LEARNING AC OPTIMAL POWER FLOW BY A CAPPED REGRESSOR WITH TOPOLOGICAL ATTENUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficiently solving the Alternating Current Optimal Power Flow (AC-OPF) problem is essential in power systems. Despite substantial progress in learning-based AC-OPF, an important question remains: Which physics-motivated patterns can a learning-based load-to-generation predictor encode under the cost-minimization objective? To answer this question, we investigate the physical patterns of load-to-generation relationships: boundedness of total active-power generation and topological attenuation of generator response. Accordingly, we propose a capped regressor with topological attenuation (CReT). To encode the boundedness of total active-power generation, CReT controls the load-relative scale of total active-power generation by a capping operation, and then employs a label distribution learning model to proportionally allocate this total generation among individual generators. To encourage topological attenuation, CReT utilizes a novel physics-informed graph convolution encoding the nonlinear couplings in power flow, where the locality of message-passing mechanism is consistent with topological attenuation. Finally, CReT is evaluated against recent baselines on multiple bus systems.

1 INTRODUCTION

The Alternating Current Optimal Power Flow (AC-OPF) problem is a fundamental challenge in modern power systems. It is formulated as an optimization problem that aims to minimize generation costs while satisfying physical and operational constraints. Conventionally, experts rely on numerical optimization, such as interior-point methods, to determine the dispatch strategies of generators under given load conditions. However, the increasing penetration of renewable energy sources and the growing fluctuation in load demand have led to the modern power systems with more frequent state changes. While capable of yielding accurate solutions, repeatedly solving a nonconvex numerical optimization problem can become a latency bottleneck for modern power systems that require real-time responses. Consequently, researchers have turned to machine learning (ML) methods, which transform the intricate relationship between power load and optimal dispatch decisions into a learnable model. In this paradigm, a set of load data is first utilized to offline train a load-to-generation predictor, i.e., a mapping from load variables to generator outputs, and then the learned model can output candidate generator setpoints in real time by simply feeding in the current bus load data. This strategy transfers much of the computational burden from online optimization to offline training, thereby reducing online latency when approximation error and constraint violations are controlled.

Current learning-based AC-OPF works can broadly be categorized into hybrid and model-tailored approaches. Hybrid methods integrate ML as an auxiliary module within conventional numerical optimization methods, using it to reduce constraint sets Falconer & Mones (2023), accelerate solution pipelines Dong et al. (2020), or repair predicted solutions Lotfi & Pirnia (2022). Model-tailored approaches directly learn the load-to-generation model by physical penalties Fioretto et al. (2020), constrained layers Kim & Kim (2022), generative models Wang & Srikantha (2023), and reinforcement learning Yan & Xu (2020). Although these works have made great strides in the heuristic exploration of empirically effective model architectures, they typically overlook a more principled question: *Which physics-motivated patterns can a learning-based load-to-generation predictor encode under the cost-minimization objective?*

To address this question, we investigate the physical patterns of load-to-generation relationships, and accordingly propose a physics-motivated load-to-generation predictor. The overview of our proposal

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

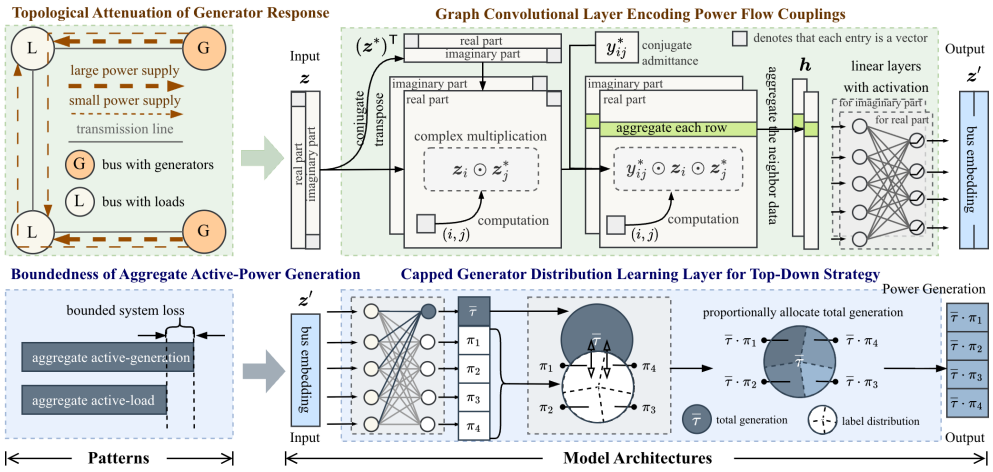

Figure 1: Overview of CReT: Physics-Motivated Patterns and Model Architectures.

is visualized in Figure 1. From a principled perspective, we propose two physics-motivated patterns: boundedness of total active-power generation and topological attenuation of generator response. The boundedness of total active-power generation concerns the global scale of active-power generation. Specifically, for a feasible operating point, total active-power generation equals total active-power load plus nonnegative active-power loss. Under bounded voltage and angle differences, together with a positive lower bound on total active-power load, this identity yields a sufficient uniform bound on the active-power loss ratio and a load-relative bound on total active-power generation. The topological attenuation of generator response concerns distance-conditioned response locality. Specifically, we observe an overall decline in generator response as the topological distance between generator and load grows, which is empirically formulated as topological attenuation assumption. From a methodological perspective, we propose CReT (a Capped Regressor with Topological attenuation) to encode the boundedness of total active-power generation and topological attenuation of generator response. Guided by the boundedness of total active-power generation, CReT adheres to a top-down architecture: The total active-power generation is first capped to restrict the active-power loss ratio, which is then proportionally allocated among individual generators using a label distribution learning modelGeng (2016). To respect the topological attenuation, we propose a power-flow-informed graph convolution (PFConv) encoding the nonlinear couplings of power flow, where buses and branches are treated as the nodes and edges of the graph, respectively. The message-passing of multiple graph convolutions repeatedly aggregates local neighborhood information, yet distant nodes exchange information only through deeper paths, which captures the topological attenuation. Finally, CReT is evaluated against recent learning-based baselines on multiple IEEE bus systems.

2 Related Work

Machine learning (ML) research for AC-OPF can be divided into hybrid approaches and model-tailored approaches. Hybrid approaches generally keep conventional numerical optimization methods during the real-time decision phase; instead, ML is embedded as an auxiliary module within the optimization process to reduce the constraint set, provide high-quality initial points, or repair infeasible solutions. For example, Falconer & Mones (2023) utilize ML to screen or predict relevant network constraints, thereby reducing a large number of redundant constraints and shortening the computation time of traditional solvers. Dong et al. (2020) use learned predictions to accelerate an AC-OPF simulation pipeline. Compared with hybrid approaches, model-tailored methods place greater emphasis on the generalization ability of the load-to-generation model and the feasibility of the output. Early research typically uses fully-connected neural networks or convolutional networks Canyasse et al. (2017); however, such unconstrained models lack guarantees on physical constraints. Hence, subsequent works introduce physical priors through penalty or dual terms Nellikkath & Chatzivasileiadis (2022); Wang & Srikantha (2024), projection or completion layers Kim & Kim (2022). DeepOPF-AL Pan et al. (2023) addresses a multi-valued load-to-generation mapping through augmented inputs,

while NGT Huang et al. (2025) exploits network reduction. Wang & Srikantha (2023); Li et al. (2022) use generative models together with physical penalties or specialized feasibility mechanisms to improve feasibility properties. Reinforcement learning-based approaches Yan & Xu (2020); Sayed et al. (2024) formulate real-time OPF as a Markov decision process. Recent studies Sayed et al. (2023); Nazari et al. (2025) explore graph reinforcement learning and transfer learning.

In addition, another important line of research focuses on learning AC-OPF by graph neural networks. Owerko et al. (2020) introduce localized graph neural network surrogates trained by supervised learning; their later unsupervised extension directly minimizes generation cost and handles network constraints through a differentiable barrier formulation Owerko et al. (2024). Liu et al. (2023) combine topology-aware graph filters with feasibility regularization and an efficient adaptation strategy for topology changes. Gao et al. (2024) derive a physics-guided graph-convolution kernel from the nodal OPF formulation. Ghamizi et al. (2024) instead uses a heterogeneous graph to represent different grid components and incorporates node-level constraints. Collectively, these studies focus on grid topology, feasibility information, or component heterogeneity. Complementing them, the graph convolution employed in this work is derived from power-flow coupling relationships, thereby endowing the model with enhanced physical interpretability.

3 BACKGROUND

The AC-OPF problem seeks to minimize the total cost of power generation given the load data, subject to a set of physical constraints. We adopt the following definition throughout this work:

Definition 3.1 (AC-OPF Problem)**.** *Let a power grid be denoted by a graph $(\mathcal{V}, \mathcal{E})$ whose nodes are buses and whose edges are transmission lines. Given the generation cost function $\mathcal{C}$, the active load p_i° and the reactive load q_i° in each bus, along with the line parameters (g_{ij}, b_{ij}) for each transmission line, the AC-OPF problem can be formulated as:*

$$
\begin{aligned}
\arg\min_{\{p_i^+\}_{i\in\mathcal{G}}^n \cup \{q_i^+\}_{i\in\mathcal{G}}^n \cup \{\upsilon_i\}_{i=1}^n \cup \{\theta_i\}_{i=1}^n} \ & \mathcal{C}(\boldsymbol{p}^+) \\
\text{s.t.} \quad & p_{ij} = g_{ij}\upsilon_i^2 - \upsilon_i\upsilon_j b_{ij}\sin(\theta_i-\theta_j) - \upsilon_i\upsilon_j g_{ij}\cos(\theta_i-\theta_j), \\
& q_{ij} = -b_{ij}\upsilon_i^2 - \upsilon_i\upsilon_j g_{ij}\sin(\theta_i-\theta_j) + \upsilon_i\upsilon_j b_{ij}\cos(\theta_i-\theta_j), \\
& p_i^+ - p_i^\circ = \sum\nolimits_{j\in\mathbb{N}_i} p_{ij}, \quad q_i^+ - q_i^\circ = \sum\nolimits_{j\in\mathbb{N}_i} q_{ij}, \quad i\in\mathcal{V}; \\
& \underline{\upsilon}_j \le \upsilon_j \le \overline{\upsilon}_j, \quad j\in\mathcal{V}; \quad \underline{p}_i^+ \le p_i^+ \le \overline{p}_i^+, \quad \underline{q}_i^+ \le q_i^+ \le \overline{q}_i^+, \quad i\in\mathcal{G}; \\
& \underline{\theta}_{ij} \le \theta_i - \theta_j \le \overline{\theta}_{ij}, \quad p_{ij}^2 + q_{ij}^2 \le \overline{s}_{ij}^2, \quad (i,j)\in\mathcal{E}.
\end{aligned}
\tag{1}
$$

Here n is the number of buses; $\mathcal{G}$ is the index set of generator buses; $\boldsymbol{p}^+$ is the vector of active-power generation. p_i^+ and q_i^+ denote the active- and reactive-power generation values in the i-th bus, respectively;[1] *υ_i and θ_i denote the voltage magnitude and phase angle of the i-th bus, respectively; p_{ij} and q_{ij} denote the active and reactive power flows in the transmission line between the i-th and j-th buses, respectively; $\mathbb{N}_i$ denotes the index set of neighbors of bus i. The constraints include AC power flow equations (the first three rows of the constraints in Equation* (1)*), bus voltage bounds, active and reactive generator power limits, angle phase difference limits, and transmission line thermal rating limits. g_{ij} and b_{ij} are the conductance and susceptance. Notation with overline or underline denotes the upper and lower bounds of the corresponding variable, respectively.*

The optimization problem in Definition 3.1 solves for the active- and reactive-power generation values, as well as the voltage magnitude and phase angle. This problem takes the values of active and reactive load as inputs, denoted by $\{p_i^\circ\}_{i=1}^n$ and $\{q_i^\circ\}_{i=1}^n$, along with the admittance of power grid and the bounds of each variable. However, load demand fluctuates in real time in practical power grid operations, whereas repeatedly solving the AC-OPF can become a computational bottleneck for each incoming load profile. To achieve real-time responsiveness, AC-OPF approaches based on machine learning model the optimal voltage and power generations as a function of the load inputs. Once trained, this function can be evaluated at low latency for new load scenarios, providing candidate operating points without repeatedly solving the full optimization problem. Formally,

[1] We use bus-indexed generation vectors throughout. For a bus without a generator, the corresponding active- and reactive-generation bounds are fixed at zero.

let $\boldsymbol{x} = [[p_1^\circ, q_1^\circ]; [p_2^\circ, q_2^\circ]; \ldots; [p_n^\circ, q_n^\circ]] \in \mathbb{R}^{n\times 2}$ denote a data point of load, where each element represents the active or reactive loads of the corresponding bus. The learning task is to train a model $f : \boldsymbol{x} \mapsto \{(\upsilon_i, \theta_i)\}_{i=1}^n \cup \{(p_i^+, q_i^+)\}_{i\in\mathcal{G}}$ that minimizes total generation cost while satisfying all physical constraints across the training dataset.

4 METHODOLOGY

4.1 PHYSICS-MOTIVATED PATTERNS

Boundedness of Total Active-Power Generation. Boundedness involves the scale of total active-power generation. For a feasible case, total active-power generation equals total active-power load plus nonnegative active-power loss. Under bounded voltage magnitudes and line-angle differences, together with a positive lower bound on total active-power load, we formally derive a sufficient uniform bound on the active-power loss ratio and a load-relative bound on total active-power generation:

Theorem 4.1 (Boundedness of Total Active-Power Generation)**.** *Let the active-power generation and load vectors be $\boldsymbol{p}^+$ and $\boldsymbol{p}^\circ$, respectively. Let g_{ij} and b_{ij} denote the conductance and susceptance with $g_{ij} \geq 0$. Every feasible power-balanced point satisfies $\mathbf{1}^\mathsf{T}\boldsymbol{p}^+ = \mathbf{1}^\mathsf{T}\boldsymbol{p}^\circ + \delta$, where $\delta_{ij} = g_{ij}(\upsilon_i^2 + \upsilon_j^2 - 2\upsilon_i\upsilon_j\cos(\theta_i - \theta_j)) \geq 0$ and $\delta = \sum_{(i,j)\in\mathcal{E}} \delta_{ij}$. If $\ell = \mathbf{1}^\mathsf{T}\boldsymbol{p}^\circ > 0$, then the total active-power generation can be rewritten as $\mathbf{1}^\mathsf{T}\boldsymbol{p}^+ = (1 + \ell^{-1}\delta)\ell = (1+\epsilon)\ell$, where $\epsilon \triangleq \ell^{-1}\delta \geq 0$. Moreover, for any family of operating points satisfying $\ell \geq \underline{\ell} > 0$, $0 \leq \underline{\upsilon} \leq \upsilon_i \leq \overline{\upsilon}$, and $|\theta_i - \theta_j| \leq \overline{\theta} \leq \pi$, the active-power loss ratio ϵ is uniformly bounded by $0 \leq \epsilon \leq \bar{\epsilon}$, where $\bar{\epsilon} = \underline{\ell}^{-1}\sum_{(i,j)\in\mathcal{E}} g_{ij}\left[(\overline{\upsilon} - \underline{\upsilon})^2 + 2\overline{\upsilon}^2(1 - \cos\overline{\theta})\right]$. Consequently, every feasible operating point in this family obeys $\mathbf{1}^\mathsf{T}\boldsymbol{p}^+ \in \left[\mathbf{1}^\mathsf{T}\boldsymbol{p}^\circ,\ (1+\bar{\epsilon})\mathbf{1}^\mathsf{T}\boldsymbol{p}^\circ\right]$.*

The proof is provided in Appendix A.1. Theorem 4.1 demonstrates that the active-power loss ratio is bounded by $\bar{\epsilon}$ under the stated prerequisites, where $\bar{\epsilon}$ is a sufficient worst-case bound and need not be small or tight since active-power loss minimization and cost minimization may not always be fully aligned: a low-cost generator can be electrically or topologically distant from major loads, so dispatching it may reduce the total generation cost while increasing active-power transmission losses.

Guided by the boundedness, CReT first constrains the total active-power generation within a physically plausible range and then allocates it among individual generators through label distribution learning, while the cost-driven training objective further refines the dispatch toward economic optimality. Empirically, Figure 2 shows the joint distribution of the total active-power load and the total active-power generation for solved AC-OPF instances. The visualization suggests a concentrated low active-power loss ratio in the two systems. Specifically, in IEEE 30/57-bus systems, the active-power losses are approximately 2% of the total active-power load.

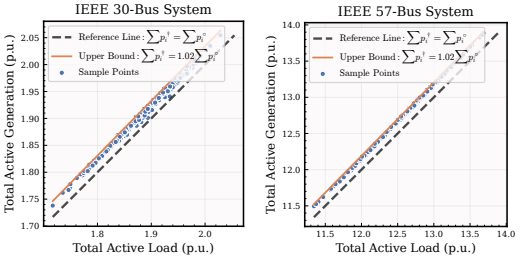

Figure 2: Joint Distribution of Total Active-Power Load and Optimal Active-Power Generation.

Topological Attenuation of Generator Response. According to the AC power-flow equations, power-flow coupling exhibits a localized pattern. The response of the optimal solution to load perturbations is jointly shaped by multiple factors, including generation costs, physical constraints, and topological distance. Hence, we characterize topological attenuation as a system-level tendency, rather than as a universal law governing the sensitivity between each individual generator-load pair.

Assumption 4.2 (Topological Attenuation of Generator Response)**.** *Let i be a load bus index, and let k be a generator bus index. Let $\mathrm{dist}(i,k)$ be the shortest-path distance on the bus graph. Define*

$$R_p(d) = \mathbb{E}\left[\left|\partial(p_k^+)^\star/\partial p_i^\circ\right| \,\middle|\, \mathrm{dist}(i,k) = d\right], \quad R_q(d) = \mathbb{E}\left[\left|\partial(q_k^+)^\star/\partial q_i^\circ\right| \,\middle|\, \mathrm{dist}(i,k) = d\right], \quad (2)$$

where $(\cdot)^\star$ denotes the optimal solution of the corresponding variables. Then $R_p(d)$ and $R_q(d)$ exhibit an overall decreasing trend as d increases.

It should be noted that Assumption 4.2 is neither strict monotonicity nor a pointwise inequality for each load–generator pair. Actually, Assumption 4.2 has potential connection to sensitivity-decay theory. An AC-OPF problem can be viewed locally as a graph-structured parametric nonlinear program. Existing sensitivity theory shows exponential decay of primal-dual solution sensitivity with graph distance under conditions including a locally unique regular solution, the linear independence constraint qualification, a strong second-order sufficient condition, and uniform conditioning of the graph-structured KKT system Shin et al. (2022). Under those hypotheses, a bound takes the form $\|\partial \boldsymbol{y}_i^\star / \partial \boldsymbol{x}_j\| \leq C\rho^{\mathrm{dist}_o(i,j)}$, where $C > 0$ and $0 < \rho < 1$. Here dist_o is the distance in the optimization graph induced by a chosen variable-constraint partition. Identifying that graph distance with bus shortest-path distance requires an additional modeling correspondence. Moreover, nonconvex AC-OPF can violate the regularity hypotheses through multiple local optima. Thus, sensitivity-decay theory motivates, but does not prove, Assumption 4.2.

To empirically assess the topological attenuation assumption, Figure 3 visualizes the response sensitivity of optimal generation to load changes for different topological distances between generators and loads. The response sensitivity is quantified by the partial derivative of optimal generation with respect to load, and approximated numerically via finite differences of generation or load. Specifically, for standard IEEE bus systems, we randomly perturb a single load and solve the corresponding optimal generation via the interior point method. Then the response sensitivity of each generator is computed as the absolute quotient of the generation variation divided by the load variation. This procedure is repeated 10,000 times to obtain the distribution presented in Figure 3, empirically validating the topological attenuation assumption.

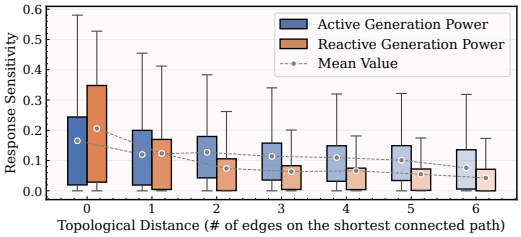

Figure 3: Distance-Conditioned Distribution of Generator Responses to Load Perturbations.

4.2 Model Architectures

Our proposed model architecture is formulated as $\tilde{\boldsymbol{p}}^+ = \phi(\psi(\boldsymbol{x}))$ and $\tilde{\boldsymbol{q}}^+ = \varphi(\psi(\boldsymbol{x}))$, where $\psi(\cdot)$ denotes the bus feature extraction network that maps raw input features ($[p_i^\circ, q_i^\circ]$) to the latent bus representation. The functions $\phi(\cdot)$ and $\varphi(\cdot)$ represent the mapping functions from bus representations to active and reactive power generations, respectively; $\tilde{\boldsymbol{p}}^+$ and $\tilde{\boldsymbol{q}}^+$ denote the predicted vectors of active and reactive power generations, respectively. Next, we illustrate the details of each module.

Physics-Informed Graph Convolution

To encode the topological attenuation into the model, we adopt a graph convolutional network (GCN) architecture. The adoption of GCN is motivated by the consistency between its inherent message-passing mechanism and the topological attenuation mechanism. In GCNs, information is aggregated from multi-hop neighbors through multiple convolution layers, which implicitly encourages the influence of neighboring nodes on a central node to diminish with topological distance. However, it should be noted that GCNs do not force the magnitude of that interaction to decrease monotonically: learned weights, nonlinearities, and the output normalization can amplify or suppress the magnitude of that interaction. To theoretically explain the consistency between GCN and topological attenuation, we derive the K-hop truncation bound.

Theorem 4.3 (K-Hop Truncation Bound)**.** *Let $\chi \subseteq \mathbb{R}^{n\times 2}$ be a convex set of load profiles, and fix a generator bus with index i. For each $\boldsymbol{x} \in \chi$, let $\boldsymbol{y}_i^\star(\boldsymbol{x}) = \left[(p_i^+)^\star(\boldsymbol{x}); (q_i^+)^\star(\boldsymbol{x})\right] \in \mathbb{R}^2$ denote the active and reactive generations at bus i in the AC-OPF solution. Assume that $\boldsymbol{y}_i^\star$ is differentiable on χ. Write $\boldsymbol{x}_j = [p_j^\circ, q_j^\circ]$ and suppose that $\sup_{\boldsymbol{x}\in\chi} \|\partial_{\boldsymbol{x}_j} \boldsymbol{y}_i^\star(\boldsymbol{x})\|_2 \leq c\rho^{\mathrm{dist}(i,j)}$ holds for each $j \in \mathcal{V}$, $c > 0$, and $0 < \rho < 1$. Let $s_i(d) = \{j \in \mathcal{V} : \mathrm{dist}(i,j) = d\}$ be an exact-distance shell. Let $\boldsymbol{x}, \boldsymbol{x}' \in \chi$ satisfy $\|\boldsymbol{x}_j - \boldsymbol{x}'_j\| = 0$ for $j \in \mathbb{N}_i^K$ and satisfy $\|\boldsymbol{x}_j - \boldsymbol{x}'_j\|_2 \leq b$ for $j \notin \mathbb{N}_i^K$, where $\mathbb{N}_i^K$ denotes the K-hop neighborhood of bus i. Then the following inequality holds:*

$$\|\boldsymbol{y}_i^\star(\boldsymbol{x}) - \boldsymbol{y}_i^\star(\boldsymbol{x}')\|_2 \leq bc \sum\nolimits_{d=K+1}^{\infty} |s_i(d)| \rho^d. \tag{3}$$

Further, if $|s_i(d)| \le a\beta^d$ *and* $\beta\rho < 1$*, then we have* $\|\boldsymbol{y}_i^\star(\boldsymbol{x}) - \boldsymbol{y}_i^\star(\boldsymbol{x}')\|_2 \le abc(\beta\rho)^{K+1}(1-\beta\rho)^{-1}$.

The proof is provided in Appendix A.2. Theorem 4.3 gives an explicit tail bound: when sensitivity decay dominates the growth of exact-distance shell, the effect of load variations beyond K hops on the generation at bus i decreases geometrically with K. Theorem 4.3 does not establish the topological attenuation assumption. Precisely, it identifies a sufficient condition under which the cumulative influence of remote load variations decays geometrically, which is consistent with the intuition underlying topological attenuation.

Next, we elaborate on the model design. Different from the traditional approach that directly utilizes the admittance matrix to construct a standard graph convolution operation, we aim to analyze the relationship between bus power injections and the neighbor power injections through the AC power flow equations, thereby developing a message-passing operation that better aligns with the physical characteristics of power grids. The AC power flow equations in complex form can be expressed as $s_i = \sum_{j\in\mathbb{N}_i} y_{ij}^* u_i u_j^*$, where s_i represents the complex power injection at node i, u denotes the complex voltage phasor, and y denotes the bus admittance. Motivated by this bilinear complex coupling, we let learned complex embeddings play the role of voltage-like latent factors. Accordingly, we define the graph convolution encoding power flow couplings:

$$\boldsymbol{h}_i^{(k)} = \sum\nolimits_{j\in\mathbb{N}_i} y_{ij}^* \boldsymbol{z}_i^{(k-1)} \odot (\boldsymbol{z}_j^{(k-1)})^*,\ \tilde{\boldsymbol{h}}_i^{(k)} = \mathfrak{D}(\boldsymbol{h}_i^{(k)}),\ \boldsymbol{z}_i^{(k)} = \xi\big(\tilde{\boldsymbol{h}}_i^{(k)}\boldsymbol{\omega}_{re}^{(k)}\big) + \mathtt{j}\xi(\tilde{\boldsymbol{h}}_i^{(k)}\boldsymbol{\omega}_{im}^{(k)}), \quad (4)$$

where $\boldsymbol{z}^{(0)} = \boldsymbol{p}^\circ + \mathtt{j}\boldsymbol{q}^\circ$ and $d_0 = 1$; $\boldsymbol{z}^{(k-1)} \in \mathbb{C}^{n\times d_{k-1}}$; $\boldsymbol{\omega}_{re}^{(k)} \in \mathbb{R}^{2d_{k-1}\times d_k}$ and $\boldsymbol{\omega}_{im}^{(k)} \in \mathbb{R}^{2d_{k-1}\times d_k}$ denote matrices of learnable parameters in the k-th layer; $\xi(\cdot)$ denotes the activation function; $\odot$ denotes the Hadamard product; $\mathfrak{D}(\boldsymbol{h}_i^{(k)}) = [\Re(\boldsymbol{h}_i^{(k)}), \Im(\boldsymbol{h}_i^{(k)})]$, $\Re(\cdot)$ and $\Im(\cdot)$ denote the operations for extracting the real and imaginary parts, respectively. $\mathbb{N}_i$ denotes the index set of neighborhood, including a self index when the corresponding diagonal admittance term is retained. Then the feature extraction network ψ can be obtained by a stack of multiple convolutional layers in Equation (4).

Capped Generator Distribution Learning

Motivated by the boundedness of total active-power generation, CReT first predicts a load-relative total active-power generation and then allocates the active-power generation among individual generator buses. Specifically, the active generation model $\phi : \boldsymbol{z} \mapsto \boldsymbol{p}^+$ is formalized as follows:

$$\alpha(\boldsymbol{x}) = \sigma(\boldsymbol{\omega}_r^\mathsf{T}\boldsymbol{p}^\circ) \in (0,1), \quad \tilde{\boldsymbol{z}} = (1+\alpha(\boldsymbol{x})\bar{t})\mathbf{1}^\mathsf{T}\boldsymbol{p}^\circ\pi(\bar{\boldsymbol{z}}\boldsymbol{\omega}_\phi), \quad \phi(\boldsymbol{z}) = \max(\underline{\boldsymbol{p}}^+, \min(\overline{\boldsymbol{p}}^+, \tilde{\boldsymbol{z}})), \quad (5)$$

where total active-power generation is estimated via a learnable factor $(1+\alpha(\boldsymbol{x})\bar{t})$; $\bar{t} \ge 0$ is a hyperparameter; $\sigma(\boldsymbol{\omega}_r^\mathsf{T}\boldsymbol{p}^\circ)$ denotes a sigmoid-activated linear layer with parameters of $\boldsymbol{\omega}_r$; $\max(\cdot,\cdot)$ and $\min(\cdot,\cdot)$ perform the element-wise maximum and minimum between the two input vectors, respectively. $\tilde{\boldsymbol{z}} = [\tilde{z}_1; \ldots; \tilde{z}_n] \in \mathbb{R}^n$ is the active-power generation vector before clipping. Subsequently, a label distribution learning model $\pi(\bar{\boldsymbol{z}}\boldsymbol{\omega}_\phi)$ allocates the total active-power generation to individual generators, where $\bar{\boldsymbol{z}}$ is obtained by a feature extraction network $\psi(\boldsymbol{x})$, $\boldsymbol{\omega}_\phi$ denotes a vector of learnable parameters, and $\pi(\cdot)$ denotes the softmax function. It should be noted that the clipping operation forcibly modifies the active-power outputs of certain generator buses, potentially resulting in an unexpected total active-power generation. We next analyze the role of clipping operation.

Proposition 4.4 (Clipping-Induced Deviation on Total Active-Power Generation)**.** *Consider a load input with positive total active-power load* $\ell \triangleq \mathbf{1}^\mathsf{T}\boldsymbol{p}^\circ > 0$*, the output* $\tilde{\boldsymbol{z}}$ *before clipping satisfies* $\mathbf{1}^\mathsf{T}\tilde{\boldsymbol{z}} = \big(1+\alpha(\boldsymbol{x})\bar{t}\big)\ell$*, i.e.,* $\ell^{-1}(\mathbf{1}^\mathsf{T}\tilde{\boldsymbol{z}} - \ell) = \alpha(\boldsymbol{x})\bar{t} \in [0,\bar{t}]$*. Moreover,* $\mathbf{1}^\mathsf{T}\tilde{\boldsymbol{p}}^+ = \big(1+\alpha(\boldsymbol{x})\bar{t}\big)\ell + \eta^- - \eta^+$*, where* $\eta^- = \sum_{i\in\mathcal{V}}(\underline{p}_i^+ - \tilde{z}_i)_+$ *and* $\eta^+ = \sum_{i\in\mathcal{V}}(\tilde{z}_i - \overline{p}_i^+)_+$*;* $(\cdot)_+ = \max(\cdot, 0)$.

The proof is provided in Appendix A.3. The capping operation controls the total active-power generation exactly before clipping. If no element is clipped, then $\eta^- = \eta^+ = 0$ and the same total active-power generation is preserved by the final output. Otherwise, η^- and η^+ show how much clipping changes the total active-power generation. Intuitively, capping operation is motivated by the identity that active-power generation surplus (total active-power generation minus total active-power load) equals active-power loss. But clipping operation and imperfect voltage recovery can break the correspondence. Hence, we study when predicted surplus is a reliable proxy for active-power loss.

Theorem 4.5 (Correspondence between Active-Power Loss and Predicted Surplus)**.** *Consider a scenario with* $\ell \triangleq \mathbf{1}^\mathsf{T}\boldsymbol{p}^\circ > 0$*. Let* $\tilde{u}$ *denote the recovered voltage. For a transmission line* (i,j)*,*

let $\tilde{p}_{ij}$ be the active power flowing from bus i to j, based on AC power flow equations. Let $\tilde{\delta} = \sum_{(i,j)\in\mathcal{E}} (\tilde{p}_{ij} + \tilde{p}_{ji})$ be the total active-power loss implied by $\tilde{u}$; let $r_i = \tilde{p}_i^+ - p_i^\circ - \sum_{j:\{i,j\}\in\mathcal{E}} \tilde{p}_{ij}$ be the active-power mismatch at bus i, $\boldsymbol{d} = \tilde{\boldsymbol{p}}^+ - \tilde{\boldsymbol{z}}$ be the change made by clipping, and $\mathfrak{b} = \{\boldsymbol{a} \in \mathbb{R}^n : \underline{p}_i^+ \le a_i \le \overline{p}_i^+,\ i \in \mathcal{V}\}$ be the set of outputs within active-generation constraints. We have $\ell^{-1}\tilde{\delta} = \alpha(\boldsymbol{x})\bar{t} + \ell^{-1}\mathbf{1}^\mathsf{T}\boldsymbol{d} - \ell^{-1}\mathbf{1}^\mathsf{T}\boldsymbol{r}$. Define the clipping-induced movement size by $\mathrm{dist}_2(\tilde{\boldsymbol{z}}, \mathfrak{b}) \triangleq \inf_{\boldsymbol{a}\in\mathfrak{b}} \|\tilde{\boldsymbol{z}} - \boldsymbol{a}\|_2$. Then $|\ell^{-1}\tilde{\delta} - \alpha(\boldsymbol{x})\bar{t}| \le \ell^{-1}(\sqrt{n}\ \mathrm{dist}_2(\tilde{\boldsymbol{z}}, \mathfrak{b}) + \|\boldsymbol{r}\|_1)$. Since $\ell \ge 0$ and $\alpha(\boldsymbol{x})\bar{t} \le \bar{t}$, we further have $0 \le \ell^{-1}\tilde{\delta} \le \bar{t} + \ell^{-1}(\sqrt{n}\ \mathrm{dist}_2(\tilde{\boldsymbol{z}}, \mathfrak{b}) + \|\boldsymbol{r}\|_1)$.

The proof is in Appendix A.4. Theorem 4.5 shows that active-power generation surplus approximately equals the active-power loss when clipping slightly changes the output and the predicted solution has small active-power mismatch. Finally, we analyze how this module affects the generalization.

Theorem 4.6 (Generalization Gap)**.** *Fix the feature extraction network ψ, and let ν denote the model in Equation* (5)*. Let q_ν be the number of learnable parameters in ν, and let e_ν and $\hat{e}_\nu$ be its expected and m-sample average approximation errors. Assume that the per-sample error is bounded by b and is κ-Lipschitz with respect to the learnable parameters, and the norm of learnable parameters is bounded by ϱ. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the draw of $m \ge 1$ independent samples, the following inequality holds over the mapping parameters:*

$$|e_\nu - \hat{e}_\nu| \le 2\kappa\varrho m^{-\frac{1}{2}} + b(2m)^{-\frac{1}{2}}\sqrt{q_\nu \log(1 + 2\sqrt{m}) + \log(4/\delta)}. \tag{6}$$

The proof is provided in Appendix A.5. Theorem 4.6 establishes a high-probability upper bound on the generalization gap. This bound is nondecreasing in the per-sample error bound b, the parameter-space Lipschitz constant κ, the parameter-norm radius ϱ, and the number q_ν of learnable parameters. Under a fixed loss function and a bounded target range, capping the total active-power generation can yield a smaller value of b, thereby tightening the resulting generalization gap.

In terms of the reactive generation, we directly employ a neural network to establish the mapping $\varphi : \bar{\boldsymbol{z}} \mapsto \boldsymbol{q}^+$, since there is no significant correlation between total reactive-power load and generation. $\varphi(\cdot)$ can be formalized as: $\varphi(\bar{\boldsymbol{z}}) = \underline{\boldsymbol{q}}^+ + (\overline{\boldsymbol{q}}^+ - \underline{\boldsymbol{q}}^+) \odot \sigma(\bar{\boldsymbol{z}}\boldsymbol{\omega}_\varphi)$, where $\boldsymbol{\omega}_\varphi$ denotes a vector of learnable parameters, and the calculation ensures the reactive generation within physical limits.

Training and Inference

Once the active- and reactive-power outputs of all generators are modeled, we can solve the AC power flow equations for the voltage using iterative methods, such as the Newton-Raphson method and the fast decoupled method. We denote the magnitude and phase angle of the solved voltage by $\tilde{v}$ and $\tilde{\theta}$, respectively. In this paper, we employ the Newton-Raphson method to obtain the voltage. Note that this paper does not focus on voltage prediction. Instead, we leverage existing methods for this purpose. Beyond the Newton-Raphson method, a neural network can also be directly used to produce the voltage, with a voltage calibration loss term added to the loss function. Finally, the following minimization objective is utilized to train the proposed models:

$$\min\ \mathcal{C}(\tilde{\boldsymbol{p}}^+) + \lambda_v \sum\nolimits_{i=1}^{n} \mathcal{L}(\tilde{v}_i) + \lambda_\theta \sum\nolimits_{(i,j)\in\mathcal{E}} \mathcal{L}(\tilde{\theta}_i - \tilde{\theta}_j) + \lambda_s \sum\nolimits_{(i,j)\in\mathcal{E}} \max(\tilde{p}_{ij}^2 + \tilde{q}_{ij}^2 - \overline{s}_{ij}^2, 0), \tag{7}$$

where $\mathcal{C}(\boldsymbol{p}^+)$ denotes the generation cost; $\mathcal{L}(v) = \max(v - \overline{v}, 0) + \max(\underline{v} - v, 0)$ measures the distance from v to the feasible interval $[\underline{v}, \overline{v}]$; $\tilde{p}_{ij}$ and $\tilde{q}_{ij}$ can be obtained through the power flow equations in Definition 3.1. During the inference stage, a candidate AC-OPF solution can be obtained through a single forward pass of the model. In terms of efficiency, given the L-layer graph convolution followed by capped generator distribution learning, where the k-th graph convolution layer maps d_{k-1} complex channels to d_k complex channels, the time complexity of the load-to-generation predictor is $\mathcal{O}(\sum_{k=1}^{L}[(n + nd_k + \sum_i^n |\mathbb{N}_i|)d_{k-1} + nd_k] + n + |\mathcal{G}|d_L)$. The derivation is provided in Appendix B.

5 Experiments

5.1 Experimental Configurations

The performance of the AC-OPF algorithms is evaluated on IEEE bus systems, including 9-bus, 14-bus, 24-bus, 30-bus, 57-bus, and 118-bus systems. To simulate realistic load variations, the active

Table 1: Algorithm Performance Based on Optimality and Feasibility on IEEE Bus Systems.

	Cost	$\mathcal{V}_{p^+}$	$\mathcal{V}_{q^+}$	$\mathcal{V}_v$	$\mathcal{V}_\theta$	$\mathcal{V}_\mathrm{S}$	$\mathcal{V}_{p^\circ}$	$\mathcal{V}_{q^\circ}$
	$(\times 10^3)$	IEEE 9-Bus System (Reference Cost: $5329.425_{\pm 26.570}$)						
AL	$4.583_{\pm 0.179}$	$0.000_{\pm 0.001}$	$0.000_{\pm 0.000}$	$0.005_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.145_{\pm 0.017}$	$0.018_{\pm 0.001}$	$0.215_{\pm 0.004}$
PTL	$4.580_{\pm 0.187}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.005_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.188_{\pm 0.022}$	$0.023_{\pm 0.001}$	$1.153_{\pm 0.005}$
PIL	$4.684_{\pm 0.250}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.005_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.147_{\pm 0.046}$	$0.025_{\pm 0.002}$	$1.150_{\pm 0.005}$
NGT	$4.566_{\pm 0.169}$	$0.006_{\pm 0.005}$	$0.000_{\pm 0.000}$	$0.005_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.183_{\pm 0.021}$	$0.020_{\pm 0.007}$	$0.169_{\pm 0.021}$
HGN	$4.650_{\pm 0.071}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.005_{\pm 0.000}$	$0.000_{\pm 0.000}$	$2.463_{\pm 0.345}$	$0.159_{\pm 0.003}$	$0.212_{\pm 0.001}$
CReT	$\mathbf{4.466}_{\pm 0.166}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\boldsymbol{0.005}_{\pm 0.000}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\mathbf{0.140}_{\pm 0.041}$	$\mathbf{0.009}_{\pm 0.001}$	$\mathbf{0.130}_{\pm 0.016}$
	$(\times 10^3)$	IEEE 14-Bus System (Reference Cost: $8105.961_{\pm 30.269}$)						
AL	$7.604_{\pm 0.449}$	$0.401_{\pm 0.294}$	$0.448_{\pm 0.084}$	$0.004_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.054_{\pm 0.006}$	$0.056_{\pm 0.004}$
PTL	$7.646_{\pm 0.712}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.004_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.048_{\pm 0.004}$	$0.101_{\pm 0.008}$
PIL	$9.167_{\pm 0.341}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.004_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.044_{\pm 0.001}$	$0.090_{\pm 0.001}$
NGT	$7.676_{\pm 0.478}$	$0.040_{\pm 0.047}$	$0.205_{\pm 0.014}$	$0.004_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.038_{\pm 0.001}$	$0.053_{\pm 0.006}$
HGN	$7.635_{\pm 0.198}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.004_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.031_{\pm 0.002}$	$0.052_{\pm 0.003}$
CReT	$\mathbf{7.512}_{\pm 0.547}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\boldsymbol{0.004}_{\pm 0.000}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\boldsymbol{0.031}_{\pm 0.003}$	$\mathbf{0.050}_{\pm 0.002}$
	$(\times 10^4)$	IEEE 24-Bus System (Reference Cost: $63348.877_{\pm 212.842}$)						
AL	$7.257_{\pm 0.105}$	$0.890_{\pm 0.025}$	$0.685_{\pm 0.021}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.002_{\pm 0.001}$	$0.418_{\pm 0.012}$	$0.638_{\pm 0.002}$
PTL	$6.376_{\pm 0.076}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.002_{\pm 0.001}$	$0.428_{\pm 0.014}$	$0.483_{\pm 0.008}$
PIL	$8.483_{\pm 0.018}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.192_{\pm 0.016}$	$0.478_{\pm 0.014}$	$0.395_{\pm 0.001}$
NGT	$6.404_{\pm 0.339}$	$0.866_{\pm 0.052}$	$0.965_{\pm 0.257}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.089_{\pm 0.057}$	$0.451_{\pm 0.026}$	$0.498_{\pm 0.078}$
HGN	$6.738_{\pm 0.119}$	$0.001_{\pm 0.001}$	$0.000_{\pm 0.001}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.438_{\pm 0.010}$	$0.538_{\pm 0.005}$
CReT	$\mathbf{6.273}_{\pm 0.021}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.405}_{\pm 0.012}$	$\mathbf{0.173}_{\pm 0.002}$
	$(\times 10^2)$	IEEE 30-Bus System (Reference Cost: $608.496_{\pm 3.555}$)						
AL	$5.331_{\pm 0.743}$	$0.022_{\pm 0.008}$	$0.055_{\pm 0.011}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.174_{\pm 0.026}$	$0.045_{\pm 0.007}$	$0.041_{\pm 0.011}$
PTL	$6.772_{\pm 0.041}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.196_{\pm 0.032}$	$0.059_{\pm 0.011}$	$0.068_{\pm 0.000}$
PIL	$6.833_{\pm 0.062}$	$0.002_{\pm 0.001}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.191_{\pm 0.035}$	$0.052_{\pm 0.013}$	$0.077_{\pm 0.000}$
NGT	$4.361_{\pm 0.733}$	$0.013_{\pm 0.009}$	$0.004_{\pm 0.002}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.172_{\pm 0.011}$	$0.044_{\pm 0.005}$	$0.039_{\pm 0.004}$
HGN	$6.860_{\pm 0.047}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.179_{\pm 0.002}$	$0.041_{\pm 0.000}$	$0.056_{\pm 0.001}$
CReT	$\mathbf{3.028}_{\pm 0.131}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\mathbf{0.164}_{\pm 0.024}$	$\mathbf{0.038}_{\pm 0.001}$	$0.060_{\pm 0.002}$
	$(\times 10^4)$	IEEE 57-Bus System (Reference Cost: $41842.125_{\pm 104.436}$)						
AL	$1.542_{\pm 0.075}$	$0.116_{\pm 0.041}$	$0.181_{\pm 0.041}$	$0.002_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.206_{\pm 0.029}$	$0.303_{\pm 0.046}$
PTL	$1.443_{\pm 0.081}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.002_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.204_{\pm 0.041}$	$0.296_{\pm 0.031}$
PIL	$1.441_{\pm 0.081}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.002_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.203_{\pm 0.036}$	$0.282_{\pm 0.031}$
NGT	$1.495_{\pm 0.042}$	$0.300_{\pm 0.217}$	$0.324_{\pm 0.118}$	$0.010_{\pm 0.002}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.138_{\pm 0.015}$	$0.160_{\pm 0.017}$
HGN	$1.697_{\pm 0.040}$	$0.000_{\pm 0.000}$	$0.010_{\pm 0.009}$	$0.002_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.116_{\pm 0.003}$	$0.178_{\pm 0.001}$
CReT	$\mathbf{1.321}_{\pm 0.038}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\boldsymbol{0.000}_{\pm 0.000}$	$0.005_{\pm 0.002}$	$\boldsymbol{0.000}_{\pm 0.000}$	$0.082_{\pm 0.198}$	$0.146_{\pm 0.084}$	$0.283_{\pm 0.152}$
	$(\times 10^5)$	IEEE 118-Bus System (Reference Cost: $129686.480_{\pm 106.848}$)						
AL	$1.304_{\pm 0.039}$	$0.220_{\pm 0.033}$	$0.569_{\pm 0.047}$	$0.002_{\pm 0.001}$	$0.000_{\pm 0.000}$	$0.007_{\pm 0.021}$	$0.360_{\pm 0.032}$	$0.528_{\pm 0.054}$
PTL	$1.221_{\pm 0.011}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.140_{\pm 0.008}$	$0.686_{\pm 0.001}$
PIL	$1.444_{\pm 0.096}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.143_{\pm 0.015}$	$0.670_{\pm 0.006}$
NGT	$1.145_{\pm 0.165}$	$0.139_{\pm 0.207}$	$0.380_{\pm 0.062}$	$0.002_{\pm 0.001}$	$0.000_{\pm 0.000}$	$0.067_{\pm 0.095}$	$0.289_{\pm 0.025}$	$0.366_{\pm 0.033}$
HGN	$1.306_{\pm 0.015}$	$0.000_{\pm 0.000}$	$0.001_{\pm 0.002}$	$0.001_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.049_{\pm 0.128}$	$0.381_{\pm 0.123}$	$0.451_{\pm 0.134}$
CReT	$\mathbf{1.006}_{\pm 0.029}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\boldsymbol{0.000}_{\pm 0.000}$	$\mathbf{0.134}_{\pm 0.004}$	$\mathbf{0.274}_{\pm 0.009}$

and reactive load powers at each bus are uniformly sampled from a range of 80% to 120% of their values specified in the respective benchmark cases. For each bus system, a dataset is constructed by generating 1,000 independent load scenarios, which is then partitioned into 700 samples for model training and 300 samples for testing. We assess the AC-OPF method from both optimality and feasibility perspectives. Optimality is assessed by the generation cost $\mathcal{C}(\boldsymbol{p}^+)$. Feasibility is assessed based on two components: the physical constraint violation, defined as the absolute deviation of a prediction from the corresponding feasible range (e.g., $\mathcal{L}(\tilde{v}_i)$ in Equation (7)), and the power imbalance, quantified by the residuals of the AC power-flow equations. The complete formulas are given in Appendix C. We employ five recent AC-OPF learning methods, including AL Pan et al. (2023), PTL Nazari et al. (2025), PIL Wang & Srikantha (2024), NGT Huang et al. (2025), HGN Ghamizi et al. (2024). For each comparison algorithm, the trade-off hyper-parameters are selected from $\{10^0, 10^1, \ldots, 10^7\}$ by a validation set separated from the training set, and the backbone network architectures align with their recommended implementations. In terms of our proposed CReT, the trade-off hyper-parameters λ_v, λ_θ, λ_s are also tuned from 10^0 to 10^7; the hyper-parameter $\bar{t}$ is selected from $\{0.01, 0.03, \ldots, 0.09\}$. The number of graph convolutional layers is selected from 4 to 10. To ensure a fair comparison, no post-calibration method is employed.

5.2 Results and Discussion

Table 1 presents the performance of the compared algorithms on the IEEE bus systems. Each performance metric cell contains the mean value and standard deviation of the corresponding algorithm,

with specific formatting used to highlight the performance rankings of our proposed CReT method. Specifically, bold text indicates that CReT achieves the best performance among all compared methods, while bold and italic text denotes a tied best performance. The reference cost is the generation cost returned by MIPS (MATPOWER Interior Point Solver). The experimental results demonstrate that CReT exhibits superior performance across the majority of the bus systems. Specifically, in the IEEE 9/14/24/30/118-bus systems, CReT attains the best or tied-best results on most reported metrics. However, the performance of CReT is less competitive on the IEEE 57-bus system. Specifically, we observe a higher degree of constraint violations in terms of voltage magnitude and line thermal limits. One plausible explanation is the challenges encountered during the training phase. On this bus system, the Newton-Raphson solver for voltage reconstruction typically requires a higher number of iterations to converge, and at times struggles to reach the desired convergence tolerance. This suboptimal reconstruction of the voltage during training directly hinders the learning process, making CReT prone to violate the voltage-related constraints such as voltage magnitude and line flow limits.

5.3 Ablation Study

We conduct ablation studies to verify the effectiveness of each component in the proposed CReT architecture, and present the experimental results in Figure 4. Owing to space limits, experimental results are reported on four representative bus systems, including two small-scale and two large-scale bus systems; four additional evaluation metrics are not shown in the main figure as all ablation algorithms demonstrate highly similar performance on these metrics. In each subfigure, "Baseline" employs a standard multi-layer perceptron (MLP) as the load-to-generation predictor, with voltage recovered via the Newton-Raphson method; "GConv" substitutes the MLP with two parallel conventional graph convolutional networks Kipf & Welling (2017), where the adjacency matrices are constructed by the real and imaginary parts of the admittance matrix; "PFConv" replaces the MLP with our proposed graph convolutional layer based on power flow equations, i.e., Equation (4); "CGDL" integrates our proposed capped generator distribution learning as the output layer of the MLP. The experimental results demonstrate that both "GConv" and "CGDL" typically enhance "Baseline" performance to varying degrees. "PFConv" typically outperforms "GConv", indicating the effectiveness of our proposed graph convolution. Finally, the integration of both "PFConv" and "CGDL" (i.e., the complete CReT algorithm) yields additional performance gains, albeit less pronounced than the individual contributions of "PFConv" or "CGDL".

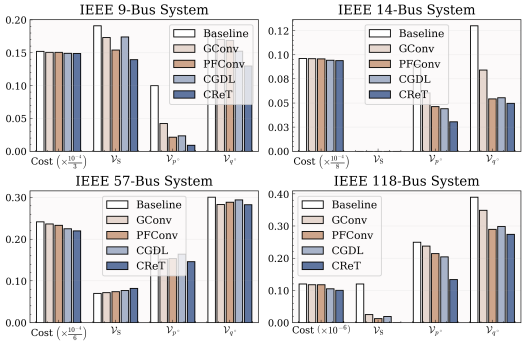

Figure 4: Ablation Performance of CReT.

6 Conclusion and Limitations

This work strives to answer a principled question: Given the cost-minimization objective, which physically motivated patterns can a learning-based load-to-generation predictor encode? To answer this question, we propose the boundedness of total active-power generation and topological attenuation of generator response, which are then utilized to establish CReT for learning AC-OPF candidate solutions. CReT constrains the total active-power generation through a two-stage prediction-and-allocation architecture and utilizes physical graph convolution to model topological attenuation, thereby improving the physical plausibility of predictions and the feasibility of constraints. Extensive experiments on standard IEEE bus systems demonstrate that CReT achieves competitive performance in generation cost while significantly reducing constraint violations. Despite its effectiveness, CReT has two potential scalability limitations: capturing rare long-range generator–load dependencies may require unnecessarily deep graph convolutions for predominantly local interactions, while its inference complexity may grow quadratically with the number of buses in densely connected networks. A more detailed limitation discussion is provided in Appendix D.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the writing of proofs, assisting with translation. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, supporting qualitative and thematic data analysis, and interpreting results, and cleaning and reformatting dataset are not applicable to this work. Additionally, we used generative AI tools for editing a research paper to improve readability, identifying relevant literature. We have reviewed all AI-assisted work. For example, we have checked whether the AI's translation has deviated from the actual meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Raphaël Canyasse, Gal Dalal, and Shie Mannor. Supervised learning for optimal power flow as a real-time proxy. In *IEEE Power & Energy Society Innovative Smart Grid Technologies Conference*, pp. 1–5, 2017.

Wenqian Dong, Zhen Xie, Gokcen Kestor, and Dong Li. Smart-PGSim: Using neural network to accelerate AC-OPF power grid simulation. In *International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–15, 2020.

Thomas Falconer and Letif Mones. Leveraging power grid topology in machine learning assisted optimal power flow. *IEEE Transactions on Power Systems*, 38(3):2234–2246, 2023.

Ferdinando Fioretto, Terrence W. K. Mak, and Pascal Van Hentenryck. Predicting AC optimal power flows: Combining deep learning and Lagrangian dual methods. In *AAAI Conference on Artificial Intelligence*, pp. 630–637, 2020.

Maosheng Gao, Juan Yu, Zhifang Yang, and Junbo Zhao. A physics-guided graph convolution neural network for optimal power flow. *IEEE Transactions on Power Systems*, 39(1):380–390, 2024.

Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28 (7):1734–1748, 2016.

Salah Ghamizi, A. Ma, J. Cao, and P. R. Cortes. OPF-HGNN: Generalizable heterogeneous graph neural networks for AC optimal power flow. In *IEEE Power & Energy Society General Meeting*, pp. 1–5, 2024.

Wanjun Huang, Minghua Chen, and Steven Low. Unsupervised learning for solving AC optimal power flows: Design, analysis, and experiment. In *IEEE Power & Energy Society General Meeting*, 2025.

Minsoo Kim and Hongseok Kim. Projection-aware deep neural network for DC optimal power flow without constraint violations. In *IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids*, pp. 116–121, 2022.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Yuxuan Li, Chaoyue Zhao, and Chenang Liu. Solving non-linear optimization problem in engineering by model-informed generative adversarial network (mi-gan). In *IEEE International Conference on Data Mining Workshops*, pp. 198–205, 2022.

Shaohui Liu, Chengyang Wu, and Hao Zhu. Topology-aware graph neural networks for learning feasible and adaptive AC-OPF solutions. *IEEE Transactions on Power Systems*, 38(6):5660–5670, 2023.

Amir Lotfi and Mehrdad Pirnia. Constraint-guided deep neural network for solving optimal power flow. *Electric Power Systems Research*, 211:108353, 2022.

Dorsa Nazari, Pirathayini Srikantha, Marzia Zaman, and Darshana Upadhyay. Unsupervised physics-informed transfer learning for power flow. In *IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids*, pp. 1–6, 2025.

Rahul Nellikkath and Spyros Chatzivasileiadis. Physics-informed neural networks for AC optimal power flow. *Electric Power Systems Research*, 212:108412, 2022.

Damian Owerko, Fernando Gama, and Alejandro Ribeiro. Optimal power flow using graph neural networks. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5930–5934, 2020. doi: 10.1109/ICASSP40776.2020.9053140.

Damian Owerko, Fernando Gama, and Alejandro Ribeiro. Unsupervised optimal power flow using graph neural networks. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6885–6889, 2024. doi: 10.1109/ICASSP48485.2024.10446827.

Xiang Pan, Wanjun Huang, Minghua Chen, and Steven H. Low. DeepOPF-AL: Augmented learning for solving AC-OPF problems with a multi-valued load-solution mapping. In *ACM International Conference on Future Energy Systems*, pp. 42–47, 2023.

Ahmed Rabee Sayed, Cheng Wang, Hussein I. Anis, and Tianshu Bi. Feasibility constrained online calculation for real-time optimal power flow: A convex constrained deep reinforcement learning approach. *IEEE Transactions on Power Systems*, 38(6):5215–5227, 2023.

Ahmed Rabee Sayed, Xian Zhang, Guibin Wang, Cheng Wang, and Jing Qiu. Optimal operable power flow: Sample-efficient holomorphic embedding-based reinforcement learning. *IEEE Transactions on Power Systems*, 39(1):1739–1751, 2024.

Sungho Shin, Mihai Anitescu, and Victor M. Zavala. Exponential decay of sensitivity in graph-structured nonlinear programs. *SIAM Journal on Optimization*, 32(2):1156–1183, 2022. doi: 10.1137/21M1391079.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2 edition, 2026.

Junfei Wang and Pirathayini Srikantha. Fast optimal power flow with guarantees via an unsupervised generative model. *IEEE Transactions on Power Systems*, 38(5):4593–4604, 2023.

Junfei Wang and Pirathayini Srikantha. Data-driven ac optimal power flow with physics-informed learning and calibrations. In *IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pp. 289–294, 2024.

Ziming Yan and Yan Xu. Real-time optimal power flow: A Lagrangian based deep reinforcement learning approach. *IEEE Transactions on Power Systems*, 35(4):3270–3273, 2020.

A Proof of Theorems

A.1 Proof of Theorem 4.1

Proof. The power injected from bus i is $p_{ij} = g_{ij}v_i^2 - v_iv_j\,(g_{ij}\cos(\theta_i-\theta_j) + b_{ij}\sin(\theta_i-\theta_j))$, and $p_{ji} = g_{ij}v_j^2 - v_iv_j\,(g_{ij}\cos(\theta_i-\theta_j) - b_{ij}\sin(\theta_i-\theta_j))$. Thus, the active-power loss on line (i,j) is $\delta_{ij} = p_{ij}+p_{ji} = g_{ij}[v_i^2+v_j^2-2v_iv_j\cos(\theta_i-\theta_j)] = g_{ij}|u_i-u_j|^2 \geq 0$, where the last inequality follows from $g_{ij} \geq 0$. At each bus, feasibility gives the active-power balance $p_i^+ - p_i^\circ = \sum_{j:\{i,j\}\in\mathcal{E}} p_{ij}$. Summing over all buses yields $\mathbf{1}^\mathsf{T}\boldsymbol{p}^+ - \mathbf{1}^\mathsf{T}\boldsymbol{p}^\circ = \sum_{(i,j)\in\mathcal{E}}(p_{ij}+p_{ji}) = \sum_{(i,j)\in\mathcal{E}}\delta_{ij} = \delta$. When $\ell = \mathbf{1}^\mathsf{T}\boldsymbol{p}^\circ > 0$, $\epsilon = \delta/\ell \geq 0$. For each line, $v_i^2+v_j^2-2v_iv_j\cos(\theta_i-\theta_j) = (v_i-v_j)^2 + 2v_iv_j\,(1-\cos(\theta_i-\theta_j))$. Under the assumed operating bounds, $(v_i-v_j)^2 \leq (\overline{v}-\underline{v})^2$ and $v_iv_j \leq \overline{v}^2$. Since $1-\cos x$ depends only on $|x|$ and is nondecreasing for $|x|\in[0,\pi]$, the assumption $|\theta_i-\theta_j| \leq \overline{\theta} \leq \pi$ gives $1-\cos(\theta_i-\theta_j) \leq 1-\cos\overline{\theta}$. Combining these inequalities gives $\delta_{ij} \leq g_{ij}\left[(\overline{v}-\underline{v})^2 + 2\overline{v}^2(1-\cos\overline{\theta})\right]$. Summing over all lines yields $\delta \leq \zeta$, where $\zeta = \sum_{(i,j)\in\mathcal{E}} g_{ij}\left[(\overline{v}-\underline{v})^2 + 2\overline{v}^2(1-\cos\overline{\theta})\right]$. Because $\ell \geq \underline{\ell} > 0$, we have $0 \leq \epsilon = \ell^{-1}\delta \leq \ell^{-1}\zeta \leq \underline{\ell}^{-1}\zeta = \bar{\epsilon}$. □

A.2 Proof of Theorem 4.3

Proof. Let $\boldsymbol{x}(t) = \boldsymbol{x}' + t(\boldsymbol{x} - \boldsymbol{x}')$ for $t \in [0, 1]$. Convexity of χ ensures $\boldsymbol{x}(t) \in \chi$. Because the two load profiles agree within K hops of generator bus i, $\boldsymbol{x}_j - \boldsymbol{x}'_j = 0$ whenever $\operatorname{dist}(i, j) \leq K$. Applying the fundamental theorem of calculus to the two-dimensional generation map defined in Theorem 4.3 therefore gives

$$\boldsymbol{y}_i^\star(\boldsymbol{x}) - \boldsymbol{y}_i^\star(\boldsymbol{x}') = \int_0^1 \sum\nolimits_{j:\operatorname{dist}(i,j)>K} \partial_{\boldsymbol{x}_j} \boldsymbol{y}_i^\star(\boldsymbol{x}(t))(\boldsymbol{x}_j - \boldsymbol{x}'_j)\,\mathrm{d}t.$$

Here $\partial_{\boldsymbol{x}_j} \boldsymbol{y}_i^\star(\boldsymbol{x}(t)) \in \mathbb{R}^{2\times 2}$ is the Jacobian from the active and reactive loads at bus j to the active and reactive generations at bus i. Taking Euclidean norms, using the induced matrix 2-norm, applying the pointwise sensitivity bound and $\|\boldsymbol{x}_j - \boldsymbol{x}'_j\|_2 \leq b$ yield

$$\|\boldsymbol{y}_i^\star(\boldsymbol{x}) - \boldsymbol{y}_i^\star(\boldsymbol{x}')\|_2 \leq \int_0^1 \sum_{j:\operatorname{dist}(i,j)>K} \left\|\partial_{\boldsymbol{x}_j} \boldsymbol{y}_i^\star(\boldsymbol{x}(t))\right\|_2 \left\|\boldsymbol{x}_j - \boldsymbol{x}'_j\right\|_2 \mathrm{d}t \leq bc \sum_{d=K+1}^{\infty} |s_i(d)|\rho^d$$

If $|s_i(d)| \leq a\beta^d$ and $\beta\rho < 1$, we have $\|\boldsymbol{y}_i^\star(\boldsymbol{x}) - \boldsymbol{y}_i^\star(\boldsymbol{x}')\|_2 \leq abc(\beta\rho)^{K+1}(1-\beta\rho)^{-1}$. □

A.3 Proof of Proposition 4.4

Proof. The softmax weights sum to one, which gives $\mathbf{1}^\mathsf{T}\tilde{\boldsymbol{z}} = \big(1 + \alpha(\boldsymbol{x})\bar{t}\big)\ell$. For one coordinate, clipping any $r \in \mathbb{R}$ to an interval $[a, b]$ satisfies $\max(a, \min(b, r)) = r + (a - r)_+ - (r - b)_+$ for $a \leq b$. Summing this identity over all coordinates gives the final identity. □

A.4 Proof of Theorem 4.5

Proof. For each line, we have $\tilde{p}_{ij} + \tilde{p}_{ji} = g_{ij}|\tilde{u}_i - \tilde{u}_j|^2 \geq 0$. Summing the mismatch r_i over all buses gives $\mathbf{1}^\mathsf{T}\boldsymbol{r} = \mathbf{1}^\mathsf{T}\tilde{\boldsymbol{p}}^+ - \ell - \tilde{\delta}$. By the definition of $\boldsymbol{d}$, we have $\mathbf{1}^\mathsf{T}\tilde{\boldsymbol{p}}^+ - \ell = \mathbf{1}^\mathsf{T}\tilde{\boldsymbol{z}} - \ell + \mathbf{1}^\mathsf{T}\boldsymbol{d} = \alpha(\boldsymbol{x})\bar{t}\ell + \mathbf{1}^\mathsf{T}\boldsymbol{d}$. Substitution and division by ℓ prove the identities. Because element-wise clipping moves $\tilde{\boldsymbol{z}}$ to the nearest vector satisfying the output intervals, $\|\boldsymbol{d}\|_2 = \operatorname{dist}_2(\tilde{\boldsymbol{z}}, \mathfrak{b})$. The Cauchy-Schwarz and triangle inequalities imply $|\mathbf{1}^\mathsf{T}\boldsymbol{d}| \leq \sqrt{n}\,\|\boldsymbol{d}\|_2$ and $|\mathbf{1}^\mathsf{T}\boldsymbol{r}| \leq \|\boldsymbol{r}\|_1$. Applying these inequalities to the identity gives the stated upper bound. □

A.5 Proof of Theorem 4.6

Proof. For a model ν, let $\boldsymbol{\vartheta} \in \mathbb{R}^{q_\nu}$ be its parameter vector with $\|\boldsymbol{\vartheta}\|_2 \leq \varrho$. For a random load example ζ, let $f_{\nu,\boldsymbol{\vartheta}}(\zeta)$ be the per-sample approximation error. The parameter-sensitivity assumptions mean that $0 \leq f_{\nu,\boldsymbol{\vartheta}}(\zeta) \leq b$ and $\big|f_{\nu,\boldsymbol{\vartheta}}(\zeta) - f_{\nu,\boldsymbol{\vartheta}'}(\zeta)\big| \leq \kappa\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\|_2$ hold for any admissible $\boldsymbol{\vartheta}$ and $\boldsymbol{\vartheta}'$. For independent samples $\zeta_1, \ldots, \zeta_m$, the expected and average training errors are formally defined as $e_\nu(\boldsymbol{\vartheta}) = \mathbb{E}[f_{\nu,\boldsymbol{\vartheta}}(\zeta)]$ and $\hat{e}_\nu(\boldsymbol{\vartheta}) = \frac{1}{m}\sum_{k=1}^m f_{\nu,\boldsymbol{\vartheta}}(\zeta_k)$. Set $\tau = \varrho/\sqrt{m}$. Let $\mathcal{N}_\nu$ denote a finite set such that every admissible parameter vector is within Euclidean distance τ of some vector in $\mathcal{N}_\nu$. The covering numbers of the Euclidean ball Vershynin (2026) gives $|\mathcal{N}_\nu| \leq (1 + 2\sqrt{m})^{q_\nu}$. Indeed, take a maximal set of parameter vectors separated by more than τ, the disjoint Euclidean balls of radius $\tau/2$ around these vectors fit inside the ball of radius $\varrho + \tau/2$. Let $\boldsymbol{\vartheta}_0$ denote any fixed parameter vector in $\mathcal{N}_\nu$, and let $u > 0$ denote a possible difference between its expected and training approximation errors. Hoeffding's inequality and the displayed error range give $\Pr(|e_\nu(\boldsymbol{\vartheta}_0) - \hat{e}_\nu(\boldsymbol{\vartheta}_0)| > u) \leq 2\exp\big(-2mu^2/b^2\big)$. Based on the union bound, if a failure probability of $\delta/2$ is assigned to the mapping, the following inequality holds with probability at least $1 - \delta$:

$$|e_\nu(\boldsymbol{\vartheta}_0) - \hat{e}_\nu(\boldsymbol{\vartheta}_0)| \leq b\sqrt{\frac{q_\nu \log(1 + 2\sqrt{m}) + \log(4/\delta)}{2m}}. \tag{8}$$

For an arbitrary admissible $\boldsymbol{\vartheta}$, choose $\boldsymbol{\vartheta}_0 \in \mathcal{N}_\nu$ with $\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2 \leq \tau$. The parameter-sensitivity conditions bound both the expected-error change and the training-error change by $\kappa\tau = \kappa\varrho/\sqrt{m}$. The triangle inequality therefore adds $2\kappa\varrho/\sqrt{m}$ to Equation (8) and proves Equation (6). □

B Load-to-Generation Predictor Complexity of CReT

Proposition B.1 (Time Complexity of the Load-to-Generation Predictor in CReT)**.** *Let n be the number of buses, g the number of generator buses, and L the number of graph convolution layers. The input has width d_0, and the k-th graph convolution layer maps d_{k-1} complex channels to d_k complex channels. Let $s = \sum_{i=1}^{n} |\mathbb{N}_i|$ denote the number of retained nonzero admittance entries, including diagonal entries. For one load profile, the L-layer graph convolution feature extractor followed by capped generator distribution learning has time complexity*

$$\mathcal{O}\left(\sum_{k=1}^{L} [(s+n)d_{k-1} + nd_{k-1}d_k + nd_k] + n + gd_L\right). \tag{9}$$

Given $d_0 = 1$, if all hidden layers have width d, then the single-profile complexity can be written as

$$\mathcal{O}\big(L(sd + nd^2) + n + gd\big). \tag{10}$$

Proof. Consider graph convolution layer k. Multiplying the admittance matrix by all d_{k-1} feature channels costs $\mathcal{O}(sd_{k-1})$. The node-wise complex product, complex conjugation, and real-imaginary conversion cost $\mathcal{O}(nd_{k-1})$. Applying each of the real and imaginary linear maps to all buses costs $\mathcal{O}(nd_{k-1}d_k)$, while the activation and normalization operations cost $\mathcal{O}(nd_k)$. Thus, layer k has complexity $\mathcal{O}((s+n)d_{k-1} + nd_{k-1}d_k + nd_k)$. Summing this expression over the L graph convolution layers gives the first term in Equation (9). For capped generator distribution learning, summing the active loads and evaluating the total active-power generation scaling factor cost $\mathcal{O}(n)$. Evaluating the active-allocation and reactive-generation linear maps on the g generator embeddings costs $\mathcal{O}(gd_L)$. The sigmoid, softmax, elementwise scaling, and clipping operations cost $\mathcal{O}(g)$, which is absorbed by $\mathcal{O}(gd_L)$ because $d_L \geq 1$. Adding these output operations proves Equation (9). Substituting $d_1 = \cdots = d_L = d$ and collecting dominated linear terms gives Equation (10). □

Remark. When the admittance matrix is sparse and both the network depth and hidden dimensions remain moderate, power generation inference is highly efficient. For a power grid with e transmission lines, the number of nonzero admittance entries satisfies $s \leq n + 2e$. Therefore, for a family of sparse grids with $e = \mathcal{O}(n)$, we have $s = \mathcal{O}(n)$. If L and $d_1, d_2, \ldots, d_L$ are fixed, the complexity in Equation (9) reduces to $\mathcal{O}(n)$. In this regime, the computational cost of generation inference grows linearly with the number of buses, where sparse graph convolutions avoid the $\mathcal{O}(n^2)$ cost associated with dense $n \times n$ propagation. When all hidden layers have a common width d, the dense feature-transformation operations scale as $\mathcal{O}(Lnd^2)$. Consequently, increasing d can incur substantial computational overhead. Increasing the number of graph convolutional layers likewise raises the cost approximately linearly with L. Thus, if larger power grids require a network depth that grows with the graph diameter to capture long-range load–generation dependencies, the computational cost of generation prediction may no longer scale linearly with the number of buses. Efficiency may also deteriorate for densely connected systems. If $s = \mathcal{O}(n^2)$, the aggregation cost becomes $\mathcal{O}(n^2 d_{k-1})$. In this case, even with fixed depth and width, the computational cost of generation inference grows quadratically with the number of buses. Similarly, explicitly constructing the admittance matrix as a dense propagation operator eliminates the computational benefits of sparsity, even when the underlying grid is sparse. Overall, CReT achieves its greatest computational efficiency on sparse power grids with fixed or slowly growing network depth and moderate hidden width. However, its efficiency may decrease in highly interconnected grids, under wide hidden representations, or when the required network depth grows substantially with system size.

C Evaluation Measures

The feasibility of an AC-OPF method can be quantified through the following violation measures:

$$\mathcal{V}_{p^+} = \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} \max(p_i^+ - \overline{p}_i^+, 0) + \max(\underline{p}^+ - p_i^+, 0), \quad \mathcal{V}_{q^+} = \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} \max(q_i^+ - \overline{q}_i^+, 0) + \max(\underline{q}_i^+ - q_i^+, 0),$$

$$\mathcal{V}_v = \frac{1}{n}\sum_{i=1}^{n} \max(v_i - \overline{v}_i, 0) + \max(\underline{v}_i - v_i, 0), \quad \mathcal{V}_\theta = \frac{1}{|\mathcal{E}|}\sum_{(i,j)\in\mathcal{E}} \max(\theta_{ij} - \overline{\theta}_{ij}, 0) + \max(\underline{\theta}_{ij} - \theta_{ij}, 0),$$

$$\mathcal{V}_{\mathrm{S}} = \frac{1}{|\mathcal{E}|}\sum_{(i,j)\in\mathcal{E}} \max\left(\sqrt{p_{ij}^2 + q_{ij}^2} - \overline{s}_{ij}, 0\right), \quad \mathcal{V}_{p^\circ} = \frac{1}{n}\sum_{i=1}^{n} \Big|p_i^\circ - p_i^+ + \sum_{j=1}^{n} p_{ij}\Big|, \quad \mathcal{V}_{q^\circ} = \frac{1}{n}\sum_{i=1}^{n} \Big|q_i^\circ - q_i^+ + \sum_{j=1}^{n} q_{ij}\Big|.$$

D LIMITATION DISCUSSION

First, CReT presents a limitation in scenarios where the optimal power supply to a load bus must depend entirely on a distant generator bus. In such scenarios, CReT requires sufficiently deep graph convolutional layers to capture such long-distance dependencies. While such a deep architecture is essential for addressing this specific case, it may introduce considerable computational redundancy for the majority of other load–generator interactions that adhere to the topological attenuation within the same power grid. This highlights a fundamental trade-off between capturing rare patterns and maintaining overall computational efficiency. Second, the inference complexity of CReT scales with the number of nonzero entries in the bus admittance matrix. For typical sparse power grids, where $|\mathcal{E}| = \mathcal{O}(n)$, the admittance matrix satisfies that the number of nonzero entries is $\mathcal{O}(n)$; consequently, the inference complexity of CReT grows approximately linearly with the number of buses. Only in the worst-case scenario of a densely connected network, where $|\mathcal{E}| = \mathcal{O}(n^2)$, can both the number of nonzero entries and the corresponding inference complexity increase to $\mathcal{O}(n^2)$, potentially compromising computational efficiency.