# STRUCTURE-GROUNDED TEST-TIME ADAPTATION OF TABULAR FOUNDATION MODELS FOR POWER OUTAGE PREDICTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Weather-induced power outages disrupt critical infrastructure, making accurate forecasting essential for proactive response. Existing methods are typically trained once on heterogeneous historical weather events, limiting their ability to adapt to rare high-impact events. Tabular foundation models such as TabPFN instead leverage in-context labeled examples at inference time, enabling event-specific prediction without retraining. However, their native inference does not explicitly capture event-specific spatial dependencies among neighboring locations, which may vary substantially across events, particularly during extreme events. To resolve this limitation, we propose Structure-Grounded Test-Time Adaptation (SG-TTA), which uses spatial structure as test-time feedback to adapt a lightweight module while keeping the tabular in-context learner frozen. For each incoming event, SG-TTA retrieves relevant locations in historical events and their neighborhoods to augment in-context support and provide cross-event spatial references for alignment, while maintaining within-event consistency via neighbor reconstruction. Experiments across four real-world service territories show that SG-TTA consistently improves both TabPFN and TabICL, with the TabPFN variant achieving the lowest MAE in every territory. Further analysis in Connecticut (CT) confirms performance gains across outage-severity levels, especially on high-impact events.

## 1 INTRODUCTION

Weather-induced power outages can cause substantial economic losses and disrupt critical infrastructure and essential services (Flores et al., 2023; Udeh et al., 2024). Accurate outage prediction is therefore important for proactive crew allocation, resource planning and emergency response. A key challenge of weather-induced power outage prediction is the substantial cross-event heterogeneity of weather events. Historical events vary widely in severity and meteorological conditions, and models trained on mixed events often overestimate frequent low-impact cases while underestimating rare high-impact extremes (Yang et al., 2020). Existing approaches typically put different types of historical events together to train a unified predictor. While models such as XGBoost (Hughes et al., 2024) can capture nonlinear covariate relationships and GNNs (Shen et al., 2026) can model spatial dependencies, they are still trained on heterogeneous historical distributions and do not naturally adapt to each incoming event. This limitation is particularly challenging for rare, high-impact events, which are underrepresented in historical data and can differ substantially from the more common low-impact events seen during training.

Tabular foundation models (Hollmann et al., 2025; Qu et al., 2025; 2026) offer a natural alternative through in-context learning. Pretrained across large collections of tabular tasks, models such as TabPFN and TabICL provide a strong general predictive prior and leverage labeled examples provided at inference time to adapt their predictions. This allows the historical evidence used for prediction to be tailored to each target event without retraining the pretrained backbone. Recent extensions further show that retrieving or partitioning locally relevant context can improve tabular in-context learning (Thomas et al., 2024; Xu et al., 2025). This makes retrieval-augmented tabular foundation models well suited to outage prediction, combining a strong pretrained prior with event-specific inference from localized historical context.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Retrieval tailors the historical evidence used for each prediction, but the pretrained backbone itself remains fixed and reflects a broad prior learned across diverse tabular tasks. This motivates adapting the model, rather than only its context, to the incoming outage event. Naively fine-tuning on pooled historical data is ineffective, however, because it does not account for event-specific distribution shifts. We therefore pursue event-specific adaptation, where a lightweight module is specialized to each incoming event while keeping the pretrained backbone frozen.

The key challenge is how to guide event-specific adaptation without outage labels for the incoming event. Spatial structure provides a natural source of label-free feedback, i.e., nearby locations within a service territory often exhibit correlated outage responses, providing a within-event consistency signal. However, topology alone is insufficient, as the same neighborhood can respond differently under different weather conditions. We therefore leverage similar historical events as complementary cross-event references to characterize the spatial relationships most relevant to the current event.

To this end, we propose **S**tructure-**G**rounded **T**est-**T**ime **A**daptation (**SG-TTA**) for power outage prediction, which leverages spatial structure as test-time feedback to specialize a tabular foundation model to each incoming weather event. SG-TTA retrieves locally relevant historical subgraphs, using them both as in-context support and as cross-event references for spatial relationships grounding. In the absence of target-event labels, it combines cross-event relational alignment with within-event consistency via neighbor reconstruction over the target graph. These complementary signals will guide the adaptation toward spatial patterns relevant to the current event while allowing dependencies among target locations to influence prediction. Adaptation is restricted to a lightweight event-specific module, with the pretrained backbone kept frozen and the adaptation state reset after each incoming event.

In summary, our contributions are as follows:

- We study weather-induced power outage prediction through the lens of tabular foundation models and show how in-context inference enables event-specific prediction under substantial cross-event heterogeneity, providing an alternative to conventional outage prediction models trained on heterogeneous historical distributions and cannot naturally adapt to each incoming event.
- We propose **SG-TTA**, a structure-grounded test-time adaptation framework that specializes a frozen tabular foundation model to each incoming weather event without target-event outage labels. SG-TTA combines cross-event relational grounding from retrieved historical neighborhoods with within-event spatial consistency to provide complementary adaptation signals for a lightweight event-specific module.
- Extensive experiments across four real-world service territories and two tabular foundation models demonstrate consistent improvements from SG-TTA. The TabPFN-based variant achieves the lowest MAE across all four territories, while severity-stratified analysis in Connecticut (CT) shows substantial improvement for high-impact events.

## 2 Related Work

**Power Outage Prediction.** Weather-induced outage prediction has been studied using statistical and machine-learning models that combine meteorological, vegetation and infrastructure information. Early work developed regression-based and Random Forest models for hurricane and storm outage estimation (Han et al., 2009; Guikema et al., 2010; Nateghi et al., 2014), while later studies incorporated richer features to improve outage forecasting across different storm types (Taylor et al., 2022). Recent work has further explored severity-aware modeling and hybrid machine learning approaches to better capture rare or high-impact events (Yang et al., 2020; Hughes et al., 2024). Beyond independently modeling spatial units, graph-based methods explicitly encode geographical relationships among grid locations. For example, GCN-based outage models incorporate circuit-level spatial connectivity and vegetation risk (Azizi et al., 2025), while SA-HGNN models spatial interactions among service locations together with event-specific weather and static infrastructure features (Shen et al., 2026). Despite these advances, existing approaches remain predominantly task-specific supervised models trained over historical events, whereas our work investigates retrieval-augmented tabular foundation models and event-specific test-time adaptation.

**Tabular Foundation Models.** Prior-data Fitted Networks (PFNs) (Müller et al., 2022) are meta-trained on synthetic datasets to approximate Bayesian posterior predictive inference in a single forward pass. Building on this idea, TabPFN (Hollmann et al., 2023; 2025) is pretrained on large-scale synthetic tasks and performs prediction via in-context learning. Related in-context learners such as TabICL (Qu et al., 2025; 2026), Mitra (Zhang et al., 2026; Tao et al., 2026) and TabDPT (Ma et al., 2026) further improve scalability and pretraining strategies. To adapt these models to specific downstream data, prior work has explored retrieval-based local context construction (Thomas et al., 2024; Xu et al., 2025), fine-tuning of the backbone (Rubachev et al., 2025), and prompt tuning with learnable context tokens (Feuer et al., 2024). Existing methods specialize tabular foundation models through localized context or parameter adaptation. In contrast, SG-TTA uses the spatial organization of each target event to guide event-specific test-time adaptation of a frozen backbone.

**Test-Time Adaptation and Structural Feedback.** Test-time adaptation (TTA) adjusts models using information available at inference, through objectives such as self-supervised test-time training (Sun et al., 2020; Gandelsman et al., 2022) and entropy minimization (Wang et al., 2020). Retrieved examples can also support test-time updates (Hardt & Sun, 2024). More closely related to tabular in-context learning, Gozeten et al. (2025) update TabPFN using labeled in-context demonstrations. Structural feedback has been studied in graph adaptation and test-time prompt tuning (Hsu et al., 2025; Jiang et al., 2026). SG-TTA instead uses structural feedback specific to each incoming weather event: retrieved historical neighborhoods provide relational references, while the unlabeled target graph provides a within-event consistency signal.

# 3 METHODOLOGY

## 3.1 PROBLEM STATEMENT AND OVERVIEW

**Problem Statement.** Consider a service territory with $N$ locations $V = \{1, \dots, N\}$ connected by a static spatial graph $G = (V, A)$, where the adjacency matrix $A$ is shared across weather events and encodes the spatial connectivity among locations. For each location $i$, $\mathcal{N}(i)$ denotes its set of spatial neighbors. Each weather event is represented by a covariate matrix $X \in \mathbb{R}^{N \times d}$ and an outage vector $Y \in \mathbb{R}^N$. The $i$-th row $x_i$ contains the meteorological, vegetation, and infrastructure features associated with location $i$, while $y_i$ denotes the corresponding outage count. Given the covariates $X^\star$ of an unseen target event, our goal is to predict its outage vector $\hat{Y}^\star$.

We formulate this task under the in-context learning paradigm of a pretrained tabular foundation model. Given a labeled context constructed from historical events $\mathcal{C}^\star = \{(x_j, y_j)\}_{j=1}^M$ and a target query $x_i^\star$, the frozen network $f_\theta$ parameterizes the posterior predictive distribution $p_\theta(y_i^\star \mid x_i^\star, \mathcal{C}^\star)$. The corresponding point prediction is obtained from this distribution, e.g., $\hat{y}_i^\star = \mathbb{E}_{p_\theta}[y_i^\star \mid x_i^\star, \mathcal{C}^\star]$.

**Overview.** In this section, we introduce the proposed SG-TTA framework for power outage prediction, as shown in Figure 1. First, *location-anchored subgraph retrieval* (Section 3.2) selects, for each target location, historical events whose local neighborhoods are most similar in a task-aligned representation space; the retrieved labeled subgraphs form the in-context support $\mathcal{C}^\star$ and serve as historical relational references. Second, *structure-grounded test-time adaptation* (Section 3.3) updates a lightweight module $\psi$ using two label-free signals, cross-event relational grounding and within-event self-consistency, while keeping the backbone parameters $\theta$ fixed. The adapted model then predicts the outage vector $\hat{Y}^\star$, and $\psi$ is reset before the next event.

## 3.2 LOCATION-ANCHORED SUBGRAPH RETRIEVAL

Given an incoming event, we first construct a localized historical context for the frozen in-context learner. Selecting local contexts through nearest neighbors in feature space has been shown to improve the scalability and expressivity of tabular in-context learning (Thomas et al., 2024; Xu et al., 2025). However, proximity in the raw covariate space is task-agnostic: not all features are equally predictive of outage response, so covariate similarity may not reflect true predictive relevance. We therefore learn a lightweight task-aligned encoder $g_\phi$ on the historical training events and represent each location by $r_i = g_\phi(x_i)/\|g_\phi(x_i)\|_2$. The encoder is used only for retrieval, while the original covariates are retained as inputs to the frozen foundation model. Details of the encoder and its training objective are provided in Appendix D.

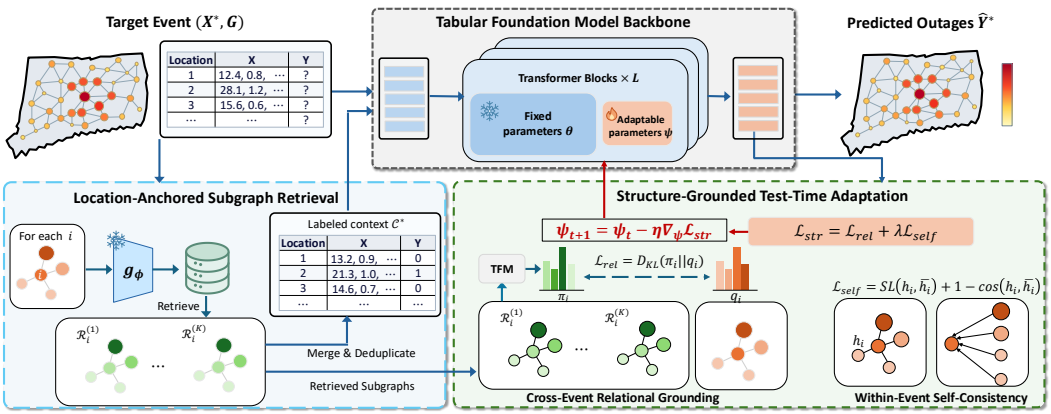



Figure 1: Overview of SG-TTA framework for power outage prediction.

Since the outage at a location can also depend on its surrounding spatial environment, we compare not only the anchor itself but also its local neighborhood. For location $i$, let $\mathcal{V}_i = \{i\} \cup \mathcal{N}_\ell(i)$ denote its $\ell$-hop neighborhood. To preserve spatial correspondence, retrieval is restricted to the same location across historical events. We then rank historical training events by the mean task-aligned similarity over the aligned neighborhood,

$$s_i^{(m)} = \frac{1}{|\mathcal{V}_i|} \sum_{v \in \mathcal{V}_i} \cos\left(\mathbf{r}_v^\star, \mathbf{r}_v^{(m)}\right), \tag{1}$$

where the small radius $\ell$ keeps the comparison localized around the target location. We select the top-$K$ historical events and retain their corresponding labeled local subgraphs. The resulting subgraphs form $\mathcal{R}_i^\star = \{\mathcal{R}_i^{(k)}\}_{k=1}^K$. Across all target locations, their deduplicated labeled observations constitute the in-context support $\mathcal{C}^\star$, while the retained local outage patterns provide historical relational references for subsequent test-time adaptation.

### 3.3 Structure-Grounded Test-time Adaptation

The retrieved context provides relevant historical observations, but the standard tabular inference does not explicitly use the spatial relationships among locations in the target event. We therefore adapt a small set of parameters $\psi$ to each incoming event while keeping all other pretrained parameters $\theta$ fixed. The event-specific parameters are obtained from the target event and retrieved relational evidence:

$$\psi^\star = \arg\min_\psi \mathcal{L}_{\text{str}}\left(\psi; X^\star, G, \mathcal{R}^\star\right). \tag{2}$$

The adapted predictor is thus parameterized by $(\theta, \psi^\star)$, with structural evidence affecting prediction only through the shared lightweight state.

Alternative ways of injecting structure either leave the model state unchanged (e.g., neighborhood-concatenate input features (Eremeev et al., 2025)), require an additional graph operator trained on historical events (graph adapters) or freely reshape target representations (direct hidden-state refinement), which risks suppressing task-relevant variation. In contrast, SG-TTA updates only lightweight adaptation parameters. Letting $H(\psi)$ denote the target representations, an update of $\psi$ induces, to first order,

$$\Delta H \approx -\eta\, J_\psi J_\psi^\top \nabla_H \mathcal{L}_{\text{str}}, \qquad J_\psi = \frac{\partial\, \text{vec}(H)}{\partial \psi}, \tag{3}$$

where $\eta$ is the step size and $J_\psi$ is the Jacobian of the target representations with respect to the adaptation parameters. The factor $J_\psi J_\psi^\top$ maps the structural gradient to a representation change within the column space of $J_\psi$, so the update follows directions induced by the shared adaptation parameters rather than directly optimizing hidden representations.

We then instantiate $\mathcal{L}_{\text{str}}$ with two complementary signals: cross-event relational grounding from retrieved historical neighborhoods and within-event self-consistency on the target graph. The former

anchors adaptation to relational patterns observed in retrieved relevant subgraphs, whereas the latter constrains the target representation field using its own topology structure.

**Cross-Event Relational Grounding.** Spatial dependencies among neighboring locations provide informative structure for outage prediction (Shen et al., 2026), while relational knowledge transfer has shown that inter-sample relations can convey structural information beyond pointwise representations (Park et al., 2019; Guo et al., 2023; Yang et al., 2022). Accordingly, the retrieved subgraphs serve not only as labeled examples for in-context prediction, but also as historical references for how outage responses are locally organized under similar conditions.

To make historical and target relations comparable, we replay the retrieved historical covariates through the frozen backbone as queries, using the same labeled context $\mathcal{C}^\star$ as for the target event. Historical and target relations are thus constructed from query representations extracted at the same layer under a shared context. For the $k$-th retrieved subgraph $\mathcal{R}_i^{(k)}$, let $\mathbf{h}_i^{(k)}$ and $\mathbf{h}_j^{(k)}$ denote the resulting representations of the anchor $i$ and its neighbor $j \in \mathcal{N}(i)$, respectively. We measure the historical affinity between the anchor and each of its neighbors by averaging their cosine similarity over the $K$ retrieved subgraphs,

$$a_{ij} = \frac{1}{K} \sum_{k=1}^{K} \cos\left(\mathbf{h}_i^{(k)}, \mathbf{h}_j^{(k)}\right), \quad j \in \mathcal{N}(i). \tag{4}$$

The resulting affinity profile defines a frozen reference distribution,

$$\pi_i(j) = \frac{\exp(a_{ij}/\tau)}{\sum_{u \in \mathcal{N}(i)} \exp(a_{iu}/\tau)}. \tag{5}$$

Let $\mathbf{h}_i(\psi)$ denote the target representation under the current lightweight state $\psi$. The target event induces the corresponding relation distribution

$$q_i(j;\psi) = \frac{\exp\left(\cos(\mathbf{h}_i(\psi), \mathbf{h}_j(\psi))/\tau\right)}{\sum_{u \in \mathcal{N}(i)} \exp\left(\cos(\mathbf{h}_i(\psi), \mathbf{h}_u(\psi))/\tau\right)}. \tag{6}$$

We align the target relational geometry with the retrieved historical reference by minimizing

$$\mathcal{L}_{\mathrm{rel}}(\psi) = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} D_{\mathrm{KL}}\left(\pi_i \,\|\, q_i(\psi)\right), \tag{7}$$

where $D_{\mathrm{KL}}$ denotes the Kullback–Leibler divergence that encourages the target neighborhood relation distribution $q_i(\psi)$ to match the fixed historical reference $\pi_i$.

**Within-Event Self-Consistency.** Cross-event relational grounding specifies how neighboring locations should be organized relative to one another based on historically relevant events, but does not explicitly enforce local coherence within the incoming event itself. We therefore introduce a complementary within-event objective that encourages each target representation to remain consistent with its spatial neighborhood. For location $i$, we reconstruct its representation by averaging those of its neighbors,

$$\hat{\mathbf{h}}_i(\psi) = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \mathbf{h}_j(\psi). \tag{8}$$

We treat this reference as a detached target using the stop-gradient operator $\mathrm{sg}(\cdot)$ and define

$$\mathcal{L}_{\mathrm{self}}(\psi) = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \left[\mathrm{SL}\Big(\mathbf{h}_i(\psi), \mathrm{sg}(\hat{\mathbf{h}}_i(\psi))\Big) + 1 - \cos\Big(\mathbf{h}_i(\psi), \mathrm{sg}(\hat{\mathbf{h}}_i(\psi))\Big)\right], \tag{9}$$

where SL is the Smooth-L1 loss. The Smooth-L1 term penalizes coordinate-wise differences, while the cosine term encourages directional agreement within each neighborhood.

**Test-Time Adaptation Procedure.** The pretrained backbone $f_\theta$ remains frozen throughout the entire procedure. We combine the two structural signals as

$$\mathcal{L}_{\mathrm{str}}(\psi) = \mathcal{L}_{\mathrm{rel}}(\psi) + \lambda \mathcal{L}_{\mathrm{self}}(\psi), \tag{10}$$

where $\lambda$ controls the contribution of within-event self-consistency. Starting from a fixed initialization $\psi_0$, we perform $T$ gradient steps:

$$\psi_{t+1} = \psi_t - \eta \nabla_\psi \mathcal{L}_{\text{str}}(\psi_t), \qquad t = 0, \ldots, T-1, \tag{11}$$

where $\eta$ is the adaptation learning rate. Predictions are obtained using the adapted state $\psi^\star = \psi_T$. Each target event is specialized independently from the shared initialization $\psi_0$. After prediction, the event-specific state $\psi^\star$ is discarded, and adaptation for the next event restarts from $\psi_0$. This ensures that the adapted state reflects only the structural evidence of the current event.

# 4 Experiments

## 4.1 Experimental Setup

**Datasets.** To evaluate the effectiveness of the proposed SG-TTA to predict power outages, we compare SG-TTA with baseline methods over four utility service territories of Eversource, i.e., Connecticut (CT), Western Massachusetts (WMA), Eastern Massachusetts (EMA) and New Hampshire (NH). The detailed statistics and descriptions are provided in Appendix A.

**Baselines.** We compare SG-TTA with eight baselines in three groups: (1) Traditional machine learning methods: Random Forest (Breiman, 2001), XGBoost (Chen & Guestrin, 2016); (2) GNN-based methods: GraphSAGE (Hamilton et al., 2017), SA-HGNN (Shen et al., 2026); (3) Tabular foundation models: TabICL v2 (Qu et al., 2026), TabPFN v3 (Grinsztajn et al., 2026), LoCalPFN (Thomas et al., 2024), G2T-FM (Eremeev et al., 2025). Baseline details are given in Appendix B.

**Evaluation Metrics.** We evaluate the performance of outage predictions by using absolute error (AE), mean absolute error (MAE), absolute percentage error (APE), mean absolute percentage error (MAPE), centered root mean square error (CRMSE) and coefficient of determination ($R^2$). Detailed metric definitions are provided in Appendix C.

**Implementation Details.** We implement SG-TTA on TabPFN v3 and TabICL v2. All vanilla backbones use 50K randomly sampled training rows as context for a controlled comparison. We inject rank-16 LoRA adapters (Hu et al., 2022) into the query, key, value, and output projections of the last four transformer layers (approximately 230K parameters, or 0.39% of the backbone). For each test event, the adapters are optimized using a learning rate of 1e-2 for 5 steps. We set $\lambda = 1$, $\tau = 0.1$, and retrieve $K = 3$ historical subgraphs per target location using one-hop neighborhoods ($\ell = 1$). Evaluation uses five-fold cross-validation with event-disjoint splits.

## 4.2 Main Results

Table 1 shows that SG-TTA consistently improves outage prediction across all four service territories. TabPFN + SG-TTA achieves the lowest MAE on every dataset, while TabICL + SG-TTA attains the second-lowest MAE in all four cases. For example, on CT, TabPFN + SG-TTA achieves the best or tied-best results on seven of the eight metrics, reducing MAE and CRMSE by 9.6% and 3.3% over the corresponding second-best results. Beyond these absolute comparisons, SG-TTA also improves both vanilla backbones in every region. Specifically, it reduces the MAE of TabPFN by 15.8%, 11.4%, 11.5%, and 14.0%, and that of TabICL by 5.1%, 9.5%, 0.1%, and 8.1% on CT, NH, WMA, and EMA, respectively. It further outperforms LoCalPFN and G2T-FM across all four regions, suggesting that the gains extend beyond existing context-localization and structure-aware adaptation strategies. Overall, these results show that the benefit of structure-grounded test-time adaptation extends across different service territories and tabular foundation model backbones.

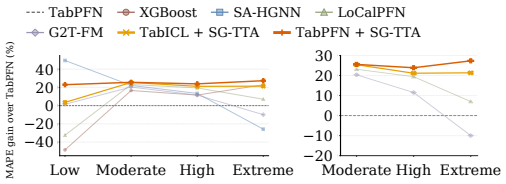



Figure 2: Relative MAPE gain over vanilla TabPFN across outage-severity groups, with a zoomed view of the Moderate–Extreme regimes.

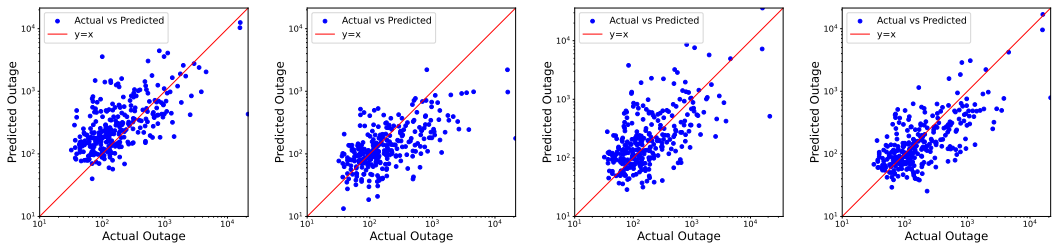
**Grouped Comparison.** Following prior outage-prediction studies that characterize event severity by outage magnitude for operational planning (Yang et al., 2020; Walsh et al., 2018), we stratify CT

Table 1: Outage prediction results across four utility service territories. Best and second-best results are highlighted in **bold** and underlined, respectively.

| Dataset | Method | AE Q25 | AE Q50 | MAE | APE Q25 | APE Q50 | MAPE | CRMSE | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|
| **Connecticut** | Random Forest | 83.00 | 162.00 | 412.99 | 39.50 | 100.00 | 175.54 | 1443 | 0.39 |
| | XGBoost | 55.25 | 120.00 | 399.56 | 31.54 | 69.36 | 126.36 | 1425 | 0.42 |
| | GraphSAGE | 31.25 | 77.00 | 392.54 | 29.61 | 58.82 | 69.31 | 1776 | 0.08 |
| | SA-HGNN | 25.10 | 67.44 | 391.51 | 22.59 | 46.58 | **61.33** | 1756 | 0.07 |
| | TabICL | 45.27 | 91.12 | 361.02 | 26.99 | 60.05 | 115.41 | 1424 | 0.40 |
| | + SG-TTA | 37.46 | 68.34 | 342.73 | 24.63 | 48.93 | 86.56 | 1376 | 0.48 |
| | TabPFN | 31.13 | 80.36 | 367.82 | 24.86 | 52.81 | 98.62 | 1452 | 0.38 |
| | + LoCalPFN | 28.71 | 73.67 | 368.09 | 23.77 | 48.39 | 112.90 | 1460 | 0.37 |
| | + G2T-FM | 23.90 | 73.16 | 372.22 | 22.63 | 47.10 | 89.49 | 1554 | 0.29 |
| | + SG-TTA | **22.87** | **65.03** | **309.84** | **21.30** | **45.04** | 66.78 | **1330** | **0.48** |
| **New Hampshire** | Random Forest | 41.00 | 90.00 | 149.72 | 33.77 | 98.62 | 161.42 | 302 | 0.45 |
| | XGBoost | 24.50 | 51.00 | 123.93 | 29.96 | 55.56 | 108.71 | 284 | 0.48 |
| | GraphSAGE | 17.00 | 39.00 | 115.41 | 29.36 | 53.94 | 61.17 | 300 | 0.38 |
| | SA-HGNN | 14.88 | 37.61 | 121.84 | 24.50 | 49.99 | 63.42 | 368 | 0.32 |
| | TabICL | 21.09 | 48.94 | 113.58 | 24.67 | 52.34 | 94.53 | 262 | 0.52 |
| | + SG-TTA | 15.61 | 39.30 | 102.83 | 22.28 | 46.71 | 79.49 | 251 | **0.56** |
| | TabPFN | 16.24 | 41.02 | 114.98 | 24.66 | 50.58 | 75.91 | 280 | 0.45 |
| | + LoCalPFN | 17.40 | 39.95 | 118.94 | 26.27 | 49.32 | 81.74 | 286 | 0.43 |
| | + G2T-FM | 17.32 | 39.06 | 104.50 | 25.32 | 47.53 | 72.62 | 256 | 0.54 |
| | + SG-TTA | **14.38** | **32.85** | **101.89** | **22.19** | **45.40** | **53.30** | **246** | 0.53 |
| **Western Massachusetts** | Random Forest | 11.00 | 26.00 | 50.18 | 38.05 | 100.00 | 173.07 | 116 | -0.05 |
| | XGBoost | 8.00 | 21.00 | 50.60 | 28.02 | 79.43 | 167.74 | 128 | -0.02 |
| | GraphSAGE | 6.00 | 15.00 | 40.16 | 36.24 | 64.86 | 66.72 | 115 | 0.01 |
| | SA-HGNN | 5.34 | 14.46 | 37.61 | 26.83 | 56.92 | 74.81 | 105 | 0.05 |
| | TabICL | 7.77 | 15.22 | 37.07 | 26.99 | 56.71 | 105.80 | 100 | 0.14 |
| | + SG-TTA | 7.19 | 15.58 | 37.02 | 24.90 | 52.20 | 90.16 | **100** | 0.13 |
| | TabPFN | 4.95 | 13.41 | 39.87 | 30.75 | 50.67 | 97.08 | 102 | 0.12 |
| | + LoCalPFN | 5.03 | 12.88 | 38.13 | 25.30 | 49.39 | 74.83 | 109 | 0.08 |
| | + G2T-FM | 5.08 | 11.79 | 37.52 | 24.69 | 49.87 | 71.95 | 111 | 0.03 |
| | + SG-TTA | **4.52** | **11.41** | **35.30** | **24.17** | **47.67** | **61.23** | 108 | **0.14** |
| **Eastern Massachusetts** | Random Forest | 27.50 | 58.00 | 153.01 | 36.99 | 84.00 | 146.45 | 514 | 0.18 |
| | XGBoost | 20.00 | 45.00 | 165.78 | 29.96 | 67.01 | 164.06 | 506 | 0.22 |
| | GraphSAGE | 16.00 | 36.00 | 143.61 | 32.99 | 59.57 | 76.80 | 537 | 0.13 |
| | SA-HGNN | **10.35** | 26.09 | 137.40 | 22.58 | 47.84 | **59.49** | 542 | 0.08 |
| | TabICL | 14.90 | 33.22 | 144.33 | 25.30 | 49.88 | 121.37 | 508 | 0.18 |
| | + SG-TTA | 14.92 | 31.38 | 132.60 | **20.22** | 45.25 | 98.02 | **487** | **0.25** |
| | TabPFN | 12.63 | 28.81 | 152.85 | 23.53 | 44.32 | 115.13 | 514 | 0.20 |
| | + LoCalPFN | 10.60 | 26.28 | 145.77 | 22.19 | 42.57 | 102.20 | 519 | 0.18 |
| | + G2T-FM | 10.55 | 28.21 | 145.02 | 22.89 | 42.95 | 83.88 | 498 | 0.22 |
| | + SG-TTA | 10.78 | **23.41** | **131.38** | 21.59 | **38.71** | 66.20 | 500 | 0.23 |

events by their total outage counts into Low ($< 300$), Moderate (300–749), High (750–4999), and Extreme ($\geq 5000$) regimes. Figure 2 further examines performance across outage-severity groups using the relative MAPE improvement over vanilla TabPFN. SG-TTA improves upon TabPFN in all four regimes, with gains of 23.0%, 25.6%, 23.9%, and 27.3% from Low to Extreme outages, respectively. The enlarged higher-severity view further shows that SG-TTA maintains the highest gain among the compared methods from Moderate through Extreme outages, while several alternative adaptation and supervised spatial methods exhibit smaller or even negative gains as severity increases. In particular, the improvement reaches 27.3% in the Extreme (high impact) group, suggesting that structure-grounded test-time feedback remains effective even for rare, high-impact outage patterns. The complete grouped results are provided in Appendix E.

**Visual Comparison.** Figure 3 plots predicted outages against actual event-level outages on the Connecticut dataset, where each point corresponds to one event and both axes are on a logarithmic scale. We compare against one representative baseline from each category. The baselines exhibit distinct failure modes. XGBoost systematically overestimates low-impact events, whereas SA-HGNN tracks them closely but flattens its response at the upper end. This regression toward the mean is consistent with a single predictor dominated by abundant low-impact observations. Vanilla TabPFN partially alleviates this bias but its predictions remain widely dispersed, with large overestimates in the intermediate range. SG-TTA yields the tightest concentration around the diagonal across the full range: the most severe events are predicted close to their actual magnitudes, and the overestimates of TabPFN are substantially reduced. These observations indicate that structure-grounded specialization improves event-level outage estimates beyond context conditioning alone.




(a) XGBoost (b) SA-HGNN (c) TabPFN (d) SG-TTA

Figure 3: Actual outages versus predicted outages on the CT dataset.

Table 3: Ablation study of structural adaptation on CT dataset. All variants use the same TabPFN backbone and retrieved context $\mathcal{C}^\star$ unless otherwise specified.

| Variant | AE Q50 | MAE | APE Q50 | MAPE | CRMSE | $R^2$ |
|---|---|---|---|---|---|---|
| Frozen TabPFN | 74.47 | 336.63 | 48.29 | 84.51 | 1395 | 0.43 |
| Supervised fine-tuning | 79.28 | 357.24 | 51.38 | 97.22 | **1321** | 0.45 |
| Supervised fine-tuning on context $\mathcal{C}^\star$ | 76.43 | 329.28 | 46.45 | 90.66 | 1346 | 0.46 |
| *(a) Structural Intervention Strategy* | | | | | | |
| Neighbor-feature concatenation | 78.82 | 345.16 | 45.68 | 84.83 | 1447 | 0.38 |
| Direct hidden refinement | 72.46 | 326.34 | 45.39 | 78.48 | 1399 | 0.42 |
| Trained graph adapter | 69.98 | 365.43 | **44.84** | 89.65 | 1501 | 0.37 |
| SG-TTA (lightweight state adaptation) | **65.03** | **309.84** | 45.04 | **66.78** | 1330 | **0.48** |
| *(b) Structural feedback objective* | | | | | | |
| $\mathcal{L}_{\text{rel}}$, shuffled reference | 77.24 | 334.52 | 50.01 | 89.94 | 1352 | 0.43 |
| $\mathcal{L}_{\text{rel}}$ only | 69.36 | 317.07 | 47.61 | 76.18 | 1328 | **0.48** |
| $\mathcal{L}_{\text{self}}$ only | 67.77 | 317.39 | 45.11 | 75.46 | 1338 | 0.46 |
| SG-TTA ($\mathcal{L}_{\text{rel}} + \lambda\mathcal{L}_{\text{self}}$) | **65.03** | **309.84** | 45.04 | **66.78** | 1330 | **0.48** |

### 4.3 Ablation Study

Table 2 isolates the effect of context construction on CT dataset. Using the full history performs substantially worse than a fixed-budget context, consistent with cross-event heterogeneity that makes broadly mixed historical observations less relevant to the incoming event. Task-aligned retrieval improves over covariate $k$NN particularly in MAE and MAPE. Our location-anchored subgraph retrieval further achieves the best performance across all four metrics, reducing MAE and CRMSE by 3.8% and 2.9% over pointwise task-aligned $k$NN, respectively. These results support retrieving locally matched historical environments for context construction.

Table 2: Ablation study of context construction on CT dataset.

| Context | MAE | MAPE | CRMSE | $R^2$ |
|---|---|---|---|---|
| Full history | 564.33 | 157.76 | 2472 | −0.78 |
| Random | 367.82 | 98.62 | 1452 | 0.38 |
| Covariate $k$NN | 369.03 | 114.77 | 1433 | 0.39 |
| Task-aligned $k$NN | 349.83 | 86.95 | 1437 | 0.37 |
| **Ours** | **336.63** | **84.51** | **1395** | **0.43** |

Table 3 examines the key design choices of structure-grounded test-time adaptation. Task-level supervised fine-tuning degrades performance, indicating that jointly adapting over heterogeneous historical events remains affected by cross-event heterogeneity. Restricting fine-tuning to the retrieved context improves MAE to 329.28, confirming the benefit of localization, but remains substantially weaker than SG-TTA. Alternative ways of injecting spatial structure, such as neighbor-feature concatenation, direct hidden-state refinement and a trained graph adapter, provide limited or inconsistent gains. In contrast, adapting a lightweight module with structural feedback yields a stronger overall improvement, supporting our choice to use structure as test-time feedback rather than directly modifying the inference representation. For the feedback objectives, $\mathcal{L}_{\text{rel}}$ and $\mathcal{L}_{\text{self}}$ are individually effective and further improve performance when combined, indicating complementary roles of cross-event relational grounding and within-event self-consistency. Importantly, shuffling the retrieved relational reference largely removes the benefit of $\mathcal{L}_{\text{rel}}$, showing that its gain comes from structurally relevant historical evidence rather than generic relational regularization.

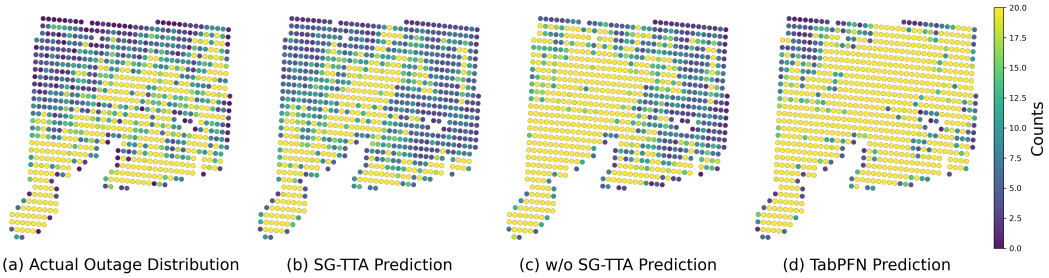



Figure 4: Outage prediction distribution of Hurricane Irene (August 28, 2011) in Connecticut.

### 4.4 In-depth Analysis

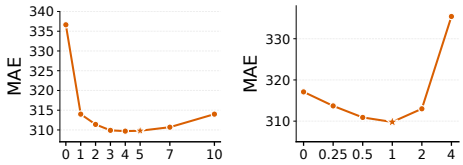
**Case Study.** Figure 4 compares the spatial outage distribution for a representative CT event. Vanilla TabPFN and the model without SG-TTA tend to produce broader high-outage regions and weaker local contrast, whereas SG-TTA more closely follows the observed spatial heterogeneity, including localized high-impact clusters and neighboring low-outage areas. The improved spatial agreement suggests that structure-grounded test-time feedback helps tailor the frozen foundation model to the event-specific spatial outage distributions than merely shifting predictions globally.

**Task Alignment of Structural Feedback.** Table 4 examines whether the structural update direction is associated with improvement in the outage prediction task. Both relational grounding and within-event self-consistency show positive associations with the observed MAE improvement. Their combination yields the highest inner-product correlation and sign agreement. The association also persists when alignment is measured by cosine similarity, which removes the effect of gradient scale from the alignment measure. These results provide empirical support for jointly using the two signals together.

Table 4: Task alignment of structural feedback. $\rho_{\text{ip}}$ and $\rho_{\cos}$ correlate MAE improvement with inner-product and cosine gradient alignment, respectively. Sign is the fraction with $\langle g_{\text{sup}}, g \rangle > 0$.

| Gradient Alignment | $\rho_{\text{ip}}$ | Sign | $\rho_{\cos}$ |
|---|---|---|---|
| $\langle g_{\text{sup}}, g_{\text{rel}} \rangle$ | 0.766 | 73.1% | 0.602 |
| $\langle g_{\text{sup}}, g_{\text{self}} \rangle$ | 0.766 | 70.7% | 0.665 |
| $\langle g_{\text{sup}}, g_{\text{str}} \rangle$ | **0.793** | **73.5%** | **0.699** |

**Hyperparameter Sensitivity.** Figure 5 examines the sensitivity of SG-TTA to the number of test-time adaptation steps $T$ and the structural loss weight $\lambda$ on CT dataset. Increasing $T$ initially reduces MAE, with performance remaining stable for $T = 3$ to 5 before deteriorating with further optimization. Similarly, MAE decreases as $\lambda$ increases to 1, but rises noticeably for larger values. Both trends indicate that structural feedback is most useful when applied with a limited update budget and a balanced contribution from the two objectives.




(a) Adaptation steps $T$ (b) $\lambda$

Figure 5: Sensitivity to adaptation steps and structural loss weight on CT.

## 5 Conclusion

In this work, we introduced SG-TTA, a structure-grounded test-time adaptation framework for weather-induced power outage prediction. SG-TTA retrieves locally matched historical subgraphs and combines cross-event relational grounding with within-event self-consistency to enable event-specific adaptation without target-event labels, while keeping the pretrained backbone frozen. Across four service territories, SG-TTA consistently improves both TabPFN and TabICL, with the TabPFN variant achieving the lowest MAE in every territory. Further analyses on CT show gains across outage-severity levels and spatial locations, while ablations confirm the complementary roles of the two structural objectives. Overall, our results demonstrate that spatial structure can provide effective test-time feedback for adapting tabular foundation models to heterogeneous weather events.

## AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing parts of the text, retrieving and discovering related work, and writing and refining parts of the software code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REPRODUCIBILITY STATEMENT

Section 3 specifies the retrieval procedure, structural objectives, and test-time adaptation protocol. Section 4.1 describes the backbones, event-disjoint data splits, and adaptation hyperparameters. Appendices A–F provide dataset descriptions, baseline details, retrieval encoder objectives, evaluation metric definitions, additional experimental results, and runtime measurements.

## REFERENCES

Morteza Azizi, Xinxuan Zhang, Feifei Yang, Kingsley Udeh, Junbo Zhao, and Emmanouil Anagnostou. Advancing outage prediction modeling: Incorporating circuit spatial relationships with graph neural networks and lidar-derived tree risk. *Reliability Engineering & System Safety*, pp. 111579, 2025.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Dmitry Eremeev, Gleb Bazhenov, Oleg Platonov, Artem Babenko, and Liudmila Prokhorenkova. Turning tabular foundation models into graph foundation models. *arXiv preprint arXiv:2508.20906*, 2025.

Benjamin Feuer, Robin T Schirrmeister, Valeriia Cherepanova, Chinmay Hegde, Frank Hutter, Micah Goldblum, Niv Cohen, and Colin White. Tunetables: Context optimization for scalable prior-data fitted networks. *Advances in Neural Information Processing Systems*, 37:83430–83464, 2024.

Nina M Flores, Heather McBrien, Vivian Do, Mathew V Kiang, Jeffrey Schlegelmilch, and Joan A Casey. The 2021 texas power crisis: distribution, duration, and disparities. *Journal of exposure science & environmental epidemiology*, 33(1):21–31, 2023.

Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems*, 35:29374–29385, 2022.

Halil Alperen Gozeten, M Emrullah Ildiz, Xuechen Zhang, Mahdi Soltanolkotabi, Marco Mondelli, and Samet Oymak. Test-time training provably improves transformers as in-context learners. *Proceedings of machine learning research*, 267:20266, 2025.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, et al. Tabpfn-3: Technical report. *arXiv preprint arXiv:2605.13986*, 2026.

Seth D Guikema, Steven M Quiring, and Seung-Ryong Han. Prestorm estimation of hurricane damage to electric power distribution systems. *Risk Analysis: An International Journal*, 30(12): 1744–1752, 2010.

Zhichun Guo, William Shiao, Shichang Zhang, Yozen Liu, Nitesh V Chawla, Neil Shah, and Tong Zhao. Linkless link prediction via relational distillation. In *International conference on machine learning*, pp. 12012–12033. PMLR, 2023.

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

Seung-Ryong Han, Seth D Guikema, and Steven M Quiring. Improving the predictive accuracy of hurricane power outage forecasts using generalized additive models. *Risk Analysis: An International Journal*, 29(10):1443–1453, 2009.

Moritz Hardt and Yu Sun. Test-time training on nearest neighbors for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 54625–54640, 2024.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=cp5PvcI6w8_.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Hans Hao-Hsun Hsu, Shikun Liu, Han Zhao, and Pan Li. Structural alignment improves graph test-time adaptation. *arXiv preprint arXiv:2502.18334*, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

William Hughes, Sita Nyame, William Taylor, Aaron Spaulding, Mingguo Hong, Xiaochuan Luo, Slava Maslennikov, Diego Cerrai, Emmanouil Anagnostou, and Wei Zhang. A probabilistic method for integrating physics-based and data-driven storm outage prediction models for power systems. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 10(2):04024021, 2024.

Yan Jiang, Ruihong Qiu, and Zi Huang. Gfmate: Empowering graph foundation models with test-time prompt tuning. *arXiv preprint arXiv:2605.14809*, 2026.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Jesse Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L Caterini, and Maks Volkovs. Tabdpt: Scaling tabular foundation models on real data. *Advances in Neural Information Processing Systems*, 38:172692–172722, 2026.

Samuel Müller, Noah Hollmann, Sebastian Pineda-Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Roshanak Nateghi, Seth Guikema, and Steven M Quiring. Power outage estimation for tropical cyclones: Improved accuracy with simpler models. *Risk analysis*, 34(6):1069–1078, 2014.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3962–3971. IEEE, 2019.

Jordan G Powers, Joseph B Klemp, William C Skamarock, Christopher A Davis, Jimy Dudhia, David O Gill, Janice L Coen, David J Gochis, Ravan Ahmadov, Steven E Peckham, et al. The weather research and forecasting model: Overview, system efforts, and future directions. *Bulletin of the American Meteorological Society*, 98(8):1717–1737, 2017.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabicl: A tabular foundation model for in-context learning on large data. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/qu25d.html.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabiclv2: A better, faster, scalable, and open tabular foundation model. *CoRR*, abs/2602.11139, 2026. doi: 10.48550/ARXIV.2602.11139. URL https://doi.org/10.48550/arXiv.2602.11139.

Ivan Rubachev, Akim Kotelnikov, Nikolay Kartashev, and Artem Babenko. On finetuning tabular foundation models. *arXiv preprint arXiv:2506.08982*, 2025.

Xuyang Shen, Zijie Pan, Diego Cerrai, Xinxuan Zhang, Christopher Colorio, Emmanouil Anagnostou, and Dongjin Song. Empowering power outage prediction with spatially aware hybrid graph neural networks and contrastive learning. *Transactions on Machine Learning Research*, 2026.

William C Skamarock, Joseph B Klemp, Jimy Dudhia, David O Gill, Dale M Barker, Michael G Duda, Xiang-Yu Huang, Wei Wang, Jordan G Powers, et al. A description of the advanced research wrf version 3. *NCAR technical note*, 475(125):10–5065, 2008.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Yefan Tao, Xiyuan Zhang, Xinyi Liu, Boran Han, Danielle Maddix, Haoyang Fang, Zhen Han, Jiading Gai, Xuanqing Liu, Michael Bohlke-Schneider, et al. Mitra-v2 technical report. *arXiv preprint arXiv:2609.04540*, 2026.

William O Taylor, Peter L Watson, Diego Cerrai, and Emmanouil N Anagnostou. Dynamic modeling of the effects of vegetation management on weather-related power outages. *Electric Power Systems Research*, 207:107840, 2022.

Valentin Thomas, Junwei Ma, Rasa Hosseinzadeh, Keyvan Golestan, Guangwei Yu, Maksims Volkovs, and Anthony Caterini. Retrieval & fine-tuning for in-context tabular models. *Advances in Neural Information Processing Systems*, 37:108439–108467, 2024.

Kingsley Udeh, David W Wanik, Diego Cerrai, Emmanouil N Anagnostou, and Derek Aguiar. Probabilistic storm and electric utility customer outage prediction. *IEEE Access*, 12:126285–126295, 2024.

Tara Walsh, Thomas Layton, David Wanik, and Jonathan Mellor. Agent based model to estimate time to restoration of storm-induced power outages. *Infrastructures*, 3(3):33, 2018.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Derek Xu, Olcay Cirit, Reza Asadi, Yizhou Sun, and Wei Wang. Mixture of in-context prompters for tabular pfns. In *International Conference on Learning Representations*, volume 2025, pp. 24630–24665, 2025.

Chenxiao Yang, Qitian Wu, and Junchi Yan. Geometric knowledge distillation: Topology compression for graph neural networks. *Advances in Neural Information Processing Systems*, 35: 29761–29775, 2022.

Feifei Yang, David W Wanik, Diego Cerrai, Md Abul Ehsan Bhuiyan, and Emmanouil N Anagnostou. Quantifying uncertainty in machine learning-based power outage prediction model training: A tool for sustainable storm restoration. *Sustainability*, 12(4):1525, 2020.

Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression. *Advances in Neural Information Processing Systems*, 36:17882–17903, 2023.

Xiyuan Zhang, Danielle Maddix Robinson, Junming Yin, Nick Erickson, Abdul Fatir Ansari, Boran Han, Shuai Zhang, Leman Akoglu, Christos Faloutsos, Michael Mahoney, et al. Mitra: Mixed synthetic priors for enhancing tabular foundation models. *Advances in neural information processing systems*, 38:15795–15840, 2026.

# A Dataset Details

We compare the proposed SG-TTA with baseline methods over four utility service territories of Eversource, i.e., Connecticut, Western Massachusetts, Eastern Massachusetts and New Hampshire. The detailed statistics of benchmark datasets are summarized in Table 5.

Each dataset integrates meteorological conditions, utility infrastructure, land-cover characteristics, vegetation information, and historical outage records for individual weather events. These variables are spatially aligned over a fixed set of locations within each service territory to characterize weather-induced power disruptions. The main data sources are described below.

**Weather.** Meteorological variables characterize the event-specific weather conditions associated with each location. Weather fields are obtained from 48-hour simulations using version 3.8.1 of the Advanced Research Weather Research and Forecasting (WRF) model (Powers et al., 2017; Skamarock et al., 2008), with a horizontal grid spacing of 4 km over the northeastern United States. The simulations are initialized and forced at the lateral boundaries using the North American Mesoscale (NAM) Forecast System at a spatial resolution of 12 km.

**Utility Infrastructure.** Infrastructure data from Eversource Energy include multiple classes of distribution assets, such as electric fuses, reclosers, and poles. These variables characterize spatial variation in the amount and composition of exposed utility infrastructure, which can influence outage susceptibility under severe weather.

**Land Cover and Vegetation.** Land-cover information is derived from the U.S. Geological Survey (USGS) National Land Cover Database (NLCD), which characterizes vegetation and urbanization patterns. We include vegetation- and development-related variables, such as the proportions of deciduous forest, other forested land, and developed areas, to capture location-specific environmental conditions relevant to weather-induced outages.

Table 5: The detailed statistics of four datasets.

| Datasets | # Events | # Locations | Feature Length | Output Length |
|---|---|---|---|---|
| Connecticut | 294 | 815 | 390 | 1 |
| New Hampshire | 227 | 1022 | 390 | 1 |
| Western Massachusetts | 271 | 312 | 390 | 1 |
| Eastern Massachusetts | 231 | 383 | 390 | 1 |

# B Baseline Details

- Random Forest (Breiman, 2001): An ensemble of decision trees trained on bootstrapped samples, serving as a robust conventional baseline for tabular regression.
- XGBoost (Chen & Guestrin, 2016): A gradient-boosted decision tree method that provides a strong baseline for nonlinear relationships in structured tabular data.
- GraphSAGE (Hamilton et al., 2017): An inductive graph neural network that learns node representations by aggregating feature information from local neighborhoods.
- SA-HGNN (Shen et al., 2026): A supervised spatial graph model that combines hybrid GNN encoders with contrastive learning to capture outage dependencies across geographically connected locations.
- TabICL (Qu et al., 2026): A scalable transformer-based tabular foundation model that performs prediction through in-context learning over labeled training examples and unlabeled queries.
- TabPFN (Hollmann et al., 2025; Grinsztajn et al., 2026): A prior-data fitted tabular foundation model that performs fast in-context Bayesian-style prediction by conditioning on labeled context examples at inference time.
- LoCalPFN (Thomas et al., 2024): A retrieval-augmented TabPFN variant that localizes inference to nearest-neighbor subsets of the training data and further adapts the model through local fine-tuning.

- G2T-FM (Eremeev et al., 2025): A graph-to-table framework that augments each node with neighborhood-aggregated features and structural encodings, enabling tabular foundation models to operate on graph-structured data.

## C Evaluation Metrics

We evaluate event-level outage predictions by aggregating location-level counts. For each of the $E$ test events, let $Y_m = \sum_{i=1}^{N} y_i^{(m)}$ and $\hat{Y}_m = \sum_{i=1}^{N} \hat{y}_i^{(m)}$ denote the observed and predicted total outage counts, respectively. Their means are $\bar{Y} = E^{-1} \sum_{m=1}^{E} Y_m$ and $\overline{\hat{Y}} = E^{-1} \sum_{m=1}^{E} \hat{Y}_m$.

Absolute error (AE) and mean absolute error (MAE) measure the magnitude of prediction errors:

$$\mathrm{AE}_m = |\hat{Y}_m - Y_m|, \qquad \mathrm{MAE} = \frac{1}{E} \sum_{m=1}^{E} |\hat{Y}_m - Y_m|. \tag{12}$$

For positive observed event totals, absolute percentage error (APE) and mean absolute percentage error (MAPE) measure relative prediction errors:

$$\mathrm{APE}_m = 100\% \times \frac{|\hat{Y}_m - Y_m|}{Y_m}, \qquad \mathrm{MAPE} = \frac{1}{E} \sum_{m=1}^{E} \mathrm{APE}_m. \tag{13}$$

We also report the 25th and 50th percentiles of event-level AE and APE, denoted by Q25 and Q50, respectively. Centered root mean square error (CRMSE) measures prediction error after removing the mean bias:

$$\mathrm{CRMSE} = \sqrt{\frac{1}{E} \sum_{m=1}^{E} \left[ (\hat{Y}_m - \overline{\hat{Y}}) - (Y_m - \bar{Y}) \right]^2}. \tag{14}$$

The coefficient of determination measures squared prediction error relative to variation in the observed totals:

$$R^2 = 1 - \frac{\sum_{m=1}^{E} (\hat{Y}_m - Y_m)^2}{\sum_{m=1}^{E} (Y_m - \bar{Y})^2}. \tag{15}$$

Lower values indicate better performance for all error metrics, whereas higher $R^2$ indicates better performance.

## D Task-Aligned Retrieval Encoder

Proximity in the raw covariate space does not necessarily reflect similarity in outage response, since different feature dimensions can have substantially different predictive relevance. We therefore learn a lightweight encoder $g_\phi : \mathbb{R}^d \to \mathbb{R}^{d_r}$ to construct a task-aligned retrieval space. For an input $x_i$, the retrieval representation is

$$\mathbf{r}_i = \frac{g_\phi(x_i)}{\|g_\phi(x_i)\|_2}, \tag{16}$$

The encoder is used exclusively for retrieval; the original covariates are retained as inputs to the frozen foundation model.

Because outage counts are zero-inflated and highly long-tailed, directly imposing metric supervision on raw target differences can overemphasize large-magnitude differences in the sparse tail. Instead, we introduce a monotone multi-resolution quantizer $b : \mathcal{Y} \to \{0, \ldots, B-1\}$ that preserves the ordering of outage severity while allocating finer resolution to densely populated low-count regions and coarser resolution to the sparse tail. For two training samples $i$ and $j$, we define their ordinal target distance as

$$\Delta_{ij} = |b(y_i) - b(y_j)|\,. \tag{17}$$

We optimize the encoder using Rank-N-Contrast (Zha et al., 2023), which encourages representation similarity to follow this ordinal target structure. For an anchor $i$ in minibatch $\mathcal{B}$, the loss is

$$\ell_i = -\frac{1}{|\mathcal{B}|-1} \sum_{\substack{j \in \mathcal{B} \\ j \neq i}} \log \frac{\exp(\mathbf{r}_i^\top \mathbf{r}_j / \tau)}{\sum_{\substack{k \in \mathcal{B} \setminus \{i\} \\ \Delta_{ik} \geq \Delta_{ij}}} \exp(\mathbf{r}_i^\top \mathbf{r}_k / \tau)}, \tag{18}$$

where $\tau$ is a temperature parameter. Thus, samples that are closer in outage severity are encouraged to remain closer in the retrieval space than samples with more distant target levels.

To prevent the dominant zero and low-count regimes from overwhelming the metric objective, we reweight anchors according to the frequency of their target level. Let $n_b$ denote the number of training samples assigned to level $b$. We use inverse-frequency smoothing,

$$w_i \propto n_{b(y_i)}^{-1/2}, \tag{19}$$

and optimize the weighted contrastive objective together with an auxiliary nonzero-discrimination loss:

$$\mathcal{L}_{\text{enc}} = \frac{\sum_{i \in \mathcal{B}} w_i \ell_i}{\sum_{i \in \mathcal{B}} w_i} + \lambda_{\text{nz}} \mathcal{L}_{\text{nz}}. \tag{20}$$

Here, $\mathcal{L}_{\text{nz}}$ is a binary cross-entropy objective for predicting whether $y_i > 0$ from $\mathbf{r}_i$. The ordinal objective captures graded similarity among outage magnitudes, while the auxiliary term explicitly separates the large zero mass from positive-outage observations.

## E Additional Results

Table 6 reports performance across four outage-severity groups on CT. SG-TTA reduces both MAE and MAPE relative to each backbone in every group, indicating that its benefits extend across severity levels. TabPFN with SG-TTA achieves the lowest MAE and MAPE in the moderate and extreme groups and the lowest MAPE in the high group, while TabICL with SG-TTA obtains the lowest MAE in the high group. SA-HGNN remains strongest in the low group, suggesting that the relative advantages of structural test-time adaptation are more pronounced for moderate-to-extreme events.

Table 6: Grouped performance comparison by true outage count range. Best results are highlighted in bold, and second-best results are underlined.

| Method | Low | | Moderate | | High | | Extreme | |
|---|---|---|---|---|---|---|---|---|
| | MAE | MAPE | MAE | MAPE | MAE | MAPE | MAE | MAPE |
| Random Forest | 175.02 | 215.85 | 314.18 | 70.32 | 1164.85 | 74.33 | 10308.00 | 53.69 |
| XGBoost | 144.37 | 150.64 | 382.16 | 56.33 | 1120.30 | 73.79 | 9911.67 | <u>51.35</u> |
| GraphSAGE | <u>74.49</u> | <u>70.79</u> | 303.86 | 62.98 | 1047.63 | 66.37 | 16528.67 | 92.51 |
| SA-HGNN | **61.32** | **51.12** | 290.66 | 52.56 | 1165.43 | 72.24 | 15974.72 | 83.84 |
| TabICL | 129.16 | 135.07 | 308.70 | 62.95 | 926.20 | 67.21 | 10919.52 | 57.73 |
| + SG-TTA | 90.87 | 97.80 | <u>239.77</u> | <u>50.37</u> | **895.53** | <u>65.77</u> | <u>9715.43</u> | 52.37 |
| TabPFN | 103.69 | 101.38 | 332.45 | 67.53 | 1189.52 | 83.42 | 12323.95 | 66.60 |
| + LoCalPFN | 138.60 | 134.30 | 242.20 | 51.90 | 951.88 | 67.20 | 11623.51 | 61.92 |
| + G2T-FM | 104.01 | 99.90 | 263.04 | 53.77 | 1015.85 | 73.80 | 13426.73 | 73.23 |
| + SG-TTA | 83.56 | 78.08 | **239.45** | **50.26** | <u>915.67</u> | **63.49** | **9450.92** | **48.39** |

Table 7 further examines prediction accuracy at the spatially resolved row level. With TabPFN as the backbone, SG-TTA reduces Row-MAE from 0.778 to 0.651 and Row-RMSE from 3.98 to 3.10, corresponding to relative reductions of 16.3% and 22.1%, respectively. Notably, SG-TTA also outperforms spatially aware supervised baselines such as GraphSAGE and SA-HGNN on both row-level metrics. These results show that the improvement is not limited to better aggregate outage estimates; it is reflected in more accurate predictions at individual spatial locations, consistent with the goal of using graph structure as test-time feedback.

The advantage remains when events are weighted equally. SG-TTA achieves the lowest Evt-RMSE of 1.19, compared with 1.57 for vanilla TabPFN, 1.30 for LoCalPFN, and 1.27 for G2T-FM. This indicates that the row-level gains are broadly distributed across target events rather than being dominated by a small number of high-outage events. Moreover, applying SG-TTA to TabICL also reduces all three errors, with Evt-RMSE decreasing from 1.30 to 1.22. Together, these results provide complementary evidence that structure-grounded adaptation improves the spatially resolved outage field at test time, rather than merely adjusting event-level prediction scale.

Table 7: Row-level prediction errors on the Connecticut dataset. Row-MAE and Row-RMSE are computed over all rows pooled across events, while Evt-RMSE first computes RMSE within each event and then averages over events, so that every event is weighted equally. Best and second-best results are highlighted in **bold** and underlined, respectively.

| **Method** | **Row-MAE** ↓ | **Row-RMSE** ↓ | **Evt-RMSE** ↓ |
|---|---|---|---|
| Random Forest | 0.828 | 3.26 | 1.30 |
| XGBoost | 0.770 | <u>3.11</u> | 1.30 |
| GraphSAGE | <u>0.676</u> | 3.55 | 1.24 |
| SA-HGNN | 0.699 | 3.79 | 1.26 |
| TabICL | 0.746 | 3.37 | 1.30 |
| + SG-TTA | 0.721 | 3.34 | <u>1.22</u> |
| TabPFN | 0.778 | 3.98 | 1.57 |
| + LoCalPFN | 0.737 | 3.33 | 1.30 |
| + G2T-FM | 0.724 | 3.40 | 1.27 |
| + SG-TTA | **0.651** | **3.10** | **1.19** |

Table 8 compares the test-time cost of SG-TTA with frozen ICL using the same backbone and retrieved context. With five adaptation steps, SG-TTA increases the average runtime from 1.08 to 3.84 seconds per event, corresponding to a $3.56\times$ cost and an additional 2.76 seconds per event. The total runtime on CT increases from 5.3 to 18.8 minutes.

Table 8: **Test-time computational cost on the CT dataset.** Both methods use the same TabPFN backbone and retrieved context $\mathcal{C}^\star$. Runtime is averaged per target event on the same GPU and excludes one-time offline training of the retrieval encoder and model loading.

| Method | Steps | Time / Event (s) | median | × frozen | Total (min) |
|---|---|---|---|---|---|
| Frozen ICL | 0 | $1.08 \pm 0.08$ | 1.07 | $1.0\times$ | 5.3 |
| SG-TTA | $T = 5$ | $3.84 \pm 0.78$ | 3.82 | $3.56\times$ | 18.8 |

## F A Local Interpretation of Gradient Alignment

We provide a local interpretation of the gradient-alignment analysis in the main text. Consider a target event and let $\psi$ denote the parameters updated during adaptation. The structural update direction is

$$g_{\mathrm{str}} = g_{\mathrm{rel}} + \lambda g_{\mathrm{self}}, \tag{21}$$

where $g_{\mathrm{rel}}$ and $g_{\mathrm{self}}$ are the gradients computed by the two structural objectives, including their prescribed stop-gradient operations. Let $\mathcal{L}_{\mathrm{sup}}(\psi)$ be a differentiable prediction loss evaluated using the target labels, and define $g_{\mathrm{sup}} = \nabla_\psi \mathcal{L}_{\mathrm{sup}}(\psi)$. This supervised loss is used only for analysis; target labels are unavailable to the adaptation procedure.

**Proposition 1 (Local benefit of aligned structural feedback).** Suppose $\mathcal{L}_{\mathrm{sup}}$ has an $L$-Lipschitz gradient in a neighborhood containing the update segment. For an update

$$\psi^+ = \psi - \alpha g_{\mathrm{str}}, \qquad \alpha > 0, \tag{22}$$

we have

$$\mathcal{L}_{\mathrm{sup}}(\psi^+) \le \mathcal{L}_{\mathrm{sup}}(\psi) - \alpha \langle g_{\mathrm{sup}}, g_{\mathrm{str}} \rangle + \frac{L\alpha^2}{2} \|g_{\mathrm{str}}\|_2^2. \tag{23}$$

Consequently, if $\langle g_{\mathrm{sup}}, g_{\mathrm{str}} \rangle > 0$, the supervised loss decreases for a sufficiently small step size.

**Proof.** Applying the smoothness inequality to the displacement $-\alpha g_{\mathrm{str}}$ yields the bound. For positive alignment, its right-hand side is strictly smaller than $\mathcal{L}_{\mathrm{sup}}(\psi)$ whenever

$$0 < \alpha < \frac{2\langle g_{\mathrm{sup}}, g_{\mathrm{str}} \rangle}{L \|g_{\mathrm{str}}\|_2^2}. \tag{24}$$

Global-norm clipping rescales the update direction by a nonnegative scalar and can therefore be incorporated into $\alpha$. □

**Connection to the empirical analysis.** The proposition motivates measuring the alignment between the supervised gradient and the full structural update:

$$\langle g_{\mathrm{sup}}, g_{\mathrm{str}} \rangle = \langle g_{\mathrm{sup}}, g_{\mathrm{rel}} \rangle + \lambda \langle g_{\mathrm{sup}}, g_{\mathrm{self}} \rangle. \tag{25}$$

The individual measurements characterize the first-order contribution of each signal, while the combined measurement characterizes their joint update direction. Cosine alignment additionally measures directional agreement independently of gradient magnitude. Neither objective is guaranteed to align with the prediction task on every event; the measurements in Table 4 assess whether such alignment occurs empirically.

This result is local and concerns the differentiable supervised loss used to compute $g_{\mathrm{sup}}$. Its association with MAE improvement after multiple adaptation steps is evaluated empirically rather than guaranteed by the proposition.