MxGPS: Multiplex Graph Transformers for a Power Grid Foundation Model

Charilaos Papaioannou, Ioannis Tsantilas, Dimitris Giannakakos, Vasilis Michalakopoulos, Sotiris Pelekis, Vangelis Marinakis, Arsam Aryandoust, Antonello Monti, Ricardo J. Bessa, Perdo P. Vergara, Jochen Cremer, Elissaios Sarmas

***Abstract*—Single-task fine-tuning of graph neural networks (GNNs) for power grid problems exhibits a systematic failure mode: models that achieve the lowest in-distribution error degrade the most under topology shift. We term this *topology overfitting*: the tendency of task-specific gradient signals to encode relational structure particular to the training topologies rather than the underlying physics, causing models to fail on unseen grids despite strong in-distribution performance. To expose and address this failure mode, we introduce MxGPS (Multiplex GPS), a multiplex graph transformer that runs K task-specialised GPS branches over a shared node encoder, jointly trained on Static State Estimation (SSE) and AC Power Flow (PF) via a self-supervised pre-training and multi-task fine-tuning protocol, with a cross-branch attention module evaluated in ablation. The joint SSE+PF objective forces the shared encoder to simultaneously satisfy complementary gradient signals, preventing it from overfitting to topology-specific relational structure. Under a 3-fold sliding-window cross-validation spanning four unseen topologies (14-, 24-, 162-, and 300-bus), MxGPS attains 0% boundary violation rate (BVR) on all four zero-shot Power Flow topologies. Critically, models with substantially lower in-distribution PF error degrade by 190%–1400% under topology shift, whereas MxGPS degrades by only 39%, an inversion that directly implicates topology overfitting as the failure mechanism rather than insufficient model capacity. With only 1.6M parameters ($12\times$ fewer than the GridFM reference baseline), MxGPS demonstrates that multi-task joint training is a principled and parameter-efficient mechanism for topology-agnostic generalisation in power grid foundation models.**

***Index Terms*—Power grid, Graph neural networks, Graph transformers, Multi-task learning, State estimation, Power flow, Foundation models**

I. Introduction

THE fundamental objective of power grid operation is the reliable, secure, and cost-effective delivery of energy. Traditionally, the core computational challenges of grid operation, namely AC power flow (PF), static state estimation (SSE), optimal power flow (OPF), and N-k contingency analysis, are addressed by iterative numerical solvers such as Newton-Raphson and interior-point methods [1]. However, the high penetration of intermittent renewables and bidirectional flexible loads has introduced unprecedented uncertainty and multi-directional information flow [2]. While these solvers are accurate, they are computationally intensive, topology-specific, and ill-suited to the rapid situational changes that characterize modern grid operation.

Machine learning approaches, and graph neural networks (GNNs) in particular, have emerged as a promising complement to traditional solvers by learning surrogate models that generalise across operating conditions [3], [4]. GNNs are naturally suited to power grid networks: buses are nodes carrying electrical state variables, transmission lines are edges carrying admittance parameters, and the physics of the grid is expressed as algebraic constraints on these graph-structured quantities [5]. Despite this alignment, existing GNN approaches for power grids train separate models per task [2], [3], [6], precluding representation sharing across operational objectives and failing to exploit the physical consistency constraints shared across SSE, PF, and contingency analysis.

GridFM [7] takes a step toward a unified model by pre-training a General, Powerful, Scalable (GPS)-based graph transformer [8] on the Masked Grid Task (MGT): a self-supervised masked feature reconstruction objective analogous to masked autoencoders in vision [9] and BERT in language [10]. GridFM demonstrates that a single model pre-trained on a diverse grid portfolio can be fine-tuned with minimal labelled data for PF and SSE, transferring across topologies. However, GridFM remains a fundamentally single-task architecture: it learns one reconstruction objective with one set of edge attention patterns, conflating the distinct relational structures that PF (requiring admittance-aligned coupling) and SSE (requiring measurement-noise-robust structure) impose on the same physical topology. We argue this conflation is a structural limitation, not merely a scaling issue: SSE should up-weight reliable measurement buses and down-weight noisy ones, while PF should align edge attention with the admittance matrix. No single graph representation can optimally serve both objectives simultaneously, a limitation noted in the GridFM paper itself [7].

We propose **MxGPS** (Multiplex GPS), an architecture that makes task-specific relational specialisation explicit by running K parallel GPS branches over the same physical node set, each steered toward a different operational task by its own gradient signal. A shared node encoder provides a common physical representation before branching; a cross-branch attention module (evaluated in ablation as MxGPS-cross) enables the task experts to mutually regularise each other during training, analogous to cross-head attention in multi-head Transformers [11]. An AC power flow consistency loss, derived from the Bus Injection Model (BIM), is applied across all branches simultaneously as a shared physics-informed self-supervised signal. Critically, this joint multi-task objective acts as a structural mechanism for *topology-agnostic* generalisation: forced to simultaneously satisfy gradient signals from both SSE and PF, the shared encoder cannot overfit its

representations to any particular training topology's relational structure, yielding embeddings that transfer substantially better to unseen grids than those learned by single-task fine-tuning.

a) GNNs for power grids: GNNs have been applied to power flow, state estimation, optimal power flow [6], and security assessment [4], [12], approximating traditional solver outputs at a fraction of the computational cost, but typically in single-task, single-topology settings.

b) Graph transformers and self-supervised learning on graphs: The Transformer architecture [11] has been extended to graphs through several lines of work [13]. GraphGPS [8] combines a local message-passing network (GatedGCN; [14]) with a global Transformer layer, achieving strong performance across benchmarks [15]; its structural encodings, Random Walk Structural Encoding (RWSE) and Laplacian Positional Encoding (LapPE) [16], capture topological position, critical for power grids where bus roles are tied to graph-theoretic properties. Heterogeneous graph transformers [17] handle multi-type node and edge relations, relevant for grids with bus and generator nodes. On the self-supervised side, masked reconstruction pre-training [9], [10] yields transferable representations; GraphMAE [18] extends it to graphs via the Scaled Cosine Error (SCE) loss, which GridFM adapts for the MGT objective in place of Mean Squared Error (MSE). GridFM [7] adopts GPS as its backbone, making this architecture the natural starting point for our work.

c) Multi-task learning and multiplex networks: Shared representations across tasks provide implicit regularisation and improve generalisation [19]. MTGC3 [20] jointly optimises task collaboration and difficulty scheduling in multi-task GNN architecture search, while Graph Mixture of Experts (MoE) [21] explicitly models expert diversity to prevent representation collapse. A multiplex network represents the same node set across multiple relational layers, each capturing a different interaction type; cross-layer attention [22] has been studied for link prediction in such networks. In MxGPS, each branch is a relational layer specialised for one operational task, paralleling the expert decomposition of MoE, and cross-branch attention is the analogue of cross-layer interaction.

d) Physics-informed learning and foundation models: Physics-informed neural networks [23] enforce known physical laws as soft constraints during training; in power grids, the AC power balance equations (Eqs. 2–3) of the BIM are the natural constraints. The success of foundation models in language [24] and vision has motivated analogous efforts in scientific domains: GridFM is the first foundation model explicitly designed for power grids, pre-trained on a diverse grid portfolio generated by `gridfm-datakit` [25] with a BIM-based physics loss as a secondary pre-training objective. MxGPS extends this direction by introducing multi-task specialisation as a structural inductive bias and by applying the BIM constraint across all task branches throughout training, not only during pre-training.

The central finding of this work is the identification of *topology overfitting* as a structural failure mode in single-task GNN fine-tuning for power grids. We argue topology-agnostic behaviour is the more relevant criterion for a power grid foundation model, as operational value compounds when the model is deployed on grids absent from the training portfolio. The primary contributions of this work are as follows:

1) We identify and characterise **topology overfitting**: the models with the lowest in-distribution PF error degrade the most under topology shift (GNS by 790%, GPS by over 1300%), an accuracy inversion exposed by our sliding-window cross-validation protocol (Section V).
2) We propose MxGPS, a multiplex graph transformer with K task-specialised GPS branches over a shared encoder, jointly trained on SSE and PF with a shared BIM physics constraint, providing implicit regularisation against topology overfitting; a cross-branch attention module is evaluated in ablation (Section III).
3) We implement a two-phase training protocol: MGT self-supervised pre-training followed by multi-task fine-tuning with a linear probe phase (Section IV).
4) We demonstrate that multi-task joint training addresses topology overfitting: MxGPS achieves 0% BVR on all four zero-shot PF topologies and degrades by only 39% in PF accuracy under topology shift, versus 190%–1400% for all single-task baselines (Sections V and VI).

II. Problem Formulation

a) Grid graph: A power grid can be modelled as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where nodes represent buses and edges are lines and transformers connecting these buses. The graph is naturally heterogeneous since each node can have either no or multiple generators connected to it. Here, we transform the heterogeneous graph into a homogeneous one by representing node features with a fixed 14-dimensional feature vector that captures aggregate generator properties wherever they are present at a given bus $i \in \mathcal{V}$ in the grid:

$$\begin{aligned}\mathbf{x}_i = [&P_d,\ Q_d,\ P_g,\ Q_g,\ V_m,\ V_a,\ \mathbf{1}_{\text{PQ}},\\ &\mathbf{1}_{\text{PV}},\ \mathbf{1}_{\text{REF}},\ V_m^{\min},\ V_m^{\max},\ G_s,\ B_s,\ V_n]\end{aligned} \tag{1}$$

where P_d, Q_d are active/reactive load; P_g, Q_g active/reactive generation; V_m, V_a voltage magnitude and angle; $\mathbf{1}_{\text{PQ/PV/REF}}$ are bus type one-hot flags; $V_m^{\min/\max}$ are per-bus operational voltage limits; G_s, B_s are shunt admittance components; and V_n is the nominal voltage. Transmission lines and transformers form edge set $\mathcal{E}$ with 11-dimensional edge features including branch power flows, admittance parameters $(Y_{ff}, Y_{ft}, Y_{tt}, Y_{tf})$, tap ratio, angle limits, and thermal rating.

b) AC power balance: At every bus i, the AC power balance imposes the following constraints:

$$P_i = |V_i| \sum_{j \in \mathcal{N}(i)} |V_j| \big(G_{ij} \cos(\theta_i - \theta_j) + B_{ij} \sin(\theta_i - \theta_j)\big) \tag{2}$$

$$Q_i = |V_i| \sum_{j \in \mathcal{N}(i)} |V_j| \big(G_{ij} \sin(\theta_i - \theta_j) - B_{ij} \cos(\theta_i - \theta_j)\big) \tag{3}$$

where $G_{ij} + jB_{ij}$ are entries of the bus admittance matrix and $\mathcal{N}(i)$ is the neighbourhood of bus i in $\mathcal{G}$. Any physically valid operating point must satisfy Eqs. (2)–(3). Note: V_a and θ_i denote the same voltage angle; G_s, B_s in $\mathbf{x}_i$ are per-bus shunt admittances, while G_{ij}, B_{ij} are mutual admittances.

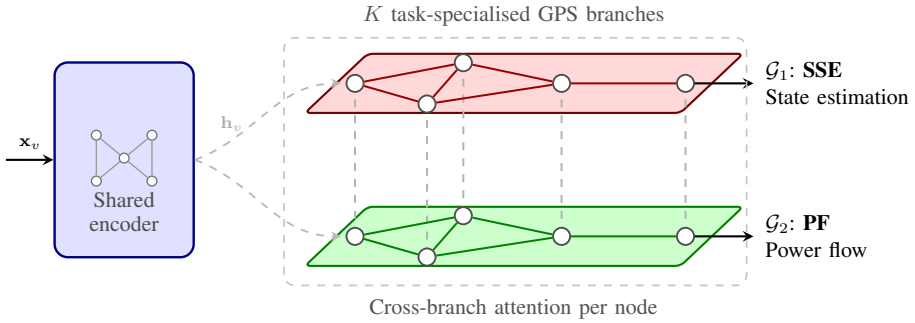

Fig. 1. MxGPS architecture with K=2 branches. A shared node encoder maps raw bus features $\mathbf{x}_v$ and structural positional encodings (RWSE, LapPE) to a common embedding $\mathbf{h}_v$, which is broadcast to both task-specialised GPS branches. Each branch processes the identical physical graph topology through its own GPS layers, conditioned by a learnable task token $\mathbf{t}_k$. Dashed vertical lines indicate the optional cross-branch attention (MxGPS-cross only), which exchanges per-node representations between branches at each GPS layer, providing mutual regularisation across tasks.

c) Tasks: We consider three tasks defined by different masking strategies applied to $\mathbf{x}_i$:

- **Masked Grid Task (MGT)**: self-supervised pre-training. Each feature dimension f of each bus i is masked independently with probability ρ; masked positions receive a learned [MASK] token. The model is trained to reconstruct the original bus features from the unmasked context.
- **Power Flow (PF)**: supervised regression. Given a known load profile and generator dispatch, predict voltage magnitudes and angles at all buses. Masking follows the physical observability structure: V_m and V_a are masked at PQ buses; Q_g and V_a at PV buses; P_g and Q_g at the reference (slack) bus, where the solver computes both to satisfy the power balance; Q_g is similarly an output at PV buses, where reactive dispatch is determined by the voltage setpoint.
- **Static State Estimation (SSE)**: supervised regression with noise. The ground truth is the clean AC power flow solution $\mathbf{x}^*$ produced by Newton-Raphson during dataset generation via `gridfm-datakit` [25]. Each measurable bus feature $(P_d, Q_d, P_g, Q_g, V_m, V_a)$ is first corrupted with additive Gaussian noise (std $\sigma_\varepsilon = 0.02$ p.u.) and then independently dropped with probability $\rho_{\text{mask}} = 0.2$; dropped positions receive a learned `[MASK]` token. These corruption levels are deliberately chosen as a stress test of robustness to noisy and missing measurements, rather than as a calibrated model of a specific metering infrastructure. The model is trained to recover the full clean state $\mathbf{x}^*$ at *all* buses, including both noisy and missing measurement positions.

III. Methodology

MxGPS has three components: a shared node encoder, K task-specialised GPS branches, and an optional cross-branch attention module evaluated in ablation (absent in MxGPS-ind, present in MxGPS-cross). In this work $K = 2$ (SSE and PF branches); the design is intended to extend to larger task portfolios by adding branches with new masking transforms and prediction heads, though we do not evaluate $K > 2$ here and treat it as future work. Figure 1 illustrates the overall architecture.

A. Shared Node Encoder

A single linear projection maps raw bus features and structural positional encodings to a shared d-dimensional embedding, providing a common physical representation before task-specific processing:

$$\mathbf{h}_i^{(0)} = \mathbf{W}_{\text{in}}\,\mathbf{x}_i + \mathbf{W}_{\text{pe}} \begin{bmatrix} \text{RWSE}_i \\ \text{LapPE}_i \end{bmatrix} \tag{4}$$

where $\text{RWSE}_i \in \mathbb{R}^r$ is the r-step Random Walk Structural Encoding [16] and $\text{LapPE}_i \in \mathbb{R}^{2p}$ is the top-p eigenvectors of the normalised graph Laplacian. Both encodings capture local and global structural roles of buses independent of their feature values, enabling cross-topology transfer [7], [8]. Both $\mathbf{W}_{\text{in}} \in \mathbb{R}^{d\times 14}$ and $\mathbf{W}_{\text{pe}} \in \mathbb{R}^{d\times(r+2p)}$ are learnable weight matrices, where d is the hidden embedding dimension (distinct from the demand subscript in P_d, Q_d). Masked input positions receive a learned [MASK] token embedding $\mathbf{m}$ in place of $\mathbf{W}_{\text{in}}\,\mathbf{x}_i$, following Devlin et al. [10] and GridFM.

B. Task Token Conditioning

A learnable task token $\mathbf{t}_k \in \mathbb{R}^d$ is added to the shared encoder output before branch k processes its input:

$$\mathbf{h}_i^{(0,k)} = \mathbf{h}_i^{(0)} + \mathbf{t}_k \tag{5}$$

This conditions each branch on its task identity without any architectural modification, analogous to task-prefix conditioning in language models; specialisation emerges from task-specific gradient steering through the branch-specific loss.

C. Task-Specialised GPS Branches

Each branch $k \in \{1, \ldots, K\}$ applies L GPS layers [8] to $\mathbf{h}^{(0,k)}$. Each GPS layer combines a local GatedGCN message-passing step [14] with a global Transformer self-attention step [11].

GatedGCN (local path). Edge gates $\hat{\eta}_{ij}$ are computed from node and edge embeddings and normalised over each node's

neighbourhood, where BN denotes Batch Normalisation and σ is the sigmoid function:

$$\mathbf{e}_{ij}^{(l+1)} = \mathbf{e}_{ij}^{(l)} + \mathrm{ReLU}\Big(\mathrm{BN}\Big(\mathbf{W}_1\mathbf{h}_i^{(l)} + \mathbf{W}_2\mathbf{h}_j^{(l)} + \mathbf{W}_3\mathbf{e}_{ij}^{(l)}\Big)\Big) \tag{6}$$

$$\hat{\eta}_{ij} = \frac{\sigma(\mathbf{e}_{ij}^{(l+1)})}{\sum_{j'\in\mathcal{N}(i)}\sigma(\mathbf{e}_{ij'}^{(l+1)}) + \varepsilon} \tag{7}$$

$$\mathbf{h}_i^{(l+\frac{1}{2})} = \mathbf{h}_i^{(l)} + \mathrm{ReLU}\left(\mathrm{BN}\left(\sum_{j\in\mathcal{N}(i)} \hat{\eta}_{ij} \odot \mathbf{W}_h\mathbf{h}_j^{(l)}\right)\right) \tag{8}$$

Global Transformer with edge bias (global path). Full $O(N^2)$ self-attention is applied over all buses in the graph; the attention logit between i and j is additively shifted by a linear projection of the edge embedding, allowing the Transformer to be directly aware of the physical connection between each bus pair:

$$a_{ij}^{(l)} = \frac{\big(\mathbf{Q}\,\mathbf{h}_i^{(l+\frac{1}{2})}\big)^\top\big(\mathbf{K}\,\mathbf{h}_j^{(l+\frac{1}{2})}\big)}{\sqrt{d_h}} + \phi\Big(\mathbf{e}_{ij}^{(l+1)}\Big) \tag{9}$$

where $\phi : \mathbb{R}^{d_e} \to \mathbb{R}^{n_{\text{heads}}}$ produces per-head additive logit biases, following the GraphGPS formulation [8].

GPS layer output. Local and global paths are fused through residual connections and LayerNorm, followed by a position-wise Feed-Forward Network (FFN) with Gaussian Error Linear Unit (GELU) activation, where MHA denotes multi-head attention:

$$\mathbf{h}_i^{(l+1)} = \mathrm{LN}\Big(\mathbf{h}_i^{(l+\frac{1}{2})} + \mathrm{MHA}\Big(\mathbf{h}^{(l+\frac{1}{2})}, a^{(l)}\Big) + \mathrm{FFN}(\cdot)\Big) \tag{10}$$

D. Cross-Branch Attention

When cross-branch attention is enabled, after each GPS layer l the representations of all K branches are exchanged per node:

$$\tilde{\mathbf{h}}_i^{(k,l)} = \mathbf{h}_i^{(k,l)} + \mathrm{CrossAttn}\Big(\mathbf{h}_i^{(k,l)},\ \big[\mathbf{h}_i^{(k',l)}\big]_{k'\neq k}\Big) \tag{11}$$

where CrossAttn uses branch k's representation as query and all other branches as keys and values. This is applied independently per node with complexity $O(K^2 \cdot d)$ per layer, lightweight for $K = 2$ and tractable as the task portfolio grows, allowing task experts to leverage complementary representations and propagate cross-task information through subsequent message-passing rounds.

E. Task Prediction Heads

Each branch k has its own two-layer MLP regression head producing per-bus predictions $\hat{\mathbf{y}}_i^{(k)} \in \mathbb{R}^4$ covering active and reactive power generation $(\hat{P}_{g,i}, \hat{Q}_{g,i})$ and voltage state $(\hat{V}_{m,i}, \hat{V}_{a,i})$.

F. Loss Function

The total training loss combines per-task supervised losses with an optional physics-informed self-supervised term and a boundary penalty:

$$\mathcal{L} = \sum_{k=1}^{K} \lambda_k \mathcal{L}_k^{\text{task}} + \lambda_{\text{phys}}\mathcal{L}_{\text{phys}} + \lambda_{\text{bnd}}\mathcal{L}_{\text{bnd}} \tag{12}$$

MGT pre-training loss. For the self-supervised pre-training task we use the Scaled Cosine Error (SCE) from GraphMAE [18] over masked bus feature positions $\mathcal{M}$:

$$\mathcal{L}_{\text{SCE}} = \frac{1}{|\mathcal{M}|}\sum_{(i,f)\in\mathcal{M}}\left(1 - \frac{\hat{x}_{if}\,x_{if}}{\|\hat{\mathbf{x}}_i\|\,\|\mathbf{x}_i\|}\right)^{\gamma} \tag{13}$$

where $\gamma = 2.0$ is a sharpening exponent that up-weights hard examples. SCE is preferred over MSE because it penalises directional rather than absolute deviation, better suiting normalised node features [18].

Supervised task loss. For PF and SSE, a dimension-weighted mean squared error is applied over the task-specific masked prediction positions, matching the gridfm-graphkit evaluation protocol:

$$\mathcal{L}_k^{\text{task}} = \sum_{f\in\{P_g,Q_g,V_m,V_a\}} w_f \cdot \frac{1}{|\mathcal{M}_k|}\sum_{i\in\mathcal{M}_k}\big(\hat{y}_{if} - y_{if}\big)^2 \tag{14}$$

Physics SSL loss (BIM). The Bus Injection Model (BIM) residual is applied to the predicted voltages of every branch simultaneously, providing a shared physics constraint:

$$\mathcal{L}_{\text{phys}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N}\Big[\big(\hat{P}_i^{(k)} - P_i^{\text{BIM}}\big)^2 + \big(\hat{Q}_i^{(k)} - Q_i^{\text{BIM}}\big)^2\Big] \tag{15}$$

where $P_i^{\text{BIM}}, Q_i^{\text{BIM}}$ are computed from predicted voltages via Eqs. (2)–(3) using the per-case bus admittance matrix. This is a key advantage over GridFM [7]: the physics constraint is applied to all task experts simultaneously, not only during the reconstruction pre-training phase.

Boundary loss. A soft penalty discourages voltage magnitudes outside operational limits $[V_{m,i}^{\min}, V_{m,i}^{\max}]$ available in the bus feature vector:

$$\begin{aligned}\mathcal{L}_{\text{bnd}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N}\Big[&\mathrm{ReLU}\big(V_{m,i}^{\min} - \hat{V}_{m,i}^{(k)}\big)^2 \\ &+ \mathrm{ReLU}\big(\hat{V}_{m,i}^{(k)} - V_{m,i}^{\max}\big)^2\Big]\end{aligned} \tag{16}$$

IV. Training Protocol

We follow a two-phase training protocol analogous to GridFM's linear probe fine-tuning curriculum [7].

Phase A: MGT pre-training. A single GPS model is pre-trained on the Masked Grid Task using the SCE loss plus a layered physics loss applied at each GPS layer, producing warm, general-purpose encoder and GPS layer weights before any task-specific gradient signal is introduced.

Phase B - MxGPS: multi-task fine-tuning. The pre-trained encoder and GPS layer weights are loaded into all K branches of MxGPS (all branches are initialised identically from the Phase A checkpoint). The shared encoder is frozen for the

first 30 epochs while only task tokens and prediction heads are trained (linear probe phase); the encoder is then unfrozen and the early-stopping budget resets for full fine-tuning, mirroring GridFM's linear probe protocol.

Phase B - Single-task baselines: task-specific fine-tuning. The same two-phase protocol is applied to all five single-task GNN baselines (GCN, GAT, GNS, GPS, and GNS-kit; see Section V-B) on PF and SSE. An MGT-pre-trained checkpoint is loaded via a weight transfer step that copies all tensors whose name and shape match the target model; the transfer is complete for all architectures evaluated here (GNS-kit's physics decoder heads are parameter-free). Single-task Phase B is trained end-to-end from the warm-started weights with no freeze phase, since all parameters are pre-trained.

V. Experiments

A. Datasets

We generate operational scenarios from standard MATPOWER test systems [26] (including grids from the PGLib-OPF benchmark [27]) using `gridfm-datakit` [25], which applies a hybrid load perturbation strategy (global temporal scaling from real load profiles combined with per-bus noise) followed by Newton-Raphson AC power flow [1]. Data is stored in Parquet format with per-bus, per-generator, and per-branch state variables. Approximately $10\,000$ scenarios are generated per case, and an $80/10/10\%$ train-validation-test split is applied deterministically by scenario ID.

Evaluation protocol. We order the eight IEEE cases [26] by size (14, 24, 30, 57, 73, 118, 162, 300 buses) and form three overlapping training windows of six consecutive cases: fold 1 trains on the 14–118-bus cases and holds out $\{162, 300\}$; fold 2 trains on 24–162 and holds out $\{14, 300\}$; fold 3 trains on 30–300 and holds out $\{14, 24\}$. Across folds this yields four unique zero-shot topologies: the 14- and 300-bus cases are zero-shot in two folds each, the 24- and 162-bus cases in one. MGT pre-training (Phase A) is repeated per fold on that fold's six training cases, so no zero-shot topology is seen at any training stage of the corresponding fold. For zero-shot cases, the per-case normaliser is fitted on a single representative scenario from that case rather than a training split, since zero-shot topologies contribute no training data and their full scenario set is reserved for evaluation; as with training cases, power features are scaled by the fixed per-case base MVA and only the nominal-voltage constant is inferred from the fitted scenario, following the per-case normaliser convention of `gridfm-datakit` [25]. Reported accuracy, BVR, and degradation figures are averaged over the folds in which each topology is zero-shot; the full scenario set of each zero-shot case is used for evaluation.

Normalisation. Bus power features are normalised by the per-case base apparent power (100 MVA for all cases). Voltage angles are converted from degrees to radians. A per-case normaliser is fitted on the training split and applied consistently to all evaluation sets [25].

B. Models and Baselines

We compare MxGPS against five single-task GNN baselines (each trained independently per task) and one analytical physics baseline. GNN baselines follow the two-phase GridFM protocol: MGT pre-training (Phase A) followed by task-specific fine-tuning (Phase B).

Physics baseline (DC PF). The linearised DC power-flow approximation: given bus active injections $(P_g - P_d)$, it solves $\mathbf{B}_{\text{red}}\,\boldsymbol{\theta} = \mathbf{P}_{\text{inj}}$ via sparse linear algebra for the voltage angles, while setting all voltage magnitudes to 1.0 p.u. The $\mathbf{B}$ matrix is reconstructed from the dataset's edge admittance features. Applied to both PF and SSE in-distribution settings. Newton-Raphson AC power flow was also implemented but excluded due to convergence failures on a subset of test cases.

GNN baselines.

- **GCN** [28]: Homogeneous graph convolutional network, 4 layers, hidden dimension 128.
- **GAT** [29]: Heterogeneous graph attention network with GATv2 convolutions [30], 4 layers, 4 heads, hidden dimension 128.
- **GNS**: Heterogeneous graph network with TransformerConv message passing over bus, generator, and edge node types; 6 layers, 8 heads, hidden dimension 256.
- **GPS** [8]: Single-task GPS (GatedGCN + global Transformer with edge-biased attention); 4 layers, 4 heads, hidden dimension 128, RWSE ($r = 16$) + LapPE ($p = 8$). GPS shares the same backbone architecture and hyperparameters as each individual MxGPS branch, making the GPS vs. MxGPS comparison a controlled ablation of single-task vs. joint multi-task training.
- **GNS-kit**: Exact replica of the gridfm-graphkit HGNS architecture [7], [25]; 12 layers, 8 heads, hidden dimension 48, trained with SCE + layered physics loss and gridfm-graphkit hyperparameters. The name distinguishes this model from the GridFM project as a whole; it is the closest proxy to GridFM we can build without access to the original pre-training corpus and checkpoints.

MxGPS is evaluated in two configurations: (i) **MxGPS-ind**, with $K = 2$ independent branches (SSE and PF) and no cross-branch attention, which we adopt as the *primary* configuration and refer to simply as MxGPS where unambiguous; and (ii) **MxGPS-cross**, which adds cross-branch attention after each GPS layer and serves as an ablation of that component. Both variants use hidden dimension 128, 4 GPS layers, 4 attention heads, and RWSE ($r = 16$) + LapPE ($p = 8$) structural encodings, trained jointly on SSE and PF batches in round-robin order.

C. Metrics

We report the following metrics per task: (1) MAE_{V_m}: mean absolute error on voltage magnitude (p.u.); (2) RMSE_{V_m}: root mean squared error on voltage magnitude; (3) MAE_{V_a}: mean absolute error on voltage angle (rad); (4) **BVR (%)**: Boundary Violation Rate, defined as the fraction of bus predictions where $\hat{V}_{m,i} \notin [V_{m,i}^{\min}, V_{m,i}^{\max}]$, measuring operational safety; (5) **Regression slope** and R^2: following Yang et al. [31], we fit a linear regression of predicted on true values per output channel and report the slope (ideally 1.0, indicating correct magnitude) and the Pearson R^2 (fraction of variance explained), detecting models that collapse toward the mean despite low MAE.

TABLE I
MGT IN-DISTRIBUTION ACCURACY AND CALIBRATION. MAE/RMSE IN P.U. EXCEPT MAE_{V_a} (RAD); SLOPE= 1, $R^2 = 1$ IDEAL. BEST PER COLUMN IN **BOLD**.

Model	MAE_{V_m}	MAE_{V_a}	RMSE_{V_m}	BVR (%)	Slope_{V_m}	$R^2_{V_m}$	Slope_{V_a}	$R^2_{V_a}$
Naive (V_m=1)	0.0333	—	0.0369	0.00	0.000	0.000	—	—
GCN	0.0230	**0.1213**	**0.0267**	**0.13**	0.202	**0.067**	0.127	**0.149**
GAT	0.0358	0.1922	0.0448	26.97	**0.378**	0.044	0.089	0.009
GPS	**0.0228**	0.2010	0.0282	14.93	0.019	0.000	−0.002	0.000
GNS	0.0445	0.1304	0.0586	19.76	−0.335	0.025	**0.252**	0.123
GNS-kit	0.0315	0.1398	0.0427	3.47	0.000	0.000	0.146	0.096

D. Implementation Details

All models are implemented in PyTorch with PyTorch Geometric [32]. Experiments are tracked with MLflow. Training uses the AdamW optimiser [33] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 10^{-4}. Learning rate is reduced on validation loss plateau (factor 0.5, patience 10 epochs). Maximum training is 500 epochs with early stopping (patience 50) for all models. GNS-kit uses graphkit-matched hyperparameters: lr $= 5\times10^{-4}$, decay factor 0.7, patience 5. All models are trained on a single GPU (L40S) with batch size 64 (32 for MxGPS). Seed is fixed for reproducibility. Single-task baselines are fine-tuned with only the supervised task loss: SSE uses dimension-weighted MAE with $w_f = [0.10, 0.10, 0.35, 0.35]$ for $[P_g, Q_g, V_m, V_a]$, and PF uses $0.90 \cdot \text{MSE} + 0.10 \cdot \mathcal{L}_{\text{phys}}$, where the physics term is a per-layer residual exposed only by GNS-kit's architecture and is therefore zero for GCN, GAT, GPS, and GNS; no boundary penalty $\mathcal{L}_{\text{bnd}}$ is applied to any single-task baseline. MxGPS is the only model trained with the full objective of Eq. (12), using per-task weights $\lambda_k = 1.0$ (SSE and PF), $\lambda_{\text{phys}} = 0.01$, and $\lambda_{\text{bnd}} = 0.01$, with the same dimension weights w_f above. Base learning rate is 1×10^{-3} for MxGPS-ind and 5×10^{-4} for MxGPS-cross. MGT pre-training (Phase A) masks each of the first six bus features independently with probability $\rho = 0.5$, matching `gridfm-graphkit`'s random masking scheme for this stage.

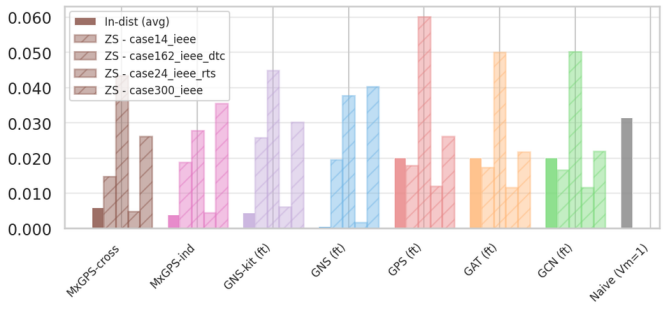

Fig. 2. SSE zero-shot MAE_{V_m} across the four unseen topologies. Among models with meaningful in-distribution SSE accuracy (GNS, GNS-kit, MxGPS), MxGPS variants degrade the least (Table V); GPS, GAT, and GCN show artificially low degradation due to in-distribution magnitude collapse (Table III).

VI. RESULTS AND DISCUSSION

A. Masked Grid Task Pre-training

Table I serves a single purpose: to show that MGT pre-training alone carries essentially no per-scenario voltage magnitude signal, motivating the supervised Phase B fine-tuning. It reports in-distribution MGT accuracy and calibration after Phase A only, averaged over the held-out test splits of each fold's training cases; the Naive (V_m=1) row is a lower-bound reference. GPS, GCN, and GNS-kit edge out the Naive predictor on MAE_{V_m}, though GNS-kit's margin is marginal (0.0315 vs. 0.0333); calibration is universally poor, confirming that the MGT objective does not directly optimise for voltage recovery. Rather than relying solely on this self-supervised signal, MxGPS introduces supervised SSE and PF gradients in Phase B, directly training each branch on the quantities it predicts at inference.

B. Power Flow

Table II reports in-distribution PF accuracy and calibration. All single-task baselines use the two-phase protocol (Phase A MGT pre-training → Phase B task-specific fine-tuning); MxGPS variants are jointly fine-tuned on SSE and PF simultaneously.

GNS and GPS achieve the lowest in-distribution MAE_{V_m} (0.0022 and 0.0023, respectively), but at the cost of substantial boundary violations (BVR 9.4% and 7.3%). MxGPS variants show higher in-distribution error ($\text{MAE}_{V_m} \approx 0.021$) but near-zero BVR ($\leq 0.33\%$), reflecting the regularising effect of joint multi-task training on operational constraint satisfaction. On calibration, single-task baselines achieve near-perfect regression slope (> 0.94), while MxGPS exhibits partial V_m magnitude suppression (slope $\approx$ 0.5–0.6), an area for future improvement. Because GPS shares the same backbone architecture and hyperparameters as each individual MxGPS branch, the GPS vs. MxGPS comparison on this table isolates the effect of multi-task joint training from any architectural difference.

C. Static State Estimation

Table III reports in-distribution SSE accuracy and calibration under the same two-phase protocol. The calibration columns reveal a critical failure mode: GPS, GAT, and GCN show near-zero V_m regression slope (< 0.1) and $R^2 < 0.1$, confirming they collapse to predicting the mean voltage despite $\text{MAE}_{V_m} \approx 0.020$ that may appear competitive. This is a low-MAE, low-information failure mode that bare accuracy statistics do not expose, directly motivating the calibration protocol of Yang et al. [31]: a model predicting the mean achieves acceptable average error while providing zero per-bus voltage information. MxGPS variants do not exhibit this

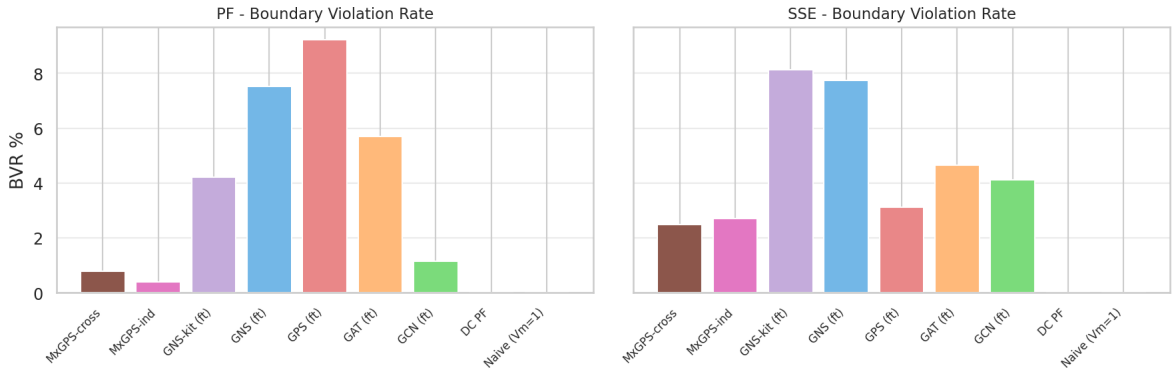

Fig. 3. In-distribution boundary violation rate per model and task. MxGPS variants consistently achieve the lowest violation rates, particularly on PF.

TABLE II

PF IN-DISTRIBUTION ACCURACY AND CALIBRATION (MEAN OVER 3 CROSS-VALIDATION FOLDS). MAE/RMSE IN P.U. EXCEPT MAE_{V_a} (RAD). CALIBRATION: REGRESSION SLOPE AND R^2 (SLOPE= 1, $R^2 = 1$ IDEAL; BEST SLOPE IS CLOSEST TO 1.0). MXGPS VARIANTS ARE JOINTLY FINE-TUNED ON SSE+PF; SINGLE-TASK MODELS INDEPENDENTLY ON PF. BEST PER COLUMN AMONG GNN AND MXGPS MODELS IN **BOLD**. DC PF SEPARATED BY RULE: SETS V_m=1.0 P.U. THROUGHOUT, SO $\text{SLOPE}_{V_m}/R^2_{V_m} = 0$; $\text{SLOPE}_{V_a}/R^2_{V_a}$ NOT COMPUTED.

Model	MAE_{V_m}	MAE_{V_a}	RMSE_{V_m}	BVR (%)	Slope_{V_m}	$R^2_{V_m}$	Slope_{V_a}	$R^2_{V_a}$
DC PF	0.0286	0.2201	0.0322	0.00	0.000	0.000	—	—
GCN	0.0043	0.0175	0.0055	1.22	0.942	0.974	0.982	0.990
GAT	0.0027	0.0161	0.0036	5.48	0.980	0.987	0.990	0.992
GPS	0.0023	**0.0057**	**0.0030**	7.29	1.009	0.989	1.008	**0.999**
GNS	**0.0022**	0.0184	0.0031	9.41	**0.996**	**0.995**	**0.995**	0.995
GNS-kit	0.0064	0.0173	0.0115	4.41	1.026	0.770	0.978	0.992
MxGPS-ind	0.0218	0.0464	0.0243	**0.23**	0.480	0.400	0.797	0.958
MxGPS-cross	0.0207	0.0530	0.0227	0.33	0.578	0.558	0.808	0.967

collapse (slope $\approx$ 0.92, $R^2 \approx$ 0.98), consistent with the joint SSE+PF gradient signal preventing mode collapse to the mean. GNS achieves both the lowest MAE_{V_m} (0.0007) and near-perfect calibration (slope = 0.999, R^2 = 0.999), reflecting its heterogeneous GNN design.

D. Zero-Shot Transfer and Topology Overfitting

Figure 3 shows the in-distribution boundary violation rate across models and tasks; MxGPS variants achieve the lowest BVR on PF ($\leq$ 0.33%). Table IV reports zero-shot results across all four unseen topologies (14-, 24-, 162-, and 300-bus) for both PF and SSE, averaged over the applicable cross-validation folds, and Figure 2 shows zero-shot SSE MAE_{V_m} per model across all four topologies.

The strongest evidence sits on the two larger zero-shot systems, which are the closest analogue of foundation-model deployment on unseen production grids. On the 162-bus case, MxGPS attains both the lowest zero-shot PF MAE_{V_m} overall (0.0230) and 0% BVR; on the 300-bus case, its error (0.0247) is within 0.006 p.u. of the best baseline (GNS-kit, 0.0185), whose predictions however violate voltage limits at 4.7% of buses. On the two small zero-shot systems (14- and 24-bus), several baselines retain lower absolute MAE, so MxGPS's advantage is concentrated where grids are larger and structurally richer. Crucially, MxGPS is the only model family with BVR = 0% on *all four* zero-shot PF topologies, while GPS incurs up to 37% BVR on case300 and GNS up to 28% on case24.

The accuracy inversion summarised in Table V identifies the mechanism: models with the lowest in-distribution PF error degrade the most under topology shift, directly evidencing *topology overfitting*. Averaged across all four zero-shot topologies, MxGPS degrades by only 39% in PF MAE_{V_m} (0.0218 $\rightarrow$ 0.0303); by contrast, GNS-kit degrades by $\approx$190% (0.0064 $\rightarrow$ 0.0187), GNS by $\approx$790%, and GPS by over 14$\times$, despite starting from substantially lower in-distribution error. Because MxGPS's in-distribution error is roughly 10$\times$ higher than the specialists', we report absolute ID and ZS errors alongside Δ% in Table V and anchor the comparison on the 162/300-bus absolutes above, so that the relative framing cannot flatter the multi-task model. The inversion is direct evidence that single-task fine-tuning encodes topology-specific relational structure rather than transferable physics. For SSE, GPS, GAT, and GCN exhibit apparent low degradation ratios (+25–44%), but this is an artifact of in-distribution magnitude collapse (Table III): their in-distribution MAE of $\approx$0.020 already reflects mean-prediction behaviour, leaving no further performance to lose under distribution shift. Among models with meaningful in-distribution SSE accuracy, the MxGPS variants degrade least (+271% to +461%, vs. GNS-kit +489% and GNS +3587%). The multi-task joint training objective, which forces the shared encoder to represent operating constraints relevant to both SSE and PF simultaneously, acts as a strong implicit regulariser against topology overfitting.

A potential objection is that the 0% BVR of MxGPS follows trivially from amplitude shrinkage toward the mean: the in-distribution PF slope is 0.48–0.58, and the Naive V_m=1

TABLE III

SSE IN-DISTRIBUTION ACCURACY AND CALIBRATION (MEAN OVER 3 CROSS-VALIDATION FOLDS). MAE/RMSE IN P.U. EXCEPT MAE_{V_a} (RAD). CALIBRATION: REGRESSION SLOPE AND R^2 (SLOPE= 1, $R^2 = 1$ IDEAL; BEST SLOPE IS CLOSEST TO 1.0). †: GPS, GAT, GCN COLLAPSE ON V_m (SLOPE AND $R^2 < 0.1$), INDICATING MEAN PREDICTION RATHER THAN GENUINE VOLTAGE ESTIMATION. GPS ALSO COLLAPSES ON V_a. BEST PER COLUMN AMONG GNN AND MXGPS MODELS IN **BOLD**. DC PF SEPARATED BY RULE: SETS V_m=1.0 P.U. THROUGHOUT; $\mathrm{SLOPE}_{V_a}/R^2_{V_a}$ NOT COMPUTED.

Model	MAE_{V_m}	MAE_{V_a}	RMSE_{V_m}	BVR (%)	Slope_{V_m}	$R^2_{V_m}$	Slope_{V_a}	$R^2_{V_a}$
DC PF	0.0315	0.8118	0.0354	0.00	0.000	0.000	—	—
GCN†	0.0201	0.0076	0.0240	2.88	0.076	0.083	0.990	0.998
GAT†	0.0202	0.0079	0.0240	2.86	0.079	0.084	0.965	0.998
GPS†	0.0201	0.0918	0.0241	2.90	0.091	0.101	0.177	0.194
GNS	**0.0007**	**0.0030**	**0.0011**	6.33	**0.999**	**0.999**	**1.000**	**1.000**
GNS-kit	0.0045	0.0132	0.0104	7.67	0.953	0.838	1.003	0.968
MxGPS-ind	0.0039	0.0258	0.0058	**1.53**	0.926	0.983	0.900	0.988
MxGPS-cross	0.0060	0.0228	0.0081	1.84	0.920	0.983	0.961	0.991

TABLE IV

ZERO-SHOT TRANSFER TO FOUR UNSEEN TOPOLOGIES FOR PF AND SSE. MAE_{V_m} IN P.U., AVERAGED OVER THE FOLDS IN WHICH EACH TOPOLOGY IS ZERO-SHOT. BEST MAE_{V_m} PER TOPOLOGY AND TASK IN **BOLD**; 0% BVR ENTRIES ALSO **BOLD**.

	14-bus		24-bus		162-bus		300-bus	
Model	MAE_{V_m}	BVR (%)	MAE_{V_m}	BVR (%)	MAE_{V_m}	BVR (%)	MAE_{V_m}	BVR (%)
Power Flow								
GCN	**0.0128**	**0.00**	0.0053	0.03	0.0304	1.13	0.0296	2.26
GAT	0.0164	12.64	0.0031	3.12	0.0358	5.81	0.0295	3.87
GPS	0.0266	**0.00**	0.0037	4.10	0.0646	14.66	0.0411	37.04
GNS	0.0158	10.91	**0.0025**	28.17	0.0312	0.39	0.0305	6.17
GNS-kit	0.0195	0.30	0.0102	8.80	0.0264	3.06	**0.0185**	4.74
MxGPS-ind	0.0383	**0.00**	0.0354	**0.00**	**0.0230**	**0.00**	0.0247	**0.00**
MxGPS-cross	0.0350	**0.00**	0.0340	**0.00**	0.0271	**0.00**	0.0342	**0.00**
State Estimation								
GCN	0.0166	**0.00**	0.0117	**0.00**	0.0502	0.49	0.0220	1.08
GAT	0.0173	**0.00**	0.0116	**0.00**	0.0500	0.55	**0.0218**	0.67
GPS	0.0179	**0.00**	0.0121	3.60	0.0601	3.70	0.0261	4.34
GNS	0.0195	3.13	**0.0017**	14.15	0.0378	1.82	0.0402	6.64
GNS-kit	0.0257	0.01	0.0062	17.02	0.0450	8.85	0.0302	10.01
MxGPS-ind	0.0188	**0.00**	0.0045	2.44	**0.0279**	0.41	0.0355	0.02
MxGPS-cross	**0.0147**	**0.00**	0.0049	2.70	0.0434	**0.00**	0.0263	0.34

predictor also posts 0% BVR. This reading does not hold: post-hoc clipping of any baseline's predictions to $[V_m^{\min}, V_m^{\max}]$ trivially yields 0% BVR but leaves the estimate elsewhere unchanged; clipping cannot produce the combination MxGPS achieves on the large zero-shot systems, where 0% BVR coincides with the lowest (162-bus) or near-lowest (300-bus) zero-shot PF error.

The GPS vs. MxGPS comparison provides a controlled ablation of multi-task training, since GPS shares the same backbone architecture and hyperparameters as each individual MxGPS branch. The outcome is unambiguous: GPS achieves competitive in-distribution PF accuracy ($\mathrm{MAE}_{V_m} = 0.0023$) but collapses on SSE calibration ($\mathrm{slope}_{V_m} = 0.091$, $R^2 = 0.101$) and degrades by over $14\times$ under topology shift. MxGPS retains strong SSE calibration and achieves zero boundary violations on all four zero-shot PF topologies, at the cost of higher in-distribution PF error. The trade-off is a direct consequence of the shared encoder being pulled by two task-specific gradient signals simultaneously: it cannot overfit to PF-specific relational structure, preventing both the magnitude collapse that pure PF fine-tuning induces on SSE and the topology overfitting visible in the degradation table.

Cross-branch attention ablation. Comparing the two variants, MxGPS-ind emerges as the stronger overall configuration: it degrades less on zero-shot PF (+39% vs. +57%), is more accurate on in-distribution SSE (0.0039 vs. 0.0060), and is smaller and faster (1.6 M parameters, 0.13 ms/graph vs. 1.9 M, 0.27 ms). MxGPS-cross offers a modest in-distribution PF advantage (0.0207 vs. 0.0218) and its only consistent zero-shot advantage is a lower SSE degradation ratio (+271% vs. +461%); zero-shot SSE BVR is comparable between the variants (0.76% vs. 0.72%), both well below GNS-kit (8.97%). We attribute the limited benefit of cross-branch attention at $K = 2$ to redundancy: with only two branches, initialised from the same Phase-A checkpoint and already coupled through the shared encoder and the common physics loss, most cross-task information flows through shared parameters, so the explicit per-node exchange adds capacity, and a risk of diluting task specialisation, without opening genuinely new information pathways. We expect the mechanism to become more valuable as the task portfolio grows, where pairwise task relations are richer and the shared encoder alone cannot mediate all of them; analysing the learned cross-branch attention weights also offers a natural explainability probe of which tasks

TABLE V

IN-DISTRIBUTION TO ZERO-SHOT PERFORMANCE DECAY, AVERAGED OVER ALL FOUR ZERO-SHOT TOPOLOGIES. ID AND ZS MAE_{V_m} ARE 3-FOLD CROSS-VALIDATION MEANS (P.U.); Δ (%) IS THE RELATIVE INCREASE; ZS BVR IS THE AVERAGE OVER THE FOUR ZERO-SHOT TOPOLOGIES. BEST Δ PER TASK IN **BOLD**; 0% ZS BVR ENTRIES ALSO **BOLD**. †: GPS, GAT, GCN EXHIBIT IN-DISTRIBUTION MAGNITUDE COLLAPSE ON SSE (SLOPE$_{V_m} < 0.1$, $R^2 < 0.1$); THEIR LOW Δ_{SSE} REFLECTS A DEGRADED IN-DISTRIBUTION BASELINE, NOT GENUINE ZERO-SHOT ROBUSTNESS.

	Power Flow				State Estimation			
Model	ID MAE_{V_m}	ZS MAE_{V_m}	Δ (%)	ZS BVR (%)	ID MAE_{V_m}	ZS MAE_{V_m}	Δ (%)	ZS BVR (%)
GCN	0.0043	0.0195	+359%	0.86	0.0201	0.0251	+25%†	0.39
GAT	0.0027	0.0212	+680%	6.36	0.0202	0.0252	+25%†	0.30
GPS	0.0023	0.0340	+1364%	13.95	0.0201	0.0290	+44%†	2.91
GNS	0.0022	0.0200	+790%	11.41	0.0007	0.0248	+3587%	6.43
GNS-kit	0.0064	0.0187	+190%	4.22	0.0045	0.0268	+489%	8.97
MxGPS-ind	0.0218	0.0303	**+39%**	**0.00**	0.0039	0.0217	+461%	**0.72**
MxGPS-cross	0.0207	0.0326	+57%	**0.00**	0.0060	0.0223	**+271%**	0.76

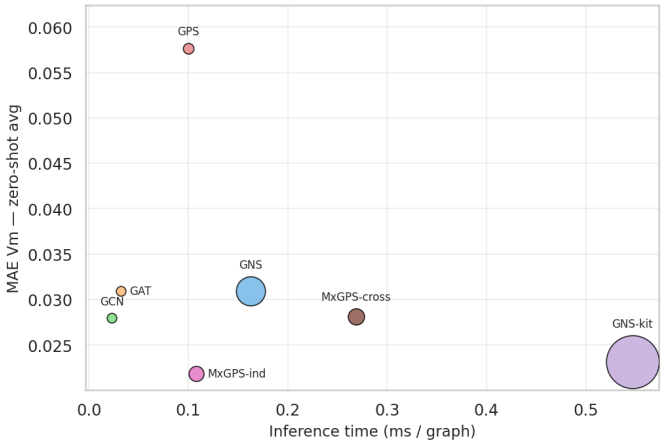

Fig. 4. Zero-shot PF accuracy–efficiency Pareto frontier: average MAE_{V_m} across the four unseen topologies vs. inference time per graph (ms); bubble size proportional to parameter count. MxGPS-ind achieves the best zero-shot PF accuracy among models with 0% BVR on all four topologies, at 12× lower parameter cost than GNS-kit.

exchange information, which we leave to future work. On the strength of this comparison, we recommend MxGPS-ind as the default configuration at $K = 2$.

TABLE VI

MODEL EFFICIENCY: TOTAL PARAMETERS, AVERAGE WALL-CLOCK TIME PER TRAINING EPOCH, AND INFERENCE LATENCY PER GRAPH (GPU, SINGLE BATCH). RANGES COVER PF AND SSE TASKS FOR SINGLE-TASK MODELS; MXGPS VALUES ARE FOR JOINT SSE+PF TRAINING. BEST PER COLUMN IN **BOLD**.

Model	Parameters	Epoch (s)	Inference (ms/graph)
GCN	69 K	18–33	**0.02–0.03**
GAT	219 K	19–40	0.03–0.04
GPS	820 K	37–71	0.09–0.13
GNS	6.1 M	45–104	0.16–0.19
GNS-kit	20.1 M	132–174	0.43–0.68
MxGPS-ind	1.6 M	614	0.13
MxGPS-cross	1.9 M	734	0.27

E. Parameter Efficiency and Computational Cost

Table VI summarises parameter counts, per-epoch training time, and per-graph inference latency. For single-task baselines the ranges cover the PF and SSE fine-tuning phases (which use different batch sizes); for MxGPS the epoch time reflects joint SSE+PF training, with the linear probe stage training only 34 K parameters for MxGPS-ind (300 K for MxGPS-cross) before full unfreezing; single-task baselines train all parameters end-to-end throughout Phase B.

MxGPS-ind (1.6 M parameters, 12.6× smaller than GNS-kit) achieves inference latency comparable to GPS (0.13 ms/graph). Figure 4 shows the zero-shot PF accuracy–efficiency Pareto frontier: MxGPS-ind occupies the accuracy–efficiency sweet spot, matching or exceeding GNS-kit's zero-shot PF accuracy at 12× lower parameter cost and 5× lower inference latency. Critically, the efficiency advantage is not traded against zero-shot performance, a consequence of MxGPS achieving generalisation through the training objective rather than through scale.

F. Limitations

The current experiments cover IEEE systems up to 300 buses; scaling to production grids requires replacing full $O(N^2)$ Transformer attention with a scalable approximation such as Exphormer. While both MxGPS variants achieve 0% BVR on zero-shot PF, SSE zero-shot BVR remains non-zero on some topologies (up to 2.7% for MxGPS-cross on case24). The multi-task portfolio currently covers SSE and PF only; extending to $K = 3$ with N-1 contingency analysis requires only a new masking transform and prediction head, but the benefit of larger portfolios remains to be demonstrated empirically. Finally, the minimum training-portfolio size needed for reliable zero-shot generalisation remains an open question that warrants systematic study.

VII. CONCLUSIONS

We have presented MxGPS, a multiplex graph transformer for multi-task power grid analysis, and identified *topology overfitting*, the systematic failure of single-task GNN fine-tuning to generalise across grid topologies despite strong in-distribution accuracy, as the central failure mode that motivates its design. Models with the lowest in-distribution PF error degrade by 190%–1400% under topology shift, while MxGPS degrades by only 39%, maintains 0% BVR on all four zero-shot PF topologies, and attains the lowest zero-shot PF error on the 162-bus system outright. The multi-task gradient signal also prevents the voltage magnitude collapse observed in single-task SSE fine-tuning, a failure mode that

bare accuracy metrics conceal. Both results share a common mechanism: the shared encoder, forced to simultaneously satisfy complementary gradient signals from SSE and PF, cannot overfit to topology-specific or task-specific relational patterns, at a fraction of the parameter cost of the GridFM reference baseline.

Our ablation indicates that at $K = 2$ the independent-branch configuration is the stronger default, with explicit cross-branch attention expected to become more valuable as the task portfolio grows. Future directions include extending the branch portfolio to optimal power flow and N-1 contingency analysis, where the value of cross-branch attention and its interpretability through learned attention weights can be assessed at larger K; replacing $O(N^2)$ full attention with a scalable approximation for production-scale grids; multi-seed replication of the cross-validation protocol; and per-topology adaptation of the shared encoder to close the remaining gap on in-distribution accuracy.

ACKNOWLEDGMENTS

The work presented is based on research conducted within the framework of the Horizon Europe European Commission project EnerTEF (Grant Agreement No. 101172887). The content of the paper is the sole responsibility of its authors and does not necessarily reflect the views of the European Commission.

REFERENCES

[1] B. Stott, "Review of load-flow calculation methods," *Proceedings of the IEEE*, vol. 62, no. 7, pp. 916–929, 1974.

[2] W. Liao, B. Bak-Jensen, J. R. Pillai, Y. Wang, and Y. Wang, "A review of graph neural networks and their applications in power systems," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 2, pp. 345–360, 2021.

[3] B. Donon, B. Donnot, I. Guyon, and A. Marot, "Graph neural solver for power systems," in *2019 international joint conference on neural networks (ijcnn)*, pp. 1–8, IEEE, 2019.

[4] S. Ghamizi, A. Bojchevski, A. Ma, and J. Cao, "Safepowergraph: Safety-aware evaluation of graph neural networks for transmission power grids," *arXiv preprint arXiv:2407.12421*, 2024.

[5] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI open*, vol. 1, pp. 57–81, 2020.

[6] T. Zhao, X. Pan, M. Chen, A. Venzke, and S. H. Low, "Deepopf+: A deep neural network approach for dc optimal power flow for ensuring feasibility," in *2020 IEEE international conference on communications, control, and computing technologies for smart grids (SmartGridComm)*, pp. 1–6, IEEE, 2020.

[7] H. F. Hamann, B. Gjorgiev, T. Brunschwiler, L. S. Martins, A. Puech, A. Varbella, J. Weiss, J. Bernabe-Moreno, A. B. Massé, S. L. Choi, *et al.*, "Foundation models for the electric power grid," *Joule*, vol. 8, no. 12, pp. 3245–3258, 2024.

[8] L. Rampášek, M. Galkin, V. P. Dwivedi, A. T. Luu, G. Wolf, and D. Beaini, "Recipe for a general, powerful, scalable graph transformer," vol. 35, pp. 14501–14515, 2022.

[9] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

[10] 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 (long and short papers)*, pp. 4171–4186, 2019.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] K. Zhang, G. Yang, F. Shi, S. He, and Y. Zhang, "Moe-graphsage-based integrated evaluation of transient rotor angle and voltage stability in power systems," *arXiv preprint arXiv:2511.08610*, 2025.

[13] L. Müller, M. Galkin, C. Morris, and L. Rampášek, "Attending to graph transformers," *arXiv preprint arXiv:2302.04181*, 2023.

[14] X. Bresson and T. Laurent, "Residual gated graph convnets," *arXiv preprint arXiv:1711.07553*, 2017.

[15] V. P. Dwivedi, C. K. Joshi, A. T. Luu, T. Laurent, Y. Bengio, and X. Bresson, "Benchmarking graph neural networks," *Journal of Machine Learning Research*, vol. 24, no. 43, pp. 1–48, 2023.

[16] V. P. Dwivedi, A. T. Luu, T. Laurent, Y. Bengio, and X. Bresson, "Graph neural networks with learnable structural and positional representations," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022.

[17] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous graph transformer," in *Proceedings of the web conference 2020*, pp. 2704–2710, 2020.

[18] 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*, pp. 594–604, 2022.

[19] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[20] Y. Qin, X. Wang, Z. Zhang, H. Chen, and W. Zhu, "Multi-task graph neural architecture search with task-aware collaboration and curriculum," *Advances in neural information processing systems*, vol. 36, pp. 24879–24891, 2023.

[21] H. Wang, Z. Jiang, Y. You, Y. Han, G. Liu, J. Srinivasa, R. Kompella, Z. Wang, *et al.*, "Graph mixture of experts: Learning on large-scale graphs with explicit diversity modeling," *Advances in neural information processing systems*, vol. 36, pp. 50825–50837, 2023.

[22] D. Sharma, A. Kishore, A. Garg, D. Mazumder, D. Mohapatra, and J. Patro, "Mind the links: Cross-layer attention for link prediction in multiplex networks," in *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, pp. 1243–1247, 2026.

[23] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics*, vol. 3, no. 6, pp. 422–440, 2021.

[24] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[25] A. Puech, M. Mazzonelli, C. Cintas, T. R. Govindasamy, M. Mngomezulu, J. Weiss, M. Baù, A. Varbella, F. Mirallès, K. Kim, *et al.*, "gridfm-datakit-v1: A python library for scalable and realistic power flow and optimal power flow data generation," *arXiv preprint arXiv:2512.14658*, 2025.

[26] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "Matpower: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Transactions on power systems*, vol. 26, no. 1, pp. 12–19, 2010.

[27] S. Babaeinejadsarookolaee, A. Birchfield, R. D. Christie, C. Coffrin, C. DeMarco, R. Diao, M. Ferris, S. Fliscounakis, S. Greene, R. Huang, *et al.*, "The power grid library for benchmarking ac optimal power flow algorithms," *arXiv preprint arXiv:1908.02788*, 2019.

[28] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, OpenReview.net, 2017.

[29] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net, 2018.

[30] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022.

[31] W. Yang, A. Britto Mattos Lima, T. V. Spina, S. Fowers, B. Zhang, and C. White, "Gridsfm: A foundation model for ac optimal power flow." May 2026.

[32] M. Fey and J. E. Lenssen, "Fast graph representation learning with pytorch geometric," *arXiv preprint arXiv:1903.02428*, 2019.

[33] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019.