APPENDIX

Proofs

Proof of Lemma 2

Lemma 2. If (1) h_2 is trained with data sampled from $\mathcal{X}_s$ such that assumption 2 is true, (2) the loss function L is the L1-norm or MSE, then $h_2^* = F$.

Proof. We prove this by contradiction. If $h_2^* \neq F$, there must exist $\hat{h}_2^*(X_t) \neq 0$ such that $h_2^*(X_t) = F(O(X(t), \mathcal{A}); \Theta_1) + \hat{h}_2^*(X_t)$ and $\hat{h}_2^*$ minimizes the following expectation:

$$\hat{h}_2^* = \min_{\hat{h}_2 : h_2 - F} \mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[L(\hat{h}_2(X(t)), G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))]. \tag{19}$$

If the loss function is the L1-norm, Problem (19) is minimized when $\hat{h}_2^*(X_t)$ equals the median of $G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2)$. Assumption 2 in Section 5 implies

$$Median_{X_{t-T:t} \sim \mathcal{D}}[G] = \mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[U] = 0. \tag{20}$$

Thus $\hat{h}_2^*(X_t) = 0$ is the optimal solution of Problem (19), which contradicts the fact that $\hat{h}_2^*(X_t) \neq 0$.
If the loss function is the MSE, there must exist $\hat{h}_2(X_t) \neq 0$ such that $h_2^*(X_t) = F(O(X(t), \mathcal{A}); \Theta_1)$ and $\hat{h}_2(X_t)$ minimize the following expectation

$$\mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[(\hat{h}_2(X_t) - G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))^2]. \tag{21}$$

Since Assumption 2 in Section 5 implies

$$\mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2)] = 0, \tag{22}$$

the expectation $\mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[(\hat{h}_2(X_t) - G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))^2]$ is minimized when the derivative $2(\hat{h}_2(X(t)) - \mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2)]) = 0$, hence $\hat{h}_2(X_t) = 0$ must minimize the expectation in Eq. (21), which contradicts the fact that $\hat{h}_2(X_t) \neq 0$. Therefore $h_2^* = F$. □

Proof of Theorem 3

Theorem 3. If (1) the training data is sampled from the source domain where assumption 2 is true, (2) the loss function $L(h, l)$ obeys the triangular equality, then the discrepancy with any triangular equality loss should satisfy

$$disc(\mathcal{H}_2^*) \leq disc(\mathcal{H}_1^*). \tag{23}$$

Proof. By the definition of discrepancy in Eq.(2), we know

$$\begin{aligned}
disc(\mathcal{H}_1^*) &= \sup_{h_1 \in \mathcal{H}_1^*} |\mathcal{L}_{(\mathcal{D}, F+G_s)}(h_1) - \mathcal{L}_{(\mathcal{D}, F+G_\tau)}(h_1)| \\
&= \sup_{h_1 \in \mathcal{H}_1^*} |\mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[L(h_1(X_{t-T:t}), F(X(t)) \\
&\qquad + G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2)) \\
&\qquad - L(h_1(X_{t-T:t}), F(X(t)) \\
&\qquad + G_\tau(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))]| \\
&\overset{(a)}{\leq} \sup_{h_1 \in \mathcal{H}_1^*} \mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[|L(h_1(X_{t-T:t}), F(X(t)) \\
&\quad + G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2)) \\
&\qquad - L(h_1(X_{t-T:t}), F(X(t)) \\
&\qquad + G_\tau(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))|] \\
&\overset{(b)}{\leq} \mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[L(G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2), \\
&\qquad G_\tau(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))],
\end{aligned} \tag{24}$$

where (a) follows from Jensen's equality ($|\cdot|$ is convex) and (b) follows from the triangle inequality (which implies $|L(x, y)| \geq |L(x, z) - L(y, z)|$, for any $x, y, z \in \mathbb{R}$). By Assumption 1 in Section 5, we can set $h_1^* = F + G_s$ where $\mathcal{L}_{(\mathcal{D}, F+G_s)}(h_1^*) = 0$. Then the discrepancy of $\mathcal{H}_1$ is

$$\begin{aligned}
disc(\mathcal{H}_1^*) &\overset{(c)}{\geq} \mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[L(F(X(t)) \\
&\qquad + G_\tau(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2), F(X(t)) \\
&\qquad + G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))] \\
&= \mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[L(G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2), \\
&\qquad G_\tau(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))],
\end{aligned} \tag{25}$$

where (c) follows from the definition that the supremum (the least element that is greater than or equal to each element in the set). Thus from (24) and (25) together

$$\begin{aligned}
disc(\mathcal{H}_1^*) = \mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[&L(G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2), \\
&G_\tau(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))].
\end{aligned} \tag{26}$$

For $\mathcal{H}_2$, by the triangle inequality,

$$\begin{aligned}
disc(\mathcal{H}_2^*) &= \sup_{h_2 \in \mathcal{H}_2^*} |\mathcal{L}_{(\mathcal{D}, F+G_s)}(h_2) - \mathcal{L}_{(\mathcal{D}, F+G_\tau)}(h_2)| \\
&\leq \mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[L(G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2), \\
&\qquad G_\tau(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))].
\end{aligned} \tag{27}$$

Hence we have shown that $disc(\mathcal{H}_2^*) \leq disc(\mathcal{H}_1^*)$. □

Discrepancy using MSE

Assumption 3. Let $U' = G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2) G_\tau(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2)$ be a random variable where $X_{t-T:t} \sim \mathcal{D}$, $\mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[U'] \leq 0$.

Corollary 1. If (1) the training data is sampled from the source domain where assumption 2 is true, (2) the labeling function in the source and target domain satisfy Assumption 3, (3) the loss function $L(h, l)$ is MSE, (4) $\mathcal{L}_{(\mathcal{D}, F)}(h_2^*) = 0$, then then $disc(\mathcal{H}_2^*) \leq disc(\mathcal{H}_1^*)$.

Proof. By Assumption 1, we can set $h_1^* = F + G_s$ where

$\mathcal{L}_{(\mathcal{D},F+G_s)}(h_1^*) = 0$, then the discrepancy of $\mathcal{H}_1$ is

$$
\begin{aligned}
disc(\mathcal{H}_1^*) = \sup_{h_1 \in \mathcal{H}_1^*} & |\mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[L(h_1(X_{t-T:t}), F(X(t)) \\
& + G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2)) \\
& - L(h_1(X_{t-T:t}), F(X(t)) \\
& + G_\tau(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))]| \\
\overset{(d)}{\geq} & \mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[L(F(X(t)) + G_\tau(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2), \\
& \qquad F(X(t)) + G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))] \\
= & \mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[(G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2) \\
& - G_\tau(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))^2],
\end{aligned}
\tag{28}
$$

where (d) follows from the definition of the supremum (the least element that is greater than or equal to each element in the set). We note that the triangular equality is not necessarily true in this case thus we cannot find the upper bound of $disc(\mathcal{H}_1^*)$.

Since $\mathcal{L}_{(\mathcal{D},F)}(h_2^*) = \mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[((h_2^*(X(t)) - F(O(X(t), \mathcal{A}); \Theta_1))^2] = 0$ implies that $h_2^* = F(O(X(t), \mathcal{A}); \Theta_1)$.

Then the discrepancy of $\mathcal{H}_2^*$ is:

$$
\begin{aligned}
disc(\mathcal{H}_2^*) = \sup_{h_2^* \in \mathcal{H}_2^*} & |\mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[(h_2^*(X(t)) - F(O(X(t), \mathcal{A}); \Theta_1) \\
& - G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))^2 \\
& - (h_2^*(X(t)) - F(O(X(t), \mathcal{A}); \Theta_1) \\
& - G_\tau(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))^2]| \\
\overset{(e)}{\leq} & \sup_{h_2^* \in \mathcal{H}_2^*} \mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[|(h_2^*(X(t)) - F(O(X(t), \mathcal{A}); \Theta_1) \\
& - G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))^2 - \\
& (h_2^*(X(t)) - F(O(X(t), \mathcal{A}); \Theta_1) \\
& - G_\tau(O_2(H_{t,T}, \mathcal{G}); \Theta_2))^2|] \\
= & \mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[|(G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))^2 \\
& - (G_\tau(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))^2|] \\
\leq & \mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[(G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))^2 \\
& + (G_\tau(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))^2] \\
\overset{(f)}{\leq} & \mathbb{E}_{X_{t-T:t} \sim \mathcal{D}}[|(G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2) \\
& - G_\tau(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2))^2|],
\end{aligned}
\tag{29}
$$

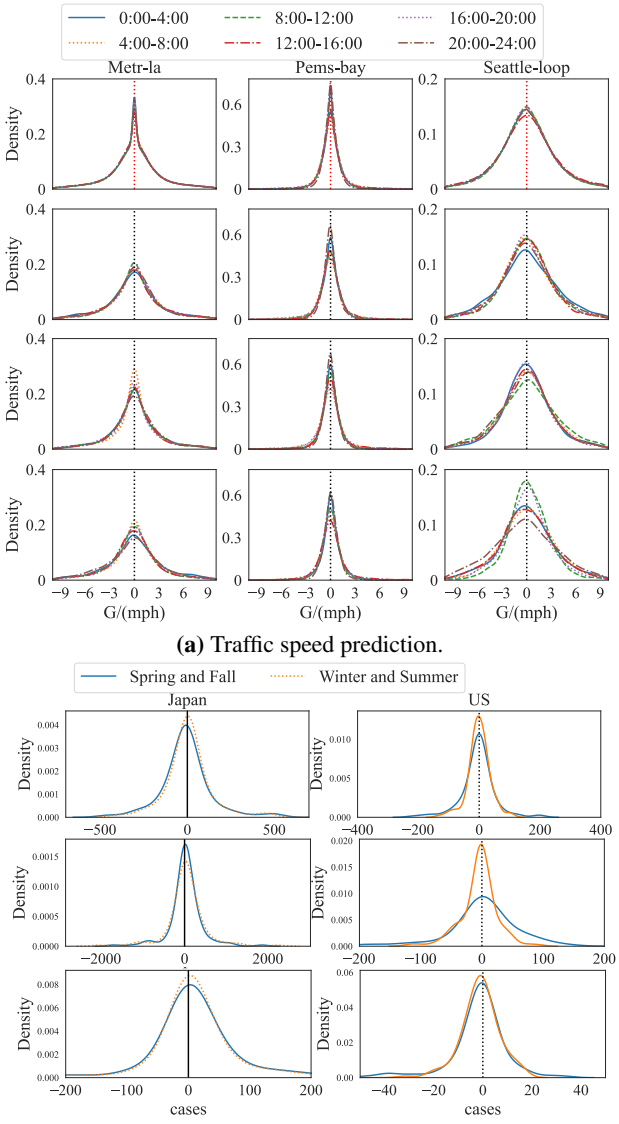
where (e) follows from Jensen's inequality, (f) follows from Assumption 3. Hence by Eq. (28)(29), we know $disc(\mathcal{H}_2^*) \leq disc(\mathcal{H}_1^*)$

□

Data Support

We empirically verify the condition in Assumption 2, in the scenario that the Reaction-Diffusion traffic model is the underlying physical law, and consequently, RDGCN, when trained well, perfectly models function f. Then,

$$
\begin{aligned}
& G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2) \\
= & X_{t+1} - F(O(X(t), \mathcal{A}); \Theta_1)) \\
\approx & X_{t+1} - RDGCN(X(t)).
\end{aligned}
\tag{30}
$$



(a) Traffic speed prediction.

(b) ILI prediction.

Figure 3: (a) The pdf of the random variable, G is symmetric about 0 for all the time periods. Figures in the first row are the mixed distribution of all sensors. Figures in the following three rows are the distribution of three randomly selected sensors in each dataset. (b) The pdf of the random variable, G is symmetric about 0 for all seasons. We randomly select 3 vertices in each data set.

The probability density function (pdf) of the random variable $G = G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2)$ in all the six periods are shown in Figure 3a, which plots the empirical density of the variable G in each dataset. As can be observed, the empirical density adheres to the condition in Assumption 2. For SIRGCN, we approximate G_s by

$$
\begin{aligned}
& G_s(O(X(t), \mathbb{1} - \mathcal{A}), X_{t-T:t-1}; \Theta_2) \\
= & X_{t+1} - F(O(X(t), \mathcal{A}); \Theta_1)) \\
\approx & X_{t+1} - SIRGCN(X(t)).
\end{aligned}
\tag{31}
$$

The corresponding pdf plot of the random variable G in ILI

prediction is in Figure 3b.

Most Important Sensors under Mismatches

In this section, we provide a motivation for going beyond domain-agnostic deep learning models by illustrating a possible weakness of such a model under mismatched data. Specifically, we apply a post-hoc explanation tool GNNExplainer (Ying et al. 2019) to identify the most influential sensors contributing to a model's prediction at the target sensor. We choose the Spatio-Temporal GCN (STGCN) model which has a good performance in graph time series prediction, particularly in traffic speed prediction. The STGCN model is trained by four-hour data in a sequence of 12 consecutive weekdays, while the GNNExplainer is used to identify the 3 most influential sensors on the weekend data. We show the location of the 3 most influential sensors under matched data (train by weekday data and test by weekday data), and mismatched data (trained by weekday data and test by weekend data) in Figure 4.

Figure 4 shows that when the test distribution is mismatched with the training distribution, the most influential sensors identified by GNNExplainer are too far to drive within the prediction window, and the distances change significantly. In other words, speed measurements from vertices that are too far to influence the target vertex, and suggests a violation of domain traffic law. This forms the motivation for our approach.

Reaction-diffusion Equation

As seen in Eq. (12), the change in speed is a function of two terms. The diffusion term is a monotone linear function of speed change in the direction of traffic, and it relies on the empirical fact that in the event of congestion, drivers prefer to bypass the congestion by following one of the neighboring links (Figure 5a). The reaction term is a non-linear monotone function (tanh activation) of speed change opposite to the direction of traffic, and it relies on the empirical fact that a road surrounded by congested roads is highly likely to be congested as well (Figure 5b). The architecture of RDGCN is shown in Figure 6.

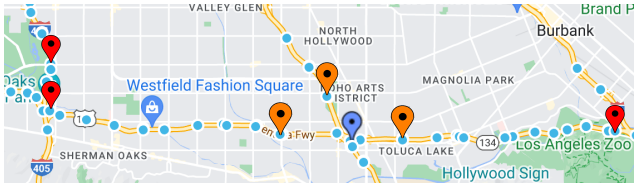
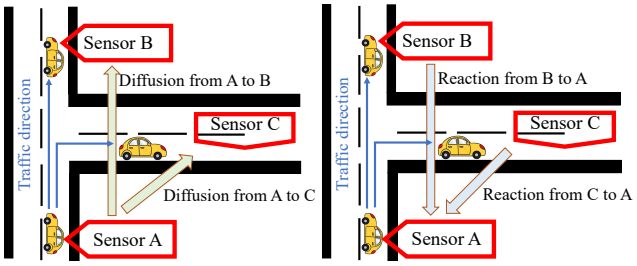


Figure 4: When an STGCN is tested on dataset from a matching distribution, the most important sensors (orange markers) are near the target sesnor, whereas the most important sensors under mismatched data (red markers) for the traffic speed prediction at target sensor (blue marker) are located far away. However, under matched data, the most important sensors are often close to the target sensor.



(a) Diffusion. **(b)** Reaction.

Figure 5: (a) Diffusion occurs in the direction of a road segment; (b) reaction occurs opposite to the direction of a road segment.

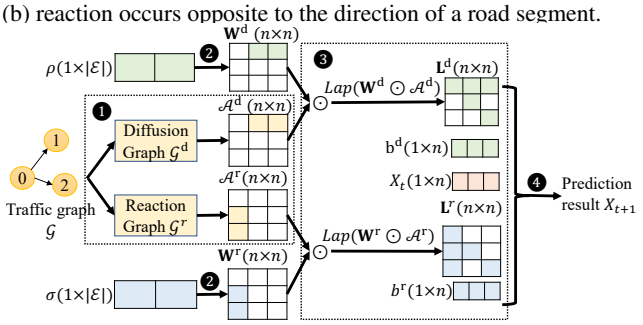


Figure 6: Reaction-diffusion GCN architecture for graph with $|\mathcal{V}| = 3$ and $|\mathcal{E}| = 2$. ❶ derives the diffusion and reaction adjacency matrices A^d and A^r; ❷ defines model weights ρ and σ for the reaction and diffusion networks, and map them to $\mathbf{W}^d$ and $\mathbf{W}^r$ with weights ρ and σ; ❸ characterizes the Graph Laplacian $\mathbf{L}^d$ and $\mathbf{L}^r$; ❹ defines the network prediction function Eq. (14).

Ablation Study

Analysis of RDGCN in Traffic Speed Prediction

Are reaction and diffusion processes essential? In this section, we investigate the prediction models incorporating the reaction equation and the diffusion equation, independently, under limited and mismatched data, to understand whether both the reaction and diffusion processes are essential. We use the same training set (i.e., 12 consecutive working days selected randomly) and test set (i.e., hourly weekend data) in Section . The curves of MAE versus time using the model incorporating the reaction equation, the diffusion equation, and the reaction-diffusion equation are shown in Figure 7.

Figure 7 indicates that the predictions of all models with the reaction-diffusion equation provide low MAE with low variance (i.e., the difference between curves with the highest MAE and lowest MAE is small) over time. However, the predictions of the reaction models only and the diffusion models only have weaker performance in at least one time period. We speculate that using only the reaction equation or the diffusion equation is not sufficient to capture the dynamics of the traffic speed change completely. Furthermore, the prediction of the model incorporating the reaction-diffusion equation is not uniformly better than the prediction of the model incorporating only the reaction or diffusion equation. One possible reason is that the reaction or diffusion process does not always exist in a specific period (e.g., if two neigh-

boring road segments are in free flow during the test period, the traffic speeds at the two segments do not affect each other. Thus there is neither diffusion nor reaction between these two road segments). These observations further strengthen that both the reaction and diffusion processes are necessary for a reliable prediction.

What is the performance under RMSE? We plot the corresponding result under RMSE loss in Figure 8, and the conclusion is consistent with the result using MAE. The RMSE of RDGCN are with low variance regardless of the peroid of the training set. We acknowledge that RDGCN is not always better than baselines under RMSE, for example, when STGCN is trained with weekday data from 16:00-20:00 in Metr-la. One possible reason is that the mismatches between the training data and test data are not significant during the corresponding time period. The prediction results of RDGCN in terms of RMSE may not always be stable. For instance, when considering the models for the 4:00 to 8:00 time period in Metr-la, we observe distinct prediction outcomes. This variation could be due to the difference between the pattern of the morning rush hour during selected weekdays and the pattern during weekends. When the training set includes all available weekday data, the predictions of RDGCN demonstrate stability.

Experimental results. The Mean and STD of prediction MAE (resp. RMSE) of each model with MAML augmentation and with full weekday training set are shown in Table 3 (i.e., the Mean and STD of all points on each subfigure in Figure 2a, Figure 2b, Figure 8b and Figure 8d), respectively. Table 3 shows that RDGCN has lower MAE (resp. RMSE) and lower variance compared with baselines under limited training set with MAML augmentation, and the gain of adding more data on RDGCN is limited, which are consistent with our observation in Figure 2 in Section .

Impact of data volume. We further investigate the influence of training data volume on the performance of baseline models and RDGCN under a mismatched setting. We focus on assessing the adequacy of training data for both morning rush hour (8:00-12:00) and evening rush hour (16:00-20:00) scenarios using the Metr-la dataset. These periods exhibit considerable patterns and exhibit relatively minor mismatches between training and test datasets. To this end, we randomly select contiguous weekdays, ranging from 20% to the entire dataset, for training the models. The MAE of speed prediction across varying quantities of training data is shown in Figure 9.

Figure 9 showcases the performance characteristics of RDGCN and baseline models over the specified time intervals. Remarkably, the performance of RDGCN remains consistent irrespective of the training dataset size. Conversely, the predictive capabilities of STGNCDE and MTGODE are notably contingent upon the amount of training data employed. The observed trend underscores increased training data volume directly correlates with enhanced prediction accuracy. In the morning rush hour, MTGODE achieves optimal performance with approximately 75% of training data (equivalent to 60 weekdays), while STGNCDE demonstrates comparable performance when trained on the entire weekday dataset. We note that the superiority of RDGCN over baseline models is not universally consistent, as elucidated earlier. Notably, integrating domain differential equations drastically reduces the hypothesis class's size, thereby filtering out erroneous hypotheses often prevalent in conventional black-box graph learning models. Consequently, domain-differential-equation-informed GCNs exhibit remarkable robustness on relatively smaller training datasets.

Analysis of SIRGCN in ILI Prediction

Do the infection rates vary among different vertices? In this section, we delve into the question of whether we require an individual infection rate for each vertex in ILI prediction. We specifically examine two approaches: one where we assign a unique infection rate, denoted as β_i, to each vertex i, resulting in a SIRGCN with n infection rates (SIRGCN-n), and another approach where we assign a single infection rate, denoted as β, to all vertices (SIRGCN-1). We report the MAE and RMSE of the prediction under mismatched data (train using Winter-Summer data and test using Spring-Fall data) in Table 4.

Table 4 shows that employing multiple infection rates leads to more accurate predictions, particularly in the case of the US-state dataset. By assigning individual infection rates to each vertex, we achieve a reduction of 2.4% in MAE (and 1.6% in RMSE). However, the advantage of utilizing multiple infection rates is less pronounced ($< 1\%$) in the ILI prediction of Japan. There could be two potential explanations for this phenomenon. First, the size of Japan's prefectures is not as substantial as that of the states in the United States. Second, the climate across Japan is relatively homogeneous, whereas the climate across different states of the United States exhibits significant variations, such as wet coastal areas and dry inland areas.

Predictions in Different Seasons. Learning patterns across different trends becomes challenging when baseline models are not trained using the same trend. For example, during Winter, the infectious number shows an increasing trend, whereas during Spring, it exhibits a decreasing trend. Figure 10 shows the predicted number of infectious cases alongside the ground truth data, revealing that SIRGCN's prediction aligns better with the ground truth. Conversely, EpiGNN's prediction performs poorly during the decline phase and when the number of infections approaches 0.

In the case of US-State ILI prediction in May 2014, both COLAGNN and EPIGNN fail to make accurate predictions around the peak, while SIRGCN demonstrates its effectiveness during the corresponding period, with the help of SIR-network model.

Model Efficiency in Computation Time

The training time and inference time (on two NVIDIA-2080ti graphic cards) of STGCN, MTGNN, GTS, and RDGCN on the Metr-la dataset are demonstrated in Table 5. It's observed that RDGCN takes less time in both training and inference than the other models, since the RDGCN contains significantly less number of parameters than the baseline models.

The training and inference time of ColaGNN, EpiGNN, and SIRGCN are shown in Table 5. SIRGCN has significantly less number of parameters than the baseline models.

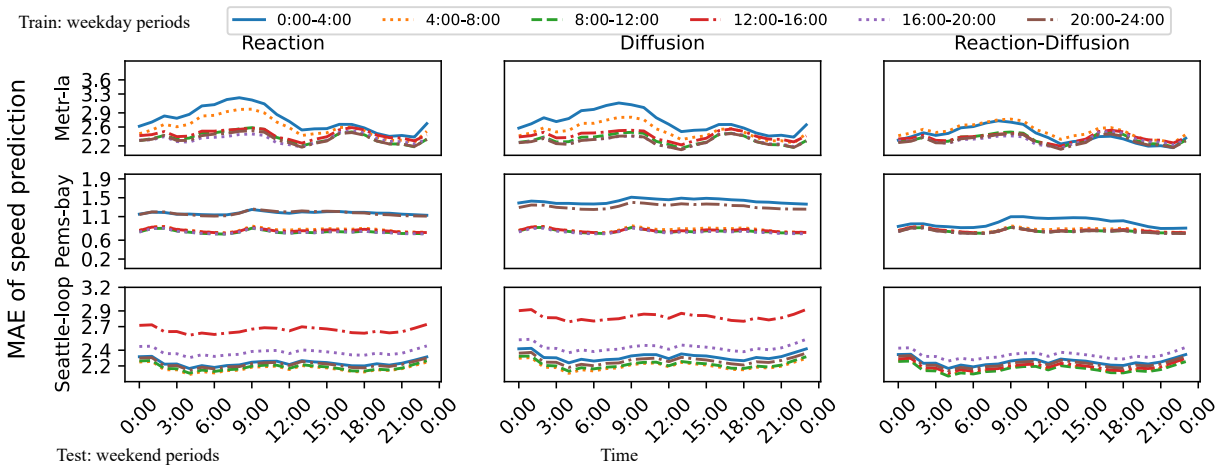


Figure 7: MAE of speed predictions on models incorporating reaction equation, diffusion equation, and reaction-diffusion equation.

Table 3: Numerical result of Figure 2: the Mean and STD of prediction MAE and RMSE of RDGCN and baselines on three real-world datasets.

	MAE						RMSE					
With MAML	STGCN	MTGNN	GTS	STGNCDE	MTGODE	RDGCN	STGCN	MTGNN	GTS	STGNCDE	MTGODE	RDGCN
Metr-la	2.47 ± 0.11	2.41 ± 0.22	2.55 ± 0.48	3.27 ± 0.47	2.82 ± 0.49	2.39 ± 0.08	5.28 ± 0.94	5.17 ± 1.16	7.55 ± 0.91	7.01 ± 1.28	5.41 ± 2.01	4.96 ± 0.83
Pems-bay	1.03 ± 0.19	0.91 ± 0.21	0.96 ± 0.03	0.77 ± 0.06	0.86 ± 0.14	0.83 ± 0.03	1.41 ± 0.05	2.86 ± 1.11	2.85 ± 0.84	1.44 ± 0.16	1.58 ± 0.44	1.40 ± 0.05
Seattle-loop	2.20 ± 0.08	2.23 ± 0.24	2.34 ± 0.15	3.20 ± 0.07	3.17 ± 0.05	2.16 ± 0.05	5.94 ± 0.14	3.92 ± 0.37	5.80 ± 0.60	6.16 ± 0.17	6.04 ± 0.19	3.44 ± 0.18
FULL	STGCN	MTGNN	GTS	STGNCDE	MTGODE	RDGCN	STGCN	MTGNN	GTS	STGNCDE	MTGODE	RDGCN
Metr-la	2.57 ± 0.68	3.11 ± 0.48	3.44 ± 0.47	2.77 ± 0.35	2.31 ± 0.43	2.38 ± 0.13	5.31 ± 0.92	4.02 ± 0.31	7.04 ± 1.20	6.43 ± 1.24	4.70 ± 1.38	3.90 ± 0.10
Pems-bay	1.38 ± 0.06	1.85 ± 0.38	2.08 ± 0.51	0.83 ± 0.09	0.79 ± 0.02	0.74 ± 0.02	1.37 ± 0.06	1.85 ± 0.38	2.08 ± 0.53	1.38 ± 0.04	1.36 ± 0.04	1.38 ± 0.04
Seattle-loop	2.90 ± 0.10	2.81 ± 0.65	3.11 ± 0.11	3.32 ± 0.07	3.21 ± 0.05	2.18 ± 0.06	3.91 ± 0.45	3.81 ± 0.65	5.33 ± 0.74	6.25 ± 0.17	6.22 ± 0.17	3.58 ± 0.05

Table 4: Evaluation models under mismatched data.

	MAE		RMSE	
	SIRGCN-1	SIRGCN-n	SIRGCN-1	SIRGCN-n
Japan-Prefectures	344 ± 22	342 ± 22	871 ± 43	863 ± 44
US-States	42 ± 4	41 ± 4	123 ± 10	121 ± 10

Table 5: The computation time on the Metr-la dataset.

		# Parameters	Training (s/epoch)	Inference (s)
Metr-la	STGCN	458865	0.5649	0.0232
	MTGNN	405452	0.5621	0.0607
	GTS	38377299	1.0632	0.1641
	STGNCDE	374904	1.7114	0.3729
	MTGODE	138636	1.6158	0.3491
	RDGCN	872	0.0308	0.0037
Japan-prefectures	ColaGNN	4272	0.0297	0.0065
	EpiGNN	16875	0.0311	0.0073
	SIRGCN	181	0.0289	0.0063

We acknowledge that the computational time of SIRGCN is similar to that of the baseline models, as the baselines are not as deep or dense as traffic prediction models and do not require a large amount of data for training.

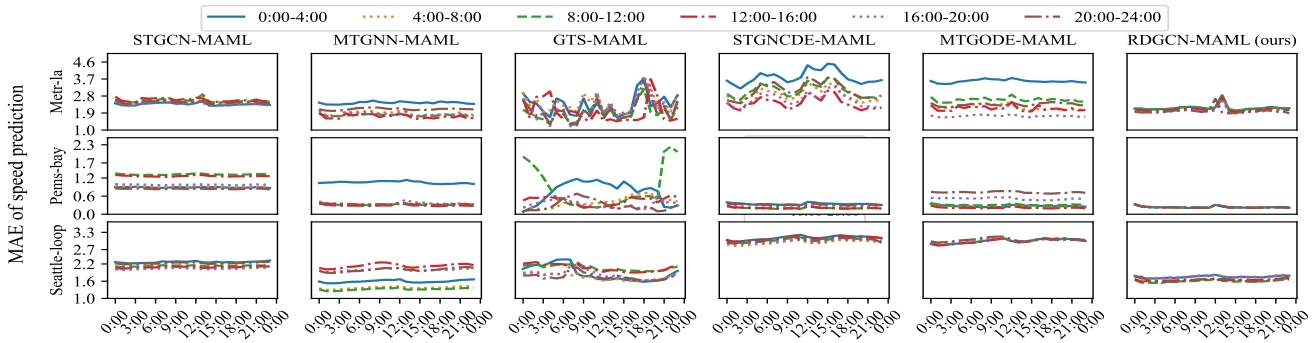
Table 6: The computation time on the Japan-Prefectures dataset.

Experimental Settings

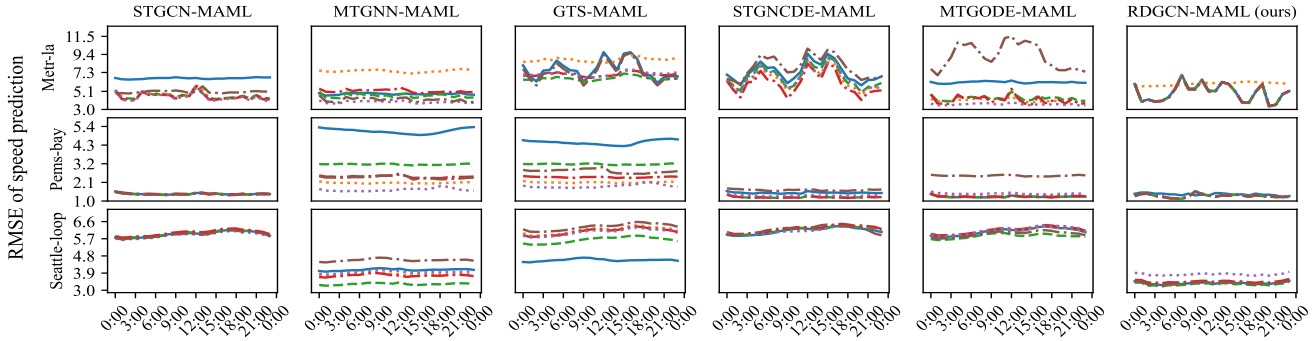
Evaluation. We assume that all zeros in the datasets are missing values, and we remove the predicted speed when the ground truth is 0, or when the last speed recorded is 0.

Hyperparameter Settings. RDGCN and SIRGCN are optimized via Adam. The batch size is set as 64. The learning rate is set as 0.001, and the early stopping strategy is used with a patience of 30 epochs. In traffic speed prediction, the training and validation set are split by a ratio of 3:1 from the weekday subset, and the test data is sampled from the weekend subset with different patterns. As for baselines, we use identical hyperparameters as released in their works. In ILI prediction, the training and validation set are split by a ratio of 5:2 from the Winter-Summer subset, and the test data is sampled from the Spring-Fall subset with different patterns. The Susceptible population at the beginning of each ILI period is 10% of the total population in each Prefectures or States. As for baselines, we also use identical hyperparameters as released in their works. We approximate the total number of populations by the average of the annual sum of infectious cases, multiplied by 10.

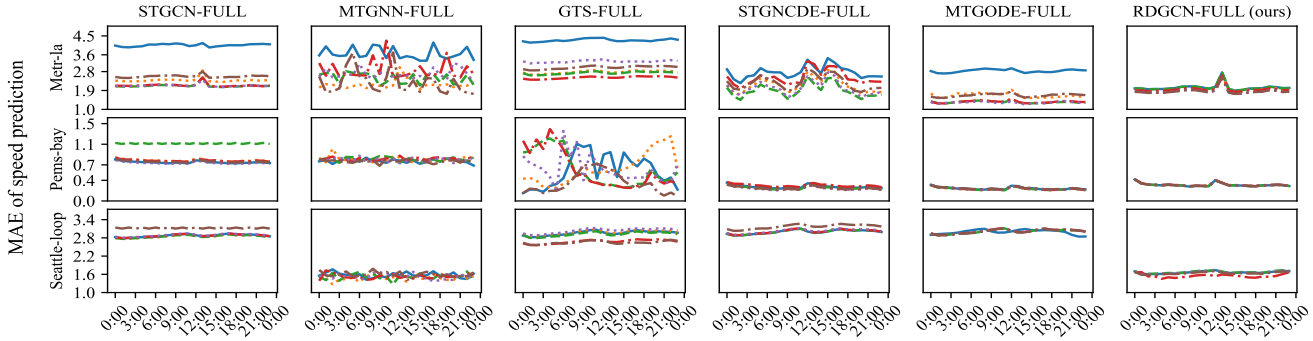
MAML Settings. Our experiment involves the following steps: (1)We randomly select sequences of 12 consecutive weekdays (same as the Limited and Mismatched Data experiment.), and sample four-hour data as the training set. We evaluate the model with hourly data on weekends. (2) We divide the training set into two equal parts: the support set and



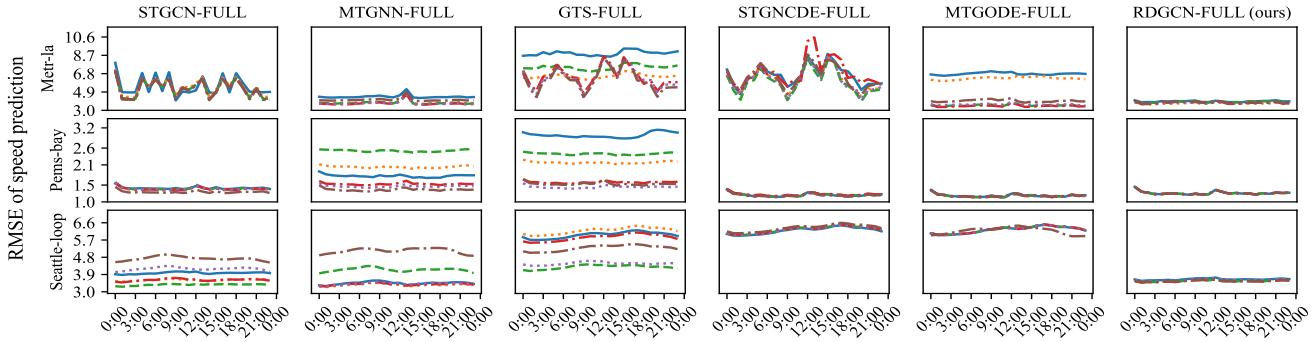
(a) Baseline models and RDGCN trained on 12 consecutive weekdays and augmented by MAML.



(b) Baseline models and RDGCN trained on 12 consecutive weekdays and augmented by MAML.



(c) Baseline models and RDGCN trained on more than half a year of weekdays.



(d) Baseline models and RDGCN trained on more than half a year of weekdays.

Figure 8: (a)(b) The results of RDGCN are very close regardless of the period of the training set. (c)(d) Even though all the models are trained using all available weekdays, the results of RDGCN are still closer, regardless of the period, compared to baseline models.

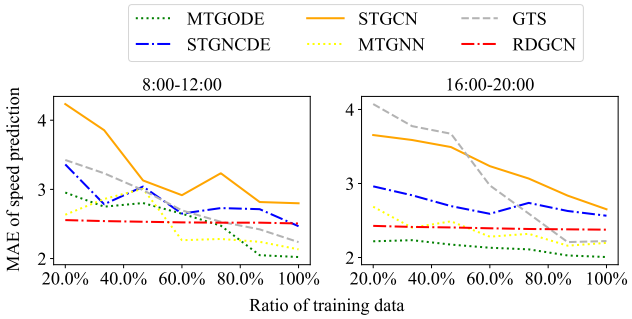


Figure 9: Feeding more training data does not lead to a significant change in the MAE of RDGCN's prediction.

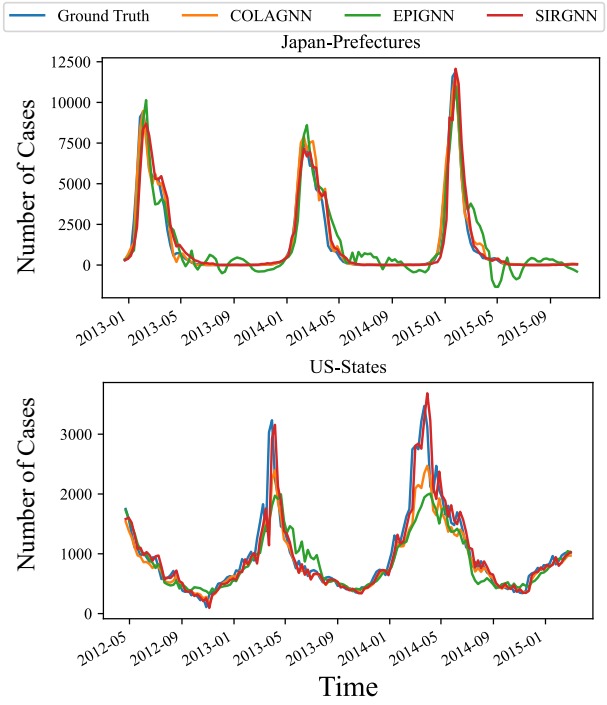


Figure 10: SIRGCN can make accurate predictions in the decreasing phase, while EpiGNN makes bad predictions in the corresponding phase.

the query set. (3) We use the support set to compute adapted parameters. (4) We use the adapted parameters to update the MAML parameters on the query set. (5) We repeat this process 200 times to obtain initial parameters for the baseline model. (6) We train baselines using the obtained initial parameters. The learning rate for the inner loop is 0.00005, and for the outer loop is 0.0005, and MAML is trained for 200 epochs.