

## A Technical Appendices and Supplementary Material

### A.1 Close-to-Infeasible Case Generation

Close-to-infeasible cases correspond to operating conditions near the steady-state voltage stability limit. Beyond this limit, no power flow solution exists, and the system is susceptible to voltage collapse, in which the whole system goes down. At this critical point, the power flow Jacobian becomes singular; solvers such as Newton-Raphson are prone to divergence [1, 25].

A technique commonly used to find the steady-state stability limit is *continuation power flow*, which traces a path of solutions to the power flow equations employing a predictor-corrector scheme. Specifically, the basic power flow equations  $g(x)$ , shown in Equation A.1 are reparameterized with a continuation parameter  $\lambda \in \mathbb{R}_{>0}$ , resulting in the system in Equation A.2, where  $x = [\theta^T \ |v|^T]^T$ . The notation used here is consistent with that used in [26], where  $P(x)$  and  $Q(x)$  corresponds to the second term in Equations 1 and 2 respectively, and  $P^{\text{inj}}$  and  $Q^{\text{inj}}$  are the net active and reactive power injections, respectively.

$$g(x) = \begin{bmatrix} P(x) - P^{\text{inj}} \\ Q(x) - Q^{\text{inj}} \end{bmatrix} = 0, \quad (\text{A.1})$$

$$f(x, \lambda) = g(x) - \lambda \begin{bmatrix} P_{\text{target}}^{\text{inj}} - P_{\text{base}}^{\text{inj}} \\ Q_{\text{target}}^{\text{inj}} - Q_{\text{base}}^{\text{inj}} \end{bmatrix} = 0. \quad (\text{A.2})$$

For a current solution  $(x^j, \lambda^j)$ , the predictor estimates the next point  $(\hat{x}^j, \hat{\lambda}^j)$ , typically by taking a step along the tangent direction of the solution trajectory. The corrector then uses this as a warm-start point for Newton’s method in order to find the next solution  $(x^{j+1}, \lambda^{j+1})$ . If the corrector fails, the prediction has surpassed the power flow solvability boundary. The continuation path bends back at this point, forming a characteristic ‘nose’ shape (see Figure A.1).

We use MATPOWER’s [26] continuation power flow implementation to trace the solution path. Specifically, we define  $(P_{\text{base}}^{\text{inj}}, Q_{\text{base}}^{\text{inj}})$  as the setpoints of the load and generator of a randomly selected sample. The target injections are scaled as  $(P_{\text{target}}^{\text{inj}}, Q_{\text{target}}^{\text{inj}}) = (2.5 \times P_{\text{base}}^{\text{inj}}, 2.5 \times Q_{\text{base}}^{\text{inj}})$ . For each base case, a close-to-infeasible case is saved when the continuation method reaches the steady-state stability limit, identified by MATPOWER’s “NOSE” event. To enrich the training set, we also include samples “approaching infeasibility” which correspond to the last four samples before the nose point was triggered. Only samples for which the NOSE event occurred or the continuation power flow converged successfully are retained. Figure A.1 illustrates an example path of power flow solutions traced for the IEEE-118 case, where the base case ( $\lambda = 0$ ) corresponds to a sample drawn from the training set. Figure A.2 shows the condition numbers of the power flow Jacobian for the last 10 points of the curve. The last point corresponds to the steady-stability limit, which shows a much higher condition number, thus exhibiting the singularity of the Jacobian at this operating condition.

### A.2 Dataset Characteristics Comparison

Table A.1 compares **PF $\Delta$**  with existing power flow datasets, including large-scale benchmarks and datasets designed for training specific GNN architectures. The comparison focuses on how these datasets meet key real-world deployability criteria, such as inclusion of load profile distributions, generator setpoint variations, varying grid sizes, N-1 and N-2 topological perturbations, close-to-infeasible cases, and sufficiently complex and realistic network sizes ( $>1000$  buses).

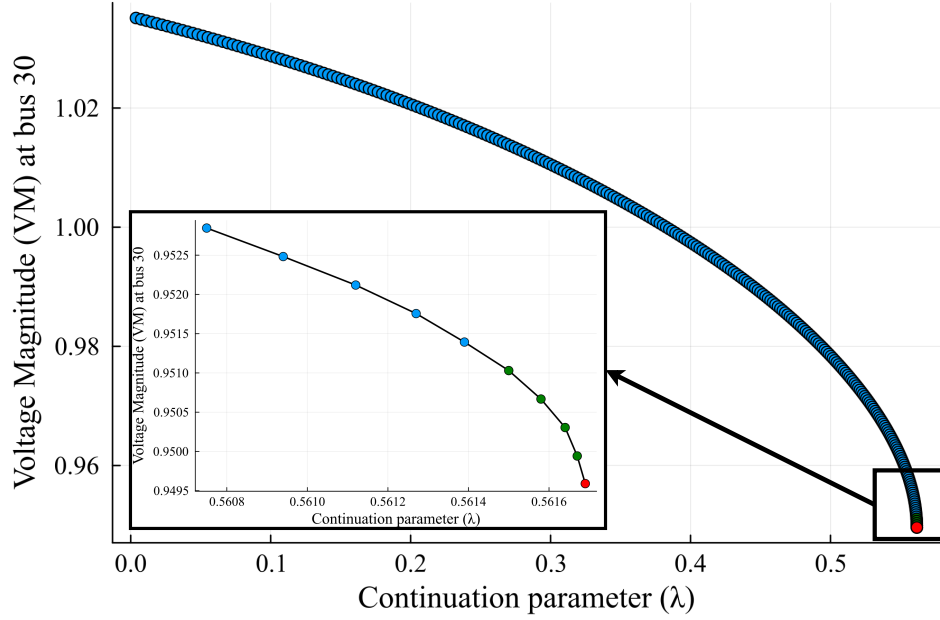


Figure A.1: Voltage magnitude at a load bus as a function of the continuation parameter  $\lambda$ . The point in red corresponds to the sample saved as close-to-infeasible, while the green points are samples labeled as "approaching infeasibility" and used to augment the training data.

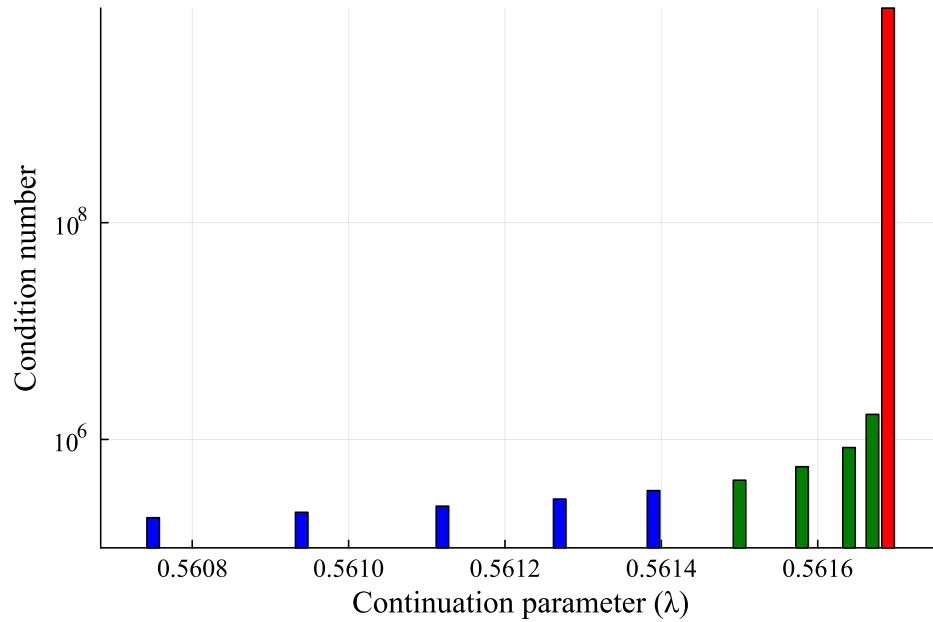


Figure A.2: Condition number of the power flow Jacobian as a function of the continuation parameter  $\lambda$  (y-axis is on a log scale). We see increasing numerical sensitivity as the voltage stability limit is approached, with a sharp rise signaling the onset of ill-conditioning near the solvability boundary.

| Dataset                      | Load Profiles | Generator Profiles | Grid Sizes | N-1 | N-2 | Close to Infeasible | >1000 Buses |
|------------------------------|---------------|--------------------|------------|-----|-----|---------------------|-------------|
| OPFData                      | ✓             | ×                  | ✓          | ✓   | ×   | ×                   | ✓           |
| OPFLearn                     | ✓             | ×                  | ✓          | ×   | ×   | ×                   | ✓           |
| PowerGraph                   | ×             | ×                  | ✓          | ×   | ×   | ×                   | ×           |
| GraphNeuralSolver            | ✓             | ✓                  | ✓          | ×   | ×   | ×                   | ×           |
| PowerFlowNet                 | ✓             | ✓                  | ✓          | ×   | ×   | ×                   | ✓           |
| CANOS                        | ✓             | ×                  | ✓          | ✓   | ×   | ×                   | ✓           |
| <b>PF<math>\Delta</math></b> | ✓             | ✓                  | ✓          | ✓   | ✓   | ✓                   | ✓           |

Table A.1: Comparative assessment of benchmark datasets (OPFData, OPFLearn), custom datasets used to train GNN models, and **PF $\Delta$**  in considering key criteria for real-world deployment.

### A.3 Custom ACOPF Formulation for Data Generation

For our data generation, we consider a custom formulation of ACOPF tailored to the power flow problem setting described in Equation (A.3). We adopt notation consistent with [16] where applicable. Let  $\mathcal{B}$  denote the overall set of all buses,  $\mathcal{D} \subseteq \mathcal{B}$  the set of PQ buses,  $\mathcal{G} \subseteq \mathcal{B}$  the set of PV buses, and  $\mathcal{R} \subseteq \mathcal{B}$  the set of slack (or reference) buses. The set  $\mathcal{L}$  denotes the directed set of branches, where each  $(i, j) \in \mathcal{L}$  indicates a branch with “from” bus  $i$  and “to” bus  $j$ . The reverse orientation of the branches is captured by  $\mathcal{L}^R$ . For a given bus  $i$ , let  $L_i$  denote the subset of edges incident to that bus.

$$\min_{p_g, q_g, |v|, \theta, p_{ij}, q_{ij}} p_g^\top A p_g + b^\top p_g \quad (\text{A.3a})$$

$$\text{s.t. } p_{g_i}^{\min} \leq p_{g_i} \leq p_{g_i}^{\max} \quad \forall i \in \mathcal{G} \setminus \mathcal{R} \quad (\text{A.3b})$$

$$p_{g_i} \geq 0 \quad \forall i \in \mathcal{R} \quad (\text{A.3c})$$

$$|v_i|^{\min} \leq |v_i| \leq |v_i|^{\max} \quad \forall i \in \mathcal{B} \setminus \mathcal{D} \quad (\text{A.3d})$$

$$|v_i| \geq 0 \quad \forall i \in \mathcal{B} \quad (\text{A.3e})$$

$$\theta_i = 0 \quad \forall i \in \mathcal{R} \quad (\text{A.3f})$$

$$\theta_i - \theta_j \in [-\theta_{ij}^{\max}, \theta_{ij}^{\max}] \quad \forall (i, j) \in \mathcal{L} \quad (\text{A.3g})$$

$$p_{g_i} - p_{d_i} - g_s |v_i|^2 = \Re \left\{ \sum_{(i,j) \in \mathcal{L} \cup \mathcal{L}^R} S_{ij} \right\} \quad \forall i \in \mathcal{B} \quad (\text{A.3h})$$

$$q_{g_i} - q_{d_i} + b_s |v_i|^2 = \Im \left\{ \sum_{(i,j) \in \mathcal{L} \cup \mathcal{L}^R} S_{ij} \right\} \quad \forall i \in \mathcal{B} \quad (\text{A.3i})$$

$$S_{ij} = \left( Y_{ij}^* - i \frac{b_{ij}^c}{2} \right) \frac{|V_i|^2}{|T_{ij}|^2} - Y_{ij}^* \frac{V_i V_j^*}{T_{ij}} \quad \forall (i, j) \in \mathcal{L} \quad (\text{A.3j})$$

$$S_{ji} = \left( Y_{ij}^* - i \frac{b_{ij}^c}{2} \right) |V_j|^2 - Y_{ij}^* \frac{V_j V_i^*}{T_{ij}^*} \quad \forall (i, j) \in \mathcal{L} \quad (\text{A.3k})$$

$$\text{where } V_i = |v_i| e^{j\theta_i}, \quad Y_{ij} = \frac{1}{r_{ij} + ix_{ij}}, \quad T_{ij} \text{ is the complex tap ratio}$$

We consider a generation cost minimization objective. Equation (A.3b) enforces active power generation limits at PV buses only (i.e., only for power flow input-related quantities), while active power generation at slack buses and reactive power generation at both PV and slack buses (i.e., power flow output-related quantities) are left unconstrained. Equation (A.3d) bounds the voltage magnitude at PV and slack buses, while for PQ buses we only require them to be nonnegative. The rest of the constraints are as standard for ACOPF. The slack bus angle is fixed to zero in Equation (A.3f). Equation (A.3g) corresponds to voltage angle difference limits, Equations (A.3h) and (A.3i) enforce active and reactive power balance at each bus, and (A.3j) and (A.3k) ensure the branch flows follow Ohm’s Law.

#### A.4 Feature Diversity Comparison Across Benchmark Datasets

Figures A.3 - A.8 present violin plots comparing 10,000 samples from four datasets (PF $\Delta$ , PowerGraph, OPFData, and OPFLearn) for the IEEE 118-bus system. These plots illustrate the distribution of nodal variables across datasets, highlighting the diversity introduced by our data generation process. Specifically, we compare the spread of real power demand ( $p_d$ ), reactive power demand ( $q_d$ ), active power generation ( $p_g$ ), reactive power generation ( $q_g$ ), voltage magnitude ( $|v|$ ), and voltage angle ( $\theta$ ) across seven randomly selected nodes in the system.

Overall, our dataset exhibits comparable or greater variability in these quantities compared to existing large-scale benchmarks. While comparisons with the OPF datasets like OPFData and OPFLearn dataset are not entirely direct given its more constrained ACOPF formulation (which inherently limits variability in power generation values), our dataset maintains a broader distribution in several dimensions. In contrast, when compared to a power flow dataset like PowerGraph, PF $\Delta$  exhibits more variability in all the compared variables. Notably, for active power demand ( $p_d$ ), the distribution in PF $\Delta$  closely matches that of OPFLearn, suggesting similar diversity in load sampling across the two datasets.

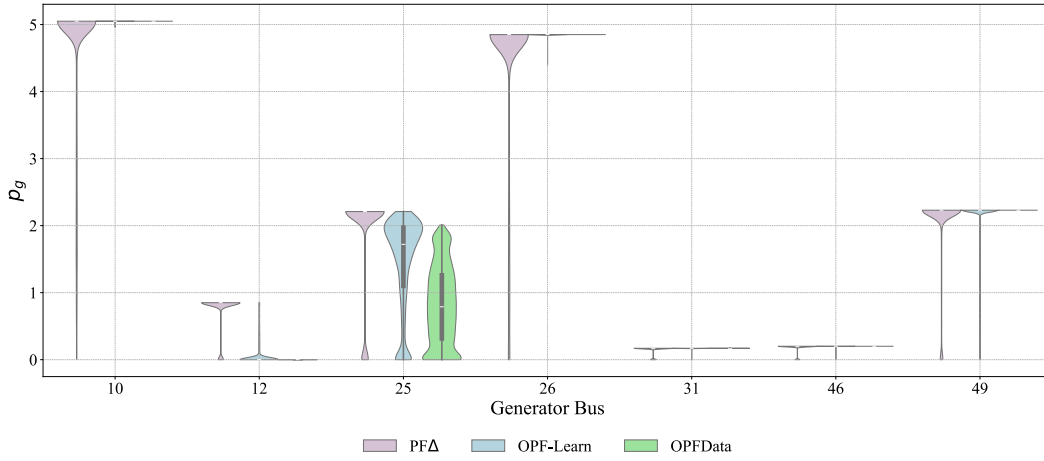


Figure A.3: Violin plots illustrating spread of  $p_g$  values in 7 randomly selected generation buses. This graphic compares feature diversity of  $p_g$  values sampled from PF $\Delta$  to other large-scale benchmark datasets. Note: PowerGraph has not been included in this plot, as this dataset does not report component-level active power generation values.

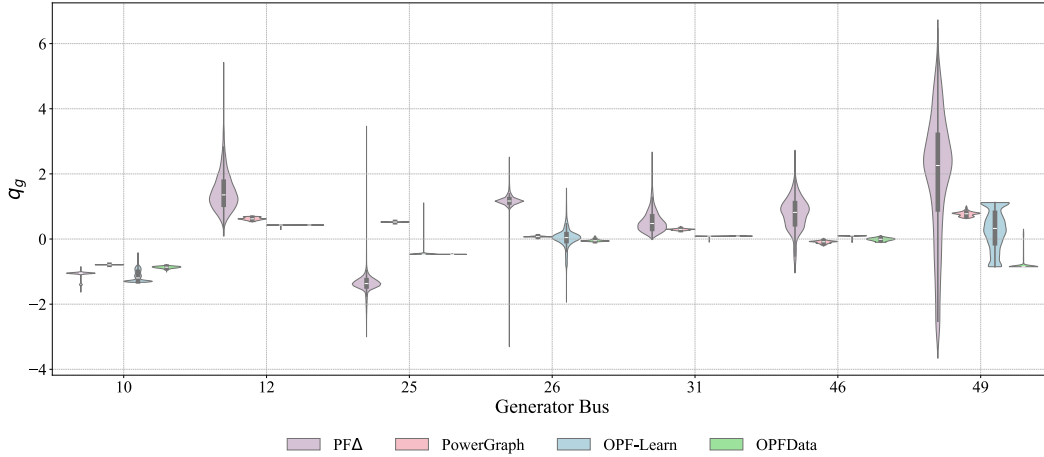


Figure A.4: Violin plots illustrating spread of  $q_g$  values in 7 randomly selected generation buses. This graphic compares feature diversity of  $q_g$  values sampled from PF $\Delta$  to other large-scale benchmark datasets.

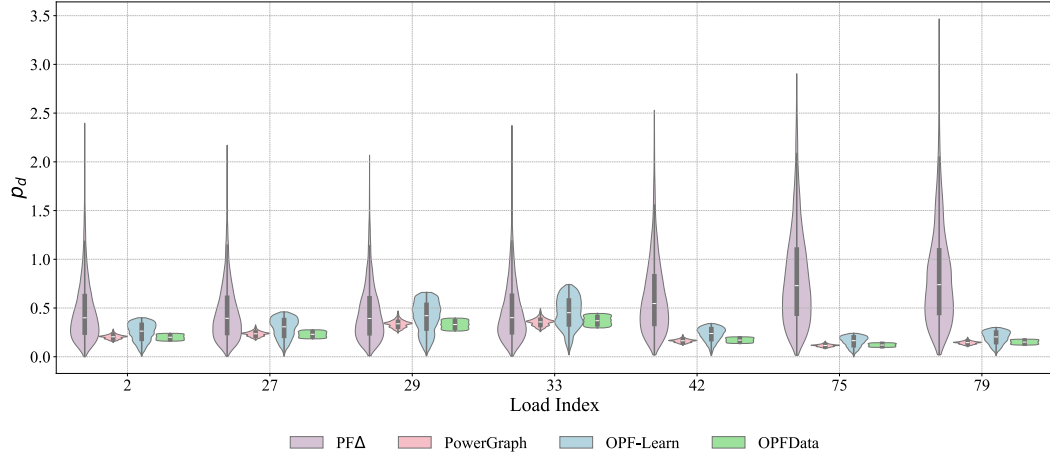


Figure A.5: Violin plots illustrating spread of  $p_d$  values in 7 randomly selected loads at PQ buses. This graphic compares feature diversity of  $p_d$  values sampled from  $PFD$  to other large-scale benchmark datasets.

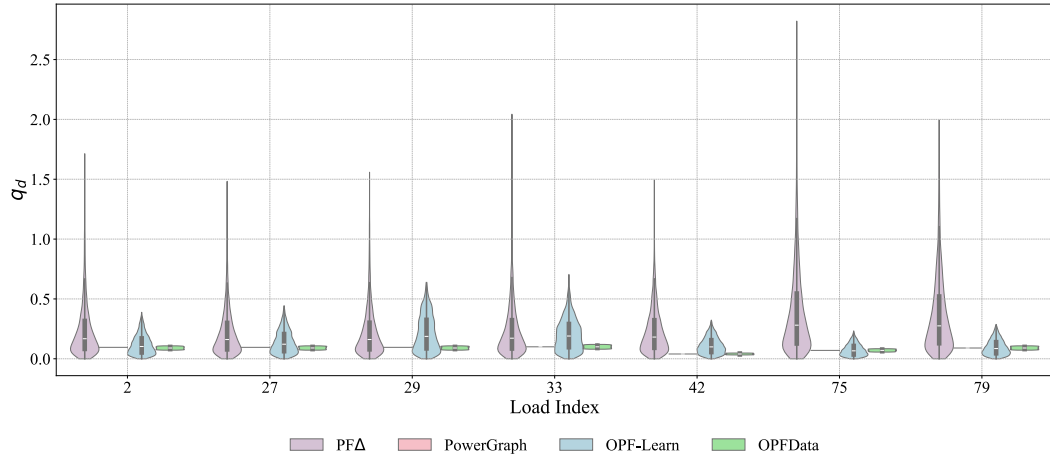


Figure A.6: Violin plots illustrating spread of  $q_d$  values in 7 randomly selected loads at PQ buses. This graphic compares feature diversity of  $q_d$  values sampled from  $PFD$  to other large-scale benchmark datasets.

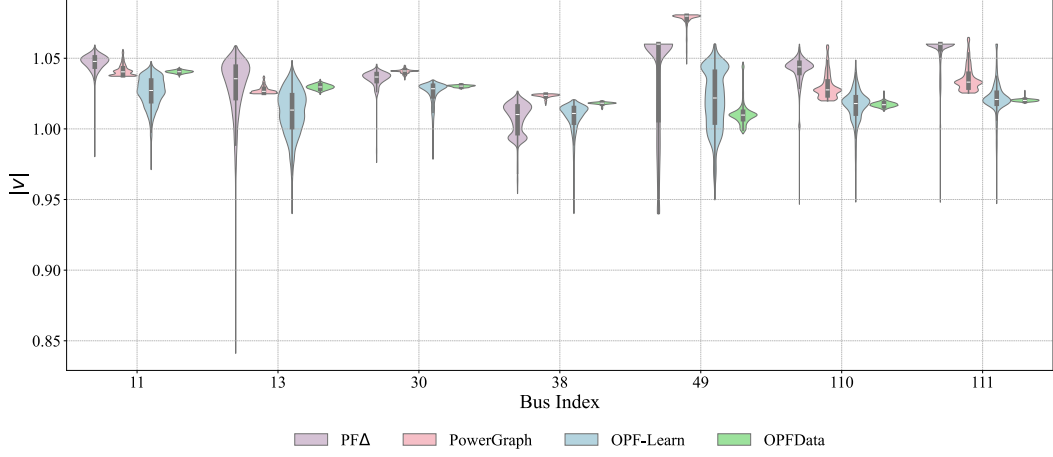


Figure A.7: Violin plots illustrating spread of  $|v|$  values in 7 randomly selected buses. This graphic compares feature diversity of  $|v|$  values sampled from  $PF\Delta$  to other large-scale benchmark datasets.

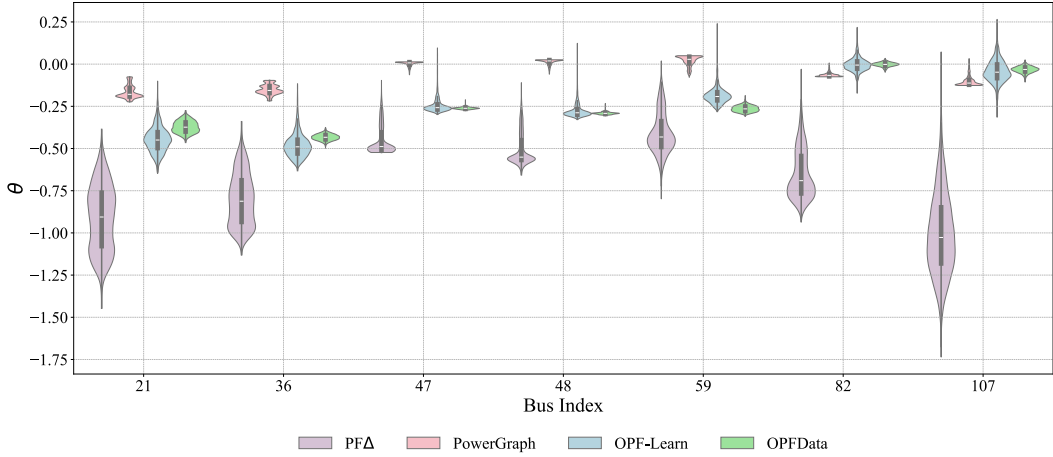


Figure A.8: Violin plots illustrating spread of  $\theta$  values in 7 randomly selected buses. This graphic compares feature diversity of  $\theta$  values sampled from  $PF\Delta$  to other large-scale benchmark datasets.

## A.5 Summary of Dataset Splits

Tables [A.2](#) and [A.3](#) summarize the breakdown of data that  $PF\Delta$  provides for each bus system size. Specifically, the tables illustrate how the data is distributed across each feasibility regime, topological perturbation type, and training and testing splits. Each bus system size contains a total of 132,000 data samples with the exception of the GOC-2000 system, which only contains 30,000 samples. We note that the data for GOC-2000 system can only be used for Task 1.3 (described in Table [2](#)). We also note that we never test on the Approaching Infeasible regime, as this subset of data is only used for data augmentation in Task 4.2 to assess whether including this subset improves testing performance on Close-to-Infeasible samples.

| Bus System Size | Feasible |        |        | Approaching Infeasible |        |        | Close-to-Infeasible |       |       |
|-----------------|----------|--------|--------|------------------------|--------|--------|---------------------|-------|-------|
|                 | N        | N-1    | N-2    | N                      | N-1    | N-2    | N                   | N-1   | N-2   |
| IEEE-14         | 54,000   | 27,000 | 18,000 | 14,400                 | 14,400 | 14,400 | 3,600               | 3,600 | 3,600 |
| IEEE-30         | 54,000   | 27,000 | 18,000 | 14,400                 | 14,400 | 14,400 | 3,600               | 3,600 | 3,600 |
| IEEE-57         | 54,000   | 27,000 | 18,000 | 14,400                 | 14,400 | 14,400 | 3,600               | 3,600 | 3,600 |
| IEEE-118        | 54,000   | 27,000 | 18,000 | 14,400                 | 14,400 | 14,400 | 3,600               | 3,600 | 3,600 |
| GOC-500         | 54,000   | 27,000 | 18,000 | 14,400                 | 14,400 | 14,400 | 3,600               | 3,600 | 3,600 |
| GOC-2000        | 27,000   | 13,500 | 9,000  | 2,400                  | 2,400  | 2,400  | 600                 | 600   | 600   |

Table A.2: Number of available *training* datapoints provided by  $\mathbf{PF}\Delta$  across bus system sizes, feasibility regimes, and grids with varying topological perturbations.

| Bus System Size | Feasible |       |       | Approaching Infeasible |     |     | Close-to-Infeasible |     |     |
|-----------------|----------|-------|-------|------------------------|-----|-----|---------------------|-----|-----|
|                 | N        | N-1   | N-2   | N                      | N-1 | N-2 | N                   | N-1 | N-2 |
| IEEE-14         | 2,000    | 2,000 | 2,000 | -                      | -   | -   | 200                 | 200 | 200 |
| IEEE-30         | 2,000    | 2,000 | 2,000 | -                      | -   | -   | 200                 | 200 | 200 |
| IEEE-57         | 2,000    | 2,000 | 2,000 | -                      | -   | -   | 200                 | 200 | 200 |
| IEEE-118        | 2,000    | 2,000 | 2,000 | -                      | -   | -   | 200                 | 200 | 200 |
| GOC-500         | 2,000    | 2,000 | 2,000 | -                      | -   | -   | 200                 | 200 | 200 |
| GOC-2000        | 1,000    | 1,000 | 1,000 | -                      | -   | -   | 100                 | 100 | 100 |

Table A.3: Number of available *testing* datapoints provided by  $\mathbf{PF}\Delta$  across bus system sizes, feasibility regimes, and grids with varying topological perturbations.

## A.6 PyTorch InMemoryDataset Format of $\mathbf{PF}\Delta$

We provide a PyTorch InMemoryDataset class called `PFDeltaDataset` to load the raw data described in Appendix A.5. This dataset class is designed to support loading data for a specified task for with a given bus system size. Each raw data instance that we load is stored as a JSON file representing the power flow solution for a specific grid configuration.

We offer two different HeteroData formulations of the dataset:

- **Component-Level Formulation.** This representation includes nodes corresponding to physical components such as buses, generators, and loads. It includes both nodal and edge attributes, all expressed in per-unit (p.u.) values. The structure of this formulation follows Figure A.9
- **PV-PQ Formulation.** Optionally-, users can use this formulation, which identifies each bus as a PV, PQ, or slack bus. This formulation is tailored to models that use unique high-dimensional projections for PV and PQ buses. This formulation contains extra edge connections between bus nodes and their specific types (PV, PQ, or slack). The structure of this formulation follows Figure A.10. This formulation can be accessed by setting the `add_bus_type` flag to True in the `PFDeltaDataset` class.

Custom `PFDeltaDataset` classes can be implemented for each model by inheriting the base `PFDeltaDataset` class and overriding the `build_heterodata` method to perform model-specific data pre-processing of the HeteroData objects.

```

HeteroData(
  bus={
    x=[14, 2],
    y=[14, 2],
    bus_gen=[14, 2],
    bus_demand=[14, 2],
    bus_voltages=[14, 2],
    bus_type=[14],
    shunt=[14, 2],
    limits=[14, 2],
  },
  gen={
    limits=[5, 4],
    generation=[5, 2],
    slack_gen=[5],
  },
  load={ demand=[11, 2] },
  (bus, branch, bus)={
    edge_index=[2, 20],
    edge_attr=[20, 8],
    edge_label=[20, 4],
    edge_limits=[20, 1],
  },
  (gen, gen_link, bus)={ edge_index=[2, 5] },
  (bus, gen_link, gen)={ edge_index=[2, 5] },
  (load, load_link, bus)={ edge_index=[2, 11] },
  (bus, load_link, load)={ edge_index=[2, 11] }
)

```

Figure A.9: Component-level formulation of a HeteroData instance of the IEEE-14 bus system available on PFDeltaDataset.

Each HeteroData instance contains nodal and edge attributes that have been preprocessed from the dataset's raw power flow solutions. Edges represent electrical or logical connections between components and buses. Each such edge includes an `edge_index` tensor that defines graph connectivity using the COO (coordinate) format, with two tensors indicating the source and destination nodes of each edge type in the graph.

Nodes contain several types of attributes that can be utilized during the GNN message passing:

- **bus**
  - **x**: Input features for all buses. This is the active and reactive power demand ( $p_d, q_d$ ) for PQ buses, the net active power and voltage magnitude ( $p_{\text{net}}, |v|$ ) for PV buses, and the voltage angle and magnitude ( $\theta, |v|$ ) for the slack bus.
  - **y**: Output targets for all buses. Voltage angles and magnitudes ( $\theta, |v|$ ) for PQ buses, the net reactive power and voltage angles ( $q_{\text{net}}, |v|$ ) for the PV buses, and the net active and reactive powers ( $p_{\text{net}}, q_{\text{net}}$ ) for the slack bus.
  - **bus\_gen**: Active and reactive power generation at each bus ( $p_g, q_g$ ).
  - **bus\_demand**: Active and reactive power demand at each bus ( $p_d, q_d$ ).
  - **bus\_voltages**: Voltage angle and magnitude at each bus ( $\theta, |v|$ ).
  - **bus\_type**: Integer type flag (1 = PQ, 2 = PV, 3 = slack)
  - **bus\_shunt**: Shunt conductance and susceptance (the real and imaginary components of shunt admittances) ( $g_s, b_s$ )
  - **limits**: Voltage limits in the original pglib case. Note that only certain voltage limits are enforced, as specified by [A.3](#).



```

HeteroData(
  bus={
    x=[14, 2],
    y=[14, 2],
    bus_gen=[14, 2],
    bus_demand=[14, 2],
    bus_voltages=[14, 2],
    bus_type=[14],
    shunt=[14, 2],
    limits=[14, 2],
  },
  gen={
    limits=[5, 4],
    generation=[5, 2],
    slack_gen=[5],
  },
  load={ demand=[11, 2] },
  PQ={
    x=[9, 2],
    y=[9, 2],
  },
  PV={
    x=[4, 2],
    y=[4, 2],
    generation=[4, 2],
    demand=[4, 2],
  },
  slack={
    x=[1, 2],
    y=[1, 2],
    generation=[1, 2],
    demand=[1, 2],
  },
  (bus, branch, bus)={
    edge_index=[2, 20],
    edge_attr=[20, 8],
    edge_label=[20, 4],
    edge_limits=[20, 1],
  },
  (gen, gen_link, bus)={ edge_index=[2, 5] },
  (bus, gen_link, gen)={ edge_index=[2, 5] },
  (load, load_link, bus)={ edge_index=[2, 11] },
  (bus, load_link, load)={ edge_index=[2, 11] },
  (PV, PV_link, bus)={ edge_index=[2, 4] },
  (bus, PV_link, PV)={ edge_index=[2, 4] },
  (PQ, PQ_link, bus)={ edge_index=[2, 9] },
  (bus, PQ_link, PQ)={ edge_index=[2, 9] },
  (slack, slack_link, bus)={ edge_index=[2, 1] },
  (bus, slack_link, slack)={ edge_index=[2, 1] }
)

```

Figure A.10: PV-PQ-level formulation of a HeteroData instance of the IEEE-14 bus system available on PFDeltaDataset. This formulation adds extra sub-node connections to the bus nodes to indicate whether they are PV, PQ, or slack nodes.

- **gen**
  - **limits**: Generator operating limits  $(p_{\min}, p_{\max}, q_{\min}, q_{\max})$  as specified in the original pglb case. Note that only certain generation limits are enforced, as specified by [A.3](#)
  - **generation**: Active and reactive power generation at generators  $(p_g, q_g)$
  - **slack\_gen**: Boolean indicator for whether the generator is located at a slack bus.
- **load**
  - **demand**: Active and reactive power demand at each load  $(p_d, q_d)$ .
- **PQ**
  - **x**: Active and reactive power demand at the bus  $(p_d, q_d)$ .
  - **y**: Voltage angles and magnitudes at the bus  $(\theta, |v|)$ .
- **PV**
  - **x**: Net active power and voltage magnitude at the bus  $(p_{\text{net}}, |v|)$ .
  - **y**: Net reactive power and voltage angle at the bus  $(q_{\text{net}}, |v|)$
  - **generation**: Active and reactive power generation at the generators connected to the bus  $(p_g, q_g)$ .
  - **demand**: Active and reactive power demand at the load connected to the bus  $(p_d, q_d)$ .
- **slack**
  - **x**: Voltage angle and magnitude  $(\theta, |v|)$ .
  - **y**: Net active and reactive powers  $(p_{\text{net}}, q_{\text{net}})$ .
  - **generation**: Active and reactive power generation at the generators connected to the bus  $(p_g, q_g)$ .
  - **demand**: Active and reactive power demand at the load connected to the bus  $(p_d, q_d)$ .
- **(bus, branch, bus)**
  - **edge\_attr**: Electrical branch characteristics:  $(r, x, g_{\text{from}}, b_{\text{from}}, g_{\text{to}}, b_{\text{to}}, \tau, \theta_{\text{shift}})$ 
    - \*  $r, x$ : Series resistance and reactance
    - \*  $g_{\text{from}}, b_{\text{from}}$ : Shunt conductance and susceptance at the "from" bus
    - \*  $g_{\text{to}}, b_{\text{to}}$ : Shunt conductance and susceptance at the "to" bus
    - \* **tap** ( $\tau$ ): Tap ratio (defaults to 1.0 if no transformer)
    - \* **shift** ( $\theta_{\text{shift}}$ ): Phase shift angle (degrees)
  - **edge\_label**: Power flow targets for each branch  $(p_{\text{from}}, q_{\text{from}}, p_{\text{to}}, q_{\text{to}})$ 
    - \*  $p_{\text{from}}, q_{\text{from}}$ : active/reactive power at the source of the edge.
    - \*  $p_{\text{to}}, q_{\text{to}}$ : active/reactive power at the destination of the edge.
  - **edge\_limits**: Branch flow limits as specified in the original pglb case. Note that these limits are not enforced in our data generation process, but are included in the dataset for analysis purposes.

## A.7 Interpretability Metrics

In addition to model performance metrics, we also include interpretability metrics to perform qualitative analysis on the solutions predicted by the ML-based power flow solvers. These metrics try to quantify the central tendencies and ranges of the predicted values of the model, such as voltage magnitudes, reactive powers, voltage phase differences, and power at the slack bus. These metrics are as follows:

**Voltage Magnitude at PQ Buses (Mean)**: Calculate the mean voltage magnitude at PQ buses for a given sample.

$$\text{Mean}_{\text{sample}}(|v|) = \frac{1}{N} \sum_{i=1}^N |v|_i \quad (\text{A.4})$$

where  $N$  is the number of PQ buses in the sample, and  $|v|_i$  is the voltage magnitude at the  $i$ -th bus. Aggregate for a dataset by calculating the mean and standard deviation across all samples.

**Voltage Magnitude at PQ Buses (Min):** Calculate the minimum voltage magnitude at PQ buses for a given sample.

$$\text{Min}_{\text{sample}}(|v|) = \min_{i=1}^N |v|_i \quad (\text{A.5})$$

where  $N$  is the number of PQ buses in the sample, and  $|v|_i$  is the voltage magnitude at the  $i$ -th bus. Aggregate for a dataset by calculating the minimum across all samples.

**Voltage Magnitude at PQ Buses (Max):** Calculate the maximum voltage magnitude at PQ buses for a given sample.

$$\text{Max}_{\text{sample}}(|v|) = \max_{i=1}^N |v|_i \quad (\text{A.6})$$

where  $N$  is the number of PQ buses in the sample, and  $|v|_i$  is the voltage magnitude at the  $i$ -th bus. Aggregate for a dataset by calculating the maximum across all samples.

**Reactive Power at PV Buses (Mean):** Calculate the mean reactive power at PV buses for a given sample.

$$\text{Mean}_{\text{sample}}(q_{\text{net}}) = \frac{1}{N} \sum_{i=1}^N q_{\text{net},i} \quad (\text{A.7})$$

where  $N$  is the number of PV buses in the sample, and  $q_{\text{net},i}$  is the reactive power at the  $i$ -th bus. Aggregate for a dataset by calculating the mean and standard deviation across all samples.

**Reactive Power at PV Buses (Min):** Calculate the minimum reactive power at PV buses for a given sample.

$$\text{Min}_{\text{sample}}(q_{\text{net}}) = \min_{i=1}^N q_{\text{net},i} \quad (\text{A.8})$$

where  $N$  is the number of PV buses in the sample, and  $q_{\text{net},i}$  is the reactive power at the  $i$ -th bus. Aggregate for a dataset by calculating the minimum across all samples.

**Reactive Power at PV Buses (Max):** Calculate the maximum reactive power at PV buses for a given sample.

$$\text{Max}_{\text{sample}}(q_{\text{net}}) = \max_{i=1}^N q_{\text{net},i} \quad (\text{A.9})$$

where  $N$  is the number of PV buses in the sample, and  $q_{\text{net},i}$  is the reactive power at the  $i$ -th bus. Aggregate for a dataset by calculating the maximum across all samples.

**Voltage Phase Difference at All Branches (Mean):** Calculate the mean absolute voltage phase difference at all branches for a given sample.

$$\text{Mean}_{\text{sample}}(|\Delta\theta|) = \frac{1}{N} \sum_{i=1}^N |\Delta\theta| \quad (\text{A.10})$$

where  $N$  is the number of branches in the sample, and  $|\Delta\theta|$  is the voltage phase difference at the  $i$ -th branch. Aggregate for a dataset by calculating the mean and standard deviation across all samples.

**Voltage Phase Difference at All Branches (Min):** Calculate the minimum absolute voltage phase difference at all branches for a given sample.

$$\text{Min}_{\text{sample}}(|\Delta\theta|) = \min_{i=1}^N |\Delta\theta| \quad (\text{A.11})$$

where  $N$  is the number of branches in the sample, and  $|\Delta\theta|$  is the voltage phase difference at the  $i$ -th branch. Aggregate for a dataset by calculating the minimum across all samples.

**Voltage Phase Difference at All Branches (Max):** Calculate the maximum absolute voltage phase difference at all branches for a given sample.

$$\text{Max}_{\text{sample}}(|\Delta\theta|) = \max_{i=1}^N |\Delta\theta| \quad (\text{A.12})$$

where  $N$  is the number of branches in the sample, and  $|\Delta\theta|$  is the voltage phase difference at the  $i$ -th branch. Aggregate for a dataset by calculating the maximum across all samples.

**Active Power at the Slack Bus (Mean):** Calculate for a dataset by computing the mean and standard deviation of active powers of slack buses across all samples.

**Active Power at the Slack Bus (Min):** Calculate for a dataset by computing the minimum active power of slack buses across all samples.

**Active Power at the Slack Bus (Max):** Calculate for a dataset by computing the maximum active power of slack buses across all samples.

**Reactive Power at the Slack Bus (Mean):** Calculate for a dataset by computing the mean and standard deviation of reactive powers of slack buses across all samples.

**Reactive Power at the Slack Bus (Min):** Calculate for a dataset by computing the minimum reactive power of slack buses across all samples.

**Reactive Power at the Slack Bus (Max):** Calculate for a dataset by computing the maximum reactive power of slack buses across all samples.

In addition to reporting these metrics for our dataset, we also report them for the predictions of PowerFlowNet, GNS-S, and CANOS-PF when trained on IEEE-118 (Task 1.3) considering topology  $N$  in tables [A.4](#), [A.5](#), and [A.6](#).

| Task 1.3 (topology $N$ ) | PV               |       |      |          |        |        |
|--------------------------|------------------|-------|------|----------|--------|--------|
|                          | $Q_{\text{net}}$ |       |      | $\theta$ |        |        |
|                          | Min              | Mean  | Max  | Min      | Mean   | Max    |
| Custom AC-OPF            | -7.66            | 0.496 | 10.1 | -1.66    | -0.688 | 0.283  |
| CANOS-PF                 | -7.36            | 0.531 | 10.4 | -1.43    | -0.675 | 0.165  |
| GNS-S                    | -5.29            | 0.123 | 10.9 | -1.70    | -1.457 | -1.021 |
| PFNet                    | -7.69            | 0.479 | 10.2 | -1.70    | -0.696 | 0.266  |

Table A.4: Interpretability metrics for PV buses in Task 1.3 considering topology  $N$ .

| Task 1.3 (topology $N$ ) | PQ       |        |        |       |      |      |
|--------------------------|----------|--------|--------|-------|------|------|
|                          | $\theta$ |        |        | $ V $ |      |      |
|                          | Min      | Mean   | Max    | Min   | Mean | Max  |
| Custom AC-OPF            | -1.59    | -0.709 | 0.194  | 0.693 | 1.02 | 1.08 |
| CANOS-PF                 | -1.35    | -0.698 | 0.080  | 0.783 | 1.02 | 1.08 |
| GNS-S                    | -1.68    | -1.487 | -1.187 | 0.804 | 1.03 | 1.11 |
| PFNet                    | -1.63    | -0.716 | 0.201  | 0.740 | 1.02 | 1.08 |

Table A.5: Interpretability metrics for PQ buses in Task 1.3 considering topology  $N$ .

| Task 1.3 (topology $N$ ) | Slack            |      |      |                  |       |      |
|--------------------------|------------------|------|------|------------------|-------|------|
|                          | $P_{\text{net}}$ |      |      | $Q_{\text{net}}$ |       |      |
|                          | Min              | Mean | Max  | Min              | Mean  | Max  |
| Custom AC-OPF            | 11.33            | 22.5 | 29.3 | -4.78            | -1.23 | 1.71 |
| CANOS-PF                 | 11.58            | 22.5 | 29.4 | -4.83            | -1.20 | 1.70 |
| GNS-S                    | -7.54            | 8.0  | 14.4 | -2.42            | -1.37 | 1.07 |
| PFNet                    | 12.63            | 22.7 | 28.7 | -5.71            | -1.21 | 1.60 |

Table A.6: Interpretability metrics for slack buses in Task 1.3 considering topology  $N$ .

## A.8 Replicating CANOS

To correctly compare the performance of CANOS-PF against that of PowerFlowNet and GNS-PF, we first re-implement the original CANOS to verify that our re-implementation works as expected. We assess the fidelity of our re-implementation by comparing its performance on its original dataset to the errors reported in its original paper [12]. Due to limited compute resources, we were unable to train CANOS using the hyperparameters specified in the main body of the original paper. Instead, we train a significantly smaller model for a smaller number of training steps. The results we report in this appendix are on CANOS with 16 message passing steps and hidden size of 256 trained on 200k training steps. In this appendix, we refer to this version of CANOS as Small CANOS.

One of the featured models in [12] is Wide CANOS, with 36 message passing steps and a hidden size 384 trained for 600k training steps. Table A.7 compares the MSE of different variables as well as their sum when trained on `pglib_opf_case500_goc`. As the table reveals, Small CANOS performs comparatively to Wide CANOS. While Small CANOS has a higher Total MSE, we attribute this discrepancy to the fact that Small CANOS is significantly smaller, and that Wide CANOS was trained for a significantly larger number of training steps. We thus conclude that our implementation of CANOS is accurate.

| Variable   | Wide CANOS - Original, TopDrop | Small CANOS - Re-implementation |
|------------|--------------------------------|---------------------------------|
| Total MSE  | 1.63e-02                       | 2.70e-02                        |
| Bus VA     | 1.59e-04                       | 5.15e-04                        |
| Bus VM     | 1.45e-06                       | 2.94e-05                        |
| Gen Pg     | 6.65e-04                       | 2.01e-03                        |
| Gen Qg     | 1.79e-04                       | 1.89e-03                        |
| Line Pf    | 2.02e-03                       | 5.46e-03                        |
| Line Pt    | 2.01e-03                       | 1.56e-03                        |
| Line Qf    | 4.43e-03                       | 5.46e-03                        |
| Line Qt    | 3.99e-03                       | 1.54e-03                        |
| Transf. Pf | 1.22e-03                       | 2.47e-03                        |
| Transf. Pt | 1.22e-03                       | 1.83e-03                        |
| Transf. Qf | 2.21e-04                       | 2.46e-03                        |
| Transf. Qt | 2.23e-04                       | 1.80e-03                        |

Table A.7: MSE comparison of Wide CANOS (36 message passing steps, 384 hidden size, 600k training steps) and Small CANOS (16 message passing steps, 256 hidden size, 200k training steps).

## A.9 Replicating PowerFlowNet Results

To correctly compare the performance of PowerFlowNet against that of CANOS-PF and GNS-PF, we adapted the implementation of PowerFlowNet [7] provided by the authors. To verify that the model works as expected within our code repository, we retrain it on its original dataset using its own pre-processing. Table 2 in [7] reports a Masked L2 loss of 0.018 when trained on `pglib_opf_case118_ieee`. When trained in our code repository, we attained a Masked L2 loss of 0.0155. Thus, we conclude the model was correctly adapted to our code repository.

## A.10 Modified GraphNeuralSolver Formulation

In this section, we present the modified components of the GraphNeuralSolver (GNS) model formulation used in our experiments. Specifically, we highlight the differences from the original architecture proposed in [4], adapting it to accommodate a single slack bus and fixed active power generation at PV buses, as in traditional power flow formulations. For consistency, we retain the same notation as in [4], and only the equations that deviate from the original formulation are shown below.

$$p_{\text{Joule}}^k = \sum_{\substack{i: \text{"from"} \\ j: \text{"to"}}} \left| -v_i^k v_j^k y_{ij} \frac{1}{\tau_{ij}} (\cos(\theta_i - \theta_j - \delta_{ij} - \theta_{\text{shift},ij}) + \cos(\theta_j - \theta_i - \delta_{ij} + \theta_{\text{shift},ij})) \right. \\ \left. + \left( \frac{v_i^k}{\tau_{ij}} \right)^2 y_{ij} \cos(\delta_{ij}) + (v_j^k)^2 y_{ij} \cos(\delta_{ij}) \right| \quad (\text{A.13a})$$

$$p_{\text{global}}^k = \sum_{i=1}^N p_{d,i} + g_{s,i} (v_i^k)^2 + p_{\text{Joule}}^k \quad (\text{A.13b})$$

$$\lambda^k = \begin{cases} \frac{p_{\text{global}}^k - \sum_{i \in \text{non-slack gens}} \bar{p}_{g,i} - \bar{p}_{g,\text{slack}}}{2(\bar{p}_{g,\text{slack}} - \underline{p}_{g,\text{slack}})}, & \text{if } p_{\text{global}}^k < \sum_{i \in \text{all gens}} \bar{p}_{g,i} \\ \frac{p_{\text{global}}^k - \sum_{i \in \text{non-slack gens}} \bar{p}_{g,i} - 2\bar{p}_{g,\text{slack}} - \bar{p}_{g,\text{slack}}}{2(\bar{p}_{g,\text{slack}} - \bar{p}_{g,\text{slack}})} & \text{otherwise} \end{cases} \quad (\text{A.13c})$$

$$p_{g,i}^k = \begin{cases} p_{g,i}^k(\lambda^k), & \text{if } i \text{ is slack} \\ \bar{p}_{g,i}, & \text{otherwise (keep original setpoint)} \end{cases} \quad (\text{A.13d})$$

$$q_{g,i}^k = (q_{d,i} - b_{s,i} (v_i^k)^2) \\ - \sum_{\substack{j \in \mathcal{N}(i) \\ i: \text{"from"}}} \left[ +v_i^k v_j^k y_{ij} \frac{1}{\tau_{ij}} \cos(\theta_i - \theta_j - \delta_{ij} - \theta_{\text{shift},ij}) + \left( \frac{v_i^k}{\tau_{ij}} \right)^2 \left( y_{ij} \sin(\delta_{ij}) + \frac{b_{ij}}{2} \right) \right] \\ - \sum_{\substack{j \in \mathcal{N}(i) \\ i: \text{"to"}}} \left[ +v_i^k v_j^k y_{ij} \frac{1}{\tau_{ij}} \sin(\theta_i - \theta_j - \delta_{ij} + \theta_{\text{shift},ij}) + (v_i^k)^2 \left( y_{ij} \sin(\delta_{ij}) + \frac{b_{ij}}{2} \right) \right] \quad (\text{A.13e})$$

$$\Delta p_i^k = (p_{g,i}^k - p_{d,i} - g_{s,i} (v_i^k)^2) \\ + \sum_{\substack{j \in \mathcal{N}(i) \\ i: \text{"from"}}} \left[ -v_i^k v_j^k y_{ij} \frac{1}{\tau_{ij}} \cos(\theta_i - \theta_j - \delta_{ij} - \theta_{\text{shift},ij}) + \left( \frac{v_i^k}{\tau_{ij}} \right)^2 y_{ij} \cos(\delta_{ij}) \right] \\ + \sum_{\substack{j \in \mathcal{N}(i) \\ i: \text{"to"}}} \left[ -v_i^k v_j^k y_{ij} \frac{1}{\tau_{ij}} \cos(\theta_i - \theta_j - \delta_{ij} + \theta_{\text{shift},ij}) + (v_i^k)^2 y_{ij} \cos(\delta_{ij}) \right] \quad (\text{A.13f})$$

$$\Delta q_i^k = (q_{g,i}^k - q_{d,i} + b_{s,i} (v_i^k)^2) \\ + \sum_{\substack{j \in \mathcal{N}(i) \\ i: \text{"from"}}} \left[ -v_i^k v_j^k y_{ij} \frac{1}{\tau_{ij}} \sin(\theta_i - \theta_j - \delta_{ij} - \theta_{\text{shift},ij}) - \left( \frac{v_i^k}{\tau_{ij}} \right)^2 \left( y_{ij} \sin(\delta_{ij}) + \frac{b_{ij}}{2} \right) \right] \\ + \sum_{\substack{j \in \mathcal{N}(i) \\ i: \text{"to"}}} \left[ -v_i^k v_j^k y_{ij} \frac{1}{\tau_{ij}} \sin(\theta_i - \theta_j - \delta_{ij} + \theta_{\text{shift},ij}) - (v_i^k)^2 \left( y_{ij} \sin(\delta_{ij}) + \frac{b_{ij}}{2} \right) \right] \quad (\text{A.13g})$$

### A.11 Hyperparameter Tuning

We hyperparameter tuned each model on Task 1.3 to maximize generalizability. We employed a grid search strategy, sweeping multiple hyperparameters for each of the three models until convergence was achieved. The best hyperparameters were identified as the ones that performed the best on a validation set (this was designated as 10% of the training dataset) based on each model’s native training loss. Each model was set to have a batch size of 64. Once an initial set of optimal hyperparameters were found, the learning rate of each model was fine-tuned based on the specific tasks we performed evaluation for (1.1, 1.2, 1.3, 2.3, 4.1, 4.2, and 4.3). This involved conducting a small sweep of learning rates on these specific tasks. Model-specific details on the hyperparameters tuned, the number of epochs trained for each model, and the final parameters are provided here:

- **PFNet:** We conducted a grid search over key hyperparameters, including `hidden_dim`, `n_gnn_layers`, `K` (the receptive field size in the TAGConv layers), and the learning rate. Each model was trained for approximately 100 epochs during tuning, consistent with the configuration in the original paper [7]. The final configuration used `hidden_dim` = 256, `n_gnn_layers` = 5, `K` = 4, and a dropout rate of 0.2. The learning rate was 0.0001 for Tasks 1.1, 1.2 and 4.3, 0.00009 for Task 1.3, and 0.0003 for Tasks 2.3, 4.1, and 4.2.
- **GNS-S:** We performed a grid search over the following hyperparameters: `K` (the depth of the network), `hidden_dim`, `gamma` (which weights the contribution of each iterative layer to the loss function), and the learning rate. The finalized values were: `K` = 10, `hidden_dim` = 20, and `gamma` = 0.01. The learning rate was 0.0003 for Tasks 1.1, 4.1, and 4.3, 0.0005 for Tasks 1.3 and 4.2, and 0.0007 for Tasks 1.2 and 2.3. The model was trained for 25 epochs, approximately consistent with the training setup in the original paper [4].
- **CANOS-PF:** We performed a grid search over the following hyperparameters: `hidden_dim`, `k_steps` (the depth of the message-passing network), and the learning rate. The learning rate scheduler was fixed to the same setup as the one defined in the original implementation and the model was trained for 50 epochs [12]. The finalized parameters were: `hidden_dim` = 128 and `k_steps` = 15. The learning rate was 0.0003 for Tasks 1.3, 4.1, and 4.2, 0.0005 for Tasks 1.1, 1.2, 4.3, and 0.0007 for Task 2.3.

### A.12 Extended Experimental Results

Tables [A.8], [A.9], [A.10], [A.11], and [A.12] present extended experimental results on  $\text{PF}\Delta$ . Model performance is evaluated on Tasks 1.1, 1.2, 1.3 (and 2.1), 3.1, 4.1, 4.2, and 4.3, with analysis primarily focused on the IEEE 118-bus system. Power Balance Loss (PBL) mean and maximum values for the IEEE 118-bus system under Tasks 1.1, 1.3, 1.3/3.1, 4.1, 4.2, and 4.3 are reported in Tables [A.8] and [A.9]. Additionally, PBL mean and maximum results across systems of varying size (IEEE-57, IEEE-118, and IEEE-500) for Task 3.1 are shown in Table [A.10].

In addition to the three GNN-based models, we report results from the PowerModels.jl Newton–Raphson (NR) solver as a point of comparison. Runtimes for all four models across the three bus systems are provided in Table [A.12]. NR calculations were performed from a flat start, and results were averaged over the percentage of samples that converged. The convergence rates for the IEEE-57, IEEE-118, and IEEE-500 systems were 95.2%, 65.7%, and 43.4%, respectively. All reported results are obtained by training each model three times per experiment with randomly initialized weights, then computing the mean and standard deviation across these runs. Runtimes for NR were calculated on an Intel Xeon Gold 6140 CPU whereas runtimes for the GNN-based models were calculated on an NVIDIA RTX 8000 GPU.

While training, the performance of the model is calculated periodically in the validation set, which is set to be a fixed random subsample of 10% of the task’s corresponding training set. Another key aspect of our training is the early stopping strategy we apply for each of the models. As described in Appendix [A.11] each model requires a different number of epochs to reach convergence; to ensure that we are not overfitting, we employ an early stopping strategy that halts training if the epoch with the best validation PBL (Mean) happened over 15 epochs ago or if this same error has not changed by more than 1% for 10 epochs consecutively.

| Experiment |          | Power Balance Loss (Mean) |                   |                   |                      |
|------------|----------|---------------------------|-------------------|-------------------|----------------------|
| Task       | Model    | N                         | N-1               | N-2               | Close-to- infeasible |
| 1.1        | PFNet    | $3.2 \pm 0.2e-1$          | $5.2 \pm 0.2e-1$  | $5.3 \pm 0.3e-1$  | $1.4 \pm 0.1e0$      |
|            | CANOS-PF | $1.9 \pm 0.3e-1$          | $7.6 \pm 0.8e-1$  | $6.6 \pm 0.5e-1$  | $1.2 \pm 0.1e0$      |
|            | GNS      | $5.3 \pm 2.7e-1$          | $6.8 \pm 2.6e-1$  | $7.1 \pm 2.5e-1$  | $1.0 \pm 0.2e0$      |
| 1.2        | PFNet    | $3.2 \pm 0.2e-1$          | $4.0 \pm 0.2e-1$  | $4.4 \pm 0.2e-1$  | $1.2 \pm 0.05e0$     |
|            | CANOS-PF | $1.8 \pm 0.06e-1$         | $2.2 \pm 0.08e-1$ | $2.4 \pm 0.09e-1$ | $8.7 \pm 0.7e-1$     |
|            | GNS      | $3.6 \pm 0.3e-1$          | $4.0 \pm 0.3e-1$  | $4.1 \pm 0.3e-1$  | $8.2 \pm 1.1e-1$     |
| 1.3, 2.1   | PFNet    | $3.4 \pm 0.08e-1$         | $4.0 \pm 0.05e-1$ | $4.2 \pm 0.09e-1$ | $1.1 \pm 0.03e0$     |
|            | CANOS-PF | $1.9 \pm 0.2e-1$          | $2.1 \pm 0.1e-1$  | $2.2 \pm 0.1e-1$  | $9.7 \pm 0.7e-1$     |
|            | GNS      | $3.5 \pm 0.4e-1$          | $3.8 \pm 0.4e-1$  | $3.9 \pm 0.4e-1$  | $7.7 \pm 0.5e-1$     |
| 2.3        | PFNet    | $4.1 \pm 0.6e-1$          | $4.9 \pm 0.7e-1$  | $5.1 \pm 0.7e-1$  | $1.1 \pm 0.08e0$     |
|            | CANOS-PF | $4.1 \pm 0.8e-1$          | $4.6 \pm 0.9e-1$  | $4.7 \pm 0.8e-1$  | $1.0 \pm 0.1e0$      |
|            | GNS      | $8.3 \pm 2.6e-1$          | $8.8 \pm 2.7e-1$  | $8.9 \pm 2.7e-1$  | $1.3 \pm 0.3e0$      |
| 4.1        | PFNet    | $4.5 \pm 0.4e-1$          | $5.2 \pm 0.09e-1$ | $5.3 \pm 0.07e-1$ | $1.3 \pm 0.04e0$     |
|            | CANOS-PF | $4.7 \pm 0.1e-1$          | $5.3 \pm 0.05e-1$ | $5.4 \pm 0.06e-1$ | $1.7 \pm 0.01e0$     |
|            | GNS      | $3.5 \pm 0.1e-1$          | $3.7 \pm 0.2e-1$  | $3.8 \pm 0.2e-1$  | $6.6 \pm 0.3e-1$     |
| 4.2        | PFNet    | $5.6 \pm 0.7e-1$          | $6.6 \pm 0.7e-1$  | $6.7 \pm 0.7e-1$  | $1.2 \pm 0.1e0$      |
|            | CANOS-PF | $3.9 \pm 0.8e-1$          | $4.3 \pm 0.8e-1$  | $4.4 \pm 0.8e-1$  | $8.8 \pm 0.9e-1$     |
|            | GNS      | $4.0 \pm 0.9e-1$          | $4.6 \pm 1.0e-1$  | $4.6 \pm 1.0e-1$  | $6.4 \pm 1.4e-1$     |
| 4.3        | PFNet    | $1.2 \pm 0.04e0$          | $1.4 \pm 0.04e0$  | $1.5 \pm 0.02e0$  | $1.1 \pm 0.07e0$     |
|            | CANOS-PF | $1.1 \pm 0.04e0$          | $1.2 \pm 0.03e0$  | $1.2 \pm 0.02e0$  | $0.8 \pm 0.05e0$     |
|            | GNS      | $6.3 \pm 2.2e-1$          | $7.3 \pm 2.4e-1$  | $7.3 \pm 2.4e-1$  | $7.0 \pm 2.8e-1$     |

Table A.8: Power Balance Loss (Mean) across different grid conditions.

| Experiment |          | Power Balance Loss (Max) |                  |                  |                      |
|------------|----------|--------------------------|------------------|------------------|----------------------|
| Task       | Model    | N                        | N-1              | N-2              | Close-to- infeasible |
| 1.1        | PFNet    | $1.5 \pm 0.2e1$          | $5.7 \pm 0.3e1$  | $5.1 \pm 0.8e1$  | $7.0 \pm 2.1e1$      |
|            | CANOS-PF | $9.4 \pm 2.6e0$          | $1.7 \pm 0.2e2$  | $1.6 \pm 0.3e2$  | $1.9 \pm 0.4e2$      |
|            | GNS      | $2.3 \pm 1.0e1$          | $8.9 \pm 4.2e1$  | $1.4 \pm 0.7e2$  | $8.4 \pm 5.6e1$      |
| 1.2        | PFNet    | $1.1 \pm 0.1e1$          | $4.2 \pm 0.7e1$  | $5.0 \pm 1.3e1$  | $3.7 \pm 0.6e1$      |
|            | CANOS-PF | $3.9 \pm 1.1e0$          | $1.2 \pm 0.06e1$ | $1.5 \pm 0.1e1$  | $2.7 \pm 0.4e1$      |
|            | GNS      | $1.6 \pm 0.2e1$          | $2.1 \pm 0.08e1$ | $1.7 \pm 0.2e1$  | $2.4 \pm 0.8e1$      |
| 1.3, 2.1   | PFNet    | $1.0 \pm 0.2e1$          | $3.5 \pm 0.3e1$  | $3.5 \pm 0.4e1$  | $3.5 \pm 0.6e1$      |
|            | CANOS-PF | $3.7 \pm 0.1e0$          | $1.1 \pm 0.1e1$  | $8.9 \pm 2.1e0$  | $3.4 \pm 0.4e1$      |
|            | GNS      | $1.1 \pm 0.3e1$          | $2.1 \pm 0.3e1$  | $1.8 \pm 0.2e1$  | $1.9 \pm 0.5e1$      |
| 2.3        | PFNet    | $1.1 \pm 0.2e1$          | $4.8 \pm 1.8e1$  | $3.9 \pm 0.2e1$  | $4.3 \pm 0.7e1$      |
|            | CANOS-PF | $0.8 \pm 0.4e1$          | $3.4 \pm 0.8e1$  | $2.9 \pm 0.2e1$  | $4.3 \pm 0.8e1$      |
|            | GNS      | $3.7 \pm 3.5e1$          | $5.4 \pm 5.1e1$  | $5.3 \pm 3.8e1$  | $4.0 \pm 2.3e1$      |
| 4.1        | PFNet    | $1.5 \pm 0.2e1$          | $4.5 \pm 0.2e1$  | $4.3 \pm 0.3e1$  | $8.5 \pm 0.8e1$      |
|            | CANOS-PF | $1.1 \pm 0.08e1$         | $3.2 \pm 0.2e1$  | $3.1 \pm 0.2e1$  | $9.1 \pm 0.8e1$      |
|            | GNS      | $1.2 \pm 0.2e1$          | $2.1 \pm 0.3e1$  | $1.5 \pm 0.07e1$ | $2.0 \pm 0.3e1$      |
| 4.2        | PFNet    | $1.6 \pm 0.3e1$          | $4.9 \pm 0.8e1$  | $4.1 \pm 1.2e1$  | $8.5 \pm 0.4e1$      |
|            | CANOS-PF | $8.5 \pm 1.5e0$          | $2.8 \pm 0.06e1$ | $2.7 \pm 0.08e1$ | $4.6 \pm 0.6e1$      |
|            | GNS      | $1.8 \pm 0.6e1$          | $2.4 \pm 0.3e1$  | $2.5 \pm 1.4e1$  | $2.2 \pm 0.2e1$      |
| 4.3        | PFNet    | $3.3 \pm 0.8e1$          | $6.0 \pm 2.2e1$  | $5.0 \pm 0.9e1$  | $8.1 \pm 0.3e1$      |
|            | CANOS-PF | $2.0 \pm 0.5e1$          | $4.0 \pm 0.3e1$  | $5.2 \pm 0.3e1$  | $2.4 \pm 0.3e1$      |
|            | GNS      | $2.4 \pm 1.9e1$          | $4.5 \pm 3.6e1$  | $4.1 \pm 2.3e1$  | $2.5 \pm 0.5e1$      |

Table A.9: Power Balance Loss (Max) across different grid conditions.



| Experiment |          | Power Balance Loss (Mean) |             |             |                     |
|------------|----------|---------------------------|-------------|-------------|---------------------|
| Case       | Model    | N                         | N-1         | N-2         | Close-to-infeasible |
| 57         | PFNet    | 2.3±0.4e0                 | 2.3±0.4e0   | 2.3±0.4e0   | 2.4±0.4e0           |
|            | CANOS-PF | 1.6±0.1e0                 | 1.6±0.1e0   | 1.6±0.1e0   | 1.7±0.1e0           |
|            | GNS      | 3.3±1.4e1                 | 8.8±3.9e1   | 1.5±0.7e2   | 0.8±0.2e0           |
|            | NR       | 1.1±0.0e-6                | 1.2±0.0e-6  | 1.1±0.0e-6  | 1.3±0.0e-6          |
| 118        | PFNet    | 3.4±0.08e-1               | 4.0±0.05e-1 | 4.2±0.09e-1 | 1.1±0.03e0          |
|            | CANOS-PF | 1.9±0.2e-1                | 2.1±0.1e-1  | 2.2±0.1e-1  | 9.7±0.7e-1          |
|            | GNS      | 3.5±0.4e-1                | 3.8±0.4e-1  | 3.9±0.4e-1  | 7.7±0.5e-1          |
|            | NR       | 3.7±0.0e-6                | 3.2±0.0e-6  | 3.3±0.0e-6  | 4.6±0.0e-6          |
| 500        | PFNet    | 8.3±3.9e1                 | 8.1±3.8e1   | 8.4±4.0e1   | 8.9±3.8e1           |
|            | CANOS-PF | 1.8±0.1e1                 | 1.8±1.00e1  | 1.8±0.1e1   | 1.9±0.1e1           |
|            | GNS      | 2.4±0.7e1                 | 2.4±0.7e1   | 2.4±0.7e1   | 2.4±0.7e1           |
|            | NR       | 1.4±0.0e-5                | 1.4±0.0e-5  | 1.3±0.0e-5  | 1.6±0.0e-5          |

Table A.10: Power Balance Loss (Mean) across different bus sizes and grid conditions.

| Experiment |          | Power Balance Loss (Max) |           |           |                     |
|------------|----------|--------------------------|-----------|-----------|---------------------|
| Case       | Model    | N                        | N-1       | N-2       | Close-to-infeasible |
| 57         | PFNet    | 1.2±0.2e1                | 1.6±0.2e1 | 1.5±0.2e1 | 2.1±0.2e1           |
|            | CANOS-PF | 3.7±0.4e1                | 3.9±0.2e1 | 3.8±0.3e1 | 3.9±0.2e1           |
|            | GNS      | 3.0±1.3e5                | 4.4±2.2e5 | 5.2±2.4e5 | 4.3±1.8e1           |
| 118        | PFNet    | 1.0±0.2e1                | 3.5±0.3e1 | 3.5±0.4e1 | 3.5±0.6e1           |
|            | CANOS-PF | 3.7±0.1e0                | 1.1±0.1e1 | 8.9±2.1e0 | 3.4±0.4e1           |
|            | GNS      | 1.1±0.3e1                | 2.1±0.3e1 | 1.8±0.2e1 | 1.9±0.5e1           |
| 500        | PFNet    | 1.9±0.6e3                | 1.9±0.6e3 | 1.9±0.6e3 | 1.9±0.6e3           |
|            | CANOS-PF | 4.7±0.5e2                | 5.1±0.4e2 | 6.1±0.2e2 | 4.9±0.7e2           |
|            | GNS      | 5.9±2.3e2                | 5.9±2.3e2 | 6.1±2.2e2 | 5.8±2.1e2           |

Table A.11: Power Balance Loss (Max) across different bus sizes and grid conditions.

| Experiment |          | Runtime (seconds) |               |               |                     |
|------------|----------|-------------------|---------------|---------------|---------------------|
| Case       | Model    | N                 | N-1           | N-2           | Close-to-infeasible |
| 57         | PFNet    | 6.6e-3±2.8e-5     | 6.6e-3±1.6e-5 | 6.6e-3±2.8e-5 | 6.7e-3±8.5e-5       |
|            | CANOS-PF | 3.1e-2±8.1e-5     | 3.1e-2±2.1e-5 | 3.1e-2±1.1e-4 | 3.2e-2±9.8e-5       |
|            | GNS      | 5.1e-2±3.1e-4     | 5.1e-2±1.5e-4 | 5.1e-2±9.6e-5 | 5.3e-2±1.0e-4       |
|            | NR       | 9.5e-3±2.0e-4     | 3.9e-3±8.4e-5 | 3.7e-3±4.1e-5 | 3.9e-3±1.4e-4       |
| 118        | PFNet    | 6.4e-3±7.7e-5     | 6.4e-3±6.3e-5 | 6.4e-3±2.8e-5 | 6.4e-3±1.3e-5       |
|            | CANOS-PF | 2.7e-2±2.1e-4     | 2.7e-2±2.1e-5 | 2.7e-2±8.8e-5 | 2.8e-2±1.8e-4       |
|            | GNS      | 4.8e-2±1.9e-4     | 4.8e-2±2.1e-4 | 4.8e-2±1.8e-4 | 4.9e-2±6.8e-5       |
|            | NR       | 4.2e-2±5.5e-4     | 1.6e-2±8.5e-4 | 1.2e-2±1.8e-4 | 1.2e-2±3.1e-4       |
| 500        | PFNet    | 7.4e-3±1.5e-4     | 7.3e-3±4.0e-6 | 7.3e-3±2.3e-5 | 7.5e-3±5.0e-5       |
|            | CANOS-PF | 3.1e-2±1.2e-4     | 3.1e-2±1.4e-4 | 3.1e-2±5.6e-4 | 3.2e-2±5.3e-4       |
|            | GNS      | 6.5e-2±1.6e-4     | 6.5e-2±2.3e-4 | 6.5e-2±3.0e-4 | 6.7e-2±5.0e-4       |
|            | NR       | 1.2e-1±3.1e-3     | 5.2e-2±2.4e-3 | 2.1e-2±1.1e-3 | 2.1e-2±1.0e-4       |

Table A.12: Runtimes across different bus sizes and grid conditions.

### A.13 Failure Modes of Models

Figure A.11 shows histograms of the maximum power balance loss (PBL Max) across all test samples, separated by model (*CANOS*, *PFNet*, *GNS-S*). Each model exhibits a left-skewed distribution with a long tail, indicating that while most samples have relatively low PBL Max values, a small subset experiences substantially larger violations. The spread and shape of these distributions differ across models: *CANOS* displays the tightest distribution with the lowest mean and standard deviation,

suggesting greater consistency and overall accuracy. *PFNet* exhibits a broader spread and a slightly higher mean, whereas *GNS-S* shows the widest spread.

Figures A.12, A.14 show histograms of power balance loss (PBL) per node, categorized by bus type (PQ, PV, Slack) for each model. Across all models, PQ buses consistently have lower PBL than PV buses. Some possible interpretations of this behavior include: (1) reactive power (predicted at PV buses) is harder to predict correctly than voltage magnitude (predicted at PQ buses), or (2) errors in reactive power contribute more heavily to overall PBL than errors in voltage magnitude. Performance on the slack bus varies distinctly across models. *CANOS-PF* achieves exactly zero PBL on the slack bus by design, due to its analytical treatment of that node. It also has the lowest PV and PQ PBL values in comparison to the other models. In contrast, *PFNet* and *GNS-S* display broad distributions of slack bus PBL, indicating greater variability and potential model instability at the slack node.

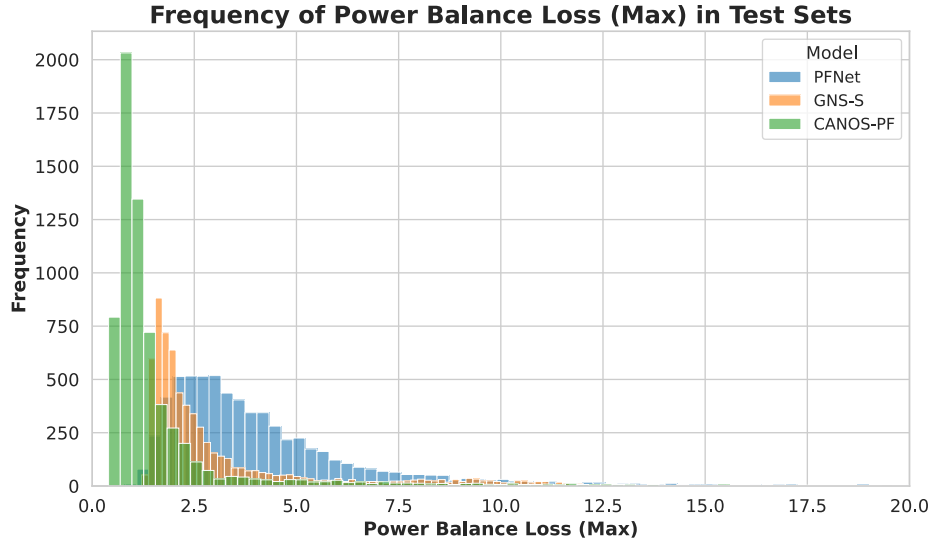


Figure A.11: Distribution of maximum power balance loss (PBL Max) across test samples for CANOS, PFNet, and GNS-S.

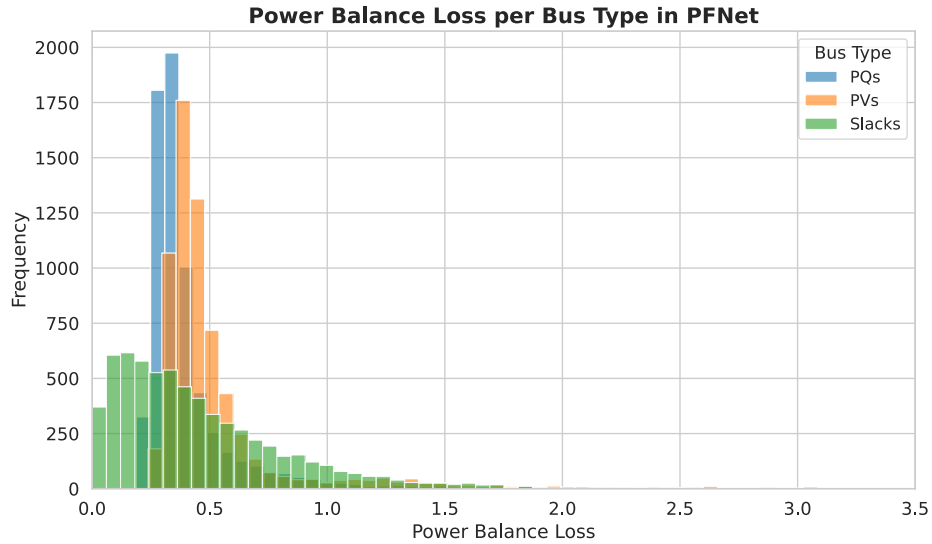


Figure A.12: Distribution of power balance loss (PBL Mean) across bus types within the test samples for PFNet.

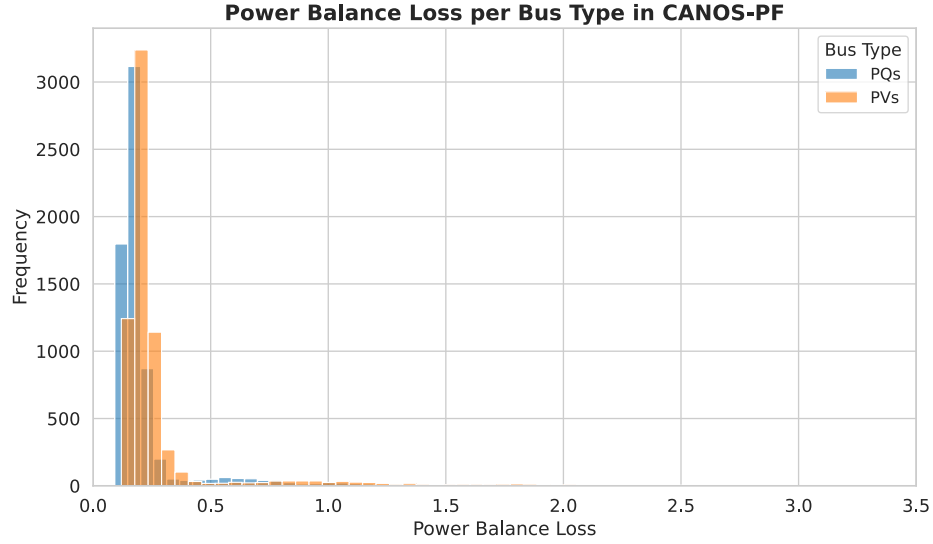


Figure A.13: Distribution of power balance loss (PBL Mean) across bus types within the test samples for CANOS-PF. All slack buses achieve a PBL of zero due to analytical enforcement of this property within the model. As a result, we do not present these results in this plot.

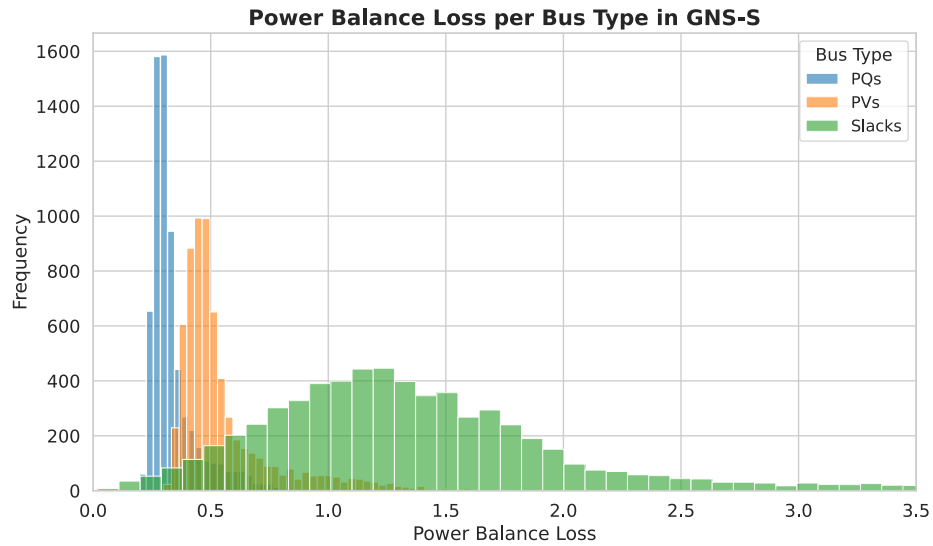


Figure A.14: Distribution of power balance loss (PBL Mean) across bus types within the test samples for GNS-S.