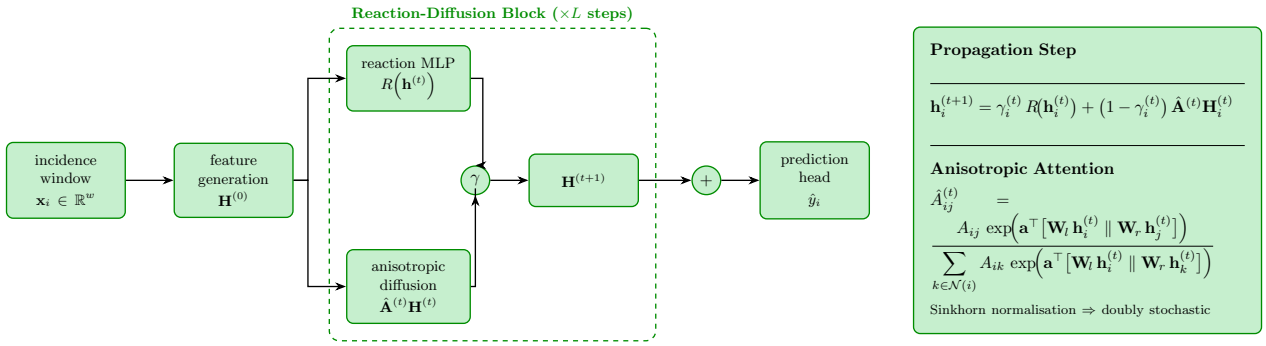

Figure 1: Architecture of EpiRDN. Each propagation step combines a reaction component modelling local infection dynamics with anisotropic diffusion spreading information across the graph, balanced by feature-conditioned damping.

4 METHODOLOGY

We propose EpiRDN, a learnable graph analogue of the reaction-diffusion system in Eq. (5). The framework operates in three stages: initial feature generation from raw epidemic time series, iterative reaction-diffusion propagation across the spatial graph, and the final prediction.

4.1 FEATURE GENERATION

Raw epidemic data associates each region i with a window of historical infection counts $\mathbf{x}_i \in \mathbb{R}^w$ covering the most recent w time steps. These scalar time series must be lifted into a latent space of sufficient dimension before graph-based propagation can capture the nuances of epidemic state. We map each region's incidence window to an initial hidden representation via a shared linear projection:

$$\mathbf{h}_i^{(0)} = W_{\text{in}}\,\mathbf{x}_i + \mathbf{b}_{\text{in}}, \tag{6}$$

where $W_{\text{in}} \in \mathbb{R}^{f\times w}$ and $\mathbf{b}_{\text{in}} \in \mathbb{R}^f$ are learned parameters shared across all regions. This projection serves two purposes: it compresses the temporal history into a fixed-size representation amenable to graph operations, and it allows the network to learn which temporal patterns in the incidence window are most predictive of future spread. The resulting matrix $\mathbf{H}^{(0)} \in \mathbb{R}^{n\times f}$ initialises the reaction-diffusion propagation described in the following subsections.

4.2 REACTION-DIFFUSION PROPAGATION

The per-node update at propagation step t is:

$$\mathbf{h}_i^{(t+1)} = \gamma_i^{(t)}\,R\Big(\mathbf{h}_i^{(t)}\Big) + \big(1-\gamma_i^{(t)}\big)\,\big(\hat{A}^{(t)}\mathbf{H}^{(t)}\big)_i, \tag{7}$$

where $R(\cdot)$ is a learned local reaction function, $\hat{A}^{(t)}$ is a learned anisotropic diffusion matrix, and $\gamma_i^{(t)} \in (0,1)$ is a state-dependent damping coefficient balancing local transformation against neighbourhood aggregation. In matrix form, this can be written as:

$$\mathbf{H}^{(t+1)} = \mathbf{\Gamma}^{(t)}R\big(\mathbf{H}^{(t)}\big) + \big(I-\mathbf{\Gamma}^{(t)}\big)\hat{A}^{(t)}\mathbf{H}^{(t)}, \tag{8}$$

with $\mathbf{\Gamma}^{(t)} = \text{diag}(\gamma_1^{(t)},\dots,\gamma_n^{(t)})$. Here $\mathbf{\Gamma}^{(t)}(R-I)$ plays the role of the reaction operator and $(I-\mathbf{\Gamma}^{(t)})(I-\hat{A}^{(t)})$ the anisotropic diffusion operator in Eq. (5).

4.3 DIRECTIONAL ANISOTROPIC DIFFUSIVITY

The propagation matrix $\hat{A}^{(t)}$ encodes directional flow strengths via learned attention:

$$\hat{A}_{ij}^{(t)} = \frac{A_{ij}\exp\Big(\mathbf{a}^\top\big[W_l\mathbf{h}_i^{(t)} \,\|\, W_r\mathbf{h}_j^{(t)}\big]\Big)}{\sum_{k\in\mathcal{N}(i)}\exp\Big(\mathbf{a}^\top\big[W_l\mathbf{h}_i^{(t)} \,\|\, W_r\mathbf{h}_k^{(t)}\big]\Big)}, \tag{9}$$

where $W_l, W_r \in \mathbb{R}^{d\times f}$ are learned source and target projections and $\mathbf{a} \in \mathbb{R}^{2d}$ is a learned scoring vector. Softmax normalisation guarantees row-stochasticity by construction.

Remark 1 (Sinkhorn Double Stochasticity)**.** *For the theoretical analysis in Section 5 we require $\hat{A}^{(t)}$ to be* doubly stochastic. *This is enforced by applying one step of Sinkhorn normalisation [Sinkhorn, 1964] after the softmax row-normalisation in Eq. (9), i.e., divide each row by its sum, then each column by its sum, repeating until convergence. A single Sinkhorn iteration suffices in practice and adds $\mathcal{O}(|E|)$ cost, preserving the overall complexity analysis. The asymmetry property (Lemma 1) is unaffected because the Sinkhorn iteration does not equalise $\hat{A}_{ij}^{(t)}$ and $\hat{A}_{ji}^{(t)}$ in general.*

The following lemma formalizes how the asymmetry of the learned attention weights is both necessary and sufficient for EpiRDN to capture directed propagation. The proof is provided in Appendix 8.

Lemma 1 (Directional Asymmetry). *$\hat{A}_{ij}^{(t)} = \hat{A}_{ji}^{(t)}$ for all i, j and all $\mathbf{H}^{(t)}$ if and only if the attention kernel is symmetric. In particular, $W_l \neq W_r$ generically implies $\hat{A}_{ij}^{(t)} \neq \hat{A}_{ji}^{(t)}$.*

Remark 2 (Role-Specific Expressivity). *When $W_l = W_r$ the attention scores become symmetric in the sense that the source and target features are projected into the same space before scoring, collapsing the role-specific expressivity that enables directed propagation. The separation $W_l \neq W_r$ allows the model to learn different representations for source and target nodes at each edge by encoding them with distinct transformations, rather than relying solely on their order in the concatenated feature vector as in existing graph attention mechanisms.*

4.4 FEATURE-CONDITIONED DAMPING

The damping coefficient is conditioned on the current node state:

$$\gamma_i^{(t)} = \sigma\Big(\mathbf{w}^\top \mathbf{h}_i^{(t)} + b\Big), \tag{10}$$

with σ the sigmoid. To guarantee a diversity-preserving floor, we learn a scalar $\underline{\gamma}$ and include a single regularisation term in the training objective:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \sum_{i,t} \max\big(0,\ \underline{\gamma} - \gamma_i^{(t)}\big)^2. \tag{11}$$

This encourages $\gamma_i^{(t)} \geq \underline{\gamma}$ during training, where $\underline{\gamma}$ is itself learned alongside all other parameters. The theoretical guarantee in Section 5.2 is stated conditionally on this holding for the trained model.

4.5 FINAL PREDICTION

After L propagation steps, each region's final hidden state $\mathbf{h}_i^{(L)}$ is mapped to a prediction by augmenting the learned graph representation with a direct readout from the most recent w_{hw} time steps of the raw incidence window:

$$\hat{y}_i = \mathbf{w}_{\text{out}}^\top \mathbf{h}_i^{(L)} + \mathbf{w}_{\text{hw}}^\top \mathbf{x}_i^{(w_{\text{hw}})} + b, \tag{12}$$

where $\mathbf{x}_i^{(w_{\text{hw}})}$ denotes the last w_{hw} entries of region i's incidence window. The high-level architecture of EpiRDN is depicted in Figure 1.

5 THEORETICAL ANALYSIS

5.1 MODEL COMPLEXITY

The per-step cost of EpiRDN is $\mathcal{O}(|E| + nf^2)$: $\mathcal{O}(|E|)$ for attention and Sinkhorn normalisation over the edge set, $\mathcal{O}(nf^2)$ for the node-wise reaction MLP, and $\mathcal{O}(n)$ for the damping coefficient. The attention cost dominates on sparse graphs and matches that of standard message-passing GNNs [Kipf and Welling, 2017, Veličković et al., 2018], with the reaction and damping components adding no asymptotic overhead. The total parameter count is $\mathcal{O}(df + f^2)$, with $d \leq f$ typically.

5.2 RESISTANCE TO FEATURE COLLAPSE

The central theoretical contribution of EpiRDN is a guarantee against feature collapse at any finite propagation depth, established under the following assumptions. The proof is provided in Appendix section 8.

Assumption 5.1 (Damping Floor). *The trained model produces $\gamma_i^{(t)} = \underline{\gamma}$ for all nodes i and steps t, where $\underline{\gamma} \in (0, 1)$ is the learned scalar from Eq. (11).*

The regularisation term in Eq. (11) penalises node-step pairs where $\gamma_i^{(t)}$ drops below $\underline{\gamma}$, encouraging all damping coefficients to remain above this floor. This ensures a minimum fraction of each node's update comes from its local reaction, maintaining its identity at every layer. The full model with feature-conditioned $\gamma_i^{(t)} \geq \underline{\gamma}$ satisfies this assumption conservatively; Theorem 2 therefore provides a lower bound on the diversity retained in practice.

Assumption 5.2 (Bi-Lipschitz Reaction). *The trained reaction satisfies $L_{\min}\|\mathbf{u} - \mathbf{v}\| \leq \|R(\mathbf{u}) - R(\mathbf{v})\|$ for some constant $L_{\min} > 0$.*

This is a mild non-degeneracy condition for the MLP R: it cannot map distinct inputs to the same output while preserving separation proportional to $L_{\min}$. Standard initialisation and weight-decay regularisation help prevent the MLP from becoming a near-constant function. In practice, we implement R as a residual MLP, which provides a structural guarantee that distinct inputs remain separated, making Assumption 5.2 an architectural property rather than a post-hoc condition on the trained model.

Assumption 5.3 (Doubly Stochastic Diffusion). *At every step t, the propagation matrix $\hat{A}^{(t)}$ is doubly stochastic: $\hat{A}^{(t)}\mathbf{1} = \mathbf{1}$ and $\hat{A}^{(t)\top}\mathbf{1} = \mathbf{1}$, i.e., all row sums and all column sums equal one.*

As noted in Remark 1, this is enforced at negligible extra cost by applying Sinkhorn normalisation after the softmax in Eq. (9). Double stochasticity ensures that diffusion preserves the population mean $\bar{\mathbf{h}}^{(t)}$ and admits the contraction bound used to establish Theorem 2. The full argument is given in Appendix Section 8.

Theorem 2 (Dirichlet Energy Retention). *Let $\pi_i = 1/n$ for all $i \in V$. Define the feature diversity measure $\widetilde{E}(\mathbf{H}) = \frac{1}{n}\sum_i \|\mathbf{h}_i - \bar{\mathbf{h}}\|^2$ where $\bar{\mathbf{h}} = \frac{1}{n}\sum_i \mathbf{h}_i$. Under Assumptions 5.1–5.3, and provided*

$$\underline{\gamma}(1 + L_{\min}) > 1, \tag{13}$$

Table 1: Epidemic forecasting performance across four datasets measured by RMSE, where lower values indicate better performance. All values are scaled by $\times 10^3$. **Blue** indicates the best result and red indicates the second best result in each column.

	Japan-Prefectures				US-Region				US-State				Australia-COVID			
Method	2	5	7	12	2	5	7	12	2	5	7	12	2	5	7	12
AR	1.670	2.440	2.561	2.485	0.626	1.079	1.280	1.496	0.172	0.264	0.300	0.341	0.509	0.540	0.556	0.587
GAR	1.484	2.428	2.561	2.530	0.607	1.054	1.288	1.568	0.162	0.247	0.292	0.355	1.343	1.394	1.430	1.519
VAR	1.577	2.566	2.530	2.504	0.695	1.113	1.234	1.380	0.298	0.333	0.397	0.422	0.581	0.592	0.607	0.619
ARMA	1.661	2.439	2.559	2.482	0.622	1.058	1.263	1.484	0.172	0.262	0.298	0.340	0.505	0.527	0.541	0.564
RNN	1.198	1.852	2.193	1.906	0.572	0.980	1.154	1.513	0.159	0.224	0.255	0.329	0.581	0.648	0.708	0.898
LSTM	1.255	1.840	2.235	1.952	0.570	1.049	1.230	1.531	0.161	0.227	0.263	0.330	1.122	1.178	1.230	1.393
GRU	1.181	1.782	2.226	1.922	0.571	1.024	1.205	1.521	0.159	0.248	0.285	0.335	1.065	1.127	1.177	1.322
RNN-Attn	1.311	2.267	2.133	1.933	0.566	0.999	1.211	1.521	0.163	0.244	0.274	0.342	1.054	1.039	1.026	1.166
CNNRNN-Res	1.401	2.570	2.603	2.067	0.595	1.006	1.153	1.316	0.228	0.320	0.360	0.280	1.484	1.488	1.491	1.476
LSTNet	1.441	2.385	2.477	2.391	0.607	1.134	1.121	1.149	0.196	0.276	0.288	0.312	0.595	0.614	0.645	0.712
GCN	1.166	1.490	1.631	1.772	0.786	0.940	1.052	1.209	0.204	0.247	0.253	0.271	1.023	1.034	0.992	1.081
GAT	1.199	1.489	1.678	1.687	0.772	1.001	1.108	1.361	0.201	0.246	0.275	0.312	0.745	0.663	0.631	0.572
GraphSAGE	1.188	1.565	1.529	1.732	0.527	0.876	0.939	1.020	0.145	0.201	0.220	0.246	0.359	0.402	0.365	0.434
GATv2	1.184	1.430	1.613	1.771	0.780	0.999	1.122	1.303	0.200	0.248	0.275	0.309	0.383	0.342	0.401	0.404
DirGNN	1.162	1.487	1.563	1.731	0.612	0.856	1.001	1.052	0.152	0.206	0.225	0.248	0.381	0.331	0.445	0.484
GRAND	1.371	2.176	2.366	2.149	0.557	0.969	1.124	1.328	0.151	0.219	0.251	0.279	0.353	0.322	0.371	0.332
GRAND++	1.374	2.178	2.365	2.149	0.556	0.969	1.128	1.335	0.150	0.220	0.255	0.287	0.352	0.327	0.379	0.332
ACMP	1.347	1.648	1.782	2.160	0.546	0.889	1.007	1.192	0.155	0.206	0.229	0.260	0.350	0.398	0.324	0.442
GREAD	1.126	1.609	2.103	1.878	0.525	0.920	1.012	1.051	0.153	0.261	0.404	0.420	0.387	0.414	0.386	0.409
ADR-GNN	1.399	2.201	2.228	2.173	0.535	0.924	1.070	1.267	0.160	0.213	0.244	0.268	0.370	0.407	0.381	0.517
STGCN	1.274	1.400	**1.478**	1.812	0.727	0.965	1.017	1.116	0.205	0.260	0.280	0.296	0.696	0.649	0.706	0.721
MGNN	1.757	1.943	2.048	2.655	1.472	1.487	1.305	1.538	0.178	0.206	0.259	0.288	0.664	0.769	0.848	0.857
TMGNN	1.500	1.559	1.977	1.839	0.656	0.916	1.043	1.228	0.179	**0.190**	0.438	0.447	0.803	0.846	0.870	0.916
Cola-GNN	1.168	1.573	1.758	1.690	0.552	0.871	1.054	1.011	0.148	0.224	0.228	0.241	0.368	0.406	0.463	0.533
Epi-GNN	1.503	1.448	1.738	1.695	0.563	0.881	0.938	1.025	0.150	**0.190**	**0.205**	0.251	**0.325**	0.381	0.416	0.491
Epi-Cola-GNN	1.118	1.544	2.549	2.164	0.547	0.872	1.160	1.365	0.148	0.219	0.268	0.303	0.803	0.854	0.912	1.124
EpiRDN	**1.113**	**1.393**	1.495	**1.683**	**0.513**	**0.834**	**0.920**	**1.001**	**0.144**	0.196	0.217	**0.237**	0.340	**0.318**	**0.291**	**0.321**

the following lower bound holds for all $T \geq 0$:

$$\widetilde{E}\left(\mathbf{H}^{(T)}\right) \geq c_{\text{low}}^{2T}\, \widetilde{E}\left(\mathbf{H}^{(0)}\right), \qquad c_{\text{low}} = \underline{\gamma}(1 + L_{\min}) - 1 > 0. \tag{14}$$

In particular, $\widetilde{E}(\mathbf{H}^{(T)}) > 0$ for all finite T whenever $\widetilde{E}(\mathbf{H}^{(0)}) > 0$, preventing feature collapse at any finite depth.

$\widetilde{E}(\mathbf{H}^{(t)})$ measures the spread of node representations around their mean, being zero only when all nodes share identical features. The theorem states this diversity cannot vanish in finite time, as it is lower-bounded by a quantity that decays at most geometrically. The condition $\underline{\gamma}(1+L_{\min}) > 1$ requires the local reaction to amplify diversity quickly enough to counteract the homogenising effect of diffusion. Higher $\underline{\gamma}$ or $L_{\min}$ increases c_{low}, strengthening the guarantee. In contrast, standard GNNs with $\underline{\gamma} = 0$ yield $c_{\text{low}} = -1 < 0$, meaning no such bound can be established, consistent with the geometric energy decay observed under pure diffusion [Chamberlain et al., 2021a].

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

Datasets We evaluate on four real-world epidemic datasets [Xie et al., 2022] covering multiple wave outbreaks.
.

- **Japan-Prefectures:** Weekly influenza-like illness(ILI) reports for 47 prefectures in Japan (Aug 2012 – Mar 2019), sourced from the Weekly Infectious Diseases Report.
- **US-Regions:** Weekly influenza patient counts across 10 HHS regions (2002–2017), sourced from the Department of Health and Human Services.
- **US-States:** Weekly influenza-positive cases across 49 US states (2010–2017), sourced from the CDC.
- **Australia-COVID:** Daily confirmed COVID-19 cases across 8 states and territories (Jan 2020 – Aug 2021), sourced from the JHU-CSSE repository.

Table 2 reports the size, temporal granularity, and descriptive statistics of the four datasets. Min and Max denote the minimum and maximum recorded case counts across all

regions and time steps; Mean is the average case count per region per time step.

Evaluation Metric We adopt Root Mean Squared Error (RMSE) as the primary evaluation metric:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{15}$$

where y_i and $\hat{y}_i$ denote the ground-truth and predicted case counts respectively. RMSE penalises large deviations more heavily than mean absolute error, making it particularly sensitive to peak outbreak episodes that are critical for public health decision-making.

Baselines We compare EpiRDN against five categories of baselines spanning a broad range of inductive biases.

- **Statistical models:** AR, VAR, GAR, and ARMA [Zhang, 2018] capture temporal autocorrelation without spatial structure.
- **Message-passing GNNs:** GCN [Kipf and Welling, 2017], GAT [Veličković et al., 2018], GraphSAGE [Hamilton et al., 2017], GATv2 [Brody et al., 2022], and DirGNN [Rossi et al., 2024] cover isotropic aggregation, attention-based weighting, inductive sampling, and directed propagation.
- **Diffusion GNNs:** GRAND [Chamberlain et al., 2021a], GRAND++ [Thorpe et al., 2022], GREAD [Choi et al., 2023], ACMP [Wang et al., 2023], and ADR-GNN [Eliasof et al., 2024] model diffusion processes on graphs.
- **Sequence models:** RNN [Medsker and Jain, 2001], LSTM [DiPietro and Hager, 2020], GRU [Dey and Salem, 2017], RNN-Attn [Cheng et al., 2016], and CNNRNN-Res [Wu et al., 2018] capture temporal dynamics without explicit spatial modelling.
- **Epidemic and spatio-temporal GNNs:** LSTNet [Lai et al., 2018], STGCN [Yu et al., 2018], MGNN [Hy et al., 2022], TMGNN [Hy et al., 2022], Cola-GNN [Deng et al., 2020], EpiGNN [Xie et al., 2022], and EpiCola-GNN [Liu et al., 2023] represent the most relevant prior work, combining spatial graph structure with temporal modelling under epidemic-specific or spatio-temporal inductive biases.

Evaluation setting We split datasets into training, validation, and test sets at a 60%–20%–20% ratio. We perform 5-fold cross-validation and report the mean RMSE across folds. While message-passing GNNs and diffusion GNNs are primarily designed for static graph tasks, we adapt them for epidemic time series prediction by feeding each node's historical incidence window as its input feature and evaluating their performance under a multi-step forecasting protocol. Baseline results are obtained from the literature when reported under the same experimental setup. Otherwise, we reproduce the baselines using the hyperparameter configurations specified in their original papers.

Model Hyperparameters Models are optimised using Adam optimizer [Kingma and Ba, 2015] with a learning rate selected from $\{10^{-3}, 2\times10^{-3}, 5\times10^{-3}\}$ and weight decay from $\{10^{-4}, 2\times10^{-4}, 5\times10^{-4}\}$. Dropout is tuned over $\{0.1, 0.2, 0.5\}$ and batch size over $\{32, 64\}$. The number of layers is selected from $\{2, 4, 8\}$. All hyperparameters are chosen via grid search [Bergstra and Bengio, 2012]. Training runs for up to 1500 epochs with early stopping if validation loss does not improve for 200 consecutive epochs.

Implementation Details All experiments were run on a Linux workstation with a 28-core Intel Xeon W-2175 CPU (2.50 GHz), 512 GB of RAM, and an NVIDIA RTX A6000 GPU. The codebase was developed in Python [Sundnes, 2020].

6.2 PREDICTION PERFORMANCE

Table 1 reports RMSE across four datasets and four forecasting horizons. EpiRDN achieves the best overall result, winning 12 of 16 settings and placing second in the remaining 4. It consistently outperforms all five baseline categories: statistical models, standard message-passing GNNs (including attention-based and direction-aware variants), sequential models, diffusion GNNs, and epidemic-specific architectures. While individual baselines perform competitively in specific settings (STGCN at short horizons on Japan-Prefectures, Epi-GNN on US-State mid-range horizons, and GRAND on Australia-COVID), no single baseline sustains strong performance across all settings simultaneously. EpiRDN is the only method that achieves consistently near-optimal results across all configurations.

Another notable pattern observed in the results is that EpiRDN's advantage tends to widen with the forecasting horizon on most datasets. Methods using symmetric aggregation, message-passing GNNs, sequential models, and epidemic-specific baselines show sharper performance drops at longer horizons, whereas EpiRDN exhibits more stable error. This stability aligns with EpiRDN's feature diversity guarantee, which maintains distinct representations.

6.3 ABLATION STUDY

To isolate the contribution of each component, we evaluate six variants of EpiRDN by removing one design choice at a time. Results are reported in Table 3.

All variants generally underperform the full model, confirming that each component contributes. Removing the reaction term causes the largest degradation at most horizons across both datasets, demonstrating that explicit local

Table 2: Full statistics of the four real-world datasets.

Dataset	Size	Granularity	Disease	Min	Max	Mean
Japan-Prefectures	47×348	Weekly	Influenza	0	26635	655
US-Regions	10×785	Weekly	Influenza	0	16526	1009
US-States	49×360	Weekly	Influenza	0	9716	223
Australia-COVID	8×556	Daily	COVID-19	0	9987	539

Table 3: Ablation study on US-Region and US-State across four forecasting horizons (RMSE $\times 10^3$ reported). Each variant removes one component of EpiRDN while keeping all others fixed.

	US-Region				US-State			
Variant	2	5	7	12	2	5	7	12
w/o anisotropy	0.528	0.886	0.942	1.018	0.151	0.218	0.266	0.279
w/o reaction	0.796	0.977	1.025	1.009	0.217	0.282	0.283	0.280
w/o damping floor	0.543	0.884	0.984	1.057	0.147	0.282	0.282	0.283
w/o Sinkhorn norm.	0.541	0.871	0.953	1.020	0.151	0.216	0.249	0.278
w/o feature cond.	0.540	0.882	0.948	1.081	0.144	0.202	0.233	0.287
w/o regularization	0.537	0.836	**0.913**	1.023	0.149	0.210	0.271	0.285
Full EpiRDN	**0.513**	**0.834**	0.920	**1.001**	**0.144**	**0.196**	**0.217**	**0.237**

state transition modelling is a critical component for capturing region-level infection dynamics. Removing the damping floor has a disproportionate effect at longer horizons, consistent with the theoretical prediction that feature diversity must be actively maintained as propagation depth grows. The anisotropy ablation produces degradation across all horizons, though its magnitude varies, confirming that directed propagation matters throughout the rollout rather than only at higher depth. The regularisation ablation reveals an asymmetric effect on the US-State dataset: short-horizon performance is largely unaffected, while longer horizons degrade noticeably, suggesting the floor penalty becomes more important when deeper propagation risks diversity loss. Sinkhorn normalisation has a comparatively modest impact, suggesting it acts as a stabiliser rather than a primary performance driver; feature-conditioned damping shows modest impact at short horizons but becomes increasingly important at longer horizons, where it produces the largest degradation of any ablation.

6.4 LONG-HORIZON ROBUSTNESS

Figure 2 evaluates prediction performance across increasing window sizes on the Australia-COVID dataset, comparing EpiRDN against baselines with diverse inductive biases: GAT employs multi-head attention for adaptive neighbour weighting, STGCN captures joint spatial and temporal dependencies via graph convolution and gated temporal convolution, GRAND adopts a continuous diffusion perspective, and EpiGNN incorporates epidemic-specific priors.

EpiRDN consistently achieves the lowest RMSE across all windows, with the gap widening substantially beyond window 14. GAT and STGCN degrade sharply at longer horizons, as their fixed aggregation schemes struggle to adapt to the evolving spatial structure of an outbreak. GRAND and EpiGNN show moderate but steady deterioration, as neither explicitly models local infection dynamics. EpiRDN's sustained advantage stems from jointly modelling spatial propagation and local reaction dynamics, allowing it to capture both inter-region transmission patterns and region-specific epidemic trajectories as the prediction window grows.

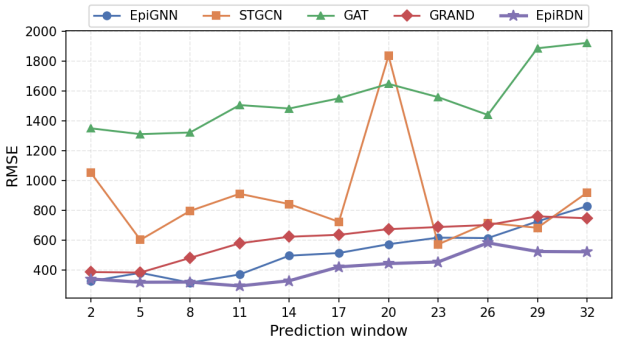

Figure 2: RMSE across increasing prediction windows on the Australia-COVID dataset.

6.5 FEATURE DIVERSITY RETENTION

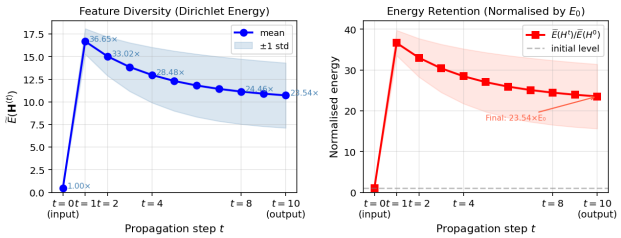

Figure 3: Dirichlet energy $\widetilde{E}(\mathbf{H}^{(t)})$ across propagation steps on the US-State dataset (horizon $= 2$, $L = 10$ layers). *Left*: absolute energy over layers. *Right*: energy normalised by E_0 over layers.

Theorem 2 guarantees that the Dirichlet energy $\widetilde{E}(\mathbf{H}^{(t)})$ is bounded below throughout propagation, preventing feature collapse. We verify this empirically by tracking $\widetilde{E}(\mathbf{H}^{(t)})$ across all propagation steps on the US-State dataset. Fig-

ure 3 reports the results.

At $t = 0$, the input projection produces relatively undifferentiated features. The reaction MLP at $t = 1$ actively amplifies inter-region differences as nodes with distinct epidemic trajectories are pushed apart in feature space. Subsequent layers exhibit gradual decay as the diffusion operator spreads contextual information across the graph, yet energy stabilises well at the final layer. This confirms that feature diversity is actively maintained throughout propagation, consistent with the theoretical guarantee of Theorem 2.

6.6 RUNTIME ANALYSIS

Figure 4 shows that EpiRDN strikes a practical balance between the lightweight standard GNNs and the expensive ODE-based models. By implementing reaction-diffusion dynamics via discrete message passing rather than continuous ODE integration, EpiRDN achieves moderate runtime while retaining its epidemiological inductive bias. GRAND and GREAD, which numerically solve a continuous diffusion equation at every epoch, incur substantially higher cost, making EpiRDN a more computationally efficient alternative without sacrificing modelling expressiveness.

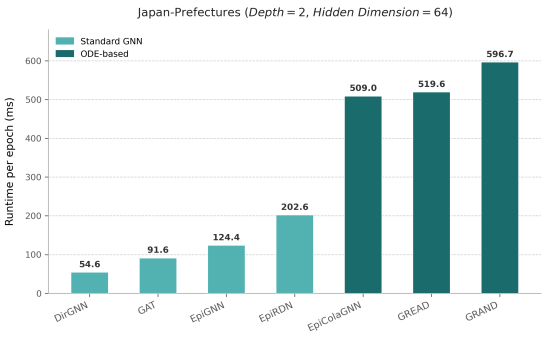

Figure 4: Runtime per epoch on Japan-Prefectures

Additional Experiments All three assumptions of Theorem 2 are empirically verified on the trained model in Appendix section 9.

7 CONCLUSION, AND LIMITATIONS

We propose EpiRDN, a novel graph neural network that discretizes the reaction-diffusion equation for epidemic time series prediction. By separating each propagation step into a local reaction component and an anisotropic diffusion operator with feature-conditioned damping, EpiRDN effectively captures directed infection flow and regional dynamics while preventing feature collapse. Experiments show consistent improvements over various baselines, especially at longer forecasting horizons.

Currently, EpiRDN assumes a fixed graph topology throughout the forecast horizon, which may not reflect the dynamic mobility patterns driving real epidemic spread. Additionally, the feature generation stage uses a simple linear projection with no explicit temporal inductive bias, potentially underutilizing domain-specific information in epidemic data. Extensions to our work include replacing the static graph with a temporally evolving mobility network learned jointly with epidemic dynamics and incorporating domain-specific priors to enhance interpretability and performance.

References

Dozdar Mahdi Ahmed, Masoud Muhammed Hassan, and Ramadhan J Mstafa. A review on deep sequential models for forecasting time series data. *Applied computational intelligence and soft computing*, 2022(1):6596397, 2022.

Ghazaleh Babanejaddehaki, Aijun An, and Manos Papagelis. Disease outbreak detection and forecasting: A review of methods and data sources. *ACM Transactions on Computing for Healthcare*, 6(2):1–40, 2025.

Apollinaire Batoure Bamana, Mahdi Shafiee Kamalabad, and Daniel L Oberski. A systematic literature review of time series methods applied to epidemic prediction. *Informatics in Medicine Unlocked*, 50:101571, 2024.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

Asha A Bharambe and Dhananjay R Kalbande. Techniques and approaches for disease outbreak prediction: A survey. In *Proceedings of the ACM Symposium on Women in Research 2016*, pages 100–102, 2016.

Fred Brauer. Compartmental models in epidemiology. In *Mathematical epidemiology*, pages 19–79. Springer, 2008.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *10th International Conference on Learning Representations, ICLR 2022*, 2022.

Ben Chamberlain, James Rowbottom, Maria I Gorinova, Michael Bronstein, Stefan Webb, and Emanuele Rossi. Grand: Graph neural diffusion. In *International conference on machine learning*, pages 1407–1418. PMLR, 2021a.

Benjamin Chamberlain, James Rowbottom, Davide Eynard, Francesco Di Giovanni, Xiaowen Dong, and Michael Bronstein. Beltrami flow and neural diffusion on graphs. *Advances in Neural Information Processing Systems*, 34: 1594–1609, 2021b.

Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 551–561, 2016.

Jeongwhan Choi, Seoyoung Hong, Noseong Park, and Sung-Bae Cho. Gread: Graph neural reaction-diffusion networks. In *International conference on machine learning*, pages 5722–5747. PMLR, 2023.

Flavio Corradini, Flavio Gerosa, Marco Gori, Carlo Lucheroni, Marco Piangerelli, and Martina Zannotti. A systematic literature review of spatio-temporal graph neural network models for time series forecasting and classification. *Neural Networks*, page 108269, 2025.

Songgaojun Deng, Shusen Wang, Huzefa Rangwala, Lijing Wang, and Yue Ning. Cola-gnn: Cross-location attention based graph neural networks for long-term ili prediction. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 245–254, 2020.

Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE, 2017.

Robert DiPietro and Gregory D Hager. Deep learning: Rnns and lstm. In *Handbook of medical image computing and computer assisted intervention*, pages 503–519. Elsevier, 2020.

Moshe Eliasof, Eldad Haber, and Eran Treister. Feature transportation improves graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 11874–11882, 2024.

Satoki Fujita and Tatsuya Akutsu. Enhancing epidemic forecasting with a physics-informed spatial identity neural network. *PLoS One*, 20(9):e0331611, 2025.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations*, 2018.

Johannes Gasteiger, Stefan Weißenberger, and Stephan Günnemann. Diffusion improves graph learning. *Advances in neural information processing systems*, 32, 2019.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *ICML*, pages 1263–1272. PMLR, 2017.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Truong Son Hy, Viet Bach Nguyen, Long Tran-Thanh, and Risi Kondor. Temporal multiresolution graph neural networks for epidemic prediction. In *Workshop on Healthcare AI and COVID-19*, pages 21–32. PMLR, 2022.

Xueting Jin, Fangwu Wei, Srinivasa Srivatsav Kandala, Tejas Umesh, Kayleigh Steele, John N Galgiani, and Manfred D Laubichler. Time series forecasting of valley fever infection in maricopa county, az using lstm. *The Lancet Regional Health–Americas*, 43, 2025.

Yufei Jin and Xingquan Zhu. Oversmoothing alleviation in graph neural networks: a survey and unified view: Y. jin, x. zhu. *Knowledge and Information Systems*, 67(12): 11259–11285, 2025.

Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 95–104, 2018.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018a.

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

Mutong Liu, Yang Liu, and Jiming Liu. Epidemiology-aware deep learning for infectious disease dynamics prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4084–4088, 2023.

Mutong Liu, Yang Liu, and Jiming Liu. Machine learning for infectious disease risk prediction: a survey. *ACM Computing Surveys*, 57(8):1–39, 2025.

Zewen Liu, Guancheng Wan, B Aditya Prakash, Max SY Lau, and Wei Jin. A review of graph neural networks in epidemic modeling. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6577–6587, 2024.

Renato R Maaliw, Zoren P Mabunga, and Frederick T Villa. Time-series forecasting of covid-19 cases using stacked long short-term memory networks. In *2021 International*

Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), pages 435–441. IEEE, 2021.

Larry R Medsker and Lakhmi Jain. Recurrent neural networks. *Design and applications*, 5(64-67):2, 2001.

JD Murray. Spatial pattern formation with reaction diffusion systems. In *Mathematical Biology: II: Spatial Models and Biomedical Applications*, pages 71–140. Springer, 2006.

Emanuele Rossi, Bertrand Charpentier, Francesco Di Giovanni, Fabrizio Frasca, Stephan Günnemann, and Michael M Bronstein. Edge directionality improves learning on heterophilic graphs. In *Learning on graphs conference*, pages 25–1. PMLR, 2024.

T Konstantin Rusch, Ben Chamberlain, James Rowbottom, Siddhartha Mishra, and Michael Bronstein. Graph-coupled oscillator networks. In *International conference on machine learning*, pages 18888–18909. PMLR, 2022.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

Joakim Sundnes. *Introduction to scientific programming with Python*. Springer, 2020.

Matthew Thorpe, Tan Minh Nguyen, Hedi Xia, Thomas Strohmer, Andrea Bertozzi, Stanley Osher, and Bao Wang. Grand++: Graph neural diffusion with a source term. In *International Conference on Learning Representations*, 2022.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Yuelin Wang, Kai Yi, Xinliang Liu, Yu Guang Wang, and Shi Jin. Acmp: Allen-cahn message passing with attractive and repulsive forces for graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023.

Yuexin Wu, Yiming Yang, Hiroshi Nishiura, and Masaya Saitoh. Deep learning for epidemiological predictions. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1085–1088, 2018.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

Feng Xie, Zhong Zhang, Liang Li, Bin Zhou, and Yusong Tan. Epignn: Exploring spatial transmission with graph neural network for regional epidemic forecasting. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 469–485. Springer, 2022.

Mingda Xu, Yang Liu, Sensen Guo, Yuxin Liu, and Chao Gao. Enhancing infectious disease forecasting via epidemiology-informed adaptive spatio-temporal graph neural networks. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1383–1390. IEEE, 2025.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 3634–3640, 2018.

Scott L Zeger, Rafael Irizarry, and Roger D Peng. On time series analysis of public health and biomedical data. *Annu. Rev. Public Health*, 27(1):57–79, 2006.

Mingda Zhang. Time series: Autoregressive models ar, ma, arma, arima. *University of Pittsburgh*, 2018.

Xingyu Zhang, Tao Zhang, Alistair A Young, and Xiaosong Li. Applications and comparisons of four time series models in epidemiological surveillance data. *Plos one*, 9 (2):e88075, 2014.

Yufan Zheng, Wei Jiang, Tong Chen, Alexander Zhou, Nguyen Quoc Viet Hung, Choujun Zhan, and Hongzhi Yin. Epidemiology-informed graph neural network for heterogeneity-aware epidemic forecasting. *arXiv preprint arXiv:2411.17372*, 2024.

Xiaofeng Zhu, Yi Zhang, Haoru Ying, Huanning Chi, Guanqun Sun, and Lingxia Zeng. Modeling epidemic dynamics using graph attention based spatial temporal networks. *Plos one*, 19(7):e0307159, 2024.

APPENDIX

8 PROOFS

Lemma 1 (Directional Asymmetry). *$\hat{A}^{(t)}_{ij} = \hat{A}^{(t)}_{ji}$ for all i, j and all $\mathbf{H}^{(t)}$ if and only if the attention kernel is symmetric. In particular, $W_l \neq W_r$ generically implies $\hat{A}^{(t)}_{ij} \neq \hat{A}^{(t)}_{ji}$.*

Proof. Partition $\mathbf{a} = [\mathbf{a}_l^\top \| \mathbf{a}_r^\top]^\top$ with $\mathbf{a}_l, \mathbf{a}_r \in \mathbb{R}^d$. The attention score decomposes as $s_{ij} = \mathbf{a}_l^\top W_l \mathbf{h}_i + \mathbf{a}_r^\top W_r \mathbf{h}_j$ and $s_{ji} = \mathbf{a}_l^\top W_l \mathbf{h}_j + \mathbf{a}_r^\top W_r \mathbf{h}_i$. These are equal for all $\mathbf{h}_i, \mathbf{h}_j$ if and only if $W_l^\top \mathbf{a}_l = W_r^\top \mathbf{a}_r$. If $W_l \neq W_r$, choose $\mathbf{h}_i = \mathbf{v}$ with $W_l \mathbf{v} \neq W_r \mathbf{v}$ and $\mathbf{h}_j = \mathbf{0}$; then $s_{ij} \neq s_{ji}$ for generic $\mathbf{a}$, yielding $\hat{A}^{(t)}_{ij} \neq \hat{A}^{(t)}_{ji}$. The Sinkhorn iteration applied afterwards rescales rows and columns by positive scalars; it does not equalise $\hat{A}^{(t)}_{ij}$ and $\hat{A}^{(t)}_{ji}$ unless the pre-Sinkhorn matrix is already symmetric, so asymmetry is preserved. □

Theorem 2 (Dirichlet Energy Retention). *Let $\pi_i = 1/n$ for all $i \in V$. Define the feature diversity measure $\widetilde{E}(\mathbf{H}) = \frac{1}{n}\sum_i \|\mathbf{h}_i - \bar{\mathbf{h}}\|^2$ where $\bar{\mathbf{h}} = \frac{1}{n}\sum_i \mathbf{h}_i$. Under Assumptions 5.1–5.3, and provided*

$$\underline{\gamma}(1 + L_{\min}) > 1, \tag{13}$$

the following lower bound holds for all $T \geq 0$:

$$\widetilde{E}\Big(\mathbf{H}^{(T)}\Big) \geq c_{\text{low}}^{2T}\, \widetilde{E}\Big(\mathbf{H}^{(0)}\Big), \qquad c_{\text{low}} = \underline{\gamma}(1 + L_{\min}) - 1 > 0. \tag{14}$$

In particular, $\widetilde{E}(\mathbf{H}^{(T)}) > 0$ for all finite T whenever $\widetilde{E}(\mathbf{H}^{(0)}) > 0$, preventing feature collapse at any finite depth.

Proof. Write $\|\cdot\|_F$ for the Frobenius norm and $\|\mathbf{X}\|^2_{\pi,F} = \frac{1}{n}\sum_i \|\mathbf{x}_i\|^2$ for the π-weighted Frobenius norm with $\pi_i = 1/n$. Let $\bar{\mathbf{h}}^{(t)} = \frac{1}{n}\sum_i \mathbf{h}_i^{(t)}$ and $\mathbf{H}_c^{(t)} = \mathbf{H}^{(t)} - \bar{\mathbf{h}}^{(t)}\mathbf{1}^\top$ denote the centred features, so that $\widetilde{E}(\mathbf{H}^{(t)}) = \|\mathbf{H}_c^{(t)}\|^2_{\pi,F}$.

Step 1: Mean preservation. By Assumption 5.3, $\hat{A}^{(t)}$ is doubly stochastic, so $\hat{A}^{(t)\top}\mathbf{1} = \mathbf{1}$. Therefore

$$\tfrac{1}{n}\mathbf{1}^\top \hat{A}^{(t)}\mathbf{H}^{(t)} = \tfrac{1}{n}(\hat{A}^{(t)\top}\mathbf{1})^\top \mathbf{H}^{(t)} = \tfrac{1}{n}\mathbf{1}^\top \mathbf{H}^{(t)} = \bar{\mathbf{h}}^{(t)\top}.$$

Thus the diffusion step preserves the mean: $\overline{(\hat{A}^{(t)}\mathbf{H}^{(t)})} = \bar{\mathbf{h}}^{(t)}$, and $(\hat{A}^{(t)}\mathbf{H}^{(t)})_c$ is well-defined with respect to the same mean. The reaction term $R(\mathbf{H}^{(t)})$ may shift the mean; let $R(\mathbf{H}^{(t)})_c$ denote the centring of $R(\mathbf{H}^{(t)})$ about $\overline{R(\mathbf{H}^{(t)})} = \frac{1}{n}\sum_i R(\mathbf{h}_i^{(t)})$. Since $\gamma_i^{(t)}$ may differ across nodes, we first re-express the centred update. The new mean is

$$\bar{\mathbf{h}}^{(t+1)} = \tfrac{1}{n}\sum_i \big[\gamma_i^{(t)} R(\mathbf{h}_i^{(t)}) + (1 - \gamma_i^{(t)})(\hat{A}^{(t)}\mathbf{H}^{(t)})_i\big],$$

and subtracting row-wise gives

$$\mathbf{H}_c^{(t+1)} = \mathbf{\Gamma}^{(t)} R(\mathbf{H}^{(t)})_c + (I - \mathbf{\Gamma}^{(t)})(\hat{A}^{(t)}\mathbf{H}^{(t)})_c + \Delta^{(t)}, \tag{16}$$

where $\Delta_i^{(t)} = \gamma_i^{(t)}\overline{R(\mathbf{H}^{(t)})} + (1 - \gamma_i^{(t)})\bar{\mathbf{h}}^{(t)} - \bar{\mathbf{h}}^{(t+1)}$ is a row-constant correction that vanishes when summed against $\mathbf{1}/n$ by construction. Since $\widetilde{E}(\mathbf{H}^{(t)}) = \|\mathbf{H}_c^{(t)}\|^2_{\pi,F}$ is the variance and variance is shift-invariant, we may without loss of generality work with any fixed reference centring. By shift-invariance of variance:

$$\widetilde{E}(\mathbf{H}^{(t+1)}) = \widetilde{E}\Big(\mathbf{\Gamma}^{(t)} R(\mathbf{H}^{(t)}) + (I - \mathbf{\Gamma}^{(t)})\hat{A}^{(t)}\mathbf{H}^{(t)}\Big).$$

We bound this from below by working with centred versions and applying the triangle inequality in the norm $\|\cdot\|_{\pi,F}$.

Step 2: Reaction lower bound. For any matrix $\mathbf{X}$ with rows $\mathbf{x}_i \in \mathbb{R}^f$, the π-weighted pairwise-variance identity

$$\sum_{i,j} \pi_i \pi_j \|\mathbf{x}_i - \mathbf{x}_j\|^2 = 2\|\mathbf{X}_c\|^2_{\pi,F} \tag{17}$$

follows by expanding $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ and using $\sum_i \pi_i = 1$. Applying the lower bi-Lipschitz condition (Assumption 5.2) pairwise,

$$\|R(\mathbf{h}_i) - R(\mathbf{h}_j)\|^2 \geq L_{\min}^2 \|\mathbf{h}_i - \mathbf{h}_j\|^2.$$

Multiplying by $\pi_i \pi_j$, summing over all (i, j), and applying (17) to both sides:

$$2\|R(\mathbf{H})_c\|^2_{\pi,F} \geq L_{\min}^2 \cdot 2\|\mathbf{H}_c\|^2_{\pi,F},$$

giving

$$\|R(\mathbf{H}^{(t)})_c\|_{\pi,F} \geq L_{\min}\|\mathbf{H}_c^{(t)}\|_{\pi,F}. \tag{18}$$

Step 3: Diffusion contraction. Since $\hat{A}^{(t)}$ is doubly stochastic (Assumption 5.3), the uniform distribution $\pi_i = 1/n$ satisfies $\pi^\top \hat{A}^{(t)} = \pi^\top$ (left stationarity follows from column sums equalling one). For any vector $\mathbf{y} \in \mathbb{R}^n$ and row-stochastic matrix P with stationary measure π, the variance inequality $\mathrm{Var}_\pi(P\mathbf{y}) \leq \mathrm{Var}_\pi(\mathbf{y})$ follows from Jensen's inequality applied to the convex function $x \mapsto x^2$. Applied coordinate-wise to the rows of $\mathbf{H}^{(t)}$:

$$\|(\hat{A}^{(t)}\mathbf{H}^{(t)})_c\|_{\pi,F} \leq \|\mathbf{H}_c^{(t)}\|_{\pi,F}. \tag{19}$$

Step 4: Inductive bound. Let $S^{(t)} = \|\mathbf{H}_c^{(t)}\|_{\pi,F}$. By shift-invariance of variance, $\Delta^{(t)}$ does not contribute to $\widetilde{E}(\mathbf{H}^{(t+1)})$, so

$$S^{(t+1)} = \big\|\mathbf{\Gamma}^{(t)} R(\mathbf{H}^{(t)})_c + (I - \mathbf{\Gamma}^{(t)})(\hat{A}^{(t)}\mathbf{H}^{(t)})_c\big\|_{\pi,F}.$$

Applying the reverse triangle inequality $\|\mathbf{u} + \mathbf{v}\|_{\pi,F} \geq \|\mathbf{u}\|_{\pi,F} - \|\mathbf{v}\|_{\pi,F}$ with $\mathbf{u} = \mathbf{\Gamma}^{(t)} R(\mathbf{H}^{(t)})_c$ and $\mathbf{v} = (I - \mathbf{\Gamma}^{(t)})(\hat{A}^{(t)}\mathbf{H}^{(t)})_c$:

$$S^{(t+1)} \geq \big\|\mathbf{\Gamma}^{(t)} R(\mathbf{H}^{(t)})_c\big\|_{\pi,F} - \big\|(I - \mathbf{\Gamma}^{(t)})(\hat{A}^{(t)}\mathbf{H}^{(t)})_c\big\|_{\pi,F}.$$

Since $\mathbf{\Gamma}^{(t)} = \mathrm{diag}(\gamma_i^{(t)})$ with $\gamma_i^{(t)} \geq \underline{\gamma}$ for all i (Assumption 5.1), and $(I - \mathbf{\Gamma}^{(t)}) = \mathrm{diag}(1 - \gamma_i^{(t)})$ with $1 - \gamma_i^{(t)} \leq 1 - \underline{\gamma}$, scaling each row of the centred matrices by these diagonal factors gives

$$\left\|\mathbf{\Gamma}^{(t)} R(\mathbf{H}^{(t)})_c\right\|_{\pi,F} \geq \underline{\gamma}\, \|R(\mathbf{H}^{(t)})_c\|_{\pi,F},$$

$$\left\|(I - \mathbf{\Gamma}^{(t)})(\hat{A}^{(t)}\mathbf{H}^{(t)})_c\right\|_{\pi,F} \leq (1 - \underline{\gamma})\, \|(\hat{A}^{(t)}\mathbf{H}^{(t)})_c\|_{\pi,F}.$$

Substituting these together with (18) and (19):

$$\begin{aligned} S^{(t+1)} &\geq \underline{\gamma}\, \|R(\mathbf{H}^{(t)})_c\|_{\pi,F} - (1 - \underline{\gamma})\, \|(\hat{A}^{(t)}\mathbf{H}^{(t)})_c\|_{\pi,F} \\ &\geq \underline{\gamma}\, L_{\min}\, S^{(t)} - (1 - \underline{\gamma})\, S^{(t)} \\ &= \big(\underline{\gamma}(1 + L_{\min}) - 1\big) S^{(t)} = c_{\mathrm{low}}\, S^{(t)}. \end{aligned}$$

Since $c_{\mathrm{low}} > 0$ by condition (13), applying this inductively over T steps and squaring:

$$\widetilde{E}(\mathbf{H}^{(T)}) = \big(S^{(T)}\big)^2 \geq c_{\mathrm{low}}^{2T} \big(S^{(0)}\big)^2 = c_{\mathrm{low}}^{2T}\, \widetilde{E}(\mathbf{H}^{(0)}). \quad \square$$

9 EMPIRICAL VALIDATION OF THEORETICAL ASSUMPTIONS

Theorem 2 establishes that feature diversity decays at most geometrically with propagation depth, provided three structural conditions hold. We verify each in turn on the trained model for Japan-Prefectures dataset.

Assumption 5.1: Damping Floor. The per-node coefficient $\gamma_i^{(t)}$ must remain above the learned floor $\bar{\gamma}$ at every node, layer, and sample. Figure 5 shows zero violations across all propagation steps, confirming that the diversity-preserving floor is active throughout, as this is enforced by out regularization term.

Assumption 5.2: Bi-Lipschitz Reaction. For the reaction MLP to counteract diffusion-induced collapse, it must preserve input separation — distinct epidemic states must map to distinct representations. This requires all singular values of each weight matrix to remain above $\sqrt{L_{\min}}$. Figure 6 confirms this holds for both layers, ensuring the reaction amplifies rather than discards inter-region differences.

Assumption 5.3: Doubly Stochastic Diffusion. Double stochasticity of $\hat{A}^{(t)}$ ensures that the diffusion step preserves the population mean of node features, which is essential for the contraction bound in the proof of Theorem 2. Figure 7 shows that a single Sinkhorn iteration reduces the maximum column-sum deviation to below 10^{-7}, with no measurable gain from further iterations.

Feature Diversity Retention. With all assumptions satisfied, the condition $c_{\mathrm{low}} = \bar{\gamma}(1 + L_{\min}) - 1 > 0$ determines whether Theorem 2 provides a non-trivial lower bound. Figure 8 places the trained model well within the region $c_{\mathrm{low}} > 0$, with substantial margin from the boundary, confirming that the reaction MLP actively counteracts diffusion-induced collapse throughout the rollout.

.

Collectively, these results confirm that the theoretical conditions of Theorem 2 are not merely assumed but hold for the trained model, with each assumption satisfied by construction and the energy retention condition met with meaningful margin.

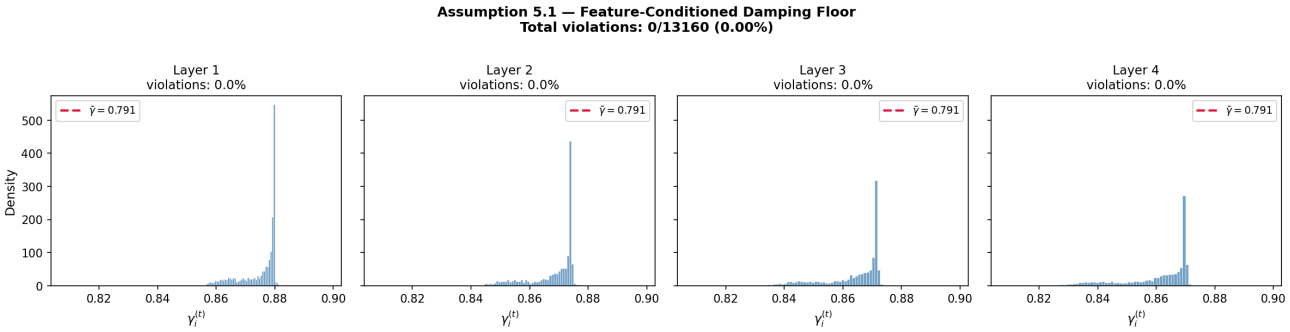

Figure 5: Assumption 5.1: distribution of $\gamma_i^{(t)}$ per layer. The learned floor $\bar{\gamma}$ (dashed) is never violated across any node or step.

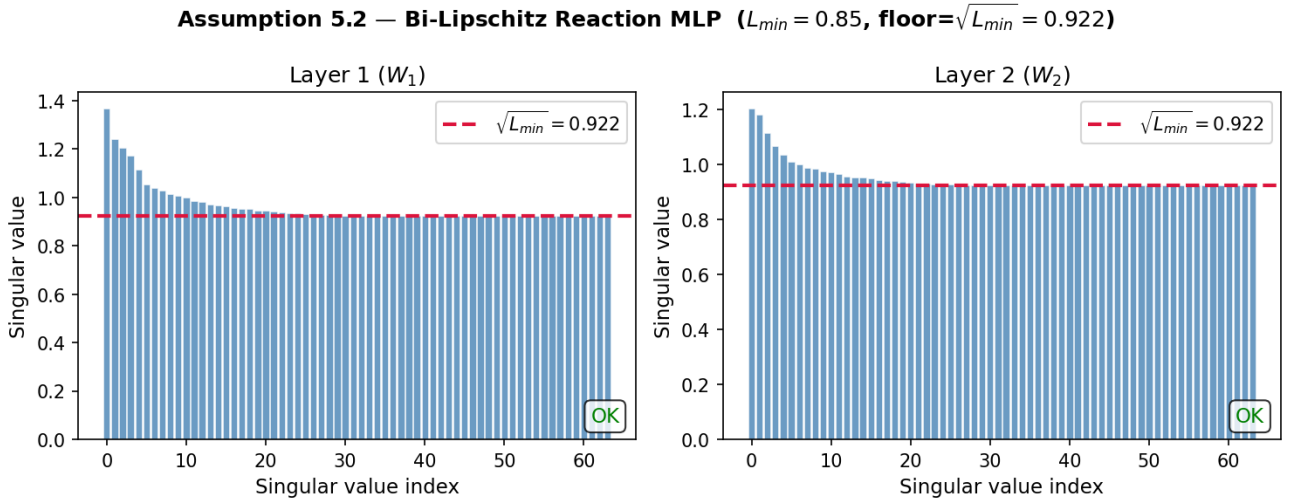

Figure 6: Assumption 5.2: singular value spectra of both reaction MLP weight matrices, sorted in descending order. The rightmost bar is $\sigma_{\min}$. All values remain above the required floor $\sqrt{L_{\min}}$ (dashed).

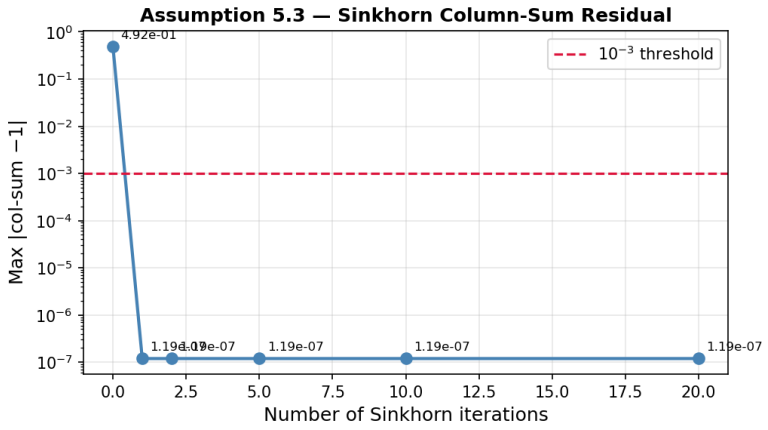

Figure 7: Assumption 5.3: Sinkhorn column-sum residual by iteration count. One iteration achieves near-exact double stochasticity.

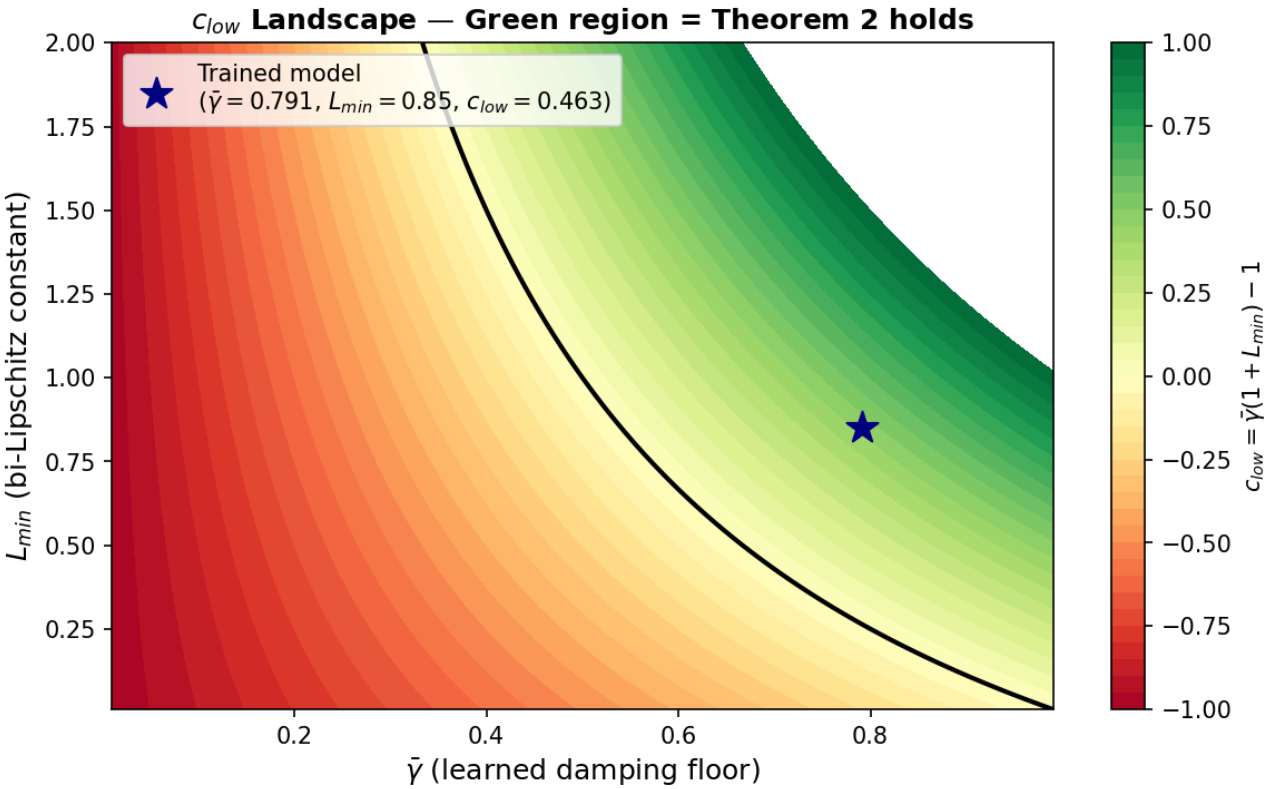

Figure 8: c_{low} landscape over $(\bar{\gamma}, L_{\min})$. The trained model (star) sits firmly in the green region, confirming the Theorem 2 condition.