WEP Warwick Evans Publishing

Academic Journal of Management Science and Engineering

ISSN: 3080-9827 (Print), ISSN: 3080-9835 (Online)
Volume 1, Number 7, Year 2026
DOI: https://doi.org/10.62051/ajmse.v1n7.12



# CF-TGNN: Counterfactual Explainable Temporal Graph Neural Networks for Supply Chain Disruption Prediction

Clara Chandra *, Theo Adler

McCombs School of Business, University of Texas at Austin, Austin, TX 78712, USA
* Corresponding author: Clara Chandra (Email: cc4840920@utexas.edu)

## ABSTRACT

Disruptions in multi-tier supply chains travel from upstream suppliers to downstream customers, and a risk model helps planners only if it can say what would have prevented a predicted disruption. Temporal graph neural networks with supplier attention forecast such ripple effects well, but attention weights are normalized over a node's suppliers: they always place blame upstream, cannot state that no supplier is responsible, and do not describe an intervention. We present CF-TGNN, a counterfactual explainable temporal graph neural network. Its backbone adopts a directed, dependency-weighted temporal attention operator and adds a region-context encoder that acts as a proxy for confounding regional shocks. Its Counterfactual Restoration Explainer (CRE) evaluates supplier-restoration interventions with do-operator graph surgery and returns a but-for estimate of the probability that upstream disruptions were necessary, a restoration Shapley decomposition of the predicted risk, an exact minimum-cardinality restoration plan, and an origin trace that follows causes backward in time. To test explanations against true counterfactuals, we turn a published cascade simulator into a structural causal model with logged exogenous noise, so that every what-if query can be answered by replay. Over five seeds, CF-TGNN raises onset AUC from 0.765 to 0.786; CRE identifies cascade origins with 44.2% accuracy versus 33.8% for layer-wise attention chaining; and graph surgery reduces the error of but-for estimates by 22%. A case study on the public SupplyGraph dataset shows that CRE returns sparse restoration plans on real production records and makes disagreements between backbones explicit, although the region context does not improve prediction on these data.

## KEYWORDS

Supply Chain Disruption; Ripple Effect; Temporal Graph Neural Networks; Counterfactual Explanation; Causal Inference; Explainable Artificial Intelligence.

## 1. INTRODUCTION

Modern supply chains are multi-tier networks in which a finished product depends, directly or through intermediate parts, on hundreds of suppliers. A local failure seldom stays local: it travels along material-flow dependencies and reaches customers several tiers downstream, a phenomenon known as the ripple effect [1], [2]. Firm-level data confirm that such propagation can multiply the economic loss of a shock far beyond its origin [3], and the COVID-19 pandemic made the mechanism visible across entire industries [4]. Carbon-aware offline reinforcement learning shows how inventory and transportation decisions can be evaluated jointly under service and emissions constraints, strengthening the methodological context for resilient supply-chain planning [31].

Machine learning is now routinely used to anticipate these events, yet practitioners keep asking for models whose outputs they can act on [5]. Recent work stresses that predicting risk and planning

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

interventions are different problems: a predictor may exploit any correlation, whereas mitigation requires statements about what would change under an action [6]. Graph neural networks (GNNs) are natural predictors for supplier networks [7], and the explainable temporal GNN X-TGNN of Wang *et al.* [8] made an important step for this setting. It pairs a directed, dependency-weighted supplier attention with a simulation testbed that records the true responsible suppliers of every propagated disruption, and its attention weights recovered root-cause suppliers about as well as gradient and perturbation explainers at a fraction of their cost.

Attention weights answer a descriptive question, namely where the model looked. A planner asks a counterfactual one: would the disruption have happened had supplier $i$ been kept running? The difference matters in three ways. First, attention is normalized over a node's suppliers, so it distributes blame upstream even when a disruption originates inside the node or its region; it cannot say that no supplier is responsible. Second, a list of weights is not an intervention: it does not tell which suppliers, and how few, must be protected to avert the predicted event. Third, chaining attention through the stacked layers of a single snapshot describes how information was aggregated, not how a disruption moved through time. Counterfactual explanations [9] and algorithmic recourse [10] target exactly these questions, and counterfactual explainers for GNNs exist [11], [12]. They perturb generic edges or features, however, ignore the temporal causal order of a cascade, and are rarely checked against true counterfactual outcomes.

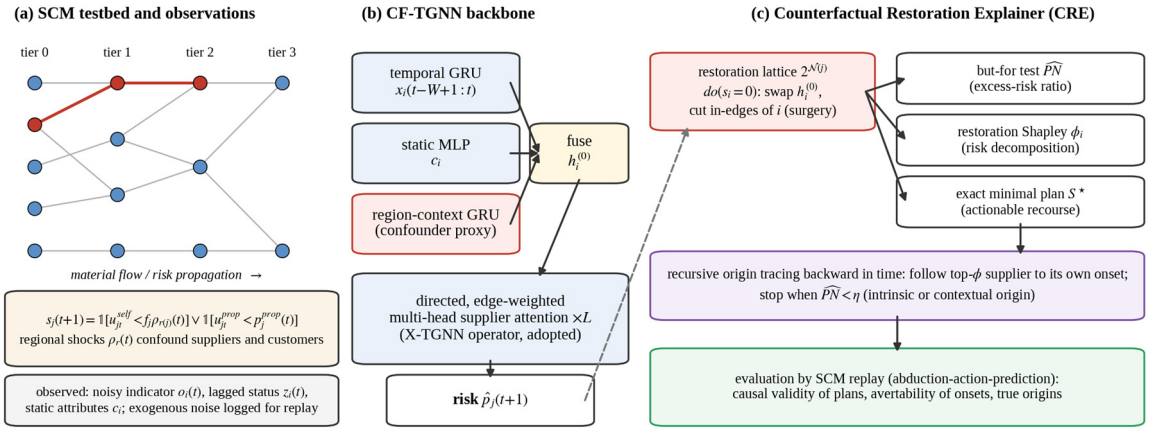



**Fig 1.** Overview of CF-TGNN. (a) The counterfactual testbed: a structural causal model of multi-tier cascades with logged exogenous noise and confounding regional shocks. (b) The backbone keeps the directed, dependency-weighted supplier attention of X-TGNN and adds a region-context encoder. (c) CRE enumerates supplier-restoration interventions with graph surgery and returns a but-for test, a Shapley risk decomposition, an exact minimal plan and a backward origin trace; all outputs are checked by replay in the structural model.

This paper develops CF-TGNN, a counterfactual explainable temporal GNN for supply chain disruption prediction (Fig. 1). Its contributions are as follows.

1) A region-aware backbone that keeps the X-TGNN supplier-attention operator and adds an encoder of regional disruption levels. The context is a proxy for regional shocks, which confound supplier and customer states, and it improves the prediction of new disruptions.

2) CRE, an explainer built on supplier-restoration interventions implemented with do-operator graph surgery. From one exactly enumerated restoration lattice it derives a but-for test grounded in probabilities of causation, a restoration Shapley decomposition of the predicted risk, an exact minimal restoration plan and a recursive tracer that locates the origin of a cascade backward in time.

3) A counterfactual testbed obtained by turning the cascade simulator of X-TGNN into a structural causal model with logged exogenous noise and regional confounding, so that the causal validity, avertability and true origin of every predicted disruption can be computed exactly.

4) An evaluation over five seeds with seven predictors and six explainers, ablations, a confounding sweep and a case study on the public SupplyGraph data.

# 2. RELATED WORK

## 2.1. Disruption Propagation and Risk Analytics

Quantitative ripple-effect research has relied mainly on simulation and optimization [2]. Basole and Bellamy used an epidemic-style diffusion model to show that network structure and visibility govern how far risk spreads [13], and digital supply chain twins have been proposed to replay disruption scenarios on a model of the real network [14]. Data-driven studies predict delays and disruptions from transactional records [5], [6], infer hidden supplier links with GNNs [7], or benchmark GNNs on real production and distribution series [15]. X-TGNN, discussed above, is the closest work: it predicts multi-tier disruptions with a temporal GNN and explains them with intrinsic attention. We keep its operator and its testbed philosophy and add counterfactual semantics on both sides. Liu et al. further demonstrate a physics-informed multi-agent formulation that incorporates operational feasibility and risk into coordinated supply-chain optimization [32].

## 2.2. Temporal Graph Neural Networks

Message-passing GNNs such as GCN [16], GAT [17] and GraphSAGE [18] learn node representations from neighbors. Temporal extensions either evolve the GNN parameters over time [19], maintain per-node memories updated by events [20], or encode each node's recent history with a gated recurrent unit (GRU) [21] before spatial aggregation. CF-TGNN follows the last design because node-local temporal encodings make supplier interventions exact (Section IV-B).

## 2.3. Factual, Counterfactual and Causal Explanations

Factual GNN explainers select subgraphs or features that preserve a prediction, as GNNExplainer does [22]; Yuan *et al.* survey the area [23], and T-GNNExplainer extends the idea to temporal event graphs [24]. Counterfactual explainers look for minimal changes that alter a prediction: CF-GNNExplainer deletes edges [11], CF² balances factual and counterfactual criteria [25], and a recent survey reviews graph counterfactuals [12]. Whether attention weights explain a model has been debated at length [26]. Shapley values [27] underlie popular attribution methods [28]. We adopt the structural view of causation [29]: an explanation is an intervention, its correctness is judged by counterfactual outcomes, and the necessity of a cause is quantified by probabilities of causation [30]. We are not aware of prior work that brings these elements together for temporal supply-network models or that evaluates them against replayed counterfactuals.

# 3. PROBLEM SETTING AND COUNTERFACTUAL TESTBED

## 3.1. Temporal Multi-Tier Supply Graph

A supply chain is a directed graph $G = (V, E, W)$ whose nodes are firms or sites arranged in tiers 0, …, $K-1$ from raw materials to final assembly. An edge $(i \to j)$ with weight $w_{ij} \in (0, 1]$ states that $j$ sources from $i$, and $\mathcal{N}(j)$ denotes the suppliers of $j$. Each node has a hidden state $s_i(t) \in \{0, 1\}$ (operational or disrupted). The learner observes static attributes $c_i$ (tier, inventory buffer, region, degrees, mean dependency weight), a noisy operational indicator $o_i(t) = s_i(t) + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, and the lagged

confirmed status $z_i(t) = s_i(t-1)$. The prediction task is to estimate $P(s_j(t+1) = 1)$ from a window of length $W$, with emphasis on onsets, that is, on nodes that are operational at time $t$.

### 3.2. Structural Causal Model

We re-implement the cascade dynamics of the X-TGNN testbed and make every mechanism explicit. The dependency-weighted pressure on $j$ and its buffer-attenuated propagation hazard are

$$P_j(t) = \frac{\sum_{i \in \mathcal{N}(j)} w_{ij} s_i(t)}{\sum_{i \in \mathcal{N}(j)} w_{ij}} \tag{1}$$

$$p_j^{prop}(t) = \theta\big(1 - b_j\big) P_j(t)^{\gamma} \tag{2}$$

where $b_j$ is the inventory buffer. Intrinsic failures occur with hazard $p_j^{self}(t) = f_j \rho_{r(j)}(t)$, where $f_j$ is an unobserved fragility and $\rho_r(t) \in \{1, \kappa\}$ is the regime of region $r$, which switches between a normal and a shock state as a two-state Markov chain. An operational node evolves as

$$s_j(t+1) = \mathbb{1}\big[u_{jt}^{self} < p_j^{self}(t)\big] \vee \mathbb{1}\big[u_{jt}^{prop} < p_j^{prop}(t)\big] \tag{3}$$

with independent exogenous noise $u \sim U(0, 1)$, and a disrupted node recovers with probability $q_j$ per step. Two design choices turn the simulator into a counterfactual testbed. First, all exogenous noise is stored, so any query of the form "what would $s_j(t+1)$ have been under intervention $do(\cdot)$?" is answered exactly by abduction, action and prediction [29]: the noise is kept, the intervened variables are changed, and (1)–(3) are recomputed. Second, customers source from suppliers in their own region with probability $q_{loc}$, so that a regional shock disrupts suppliers and customers at the same time. The shock then becomes a confounder, and a model may learn that a disrupted supplier predicts a disrupted customer even when no material ever failed to arrive.

### 3.3. Counterfactual Ground Truth

For an onset of node $j$ at $t+1$ we define three quantities. The onset is *avertable* if restoring all suppliers at $t$, $do(s_i(t) = 0, i \in \mathcal{N}(j))$, would have prevented it under the logged noise, which holds exactly when the propagation mechanism fired and the intrinsic one did not. A restoration plan $S \subseteq \mathcal{N}(j)$ is *causally valid* if $do(s_i(t) = 0, i \in S)$ averts the onset. The *origin* of the onset is found by following the dominant disrupted supplier, i.e., the one with the largest responsibility $w_{ij}/\Sigma_{k \in D} w_{kj}$ used by the original testbed, to that supplier's own onset, recursively, until an onset whose intrinsic mechanism fired.

*Proposition 1 (avertability):* given an onset of $j$ with hazards $p^{self}$ and $p^{prop}$, and $p_j = 1-(1-p^{self})(1-p^{prop})$, the probability that the onset is avertable equals

$$\mathrm{PN} = \frac{p_j - p_j^{self}}{p_j} \tag{4}$$

*Proof sketch:* restoring every supplier sets $P_j = 0$ and hence $p^{prop} = 0$, so the onset is averted if and only if the intrinsic mechanism did not fire. Conditioning on the onset gives $(1-p^{self})p^{prop}/p_j = (p_j - p^{self})/p_j$. This is the excess-risk ratio, which identifies the probability of necessity under exogeneity and monotonicity [30]; the noisy-OR mechanism (3) is monotone.

*Proposition 2 (locality):* $s_j(t+1)$ depends on interventions at time $t$ only through $s_i(t)$, $i \in \mathcal{N}(j)$, and the edges into $j$; an intervention further upstream at time $t$ cannot affect $j$ before $t+2$. The proposition follows directly from (1)–(3). It limits a single-step counterfactual search to the suppliers of the target and implies that multi-hop explanations must be traced backward in time rather than through the layers of one snapshot.

# 4. METHOD: CF-TGNN

## 4.1. Region-Aware Temporal Backbone

CF-TGNN adopts the directed, dependency-weighted, multi-head supplier attention of X-TGNN [8] together with its GRU temporal encoder [21]. For node $i$, a GRU summarizes the window $x_i(t-W+1 : t)$ into $h_i^{tmp}$, and a multilayer perceptron (MLP) embeds $c_i$. The new component is a region-context encoder. For each region $r$ we form the series $\zeta_r(\tau)$, the mean of $o$ and $z$ over the members of the region, and a second GRU encodes it into $g_r(t)$. The initial node state is

$$h_i^{(0)} = \mathrm{ReLU}\big(W_f\big[h_i^{tmp} \,\|\, \mathrm{MLP}(c_i) \,\|\, g_{r(i)}(t)\big]\big) \tag{5}$$

and the supplier attention and residual update of X-TGNN are kept unchanged:

$$e_{ij}^{(l)} = \mathrm{LeakyReLU}\Big(a_s^\top W h_i^{(l)} + a_d^\top W h_j^{(l)} + a_e w_{ij}\Big) \tag{6}$$

$$h_j^{(l+1)} = \mathrm{ReLU}\Big(h_j^{(l)} + \textstyle\sum_m^M \sum_{i\in\mathcal{N}(j)} \alpha_{ij}^{(l,m)} W^{(m)} h_i^{(l)}\Big) \tag{7}$$

where $\alpha_{ij}^{(l,m)}$ is the softmax of $e_{ij}^{(l)}$ over $\mathcal{N}(j)$ for head $m$, and the risk is $\hat{p}_j(t+1) = \sigma(w_o^\top h_j^{(L)} + b)$. Training minimizes a class-weighted binary cross-entropy with positive weight $\beta$. Because the weighting shifts the logit by $\log\beta$, calibrated probabilities are recovered as $\tilde{p} = \sigma(\mathrm{logit} - \log\beta)$; CRE uses them whenever risk is needed on a probability scale.

The region context serves causality as much as accuracy. In the testbed, and plausibly in practice, the regional regime is a common cause of a supplier's and a customer's state. Conditioning on a proxy of it blocks the back-door path supplier ← shock → customer, so that the share of risk the model routes through suppliers moves closer to the share that actually travels through them.

## 4.2. Restoration Interventions with Graph Surgery

The atomic action of CRE restores a supplier set $S$, $do(s_i = 0 \text{ for } i \in S)$. It stands for any measure that neutralizes a supplier's disruption from the customer's viewpoint, such as expediting, releasing safety stock or opening a temporary alternative lane. We implement it on the model in two steps. *Input restoration:* the last $\delta$ steps of the window of $i$ are set to the operational state ($o = 0$, $z = 0$) and its encoding is recomputed as $\tilde{h}_i^{(0)}$. Because $h_i^{(0)}$ depends only on the window of $i$, its static attributes and the regional context, the swap $h_i^{(0)} \to \tilde{h}_i^{(0)}$ is exact, and the restored encodings of all nodes come from one batched pass. The regional context is held fixed because it proxies an exogenous parent of the state of $i$. *Graph surgery:* all edges into $i$ are removed in every attention layer. In the structural model, $do(s_i = 0)$ disconnects $i$ from its own causes; without surgery, a two-layer model would still read the upstream suppliers of $i$ as evidence that $i$ is disrupted and would mix observation with intervention. We write $f_j(S)$ for the resulting logit of $j$ and $\tilde{p}_j(S)$ for its calibrated risk.

## 4.3. Counterfactual Restoration Explainer

For a target onset $(j, t)$ with $n = |\mathcal{N}(j)|$ suppliers, CRE evaluates the restoration lattice $\{f_j(S) : S \subseteq \mathcal{N}(j)\}$, which by Proposition 2 is the complete single-step causal search space. The $2^n$ evaluations reuse cached encodings and run as one batch through the attention layers. With the value function $v(S) = \tilde{p}_j(\emptyset) - \tilde{p}_j(S)$, CRE reads four explanations from the lattice.

*But-for test.* Following Proposition 1,

$$\hat{\mathrm{PN}} = \frac{\tilde{p}_j(\emptyset) - \tilde{p}_j(\mathcal{N}(j))}{\tilde{p}_j(\emptyset)} \tag{8}$$

estimates the probability that upstream disruptions were necessary for the predicted onset. CRE labels the risk as propagated if $\hat{\text{PN}} \geq \eta$ and as intrinsic or contextual otherwise; we use $\eta = 0.5$, i.e., "more likely than not".

*Restoration Shapley decomposition.* The Shapley value [27] of supplier $i$ in the restoration game is

$$\varphi_i = \sum_{S \subseteq \mathcal{N}(j) \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} [v(S \cup \{i\}) - v(S)] \quad (9)$$

By efficiency, $\tilde{p}_j(\emptyset) = \tilde{p}_j(\mathcal{N}(j)) + \Sigma_i \varphi_i$: the predicted risk splits into an intrinsic and contextual residual plus one additive share per supplier. Unlike attention, the supplier shares can sum to almost nothing.

*Minimal restoration plan.* The plan $S^\star$ is the smallest $S$ with $f_j(S) < \tau$, where $\tau$ is the decision threshold, with ties broken by the lowest residual risk. Because the lattice contains every candidate, $S^\star$ is exact rather than heuristic.

*Origin tracing.* By Proposition 2, a disruption reaches $j$ from two tiers upstream only through an earlier onset of an intermediate supplier. Algorithm 1 therefore moves backward in time: it attributes the onset to the supplier with the largest $\varphi_i$, reads the onset time of that supplier from its observed status history, and repeats until the but-for test reports an intrinsic or contextual cause.

**Algorithm 1:** Counterfactual origin tracing. Input: predicted onset $(j, t)$, model $f$, threshold $\eta$, depth limit $D$.

1: compute lattice $\{f_j(S)\}_{S \subseteq \mathcal{N}(j)}$ of $(j, t)$ with surgery
2: if $\mathcal{N}(j) = \emptyset$ or $\hat{\text{PN}} < \eta$ (8): **return** $j$
3: $i \leftarrow \text{argmax}_k \varphi_k$ (9)
4: $t' \leftarrow$ decision time of the onset of $i$, read from $z_i(\cdot)$
5: $(j, t) \leftarrow (i, t')$; repeat from line 1 (at most $D$ times)

*Complexity.* Encodings are computed once per time step. A query costs $2^n$ evaluations of the attention layers on the restored graph, 12.2 on average in our testbed, where $n \leq 5$. For high in-degree, the lattice can be restricted to the $k$ suppliers ranked highest by attention, which Section V-D shows to be an almost lossless search order, and $\varphi$ can be estimated by permutation sampling.

# 5. EXPERIMENTS

## 5.1. Setup

*Testbed.* Following the default instance of X-TGNN [8], each run generates a four-tier network of 260 nodes by preferential attachment, to which we add the regional sourcing affinity of Section III-B ($q_{\text{loc}} = 0.6$); each run simulates $T = 400$ steps; Table 1 lists the statistics, averaged over the five seeds used throughout. About 59% of all onsets involve the propagation mechanism, close to the 60.5% of the original testbed. Decision times are split chronologically into 60% training, 20% validation and 20% test data, so models are always evaluated on future periods.

*Predictors.* MLP; GRU (temporal encoder only); GRU+RC (with the region context but no graph); GCN-T and GraphSAGE-T (the same temporal encoder with dependency-weighted mean and mean aggregation); X-TGNN (our re-implementation); and CF-TGNN. All models use hidden size 32, window $W = 8$, $L = 2$ graph layers and $M = 4$ heads where applicable, Adam with learning rate $3 \times 10^{-3}$,

and batches of eight snapshots. Training runs for at most 30 epochs; the checkpoint with the best validation onset AUC is kept (minimum 12 epochs, patience 8), and the decision threshold maximizes validation F1. The code is written in JAX and runs on a single CPU core.

**Table 1.** Counterfactual testbed statistics (mean over five seeds)

| Property | Value |
|---|---|
| Tiers / regions / time steps | 4 / 6 / 400 |
| Nodes / directed edges | 260 / 453.2 |
| Max. out-degree / mean in-degree | 14.8 / 3.02 |
| Same-region dependency edges | 66.4% |
| Disruption prevalence | 16.5% |
| Onsets per run | 5186 |
| Propagation mechanism fired | 58.9% |
| Shock level κ / on, off probabilities | 8 / 0.015, 0.08 |

*Explainers.* Unless stated otherwise, all explainers act on the same trained CF-TGNN. Random; Attention, the layer-averaged attention of X-TGNN; Gradient × Δ, the first-order restoration effect $\nabla_x f \cdot (x - \tilde{x})$ over the restored window; Occlusion, the single-supplier restoration effect $v(\{i\})$; CF-GNNExplainer [11], adapted to directed supplier edges with sigmoid edge masks, 100 Adam steps and the sparsest valid deletion retained; and CRE. Ranking explainers form plans by restoring suppliers in rank order until the decision flips.

*Metrics.* Root-cause attribution follows the original protocol: onsets whose propagation mechanism fired and that have at least two disrupted suppliers, scored by precision@1 (P@1), NDCG and Spearman's ρ against the responsibility $w_{ij}/\Sigma_{k \in D} w_{kj}$. We add Comp@1, the relative drop of calibrated risk when the top-ranked supplier is restored, as a faithfulness measure. Plans are evaluated on true-positive onsets by coverage (share of events that receive a valid plan), size $|S|$, share of minimum-cardinality plans, causal validity (CV) under replay, wasted actions (restored suppliers that were in fact operational) and model evaluations per query. The but-for test is scored by AUC against avertability and by its mean absolute error (MAE) and bias against the oracle PN of Proposition 1. Origin tracing is scored by exact-match accuracy on true-positive onsets.

## 5.2. Disruption Prediction

Table 2 reports test performance. Graph models predict new disruptions far better than models that see only a node's own history: the onset AP of X-TGNN is 1.8× that of the GRU, the same pattern as in the original study. The region context helps with and without a graph. It raises the onset AUC of the GRU from 0.697 to 0.731 and that of the attention backbone from 0.765 to 0.786 (higher in each of the five seeds), because regional shocks drive a large share of intrinsic onsets that supplier signals cannot anticipate. GRU+RC nevertheless stays well below every graph model on onset AP, so context complements, rather than replaces, propagation along dependencies.

## 5.3. Root-Cause Attribution

Table 3 follows the ground-truth protocol of the original testbed on 193 events per seed. All principled methods clearly beat random ranking, and the attention weights of the backbone remain a strong

ranking signal, consistent with the X-TGNN study. Attention attains the highest mean P@1, but on P@1, NDCG and ρ the four methods lie within about one standard deviation of each other. They differ on faithfulness: restoring the supplier ranked first by the restoration Shapley values lowers the calibrated risk by 41.6% on average, on par with single-supplier occlusion (41.7%) and above attention (38.7%) and gradients (38.5%). In other words, the Shapley ranking says what the model would do under the intervention, while attention says where the model looked.

**Table 2.** Disruption prediction on the testbed (mean ± std over five seeds)

| Model | AUC-ROC | AUC-PR | F1 | Onset AUC | Onset AP |
|---|---|---|---|---|---|
| MLP | 0.841±0.008 | 0.549±0.009 | 0.585±0.007 | 0.692±0.015 | 0.132±0.010 |
| GRU | 0.843±0.009 | 0.551±0.011 | 0.589±0.008 | 0.697±0.016 | 0.134±0.011 |
| GRU+RC | 0.856±0.006 | 0.562±0.010 | 0.598±0.008 | 0.731±0.013 | 0.155±0.013 |
| GCN-T | 0.868±0.010 | **0.579±0.013** | **0.626±0.011** | 0.766±0.019 | 0.241±0.025 |
| GraphSAGE-T | 0.867±0.011 | 0.579±0.015 | 0.624±0.011 | 0.762±0.019 | 0.224±0.023 |
| X-TGNN | 0.867±0.009 | 0.575±0.009 | 0.622±0.009 | 0.765±0.017 | 0.238±0.016 |
| CF-TGNN (ours) | **0.875±0.006** | 0.578±0.009 | 0.625±0.009 | **0.786±0.013** | **0.246±0.020** |

**Table 3.** Root-cause attribution on the CF-TGNN backbone

| Method | P@1 | NDCG | ρ | Comp@1 | Evals |
|---|---|---|---|---|---|
| Random | 0.263±0.023 | 0.775±0.012 | -0.029±0.059 | 0.203±0.022 | 0 |
| Attention (X-TGNN) | **0.573±0.066** | 0.933±0.007 | 0.623±0.039 | 0.387±0.017 | 1 |
| Gradient × Δ | 0.554±0.042 | 0.926±0.014 | 0.568±0.083 | 0.385±0.023 | 2 |
| Occlusion | 0.541±0.022 | 0.931±0.003 | 0.599±0.028 | **0.417±0.013** | $n$+1 |
| Restoration Shapley | 0.551±0.023 | **0.935±0.003** | **0.623±0.023** | 0.416±0.014 | $2^n$ |

## 5.4. Counterfactual Restoration Plans

Table 4 compares the plans on 399 true-positive onsets per seed. CRE returns a valid plan whenever one exists and every plan has minimum cardinality, at an average of 12.2 lattice evaluations. Greedy restoration in attention order is a remarkably good approximation: 94% of its plans are minimal and it needs only 1.5 evaluations, which makes the X-TGNN attention a natural search order when the lattice is too large to enumerate. Random order needs 1.5× more restorations and wastes 31% of them on suppliers that were never disrupted. Edge deletion, the intervention of CF-GNNExplainer, flips the decision for 58% of the events, against 90% for restorations, at a much higher cost. Deleting a dependency also reroutes demand to the remaining suppliers, which in the structural model raises their share of the pressure; a restoration expresses the intended action more directly.

Causal validity is similar across restoration methods at roughly 72%, and mainly reflects plan size: a plan only needs to push the model below its decision threshold, whereas replay asks whether the realized propagation noise is beaten. With the but-for gate, CRE withholds plans for onsets that it judges intrinsic; coverage falls from 90.0% to 80.7% and the share of wasted actions drops from 7.5% to 5.5%, while causal validity stays at 73%. The gate thus makes plans leaner rather than more often correct, a consequence of the moderate discrimination analysed next.

**Table 4.** Counterfactual plans on true-positive onsets (mean over seeds)

| Method | Cover. | \|S\| | Min. | CV | Waste | Evals |
|---|---|---|---|---|---|---|
| Random greedy | 0.89 | 1.80 | 0.61 | 0.68 | 0.31 | 2.0 |
| Attention greedy | 0.90 | 1.29 | 0.94 | 0.73 | 0.09 | 1.5 |
| Gradient greedy | 0.90 | 1.31 | 0.93 | 0.72 | 0.10 | 3.5 |
| Occlusion greedy | 0.90 | 1.25 | 0.99 | 0.73 | 0.08 | 4.4 |
| CF-GNNExplainer | 0.58 | 1.45 | – | 0.71 | 0.08 | 300.0 |
| CRE minimal | 0.90 | 1.23 | 1.00 | 0.72 | 0.07 | 12.2 |
| CRE (+ but-for) | 0.81 | 1.19 | 1.00 | 0.73 | 0.05 | 12.2 |

## 5.5. But-For Test and Confounding

Among true onsets with at least one disrupted supplier, 88% are avertable by restoring the suppliers; the rest stem from fragility or regional shocks. Raw attention weights cannot separate the two groups because they sum to one in both. Fig. 2(a) shows that $\widehat{\mathrm{PN}}$ separates them only moderately: its AUC is 0.592 on CF-TGNN and 0.536 on X-TGNN, while the oracle PN computed from the true hazards reaches 0.826. A heuristic that weights attention by the observed disruption indicators attains 0.570, comparable to $\widehat{\mathrm{PN}}$, but it is not a probability and has no principled threshold. Much of the gap to the oracle cannot be closed with the available data: node fragility is never observed, and avertability is random even when the hazards are known. The comparison between backbones is more telling. The region context raises the AUC of $\widehat{\mathrm{PN}}$ by 5.6 points, as the back-door argument of Section IV-A predicts. Both backbones understate the role of suppliers (mean bias -0.136 and -0.141, Fig. 2(b)), so $\widehat{\mathrm{PN}} \geq 0.5$ is a conservative criterion for declaring a disruption upstream-driven.

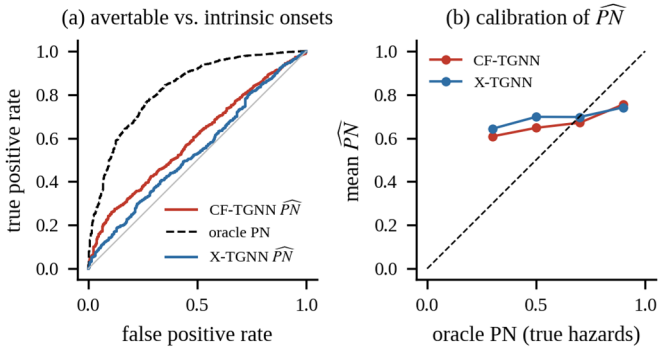



**Fig 2.** But-for test on true onsets with at least one disrupted supplier, pooled over five seeds. (a) ROC of $\widehat{\mathrm{PN}}$ for separating avertable from intrinsic onsets. (b) Mean $\widehat{\mathrm{PN}}$ per bin of the oracle PN of Proposition 1.

Fig. 3 varies the strength of regional confounding. Without shocks ($\kappa = 1$) the region context carries no information: both backbones reach the same onset AUC (0.794 and 0.795), and the $\widehat{\mathrm{PN}}$ of CF-TGNN is even slightly less discriminative (0.517 versus 0.537). As $\kappa$ grows, intrinsic onsets become more frequent and more clustered. The advantage of CF-TGNN in onset AUC is 2.1 points at $\kappa = 8$ and 2.5 points at $\kappa = 12$, and its advantage in the AUC of $\widehat{\mathrm{PN}}$ grows from 5.6 to 8.9 points. The context therefore helps only when there is a confounder to adjust for, and its benefit grows with the strength of that confounder, as the back-door argument of Section IV-A predicts.

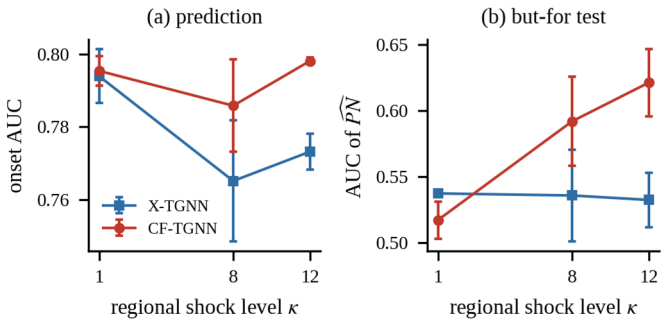



**Fig 3.** Effect of the regional shock level κ (κ = 1, 12: two seeds; κ = 8: five seeds). (a) Onset AUC. (b) AUC of $\widehat{PN}$ for avertable versus intrinsic onsets.

## 5.6. Multi-Hop Origin Tracing

True-positive onsets have origins up to three tiers away; over all seeds, 49% lie two or more hops upstream and 15% are self-originated. Fig. 4 compares four tracers. Chaining the highest attention weight through the two layers of one snapshot, as suggested for X-TGNN, reaches 33.8%: it can never answer "self", and its path length is fixed by the number of layers rather than by the cascade. Following attention backward in time instead of across layers raises the accuracy to 38.7%. Full CRE, which adds the but-for stopping rule, reaches 44.2%, a gain of 10.4 points over the snapshot chain. Replacing the Shapley ranking of CRE by attention while keeping the backward recursion and the but-for stop gives 43.6%, so the gain comes from the temporal recursion and the stopping rule rather than from the ranking; used inside a counterfactual procedure, the attention of X-TGNN remains a good guide. The stopping rule is what allows CRE to recognize self-originated onsets (33% versus 0%) and one-hop cascades (51% versus 44% without it). It occasionally stops early on long chains, where the tracer without stopping is slightly better.

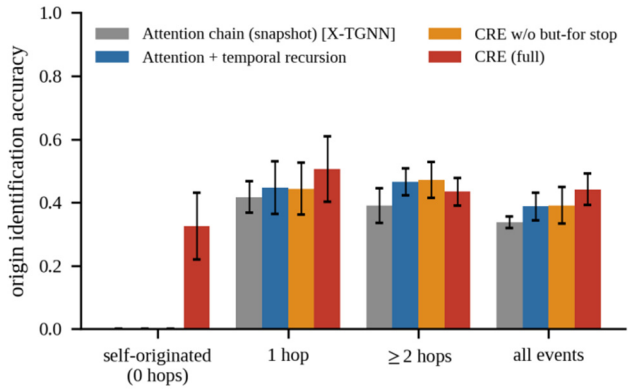



**Fig 4.** Origin identification accuracy by true chain length (mean ± std over five seeds).

## 5.7. Ablation Study

Table 5 isolates the design choices of CRE. Graph surgery has the largest effect on the but-for estimate: without it, upstream messages keep signalling that a restored supplier is still at risk, $\widehat{PN}$ is biased downward by -0.200, and its MAE grows from 0.197 to 0.252. Removing the region context, i.e., running CRE on X-TGNN, leaves the trace accuracy almost unchanged (0.481 versus 0.483) but lowers the AUC of $\widehat{PN}$ from 0.571 to 0.517. The restoration window δ matters less; Table 5 reports one and four steps for comparison.

## 5.8. Qualitative Example

Fig. 5 decomposes one test onset of a tier-3 node with predicted risk 0.71 (calibrated). The restoration Shapley values attribute most of the risk to the disrupted suppliers and almost none to the operational

ones, the intrinsic and contextual residual is 0.01, and $\widehat{PN} = 0.98$. The minimal plan restores 2 suppliers, and replay in the structural model confirms that this plan averts the onset. Attention points to the same top supplier but, being normalized, it gives no indication of how much risk would remain after an intervention.

**Table 5.** Ablation of CRE (matched seeds 0–2)

| Variant | $\widehat{PN}$ AUC | MAE | Bias | Trace |
|---|---|---|---|---|
| CRE (CF-TGNN, δ=2, surgery) | 0.571 | 0.197 | -0.127 | 0.481 |
| w/o graph surgery | 0.550 | 0.252 | -0.200 | 0.453 |
| w/o region context (X-TGNN) | 0.517 | 0.210 | -0.135 | 0.483 |
| restoration window δ=1 | 0.573 | 0.197 | -0.128 | 0.473 |
| restoration window δ=4 | 0.569 | 0.201 | -0.130 | 0.479 |
| oracle PN (true hazards) | 0.833 | 0 | 0 | – |

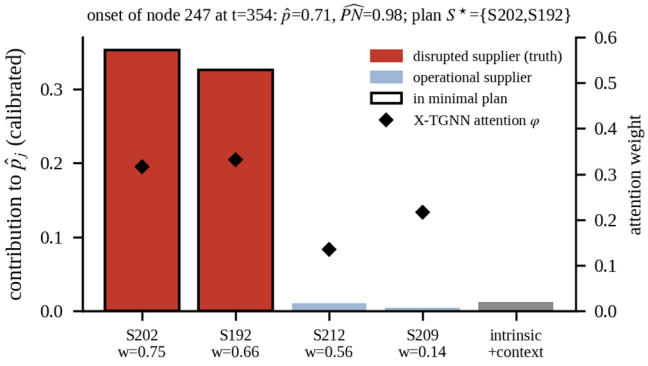



**Fig 5.** Counterfactual risk decomposition of one predicted onset: restoration Shapley values per supplier (bars), intrinsic and contextual residual (grey) and X-TGNN attention (diamonds). Outlined bars form the minimal restoration plan.

## 5.9. Real-World Case Study: SupplyGraph

To test the framework on real records we use SupplyGraph [15], which provides daily production, sales-order and delivery data for 41 FMCG products of one manufacturer over 221 days in 2023. A product is in shortfall when its three-day delivery-to-order ratio falls below half of its training-period median; the task is to predict next-day shortfalls (prevalence 9.1%, 331 onsets). Products of the same sub-group share their base product and production resources and are linked in both directions (96 edges); the plant of each product plays the role of the region. The split, models and hyperparameters are those of Section V-A, with five seeds. No causal ground truth exists here, so explanations are judged by model-level criteria.

Table 6 shows that all temporal models perform within about 0.02 onset AUC of each other. The sub-group graph gives a small gain in onset AP (X-TGNN 0.284 versus 0.266 for the GRU), while the plant context does not help: with 12 plants for 41 products, it mostly repeats what the graph already carries and adds variance. Table 7 summarizes CRE on the predicted onsets of the test period. For X-TGNN, 97% of them can be averted by restoring sister products, typically by restoring a single one ($|S^\star| = 1.12$). CF-TGNN attributes much more risk to plant-level conditions: only 54% of its predicted onsets pass the but-for test, against 88% for X-TGNN. Without ground truth we cannot tell which account is right, but the test makes the disagreement explicit, which attention alone would hide. On

both backbones the Shapley-ranked restoration removes more predicted risk than the attention-ranked one (Comp@1), consistent with Table 3.

**Table 6.** Next-day shortfall prediction on SupplyGraph (five seeds)

| Model | AUC-ROC | AUC-PR | Onset AUC | Onset AP |
|---|---|---|---|---|
| GRU | 0.922±0.001 | 0.573±0.015 | 0.882±0.003 | 0.266±0.006 |
| GRU+RC | 0.922±0.004 | 0.574±0.025 | 0.884±0.006 | 0.260±0.015 |
| GraphSAGE-T | **0.925±0.002** | **0.597±0.012** | **0.888±0.005** | 0.274±0.020 |
| X-TGNN | 0.919±0.004 | 0.574±0.025 | 0.879±0.006 | **0.284±0.027** |
| CF-TGNN | 0.912±0.010 | 0.557±0.022 | 0.866±0.018 | 0.266±0.027 |

**Table 7.** CRE on predicted SupplyGraph onsets (mean over seeds)

| Backbone | Events | Valid | \|S\| | $P\hat{N}\geq 0.5$ | Comp@1 rand. | att. | Shap. |
|---|---|---|---|---|---|---|---|
| X-TGNN | 50.4 | 0.97 | 1.12 | 0.88 | 0.57 | 0.62 | 0.63 |
| CF-TGNN | 58.6 | 0.67 | 1.28 | 0.54 | 0.39 | 0.41 | 0.45 |

## 6. DISCUSSION AND LIMITATIONS

For a planner, CF-TGNN turns a risk score into three answers: whether the predicted disruption is upstream-driven at all ($P\hat{N}$), which suppliers carry how much of the risk ($\varphi$) and the smallest set of suppliers to protect ($S^{\star}$). The origin tracer adds where the cascade started, which is where a mitigation budget has the widest effect. Several limitations remain. The causal metrics rely on a simulator; the testbed is grounded in the published risk-diffusion dynamics, but real disruptions involve inventories, capacity reallocation and contracts that a one-step hazard model omits. Avertability is only partly predictable because fragility is latent, so even a perfect predictor would not reach the oracle AUC of Fig. 2. The exact lattice grows as $2^n$; in-degrees above about 12 call for the attention-ordered or sampled variants described in Section IV-C. Finally, restorations assume that the state of a supplier can be acted on, which is realistic for expediting or safety stock but less so for long-term sourcing changes.

## 7. CONCLUSION

We introduced CF-TGNN, a counterfactual explainable temporal GNN for multi-tier supply chain disruption prediction. Building on the supplier-attention operator and the simulation philosophy of X-TGNN, it adds a region context that proxies confounding shocks and an explainer that reasons with restoration interventions rather than with weights. Turning the cascade simulator into a structural causal model let us check explanations against true counterfactuals. The results show that counterfactual reasoning adds what attention cannot provide: a but-for test, an exact minimal intervention and a temporally consistent origin trace, at the cost of a few batched model evaluations. Future work will scale the lattice search to dense networks, add inventory dynamics to the structural model, and validate the explanations against documented disruption events in industrial data.

# REFERENCES


[1] Ivanov, D., Sokolov, B., & Dolgui, A. (2014). The ripple effect in supply chains: Trade-off "efficiency-flexibility-resilience" in disruption management. International Journal of Production Research, 52(7), 2154–2172. https://doi.org/10.1080/00207543.2013.858836.

[2] Dolgui, A., Ivanov, D., & Sokolov, B. (2018). Ripple effect in the supply chain: An analysis and recent literature. International Journal of Production Research, 56(1–2), 414–430. https://doi.org/10.1080/00207543.2017.1387680.

[3] Inoue, H., & Todo, Y. (2019). Firm-level propagation of shocks through supply-chain networks. Nature Sustainability, 2, 841–847. https://doi.org/10.1038/s41893-019-0351-x.

[4] Ivanov, D. (2020). Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case. Transportation Research Part E: Logistics and Transportation Review, 136, 101922. https://doi.org/10.1016/j.tre.2020.101922.

[5] Baryannis, G., Dani, S., & Antoniou, G. (2019). Predicting supply chain risks using machine learning: The trade-off between performance and interpretability. Future Generation Computer Systems, 101, 993–1004. https://doi.org/10.1016/j.future.2019.07.059.

[6] Wyrembek, M., Baryannis, G., & Brintrup, A. (2025). Causal machine learning for supply chain risk prediction and intervention planning. International Journal of Production Research, 63(15), 5629–5648. https://doi.org/10.1080/00207543.2025.2458121.

[7] Kosasih, E. E., & Brintrup, A. (2022). A machine learning approach for predicting hidden links in supply chain with graph neural networks. International Journal of Production Research, 60(17), 5380–5393. https://doi.org/10.1080/00207543.2021.1956697.

[8] Wang, Y., Bian, M., & Lin, Y. (2022). An explainable temporal graph neural network for disruption prediction and risk propagation in multi-tier supply chains. Journal of Computational Science and Electrical Engineering, 4(1), 17–25. https://doi.org/10.61784/jcsee4151.

[9] Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.

[10] Zhou, F. (2026). Moment-constrained side-force allocation for wind-assisted merchant ships: Hydrodynamic performance of retractable anti-leeway fins and flap rudders. Innovation and Technology Studies, 3(1), 27–41.

[11] Lucic, A., ter Hoeve, M. A., Tolomei, G., de Rijke, M., & Silvestri, F. (2022). CF-GNNExplainer: Counterfactual explanations for graph neural networks. Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, PMLR, 151, 4499–4511.

[12] Prado-Romero, M. A., Prenkaj, B., Stilo, G., & Giannotti, F. (2024). A survey on graph counterfactual explanations: Definitions, methods, evaluation, and research challenges. ACM Computing Surveys, 56(7), 171. https://doi.org/10.1145/3618105.

[13] Basole, R. C., & Bellamy, M. A. (2014). Supply network structure, visibility, and risk diffusion: A computational approach. Decision Sciences, 45(4), 753–789. https://doi.org/10.1111/deci.12099.

[14] Ivanov, D., & Dolgui, A. (2021). A digital supply chain twin for managing the disruption risks and resilience in the era of Industry 4.0. Production Planning & Control, 32(9), 775–788. https://doi.org/10.1080/09537287.2020.1768450.

[15] Wasi, A. T., Islam, M. S., & Akib, A. R. (2024). SupplyGraph: A benchmark dataset for supply chain planning using graph neural networks. arXiv:2401.15299.

[16] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations.

[17] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. International Conference on Learning Representations.

[18] Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in Neural Information Processing Systems, 30, 1024–1034.

[19] Pareja, A., et al. (2020). EvolveGCN: Evolving graph convolutional networks for dynamic graphs. Proceedings of the AAAI Conference on Artificial Intelligence, 34(4), 5363–5370. https://doi.org/10.1609/aaai.v34i04.5984.

[20] Yue, X., & Zhu, R. (2024). An explainable hybrid machine learning framework for cash-flow-conditioned multi-horizon liquidity risk early warning: An SME-oriented benchmark study. World Journal of Management Science, 2(1), 63–72.

[21] Cho, K., et al. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 1724–1734. https://doi.org/10.3115/v1/D14-1179.

[22] Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. Advances in Neural Information Processing Systems, 32, 9240–9251.

[23] Yuan, H., Yu, H., Gui, S., & Ji, S. (2023). Explainability in graph neural networks: A taxonomic survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(5), 5782–5799. https://doi.org/10.1109/TPAMI.2022.3204236.

[24] Xia, W., Lai, M., Shan, C., Zhang, Y., Dai, X., Li, X., & Li, D. (2023). Explaining temporal graph models through an explorer-navigator framework. International Conference on Learning Representations.

[25] Tan, J., Geng, S., Fu, Z., Ge, Y., Xu, S., Li, Y., & Zhang, Y. (2022). Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning. Proceedings of the ACM Web Conference, 1018–1027. https://doi.org/10.1145/3485447.3511948.

[26] Jain, S., & Wallace, B. C. (2019). Attention is not explanation. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 3543–3556. https://doi.org/10.18653/v1/N19-1357.

[27] Wang, J., Ge, Y., Liu, J., & Jiang, B. (2024). Early warning of liquidity stress in U.S. banks with seasonally aligned frailty hazards and delayed-feedback conformal alarms. Journal of Trends in Finance and Economics, 1(1), 85–100.

[28] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[29] Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161.

[30] Tian, J., & Pearl, J. (2000). Probabilities of causation: Bounds and identification. Annals of Mathematics and Artificial Intelligence, 28, 287–313. https://doi.org/10.1023/A:1018912507879.

[31] Li, P., Bian, M., & Lin, Y. (2026). Carbon-aware offline reinforcement learning for joint inventory replenishment and transportation scheduling in multi-echelon supply chains. Highlights in Business, Economics and Management, 69, 251–262.

[32] Liu, J., Wang, J., & Lin, H. (2025). Coordinated physics-informed multi-agent reinforcement learning for risk-aware supply chain optimization. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3629716.