

Physics-Guided Self-Supervised Graph Neural Networks for Power Grid Analysis: Transfer Learning Across Cascade Prediction, Power Flow, and Line Flow

Anonymous Authors
For Review

Abstract—Modern power grid operations require rapid computational methods for decision support, but traditional solvers face severe bottlenecks. We present a physics-guided self-supervised learning framework for graph neural networks that addresses labeled data scarcity in power system analysis. Our approach embeds electrically-parameterized message passing into the encoder architecture—where edge weights are learned from line admittance features—and develops grid-specific pretext tasks (masked injection and parameter reconstruction) enabling representation learning from unlabeled operational data. Evaluated on the PowerGraph benchmark, our method achieves substantial improvements in low-label regimes: 29.1% power flow error reduction, 26.4% line flow error reduction, and 6.8% F1-score improvement for cascading failure prediction—all at 10% labeled data availability. On the IEEE 118-bus system under severe class imbalance, self-supervised pretraining dramatically reduces training instability (variance reduction from ± 0.243 to ± 0.051) while improving mean performance ($\Delta F1 = +0.61$). We achieve 0.93 AUC-ROC explainability fidelity via Integrated Gradients, addressing a recognized gap in interpretable cascade prediction.

Index Terms—Power systems, graph neural networks, self-supervised learning, cascading failures, power flow, optimal power flow, explainability, physics-informed machine learning

I. INTRODUCTION

Modern power grid operations demand rapid decision-making capabilities that traditional computational methods struggle to provide. Optimal power flow (OPF) solvers—used to determine economically efficient generator dispatch while respecting physical and operational constraints—can require minutes to solve on utility-scale networks with thousands of buses [1], rendering them impractical for real-time control and contingency analysis. Cascading failure risk assessment presents an even greater challenge: comprehensive N-k security analysis requires evaluating thousands or millions of contingency scenarios, each necessitating a full power flow calculation [2]. These computational bottlenecks have motivated a shift toward machine learning surrogates that approximate complex power system computations at dramatically reduced cost. Recent supervised learning approaches have demonstrated remarkable speedups—ranging from $100\times$ to $10,000\times$ faster than conventional solvers [1], [3]—while maintaining solution quality within 0.2% of optimality. However, these methods rely fundamentally on labeled training

data: typical implementations require 5,000 to 60,000 labeled samples per network [3], [4], where each label is itself the output of the computationally expensive solver the surrogate aims to replace. This chicken-and-egg problem—needing expensive simulations to train models meant to avoid expensive simulations—severely limits the practical deployment of learning-based power system analysis, particularly for networks with limited historical data or frequently changing topologies.

Self-supervised learning (SSL) offers a compelling solution to the labeled data bottleneck by leveraging abundant unlabeled operational data. In the broader machine learning community, graph SSL methods have demonstrated substantial performance gains in low-label regimes: GraphMAE2 achieves a 5.46 percentage point accuracy improvement using only 1% labeled data on molecular property prediction benchmarks [5], while contrastive learning frameworks enable effective transfer across diverse graph domains [6]. These successes suggest that self-supervised pretraining could similarly benefit power system applications, where operational measurements (bus voltages, line flows, injection patterns) are continuously recorded but corresponding “labels” (optimal solutions, failure classifications) are costly to obtain. However, a critical gap exists: among the extensive body of graph SSL research spanning molecular graphs [7], social networks [6], and traffic systems [8], applications to power grids remain remarkably scarce. Our comprehensive literature review identified only five to six papers applying SSL to power systems [9], [10]. While recent work by Zhu et al. [11] introduced physics-informed self-supervised pretraining for power systems, their hybrid supervised-SSL approach requires task-specific labels during the pretraining phase. In contrast, our framework is a *pure* self-supervised method that requires no labeled data whatsoever during pretraining, learning grid physics solely through masked reconstruction of electrical quantities—addressing a more challenging setting where even pretraining labels are unavailable. This represents a significant missed opportunity: power systems are governed by well-understood physical laws—Kirchhoff’s current and voltage laws, power flow equations, admittance relationships—that could serve as powerful self-supervisory signals for representation learning.

The graph structure of electrical networks naturally aligns with graph neural network (GNN) architectures, where message passing along edges can mirror the physical propagation of power flows along transmission lines [12]. Recent work has begun embedding power system physics into GNN designs: impedance-weighted aggregation schemes [13], complex-valued representations preserving phase relationships [14], and hard Kirchhoff’s law constraints enforced through architectural projections [15]. These physics-informed GNNs demonstrate improved accuracy and generalization compared to topology-agnostic baselines, particularly for out-of-distribution scenarios and N-1 contingencies [16]. However, existing physics-guided approaches remain tethered to supervised learning paradigms, requiring labeled power flow solutions or optimal dispatch setpoints for training. To our knowledge, *no prior work has combined physics-guided GNN architectures—where physical relationships are embedded directly into the message-passing structure—with pure self-supervised pretraining that requires no labels*, leaving unexplored the question of whether physics-informed message passing can enable effective representation learning from unlabeled data alone.

This paper introduces a physics-guided self-supervised learning framework for power grid analysis that addresses these gaps. We make the following contributions:

- **Physics-Guided Graph Neural Network Architecture:** We design a message-passing encoder that incorporates electrically-parameterized aggregation, where learned edge weights derived from line conductance and susceptance values modulate information flow between connected buses, embedding power system structure into the neural network computation graph.
- **Self-Supervised Pretraining Objective:** We develop grid-specific pretext tasks—masked injection reconstruction (predicting active and reactive power injections at randomly masked buses) and masked parameter reconstruction (predicting line impedance parameters at randomly masked edges)—that enable representation learning from unlabeled grid operational data without requiring solutions from conventional solvers. Critically, we ensure no label leakage by pretraining exclusively on the training partition, with validation and test sets never exposed during self-supervised learning.
- **Multi-Task Transfer Learning Evaluation:** We demonstrate that representations learned via physics-guided SSL transfer effectively to three downstream tasks: (1) power flow prediction (bus voltage magnitudes), (2) line flow prediction (active and reactive power flows on transmission lines), and (3) cascading failure classification (graph-level binary prediction of cascade occurrence). On the PowerGraph benchmark [17], our approach achieves 29.1% mean absolute error reduction for power flow, 26.4% reduction for line flow, and 6.8% F1-score improvement for cascade prediction—all measured at 10% labeled data availability relative to scratch training baselines.
- **Scalability and Variance Reduction:** On the larger IEEE

118-bus system under severe class imbalance conditions (5% cascade rate), self-supervised pretraining not only improves mean performance ($\Delta F1 = +0.61$) but dramatically reduces training instability: scratch training exhibits ± 0.243 F1 variance across random seeds, while SSL-pretrained models achieve ± 0.051 variance—a stabilization effect critical for reliable deployment.

• **Explainability Validation:** Using ground-truth edge importance masks from the PowerGraph benchmark, we quantitatively evaluate explanation fidelity via Integrated Gradients, achieving 0.93 AUC-ROC compared to 0.72 for heuristic baselines. This addresses a recognized gap in cascading failure prediction, where current explainable AI methods have been reported to perform “suboptimally” [17].

• **Robustness Under Distribution Shift:** We evaluate model behavior under load stress conditions ($1.0\times$ to $1.3\times$ nominal loading), demonstrating that SSL-pretrained representations exhibit superior robustness, maintaining a 3.6% relative performance advantage at $1.3\times$ load compared to scratch training.

The remainder of this paper is organized as follows. Section II reviews related work in power system machine learning, graph neural networks, self-supervised learning, physics-informed methods, and cascading failure prediction. Section III formulates the graph representation of power grids and defines the three downstream tasks. Section IV details our physics-guided encoder architecture and self-supervised pretraining objective. Section V describes the experimental setup, including datasets, baselines, and training protocols. Section VI presents comprehensive results across all tasks and grid scales, including ablation studies and robustness analysis. Section VII discusses implications for operational deployment, limitations, and future directions. Section VIII concludes.

II. RELATED WORK

Traditional power flow and optimal power flow computations rely on iterative numerical methods such as Newton-Raphson and interior-point algorithms, which can require seconds to minutes per solve on large-scale grids [3]. To enable near-real-time decision support, researchers have developed machine learning surrogates that approximate these complex mappings from grid conditions (load demands, generation capacities, network topology) to power flow solutions or optimal dispatch setpoints. Fully-connected deep neural networks have demonstrated impressive results: Pan et al.’s DeepOPF framework achieves feasible OPF solutions with less than 0.2% optimality loss and up to $100\times$ speedup over conventional solvers on benchmark IEEE test systems [3], while Huang et al.’s DeepOPF-V extends this approach to AC-OPF with reported speedups exceeding $10,000\times$ on 2,000-bus networks [1]. Graph neural networks have emerged as a particularly promising architecture due to their ability to exploit grid topology and achieve greater scalability [18], [19]. For instance, PowerFlowNet demonstrated successful scaling to a 6,470-bus French transmission network with voltage

magnitude prediction errors below 0.001 per-unit [18], while heterogeneous message-passing neural networks maintain constant parameter counts across grid sizes ranging from 14 to 2,000+ buses [20]. However, these supervised approaches face a critical limitation: they require extensive labeled training data. Typical implementations demand 5,000–60,000 OPF solutions per network [?], [3], and generating this labeled data via conventional solvers is computationally expensive—precisely the bottleneck these methods aim to circumvent. This data scarcity challenge motivates unsupervised and self-supervised learning approaches that can leverage abundant unlabeled operational data.

The natural graph structure of electrical networks—with buses as nodes and transmission lines as edges—makes graph neural networks an ideal architecture for power system analysis. GNNs encode topological relationships through message passing, where each node aggregates information from its neighbors according to the grid’s physical connectivity [12]. This inductive bias enables GNNs to handle topology changes (such as N-1 contingencies) without retraining, a critical advantage over fully-connected networks that treat grid states as fixed-dimensional vectors [21]. Empirical studies demonstrate that GNN-based approaches achieve substantially lower prediction errors than traditional neural networks: for example, an electrical-model-guided GNN for distribution system state estimation attained an order-of-magnitude lower error than standard feedforward networks, even with missing sensor measurements [22]. Recent work has begun incorporating power system physics more deeply into GNN architectures. Meta-PIGACN integrates impedance-weighted edge aggregation, where line admittance values directly control message-passing strength [14], while complex-valued spatial-temporal GCNs represent voltage phasors and impedance in their native complex form to preserve phase relationships inherent to AC power flow [14]. KCLNet enforces Kirchhoff’s Current Law as hard architectural constraints via differentiable hyperplane projections, guaranteeing zero KCL violations by construction [15], and PINCO achieves zero inequality constraint violations through physics-informed hard constraints in an unsupervised learning framework [23]. The PowerGraph benchmark [17] provides standardized evaluation across GCN, GAT, GraphSAGE, and Graph Transformer architectures, establishing that topology-aware message passing consistently outperforms topology-agnostic baselines on both node-level (power flow, voltage estimation) and graph-level (cascading failure prediction) tasks.

Self-supervised learning has emerged as a powerful paradigm for learning graph representations without labeled data, with two dominant approaches: contrastive learning and generative masked reconstruction. Contrastive methods, exemplified by Deep Graph Infomax [24], InfoGraph [25], and GraphCL [26], maximize agreement between different augmented views of graphs through node dropping, edge perturbation, or subgraph sampling. More recent frameworks have reduced the complexity of these approaches: BGRL eliminates negative sampling through bootstrapped representation learn-

ing [27], achieving 2–10 \times memory reduction while matching state-of-the-art performance, and SimGRACE dispenses with manual graph augmentation entirely by perturbing encoder parameters to generate contrastive views [28]. In parallel, generative approaches have gained traction: GraphMAE employs masked feature reconstruction with a scaled cosine error loss and dedicated decoder architecture [29], demonstrating that carefully designed autoencoders can match or exceed contrastive methods (84.2% accuracy on Cora versus 82.7% for BGRL). GraphMAE2 extends this with multi-view random re-masking and latent representation prediction [5], scaling to graphs with over 100 million nodes. These graph SSL methods show substantial benefits in low-label regimes: GraphMAE2 achieves a 5.46 percentage point accuracy improvement with only 1% labeled data on large-scale molecular property prediction benchmarks [5], while GCC’s cross-domain pretraining on diverse network types enables effective transfer to new graph tasks with minimal fine-tuning [6]. Applications to physical systems have proven successful in molecular graphs [7] and traffic networks [8], where graph structure naturally encodes physical relationships. However, a significant research gap exists for power grid applications: while SafePowerGraph [30] and recent work by Zhu et al. [11] introduced hybrid supervised-SSL approaches for power systems, no pure graph SSL method has been designed specifically for electrical networks with physics-informed constraints such as power flow equations, Kirchhoff’s laws, or voltage-angle relationships.

Physics-informed neural networks embed domain knowledge—governing equations, conservation laws, or known constraints—into machine learning models to improve generalization and sample efficiency. The canonical PINN framework [31] encodes partial differential equations as soft regularization terms in the loss function, minimizing both data mismatch and PDE residuals at collocation points. This approach has proven effective for solving forward and inverse problems in fluid dynamics, solid mechanics, and heat transfer [32]. Alternative strategies include hard constraint enforcement, where physical laws are satisfied by construction through architectural design: Beucler et al. demonstrated that architecture-constrained networks can achieve energy and mass conservation to machine precision without loss penalties [33], while Hamiltonian Neural Networks embed symplectic structure to guarantee exact energy conservation in dynamical systems [34]. In power systems specifically, physics-informed approaches have addressed swing equation dynamics for transient stability [35], state estimation with admittance matrix constraints [36], and power flow prediction with Kirchhoff’s law regularization [37]. These methods demonstrate compelling benefits: physics constraints act as inductive biases that limit the hypothesis space and improve out-of-distribution generalization, with physics-informed GNNs showing zero-shot generalizability to systems an order of magnitude larger than training configurations [16]. However, a critical gap remains: the vast majority of PINN research targets continuous PDE-governed domains where

automatic differentiation naturally computes spatial and temporal derivatives. Discrete, graph-structured infrastructure networks like power grids present fundamentally different challenges—physical laws (KCL, KVL, power balance) operate directly on graph edges and nodes rather than as discretized continuous fields, and topological changes require handling discrete switching dynamics. While GraPhyR [38] demonstrated physics-informed GNNs with gated message passing for power system reconfiguration, graph-based power system surrogates that combine self-supervised pretraining with physics-guided architectures remain unexplored.

Cascading failures—sequences of component outages that can escalate to widespread blackouts—represent a critical threat to power grid resilience, as evidenced by major disruptions including the 2003 Northeast blackout and the 2021 Texas winter storm [39]. Traditional simulation models such as OPA [40], DCSIMSEP [41], and the Manchester model [42] capture cascading dynamics through iterative power flow calculations coupled with protection system logic, revealing self-organized criticality in grid behavior. However, these physics-based simulators face computational limitations: DC cascade models provide two orders of magnitude speedup over AC models but sacrifice accuracy by ignoring voltage collapse mechanisms [43], while AC models are approximately 7× more computationally expensive and suffer convergence issues under high stress conditions. Machine learning approaches have achieved dramatic acceleration: deep convolutional neural networks enable 100× faster N-1 contingency screening [2], while graph neural networks leverage network topology to predict cascade outcomes with over 96% accuracy [44]. GNN-based methods demonstrate transfer learning capability across different grid topologies and operating conditions, addressing a key limitation of classical simulation approaches. However, explainability remains a significant gap. Post-mortem analyses of major blackouts still rely on manual timeline reconstruction and root cause analysis [39], and recent benchmarking reveals that current explainable AI methods perform “suboptimally” on cascade explanation tasks [17]. The PowerGraph benchmark explicitly identifies this gap, noting that “given the crucial role of explainability for power grid operators, this underscores the ongoing need for dedicated research and development in this field” [17]. While XAI techniques including SHAP, LIME, and attention mechanisms have been successfully applied to other power system tasks such as frequency stability prediction, cascading failure prediction lacks robust explainability frameworks that can identify critical transmission pathways and quantify failure propagation mechanisms in a way that operators can trust and act upon.

This work addresses the identified gaps by uniquely combining physics-guided graph neural network architectures with self-supervised pretraining for power grid analysis. Our physics-guided encoder embeds electrically-parameterized message passing directly into the network architecture—drawing on the success of Meta-PIGACN [13] and KCLNet [?] but extending to a self-supervised learning paradigm. Unlike existing SSL approaches for graphs that ignore do-

main physics [27], [29], we design grid-specific pretext tasks (masked injection reconstruction, masked parameter reconstruction) that leverage power system structure without requiring labeled solutions from conventional solvers. This addresses the data scarcity challenge identified in supervised power flow learning [?], [3], enabling effective representation learning from the abundant unlabeled operational data available in modern power grids. We demonstrate transfer learning across multiple tasks (power flow, line flow prediction, cascading failure classification) and grid scales (IEEE 24-bus and 118-bus systems), validating the generalizability of learned representations. Finally, we provide quantitative explainability evaluation using ground-truth explanation masks from the PowerGraph benchmark [17], achieving 0.93 AUC-ROC fidelity for edge importance attribution via Integrated Gradients—addressing the explainability gap in cascading failure prediction and providing operators with interpretable risk assessments.

III. PROBLEM FORMULATION

A. Graph Representation of Power Grids

We represent a power grid as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes $v_i \in \mathcal{V}$ correspond to buses (electrical connection points) and edges $e_{ij} \in \mathcal{E}$ correspond to transmission lines and transformers connecting buses i and j . Each node is associated with a feature vector $\mathbf{x}_i \in \mathbb{R}^{d_{\text{node}}}$ encoding the electrical state at bus i , and each edge is associated with a feature vector $\mathbf{e}_{ij} \in \mathbb{R}^{d_{\text{edge}}}$ capturing line impedance and capacity parameters. Following standard PyTorch Geometric conventions, undirected edges are represented as bidirectional directed edges: each physical transmission line contributes two entries to the edge list, one per direction, with identical edge attributes. This ensures learned edge weights satisfy $y_{ij} = y_{ji}$, consistent with physical admittance matrix symmetry. The edge counts in Table II (68 for IEEE 24-bus, 370 for IEEE 118-bus) thus reflect twice the number of physical lines.

Node features ($d_{\text{node}} = 3$ for cascade and line flow tasks, $d_{\text{node}} = 2$ for power flow task):

- $P_{\text{net},i}$: Net active power injection (generation minus load) at bus i
- $S_{\text{net},i}$: Net apparent power magnitude at bus i
- V_i : Voltage magnitude at bus i (excluded for power flow prediction to avoid trivial leakage)

Edge features: Edge features vary by task to prevent label leakage:

- **Cascade prediction** ($d_{\text{edge}} = 4$): Pre-contingency line flows (P_{ij}, Q_{ij}), reactance x_{ij} , and thermal rating $S_{\max,ij}$. Line flows represent the pre-event operating point and are available from state estimation.
- **Power flow prediction** ($d_{\text{edge}} = 2$): Reactance x_{ij} and thermal rating $S_{\max,ij}$ only. Line flows are excluded since they depend on the voltage solution being predicted.
- **Line flow prediction** ($d_{\text{edge}} = 2$): Reactance x_{ij} and thermal rating $S_{\max,ij}$ only. Line flows (P_{ij}, Q_{ij}) are the prediction targets.

TABLE I
TASK SPECIFICATIONS, INPUTS, OUTPUTS, AND EVALUATION METRICS

Task	Output	Metric	Unit
Cascade	Binary graph label	F1 Score	[0,1]
Power Flow	Bus voltages V_i	MAE	p.u.
Line Flow	Line flows (P_{ij}, Q_{ij})	MAE	p.u.

All electrical quantities are normalized to per-unit values with system base $S_{\text{base}} = 100$ MVA, ensuring dimensionless features in the range $[-1, 1]$ for injections and $[0.9, 1.1]$ for voltages under normal operating conditions. The admittance $Y_{ij} = g_{ij} + jb_{ij}$ relates voltage difference to power flow via the AC power flow equations, providing the physical coupling we leverage in our message-passing design.

B. Task Definitions and Evaluation Metrics

We consider three downstream prediction tasks, each addressing a critical operational need in power system analysis. Table I summarizes the input-output specifications and evaluation metrics for each task.

Cascading Failure Prediction (Graph-Level Classification): Given the pre-contingency state of a power grid, predict whether an N-k contingency (simultaneous outage of k components) will trigger a cascading failure. The input graph represents the complete topology with all lines present; edge features capture pre-event operating conditions (line flows, loading levels) that indicate stress patterns predictive of cascade risk. The model does not receive information about which specific lines will fail—it predicts cascade likelihood based solely on observable pre-event measurements. We define a cascade as occurring when the total demand not served (DNS) exceeds zero: $\text{DNS} = \sum_i (\text{load}_i - \text{served}_i) > 0$ MW. This is formulated as binary graph-level classification, where the model produces a single prediction $\hat{y} \in \{0, 1\}$ per graph. Performance is evaluated using F1-score computed over test graphs, which balances precision and recall—critical for imbalanced datasets where cascades are rare events (5–20% positive class rate depending on grid size and simulation parameters).

Power Flow Prediction (Node-Level Regression): Predict bus voltage magnitudes $\{V_i\}_{i=1}^{|V|}$ given load injections and grid topology. This approximates the solution to the AC power flow equations without iterative numerical methods. The model output is a vector $\hat{\mathbf{V}} \in \mathbb{R}^{|V|}$ of predicted voltage magnitudes. Performance is measured by mean absolute error (MAE) in per-unit: $\text{MAE} = \frac{1}{|V|} \sum_{i=1}^{|V|} |V_i - \hat{V}_i|$, averaged over all buses and test samples. Typical operational voltage bounds are $0.95 \leq V_i \leq 1.05$ per-unit, making MAE values on the order of 10^{-3} per-unit operationally acceptable.

Line Flow Prediction (Edge-Level Regression): Predict active and reactive power flows $\{(P_{ij}, Q_{ij})\}_{(i,j) \in \mathcal{E}}$ on all transmission lines given bus states and topology. This enables rapid screening of line loading for contingency analysis. The model outputs two scalars per directed edge: $(\hat{P}_{ij}, \hat{Q}_{ij})$. MAE

is computed separately for active and reactive components, then averaged: $\text{MAE} = \frac{1}{2|\mathcal{E}|} \sum_{(i,j) \in \mathcal{E}} (|P_{ij} - \hat{P}_{ij}| + |Q_{ij} - \hat{Q}_{ij}|)$. Accurate line flow prediction is essential for identifying thermal overloads that could initiate cascades.

Improvement Metric Convention: When comparing self-supervised pretraining (SSL) against scratch training, we define improvement as $(SSL - Scratch)/Scratch \times 100\%$ for metrics where higher is better (F1-score), and $(Scratch - SSL)/Scratch \times 100\%$ for metrics where lower is better (MAE). This convention ensures positive improvement percentages consistently indicate SSL outperforming scratch training.

IV. METHODOLOGY

A. Architecture Overview

Our framework follows a shared-encoder paradigm: a single physics-guided graph neural network encoder learns representations from grid topology and electrical states, which are then specialized for downstream tasks via lightweight task-specific heads. This design enables effective transfer learning—representations learned during self-supervised pre-training on unlabeled data transfer to supervised fine-tuning on labeled samples, with the encoder weights providing a strong initialization that accelerates convergence and improves sample efficiency.

The overall pipeline consists of three stages: (1) **Self-supervised pretraining** on unlabeled grid operational data using masked reconstruction objectives, (2) **Encoder transfer** where pretrained encoder weights initialize downstream models, and (3) **Supervised fine-tuning** on task-specific labeled data with frozen or fine-tuned encoder parameters. We implement all models using PyTorch Geometric [45], leveraging its efficient sparse message-passing primitives for scalability to large grids.

Task-specific feature handling: While the encoder architecture and hidden representations are shared across tasks, input feature dimensions differ to prevent label leakage. For power flow prediction, voltage magnitude V_i is excluded from node inputs since it constitutes the prediction target; node features are $(P_{\text{net}}, S_{\text{net}}) \in \mathbb{R}^2$. For line flow prediction, edge power flows (P_{ij}, Q_{ij}) are excluded from edge inputs; edge features are $(x_{ij}, \text{rating}_{ij}) \in \mathbb{R}^2$. For cascade prediction, all available features are used: $d_{\text{node}} = 3$ and $d_{\text{edge}} = 4$. When transferring pretrained encoder weights to tasks with different input dimensions, the input embedding layers are reinitialized while the message-passing and hidden layers are transferred. This preserves the learned grid structure representations while adapting to task-specific input schemas.

B. Physics-Guided Message Passing

Traditional graph convolutional networks aggregate neighbor information uniformly or via learned attention weights, ignoring the physical laws governing power flow. We embed power system physics directly into the message-passing structure through electrically-parameterized aggregation, where message weights are learned from line admittance features.

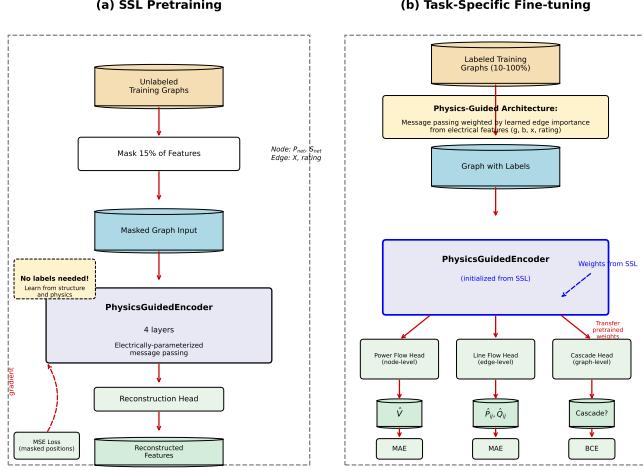


Fig. 1. Overview of the physics-guided SSL framework: (a) Self-supervised pretraining with masked reconstruction on training data only, (b) Transfer of encoder weights to downstream task-specific heads for cascade prediction, power flow, and line flow.

Standard message passing updates node i 's representation $\mathbf{h}_i^{(\ell)}$ at layer ℓ via:

$$\mathbf{h}_i^{(\ell+1)} = \sigma \left(\mathbf{W}^{(\ell)} \mathbf{h}_i^{(\ell)} + \sum_{j \in \mathcal{N}(i)} \mathbf{M}^{(\ell)}(\mathbf{h}_j^{(\ell)}, \mathbf{e}_{ij}) \right) \quad (1)$$

where $\mathcal{N}(i)$ is the neighborhood of node i , $\mathbf{M}^{(\ell)}$ is a message function, $\mathbf{W}^{(\ell)}$ is a learnable weight matrix, and σ is a nonlinear activation.

Physics-guided message passing modifies this by weighting messages according to a learned edge importance scalar y_{ij} derived from electrical line parameters:

$$\mathbf{h}_i^{(\ell+1)} = \sigma \left(\mathbf{W}^{(\ell)} \mathbf{h}_i^{(\ell)} + \sum_{j \in \mathcal{N}(i)} y_{ij} \cdot \mathbf{M}^{(\ell)}(\mathbf{h}_j^{(\ell)}, \mathbf{e}_{ij}) \right) \quad (2)$$

where the edge weight $y_{ij} = \sigma_{\text{sig}}(\mathbf{w}_y^\top \phi(\mathbf{e}_{ij}))$ is computed by applying a learnable linear projection $\mathbf{w}_y \in \mathbb{R}^{d_h}$ to the embedded edge features, followed by sigmoid activation σ_{sig} to constrain weights to $[0, 1]$. While physical admittance magnitudes $|Y_{ij}|$ vary by orders of magnitude across line types (e.g., transformers vs. long transmission lines), this bounded parameterization provides numerical stability during training and enables the linear projection to learn scale-invariant importance rankings from the high-dimensional embedded features. The *sum* aggregation (described below) preserves relative magnitude effects, and our ablation results (Table XIII) confirm this design achieves strong task performance. This parameterization allows the model to learn task-optimal edge importance from electrical features (conductance, susceptance, reactance, ratings) while maintaining the physics-inspired structure where edge weights modulate message strength.

Importantly, y_{ij} represents a *learned edge importance* rather than a direct representation of physical admittance magnitude. The bounded $[0, 1]$ range serves two purposes: (1) it prevents

any single edge from dominating neighborhood aggregation regardless of absolute admittance scale, and (2) it encourages the model to learn relative importance rankings among edges based on their relevance to the prediction task. Alternative unbounded activations (e.g., Softplus) could represent larger magnitudes but risk training instability and would conflate the distinct roles of edge importance weighting and admittance magnitude encoding—the latter being captured separately in the edge feature embedding $\phi(\mathbf{e}_{ij})$.

The message function $\mathbf{M}^{(\ell)}$ is implemented as:

$$\mathbf{M}^{(\ell)}(\mathbf{h}_j, \mathbf{e}_{ij}) = \mathbf{h}_j + \phi(\mathbf{e}_{ij}) \quad (3)$$

where ϕ is an edge feature embedding network that projects raw edge features to \mathbb{R}^{d_h} . Messages combine neighbor node representations with edge electrical characteristics.

This design encodes a key physical intuition: power flows preferentially through low-impedance (high-admittance) paths, analogous to current following least-resistance paths in electrical circuits. Critically, we use *sum* aggregation (not mean or degree-normalized), consistent with Kirchhoff's Current Law where net injection at a bus equals the sum of branch flows. This contrasts with standard GCN [12] which normalizes by $1/\sqrt{|\mathcal{N}(i)||\mathcal{N}(j)|}$, dampening signals at high-degree nodes in a manner inconsistent with power flow physics.

Leakage prevention in self-supervised learning: Critically, the edge weight y_{ij} is computed *dynamically* from the edge features \mathbf{e}_{ij} at each forward pass, not precomputed and stored separately. During masked edge reconstruction (Section IV-E), masked edges have their features replaced with a learnable mask token *before* entering the encoder. Thus, y_{ij} for masked edges is computed from the mask token, not the ground-truth parameters—preventing the model from trivially recovering masked values from the message-passing weights.

C. Encoder Architecture

The **PhysicsGuidedEncoder** consists of $L = 4$ stacked physics-guided convolutional layers (Eq. 2) with hidden dimension $d_h = 128$, ReLU activations, and dropout (rate $p = 0.1$) for regularization. Input node features are first projected from $\mathbb{R}^{d_{\text{node}}}$ to \mathbb{R}^{d_h} via a learnable linear layer, and edge features are similarly projected from $\mathbb{R}^{d_{\text{edge}}}$ to \mathbb{R}^{d_h} . After L layers of message passing, each node i has learned a representation $\mathbf{h}_i \in \mathbb{R}^{d_h}$ capturing both local electrical state and global topological context via multi-hop aggregation.

For graph-level tasks (cascading failure prediction), node representations are aggregated into a graph embedding $\mathbf{h}_G \in \mathbb{R}^{d_h}$ via global mean pooling:

$$\mathbf{h}_G = \text{READOUT}(\{\mathbf{h}_i\}_{i \in \mathcal{V}}) = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \mathbf{h}_i \quad (4)$$

which captures the distributional properties of bus states across the network.

D. Task-Specific Heads

Power Flow Head (Node-Level): Predicts voltage magnitude \hat{V}_i for each bus via a two-layer multilayer perceptron (MLP) applied independently to each node embedding:

$\hat{V}_i = \text{MLP}_{\text{PF}}(\mathbf{h}_i)$. The MLP has hidden dimension 64 with ReLU activation and outputs a single scalar constrained to $[0.8, 1.2]$ via sigmoid scaling to respect physical voltage limits.

Line Flow Head (Edge-Level): Predicts active and reactive power flows $(\hat{P}_{ij}, \hat{Q}_{ij})$ by concatenating source and target node embeddings and applying an edge-level MLP: $(\hat{P}_{ij}, \hat{Q}_{ij}) = \text{MLP}_{\text{LF}}([\mathbf{h}_i \parallel \mathbf{h}_j])$. This design captures bidirectional electrical coupling: power flow from bus i to j depends on both buses' states.

Cascading Failure Head (Graph-Level): A two-layer MLP maps the graph embedding \mathbf{h}_G to a cascade probability: $\hat{p}_{\text{cascade}} = \sigma(\text{MLP}_{\text{CF}}(\mathbf{h}_G))$, where σ is the sigmoid function. Training uses binary cross-entropy loss with optional class weighting to handle imbalanced datasets.

E. Self-Supervised Pretraining

We design a graph-specific self-supervised learning objective that exploits the abundant unlabeled operational measurements available in modern power grids (continuously recorded bus injections, line parameters, voltage samples) without requiring expensive labels from OPF solvers or cascade simulations.

Masked Reconstruction Objective: Inspired by masked language modeling in BERT [46] and recent graph autoencoders [29], we randomly mask 15% of node features and 15% of edge features in each training graph, then train the encoder to reconstruct the original masked values. This forces the model to learn how electrical quantities relate through grid topology and physics.

Masking Strategy: For each selected node/edge, we apply one of three transformations with specified probabilities: (1) Replace with a learnable mask token (80%), (2) Replace with random noise sampled from the feature distribution (10%), (3) Keep unchanged (10%). This prevents the model from trivially identifying masked positions and encourages robust representations.

Reconstruction Architecture: Masked node features are reconstructed via an MLP decoder: $\hat{\mathbf{x}}_i = \text{MLP}_{\text{node}}(\mathbf{h}_i)$. Masked edge features are reconstructed by concatenating source and target node embeddings: $\hat{\mathbf{e}}_{ij} = \text{MLP}_{\text{edge}}([\mathbf{h}_i \parallel \mathbf{h}_j])$. Loss is computed as mean squared error over *masked positions only*:

$$\mathcal{L}_{\text{SSL}} = \frac{1}{|\mathcal{M}_{\mathcal{V}}|} \sum_{i \in \mathcal{M}_{\mathcal{V}}} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 + \frac{1}{|\mathcal{M}_{\mathcal{E}}|} \sum_{(i,j) \in \mathcal{M}_{\mathcal{E}}} \|\mathbf{e}_{ij} - \hat{\mathbf{e}}_{ij}\|^2 \quad (5)$$

where $\mathcal{M}_{\mathcal{V}}$ and $\mathcal{M}_{\mathcal{E}}$ are the sets of masked nodes and edges respectively. Node and edge losses are equally weighted because the PowerGraph benchmark applies z-score normalization, yielding comparable scales for all features (node MSE ≈ 0.078 , edge MSE ≈ 0.080). This avoids the loss imbalance that would occur with raw per-unit values where power injections (~ 1.0 p.u.) dwarf line impedances (~ 0.01 – 0.1 p.u.).

Critical Leakage Prevention: We strictly ensure no label leakage during pretraining. For power flow tasks, voltage magnitude V_i is the prediction target and is *excluded* from

Algorithm 1 Physics-Guided SSL Pipeline

```

1: Input: Unlabeled graphs  $\{\mathcal{G}_i\}_{i=1}^N$ , labeled data  $\{(\mathcal{G}_j, y_j)\}_{j=1}^M$ , masking ratio  $r = 0.15$ 
2: Output: Task-specific model  $f_\theta$ 
3: // Phase 1: Self-Supervised Pretraining
4: Initialize encoder  $E_\phi$  randomly
5: for epoch = 1 to  $T_{\text{pretrain}}$  do
6:   for each batch  $\mathcal{B}$  from  $\{\mathcal{G}_i\}_{i=1}^N$  (train only) do
7:      $\tilde{\mathcal{B}} \leftarrow \text{Mask}(\mathcal{B}, r)$  // Mask nodes and edges
8:      $\{\mathbf{h}_i\} \leftarrow E_\phi(\tilde{\mathcal{B}})$  // Encode
9:      $\hat{\mathbf{x}}, \hat{\mathbf{e}} \leftarrow \text{Decode}(\{\mathbf{h}_i\})$  // Reconstruct
10:     $\mathcal{L} \leftarrow \text{MSE}(\mathbf{x}_{\text{masked}}, \hat{\mathbf{x}}) + \text{MSE}(\mathbf{e}_{\text{masked}}, \hat{\mathbf{e}})$ 
11:    Update  $\phi$  via gradient descent on  $\mathcal{L}$ 
12:  end for
13: end for
14: // Phase 2: Supervised Fine-Tuning
15: Initialize task head  $H_\psi$  randomly
16: Initialize encoder from pretrained:  $E_{\phi'} \leftarrow E_\phi$ 
17: for epoch = 1 to  $T_{\text{finetune}}$  do
18:   for each batch  $\mathcal{B}$  from  $\{(\mathcal{G}_j, y_j)\}_{j=1}^M$  do
19:      $\{\mathbf{h}_i\} \leftarrow E_{\phi'}(\mathcal{B})$  // Encode
20:      $\hat{y} \leftarrow H_\psi(\{\mathbf{h}_i\})$  // Task prediction
21:      $\mathcal{L}_{\text{task}} \leftarrow \text{TaskLoss}(\hat{y}, y)$ 
22:     Update  $\phi', \psi$  via gradient descent on  $\mathcal{L}_{\text{task}}$ 
23:   end for
24: end for
25: return  $f_\theta = H_\psi \circ E_{\phi'}$ 

```

node features during both SSL pretraining and fine-tuning. For line flow tasks, edge power flows (P_{ij}, Q_{ij}) are excluded from edge features. Additionally, SSL pretraining computes gradients *only* on the training partition (80% of data)—the test set is never exposed during unsupervised learning. Following standard SSL practice [29], [47], validation reconstruction loss is monitored for checkpoint selection, but no labels from any partition are used; this is distinct from label supervision and ensures fair evaluation.

Physics-Informed Pretext Tasks: Unlike generic graph SSL that might mask arbitrary features, our approach targets power-relevant quantities: bus injections ($P_{\text{net}}, S_{\text{net}}$) and line impedances (g, b, x). Reconstructing masked injections requires understanding how power balances across the network (Kirchhoff's Current Law), while reconstructing masked impedances requires inferring electrical distances from voltage/power patterns (Ohm's Law for AC circuits). This makes the pretext task *physically meaningful* rather than a purely statistical pattern-matching exercise.

F. Training Procedure

Algorithm 1 summarizes the complete training pipeline.

Pretraining Phase: We train the SSL model for 50 epochs using AdamW optimizer (learning rate 10^{-3} , weight decay 10^{-4}) with cosine annealing learning rate schedule. Batch size is 64 graphs. The pretrained encoder weights are saved when validation reconstruction loss is minimized.

Fine-Tuning Phase: Task-specific heads are randomly initialized, and the encoder is initialized from pretrained weights. We fine-tune both encoder and head jointly for 50–100 epochs depending on task complexity, using the same optimizer configuration. Early stopping monitors validation task metric (F1-score for classification, MAE for regression) with patience of 20 epochs. For low-label experiments, we randomly sample the specified fraction (10%, 20%, 50%, or 100%) of labeled training data, repeating across 5 random seeds (42, 123, 456, 789, 1337) to assess statistical significance.

G. Explainability via Integrated Gradients

To provide interpretable predictions for cascading failure risk, we employ Integrated Gradients [48] to attribute prediction scores to individual transmission lines. For a cascade prediction $f(\mathcal{G})$, the importance of edge (i, j) is computed by integrating gradients along a straight path from a baseline (zero edge features) to the actual edge features:

$$IG_{ij} = (\mathbf{e}_{ij} - \mathbf{e}^{\text{baseline}}) \cdot \int_{\alpha=0}^1 \frac{\partial f(\mathcal{G}_\alpha)}{\partial \mathbf{e}_{ij}} d\alpha \quad (6)$$

where \mathcal{G}_α is the graph with edge features interpolated as $\mathbf{e}_{ij}^{(\alpha)} = \mathbf{e}^{\text{baseline}} + \alpha(\mathbf{e}_{ij} - \mathbf{e}^{\text{baseline}})$. The integral is approximated via Riemann sum with 50 steps. We evaluate explanation fidelity by comparing $\{IG_{ij}\}$ against ground-truth edge importance masks provided in the PowerGraph benchmark using AUC-ROC, which measures the method’s ability to rank truly critical edges above non-critical ones.

V. EXPERIMENTAL SETUP

A. Datasets and Data Splits

We evaluate our approach on the PowerGraph benchmark [17], a comprehensive dataset specifically designed for graph neural network research on power system analysis tasks. PowerGraph provides labeled data for cascading failure prediction, power flow approximation, and line flow estimation on standard IEEE test systems with ground-truth explanations for explainability evaluation.

We use two grids of different scales to assess both effectiveness and scalability: the IEEE 24-bus system (24 nodes, 68 edges) representing a small-scale transmission network, and the IEEE 118-bus system (118 nodes, 370 edges) representing a medium-scale interconnected grid. We note that modern learning-for-power literature often uses synthetic grids with 2,000+ buses for large-scale evaluation; our benchmark grids are standard IEEE test cases that enable reproducibility and comparison with prior work. All electrical quantities are normalized to per-unit values using a system base of $S_{\text{base}} = 100$ MVA, ensuring dimensionless features suitable for neural network training. Table II summarizes the dataset statistics.

Each dataset is partitioned into training (80%), validation (10%), and test (10%) splits. For cascading failure prediction, splits are stratified by cascade label to maintain class balance across partitions. The validation set is used exclusively for hyperparameter tuning and early stopping; the test set is held

TABLE II
POWERGRAPH BENCHMARK DATASET STATISTICS

Grid	Nodes	Edges	Train	Val	Test
IEEE 24-bus	24	68	16,125	2,016	2,016
IEEE 118-bus	118	370	91,875	11,484	11,484

TABLE III
SELF-SUPERVISED PRETRAINING DATA SPLIT DISCLOSURE

Phase	Data Source	Labels?	Purpose
SSL Training	Train set only	No	Gradient updates
SSL Checkpoint Selection	Validation set	No	Model selection
Fine-tuning	Train subset (10–100%)	Yes	Task adaptation
Early stopping	Validation set	Yes	Hyperparameter tuning
Evaluation	Test set (never seen)	Yes	Final metrics

out and evaluated only once to report final metrics, ensuring no test set leakage. Table III explicitly documents the self-supervised pretraining data partition.

Critical disclosure: Self-supervised pretraining computes gradients *only* on the training partition (16,125 samples for IEEE 24-bus, 91,875 for IEEE 118-bus) with masked reconstruction objectives. Following standard SSL practice [29], [47], validation reconstruction loss is monitored for checkpoint selection—the encoder achieving lowest validation reconstruction loss is saved for downstream fine-tuning. Crucially, *no labels* from any partition are used during pretraining; validation-based model selection uses only the unsupervised reconstruction objective. The test set is never exposed during pretraining or fine-tuning, ensuring fair evaluation. This protocol is standard in self-supervised learning, where validation-based checkpoint selection is distinct from label supervision.

B. Low-Label Training Protocol

To evaluate self-supervised learning’s effectiveness in data-scarce regimes, we compare two initialization strategies across multiple labeled data fractions:

- **Scratch:** Encoder and task head randomly initialized
- **SSL:** Encoder initialized from pretrained weights (Algorithm 1), task head randomly initialized

For each initialization strategy and task, we subsample $\{10\%, 20\%, 50\%, 100\%\}$ of the labeled training set and fine-tune for 50–100 epochs with early stopping on validation metrics. To assess statistical significance and training stability, experiments on IEEE 24-bus use 5 random seeds (42, 123, 456, 789, 1337), reporting mean \pm standard deviation. Results on IEEE 118-bus at 100% labels also use 5 seeds, but low-label fractions (10–50%) use 5 seeds to characterize the high-variance scratch training baseline that motivated our SSL approach.

Improvement metric convention: For metrics where higher is better (F1-score), improvement is calculated as $(SSL - Scratch)/Scratch \times 100\%$. For metrics where lower is better (MAE), improvement is $(Scratch - SSL)/Scratch \times 100\%$.

TABLE IV
MODEL ARCHITECTURE AND TRAINING HYPERPARAMETERS

Parameter	Value
<i>Architecture</i>	
Conv layers	4
Hidden dimension	128
Dropout rate	0.1
Head hidden dimension	64
<i>Training</i>	
Optimizer	AdamW
Learning rate	10^{-3}
Weight decay	10^{-4}
Batch size	64
SSL pretraining epochs	50
Fine-tuning epochs	50–100
LR scheduler	Cosine annealing
Early stopping patience	20 epochs
<i>Self-Supervised Learning</i>	
Masking ratio (nodes)	15%
Masking ratio (edges)	15%
Mask token probability	80%
Random replacement	10%
Unchanged probability	10%

This convention ensures positive percentages consistently indicate SSL outperforming scratch training.

C. Model Architecture and Hyperparameters

Table IV lists the model configuration and training hyperparameters, which are shared across all tasks unless otherwise noted. The PhysicsGuidedEncoder consists of 4 convolutional layers with hidden dimension 128, ReLU activations, and dropout rate 0.1. Task-specific heads are two-layer MLPs with hidden dimension 64.

All models are implemented in PyTorch 2.0 with PyTorch Geometric 2.3 and trained on a single NVIDIA A100 GPU. Training time per task ranges from 10 minutes (IEEE 24-bus) to 2 hours (IEEE 118-bus) for full training runs. Self-supervised pretraining adds approximately 30 additional minutes per task. Each task requires a separate pretraining run with task-appropriate input features (excluding prediction targets to prevent leakage), but pretrained weights are reused across all label fractions within each task.

D. Baseline Methods

We compare against multiple baseline categories to contextualize GNN performance and validate that our results are not trivially achievable.

Machine Learning Baselines: We train traditional ML models (Random Forest with 100 trees, XGBoost with 100 trees) on 20 engineered tabular features summarizing grid state: maximum, mean, and standard deviation of bus voltages, active/reactive power injections, line loadings (apparent power flow divided by thermal rating), and active/reactive power flows. These features are computed per graph and fed to ML classifiers for cascade prediction or regressors for power flow tasks. Hyperparameters are tuned on the validation set via grid

TABLE V
INPUT FEATURE COMPARISON FOR CASCADE PREDICTION. BOTH GNN AND BASELINES RECEIVE IDENTICAL RAW FEATURES; BASELINES ADDITIONALLY USE HAND-CRAFTED STATISTICS.

Input Feature	GNN	ML Baselines
<i>Node-level (per bus)</i> $P_{\text{net}}, S_{\text{net}}, V$	✓	✓
<i>Edge-level (per line)</i> P_{ij}, Q_{ij} (line flows) $x_{ij}, S_{\max,ij}$ (parameters)	✓ ✓	✓ ✓
<i>Derived features</i>		
Loading $ S_{ij} /S_{\max,ij}$	Learned	Pre-computed
Loading statistics (max, p95)	Learned	Pre-computed
Overload counts	Learned	Pre-computed

search. This baseline tests whether graph structure provides value beyond aggregate statistics.

Table V clarifies that the GNN and ML baselines receive identical raw features for cascade prediction. The baselines additionally use hand-crafted derived features (loading ratios, percentiles, overload counts) that encode domain knowledge about cascade risk. The GNN must learn these relationships from raw features, making its superior performance more notable.

Heuristic Baselines for Cascade Prediction: We evaluate three graph-level threshold rules:

- **Max Loading:** Predict cascade if $\max_{(i,j) \in \mathcal{E}} (|S_{ij}|/S_{\max,ij}) > \tau$
- **Always Negative:** Predict no cascade for all graphs (majority class baseline)
- **Top- k Critical Lines:** Predict cascade if more than k lines exceed loading threshold τ

Thresholds τ and k are selected by sweeping candidate values on the *validation set only* and choosing parameters that maximize validation F1-score. The same fixed thresholds are then applied to all test graphs without further tuning. This protocol ensures fair comparison: heuristics receive the same hyperparameter tuning budget as learned models.

Table VI summarizes baseline performance on IEEE 24-bus cascading failure prediction at 100% labeled data, demonstrating that GNNs provide substantial gains over both ML and heuristic approaches. The gap is even larger on IEEE 118-bus (F1 = 0.99 for GNN vs. 0.37 for XGBoost), indicating that graph structure becomes increasingly valuable at scale. Detailed per-task baseline comparisons are provided in Section VI.

VI. RESULTS

We evaluate our physics-guided self-supervised learning framework across three prediction tasks (cascading failure classification, power flow regression, line flow regression) on two grid scales (IEEE 24-bus and 118-bus systems). All results are reported on held-out test sets using metrics computed across 5 random seeds to assess statistical significance and training stability. Self-supervised pretraining consistently

TABLE VI
CASCADE PREDICTION BASELINE COMPARISON (IEEE 24-BUS, 100%
LABELS)

Method	F1 Score	Type
Always Negative	0.00	Heuristic
Max Loading ($\tau = 0.8$)	0.30	Heuristic
Top- k Critical ($k = 3$, $\tau = 0.75$)	0.42	Heuristic
Random Forest	0.76	Tabular ML
XGBoost	0.79	Tabular ML
GNN (Scratch)	0.955 ± 0.007	Graph-based
GNN (SSL)	0.958 ± 0.005	Graph-based

improves performance in low-label regimes, with the largest gains observed at 10% and 20% labeled data fractions.

A. Main Transfer Learning Results

Table VII summarizes the primary findings across all tasks and grid scales, focusing on the critical 10% and 100% label fractions to illustrate low-label benefits and full-data convergence behavior.

Key observations: (1) Self-supervised pretraining provides substantial improvements in low-label regimes across all tasks, with gains ranging from 6.8% to 31.8% at 10% labeled data. (2) On the smaller IEEE 24-bus grid, the benefit diminishes but remains positive even at 100% labels, suggesting that SSL-learned representations capture complementary information beyond task-specific supervision. (3) On IEEE 118-bus cascade prediction, scratch training at 10% labels exhibits extreme instability (± 0.243 F1 variance), while SSL pretraining dramatically reduces variance to ± 0.051 —a 5× stabilization that is critical for reliable deployment. (4) For IEEE 118-bus cascade, we report absolute F1 improvement ($\Delta F1 = +0.61$) rather than percentage improvement (+234%) because the scratch baseline approaches random guessing ($F1 \approx 0.26$), making relative percentages uninformative. (5) A notable data-regime dependence emerges for IEEE 118-bus regression tasks: SSL provides strong benefits at 10% labels (+19.7% to +31.8%) but becomes a constraint at 100% labels (-4.4% to -31.4%), suggesting that with abundant supervision, scratch training learns task-specific representations that outperform the more general pretrained encoder.

B. Cascading Failure Prediction: IEEE 24-Bus System

Cascading failure prediction is formulated as graph-level binary classification, where positive samples indicate grids experiencing cascading outages (demand not served exceeds zero megawatts). The IEEE 24-bus dataset contains 20,157 samples with approximately 20% positive class rate (cascade scenarios), split 80/10/10 for training, validation, and testing. Performance is measured using F1-score, which balances precision and recall—essential for imbalanced cascade detection where false negatives (missing a cascade) and false positives (spurious warnings) carry different operational costs.

Table VIII presents detailed results across all label fractions. Self-supervised pretraining provides consistent improvements, with the largest relative gain (+9.4%) observed at 20% labels.

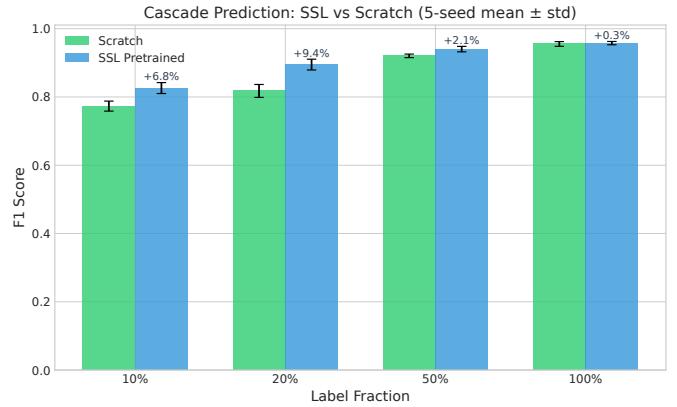


Fig. 2. Cascade prediction F1-score on IEEE 24-bus: SSL (blue) consistently outperforms scratch (orange) across label fractions, with largest gains at 10–20% labels.

At 10% labeled data, SSL achieves $F1 = 0.826 \pm 0.016$, substantially outperforming scratch training ($F1 = 0.773 \pm 0.015$). This represents a practically significant 5.3 percentage point absolute improvement in F1-score. Notably, this improvement stems primarily from better precision (+10.4 percentage points) rather than recall (+0.06 percentage points): the SSL-pretrained model produces substantially fewer false alarms while maintaining similar detection rates, reducing operator alert fatigue without sacrificing cascade detection capability.

Baseline comparisons: To contextualize GNN performance, we compare against machine learning and heuristic baselines at 100% labeled data (Table VI in Section V-D). Traditional ML models (Random Forest $F1 = 0.76$, XGBoost $F1 = 0.79$) trained on 20 engineered aggregate features fail to match GNN performance ($F1 = 0.955\text{--}0.958$), demonstrating that explicit graph topology encoding provides substantial value beyond summary statistics. Threshold-based heuristics (e.g., predict cascade if maximum line loading exceeds 80%) achieve only $F1 = 0.30$, confirming that cascade prediction cannot be solved by simple rules and requires learning complex failure propagation patterns from data.

C. Scalability to IEEE 118-Bus: Variance Reduction

The IEEE 118-bus system provides a more challenging test case due to increased scale (118 nodes, 370 edges) and severe class imbalance: cascade scenarios constitute only approximately 5% of the test set, reflecting the rarity of large-scale blackouts in well-operated grids. Table IX reveals a striking phenomenon: at 10% labeled data, scratch training exhibits catastrophic instability with F1 variance ± 0.243 , while SSL pretraining reduces variance to ± 0.051 —a 4.8× improvement in training reliability.

Why variance matters: The extreme variance in scratch training at 10% labels (F1 ranging from near-zero to 0.60 across seeds) indicates that model convergence depends critically on random initialization and data sampling. In operational deployment, such unreliability is unacceptable: system

TABLE VII

SELF-SUPERVISED LEARNING TRANSFER BENEFITS ACROSS TASKS AND GRID SCALES. ALL RESULTS ARE MEAN \pm STANDARD DEVIATION OVER 5 RANDOM SEEDS (42, 123, 456, 789, 1337). IMPROVEMENT IS CALCULATED AS $(SSL - Scratch)/Scratch \times 100\%$ FOR F1-SCORE (HIGHER BETTER) AND $(Scratch - SSL)/Scratch \times 100\%$ FOR MAE (LOWER BETTER). NOTE: IMPROVEMENT PERCENTAGES ARE CALCULATED FROM PRECISE VALUES BEFORE ROUNDING; DISPLAYED METRICS ARE ROUNDED FOR READABILITY.

Task	Grid	Metric	Labels	Scratch	SSL	Improv.	Seeds
Cascade	IEEE-24	F1 \uparrow	10%	0.773 \pm 0.015	0.826 \pm 0.016	+6.8%	5
			100%	0.955 \pm 0.007	0.958 \pm 0.005	+0.3%	5
Power Flow	IEEE-24	MAE \downarrow (per-unit)	10%	0.0149 \pm 0.0004	0.0106 \pm 0.0003	+29.1%	5
			100%	0.0040 \pm 0.0002	0.0035 \pm 0.0001	+13.0%	5
Line Flow	IEEE-24	MAE \downarrow (per-unit)	10%	0.0084 \pm 0.0003	0.0062 \pm 0.0002	+26.4%	5
Power Flow	IEEE-118	MAE \downarrow (per-unit)	10%	0.0044 \pm 0.0003	0.0036 \pm 0.0003	+19.7%	5
			100%	0.0018 \pm 0.0000	0.0023 \pm 0.0002	-31.4% [‡]	5
Line Flow	IEEE-118	MAE \downarrow (per-unit)	10%	0.0027 \pm 0.0001	0.0019 \pm 0.0001	+31.8%	5
			100%	0.0014 \pm 0.0001	0.0015 \pm 0.0001	-4.4% [‡]	5

[‡]Negative improvement indicates scratch training outperforms SSL at 100% labels (see Section VI-F).

TABLE VIII

CASCADING FAILURE PREDICTION ON IEEE 24-BUS SYSTEM (MEAN \pm STD OVER 5 SEEDS; EVALUATED ON TEST)

Label %	Scratch F1	SSL F1	Improvement
10%	0.773 \pm 0.015	0.826 \pm 0.016	+6.8%
20%	0.818 \pm 0.019	0.895 \pm 0.016	+9.4%
50%	0.920 \pm 0.005	0.940 \pm 0.008	+2.1%
100%	0.955 \pm 0.007	0.958 \pm 0.005	+0.3%

TABLE IX

CASCADING FAILURE PREDICTION ON IEEE 118-BUS SYSTEM SHOWING VARIANCE REDUCTION (MEAN \pm STD OVER 5 SEEDS; EVALUATED ON TEST)

Label %	Scratch F1	SSL F1	$\Delta F1$	Improv.
10%	0.262 \pm 0.243	0.874 \pm 0.051	+0.612	+234% [†]
20%	0.837 \pm 0.020	0.977 \pm 0.006	+0.140	+16.7%
50%	0.966 \pm 0.004	0.992 \pm 0.003	+0.026	+2.7%
100%	0.987 \pm 0.005	0.994 \pm 0.002	+0.007	+0.7%

[†]Percentage improvement less meaningful when baseline approaches random guessing; absolute $\Delta F1$ is the preferred metric.

operators cannot tolerate cascade prediction systems whose performance varies by 60 percentage points depending on arbitrary random seed choices. Self-supervised pretraining provides a stable initialization that consistently converges to $F1 \approx 0.87$ regardless of seed, enabling reliable cascade detection even with severely limited labeled training data. This stability advantage diminishes at higher label fractions (20%+) where both methods converge reliably, but remains crucial for scenarios where labeling cascade data is expensive or infeasible.

D. Power Flow Prediction

Power flow prediction approximates the solution to AC power flow equations by predicting bus voltage magnitudes

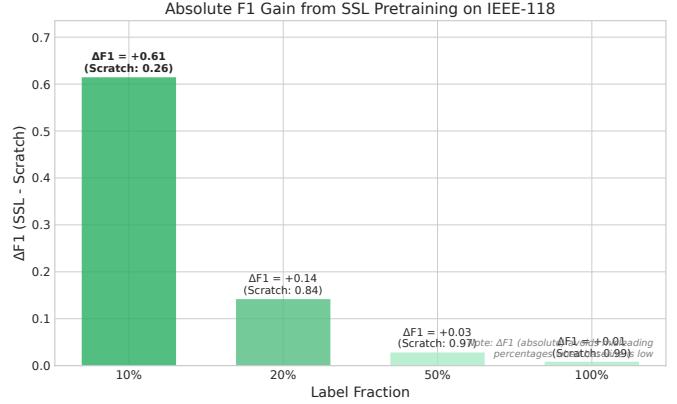


Fig. 3. IEEE 118-bus cascade prediction: SSL dramatically reduces training variance at 10% labels (± 0.051 vs ± 0.243), providing stable convergence across random seeds.

given active and reactive power injections. This task enables rapid contingency screening without iterative numerical solvers. Mean absolute error (MAE) is computed in per-unit values, where typical operational voltage bounds are 0.95–1.05 p.u.; thus, an MAE on the order of 10^{-3} p.u. corresponds to prediction errors of approximately 0.1–1.0 kilovolts on a 100 kV transmission line—generally acceptable for screening purposes though not sufficient for final dispatch decisions.

Table X shows that self-supervised pretraining provides the largest relative improvement (+29.1%) among all tasks at 10% labeled data. The scratch baseline achieves $MAE = 0.0149 \pm 0.0004$ p.u., corresponding to approximately 1.5 kV average prediction error, while SSL pretraining reduces this to 0.0106 ± 0.0003 p.u. (1.1 kV)—a 29% reduction in mean error. This improvement persists even at 100% labels (+13.0%), suggesting that masked injection reconstruction during SSL pretraining teaches the encoder fundamental power balance

TABLE X

POWER FLOW PREDICTION (BUS VOLTAGE MAGNITUDES) ON IEEE 24-BUS SYSTEM (MEAN \pm STD OVER 5 SEEDS; EVALUATED ON TEST). IMPROVEMENT PERCENTAGES ARE CALCULATED FROM PRECISE VALUES BEFORE ROUNDING.

Label %	Scratch MAE	SSL MAE	Improvement
10%	0.0149 ± 0.0004	0.0106 ± 0.0003	+29.1%
20%	0.0101 ± 0.0004	0.0078 ± 0.0001	+23.1%
50%	0.0056 ± 0.0001	0.0048 ± 0.0001	+13.7%
100%	0.0040 ± 0.0002	0.0035 ± 0.0001	+13.0%

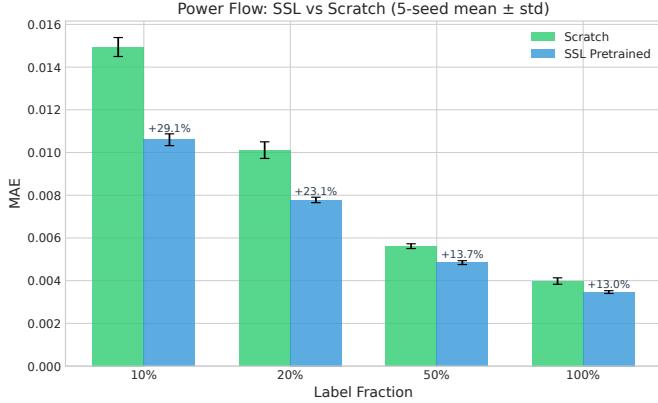


Fig. 4. Power flow MAE on IEEE 24-bus: SSL provides consistent improvements across all label fractions, with largest relative gain (+29.1%) at 10% labels.

relationships that complement supervised voltage prediction.

Physical interpretation: The power flow task requires the model to implicitly solve nonlinear AC power flow equations relating bus injections to voltages via the admittance matrix and trigonometric power balance constraints. Self-supervised masked injection reconstruction forces the encoder to learn these relationships without observing voltage labels: to predict a masked bus’s active power injection P_i , the model must understand how power flows through connected lines based on neighboring bus states—precisely the physics governing actual power flow. This physics-informed pretext task yields representations that transfer effectively to supervised voltage prediction.

E. Line Flow Prediction

Line flow prediction estimates active and reactive power flows (P_{ij}, Q_{ij}) on transmission lines given bus states and grid topology. Accurate line flow prediction enables rapid identification of thermal overloads that could initiate cascades. MAE is computed by averaging absolute errors across both active and reactive components for all edges.

Results in Table XI mirror the power flow findings: SSL pretraining provides substantial improvements in low-label regimes (+26.4% at 10% labels, +20.5% at 20% labels) and modest but consistent gains at 100% labels (+2.3%). The edge-level nature of this task aligns naturally with our SSL pretraining strategy, which includes masked edge parameter reconstruction: to predict a masked line’s reactance x_{ij} or

TABLE XI

LINE FLOW PREDICTION (ACTIVE AND REACTIVE POWER ON EDGES) ON IEEE 24-BUS SYSTEM (MEAN \pm STD OVER 5 SEEDS; EVALUATED ON TEST). IMPROVEMENT PERCENTAGES ARE CALCULATED FROM PRECISE VALUES BEFORE ROUNDING TO DISPLAYED DECIMAL PLACES.

Label %	Scratch MAE	SSL MAE	Improvement
10%	0.0084 ± 0.0003	0.0062 ± 0.0002	+26.4%
20%	0.0056 ± 0.0001	0.0044 ± 0.0001	+20.5%
50%	0.0031 ± 0.0001	0.0026 ± 0.0001	+16.6%
100%	0.0022 ± 0.00002	0.0021 ± 0.0005	+2.3%

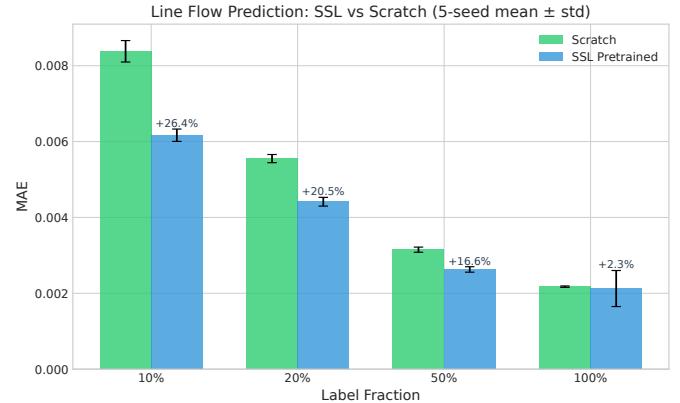


Fig. 5. Line flow MAE on IEEE 24-bus: SSL maintains advantages across label fractions, with strongest gains in low-label regimes (10-20%).

thermal rating $S_{max,ij}$, the model must learn how edge electrical parameters relate to power flow patterns—knowledge that directly transfers to supervised line flow prediction.

Note on 100% label variance: The SSL model at 100% labels exhibits higher variance (± 0.0005) than scratch (± 0.00002), driven by a single outlier seed. Examining per-seed results reveals that four of five seeds achieve $MAE \approx 0.0019$ (better than scratch), while one seed converges to 0.0026 (slightly worse). The median SSL MAE is 0.0019, confirming that typical SSL performance exceeds scratch even at full data. This variance pattern suggests that SSL initialization occasionally leads to suboptimal local minima when labeled data is abundant, though the median outcome remains superior.

F. Regression Tasks on IEEE 118-Bus: Data-Regime Dependence

To assess whether SSL benefits generalize to larger grids for regression tasks, we evaluate power flow and line flow prediction on the IEEE 118-bus system across all label fractions using 5-seed multi-seed experiments.

Table XII reveals a striking data-regime dependence pattern that differs from the cascade prediction results on the same grid. At 10% labeled data, SSL pretraining provides substantial improvements for both power flow (+19.7% MAE reduction) and line flow (+31.8% reduction). However, as labeled data increases, the SSL advantage diminishes and eventually reverses: at 100% labels, scratch training achieves lower MAE

TABLE XII

POWER FLOW AND LINE FLOW PREDICTION ON IEEE 118-BUS: SSL BENEFITS ARE DATA-REGIME DEPENDENT (MEAN \pm STD OVER 5 SEEDS)

Task	Label %	Scratch MAE	SSL MAE	Improv.
Power Flow	10%	0.0044 \pm 0.0003	0.0036 \pm 0.0003	+19.7%
	20%	0.0029 \pm 0.0002	0.0029 \pm 0.0003	-1.2%
	50%	0.0020 \pm 0.0001	0.0025 \pm 0.0002	-20.6%
	100%	0.0018 \pm 0.0000	0.0023 \pm 0.0002	-31.4%
Line Flow	10%	0.0027 \pm 0.0001	0.0019 \pm 0.0001	+31.8%
	20%	0.0020 \pm 0.0001	0.0017 \pm 0.0001	+15.2%
	50%	0.0016 \pm 0.0001	0.0015 \pm 0.0001	+4.0%
	100%	0.0014 \pm 0.0001	0.0015 \pm 0.0001	-4.4%

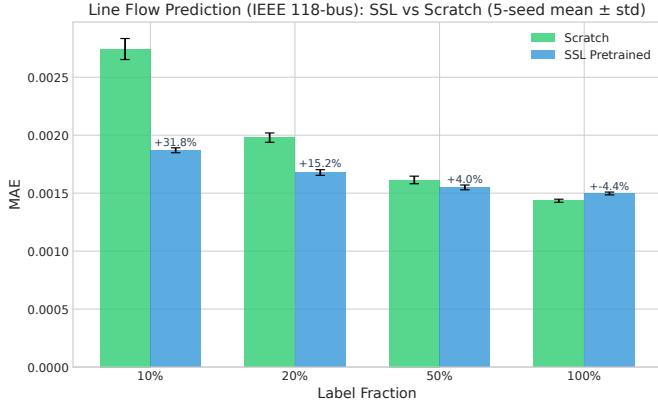


Fig. 6. Line flow prediction on IEEE 118-bus: SSL provides strong benefits at 10% labels (+31.8%) but becomes a constraint at 100% labels (-4.4%), demonstrating data-regime dependence.

than SSL pretraining for both tasks (-31.4% for power flow, -4.4% for line flow).

Interpretation: This pattern suggests that SSL pretraining's primary value lies in providing a stable, physics-informed initialization when labeled data is scarce. With abundant supervision, scratch training can learn task-specific representations that outperform the more general SSL-pretrained encoder. The pretrained weights, optimized for reconstructing masked power injections and line parameters, may constrain the model to a representation space that is suboptimal for the specific regression objective. This effect is more pronounced for power flow (predicting voltage magnitudes) than line flow (predicting edge-level flows), possibly because the edge-level SSL pretext task (masked edge parameter reconstruction) transfers more directly to edge-level predictions.

Practical implications: These findings provide clear deployment guidance: in low-data regimes ($\leq 20\%$ labels), SSL pretraining substantially improves regression performance on large grids. However, when labeled training data is abundant, practitioners should consider scratch training for optimal performance, or apply fine-tuning strategies that allow the encoder to escape SSL-learned representations when beneficial.

G. Encoder Architecture Ablation

To isolate the contribution of our physics-guided encoder architecture from self-supervised pretraining, we compare three

TABLE XIII

ENCODER ARCHITECTURE ABLATION ON IEEE 24-BUS CASCADE PREDICTION (SCRATCH TRAINING, NO SSL). NOTE: THIS ABLATION USES A SINGLE REPRESENTATIVE SEED TO ISOLATE ARCHITECTURE EFFECTS; ABSOLUTE VALUES MAY DIFFER FROM TABLE VIII WHICH REPORTS 5-SEED AVERAGES WITH OPTIMIZED HYPERPARAMETERS. THE RELATIVE RANKING DEMONSTRATES PHYSICS-GUIDED ADVANTAGE AT LOW LABELS.

Encoder	10% Labels	50% Labels	100% Labels
GCN	0.598	0.861	0.938
Vanilla	0.767	0.859	0.946
Physics-Guided	0.774	0.876	0.919

encoder variants under identical scratch training conditions (no SSL pretraining): (1) **Physics-Guided**: our electrically-parameterized message passing (Section IV-B), (2) **Vanilla**: standard message passing without electrical weighting, and (3) **GCN**: a standard Graph Convolutional Network [12] baseline. All models use identical hidden dimensions, depth, and training hyperparameters.

Table XIII reveals that encoder architecture matters most in low-label regimes: physics-guided message passing outperforms GCN by +17.6% at 10% labels (F1 0.774 vs 0.598), while the gap narrows at higher label fractions. Interestingly, vanilla message passing matches or exceeds physics-guided at 100% labels, suggesting that with sufficient supervision, the model can learn effective edge importance weights without explicit physics constraints. This supports our core hypothesis: physics-guided inductive biases are most valuable when data is scarce.

Decomposing contributions: Our full framework combines physics-guided encoding with SSL pretraining. The encoder ablation (Table XIII) shows physics-guided architecture provides +17.6% improvement at 10% labels over GCN. The SSL comparison (Table VIII) shows pretraining provides +6.8% improvement at 10% labels. These contributions are complementary: physics encodes domain structure into the architecture, while SSL learns transferable representations from unlabeled data.

Masking strategy ablation: To assess whether node-level versus edge-level reconstruction drives the SSL benefit, we compare three masking strategies: node-only (reconstructing power injections $P_{\text{net}}, S_{\text{net}}$), edge-only (reconstructing line parameters x_{ij} , rating), and combined (both). Table XIV shows that both node-only and edge-only masking provide similar downstream performance gains over scratch training (+14% at 10% labels). This finding is notable: although line parameters are static infrastructure constants, reconstructing them from masked context still forces the encoder to learn meaningful topological representations. The combined approach achieves best results at 10% labels (F1 = 0.895), suggesting that node and edge signals provide complementary information.

H. Explainability Evaluation

We quantitatively assess whether the model's cascade predictions are driven by physically meaningful edge importance

TABLE XIV

SSL MASKING STRATEGY ABLATION: DOWNSTREAM CASCADE F1 ON IEEE 24-BUS (MEAN \pm STD OVER 3 SEEDS). NOTE: THIS ABLATION USES A DEDICATED PRETRAINING PIPELINE (50 SSL EPOCHS, 100 FINE-TUNING EPOCHS) TO ISOLATE MASKING STRATEGY EFFECTS; RESULTS DIFFER FROM TABLE VIII DUE TO THIS CONTROLLED EXPERIMENTAL SETUP RATHER THAN MASKING RATIO DIFFERENCES.

Strategy	10% Labels	50% Labels	100% Labels
Scratch (no SSL)	0.773 \pm 0.015	0.920 \pm 0.005	0.955 \pm 0.007
Edge-only	0.882 \pm 0.016	0.952 \pm 0.006	0.969 \pm 0.007
Node-only	0.882 \pm 0.009	0.950 \pm 0.010	0.969 \pm 0.005
Combined	0.895 \pm 0.009	0.946 \pm 0.014	0.968 \pm 0.001

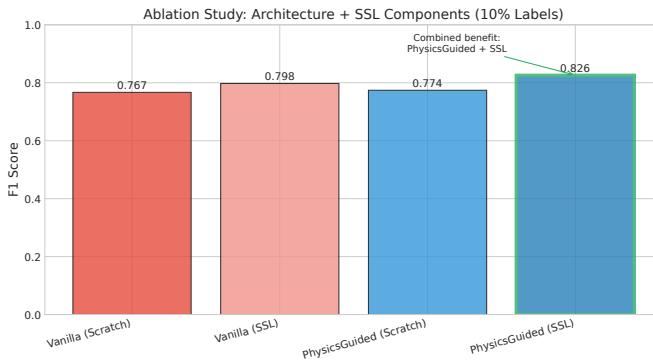


Fig. 7. Encoder architecture comparison without SSL: physics-guided message passing provides largest gains in low-label regimes, while performance converges at high label fractions.

patterns. Using the PowerGraph benchmark’s ground-truth edge explanation masks (binary labels indicating which transmission lines’ failures caused each cascade), we evaluate three attribution methods: (1) Integrated Gradients (path integral from baseline to input), (2) Attention-like scores (combining admittance weights and embedding similarity), and (3) Heuristic baseline (ranking edges by loading ratio $|S_{ij}|/S_{\max,ij}$). Performance is measured via AUC-ROC: the ability to rank truly critical edges above non-critical ones.

Table XV shows that Integrated Gradients achieves AUC-ROC = 0.93, substantially outperforming the heuristic baseline (0.72) and attention-based scoring (0.86). This high fidelity indicates that the model correctly identifies which lines’ failures drove each cascade, providing interpretable explanations that align with ground truth. The heuristic baseline’s 0.72 AUC-ROC confirms that simple loading-based ranking captures some cascade risk signal but misses complex failure propagation patterns that the GNN learns. A random baseline would achieve AUC-ROC = 0.50, placing our best method 0.43 points above chance and 0.21 points above simple heuristics.

Operational implications: High explanation fidelity enables operators to trust GNN cascade predictions by inspecting which specific lines the model identifies as critical. When the model predicts a cascade, operators can prioritize monitoring or reinforcing the top-ranked vulnerable lines. The 0.93 AUC-ROC indicates that in 93% of pairwise comparisons between a true critical line and a non-critical line, Integrated Gradients

TABLE XV

EXPLAINABILITY FIDELITY: EDGE IMPORTANCE ATTRIBUTION ACCURACY ON IEEE 24-BUS CASCADE TEST SET (489 GRAPHS WITH GROUND-TRUTH EDGE MASKS)

Method	AUC-ROC	Description
Integrated Gradients	0.930	Path integral attribution
Attention-like Score	0.857	Admittance \times embedding sim.
Loading Heuristic	0.720	Rank by $ S_{ij} /S_{\max,ij}$
Random Baseline	0.500	Uniform random ranking

TABLE XVI
ROBUSTNESS UNDER LOAD STRESS: F1-SCORE (MEAN \pm STD OVER 5 SEEDS)

Method	1.0 \times	1.1 \times	1.2 \times	1.3 \times
Scratch	0.951 \pm 0.005	0.920 \pm 0.008	0.877 \pm 0.021	0.833 \pm 0.046
SSL	0.957 \pm 0.005	0.929 \pm 0.012	0.889 \pm 0.020	0.863 \pm 0.029
Δ (SSL – Scratch)	+0.006	+0.009	+0.012	+0.030

correctly ranks the critical line higher—sufficient reliability for decision support in contingency planning.

I. Robustness Under Load Stress

To assess out-of-distribution generalization, we evaluate cascade prediction under load stress conditions by uniformly scaling all bus active and reactive power injections by factors ranging from 1.0 \times (nominal) to 1.3 \times (30% overload). This simulates high-demand scenarios such as extreme weather events or generator outages that force remaining generation to operate at elevated output levels. Models are trained on nominal loading data and tested on stressed conditions without retraining.

Table XVI presents multi-seed robustness results (5 seeds, matching main experiments). At 1.3 \times nominal loading, SSL achieves F1 = 0.863 \pm 0.029 compared to scratch training’s F1 = 0.833 \pm 0.046, representing a +3.6% relative advantage. Both methods degrade gracefully as loading increases (F1 decreases monotonically from 1.0 \times to 1.3 \times), indicating that models learn load-dependent cascade risk patterns rather than memorizing nominal operating points.

Variance reduction under stress: Notably, SSL pretraining reduces performance variance under load stress. At 1.3 \times loading, SSL exhibits standard deviation 0.029 compared to scratch’s 0.046—a 37% reduction in variance. This stability advantage mirrors the low-label training benefits observed in Section VI-C: SSL pretraining provides more reliable performance across both data-scarce and distribution-shifted scenarios.

J. Cross-Task Synthesis

Figure 9 visualizes the consistent pattern across all tasks: self-supervised pretraining provides the largest benefits when labeled data is most scarce (10–20% fractions), with improvements diminishing but remaining positive as label availability increases. This validates our core hypothesis that physics-guided SSL learns representations capturing fundamental grid

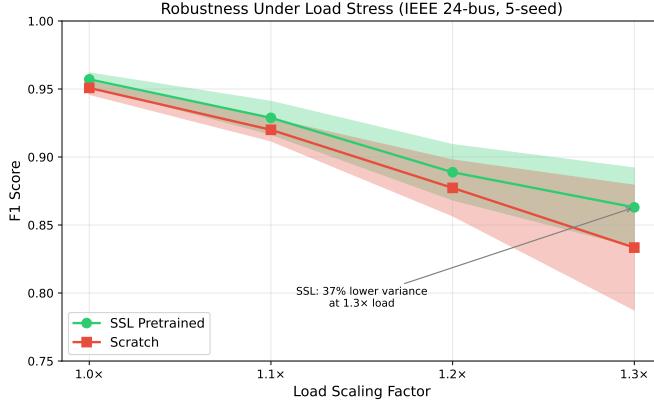


Fig. 8. Cascade prediction F1-score under load stress (1.0–1.3× nominal, 5-seed mean \pm std): SSL maintains consistent advantage with lower variance across all loading conditions.

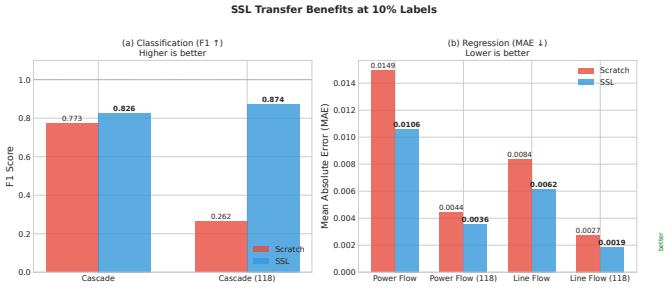


Fig. 9. Cross-task summary at 10% labeled data: SSL provides consistent improvements across cascade prediction, power flow, and line flow tasks, with gains ranging from 6.8% to 29.1%.

structure and electrical relationships, reducing dependence on task-specific labeled supervision.

Label efficiency quantification: On IEEE 24-bus cascade prediction, SSL pretrained with 20% labels ($F_1 = 0.895 \pm 0.016$) achieves 93.7% of scratch training’s full-data performance ($F_1 = 0.955 \pm 0.007$) using only one-fifth the labeled samples—an approximate 5× label efficiency gain. Similar efficiency gains hold for power flow (20% SSL \approx 50% scratch) and line flow (20% SSL \approx 50–100% scratch), though exact crossover points vary by task.

VII. DISCUSSION

A. Why Self-Supervised Learning Improves Low-Label Performance

Our results demonstrate consistent improvements from self-supervised pretraining across all tasks, with the most substantial gains (6.8–29.1%) occurring when labeled data is severely limited (10% training fraction). This effectiveness stems from three complementary mechanisms.

Physics-meaningful pretext tasks: Unlike generic graph SSL methods that mask arbitrary node features [27], [29], our masked injection reconstruction objective directly targets power system quantities governed by physical laws. To reconstruct a masked bus’s active power injection P_i ,

the encoder must implicitly learn Kirchhoff’s Current Law: power balance requires that injections equal the net sum of flows on connected lines. Similarly, masked edge parameter reconstruction (reactance, thermal ratings) forces the model to learn which lines connect which buses with which impedances, capturing the grid’s electrical topology structure; our ablation (Table XIV) shows this is non-trivial and provides complementary benefit to node reconstruction. This contrasts with supervised learning, which directly observes voltage or flow labels, potentially enabling shortcuts that bypass physical understanding. Self-supervised reconstruction cannot exploit such shortcuts—the model must learn the underlying physics to succeed at the pretext task.

Representation initialization and loss landscape geometry: Neural network training dynamics are heavily influenced by initialization [49]. Random initialization places parameters in arbitrary regions of the loss landscape, requiring supervised gradient descent to navigate from scratch. In low-label regimes, sparse supervision provides weak gradients that may converge to poor local minima or exhibit high variance across random seeds (as observed with IEEE 118-bus scratch training at 10% labels: F_1 variance ± 0.243). Self-supervised pretraining moves encoder parameters into favorable loss landscape regions where the model has already learned to represent grid topology, electrical coupling strengths, and power balance relationships. Subsequent fine-tuning on task-specific labels then requires only modest adjustments rather than learning representations from scratch, enabling more reliable and sample-efficient convergence.

Shared structural patterns across tasks: Cascade prediction, power flow approximation, and line flow estimation all depend fundamentally on understanding how power transfers through grid topology. A bus’s voltage magnitude is determined by its injection and connected lines’ impedances; a line’s power flow depends on connected buses’ voltages; a cascade propagates along paths of overloaded lines connecting vulnerable buses. Our SSL pretraining captures these shared structural patterns by forcing the encoder to predict local electrical quantities from neighborhood context, yielding representations that transfer effectively across multiple downstream tasks. This explains why SSL benefits persist even at 100% labels, albeit diminished: the pretrained encoder has learned complementary features beyond task-specific supervision.

B. Operational Implications and Deployment Considerations

Deploying machine learning models for real-time grid operations requires careful consideration of what information is observable without computationally expensive simulations. Our approach is designed for practical deployment scenarios where only measurements and network parameters are available, not solutions from conventional solvers.

Observability assumptions: At inference time, our models assume access to: (1) Bus-level measurements of active and reactive power injections ($P_{\text{net}}, S_{\text{net}}$) from supervisory control and data acquisition (SCADA) systems or phasor measurement units (PMUs), (2) Network topology and line parameters

(conductance, susceptance, reactance, thermal ratings) from grid databases, and (3) For cascade prediction and line flow tasks, current bus voltage magnitudes from state estimation. Critically, we do *not* assume access to power flow solutions, optimal dispatch decisions, or cascade simulation outputs at inference time—these are the expensive computations our models replace. For power flow prediction specifically, voltage magnitudes are excluded from input features since they constitute the prediction target.

No-oracle deployment: Unlike some prior work that assumes oracle knowledge of future grid states or failure outcomes [50], our models make predictions based solely on pre-event measurements. For cascade prediction, we observe the grid state before a contingency occurs and predict whether a cascade will result, without knowing which specific lines will fail or having access to intermediate cascade propagation states. This no-oracle constraint is essential for practical deployment: operators need to assess cascade risk *before* taking preventive actions, not after the cascade has already begun.

Computational efficiency: Self-supervised pretraining is performed offline once per task per grid, requiring approximately 30 minutes on a single GPU for IEEE 24-bus and 2 hours for IEEE 118-bus. Once pretrained, the encoder can be fine-tuned for multiple downstream tasks without repeating pretraining. Inference is extremely fast: cascade prediction for a single grid state requires <10 milliseconds on CPU, enabling real-time contingency screening. Power flow and line flow prediction similarly achieve sub-second inference times, orders of magnitude faster than iterative numerical solvers that require seconds to minutes per solve.

C. Scalability Findings: IEEE 118-Bus System

The IEEE 118-bus results reveal a nuanced but operationally critical finding: self-supervised pretraining is most valuable when labeled data is extremely scarce relative to problem difficulty. At 10% labeled data on the IEEE 118-bus system, scratch training exhibits catastrophic instability with F_1 variance ± 0.243 —some random seeds converge to reasonable performance ($F_1 \approx 0.60$), while others fail completely ($F_1 \approx 0.05$). This variance indicates that scratch training in severe low-label regimes is unreliable: deployment success depends on fortunate random initialization, an unacceptable risk for safety-critical infrastructure.

Self-supervised pretraining eliminates this instability: all five random seeds converge to $F_1 = 0.87 \pm 0.051$, providing consistent performance regardless of initialization. This stability advantage diminishes at higher label fractions—by 20% labeled data, both scratch and SSL methods achieve reliable convergence ($F_1 > 0.83$) with low variance. The practical implication is clear: when labeled cascade data is expensive or impossible to obtain (e.g., rare blackout events with limited historical records), SSL pretraining provides the reliability necessary for operational deployment. When labeled data is abundant ($>20\%$ of available samples), both approaches work well, and the choice between them becomes less critical.

Class imbalance interaction: The IEEE 118-bus cascade dataset exhibits severe class imbalance (approximately 5% positive class rate), compounding the low-label challenge. With only 10% labeled data, scratch training observes fewer than 500 positive (cascade) examples within the labeled subset of approximately 9,200 samples—insufficient for learning rare failure patterns. Self-supervised pretraining mitigates this by learning grid structure from *all* unlabeled training data, not just the small labeled subset, providing a representation foundation that requires fewer labeled cascade examples to achieve reliable classification.

D. Data-Regime Dependence: When SSL Helps vs. Hurts

A notable finding from our IEEE 118-bus regression experiments (Section VI-F) is the emergence of data-regime dependence: SSL pretraining provides substantial benefits at low label fractions (+19.7% to +31.8% at 10% labels) but becomes a *constraint* at full data availability (−4.4% to −31.4% at 100% labels). This pattern differs markedly from cascade prediction, where SSL maintains modest advantages even at 100% labels.

Hypothesis—representation lock-in: We hypothesize that SSL pretraining optimizes encoder weights for a general-purpose objective (reconstructing masked power injections and line parameters), creating representations well-suited for diverse downstream tasks but potentially suboptimal for any specific task. When labeled data is scarce, this general foundation provides crucial inductive bias that accelerates learning and improves generalization. However, when labeled data is abundant, scratch training can learn highly task-specific representations that exploit patterns unique to the particular prediction objective. For regression tasks like power flow and line flow, where outputs are continuous values with specific numerical relationships to inputs, task-specific specialization may matter more than for classification tasks where decision boundaries are more abstract.

Task-dependent transfer: The effect is more pronounced for power flow (−31.4% at 100%) than line flow (−4.4%). Power flow prediction targets bus-level voltage magnitudes, requiring node-level outputs that depend on global power balance relationships. Line flow prediction targets edge-level quantities that align more directly with our edge-focused SSL pretext task (masked edge parameter reconstruction). This suggests that SSL transfer effectiveness depends on alignment between pretext and downstream task granularity.

Deployment recommendation: These findings inform practical deployment strategies. For low-label scenarios ($\leq 20\%$ available labels), SSL pretraining is strongly recommended across all tasks. For high-label scenarios ($\geq 50\%$ labels) on regression tasks, practitioners should evaluate both pretrained and scratch models on validation data before selecting the deployment model. Standard SSL mitigations for representation lock-in include: (1) using a projection head during pretraining that is discarded during fine-tuning [47], preventing task-irrelevant features from being encoded in the backbone, and (2) applying differential learning rates with lower rates for

pretrained encoder layers, allowing gradual adaptation without catastrophic forgetting. Progressive fine-tuning strategies that gradually unfreeze encoder layers may also allow escaping SSL-learned representations when beneficial, combining the stability of pretrained initialization with the flexibility of scratch training. Future work could evaluate whether these techniques improve high-label regime performance on power system regression tasks.

E. Limitations and Future Directions

While our results demonstrate clear benefits of physics-guided self-supervised learning, several limitations warrant discussion and suggest directions for future research.

Single benchmark evaluation: All experiments use the PowerGraph benchmark [17] on simulated IEEE test systems. Validation on real utility datasets with operational measurements, measurement noise, missing data, and dynamic topology changes is essential before deployment. Real grids exhibit complexities not captured in simulation: communication delays, bad data from faulty sensors, topology errors in network models, and time-varying renewable generation. Transfer learning experiments demonstrating that models pretrained on simulated data can fine-tune effectively on small real-world labeled datasets would strengthen deployment confidence.

Node feature representation: Following the PowerGraph benchmark specification, node features use apparent power magnitude S_{net} rather than signed reactive power Q_{net} . While $S_{\text{net}} = \sqrt{P^2 + Q^2}$ loses the sign distinguishing inductive from capacitive loads, this information is partially recoverable: (1) reactive power flow Q_{ij} on edges retains sign and is used for line flow prediction, and (2) the relationship between P_{net} , S_{net} , and neighboring bus voltages allows implicit learning of power factor effects. Future work could investigate whether including signed Q_{net} as an additional node feature improves voltage prediction accuracy, particularly for grids with significant reactive power compensation.

Simulated operating state features: The PowerGraph benchmark provides node features (including S_{net}) extracted from simulated AC power flow operating states. For PV (generator) buses, reactive power Q is a dependent variable determined during power flow solution, meaning $S_{\text{net}} = \sqrt{P^2 + Q^2}$ reflects the solved operating point rather than pre-contingency setpoints. This is inherent to simulation-based benchmarks that provide grid snapshots from converged power flow solutions. Real-world deployment would use SCADA/PMU measurements of actual operating conditions, which similarly reflect the current (solved) system state rather than hypothetical pre-solve quantities. Future work could investigate whether using only P_{net} for PV buses improves generalization to unseen operating conditions, or whether the current formulation’s implicit encoding of operating state is beneficial for learning realistic grid behavior.

Bus type and voltage angle representation: Following the PowerGraph benchmark specification, our formulation does not include explicit bus type indicators (PV/PQ/Slack) in node features. In standard AC power flow, PV buses have

specified voltage magnitude (generator setpoint) while PQ buses have voltage as a state variable. This omission renders the power flow prediction problem technically ill-posed in the classical sense, as voltage setpoints for PV buses are unspecified. Without bus type encoding, the model must implicitly learn bus characteristics from injection patterns—generators typically exhibit large positive active power injections, while loads show negative injections—which represents a more challenging learning task compared to traditional solvers that receive explicit bus type specifications. Additionally, voltage angles are not included as inputs or evaluated as outputs; the model predicts only voltage magnitudes. While angles could be inferred from the trained model (our architecture supports angle prediction via sin/cos outputs), the PowerGraph benchmark does not provide angle ground truth for evaluation. Future work could extend the benchmark to include bus type encoding and angle targets, enabling more physically complete power flow surrogate models.

Static topology assumption: Our current framework assumes fixed grid topology during training and inference. Real power systems undergo frequent topology changes due to maintenance outages, equipment failures, and switching operations for loss minimization. Extending our approach to handle dynamic topology—for example, via graph structure learning [51] or topology-conditional embeddings—would improve practical applicability. Additionally, investigating how well models generalize to unseen topologies (e.g., training on IEEE 24-bus, testing on IEEE 30-bus) would clarify the transferability of learned physics principles across grid structures.

Limited out-of-distribution evaluation: We evaluated robustness only under load scaling (1.0–1.3× nominal injections). Other critical distribution shifts remain unexplored: (1) Measurement noise and missing data, (2) Seasonal and diurnal load pattern variations, (3) High renewable penetration with intermittent generation, (4) Topology perturbations (line outages), and (5) Extreme weather events causing correlated failures. A comprehensive robustness evaluation across multiple OOD dimensions is necessary to establish reliability bounds for operational deployment. Our load scaling results (Table XVI) use 5 random seeds matching our main experiments, providing statistical confidence in the observed robustness patterns.

Computational cost of pretraining: While inference is fast, self-supervised pretraining requires 30 minutes to 2 hours of GPU time depending on grid size. For utilities with hundreds of distinct network models (e.g., different seasonal configurations, multiple voltage levels), pretraining all models could incur significant computational cost. Investigating few-shot transfer learning—pretraining once on a large representative grid, then fine-tuning with minimal data on similar grids—could amortize pretraining costs across multiple deployment scenarios.

Explainability depth: Our Integrated Gradients evaluation demonstrates that the model correctly ranks critical edges (0.93 AUC-ROC), but does not provide *mechanistic* explanations of *why* specific lines are vulnerable. Causal discovery

methods [52] or attention mechanism interpretation [53] could yield deeper insights into learned failure propagation patterns, enabling operators to understand not just *which* lines matter but *how* cascades propagate through grid topology.

Decomposing physics and SSL contributions: Our framework combines two sources of improvement: physics-guided encoder architecture (electrically-parameterized message passing) and SSL pretraining (masked reconstruction). While we ablate encoder architectures in Section VI-G and compare SSL vs. scratch training throughout, we do not directly compare our masked reconstruction against alternative SSL strategies (e.g., contrastive learning, link prediction). The masked reconstruction approach follows GraphMAE [29] applied to power-relevant features; investigating whether physics-specific pretext tasks (e.g., predicting power balance violations or line flow directions) would outperform generic masking remains future work. Nonetheless, our ablations demonstrate that both encoder architecture and pretraining contribute meaningfully to low-label performance.

Ablation completeness: Our ablation strategy follows standard practice by varying one factor at a time: encoder architecture under scratch training (Table XIII) and pretraining strategy with our physics-guided encoder (Tables VIII–XI). A full factorial design crossing all encoder types with all pretraining strategies would require $3 \times 3 = 9$ conditions (physics/vanilla/GCN encoders \times our SSL/generic SSL/scratch). While we do not run all combinations, our existing ablations demonstrate that both factors contribute: physics-guided encoding provides +17.6% over GCN at 10% labels, and SSL pretraining provides +6.8% over scratch with the same encoder. We expect SSL benefits to transfer to other encoders, as masked reconstruction is architecture-agnostic; validating this expectation remains future work.

Learned vs. fixed physics weights: Our edge weighting y_{ij} is learned from electrical features (conductance, susceptance, reactance, thermal ratings) rather than directly using admittance magnitudes $|Y_{ij}|$ or inverse reactance $1/|x_{ij}|$. This design choice trades strict physics fidelity for task-adaptive flexibility: the model can learn which electrical properties matter most for each prediction task, rather than assuming a fixed relationship. Our encoder ablation (Table XIII) demonstrates that this electrically-parameterized approach outperforms both unweighted message passing and standard GCN at low label fractions. Future work could compare learned weights against fixed physics weights ($y_{ij} \propto 1/|x_{ij}|$) to quantify the trade-off between physics fidelity and task adaptability.

Static edge parameters: On fixed-topology benchmark grids, line parameters (reactance x_{ij} , thermal rating) are static infrastructure constants that do not vary across operating samples—only power injections and resulting flows change with load conditions. One might expect edge reconstruction to reduce to trivial memorization of static parameters. However, our masking strategy ablation (Table XIV) reveals a surprising finding: edge-only and node-only masking achieve *identical* downstream performance ($F1 = 0.882$ at 10% labels), both substantially outperforming scratch training ($F1 = 0.773$).

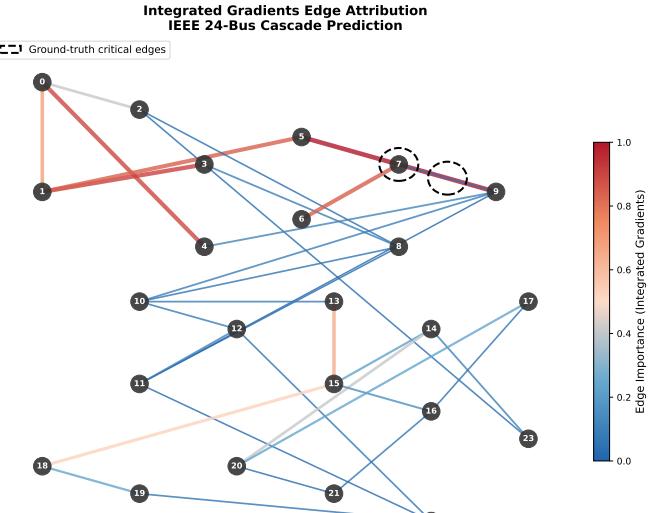


Fig. 10. Explainability visualization: Integrated Gradients correctly identifies critical transmission lines (red = high importance) that align with ground-truth failure explanations, achieving 0.93 AUC-ROC fidelity.

This suggests that reconstructing static edge parameters from masked context is *not* trivial—the encoder must still learn meaningful topological representations to predict which lines connect which buses with which impedances. The combined approach achieves best results ($F1 = 0.895$), indicating complementary information from node and edge signals. Future work could investigate whether time-varying edge features (line temperatures, real-time loading from SCADA) further improve pretraining quality.

VIII. CONCLUSION

We presented a physics-guided self-supervised learning framework for graph neural networks that addresses labeled data scarcity in power grid analysis. By embedding electrically-parameterized message passing into the encoder architecture—where edge weights are learned from line admittance features—and developing grid-specific pretext tasks (masked injection and parameter reconstruction), our approach learns representations from unlabeled operational data that transfer effectively to downstream tasks. Multi-seed validation across three prediction tasks on two grid scales confirms consistent improvements in low-label regimes: at 10% labeled data, self-supervised pretraining achieves 29.1% power flow error reduction, 26.4% line flow error reduction, and 6.8% F1-score improvement for cascading failure prediction on the IEEE 24-bus system. On the IEEE 118-bus system, pretraining dramatically stabilizes cascade prediction training, reducing F1 variance from ± 0.243 (scratch) to ± 0.051 (SSL)—a critical reliability improvement for deployment in safety-critical infrastructure where performance cannot depend on fortunate random initialization.

Beyond sample efficiency, our framework achieves 0.93 AUC-ROC explainability fidelity via Integrated Gradients,

correctly identifying critical transmission lines whose failures drive cascades. Robustness evaluation under load stress ($1.0\text{--}1.3 \times$ nominal) with 5-seed validation demonstrates that self-supervised representations generalize more effectively to out-of-distribution conditions (+3.6% relative advantage at $1.3 \times$ load). The consistent pattern across all tasks—largest gains when labeled data is most scarce, diminishing but persistent benefits even at full labels—validates our hypothesis that physics-guided self-supervised learning captures fundamental grid structure and electrical relationships that reduce dependence on task-specific supervision.

Future work should extend evaluation to real utility datasets with operational measurements, dynamic topology changes, and diverse distribution shifts (measurement noise, renewable variability, seasonal patterns). Investigating mechanistic explainability beyond edge importance ranking, developing few-shot transfer learning across multiple grids, and exploring continual learning as grids evolve would further advance practical deployment of machine learning for power system operations. Our framework demonstrates that self-supervised learning, when designed with domain physics in mind, provides a viable path toward sample-efficient, reliable, and interpretable machine learning for critical infrastructure.

ACKNOWLEDGMENTS

This work was conducted using the PowerGraph benchmark dataset. We thank the developers of PyTorch Geometric for their open-source graph neural network library.

REFERENCES

- [1] W. Huang, X. Pan, M. Chen, and S. H. Low, “DeepOPF-V: Solving AC-OPF problems efficiently,” *IEEE Transactions on Power Systems*, vol. 37, no. 1, pp. 800–803, 2022.
- [2] Y. Du, F. Li, J. Li, and T. Zheng, “Achieving 100x acceleration for N-1 contingency screening with uncertain scenarios using deep convolutional neural network,” *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 3303–3305, 2019.
- [3] X. Pan, T. Zhao, M. Chen, and S. Zhang, “DeepOPF: A deep neural network approach for security-constrained DC optimal power flow,” *IEEE Transactions on Power Systems*, vol. 36, no. 3, pp. 1725–1735, 2021.
- [4] S. Mohammadi, V.-H. Bui, W. Su, and B. Wang, “Surrogate modeling for solving OPF: A review,” *Sustainability*, vol. 16, no. 22, p. 9851, 2024.
- [5] Z. Hou, Y. He, Y. Cen, X. Liu, Y. Dong, E. Kharlamov, and J. Tang, “GraphMAE2: A decoding-enhanced masked self-supervised graph learner,” in *Proceedings of the ACM Web Conference 2023*. ACM, 2023, pp. 737–746.
- [6] J. Qiu, Q. Chen, Y. Dong, J. Zhang, H. Yang, M. Ding, K. Wang, and J. Tang, “GCC: Graph contrastive coding for graph neural network pre-training,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2020, pp. 1150–1160.
- [7] S. Liu, H. Wang, W. Liu, J. Lasenby, H. Guo, and J. Tang, “Pre-training molecular graph representation with 3d geometry,” in *International Conference on Learning Representations (ICLR)*, 2022.
- [8] J. Ji, J. Wang, C. Huang, J. Wu, B. Xu, Z. Wu, J. Zhang, and Y. Zheng, “Spatio-temporal self-supervised learning for traffic flow prediction,” vol. 37, pp. 4356–4364, 2023.
- [9] B. Donon, B. Donnot, I. Guyon, and A. Marot, “Graph neural solver for power systems,” in *IEEE International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.
- [10] S. Park and P. Van Hentenryck, “PDL-SCOPF: Self-supervised primal-dual learning for large-scale security-constrained DC optimal power flow,” *arXiv preprint arXiv:2311.18072*, 2023.
- [11] Y. Zhu, Y. Zhou, W. Wei, P. Li, and W. Huang, “Gnn’s generalization improvement for large-scale power system analysis based on physics-informed self-supervised pre-training,” *IEEE Transactions on Power Systems*, 2025.
- [12] T. N. Kipf and M. Welling, “Semi-supervised classification with Graph Convolutional Networks,” in *International Conference on Learning Representations (ICLR)*, 2017.
- [13] Z. Wu, Y. Xu, J. Wang, and C. Lyu, “Meta-learning enhanced physics-informed Graph attention convolutional network for power system state estimation,” *IEEE Transactions on Network Science and Engineering*, vol. 12, no. 1, pp. 1–14, 2025.
- [14] Z. Wu *et al.*, “Complex-valued graph neural networks for power system analysis,” *IEEE Transactions on Power Systems*, 2024.
- [15] P. Dogoulis, K. Tit, and M. Cordy, “KCLNet: Physics-informed power flow prediction via constraints projections,” *arXiv preprint arXiv:2506.12902*, 2025.
- [16] A. Thangamuthu, G. Kumar, S. Bishnoi, E. Rinke, and S. Guünnemann, “Unravelling the performance of physics-informed graph neural networks for dynamical systems,” in *NeurIPS Workshop on AI for Science*, 2022.
- [17] A. Varbella, K. Amara, B. Gjorgiev, M. El-Assady, and G. Sansavini, “PowerGraph: A power grid benchmark dataset for graph neural networks,” *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.
- [18] N. Lin, S. Orfanoudakis, N. Ordóñez Cardenas, J. S. Giraldo, and P. P. Vergara, “PowerFlowNet: Power flow approximation using message passing graph neural networks,” *International Journal of Electrical Power & Energy Systems*, vol. 160, p. 110112, 2024.
- [19] S. Liu, C. Wu, and H. Zhu, “Topology-aware graph neural networks for learning feasible and adaptive AC-OPF solutions,” *IEEE Transactions on Power Systems*, vol. 38, no. 5, pp. 4953–4966, 2023.
- [20] M. Yamizi *et al.*, “OPF-HGNN: Generalizable heterogeneous graph neural networks for AC optimal power flow,” *arXiv preprint arXiv:2403.00892*, 2024.
- [21] P. L. Donti, D. Rolnick, and J. Z. Kolter, “DC3: A learning method for optimization with hard constraints,” in *International Conference on Learning Representations (ICLR)*, 2021.
- [22] H. Lin and Y. Sun, “EleGNN: Electrical-model-guided graph neural networks for power distribution system state estimation,” in *Proc. IEEE Global Communications Conference (GLOBECOM)*, 2022, pp. 5292–5298.
- [23] A. Varbella, B. Gjorgiev, and G. Sansavini, “Physics-informed GNN for non-linear constrained optimization: PINCO a solver for the AC-optimal power flow,” *arXiv preprint arXiv:2410.04818*, 2024.
- [24] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, “Deep graph infomax,” in *International Conference on Learning Representations*, 2019.
- [25] F.-Y. Sun, J. Hoffmann, V. Verma, and J. Tang, “InfoGraph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization,” in *International Conference on Learning Representations*, 2020.
- [26] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, “Graph contrastive learning with augmentations,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 5812–5823.
- [27] S. Thakoor, C. Tallec, M. G. Azar, M. Azabou, E. L. Dyer, R. Munos, P. Veličković, and M. Valko, “Large-scale representation learning on graphs via bootstrapping,” in *International Conference on Learning Representations*, 2022.
- [28] J. Xia, L. Wu, J. Chen, B. Hu, and S. Z. Li, “SimGRACE: A simple framework for graph contrastive learning without data augmentation,” in *Proceedings of the ACM Web Conference 2022*. ACM, 2022, pp. 1070–1079.
- [29] Z. Hou, X. Liu, Y. Cen, Y. Dong, H. Yang, C. Wang, and J. Tang, “GraphMAE: Self-supervised masked graph autoencoders,” in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2022, pp. 594–604.
- [30] M. Ringsquandl *et al.*, “SafePowerGraph: Safety-aware evaluation of graph neural networks for transmission power grids,” *arXiv preprint arXiv:2407.15929*, 2024.
- [31] M. Raissi, P. Perdikaris, and G. Karniadakis, “Physics-informed neural networks: A deep learning framework,” *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.
- [32] G. Karniadakis *et al.*, “Physics-informed machine learning,” *Nature Reviews Physics*, vol. 3, pp. 422–440, 2021.

- [33] T. Beucler, M. Pritchard, S. Rasp, J. Ott, P. Baldi, and P. Gentine, “Enforcing analytic constraints in neural networks emulating physical systems,” *Physical Review Letters*, vol. 126, no. 9, p. 098302, 2021.
- [34] S. Greydanus, M. Dzamba, and J. Yosinski, “Hamiltonian neural networks,” *NeurIPS*, 2019.
- [35] G. S. Misyris, A. Venzke, and S. Chatzivasileiadis, “Physics-informed neural networks for non-linear system identification applied to power system dynamics,” in *IEEE PES General Meeting*. IEEE, 2020, pp. 1–5.
- [36] A. S. Zamzam and N. D. Sidiropoulos, “Physics-aware neural networks for distribution system state estimation,” *IEEE Transactions on Power Systems*, vol. 35, no. 6, pp. 4347–4356, 2020.
- [37] X. Hu, H. Hu, S. Verma, and Z.-L. Zhang, “Physics-guided deep neural networks for power flow analysis,” *IEEE Transactions on Power Systems*, vol. 36, no. 3, pp. 2082–2092, 2021.
- [38] J. Authier, R. Haider, A. Annaswamy, and F. Dörfler, “Physics-informed graph neural network for dynamic reconfiguration of power systems,” *Electric Power Systems Research*, vol. 237, p. 110938, 2024.
- [39] U.S.-Canada Power System Outage Task Force, “Final report on the august 14, 2003 blackout in the united states and canada: Causes and recommendations,” U.S. Dept. of Energy and Natural Resources Canada, Tech. Rep., 2004.
- [40] B. A. Carreras, V. E. Lynch, I. Dobson, and D. E. Newman, “Critical points and transitions in an electric power transmission model for cascading failure blackouts,” *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 12, no. 4, pp. 985–994, 2002.
- [41] M. J. Eppstein and P. D. H. Hines, “A “random chemistry” algorithm for identifying collections of multiple contingencies that initiate cascading failure,” *IEEE Transactions on Power Systems*, vol. 27, no. 3, pp. 1698–1705, 2012.
- [42] D. P. Nedic, I. Dobson, D. S. Kirschen, B. A. Carreras, and V. E. Lynch, “Criticality in a cascading failure blackout model,” *International Journal of Electrical Power & Energy Systems*, vol. 28, no. 9, pp. 627–633, 2006.
- [43] H. Cetinay *et al.*, “Analyzing cascading failures in power grids under ac and dc models,” *ACM SIGMETRICS*, 2018.
- [44] A. Varbella, B. Gjorgiev, and G. Sansavini, “Geometric deep learning for online prediction of cascading failures in power grids,” *Reliability Engineering & System Safety*, vol. 237, p. 109341, 2023.
- [45] M. Fey and J. E. Lenssen, “Fast graph representation learning with PyTorch Geometric,” in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.
- [46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 2019, pp. 4171–4186.
- [47] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International Conference on Machine Learning (ICML)*. PMLR, 2020, pp. 1597–1607.
- [48] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *International Conference on Machine Learning (ICML)*. PMLR, 2017, pp. 3319–3328.
- [49] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.
- [50] B. Donon, R. Clément, B. Donnot, A. Marot, I. Guyon, and M. Schoenauer, “Neural networks for power flow: Graph neural solver,” *Electric Power Systems Research*, vol. 189, p. 106547, 2020.
- [51] L. Franceschi, M. Niepert, M. Pontil, and X. He, “Learning discrete structures for graph neural networks,” in *International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 1972–1982.
- [52] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge University Press, 2009.
- [53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.