Life of Research · Published 22 November 2026 · DOI 10.5281/zenodo.24430423

A Recovered Average Is Not a Defended Corridor: What Adversarial Training Buys Spatiotemporal Traffic Forecasters, Why the Case That Static Methods Fail Rests on Node Selection While the Evidence of Failure Comes From Attack Shape, and the Evaluation That Would Tell Them Apart

Pranay Mahendrakar
SONYTECH · ORCID 0009-0003-7224-029X
pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Spatiotemporal graph networks forecast speed and flow across road sensor networks, and a line of work since 2021 shows that a few perturbed sensors can wreck those forecasts. Whether adversarial training repairs the damage is disputed. One position, stated most sharply by the reinforced dynamic adversarial training (RDAT) paper at KDD 2023, holds that adversarial training designed for static domains does not carry over to traffic, because the attacker picks different victim sensors at each moment, and that defence needs a learned, time-varying choice of training nodes. A second position, from attack papers, time-series robustness studies and a 2025 study of forecasting foundation models, reports that ordinary adversarial training, or for the foundation models lightweight latent adversarial fine-tuning, already recovers most of the lost accuracy. A third, from a 2026 re-evaluation, holds that adversarial training tuned to norm-bounded perturbations barely transfers to targeted, congestion-shaped attacks, and that even training on the attack itself loses ground when the attack's reach changes. In this analysis the author reviews and synthesises published results to show that the three positions are answering different questions under one word. A "dynamic attack" can vary in at least five things: which sensors it touches over time, the shape and size of the perturbation, the objective, the attacker's knowledge, and the metric used to score it. The dynamic-domain thesis concerns only the first. In RDAT's own main tables, adversarial training with randomly chosen nodes, the baseline RDAT presents as traditional adversarial training and elsewhere labels a dynamic strategy, closes about 93 percent of the attack-induced error on both benchmarks with Graph WaveNet, including against the time-dependent attack. Against that attack the learned selector adds less than two points of the gap, within run-to-run spread; under the other four attacks it stays within the spread on PeMS-D4 but adds 3.3 to 3.8 points on PeMS-BAY, slightly more than AT's seed spread. In the appendix, under the time-dependent attack, the ordering of the two methods flips with architecture while both close only 52 to 71 percent. Node selection is also the smaller lever in the attack literature's own ablations. The failures that are large come from changes in perturbation shape, target and metric. The paper consolidates the measured values in a table and a figure, states a procedure for crediting a robustness gain to a defence's dynamic component, and names the evaluations that would settle what remains open.

Keywords: traffic forecasting; spatiotemporal graph neural networks; adversarial training; adversarial robustness; threat models; time-series forecasting; sensor spoofing; evaluation validity

1. Introduction

A traffic forecaster takes the last hour of readings from a few hundred road sensors and predicts the next hour at every one of them. The models that lead the standard benchmarks are spatiotemporal graph neural networks: diffusion convolution combined with recurrence (Li, Y. et al., 2017), graph convolution combined with temporal convolution (Yu, Yin and Zhu, 2017), and adaptive learned adjacency with dilated temporal convolutions (Wu, Z. et al., 2019). Their forecasts feed route guidance, signal timing and incident response, which is why their failure under deliberate manipulation has drawn attention. Falsified data fed to signal controllers is already known to change delay on real corridors in simulation (Feng, Y. et al., 2018; Huang, S. E. et al., 2020), and analog sensors are known to accept injected signals (Foo Kune et al., 2013). Learned forecasters add a new surface: an attacker who changes readings at a few sensors can move predictions at many. Adversarial examples were first described for image classifiers (Szegedy et al., 2013; Goodfellow, Shlens and Szegedy, 2014), and graph neural networks were later shown to be vulnerable to small changes in node features and edges (Zugner, Akbarnejad and Gunnemann, 2018; Dai, H. et al., 2018; Xu, K. et al., 2019), a literature reviewed by Jin et al. (2020). Spatiotemporal forecasters inherit both weaknesses.

That surface has been mapped in some detail. Liu, F., Miranda-Moreno and Sun (2021) perturbed a single vertex and watched the error spread across much of the graph. Zhu, L. et al. (2021) selected a small attack set under a cost budget with black-box gradient estimates. Poudel and Li (2021) trained a substitute model from queries and transferred its perturbations. Liu, F., Liu, H. and Jiang (2022) chose victim nodes afresh at each time step by gradient saliency and reported that attacking 10 percent of the nodes raised global mean absolute error on PeMS-BAY from 1.975 to 6.13. Most of these papers close by discussing what would defend against the attack they describe.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

The most direct answer is Liu, F., Zhang, W. and Liu, H. (2023), hereafter RDAT. Its abstract makes a claim with two parts. The first is negative: adversarial training methods designed for static domains such as images and graphs cannot be applied directly to traffic forecasting, because they do not defend against dynamic attacks. The second is constructive: a reinforcement-learned policy that chooses which nodes to perturb during training, plus a self-distillation term against what the authors call a forgetting issue, does defend. The negative part has become the field's premise. A 2026 re-evaluation describes the prevailing message as that adversarially trained forecasters are robust, with RDAT as its culmination (Zhang, Q., 2026), and then argues the opposite: adversarial training barely dents a targeted attack shaped like real congestion.

So the literature now holds three positions. Static adversarial training fails and dynamic training succeeds (RDAT). Ordinary adversarial training largely succeeds (the attack papers' own defence sections, multivariate forecasting work, and a study of time-series foundation models that uses lightweight latent adversarial fine-tuning). Adversarial training fails once the attack's shape differs from the one it was trained on (Zhang, Q., 2026). These cannot all be right about one quantity. In this analysis the author reviews and synthesises published results to show that they are not about one quantity, and that the variable the dynamic-domain thesis emphasises is the least consequential of the ones in play.

The argument has four steps. First, Section 3 separates five things a "dynamic" attack can vary, and shows that the dynamic-domain thesis is a claim about only one of them, the time-varying choice of victim nodes. Second, Section 5 re-reads RDAT's own tables: the baseline it labels as traditional adversarial training already samples nodes at random each step, and it closes most of the error the attack opens; the evidence that a static method fails is a single bar chart comparing training on high-degree or high-PageRank nodes against random sampling. Third, Sections 6 and 7 show that, across the attack literature, node selection moves attacked error by roughly 3 to 35 percent in the two published ablations, and defended error by a few to about ten percent, while changes in the perturbation's shape and the attack's target move them by multiples. Fourth, Sections 4, 8 and 9 show that the metric and the strength of the evaluating attacker decide what "defended" means, and that the benchmark protocol is generous to the attacker in some respects and stingy in others. Section 10 turns to defences that do not depend on knowing the attack, and Section 11 states the strongest version of the case that traffic really is different, and what it predicts.

The paper's contribution is the reconciliation, the re-reading of the anchor from its own numbers, and an explicit procedure (Algorithm 1) for crediting a robustness gain to the dynamic component of a defence. It does not adjudicate which defence is best. The evidence base is too thin for that, and Section 15 says where it runs out.

Three conventions hold throughout. "Adversarial training" (AT) means the min-max procedure of Madry et al. (2017), in which a model is trained on inputs perturbed to maximise its loss. "Gap closure" is this paper's arithmetic on published numbers: the share of the error an attack adds to an undefended model that a defence removes, computed as (undefended attacked error minus defended attacked error) divided by (undefended attacked error minus undefended clean error). Every value in the table and the figure is a published one, credited to its source on its row or in its caption, and every derived ratio is labelled as derived.

2. What each side claims

2.1 The dynamic-domain claim

RDAT's argument begins from a diagnosis. In images, a perturbation can touch every pixel; in a sensor network, an attacker reaches only some sensors, and a practical attacker, following Liu, F., Liu, H. and Jiang (2022), chooses which ones at each time step. RDAT's introduction reports two exploratory results. In one (its Figure 2a), training with adversarial examples placed on nodes chosen by degree or PageRank produced higher error under attack than placing them on randomly sampled nodes. In the other (its Figure 2b), raising the share of nodes perturbed during training from 20 toward 100 percent made the trained model worse rather than better; the authors call this counterintuitive and relate it to overfitting, illustrated by training curves at 80 and 10 percent (its Figure 4). They add a third observation, that a model trained against a continually changing set of adversarial nodes loses robustness to stronger attacks, and borrow the name forgetting from the memorisation literature (Dong, Y. et al., 2021).

The proposed remedy has three parts. A policy network with a Graph WaveNet-style encoder and an attention decoder picks a fixed number of nodes per step to maximise the inner loss, trained with REINFORCE against a reward measured relative to a heuristic baseline selector (the time-dependent node saliency method of the 2022 attack paper, TNDS). The forecaster is then trained by projected gradient descent on those nodes. A self-distillation term pulls the current model's predictions on perturbed inputs toward the previous epoch's predictions on clean inputs. The paper concludes that traditional static methods are unsuitable for traffic and that its framework fixes this.

2.2 The sufficiency claim

The attack papers that RDAT builds on report that ordinary adversarial training works. Liu, F., Liu, H. and Jiang (2022) end with a defence experiment on PeMS-BAY in which conventional adversarial training, with random victim nodes, takes global mean absolute error under their random-node attack from 6.15 to 2.54, against a clean error of 2.03; training on their own saliency-chosen nodes gives 2.47. They write that adversarial training or Mixup can successfully resist the attack.

Outside the graph setting the same pattern recurs. For multivariate probabilistic forecasting, Liu, L. et al. (2023) designed sparse, indirect attacks that perturb other series in order to move a target series, and a min-max defence that trains against a learned sparse perturbation. On the UCI Traffic dataset their defence, trained at one sparsity level, held weighted quantile loss near its clean value at a sparsity nearly twice as large, where the undefended model's loss had more than doubled. For time-series foundation models, Zhang, J. et al. (2025) found that lightweight latent adversarial fine-tuning cut worst-case error under white-box attack by a factor of roughly four to ten, and that fine-tuning on an unrelated dataset kept most of that gain. Single-model studies of predictive maintenance (Siddique et al., 2023) and smart-infrastructure forecasting (Krishan et al., 2024) report large reductions in attacked error after adversarial training. On graphs, adversarial training on perturbed structure (Xu, K. et al., 2019) and on features of connected nodes (Feng, F. et al., 2019) predates the traffic work, and the latter, GraphAT, is one of RDAT's own baselines. Earlier work had already shown that regression forecasters inherit the vulnerability of classifiers (Mode and Hoque, 2020; Dang-Nhu et al., 2020) and that time-series classifiers do too (Fawaz et al., 2019; Karim, Majumdar and Darabi, 2019); forecasters built on large language models have since been attacked in the same way (Liu, F., Jiang et al., 2024).

2.3 The attack-specificity claim

Zhang, Q. (2026) argues that the benchmark threat model is unrealistic in five ways. It grants the attacker white-box knowledge, a large share of the sensors, full observation, a network-wide objective and an l-infinity budget borrowed from imperceptible image perturbations. A deployable attacker, on this account, knows less and reaches fewer sensors, but wants something local, such as a slower predicted travel time on one corridor, and is limited not by a norm but by what a sensor can plausibly report. Zhang builds such an attack. It targets a five-node corridor, drops the norm ball, and penalises spatial inconsistency with neighbours and temporal jerk so that the injection resembles real congestion. Against it, RDAT's forecasters improve on no defence in some cells and are worse in others, while adversarial training on the physics-aware attack itself does much better, but degrades as the attack is widened beyond the reach it was trained on. Zhang's conclusion is that adversarial-training robustness is earned per threat and does not transfer across attack shapes.

This is the image-domain finding in traffic form. Tramer and Boneh (2019) proved that robustness to different perturbation types must trade off in a simple statistical model, showed the trade-off on MNIST and CIFAR10, and observed that training against one type can increase vulnerability to another. Kaufmann et al. (2019) built eighteen attacks outside the l-p family and found that standard robustness measures did not predict robustness to unforeseen adversaries. Dai, S. et al. (2023) benchmarked sixteen defended models against 180 attack settings and found worst-case robustness across attack types still largely unsolved.

2.4 The anchor, and why it matters which claim is right

The three claims make different recommendations. If the dynamic-domain claim holds, the engineering priority is the training-time node selector. If the sufficiency claim holds, standard adversarial training is enough and attention should go elsewhere. If the attack-specificity claim holds, no training-time recipe tuned to one threat will carry the defence, and the priority is a threat model that matches deployment, plus defences that do not depend on knowing the attack. The rest of this paper argues that each claim is right about something and wrong about what the others measured.

3. Five things a "dynamic attack" can vary

The phrase dynamic adversarial attack carries more than one meaning in this literature. A threat model for a sensor-network forecaster can be written as five components.

S, the node-selection rule. Which sensors the attacker perturbs, and whether that set changes over time. A static rule fixes the set (for example, the highest-degree nodes); a dynamic rule recomputes it per input window (for example, by gradient saliency on the current window, as in TNDS). The size of the set is the node budget.

Delta, the perturbation set. What values the attacker may write at the selected sensors. The benchmark choice is an l-infinity ball of radius epsilon around the true reading. Alternatives include sparse l-0 patterns, structured shapes that respect spatial and temporal smoothness, and capability bounds set by the sensor's physical range.

J, the objective. What the attacker is trying to achieve. Untargeted attacks maximise network-wide forecast error. Targeted attacks push one corridor's forecast toward a chosen value, such as a fabricated jam.

K, knowledge. Whether the attacker has the model's gradients, a surrogate, queries only, or partial observation of the network; and whether it has the label. In forecasting the label is the future, which a real attacker does not have at run time. Liu, F., Liu, H. and Jiang (2022) build a surrogate label from the model's own prediction for this reason and call that setting grey-box.

M, the metric. How the attack and the defence are scored. Global MAE averages over every node and horizon; local MAE averages over victims; target MAE averages over a chosen corridor; stratified metrics separate classes of observation.

Under this decomposition, the dynamic-domain thesis is a claim about S. It says that a defence trained with a static S fails against an attacker with a dynamic S, and that a defence whose training S is learned and time-

varying succeeds. It holds Delta, J, K and M fixed: in RDAT's experiments the perturbation set is an l-infinity ball of radius 0.5 on min-max-normalised data, the objective is untargeted, the attacker is white-box, and the metric is global MAE and RMSE.

The attack-specificity thesis, by contrast, is a claim about Delta and J, scored under a different M. Zhang's attack changes the perturbation set from a norm ball to a smooth congestion shape, changes the objective from global to a five-node corridor, and changes the metric to target-corridor MAE. When the two papers disagree about whether adversarial training works, they are reporting results under different values of four of the five components.

The sufficiency thesis varies little. Its evidence is mostly matched: the defence is trained against the same family of attack it is tested against, under the same metric. That is the regime in which the image-domain literature also finds adversarial training effective (Madry et al., 2017; Bai, T. et al., 2021).

Laid out this way, two predictions follow. If the dynamic-domain thesis is right, varying S alone, with Delta, J, K and M fixed, should produce large differences in defended error. If the attack-specificity thesis is right, varying Delta or J, with S fixed, should produce large differences. Sections 5 to 7 check both predictions against the published numbers.

4. The budget the benchmark grants

Before reading the defence results it is worth being exact about what the benchmark attacker is allowed to do, because it is generous in some directions and restrictive in others.

Magnitude. In both Liu, F., Liu, H. and Jiang (2022) and RDAT, traffic data are normalised to the range zero to one and epsilon is 0.5. With min-max normalisation, that allows each perturbed reading to move by up to half of the dataset's whole range of values; the 2022 paper's own Figure 1 caption describes the setting as perturbing speed by less than 50 percent. In image classification the comparable convention for CIFAR-10 is 8/255, about 3 percent of the pixel range (Wong, Rice and Kolter, 2020). The 2022 paper nonetheless calls its constraint human-imperceptible. Whatever that word means for a time series, a shift of half the full range is not small in the sense the image literature intends, and it is also not obviously plausible for a working detector. Gilmer et al. (2018) made the general point that defence papers had mostly studied abstract games with no stated security concern, and asked for threat models that state what an attacker can actually do.

Reach. The benchmark attacker perturbs 10 percent of nodes in the 2022 paper and 20 percent at test time in RDAT, chosen anew for every input window. On PeMS-BAY's 325 sensors that is 33 or 65 sensors, re-chosen every five minutes. Zhang, Q. (2026) argues that physical or network access to a sensor is costly, so that a realistic attacker controls a small fixed set; its own attack uses five nodes plus a one-hop halo.

Knowledge. RDAT evaluates white-box, with five steps of projected gradient descent and step size 0.1. The 2022 paper reports white-box, grey-box and black-box variants; its black-box attack through a surrogate degrades accuracy less than the grey-box one but still substantially.

Shape. In the other direction, the benchmark attacker is restricted. Its perturbation is any point in a box, with no requirement of temporal smoothness or spatial coherence; but it may not leave the box, however physically plausible a larger change might be. Zhang argues that this is the wrong constraint: a sensor can be made to report any value in its dynamic range, so what makes an injection hard to catch is resemblance to real traffic, not small size.

Consistency across windows. A forecaster with a twelve-step history reads each sensor value in twelve overlapping input windows. A physical attacker writes one value per sensor per time step, and that value then appears in every window that contains it. As far as the published descriptions state, the benchmark attacks of Liu, F., Liu, H. and Jiang (2022) and RDAT optimise a fresh perturbation, and in the dynamic case a fresh victim set, for each input window independently. If so, the same historical reading can take one spoofed value in one window and a different value, or no perturbation at all, in the next. That freedom is not available to an attacker in a streaming deployment, and it is part of what makes the benchmark's node selection "dynamic". This paper infers it from the method descriptions; neither paper discusses it, and a streaming-consistent version of either attack has not been reported.

The label. Benchmark adversarial training uses the clean future reading as the target. If an attacker spoofs a sensor persistently, the recorded future at that sensor is also spoofed. The forecaster is then being asked to predict the true traffic state, which is unobserved at the attacked sensor, rather than the sensor's next report. None of the papers reviewed here discusses this; it is a point about what robustness should mean in forecasting that the classification framing hides.

The classical traffic-engineering literature already knew that sensor data are unreliable for mundane reasons. The PeMS system itself runs diagnostics that flag malfunctioning single-loop detectors from their volume and occupancy time series, and imputes bad samples from neighbouring loops by regression (Chen, C. et al., 2003). That pipeline is a non-adversarial ancestor of the physics-informed detectors proposed in 2026, and it implies that a crude, isolated spoof would be caught before a forecaster ever saw it. A recent sensor-fault benchmark (Windmann et al., 2026) makes the same point from the robustness side: clean-error rankings of forecasting architectures can disagree with their rankings under realistic faults.

So the benchmark grants an attacker a very large per-sensor magnitude, many sensors, fresh choices every

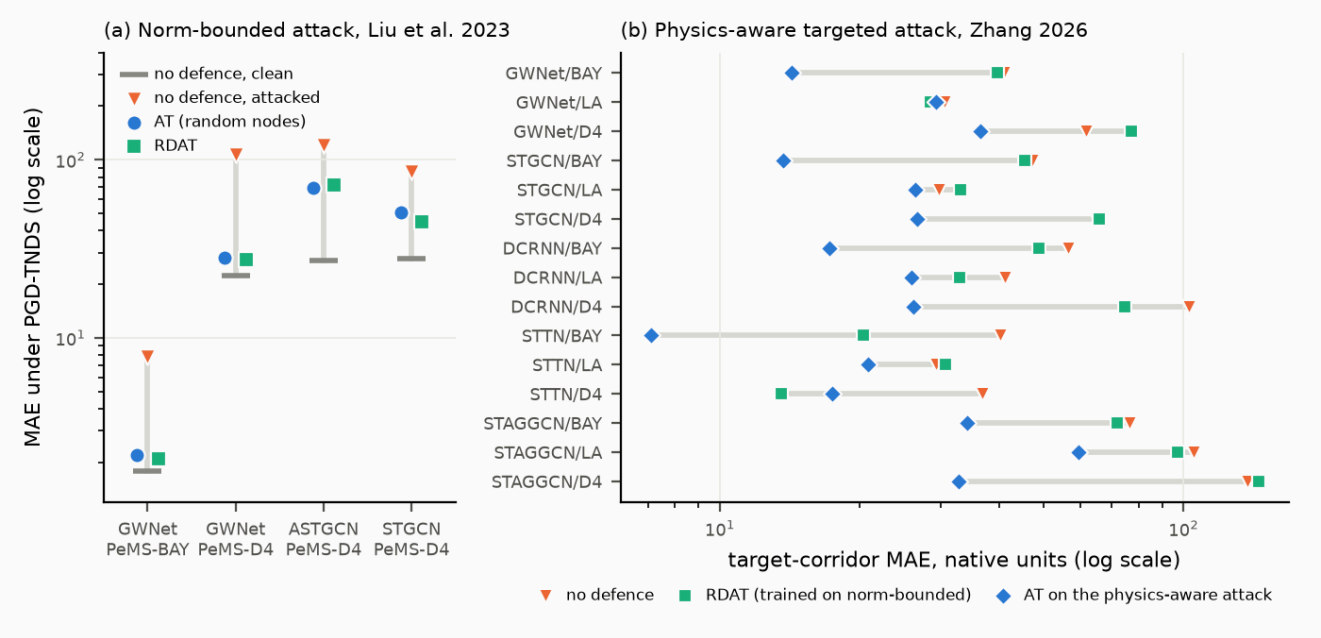

Figure 1. (a) MAE under the time-dependent PGD-TNDS attack for four model-dataset settings in Liu, F., Zhang and Liu (2023), Tables 1-4: clean undefended (bar), attacked undefended (triangle), adversarial training with random nodes (circle) and RDAT (square). (b) Target-corridor MAE under the physics-aware attack for fifteen model-dataset cells in Zhang, Q. (2026), Table 4: no defence, RDAT trained on the norm-bounded attack, and adversarial training on the physics-aware attack. All values as published; log scales.

window and full gradients, while forbidding the coherent shapes a sophisticated attacker would use. Robustness measured inside that box need not predict robustness outside it in either direction. Apruzzese et al. (2022) document the analogous gap in security practice, where real attackers favour cheap tactics over gradient computation, and Kurakin, Goodfellow and Bengio (2016) showed long ago that adversarial images can survive passage through a physical channel, a phone camera, though not all of them do, so that the channel itself is part of the threat model.

5. The anchor re-read: what RDAT's tables show

RDAT reports results in two main tables (PeMS-BAY speed and PeMS-D4 flow, both on Graph WaveNet) and two appendix tables (PeMS-D4 on ASTGCN and STGCN). Each row is a defence; each column is an attack, from no attack through PGD with random, PageRank, centrality, degree and TNDS node selection. Each cell is the mean and standard deviation over five runs.

The first thing to establish is what the baseline called AT is. RDAT's appendix describes it as traditional adversarial training that selects adversarial nodes randomly. Random selection is redrawn at each training step. It is therefore not static in the sense of Section 3: its training-time S is time-varying, merely not learned. RDAT itself classifies it this way: its Figure 2 caption contrasts "static (PageRank and Degree) and dynamic (Random) defense strategies", and its ablation describes "AT-Random, which selects nodes randomly in a dynamic manner". The comparison between AT and RDAT is a comparison between random and learned dynamic selection, not between static and dynamic.

Graph WaveNet on PeMS-BAY. Undefended, the model's clean MAE is 1.78; under the time-dependent attack it is 7.82. AT brings the attacked error to 2.20; RDAT to 2.09. By this paper's arithmetic, AT closes 93.1 percent of the gap and RDAT 94.8 percent. The difference between them, 0.10 MAE, is smaller than AT's run-to-run standard deviation in that cell, 0.15. On RMSE under the same attack, GraphAT (4.38) is lower than RDAT (4.45), as the RDAT text acknowledges.

Graph WaveNet on PeMS-D4. Clean undefended MAE is 22.42; attacked, 106.16. AT reaches 28.10 and RDAT 27.54, closing 93.2 and 93.9 percent respectively. The standard deviations are 1.23 and 1.00. On RMSE, AT with TNDS training nodes (42.46) is marginally lower than RDAT (42.53).

Under the random-node attack AT closes 91.1 percent on PeMS-BAY and 93.9 percent on PeMS-D4, RDAT 94.9 and 94.7 percent. On PeMS-D4 the difference again lies inside AT's seed spread (0.50 MAE against a standard deviation of 1.27). On PeMS-BAY it does not: the 3.8-point difference is 0.18 MAE, against an AT standard deviation of 0.15, and the PageRank, centrality and degree attacks give the same result (3.3 to 3.7 points, MAE differences of 0.15 to 0.17 against AT deviations of 0.14 to 0.16).

ASTGCN and STGCN on PeMS-D4. The appendix tells a different story, and RDAT reports it with a caveat that adversarial training shows considerable variability on these architectures. On ASTGCN, under the time-dependent attack, the undefended error rises from 27.28 to 120.17. AT reaches 69.60 and RDAT 71.80. Plain AT is the better of the two here, and both close only about

half of the gap (54.4 and 52.1 percent). On STGCN, AT reaches 50.63 and RDAT 44.97 from an undefended 86.09, closing 60.9 and 70.6 percent; the standard deviations, 11.2 and 9.9, are each larger than the difference. One caution about the source: in the STGCN table the undefended TNDS cell repeats the undefended centrality cell digit for digit, mean and deviation for both metrics, and in the AT-TNDS row the clean RMSE repeats the random-attack RMSE. These may be transcription slips; the paper does not say, and this analysis uses the values as printed.

Figure 1(a) plots these values. Read together, they support three statements, each narrower than the abstract's.

First, adversarial training with a randomly resampled node set, which the paper presents as the traditional method, closes more than nine-tenths of the attack's damage on the architecture and benchmarks the paper leads with. It does not fail to defend.

Second, under the time-dependent attack the increment from learning the node selector is 0.7 to 1.7 points of gap closure on Graph WaveNet, within run-to-run variation. Under the other four attacks it stays within that variation on PeMS-D4, but on PeMS-BAY it is 3.3 to 3.8 points and slightly exceeds AT's seed spread. On the appendix architectures, under the time-dependent attack, it is positive in one case and negative in the other, and its sign also varies across attack columns: RDAT is ahead of AT on ASTGCN under the other four attacks (for example 57.17 against 60.91 under the random-node attack), and behind it on STGCN under the random-node attack (41.52 against 39.82).

Third, the residual vulnerability depends far more on the architecture than on the defence. AT and RDAT both close more than 90 percent of the gap on Graph WaveNet and between about a half and seven-tenths on ASTGCN and STGCN; the other baselines show the same drop from the first architecture to the other two. RDAT's own appendix offers a mechanism, that Graph WaveNet learns its spatial relationships from data and may give small weights to attackable nodes, without testing it.

What, then, is the evidence that static methods fail? It is the introduction's Figure 2a: three bars showing that training with nodes chosen by PageRank or degree yields higher attacked MAE than training with random nodes, and the ablation in Figure 6, which includes a degree-based variant. These do show that a fixed node set chosen by graph centrality is a poor training distribution against an attacker who chooses nodes by saliency. That is a genuine finding. It is also what the general theory of adversarial training would predict, since a training distribution that never perturbs most nodes cannot teach the model to resist perturbations there. But it is a statement about one bad way of choosing training nodes, not about adversarial training methods from static domains in general, several of which, as implemented in the same tables (AT, Mixup, GraphAT), perform close to RDAT.

There is one result in RDAT that cuts the other way and should be stated plainly. When the test attack is widened from 20 percent of nodes to 40, 60, 80 and 100 percent, RDAT reports that its method stays best at every strength, with an improvement of 14.0 percent on PeMS-BAY and 2.9 percent on PeMS-D4 at 100 percent. The per-method values behind those percentages appear only in a figure with a logarithmic axis, so they cannot be read off exactly, and the text does not say which baseline each percentage is computed against. Still, they suggest that the learned selector's margin may grow as the test attack reaches more nodes, at least on PeMS-BAY. If that holds, the dynamic component's value lies in robustness to attacks wider than those trained on, which is a claim about transfer along the node-budget axis and one worth testing directly.

A careful reader might object that the dynamic-domain claim is about stability rather than mean error: RDAT has lower standard deviation than several baselines, and its ablation attributes much of that stability to the self-distillation term. That is a fair reading of the method's benefit. It is a smaller claim than the abstract's, and a different one.

6. Node selection is the small lever

If the dynamic-domain thesis were right, the choice of which nodes to attack, with everything else fixed, should be a large lever on both attack strength and defence. The attack literature's own ablations say it is small.

In Liu, F., Liu, H. and Jiang (2022), the full attack combines saliency-based node selection (TNDS) with a spatiotemporal projected gradient method (STPGD). Their ablation removes each in turn on PeMS-BAY. Replacing TNDS with random selection reduces attack strength by 3.91 percent in global MAE; replacing STPGD with vanilla PGD noise on the selected nodes reduces it by 33.41 percent. The authors note that the perturbation module matters more. In their defence table, training on saliency-chosen nodes (AT-TDNS, 2.47) beats training on random nodes (AT, 2.54) by 0.07 MAE under the random-node attack, with both far below the undefended 6.15.

Zhang, Q. (2026) compares four ways of choosing a five-node set: graph degree, surrogate saliency, a connected corridor, and random. Degree, which needs no model access at all, gives the largest attacked error in three of four model-dataset settings. In the fourth, Graph WaveNet on PeMS-BAY, random choice is the strongest (7.0 mph) and degree the weakest (5.2). Zhang's summary is that node choice matters least among the attacker's levers, behind budget, optimisation steps and observation coverage.

RDAT's own evidence fits this. Its node-ratio experiment (Figure 2b) varies how many nodes are perturbed in training, not which ones, and finds that more is worse. That is a statement about the strength of the training perturbation, and it has a well-studied counterpart. Adversarial training overfits: Rice, Wong and Kolter

(2020) found that robust test performance degrades with continued training across several datasets and threat models, and that early stopping matched most claimed algorithmic gains. Schmidt et al. (2018) showed that robust learning can need far more samples than standard learning. Dong, Y. et al. (2021) tied robust overfitting to memorisation of adversarial training examples. Wong, Rice and Kolter (2020) described a catastrophic form of overfitting in fast single-step training. Tsipras et al. (2018) and Zhang, H. et al. (2019) formalised the tension between clean accuracy and robustness that TRADES trades off. A training distribution that perturbs 80 percent of nodes by up to half the data range is a strong adversary on a dataset of a few hundred sensors; that it degrades the trained model is what the general literature would lead one to expect, not a property peculiar to dynamic domains.

The forgetting issue has the same shape. RDAT cites the memorisation and robust-overfitting work for it. Its remedy, distillation from the previous epoch, is a regulariser on the trajectory of training, a family the general literature has explored extensively (Bai, T. et al., 2021). Its benefit in RDAT's ablation is mainly to variance.

None of this says node selection is irrelevant. It says that, within the norm-bounded untargeted regime where the dynamic-domain thesis was tested, the published ablations put node selection at roughly 3 to 35 percent of attacked error, and the defence comparisons at a few to about ten percent of defended error, while perturbation method, architecture and training strength move results by much more.

7. Attack shape is the large lever

The prediction of the attack-specificity thesis is that changing Delta or J while holding S fixed produces large changes. The evidence for this is stronger, though it rests heavily on one recent preprint.

Zhang's physics-aware attack. Under the norm-bounded ablation (epsilon 0.5, 30 PGD steps, five-node corridor), every architecture's target-corridor MAE at least doubles, and most rise severalfold: Graph WaveNet on PeMS-BAY from 1.9 to 6.1 mph, STGCN on PeMS-D4 from 7.1 to 90.5 vehicles per five minutes, DCRNN on PeMS-D4 from 6.2 to 152.7. Against the physics-aware attack, Zhang's Table 4 reports target-corridor MAE for no defence, RDAT trained on the norm-bounded attack, and adversarial training on the physics-aware attack itself (AT-physics). Figure 1(b) plots all fifteen cells.

Counting from that table, RDAT is lower than no defence in ten of fifteen cells and higher in five. The reductions are small in most cells (Graph WaveNet on PeMS-BAY, 41.1 to 39.7) but large in a few (STTN on PeMS-BAY, 40.2 to 20.4; STTN on PeMS-D4, 36.9 to 13.6; DCRNN on PeMS-D4, 102.5 to 74.7). The increases include Graph WaveNet on PeMS-D4 (61.5 to 77.2) and STAGGCN on PeMS-D4 (137.2 to 145.0). Zhang describes norm-trained adversarial training as nearly useless against this attack. The table supports a softer version: transfer is partial, inconsistent in sign and strongly architecture-dependent. Adversarial training on the physics-aware attack, by contrast, cuts target error to between about a fifth and three-fifths of the undefended value in every PeMS-BAY and PeMS-D4 cell (Graph WaveNet on PeMS-BAY 41.1 to 14.3; DCRNN on PeMS-D4 102.5 to 26.2), with METR-LA weaker, where it keeps between about half and nearly all of the undefended error (0.56 to 0.96; Graph WaveNet 30.5 to 29.3, STAGGCN 105.0 to 59.3).

Reach as a shape parameter. Zhang then widens the attack's halo from the one hop it was trained on to two and three hops. AT-physics on STGCN with PeMS-BAY goes from 13.7 to 25.6 to 34.8, rising toward the undefended one-hop value of 47.3. The undefended error under the wider attacks is not reported, so the size of the lost benefit cannot be computed exactly, but a defence trained against one spatial footprint clearly loses much of its advantage when the footprint grows. That is a change in Delta with S, J and M fixed, and it is large.

Black-box and query attacks. Zhang, J. et al. (2025) tested six time-series foundation models with white-box PGD and with the query-based SimBA attack. Adversarial fine-tuning was highly effective against PGD; on the Loop Seattle traffic-speed data, TimesFM's error under PGD fell from 1.213 to 0.132 against a clean 0.113. Against SimBA the gains were marginal and sometimes negative: on hierarchical sales data, every adversarially fine-tuned variant had higher SimBA error (2.37 to 2.82) than the undefended model (1.82). The authors read this as a mismatch between gradient-based training and the geometry of query-based attacks. Anand and Pappas (2026) found that model-free attacks exploiting a state-space forecaster's locally linear behaviour caused at least a third more error than PGD with a small step size, which bears both on how strong a PGD-based evaluation is and on whether PGD-trained robustness covers such attacks.

Faults as a different perturbation family. Windmann et al. (2026) trained forecasters with PGD adversarial training and scored them under realistic sensor faults, from noise and bias to dropped and misaligned readings. On the single-target datasets where value faults dominated, PGD training reduced relative degradation. On the Traffic dataset, where availability faults dominated, PGD training lowered absolute error but increased relative degradation, at all three seeds; the increase is small, though, and under the source's alternative selection rule it changes sign ("PGD adversarial training on Traffic moves from 0.023 to −0.001"; Windmann et al., 2026). The perturbation family the defence was trained on was not the one that hurt.

Transfer within a family can be good. Not every shift breaks adversarial training. Liu, L. et al. (2023) trained their mini-max defence with a sparse layer at sparsity five and evaluated it from sparsity one to nine on the UCI Traffic data. Under the deterministic attack at

sparsity nine, the undefended model's loss was 0.464 and the defended model's 0.197, below the undefended model's clean loss of 0.228. Zhang, J. et al. (2025) report that cross-domain adversarial fine-tuning kept most of the in-domain gain against PGD. Perceptual adversarial training in images generalised to several unseen attack types (Laidlaw, Singla and Feizi, 2020). The picture is not that adversarial training never transfers. It transfers along the axis it was trained on, such as more nodes, more sparsity or a different dataset, and less across changes of perturbation shape or attack mechanism.

Why shape matters. Stutz, Hein and Schiele (2018) distinguished off-manifold adversarial examples, which leave the data manifold, from on-manifold ones, which stay on it and behave like generalisation errors. A norm-bounded perturbation of a speed series is typically off-manifold: it introduces high-frequency jitter no road produces. A congestion-shaped perturbation is designed to be on-manifold. Training against off-manifold examples need not teach a model anything about on-manifold ones, and resisting on-manifold ones may conflict with predicting real congestion, which is the accuracy-robustness wall Zhang describes.

Taken together, changing the attack's shape, reach or mechanism with node selection held fixed moves defended error by factors of two or more in many reported cells. Changing node selection with shape held fixed moves attacked error by roughly 3 to 35 percent and defended error by a few to about ten percent. On the published evidence, the large lever is shape.

8. The metric decides what "defended" means

The third disagreement is about scoring. Global MAE averages over hundreds of nodes and twelve horizons. An attack that ruins five sensors' forecasts and leaves the rest alone barely moves it.

The attack literature saw this early. Liu, F., Liu, H. and Jiang (2022) report global and local versions of MAE and RMSE, and their Figure 1 notes errors at victim nodes and at neighbours separately. But defence results in both their paper and RDAT are reported in global MAE and RMSE only. Zhang's attack is designed around the gap: its figure reports the attack leaving 82 percent of the network untouched, and its case study describes a corridor whose forecast falls from a true 68 miles per hour to a predicted 31 under attack while the network-wide error barely changes. A defence that restores global MAE has not thereby restored the corridor that a routing service reads.

Stratified metrics expose a different hidden cost. Bai, S. et al. (2025) studied adversarial training for spatiotemporal prediction of crime and traffic accidents, where most observations are zero. They report that AT-TNDS, the dynamic-selection adversarial training of the 2022 paper, achieves strong average robustness but widens the gap between majority and minority classes; on the NYC data it raises their recall-disparity and average-precision-disparity measures by 3.7 and 2.1 percent, mainly through lost performance on the rare non-zero events. Averaged robustness hid a cost concentrated where the stakes are highest.

Relative and absolute metrics can disagree too. In Windmann et al. (2026), PGD training on the Traffic dataset lowered absolute error under faults while increasing degradation relative to clean error, although that increase is small and changes sign under the source's alternative selection rule. Which of those is "more robust" depends on whether the user cares about the error level or the size of the drop.

The metric also decides what counts as an attack's success. Untargeted global attacks are scored by how much average error rises; targeted local attacks by whether the forecast at a chosen place moves toward a chosen value. Liu, L. et al. (2023) score a target series that the attacker never touches directly, which is a different object again. A defence evaluated on one of these does not have a result on the others, and none of the papers reviewed here reports the same defence under all three.

9. How strong was the attacker in the evaluation?

A defence is only as robust as the strongest attack it has been tested against. The image-domain literature learned this expensively. Athalye, Carlini and Wagner (2018) showed that defences relying on obfuscated gradients gave a false sense of security. Carlini et al. (2019) set out guidelines for evaluating robustness, including adaptive attacks designed against the specific defence. Tramer et al. (2020) circumvented thirteen published defences whose authors had tried to run adaptive attacks. Croce and Hein (2020) applied a parameter-free attack ensemble to more than fifty published models and obtained lower robust accuracy than reported in all but one. Croce et al. (2020) built RobustBench to standardise such evaluations. In graphs, Mujkanovic et al. (2023) re-evaluated seven popular defences with adaptive attacks and found that most gave no or marginal improvement over an undefended model.

Graph adversarial training has its own pitfall. Gosch et al. (2023) showed that in the transductive node-classification setting most graph-robustness work had used, a defence can achieve perfect robustness trivially by memorising the clean training graph, and that adversarial training only becomes meaningful in an inductive setting, where it then performs as a state-of-the-art defence against structure perturbations. Traffic forecasting is inductive in time, since test windows are new, but the node set and graph are fixed across training and test. Whether a forecaster can exploit that fixity, for instance by learning per-node priors that shrink reliance on historically noisy sensors, is not examined in the papers reviewed here. RDAT's own conjecture about Graph WaveNet's learned adjacency is a hint in that direction.

Against this standard, the traffic defence evaluations are light. RDAT evaluates with five-step PGD at step size 0.1 under five node-selection rules, all from one family, and reports no adaptive attack against its policy-selected training and no attack designed for the distillation term. Liu, F., Liu, H. and Jiang (2022) use the same family for their defence table. Zhang, Q. (2026) does run a detector-aware adaptive attack against its own defence and reports that error rises monotonically with optimisation steps, which argues against gradient masking. It attacks RDAT white-box through the full pipeline, but designs no attack aimed at RDAT's selector or distillation term, though RDAT is a baseline there, not the defence under study.

Two findings suggest the PGD family may underestimate the threat. Anand and Pappas (2026) report model-free attacks outperforming small-step PGD by at least a third. Poudel and Li (2021) report that a substitute-model black-box attack degraded graph-based forecasters by up to 54 percent while linear regression and historical-average baselines were affected by less than 3 percent or not at all, which also illustrates that robustness and accuracy trade off across model classes. Zhu, L. et al. (2021) report that their black-box diffusion attack still produced adversarial samples against models trained with dropout-style regularisation; dropout weakened it in most cases, while DropNode and DropEdge did not protect.

None of this shows that the traffic defences are broken. It shows that their evaluations do not meet the standard the adversarial-robustness field adopted after 2018, and that the reported robustness is robustness against PGD with saliency or random node choice, at five steps, under a large box.

10. Distrust instead of tolerance: detection and certification

If adversarial training tolerates perturbations it was trained on and little else, the alternatives are to detect perturbations and to bound their effect.

Detection. Ilbert et al. (2023), on telecom traffic, combined a classifier that flags adversarial inputs, a denoiser and a standard forecaster. They report that this hybrid kept up to 92 percent of the original model's clean performance in mean squared error and had errors 2.71 and 2.51 times lower than adversarially trained comparisons on clean and perturbed data. Zhang, Q. (2026) proposes a physics-informed detector that scores each node by its inconsistency with directed neighbours and with smooth dynamics, and feeds the score to a frozen, adversarially trained forecaster as an extra input. Against AT-physics, the detector-augmented model is better in 13 of 15 cells, with its largest margins on the widened attack it was not trained against; in the case study it holds the corridor forecast at 60 miles per hour where AT-physics drops to 54. GNNGuard (Zhang, X. and Zitnik, 2020) applies a related idea, down-weighting edges between dissimilar nodes, to structure attacks on static graphs.

Detection has a known weakness. Carlini and Wagner (2017) defeated ten published detectors by building new loss functions, and concluded that adversarial examples are much harder to detect than had been thought. Zhang's adaptive attack against its own detector is a step toward meeting that standard; one model-dataset pair, Graph WaveNet on PeMS-BAY, is reported in detail.

Zhang's paper also contains an internal tension worth naming. It explains adversarial training's failure by saying that a congestion-shaped injection is a valid input state, so rejecting it would also reject real congestion. It then reports that its detector separates attacks from the lowest-speed quartile of clean inputs with an area under the curve of 0.996 to 1.0, flagging none of the real slowdowns. If the injections are that separable from real congestion, they are not valid input states in the sense the first argument needs, and the failure of adversarial training against them needs another explanation, such as the training distribution never containing them. Both statements may hold for different attack strengths; the paper does not reconcile them.

Certification. Randomized smoothing (Cohen, Rosenfeld and Kolter, 2019) gives provable robustness guarantees for classifiers. Yoon et al. (2022) extended it to probabilistic forecasting, defining robustness as bounded Wasserstein deviation of the forecast distribution, and Liu, L. et al. (2023) to multivariate forecasting. For graphs, Scholten et al. (2023) derived gray-box certificates against adversaries that control entire nodes and can set their features arbitrarily, exploiting the fact that messages from distant nodes are likely to be intercepted. That threat model, an attacker who owns some sensors outright, is closer to the sensor-spoofing setting than an l-infinity ball is, and the certificates strengthen with distance from the controlled nodes. No paper reviewed here applies node-control certificates to traffic forecasters.

Certification and detection share a property adversarial training lacks: their guarantees, or their signals, do not depend on having trained against the specific attack. That makes them natural candidates in a regime where the attack shape is unknown. They have their own costs, in clean accuracy for smoothing and in false alarms and adaptive evasion for detection, and the traffic evidence on both is thin.

11. What would make traffic genuinely different

The argument so far could be misread as saying that a traffic forecaster is just an image classifier with a different input. It is not, and the dynamic-domain thesis deserves its strongest form. There are at least five ways in which the setting really does differ, and each one makes a prediction that the published evidence can be held against.

Propagation through the graph. In an image classifier, a perturbed pixel affects the output through the network's receptive field. In a spatiotemporal graph

network, a perturbed sensor also affects every neighbour's forecast through message passing, and the effect compounds over layers. Liu, F., Miranda-Moreno and Sun (2021) showed that perturbing a single well-chosen vertex spreads error over a large part of the graph. Liu, F., Liu, H. and Jiang (2022) derived an upper bound on the change in the model's final-layer embedding under attack that grows linearly with the node budget, with the square of the perturbation budget, and with a power of the graph's maximum degree (together with weight and activation bounds) that doubles the number of layers. That bound says the graph matters to vulnerability. It predicts that architecture, through depth and the way adjacency is learned or fixed, should matter a great deal to how much any defence can recover. That is what RDAT's appendix and Zhang's tables show. It does not predict that the training-time node selector should matter much, and the evidence agrees.

The label is the future. A classifier's attacker knows the true class. A forecaster's attacker does not know the future readings it is trying to make the model mispredict, and Liu, F., Liu, H. and Jiang (2022) build a surrogate label from the model's own forecast for this reason. On the defence side, adversarial training uses the true future as its target. The asymmetry predicts that white-box evaluations, which hand the attacker the true future, overstate what a deployed attacker can do, and the 2022 paper's grey-box results, slightly weaker than its white-box ones (67.8 against 68.7 percent degradation on PeMS-BAY), are consistent with that, though the gap is small. It bears on how strong an evaluation is, not on whether node selection matters.

Streaming and persistence. A deployed forecaster runs continuously on a stream, and an attacker's choices are tied across time by the consistency constraint of Section 4. That predicts that a realistic attacker is less dynamic, in the S sense, than the benchmark one, and that defences tuned to a per-window resampling attacker may be trained against a harder node-selection problem than deployment poses, and an easier shape problem.

Temporal structure of the signal. Traffic series have strong daily and weekly periodicity and smooth dynamics, outside incidents. Perturbations that ignore that structure are off-manifold in the sense of Stutz, Hein and Schiele (2018), which predicts that they are easy to detect and that training against them teaches little about on-manifold attacks. The physics-aware attack and the detector built to catch it are the two sides of exactly this prediction (Zhang, Q., 2026). Zhang, J. et al. (2025) found, for foundation models, that decomposing perturbed series still gave trend and seasonal components highly correlated with the originals, while forecasts degraded substantially; small structured deviations, not gross distortions, did the damage.

Many outputs, many horizons. A traffic forecaster emits twelve horizons at every node. A defence can look good on the average and poor at the near horizon a navigation system uses, or at the far horizon a signal planner uses. Zhang, J. et al. (2025) report that input points closest to the forecast boundary are the most vulnerable, which suggests that horizon and position should be reported separately. No traffic defence paper does so.

Each of these differences is real. Each predicts something about attack strength, evaluation or metric. None predicts that a learned, time-varying choice of training nodes is the key to defence, and the published ablations bear that out. The strongest form of the dynamic-domain thesis is therefore a thesis about propagation and architecture, and that is where the evidence points: the residual vulnerability after adversarial training tracks the architecture far more closely than the training-time node selector.

It is worth being clear about what the graph structure does and does not license on the defence side. Message passing means a defended forecaster can, in principle, cross-check a node against its neighbours, which a pixel-wise image model cannot do in the same way. That is the opportunity physics-informed detection exploits, and the reason the classical loop diagnostics of Chen, C. et al. (2003) regress each loop on its neighbours. Adversarial training uses the graph only implicitly, through whatever the trained weights learn. A defence that uses the graph explicitly, as a source of redundancy, is a different kind of object from one that resamples training nodes, and the evidence reviewed here gives it the larger margins.

12. The measured values

Table 1 collects the published numbers the argument rests on. Each row names its source. Values are as reported, in the source's units; "gap closed" is this paper's arithmetic, defined in Section 1.

Figure 1 sets the two regimes side by side. In panel (a), the two defences sit close together and close to the clean value on Graph WaveNet, and close together but well above it on the other two architectures: architecture, not node selection, separates the outcomes. In panel (b), RDAT sits near the undefended value in most cells and on the far side of it in some, while training on the attack's own shape moves the error a long way. The contrast between the panels is the paper's argument in one picture: the dynamic component of a defence is a small lever under a matched threat, and the threat's shape is a large one.

13. Crediting a robustness gain to a defence's dynamic component

The re-reading above suggests a procedure. A claim that a defence's dynamic or learned node selection is what makes it robust should survive the following checks. Algorithm 1 states them as a decision structure.

```
Algorithm 1. Crediting a gain to node selection.
Input: defence D with learned/dynamic selector;
  baselines B_rand (same D, random nodes),
  B_fix (same D, fixed central nodes);
  threat tuples T = (S, Delta, J, K, M).
Step 1. Fix Delta, J, K, M. Vary only S at test time.
```

Table 1. Published robustness results for spatiotemporal and time-series forecasters, with the source on each row. Errors are MAE unless marked. Gap closed is computed by the author from the source's numbers.

Source	Model / data	Attack at test	Undefended clean → attacked	Defended (method)	Reading
Liu, F. et al. 2023 (RDAT), T1	GWNet / PeMS-BAY	PGD-TNDS, 20% nodes	1.78 → 7.82	2.20 (AT); 2.09 (RDAT)	93.1% vs 94.8% closed
Liu, F. et al. 2023, T2	GWNet / PeMS-D4	PGD-TNDS	22.42 → 106.16	28.10 (AT); 27.54 (RDAT)	93.2% vs 93.9% closed
Liu, F. et al. 2023, App. T3	ASTGCN / PeMS-D4	PGD-TNDS	27.28 → 120.17	69.60 (AT); 71.80 (RDAT)	54.4% vs 52.1%; AT ahead
Liu, F. et al. 2023, App. T4	STGCN / PeMS-D4	PGD-TNDS	27.87 → 86.09	50.63 (AT); 44.97 (RDAT)	60.9% vs 70.6%; std ∼10
Liu, F. et al. 2023, T1	GWNet / PeMS-BAY	PGD-TNDS, RMSE	3.76 → 13.48	4.38 (GraphAT); 4.45 (RDAT)	GraphAT lower
Liu, F. et al. 2022, T4	GWNet / PeMS-BAY	PGD-Random, 10%	2.03 → 6.15 (G-MAE)	2.54 (AT); 2.47 (AT-TDNS)	87.5% vs 89.3% closed
Liu, F. et al. 2022, T2	GWNet / PeMS-BAY	ablation of attack	full attack 6.13	-	random nodes -3.91%; vanilla PGD -33.41%
Zhang, Q. 2026, T2	DCRNN / PeMS-D4	norm-bounded, 5-node target	6.2 → 152.7 (target)	-	corridor error x25
Zhang, Q. 2026, T3	GWNet / PeMS-BAY	node choice, 5 nodes	-	-	random 7.0 > saliency 6.3 > degree 5.2
Zhang, Q. 2026, T4	GWNet / PeMS-BAY	physics-aware	41.1 (target, attacked)	39.7 (RDAT); 14.3 (AT-phys)	norm AT barely helps
Zhang, Q. 2026, T4	GWNet / PeMS-D4	physics-aware	61.5	77.2 (RDAT); 36.5 (AT-phys)	norm AT worse than none
Zhang, Q. 2026, T4	STTN / PeMS-D4	physics-aware	36.9	13.6 (RDAT); 17.5 (AT-phys)	norm AT transfers here
Zhang, Q. 2026, T5	STGCN / PeMS-BAY	halo 1 → 3 hops	-	13.7 → 34.8 (AT-phys)	trained reach does not hold
Zhang, J. et al. 2025, T4	TimesFM / Loop Seattle	PGD (NMAE)	0.113 → 1.213	0.132 (in-domain LAT)	98% closed
Zhang, J. et al. 2025, T4	TimesFM / Hier. Sales	SimBA query (NMAE)	0.927 → 1.817	2.371 (in-domain LAT)	AT worse than none
Liu, L. et al. 2023, T1	DeepVAR / Traffic	sparse, kappa 9 (wQL)	0.228 → 0.464	0.197 (mini-max, trained kappa 5)	holds beyond training sparsity
Bai, S. et al. 2025	GSNet / NYC accidents	STPGD	-	AT-TNDS	recall gap +3.7%, AP gap +2.1%
Windmann et al. 2026	5 archs / Traffic	sensor faults	-	PGD AT	abs. error down, degradation up
Poudel and Li 2021	GCGRNN, DCRNN / traffic	black-box substitute	up to 54% worse	-	linear models <3% or immune

```
Step 2. Train D, B_rand, B_fix with matched budget,
  epochs, steps and model; >= 5 seeds each.
Step 3. For each test S: report clean and attacked
  error, and gap closed per method.
Step 4. If D - B_rand < seed std: credit nothing to
  learned selection; report stability only.
Step 5. If B_fix << B_rand: credit "training must
  resample nodes", not "dynamic domain".
Step 6. Repeat 1-5 on >= 3 architectures. If the
  sign of D - B_rand flips: no general claim.
Step 7. Change Delta (shape, reach) and J (target)
  with S fixed. If gains vanish: the defence
  is threat-specific; say which threat.
Step 8. Score with global, target-set and
  stratified M; report each.
Step 9. Attack D adaptively (selector- and
  regulariser-aware); report the strongest.
Output: the narrowest claim that survives.
```

Step 1 isolates the variable the dynamic-domain thesis is about. Steps 2 and 3 make the comparison fair and expressed in a unit that can be compared across datasets. Step 4 is the test RDAT's Graph WaveNet results fail on PeMS-D4 under every attack and on PeMS-BAY under the time-dependent attack, where the learned selector's improvement over random selection is within run-to-run spread; they pass it, narrowly, under the other four PeMS-BAY attacks. Step 5 is the test RDAT's Figure 2a passes, and it licenses a real but narrow conclusion. Step 6 is the test the appendix fails: under the time-dependent attack the ordering flips on ASTGCN, and the sign of the difference also varies across attack columns. Step 7 is Zhang's question. Step 8 is the metric question of Section 8. Step 9 is the evaluation standard of Section 9.

Applied to the published record, the procedure credits the following. Resampling training nodes rather than fixing them on central nodes helps (RDAT, Figure 2a). Learned selection improves on random selection on the leading architecture by an amount within run-to-run variation, except under four of the five PeMS-BAY attacks, where it slightly exceeds AT's seed spread, and by an amount of inconsistent sign on others. Norm-bounded adversarial training, dynamic or not, is specific to its threat shape, with partial and architecture-dependent transfer to targeted, coherent attacks (Zhang, Q., 2026). None of the defences has been evaluated to the adaptive standard of the image literature.

14. The strongest case against this paper

The gap-closure measure may flatter the baselines. Gap closure compresses absolute differences; a 0.1 MAE difference on PeMS-BAY is about 5 percent of the clean error. One could argue that, for a deployed system, five percent of residual error is worth a learned selector. The reply is that the difference is within the reported seed variation and does not survive the architecture change in the appendix. But the point stands that this paper's arithmetic is one way of reading the table, and RDAT's own percentage improvements over its baselines are another.

The learned selector may matter more under wider attacks. RDAT's sweep of test attack strength (Section 5) reports a 14.0 percent margin on PeMS-BAY when every node is attacked, larger than its margin at the 20 percent default. This paper's comparison uses the default strength, where the tables give exact values. If the margin at full coverage holds up under matched seeds and against plain AT specifically, the dynamic component earns a real but specific credit: better transfer along the node-budget axis. That would still be a claim about S, and still leave the shape and target results of Section 7 untouched.

RDAT's claim may be about stability, and stability matters. RDAT reports lower standard deviation than several baselines and attributes it to self-distillation. A defence that is reliably good may be worth more than one that is better on average and erratic. This paper does not dispute that; it disputes that the evidence shows static methods failing to defend.

The attack-specificity evidence rests on one preprint. Zhang, Q. (2026) is a single-author arXiv preprint, not yet peer reviewed, and its reproduction of RDAT may differ from the original. Its strongest wording (nearly useless) is softer in its own table. Its physics-aware attack has free weights that the attacker tunes. If its RDAT numbers are wrong, Section 7 loses its sharpest example. The general finding that adversarial training is threat-specific does not depend on it: the image-domain results (Tramer and Boneh, 2019; Kaufmann et al., 2019; Dai, S. et al., 2023), the black-box results for forecasting foundation models (Zhang, J. et al., 2025) and the fault results (Windmann et al., 2026) all point the same way.

The benchmark threat may be the right worst case. One can defend the large l-infinity box on the grounds that a defence robust to large untargeted perturbations of many sensors is robust to anything smaller. The physics-aware attack is a counterexample to that inference, since it is not contained in the box; it is unbounded in magnitude and coherent in shape. A box can be large and still not contain the threat.

This paper's reconciliation could be read as a truism. That adversarial training works against the threat it was trained on and less well against others is familiar from images. What is not familiar, and what the traffic literature's main defence paper asserted the opposite of, is that in traffic the time-varying choice of nodes is the small part of the threat, while the shape and target of the perturbation are the large part. That inversion of emphasis is the paper's claim.

15. What is not known

No study has compared random, saliency-based and learned node selection for adversarial training across several architectures with matched training budgets and enough seeds to resolve differences of a few tenths of MAE. RDAT's comparison is closest and is underpowered for the size of effect it reports.

No study has evaluated a traffic forecasting defence with an adaptive attack designed against that defence's specific components, in the sense of Tramer et al. (2020).

No study reports the same defence under global, target-corridor and stratified metrics at once, so the trade-offs among them for a single defence are unmeasured.

The relation between sensor physics and the perturbation set is unmeasured. What a compromised loop detector, a spoofed probe-vehicle feed or a tampered roadside unit can actually report, and how that maps onto Delta, is argued in Zhang, Q. (2026) by reference to sensor-security work rather than measured on traffic hardware.

The forecasting label problem is unaddressed: when an attacked sensor's future readings are also corrupted, what a robust forecaster should predict at that sensor, and how to score it, has not been set out.

Node-control certificates of the kind in Scholten et al. (2023) have not been applied to spatiotemporal forecasters, so whether certified robustness at useful radii is attainable for traffic is unknown.

The architecture dependence that dominates RDAT's appendix and Zhang's tables has no tested explanation. Whether learned adjacency, recurrence or attention is responsible is a conjecture in both papers.

Whether existing operational data-quality pipelines, such as the PeMS loop diagnostics (Chen, C. et al., 2003), would already catch the benchmark attacks before a forecaster saw them has not been tested.

16. Conclusion

The traffic-forecasting robustness literature has been arguing about whether adversarial training transfers, and the three published positions turn out to have measured different things. The dynamic-domain thesis is about which sensors an attacker touches over time; in its own main tables, adversarial training with randomly resampled nodes closes more than nine-tenths of the attack's damage on its leading architecture, and against the time-dependent attack the learned selector's increment is within run-to-run variation; under the other attacks it stays within that variation on PeMS-D4 and slightly exceeds it on PeMS-BAY, at 3.3 to 3.8 points of gap. The sufficiency thesis is right about matched threats. The attack-specificity thesis is right that a change of shape, reach or target defeats defences tuned

to a norm ball, though transfer is partial rather than absent. Across the attack literature's own ablations, node selection moves attacked error by roughly 3 to 35 percent and defended error by a few to about ten percent, while perturbation shape is a lever of multiples. Algorithm 1 states how a future claim about a defence's dynamic component should be tested. What a traffic forecaster needs to be robust against is still an open question about sensors and deployment, and adversarial training alone cannot answer it.

References

Anand, S. C., & Pappas, G. J. (2026). Adversarial Robustness of Deep State Space Models for Forecasting. arXiv:2604.03427. doi:10.48550/arXiv.2604.03427.

Apruzzese, G., Anderson, H. S., Dambra, S., Freeman, D., Pierazzi, F., & Roundy, K. A. (2022). "Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice. arXiv:2212.14315. doi:10.48550/arXiv.2212.14315.

Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. ICML 2018. arXiv:1802.00420. doi:10.48550/arXiv.1802.00420.

Bai, S., Ji, Y., Liu, Y., Zhang, X., Zheng, X., & Zeng, D. D. (2025). Alleviating Performance Disparity in Adversarial Spatiotemporal Graph Learning Under Zero-Inflated Distribution. arXiv:2504.00721. doi:10.48550/arXiv.2504.00721.

Bai, T., Luo, J., Zhao, J., Wen, B., & Wang, Q. (2021). Recent Advances in Adversarial Training for Adversarial Robustness. arXiv:2102.01356. doi:10.48550/arXiv.2102.01356.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., & Kurakin, A. (2019). On Evaluating Adversarial Robustness. arXiv:1902.06705. doi:10.48550/arXiv.1902.06705.

Carlini, N., & Wagner, D. (2017). Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. AISec 2017. arXiv:1705.07263. doi:10.48550/arXiv.1705.07263.

Chen, C., Kwon, J., Rice, J., Skabardonis, A., & Varaiya, P. (2003). Detecting Errors and Imputing Missing Data for Single-Loop Surveillance Systems. Transportation Research Record, 1855, 160-167. doi:10.3141/1855-20.

Cohen, J. M., Rosenfeld, E., & Kolter, J. Z. (2019). Certified Adversarial Robustness via Randomized Smoothing. ICML 2019. arXiv:1902.02918. doi:10.48550/arXiv.1902.02918.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., & Hein, M. (2020). RobustBench: a standardized adversarial robustness benchmark. arXiv:2010.09670. doi:10.48550/arXiv.2010.09670.

Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. ICML 2020. arXiv:2003.01690. doi:10.48550/arXiv.2003.01690.

Dai, H., Li, H., Tian, T., Huang, X., Wang, L., Zhu, J., & Song, L. (2018). Adversarial Attack on Graph Structured Data. ICML 2018. arXiv:1806.02371. doi:10.48550/arXiv.1806.02371.

Dai, S., Mahloujifar, S., Xiang, C., Sehwag, V., Chen, P.-Y., & Mittal, P. (2023). MultiRobustBench: Benchmarking Robustness Against Multiple Attacks. ICML 2023. arXiv:2302.10980. doi:10.48550/arXiv.2302.10980.

Dang-Nhu, R., Singh, G., Bielik, P., & Vechev, M. (2020). Adversarial Attacks on Probabilistic Autoregressive Forecasting Models. ICML 2020. arXiv:2003.03778. doi:10.48550/arXiv.2003.03778.

Dong, Y., Xu, K., Yang, X., Pang, T., Deng, Z., Su, H., & Zhu, J. (2021). Exploring Memorization in Adversarial Training. ICLR 2022. arXiv:2106.01606. doi:10.48550/arXiv.2106.01606.

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Adversarial Attacks on Deep Neural Networks for Time Series Classification. IJCNN 2019. arXiv:1903.07054. doi:10.48550/arXiv.1903.07054.

Feng, F., He, X., Tang, J., & Chua, T.-S. (2019). Graph Adversarial Training: Dynamically Regularizing Based on Graph Structure. IEEE TKDE. arXiv:1902.08226. doi:10.48550/arXiv.1902.08226.

Feng, Y., Huang, S., Chen, Q. A., Liu, H. X., & Mao, Z. M. (2018). Vulnerability of Traffic Control System Under Cyberattacks with Falsified Data. Transportation Research Record, 2672. doi:10.1177/0361198118756885.

Foo Kune, D., Backes, J., Clark, S. S., Kramer, D., Reynolds, M., Fu, K., Kim, Y., & Xu, W. (2013). Ghost Talk: Mitigating EMI Signal Injection Attacks against Analog Sensors. IEEE Symposium on Security and Privacy 2013. doi:10.1109/SP.2013.20.

Gilmer, J., Adams, R. P., Goodfellow, I., Andersen, D., & Dahl, G. E. (2018). Motivating the Rules of the Game for Adversarial Example Research. arXiv:1807.06732. doi:10.48550/arXiv.1807.06732.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and Harnessing Adversarial Examples. ICLR 2015. arXiv:1412.6572. doi:10.48550/arXiv.1412.6572.

Gosch, L., Geisler, S., Sturm, D., Charpentier, B., Zugner, D., & Gunnemann, S. (2023). Adversarial Training for Graph Neural Networks: Pitfalls, Solutions, and New Directions. NeurIPS 2023. arXiv:2306.15427. doi:10.48550/arXiv.2306.15427.

Huang, S. E., Wong, W., Feng, Y., Chen, Q. A., Mao, Z. M., & Liu, H. X. (2020). Impact Evaluation of Falsified Data Attacks on Connected Vehicle Based Traffic Signal Control. arXiv:2010.04753. doi:10.48550/arXiv.2010.04753.

Ilbert, R., Hoang, T. V., Zhang, Z., & Palpanas, T. (2023). Breaking Boundaries: Balancing Performance and Robustness in Deep Wireless Traffic Forecasting. arXiv:2311.09790. doi:10.48550/arXiv.2311.09790.

Jin, W., Li, Y., Xu, H., Wang, Y., Ji, S., Aggarwal, C., & Tang, J. (2020). Adversarial Attacks and Defenses on Graphs: A Review, A Tool and Empirical Studies. arXiv:2003.00653. doi:10.48550/arXiv.2003.00653.

Karim, F., Majumdar, S., & Darabi, H. (2019). Adversarial Attacks on Time Series. arXiv:1902.10755. doi:10.48550/arXiv.1902.10755.

Kaufmann, M., Kang, D., Sun, Y., Basart, S., Yin, X., Mazeika, M., Arora, A., Dziedzic, A., et al. (2019). Testing Robustness Against Unforeseen Adversaries. arXiv:1908.08016. doi:10.48550/arXiv.1908.08016.

Krishan, P., Mohapatra, R., Das, S., & Sengupta, S. (2024). Adversarial Attacks and Defenses in Multivariate Time-Series Forecasting for Smart and Connected Infrastructures. arXiv:2408.14875. doi:10.48550/arXiv.2408.14875.

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. arXiv:1607.02533. doi:10.48550/arXiv.1607.02533.

Laidlaw, C., Singla, S., & Feizi, S. (2020). Perceptual Adversarial Robustness: Defense Against Unseen Threat Models. ICLR 2021. arXiv:2006.12655. doi:10.48550/arXiv.2006.12655.

Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2017). Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. ICLR 2018. arXiv:1707.01926. doi:10.48550/arXiv.1707.01926.

Liu, F., Jiang, S., Miranda-Moreno, L., Choi, S., & Sun, L. (2024). Adversarial Vulnerabilities in Large Language Models for Time Series Forecasting. arXiv:2412.08099. doi:10.48550/arXiv.2412.08099.

Liu, F., Liu, H., & Jiang, W. (2022). Practical Adversarial Attacks on Spatiotemporal Traffic Forecasting Models. NeurIPS 2022. arXiv:2210.02447. doi:10.48550/arXiv.2210.02447.

Liu, F., Miranda-Moreno, L., & Sun, L. (2021). Spatially Focused Attack against Spatiotemporal Graph Neural Networks. arXiv:2109.04608. doi:10.48550/arXiv.2109.04608.

Liu, F., Zhang, W., & Liu, H. (2023). Robust Spatiotemporal Traffic Forecasting with Reinforced Dynamic Adversarial Training. KDD 2023. arXiv:2306.14126. doi:10.48550/arXiv.2306.14126.

Liu, L., Park, Y., Hoang, T. N., Hasson, H., & Huan, J. (2023). Robust Multivariate Time-Series Forecasting: Adversarial Attacks and Defense Mechanisms. ICLR 2023. arXiv:2207.09572. doi:10.48550/arXiv.2207.09572.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards Deep Learning Models Resistant to Adversarial Attacks. ICLR 2018. arXiv:1706.06083. doi:10.48550/arXiv.1706.06083.

Mode, G. R., & Hoque, K. A. (2020). Adversarial Examples in Deep Learning for Multivariate Time Series Regression. arXiv:2009.11911. doi:10.48550/arXiv.2009.11911.

Mujkanovic, F., Geisler, S., Gunnemann, S., & Bojchevski, A. (2023). Are Defenses for Graph Neural Networks Robust? NeurIPS 2022. arXiv:2301.13694. doi:10.48550/arXiv.2301.13694.

Poudel, B., & Li, W. (2021). Black-box Adversarial Attacks on Network-wide Multi-step Traffic State Prediction Models. arXiv:2110.08712. doi:10.48550/arXiv.2110.08712.

Rice, L., Wong, E., & Kolter, J. Z. (2020). Overfitting in adversarially robust deep learning. ICML 2020. arXiv:2002.11569. doi:10.48550/arXiv.2002.11569.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., & Madry, A. (2018). Adversarially Robust Generalization Requires More Data. NeurIPS 2018. arXiv:1804.11285. doi:10.48550/arXiv.1804.11285.

Scholten, Y., Schuchardt, J., Geisler, S., Bojchevski, A., & Gunnemann, S. (2023). Randomized Message-Interception Smoothing: Gray-box Certificates for Graph Neural Networks. NeurIPS 2022. arXiv:2301.02039. doi:10.48550/arXiv.2301.02039.

Siddique, A., Kundu, R. K., Mode, G. R., & Hoque, K. A. (2023). RobustPdM: Designing Robust Predictive Maintenance against Adversarial Attacks. arXiv:2301.10822. doi:10.48550/arXiv.2301.10822.

Stutz, D., Hein, M., & Schiele, B. (2018). Disentangling Adversarial Robustness and Generalization. CVPR 2019. arXiv:1812.00740. doi:10.48550/arXiv.1812.00740.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. ICLR 2014. arXiv:1312.6199. doi:10.48550/arXiv.1312.6199.

Tramer, F., & Boneh, D. (2019). Adversarial Training and Robustness for Multiple Perturbations. NeurIPS 2019. arXiv:1904.13000. doi:10.48550/arXiv.1904.13000.

Tramer, F., Carlini, N., Brendel, W., & Madry, A. (2020). On Adaptive Attacks to Adversarial Example Defenses. NeurIPS 2020. arXiv:2002.08347. doi:10.48550/arXiv.2002.08347.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2018). Robustness May Be at Odds with Accuracy. ICLR 2019. arXiv:1805.12152. doi:10.48550/arXiv.1805.12152.

Windmann, A., Wittenberg, P., Manca, G., Dix, M., Brandt, J. U., & Niggemann, O. (2026). Benchmarking Sensor-Fault Robustness in Forecasting. arXiv:2605.10822. doi:10.48550/arXiv.2605.10822.

Wong, E., Rice, L., & Kolter, J. Z. (2020). Fast is better than free: Revisiting adversarial training. ICLR 2020. arXiv:2001.03994. doi:10.48550/arXiv.2001.03994.

Wu, Z., Pan, S., Long, G., Jiang, J., & Zhang, C. (2019). Graph WaveNet for Deep Spatial-Temporal Graph Modeling. IJCAI 2019. arXiv:1906.00121. doi:10.48550/arXiv.1906.00121.

Xu, K., Chen, H., Liu, S., Chen, P.-Y., Weng, T.-W., Hong, M., & Lin, X. (2019). Topology Attack and Defense for Graph Neural Networks: An Optimization Perspective. IJCAI 2019. arXiv:1906.04214. doi:10.48550/arXiv.1906.04214.

Yoon, T., Park, Y., Ryu, E. K., & Wang, Y. (2022). Robust Probabilistic Time Series Forecasting. AISTATS 2022. arXiv:2202.11910. doi:10.48550/arXiv.2202.11910.

Yu, B., Yin, H., & Zhu, Z. (2017). Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. IJCAI 2018. arXiv:1709.04875. doi:10.48550/arXiv.1709.04875.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., El Ghaoui, L., & Jordan, M. I. (2019). Theoretically Principled Trade-off between Robustness and Accuracy. ICML 2019. arXiv:1901.08573. doi:10.48550/arXiv.1901.08573.

Zhang, J., Zhang, Z., Zheng, S., Wen, X., Li, J., & Bian, J. (2025). Are Time-Series Foundation Models Deployment-Ready? A Systematic Study of Adversarial Robustness Across Domains. arXiv:2505.19397. doi:10.48550/arXiv.2505.19397.

Zhang, Q. (2026). Revisiting the Adversarial Robustness of Graph-Based Traffic Forecasting. arXiv:2607.27604. doi:10.48550/arXiv.2607.27604.

Zhang, X., & Zitnik, M. (2020). GNNGuard: Defending Graph Neural Networks against Adversarial Attacks. NeurIPS 2020. arXiv:2006.08149. doi:10.48550/arXiv.2006.08149.

Zhu, L., Feng, K., Pu, Z., & Ma, W. (2021). Adversarial Diffusion Attacks on Graph-based Traffic Prediction Models. arXiv:2104.09369. doi:10.48550/arXiv.2104.09369.

Zugner, D., Akbarnejad, A., & Gunnemann, S. (2018). Adversarial Attacks on Neural Networks for Graph Data. KDD 2018. arXiv:1805.07984. doi:10.48550/arXiv.1805.07984.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.24430423