A UNIFIED MULTI-TASK BENCHMARK FOR DATA-DRIVEN MODELS OF POWER SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a learned representation of a structured physical system transfer across tasks, topologies, scales, and interventions? Despite growing interest in foundation models for scientific and engineering systems, existing power-grid datasets largely isolate individual tasks and operating regimes, making such questions difficult to study. We introduce a unified multi-task benchmark for power systems whose tasks are genuinely different learning problems: constrained regression (AC and distribution optimal power flow), mixed-integer structured prediction (security-constrained unit commitment), rare-event screening ($N-k$ contingency analysis), inverse problems under partial observability (transmission and distribution state estimation), and Monte-Carlo risk estimation (reliability assessment), spanning transmission and distribution networks ranging from 3 to 13659 buses. Alongside the data, we specify benchmark results on cross-task and topology transfer. Surprisingly, stock GNNs, cross-task pretraining, and publicly released foundation models frequently fail to outperform simple classical or nearest-neighbor baselines, revealing substantial headroom for models that truly generalize across structured physical systems. We release the dataset, generalization splits, and evaluation harness as an open benchmark for graph learning, scientific foundation models, and structured prediction.

1 INTRODUCTION

The electric power grid is undergoing a transformation driven by demands for energy security and affordability, large-scale integration of diverse generation sources, and widespread electrification. These developments increase the variability of system conditions and push infrastructure closer to operational limits while reliability requirements remain stringent. Operators must screen large contingency sets, enforce security constraints, and react in near real-time, yet core analyses such as AC optimal power flow (ACOPF), security-constrained unit commitment (SCUC), and reliability assessment (RA), which rely on non-convex and/or mixed-integer programming, remain computationally expensive, often requiring seconds to minutes per scenario, creating a critical computational bottleneck for real-time grid operations.

Data-driven surrogates promise large speedups on these tools, and a recurring hope is that a model trained on diverse grid data and topologies could learn representations that transfer across tasks (Hamann et al., 2024; Tu et al., 2024), as transfer of learned representations across tasks and datasets is well established in other domains (Bommasani, 2021). However, these are not variants of a single problem: as learning problems they span constrained regression (ACOPF, DOPF), mixed-integer structured prediction (SCUC), rare-event screening ($N-k$), inverse problems under partial observability (SE, DSSE), and Monte-Carlo risk estimation (RA), which differ in output space, constraint structure and supervision. Whether one representation transfers across them is an unverified empirical question.

Training high-capacity models needs millions, if not billions, of examples across operating conditions, topologies, and disturbances (Bommasani, 2021), but operational records are fragmented across utilities, commercially sensitive and regulated (Asghar et al., 2017), and cover a narrow region of the operational space with few near-limit or rare events (Hamann et al., 2024; Deng et al., 2024). Synthetic corpora have consequently become central for optimization and estimation (Zamzam & Baker, 2020; Fioretto et al., 2020): OPFData standardized ACOPF with topological pertur-

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

bations (Lovett et al., 2024), PowerGraph targeted GNNs on power-flow tasks (Varbella et al., 2024), PFΔ produced the first dataset of near-infeasible cases (Rivera et al., 2025), Gillioz et al. built validated synthetic European transmission data (Gillioz et al., 2025), and for distribution Habib et al. and the SoCal digital twin provide DSSE and phasor data (Habib et al., 2023; Xie et al., 2025). A parallel literature synthesizes the inputs themselves (load and PMU time series, distribution networks, reliability surrogates) (Li et al., 2020; Pinceti et al., 2021; Zheng et al., 2019; Dande et al., 2024; Dong et al., 2022; Li et al., 2025).

Table 1 places these resources on the axes a unified power-system benchmark needs. Existing synthetic datasets are comprehensive along their chosen axes of variation, for instance OPF-Data on topological perturbation scale (Lovett et al., 2024), PGLearn on optimization-proxy tooling with primal and dual solutions (Klamkin et al., 2025), PowerGraph on graph-level power-flow and cascading-failure supervision (Varbella et al., 2024), PFΔ on near-infeasible operating points (Rivera et al., 2025), and SimBench and Gillioz et al. on validated grid models and snapshots (Meinecke et al., 2020; Gillioz et al., 2025). However, none of these datasets couples multiple learning tasks for the same networks using a shared representation, and none provides a joint benchmark for transmission and distribution problems. The generation of such a corpus is a gap that has been identified as a necessary precursor to the development of models that exhibit cross-task transfer at scale (Hamann et al., 2024).

This work directly addresses these challenges by introducing a unified multi-task benchmark and evaluation for power grid representation learning. Specifically, it provides:

- a unified schema that standardizes input-output representations across every task, letting one model and reader serve all seven;
- a corpus on shared networks and scenarios spanning all seven tasks (Section 2.2);
- a benchmark harness separated into case-size tiers, with generalization splits, and per-task metrics against classical baselines, ranging from a tier accessible to student researchers to one requiring leadership scale computing resources.

2 Multi-task Framework and Representation

Each task operates at a distinct time horizon, uses a different physics model, and serves a different operational objective (Table 2). The corpus is organized around a *core set*: 30 transmission networks carrying data across ACOPF, SCUC, N-k data. On those core networks the same MATPOWER case file, the same generator and branch parameters, and common loading scenarios underlie both a snapshot optimization task and a day-ahead scheduling task. Additional networks holding only one task are released as single-task extensions, allowing for the direct measurement of task generalization outside of the core topology set.

Table 1: Breadth of existing reference power-systems datasets commonly used for surrogate model training. Columns denote categorical features that support multi-task surrogate training, labeled ✓ present; – absent; ○ partial or limited. "Multi-task / shared net": several learning tasks defined on one shared network set. "Typed hetero graph": a heterogeneous graph with distinct `bus`/`generator`/`load`/`shunt` node types rather than a homogeneous bus-branch graph or tabular instances. "Correlated demand": a scenario sampler that moves system-total load. "Duals": dual variables / locational prices retained. "AC round-trip": exact AC-feasibility labels or a power-flow re-solve stored per instance. Competitor facts are from the cited papers.

Resource	Tasks	Shared-topology multi-task	Typed hetero	Corr. demand	Duals	AC r.-trip	Ref. basel.
OPFData (Lovett et al., 2024)	ACOPF	–	✓	–	–	–	✓
PGLearn (Klamkin et al., 2025)	ACOPF, DC-OPF	–	○	○	✓	–	✓
PowerGraph (Varbella et al., 2024)	PF, cascade	○	–	–	–	–	✓
PFΔ (Rivera et al., 2025)	PF (near-infeas.)	–	–	–	–	○	✓
SimBench (Meinecke et al., 2020)	grid models	–	–	○	–	–	–
Gillioz et al. (Gillioz et al., 2025)	PF/OPF snapshots	–	–	○	–	–	–
(ours)	ACOPF, SCUC, N-1/N-k, DSSE, SE, DOPF, RA	✓	✓	✓	✓	✓	✓

Table 2: Benchmark tasks and corpus scope. Each task's time horizon, physics model, and learning goal, alongside its released network and sample counts. The core set is the 30 transmission networks carrying ACOPF and SCUC (25 also carry N-k); extensions add single-task networks. DSSE is one 40-feeder corpus reported under two sampling regimes (IID/QSTS; the two figures are alternative samplings of the same feeders and are not additive); DOPF ships 35 of 40 generated feeders (5 reserved for evaluation). Sampling strategies are described in Appendix D.

	Task	Time Horizon	Physics Model	Learning Problem	Networks	Samples
Trans. Core	ACOPF	Snapshot (5 min–1 hr)	Single-phase AC	Constrained regression	30	169,000
	SCUC	Day-ahead (36 hr)	Single-phase DC	Mixed-integer structured prediction	30	44,627
	SE	Real-time (15 min)	Single-phase AC	Inverse problem (partial obs.)	30	639,737
	N-k Cont.	Snapshot, post-outage	Single-phase AC	Rare-event screening	25	15,793,200
Trans. Ext.	ACOPF-Ext	Snapshot (5 min–1 hr)	Single-phase AC	Constrained regression	20	174,000
	SCUC-Ext	Day-ahead (36 hr)	Single-phase DC	Mixed-integer structured prediction	10	3,427,715
Distrib. Feeders	DSSE	Real-time (15 min)	Three-phase AC	Inverse problem (partial obs.)	40	825,600 (IID) 1,401,600 (QSTS)
	DOPF	Snapshot	Three-phase AC	Constrained regression	35	17,500
System	RA	Planning horizon	Network flow	Monte-Carlo risk estimation	1	1,418,520

Table 2 summarizes each task's horizon, physics model and learning problem; the full formulations are given in Appendix A.

2.1 Unified Data Structure

A persistent barrier in graph benchmarks for power systems is that several representational choices in the graph structure are design choices. Different researchers use alternative representations, some represent lines as graph edges (Lovett et al., 2024), some represent lines as nodes (Ringsquandl et al., 2021), generators could be bus attributes (Varbella et al., 2024) or their own nodes (Lovett et al., 2024), giving transmission a single phase with a zero reference angle is a modeling choice (Molzahn & Hiskens, 2019), and so on. We answer it by separating a *canonical model*, based on the IEEE Common Information Model standard (Uslar et al., 2012), from its *projection*. The canonical model describes what a network, a scenario, and an experiment *mean* in engineering terms; it is medium- and architecture-independent. The graph a GNN consumes is a deterministic projection

$$\Pi\big(N,\, S,\, F_\tau,\, O_\tau;\, v\big)$$

of that canonical model under a named view v, mapping a network N, a scenario S, a task formulation F_τ and an observation set O_τ to a typed graph. Each modeling choice is then a property of a particular view, instead of being baked irretrievably into the released data in such a way that a correction would require regeneration of the dataset. This is a design choice to admit a native notion of observation models in the dataset, allowing a common basis for tasks where observability is typically limited. The definitions below describe the projection included in this dataset release.

Heterogeneous Relational Graph Representation. Power systems comprise diverse entity types with distinct physical characteristics and feature dimensionalities, motivating a *heterogeneous* graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \tau, \phi)$ where $\tau : \mathcal{V} \to \mathcal{T}$ types nodes and $\phi : \mathcal{E} \to \mathcal{R}$ types edges, each relation a triplet (t_s, r, t_d) mapping the typed node t_s to node t_d with relation r. Node types then carry different feature dimensions. For instance, generators carry cost curves, buses carry voltage limits.

Typed Representation of Missing Data. The remaining profile details, including the typing of absent fields, the treatment of dual variables in optimization tasks, and the unit and normalization choices, are given in Appendix C. One detail of note is that due to state estimation tasks requiring partial observations, a field's absence in the data is typed ("not represented" $\neq$ "measured zero") in order to differentiate missing physical measurements from zero-valued measurements and from modeling omissions. This is a necessary component in defining state estimation tasks across either transmission or distribution systems, as each task is either over- or under- determined depending on the observation model according to which state is sampled, but is something typically omitted from ACOPF datasets where total observability is assumed.

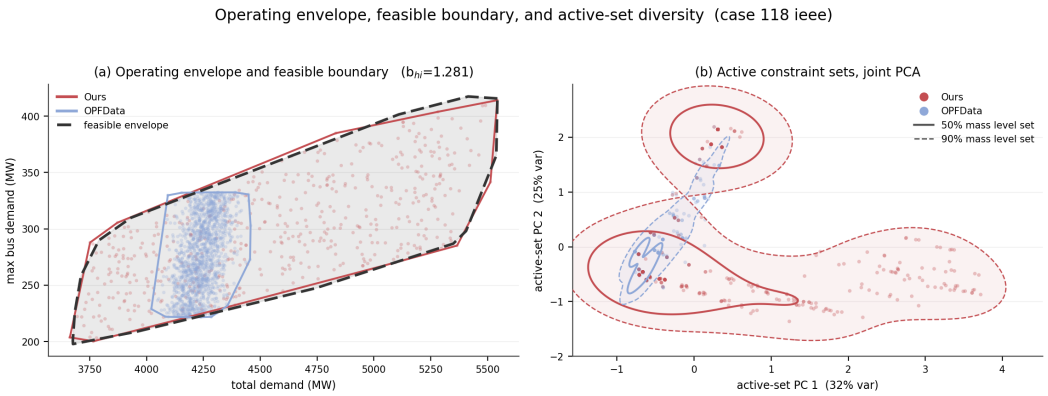


Figure 1: *Left:* Operating loads overlaid on the feasible operating window for all scenarios in case 118, as compared to those available in OPFData. Our dataset spans the range of feasible operating conditions rather than providing a dense sample of a narrow range. *Right:* PCA of the active constraint set for all scenarios in case 118, across our dataset and OPFData. OPFData constraints are limited to a narrow cluster, while ours provides greater diversity.

2.2 Corpus Construction

The corpus is generated task by task on the shared networks, with per-task construction details, the task inventories, and the operating-envelope and active-set analyses given in Appendix D. Details of the storage format are given in Appendix C.

In order to be useful during grid operations, power grid models must not only generalize to unseen network topologies, but also to unseen operating conditions. Two datasets can share the same reference networks yet differ by orders of magnitude in how far their operating points travel from the nominal solution and in how far their load distribution ranges over the grid's feasible operating envelope. In the construction of this dataset, we attempt to ensure both a diversity of tasks and of operating conditions by construction.

ACOPF is solved over PGLib (PowerModels/Ipopt) from a two-level sampler using a global loading factor $b \sim \mathcal{U}(b_{\text{lo}}, b_{\text{hi}})$ times an independent per-load $\varepsilon \sim \mathcal{U}(0.8, 1.2)$, with $[b_{\text{lo}}, b_{\text{hi}}]$ calibrated per case by bisection toward infeasibility. Under independent perturbation alone (as in OPFData), total load is near-constant and a per-unit constant predictor scores $R^2 = 0.9997$ on active power, whereas the correlated-b sampler ensures a far broader sampling of the operating envelope (view in Figure 1, App. D). Every solve stores the primal solution and the nodal-balance (KCL) duals and branch-limit multipliers ($\text{LMP} = -\lambda_{\text{kcl}}/\text{baseMVA}$). This also allows for the measurement of binding constraint diversity: by a primal-at-bound criterion, our corpus spans $13/15/2{,}080$ unique convergent active sets on case14/case30/case118 against OPFData's $5/2/1{,}231$ (active set diversity is illustrated in Figure 1, additional sampling details are provided in App. D).

Constant-predictor residual: Let y_{ic} be a solved quantity for operating point $i \in \{1, \ldots, N\}$ at component $c \in \{1, \ldots, C\}$ (component = generator for dispatch P_g, or bus for voltage magnitude $|V|$). Let $\bar{y}_c$ be the per-component mean and $\bar{y}$ the global mean. Define the *constant-predictor residual*

$$\rho(y) \;=\; \frac{\sum_{i,c}(y_{ic} - \bar{y}_c)^2}{\sum_{i,c}(y_{ic} - \bar{y})^2} \;=\; 1 - R^2_{\text{const}}, \tag{1}$$

the fraction of pooled label variance left unexplained by the predictor that ignores the input and always returns the per-component training mean. $\rho \in [0, 1]$: $\rho \to 0$ means every operating point coincides with the component mean, trivially predictable; larger ρ means genuine operating-point-to-operating-point variation.

We read ρ as a measure of component variation, a surrogate that structurally cannot beat the constant predictor is of limited use, and a dataset with ρ near zero cannot easily be used to discriminate model

Table 3: Operating-point diversity across PGLib-OPF reference cases, reported as means of 1500 samples/case. ρ = constant-predictor residual (Eq. 1); CV = total-load coefficient of variation; stressed = fraction of operating points with $\min |V| < 0.95$ p.u.

case	dataset	$\rho(P_g) \uparrow$	$\rho(\lvert V\rvert) \uparrow$	load CV $\uparrow$	stressed $\uparrow$
case14	Ours	0.0210	0.2809	0.1298	0.000
	PGLearn	0.0173	0.2632	0.1179	0.000
	OPFData	0.0036	0.0092	0.0500	0.000
case30	Ours	0.0249	0.2364	0.1513	0.000
	PGLearn	0.0240	0.2288	0.1498	0.000
	OPFData	0.0024	0.0141	0.0362	0.000
case57	Ours	0.0287	0.0993	0.1531	0.038
	PGLearn	0.0283	0.0970	0.1510	0.026
	OPFData	0.0043	0.0072	0.0416	0.487
case118	Ours	0.0390	0.2756	0.1089	0.190
	PGLearn	0.0480	0.3027	0.1178	0.035
	OPFData	0.0028	0.0499	0.0159	0.000
case6470	Ours	0.0700	0.1272	0.1510	1.000
	PGLearn	0.0656	0.1675	0.1406	1.000
	OPFData	0.0071	0.0416	0.0066	1.000

quality, regardless of how many samples it contains. We report ρ on dispatch P_g and on voltage magnitude $|V|$.

Coefficient of Variation Whereas the constant-predictor residual captures a notion of output diversity, we measure a notion of input variety with the coefficient of variation of the total system load. This is defined as $\mathrm{CV} = \frac{\sigma}{\mu}$, for mean μ and standard deviation σ of the total system load across operating points. A dataset with a low coefficient of variation represents a limited snapshot of the operating regime, and may limit a model's ability to generalize to unseen input conditions

Stress percentage Beyond measures of input space and output space coverage, there is a question of the prevalence of adverse operating conditions in the data. We define this as the fraction of solved operating points for which there exists a bus with voltage V s.t. $|V| < 0.95$ per unit (i.e., an abnormally undervoltage state as defined by ANSI C84.1), capturing a notion of how often the dataset reaches a regime where voltage and thermal limits are active. A dataset that samples only operating points where no constraints are active offers little information to the model about how the solutions behave when constraints bind.

Values are reported in Table 3. On every case, our dataset provides broader diversity of both input and outputs than OPFData, and maintains parity or compares favorably with PGLearn while encompassing a much broader range of tasks.

3 Benchmark and Evaluation Protocol

We also distribute a benchmark harness that allows for reproducible evaluation on the distributed dataset. The evaluation protocol defines the per-task input/output formats, standard splits across tasks and size tiers, and common per-task metrics, which we use to evaluate classical baseline methods. For each task, the harness reports the score of a candidate model on the task metric, the score of a designed classical baseline for that task, the score of a task-independent kNN lookup, and a skill score measuring the model's improvement over the baseline. For metrics where lower is better (e.g., error), the skill score is $s = \mathrm{clip}(1 - M_{\mathrm{model}}/M_{\mathrm{base}})$; for metrics where higher is better (e.g., accuracy), it is $s = \mathrm{clip}((A_{\mathrm{model}} - A_{\mathrm{base}})/(1 - A_{\mathrm{base}}))$. s=1 is a perfect model, s=0 ties the baseline, s=-1 is the floor. Unless noted, every reported metric and skill is a mean over the 3 seeds$\times$cases in a (task, tier) cell and $\pm$ is one standard deviation over that population. The per-task input/output contract is provided in the attached code supplement and specifies input fields, target fields, dims,

Table 4: Per-task primary metric and the classical baseline it is scored against. Skill over a classical baseline is included as a secondary summary. The ACOPF/DOPF optimality gap is computed on cost. RA reports EUE primary, LOLE, and a risk index.

Task	Raw primary metric	Classical baseline
ACOPF	cost-gap	DC-OPF
DOPF	per-unit state RMSE	mean predictor
SCUC	commitment Hamming distance	merit order economic commitment
Contingency	recall @ 5% FPR	LODF / DC screen
DSSE / SE	per-unit state RMSE	mean predictor
RA	EUE MAE	seasonal mean

masks, units, angle reference, and fields prohibited at inference. Details are given in Table 14 and App. B. Each task carries a standard classical baseline, given in Table 4.

Tiers, splits, tracks. Four case-size tiers scale compute at a fixed contract (NANO/SMALL/MEDIUM/FULL). Every submission reports one in-distribution and four generalization splits from published index files, targeting topology transfer, scale transfer (train $\leq$300 buses, test above 2,000), regime shift, and N-k perturbation.The reference results in Section 3.1 are in-distribution and cross-topology and cross-task generalization results are reported in Appendix E.1. Full evaluation protocol details are in Appendix E.

3.1 Reference Results

We run a bank of stock learned baselines (GCN (Kipf, 2016), GAT (Veličković et al., 2017), GIN (Xu et al., 2018) and GraphSAGE (Hamilton et al., 2017)) through the protocol against the classical baselines of Table 4. Table 5 reports the raw model and baseline metrics per task (mean over 3 seeds$\times$cases). Results here are provided as a reference for reproduction, not for the purposes of evaluating the quality of any specific model architecture or to propose a novel model architecture.

Table 5: Dense reference results: the best model backbone per (task, tier), raw metric, the classical baseline quality on the same metric, and a comparison to kNN. Lower is better for cost-gap/Hamming/RMSE.

Task	Tier	Model	Raw	Base	k-NN	Skill
ACOPF	NANO	GIN	0.9091	0.2309	0.02098	-1.000 ± 0.000
	SMALL	GCN	38.72	0.2125	3.678	-0.814 ± 0.511
	MEDIUM	GAT	0.6303	0.1471	0.01453	-0.592 ± 0.673
	FULL	GCN	123.1	1.191	0.07029	-0.927 ± 0.310
DOPF	NANO	GAT	0.007417	0.008128	0.0004142	$+0.090 \pm 0.030$
	SMALL	GAT	0.007411	0.008128	0.0004142	$+0.090 \pm 0.028$
	MEDIUM	GAT	0.007409	0.008128	0.0004142	$+0.091 \pm 0.023$
	FULL	GAT	0.007437	0.008128	0.0004142	$+0.088 \pm 0.028$
DSSE	NANO	SAGE	0.05291	0.03903	0.04091	-0.363 ± 0.097
	SMALL	SAGE	0.05305	0.03903	0.04091	-0.367 ± 0.090
	MEDIUM	SAGE	0.05307	0.03903	0.04091	-0.367 ± 0.089
	FULL	SAGE	0.05303	0.03903	0.04091	-0.366 ± 0.090
SCUC	SMALL	GCN	0.1261	0.2261	0.06213	$+0.100 \pm 0.817$
	MEDIUM	GCN	0.1787	0.06433	0.02301	-0.854 ± 0.232
SE	NANO	GAT	0.09581	0.07388	0.003844	-0.297 ± 0.033
	SMALL	SAGE	0.09625	0.07388	0.003844	-0.303 ± 0.031
	MEDIUM	SAGE	0.09627	0.07388	0.003844	-0.303 ± 0.031

The stock learned reference models rarely beat the task-specific baselines once those baselines are non-trivial. Across ACOPF tiers the best bank backbone trails the DC-OPF floor (skill -0.59 to -0.93), and on SCUC a hetero-GNN trails the merit-order commitment baseline on SCUC at the medium tier (-0.85; only marginal, $+0.10$, at the small tier). These baselines are trained to a val-

idation plateau with early stopping at patience 5 checks of a 1% relative-improvement threshold. Checkpoints are released with the supplement), so the weak skill is more likely due to a capacity floor rather than to undertraining. We are careful about scope: these are stock architectures without per-task tuning, and the task-independent k-NN lookup beats the classical baseline on several task/tiers (Table 5). The evidence is thus that these particular reference models underperform simple alternatives, not that data-driven methods cannot succeed here. That gap is the headroom that makes the corpus useful as a benchmark, and it motivates both stronger architectures and the non-cost rationales for learned surrogates (solve speed, modularity, uncertainty quantification).

There are many candidate machine learning models and many candidate classical methods that could be used as a baseline for each target task. In order to illustrate the generality of the finding, we evaluate a task-independent baseline across every regression task and size tier. Because each task has a common representation in our schema, a lookup can be defined within a fixed topology, and can be extended across topologies with the key alignment described below. The task-independent baseline thus works as follows: Let the training set be inputs $X_{\text{tr}} \in \mathbb{R}^{n \times d}$ with targets Y_{tr}, and let x_q be a query input. Distances are computed in the z-standardized input space, using the training mean μ and standard deviation σ (per feature),

$$\tilde{x} = (x - \mu)/\sigma, \qquad \mathcal{N}_k(x_q) = \underset{S \subseteq \{1,\dots,n\}, |S|=k}{\arg\min} \sum_{i \in S} \|\tilde{x}_q - \tilde{x}_i\|_2^2, \tag{2}$$

and the prediction is the unweighted mean of the neighbors' output fields,

$$\hat{y}(x_q) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x_q)} y_i. \tag{3}$$

We use $k = 5$ throughout.

This defines a lookup procedure for an entire network, in which x is the full load vector and y is the full solution vector. For cross-topology lookup we use a per-node table keyed on each bus's topology-independent attributes, predicting a normalized dispatch or state.

We evaluate kNN within and across topologies and tasks under the definitions above: the in-distribution reference is Table 5, and the cross-topology and cross-task results are in Appendix **??**. Within a fixed topology the learned models underperform the naive lookup, so a trained model earns its value by generalizing to unseen topologies rather than by in-distribution accuracy.

To illustrate the flexibility of the common data representation, we transfer an encoder trained on each source task (rows of Table 6) to ACOPF and SCUC, holding target labels and compute fixed across the full evaluation benchmark. On ACOPF, no pretraining source approaches from-scratch training (0.84–1.0 vs 0.59). On SCUC, every source modestly improves over the weak from-scratch encoder (0.19–0.29 vs 0.31), with the largest relative gains at the smallest (1%) label budget, but every transferred model still trails the classical merit-order baseline (0.12). We therefore make no strong positive-transfer claim, and only use this as an illustration that this corpus makes such studies controlled and cheap. The development of a generalized cross-topology and cross-task model is an open question in the field, we only demonstrate that this dataset is able to natively support cross-task and cross-topology training and evaluation.

To separate dataset difficulty from undertraining of our stock models, we evaluate two publicly released, fully-trained ACOPF foundation models under the same protocol and input contract. Both foundation models are evaluated frozen with only an input schema adapter, each using its own native preprocessing and normalization: LUMINA-2M v0.1.5a (LUMINA) and GridSFM-Open (rev. 1b41299, MIT (Microsoft Research (2026))). Both are scored on the same cases, split, and ACOPF metric as the rest of the benchmark. In a majority of cases both underperform the k-NN lookup on ACOPF (Figure 2), suggesting that the low skill shown in Table 5 is not only due to model quality and that the wider operating regime contributes to the gap.

Cross-topology transfer (Table 7). Here the encoder is pretrained on the same task's other networks and adapted to a held-out network (the topology-transfer split). Per network, transfer tracks from-scratch training closely and shows no systematic gain, and both remain well above the classical baseline and the kNN floor; on the full-tier held-out network (case8387) both diverge, marking scale generalization as the open difficulty.

Table 6: Source→target transfer, median raw metric on the full evaluation labels. Rows are the source task the encoder was pretrained on, columns are the target task it is adapted to. ACOPF is cost-gap, SCUC commitment Hamming; lower is better.

Source →	**ACOPF** (cost-gap ↓)	**SCUC** (Hamming ↓)
Random init	1.00	0.508
From scratch	0.589	0.308
Classical baseline	0.133	0.12
kNN floor	0.0214	0.0975
ACOPF	-	0.289
SCUC	1.00	-
DOPF	1.00	0.192
SE	0.835	0.196
DSSE	0.882	0.194
Contingency	1.00	0.197
Feasibility	0.836	0.195

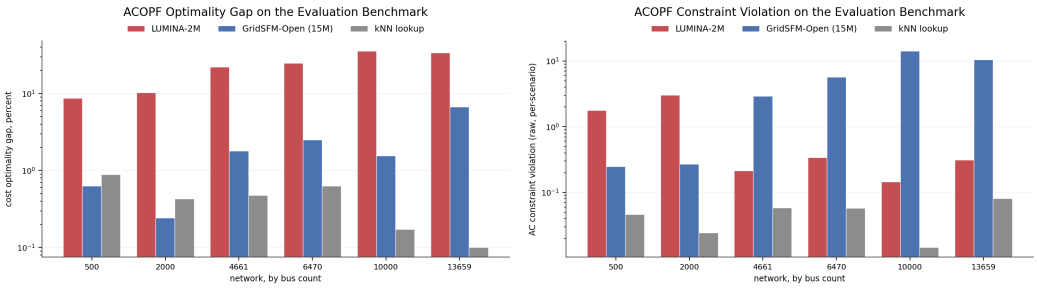


Figure 2: Public grid foundation models under the benchmark protocol. *Left:* cost optimality gap; *right:* physics-feasibility residuals. Both models track demand direction ($R^2 \approx 0.99$) but underrespond to it (regression slope ≈0.2) and over-generate by ≈13%, matching or underperforming the classical baselines and the kNN lookup on most cases. This suggests a failure to generalize to the dataset and indicates that our wider operating regime may shore up a gap in current foundation model training protocols.

The kNN floor that dominates in-distribution fares poorly on generalization ((Table 8)). We split the ACOPF lookup into a *within-topology* per-generator table (key: generator limits, cost coefficients, system loading; predict $p_g/p^{\max}$) and the identical *cross-topology* lookup with the target grid held out of the pool. Within topology the lookup reaches 0.05–1.7% cost-gaps and beats every learned model; with the target grid unseen it degrades by one to two orders of magnitude (4–80% gaps) even with local features chosen to favor it, while the topology-general foundation model holds 0.8–2.6% on its supported regime, up to $5\times$ (case500) to $\sim 96\times$ (case2000) better than the transferred lookup. Topology-aware message passing generalizes across grids in a way a local-feature lookup cannot.

4 Discussion

A key bottleneck for grid foundation models is task diversity more than raw scale. The corpora available so far (see Table 1) have concentrated the community's effort on a narrow task family, and multi-task representations on the same networks are scarce. We believe major data generation efforts should be dedicated to increasing the breadth and diversity of existing power systems datasets to represent a more complete subset of the tasks that a system operator would require from a true foundation model for the power grid. To date, optimal power flow problems (AC/DC) are heavily overrepresented in the data landscape, and while this is natural due to the ubiquity of the problems as computational primitives, we believe more attention should be devoted to unit commitment, contingency analysis, and reliability analysis. Similarly, a unified model of power grid operations requires the development of joint or co-simulated representations of transmission and distribution

Table 7: Cross-topology transfer. The encoder is pretrained on *all other networks of the same task* (leave-one-out) and adapted to the held-out network; *scratch* trains on the held-out network alone. Median over 3 seeds.

Task	Held-out network	Transfer (other nets, LOO)	Scratch	Base	kNN
ACOPF	case14 (NANO)	0.88	0.94	0.08	0.035
	case73_ieee_rts (SMALL)	0.45	0.66	0.24	0.016
	case118 (SMALL)	0.59	0.60	0.06	0.019
	case300 (SMALL)	1.80	1.81	0.13	0.024
	case2868_rte (MEDIUM)	0.66	0.79	0.12	0.016
	case8387_pegase (FULL)	715.5	715.2	6.40	0.34
SCUC	case60_c (SMALL)	0.32	0.31	0.17	0.095
	case118 (SMALL)	0.18	0.17	0.10	0.097
	case2746wp_k (MEDIUM)	0.065	0.064	0.02	0.014

Table 8: The kNN lookup fails at topology transfer (ACOPF cost-gap, lower is better). *Within*: per-generator lookup inside the target topology (the in-distribution floor). *Cross*: identical lookup with the target grid held out of the pool. GridSFM-Open and LUMINA-2M are scored frozen under identical splits.

Case	kNN within	kNN cross	GridSFM	LUMINA
case500	0.0005	0.0408	0.0076	0.0871
case2000	0.0008	0.2487	0.0026	0.1024
case6470	0.0164	0.5405	0.0257	0.2507
case4661	0.0170	0.7969	0.0185	0.2199
case2868_rte	0.0116	0.3851	0.0186	0.1662

systems. While the transmission and distribution tasks we provide are disjoint, we provide a unified representation and comparable scales across both system categories to allow for the development of transferrable models.

While this dataset represents a step-change in existing multi-task power systems benchmarks, both along the axis of task diversity and operating condition diversity, and provides the first public dataset of its size for both transmission and distribution system tasks, there are a number of caveats that apply to it. First, and most notably, is that the corpus is synthetic: load scenarios are drawn from a calibrated statistical model, and while aggregate statistics match published references, the gap to any real utility's operational data is unquantified. Second, cross-task coverage is uneven: reliability assessment is single-bus aggregated, and weather-correlated or cascading outages are not represented. Third, no time-domain simulation results are provided, as dynamics typically need machine parameters available for few of these networks and are often non-redistributable, and any time-series data are imputed from quasi-static time-series (QSTS) simulations rather than capturing true dynamics. Closing these gaps and expanding the benchmark evaluation to capture a greater diversity of candidate objectives is the natural next extension.

5 Conclusion

We present a multi-task benchmark for grid foundation models unified by a single heterogeneous graph schema and paired with an evaluation protocol. The released corpus spans ACOPF, SCUC, N-k contingency, SE, three-phase distribution state estimation, and distribution OPF. The distributed benchmark harness fixes tiers, generalization splits, compute tracks, classical-baseline-normalized metrics, and a input/output interface, and the experimental results establish reference model performance. By harmonizing networks, scenarios and representation across tasks ordinarily benchmarked in isolation, the benchmark makes multi-task and transfer questions tractable on a common, open, reproducible substrate.

REFERENCES

M. R. Asghar, G. Dán, D. Miorandi, and I. Chlamtac. Smart meter data privacy: A survey. *IEEE Communications Surveys & Tutorials*, 19(4):2820–2835, 2017.

R. Bommasani. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

C. S. C. Dande, L. Mattorolo, J. d. S. Andre, L. Lavecchia, N. Efkarpidis, and D. Toffanin. Synthetic grid generator: Synthesizing large-scale power distribution grids using OpenStreetMap. *arXiv preprint arXiv:2408.04923*, 2024.

S. Deng, L. Zhang, and D. Yue. Data-driven and privacy-preserving risk assessment method based on federated learning for smart grids. *Communications Engineering*, 3(1):154, 2024.

Z. Dong, K. Hou, H. Meng, X. Yu, and H. Jia. Data-driven power system reliability evaluation based on stacked denoising auto-encoders. *Energy Reports*, 8:920–927, 2022.

ENTSO-E. *Common grid model exchange standard (CGMES)*, 3.0 edition, 2020. Standardized as IEC 61970-600-1 and IEC 61970-600-2.

F. Fioretto, T. W. Mak, and P. Van Hentenryck. Predicting AC optimal power flows: Combining deep learning and Lagrangian dual methods. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(1):630–637, 2020.

D. M. Fobes, S. Claeys, F. Geth, and C. Coffrin. PowerModelsDistribution.jl: An open-source framework for exploring distribution power flow formulations. *Electric Power Systems Research*, 189:106664, 2020.

M. Gillioz, G. Dubuis, and P. Jacquod. A large synthetic dataset for machine learning applications in power transmission grids. *Scientific Data*, 12(1):168, 2025.

B. Habib, E. Isufi, W. van Breda, A. Jongepier, and J. L. Cremer. Deep statistical solver for distribution system state estimation. *IEEE Transactions on Power Systems*, 39(2):4039–4050, 2023.

A. Haller, K. Janowicz, S. J. D. Cox, M. Lefrançois, K. Taylor, D. Le Phuoc, J. Lieberman, R. García-Castro, R. Atkinson, and C. Stadler. The modular SSN ontology: A joint W3C and OGC standard specifying the semantics of sensors, observations, sampling, and actuation. *Semantic Web*, 10(1):9–32, 2019.

H. F. Hamann, B. Gjorgiev, T. Brunschwiler, L. S. Martins, A. Puech, A. Varbella, J. Weiss, J. Bernabe-Moreno, A. B. Massé, S. Choi, et al. Foundation models for the electric power grid. *Joule*, 8(12):3245–3258, 2024.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

International Electrotechnical Commission. *IEC 61968-11: Application integration at electric utilities—System interfaces for distribution management—Part 11: Common information model (CIM) extensions for distribution; IEC 61968-13: CIM RDF model exchange format for distribution (CDPSM)*. Geneva, Switzerland, 2013. IEC Standards.

International Electrotechnical Commission. *IEC 61970-301: Energy management system application program interface (EMS-API)—Part 301: Common information model (CIM) base*. Geneva, Switzerland, 2020. IEC Standard.

T. Joswig-Jones, K. Baker, and A. S. Zamzam. OPF-Learn: An open-source framework for creating representative AC optimal power flow datasets. *arXiv preprint arXiv:2111.01228*, 2021.

Thomas N. Kipf. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

M. Klamkin, M. Tanneau, and P. Van Hentenryck. PGLearn: An open-source learning toolkit for optimal power flow. *arXiv preprint arXiv:2505.22825*, 2025.

T. Lebo, S. Sahoo, and D. McGuinness. PROV-O: The PROV ontology. W3C recommendation, World Wide Web Consortium, 2013.

H. Li, J. H. Yeo, A. L. Bornsheuer, and T. J. Overbye. The creation and validation of load time series for synthetic electric power systems. *IEEE Transactions on Power Systems*, 36(2):961–969, 2020.

Q. Li, Y. Xu, C. Ren, and R. Zhang. Enhancing trustworthiness of data-driven power system dynamic security assessment via hybrid credibility learning. *IEEE Transactions on Power Systems*, 2025.

S. Lovett, M. Zgubic, S. Liguori, S. Madjiheurem, H. Tomlinson, S. Elster, C. Apps, S. Witherspoon, and L. Piloto. OPFData: Large-scale datasets for AC optimal power flow with topological perturbations. *arXiv preprint arXiv:2406.07234*, 2024.

LUMINA. LUMINA-2M v0.1.5a. Hugging Face `argonne/LUMINA-2M`, revision `v0.1.5a`, 2026. Apache-2.0 license; `config.json` + `model.safetensors`.

S. Meinecke, D. Sarajlić, S. R. Drauz, A. Klettke, L.-P. Lauven, C. Rehtanz, A. Moser, and M. Braun. SimBench—a benchmark dataset of electric power systems to compare innovative solutions based on power flow analysis. *Energies*, 13(12):3290, 2020.

Microsoft Research. GridSFM-Open v1.1. Hugging Face `microsoft/GridSFM_Open`, revision `1b41299`, 2026. MIT license; weights `gridsfm_open_v1.1.pt`.

Daniel K Molzahn and Ian A Hiskens. A survey of relaxations and approximations of the power flow equations. *Foundations and Trends in Electric Energy Systems*, 4(1-2):1–221, 2019.

Germán Morales-España, Jesus M. Latorre, and Andres Ramos. Tight and compact MILP formulation for the thermal unit commitment problem. *IEEE Transactions on Power Systems*, 28(4): 4897–4908, 2013.

A. Pinceti, L. Sankar, and O. Kosut. Synthetic time-series load data via conditional generative adversarial networks. In *2021 IEEE Power & Energy Society General Meeting (PESGM)*, pp. 1–5. IEEE, 2021.

Martin Ringsquandl, Houssem Sellami, Marcel Hildebrandt, Dagmar Beyer, Sylwia Henselmeyer, Sebastian Weber, and Mitchell Joblin. Power to the relational inductive bias: Graph neural networks in electrical power grids. In *Proceedings of the 30th ACM international conference on Information & Knowledge Management*, pp. 1538–1547, 2021.

A. K. Rivera, A. Bhagavathula, A. Carbonero, and P. Donti. PFδ: A benchmark dataset for power flow under load, generation, and topology variations. *arXiv preprint arXiv:2510.22048*, 2025.

S. Tu, Y. Zhang, J. Zhang, Z. Fu, Y. Zhang, and Y. Yang. PowerPM: Foundation model for power systems. In *Advances in Neural Information Processing Systems*, volume 37, pp. 115233–115260, 2024.

Mathias Uslar, Michael Specht, Sebastian Rohjans, Jörn Trefke, and José M González. *The Common Information Model CIM: IEC 61968/61970 and 62325-A practical introduction to the CIM*. Springer Science & Business Media, 2012.

A. Varbella, K. Amara, B. Gjorgiev, M. El-Assady, and G. Sansavini. PowerGraph: A power grid benchmark dataset for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 37, pp. 110784–110804, 2024. arXiv:2402.02827.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Álinson S. Xavier, Feng Qiu, and Shabbir Ahmed. Learning to solve large-scale security-constrained unit commitment problems. *INFORMS Journal on Computing*, 33(2):739–756, 2021. arXiv:1902.01697.

Y. Xie, L. Werner, K. Chen, T.-L. Le, C. Ortega, and S. Low. A digital twin of an electrical distribution grid: SoCal 28-bus dataset. *arXiv preprint arXiv:2504.06588*, 2025.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

A. S. Zamzam and K. Baker. Learning optimal solutions for extremely fast AC optimal power flow. In *2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pp. 1–6. IEEE, 2020.

X. Zheng, B. Wang, D. Kalathil, and L. Xie. Creation of synthetic networked PMU data: A generative adversarial network approach. *arXiv preprint arXiv:1908.08180*, 2019.

A Formulation

A.1 AC Optimal Power Flow

ACOPF minimizes generation cost subject to power-flow physics and operational limits. On a network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$:

$$\min_{p_g, q_g, v, \theta} \quad \sum_{g \in \mathcal{G}} c_g(p_g) \tag{4}$$

$$\text{s.t.} \quad p_i^{\text{inj}} = \sum_{j \in \mathcal{N}(i)} v_i v_j (G_{ij} \cos\theta_{ij} + B_{ij} \sin\theta_{ij}) \tag{5}$$

$$q_i^{\text{inj}} = \sum_{j \in \mathcal{N}(i)} v_i v_j (G_{ij} \sin\theta_{ij} - B_{ij} \cos\theta_{ij}) \tag{6}$$

$$p_g^{\min} \le p_g \le p_g^{\max}, \quad q_g^{\min} \le q_g \le q_g^{\max} \tag{7}$$

$$v_i^{\min} \le v_i \le v_i^{\max}, \quad |S_{ij}| \le S_{ij}^{\max} \tag{8}$$

where p_g, q_g are active and reactive generation, v_i, θ_i voltage magnitude and angle, and G_{ij}, B_{ij} branch conductance and susceptance. Constraints (2)–(3) are nonconvex, making ACOPF NP-hard in general.

A.2 Security-Constrained Unit Commitment

SCUC determines generator on/off status and dispatch over a multi-period horizon under security constraints, combining binary and continuous decisions:

$$\min_{u, p, r} \quad \sum_{t \in \mathcal{T}} \sum_{g \in \mathcal{G}} \left[c_g(p_{g,t}) + c_g^{\text{NL}} u_{g,t} + c_g^{\text{SU}} v_{g,t} \right] \tag{9}$$

$$\text{s.t.} \quad \sum_g p_{g,t} = \sum_d D_{d,t} + \text{losses}_t \tag{10}$$

$$p_g^{\min} u_{g,t} \le p_{g,t} \le p_g^{\max} u_{g,t} \tag{11}$$

$$p_{g,t} - p_{g,t-1} \le R_g^{\text{up}}, \qquad \sum_{\tau = t - T_g^{\min} + 1}^{t} v_{g,\tau} \le u_{g,t} \tag{12}$$

$$|f_\ell^{(k)}| \le f_\ell^{\max} \quad \forall \ell, t, k \in \mathcal{K} \tag{13}$$

with $u_{g,t} \in \{0,1\}$ the commitment status, $v_{g,t}$ startup, $c_g^{\text{NL}}, c_g^{\text{SU}}$ no-load and startup costs, R_g^{up} the ramp rate, $T_g^{\min}$ minimum up-time, and (10) enforcing N-1 security over $\mathcal{K}$.

SCUC adds binary on/off decisions and inter-temporal coupling to DCOPF over a horizon $\mathcal{T} = \{1, \ldots, T\}$, giving a mixed-integer linear program (MILP). We follow the formulation of Xavier et al. (2021), which builds on the tight and compact formulation of Morales-España et al. (2013). Let $\mathcal{B}$, $\mathcal{G}$ and $\mathcal{L}$ denote the sets of buses, generators and lines, and let $\mathcal{G}_b \subseteq \mathcal{G}$ be the generators at bus b. For each $g \in \mathcal{G}$ and $t \in \mathcal{T}$, the decision variables are the commitment status $x_{gt} \in \{0,1\}$, startup and shutdown indicators $z_{gt}, w_{gt} \in \{0,1\}$, power output $y_{gt} \ge 0$, the output $y_{gt}^k \ge 0$ above

$P_g^{\min}$ on each piecewise-linear cost segment $k \in \mathcal{K}$, and the spinning reserve $r_{gt} \geq 0$:

$$\min_{x,z,w,y,r} \quad \sum_{g\in\mathcal{G}}\sum_{t\in\mathcal{T}}\Big[c_g^S z_{gt} + c_g^0 x_{gt} + \textstyle\sum_{k\in\mathcal{K}} c_g^k y_{gt}^k\Big] \tag{14a}$$

$$\text{s.t.} \quad \textstyle\sum_{g\in\mathcal{G}} y_{gt} = \sum_{b\in\mathcal{B}} d_{bt}, \qquad \sum_{g\in\mathcal{G}} r_{gt} \geq R_t \qquad \forall t \tag{14b}$$

$$y_{gt} = P_g^{\min} x_{gt} + \textstyle\sum_{k\in\mathcal{K}} y_{gt}^k, \qquad 0 \leq y_{gt}^k \leq P_g^k \qquad \forall g,t,k \tag{14c}$$

$$\textstyle\sum_k y_{gt}^k + r_{gt} \leq (P_g^{\max} - P_g^{\min})\, x_{gt} - (P_g^{\max} - RU_g)\, z_{gt} - (P_g^{\max} - RD_g)\, w_{g,t+1} \qquad \forall g,t \tag{14d}$$

$$-RD_g \leq y_{gt} - y_{g,t-1} \leq RU_g \qquad \forall g,t \tag{14e}$$

$$\textstyle\sum_{i=t-UT_g+1}^{t} z_{gi} \leq x_{gt}, \qquad \sum_{i=t-DT_g+1}^{t} z_{gi} \leq 1 - x_{g,t-DT_g} \qquad \forall g,t \tag{14f}$$

$$x_{gt} - x_{g,t-1} = z_{gt} - w_{gt} \qquad \forall g,t \tag{14g}$$

$$-F_l^c \leq \textstyle\sum_{b\in\mathcal{B}} \delta_{lb}^c \Big(\sum_{g\in\mathcal{G}_b} y_{gt} - d_{bt}\Big) \leq F_l^c \qquad \forall l \in \mathcal{L},\ c \in \mathcal{L}\cup\{0\},\ t \tag{14h}$$

Here c_g^S is the startup cost, c_g^0 the cost of running at $P_g^{\min}$ for one period, and c_g^k the marginal cost on segment k (of width P_g^k), so production cost is convex piecewise-linear. R_t is the system reserve requirement, RU_g/RD_g are ramp limits, and UT_g/DT_g are minimum up/down times. Constraint equation 14d is the tight capacity bound of Morales-España et al. (2013), which also caps output in startup and shutdown periods. For units with $UT_g = 1$ it splits into two variants, and in the final period the $w_{g,t+1}$ term is dropped. Line flows use the DC approximation with injection shift factors δ_{lb}^c: $c = 0$ is the base case, and $c \in \mathcal{L}$ enforces N-1 security against the outage of line c with post-contingency limit F_l^c. Index windows are truncated at $t = 1$, and the ramping and logic constraints hold for $t \geq 2$. The number of constraints in equation 14h grows quadratically in $|\mathcal{L}|$, and most are inactive at the optimum. Once x is fixed, the other binaries follow from equation 14g and the problem reduces to a linear economic-dispatch program.

A.3 N-k Contingency Analysis

Contingency analysis evaluates the post-outage operating point after loss of one or more elements. For each contingency $k \in \mathcal{K}$ the network is modified by removing the outaged element(s) and an ACOPF is re-solved on the reduced topology. Where ACOPF varies the operating point on a fixed network, contingency analysis varies the network itself, exercising generalization across topology rather than across loading alone.

A.4 State Estimation

SE recovers the transmission network's operating state from redundant, noisy telemetry. The state is the vector of bus voltage phasors

$$x = \big[\,\theta_2, \ldots, \theta_N,\ V_1, \ldots, V_N\,\big]^\top, \tag{15}$$

the angles θ_i and magnitudes V_i, with a fixed angular reference $\theta_1 = 0$ at the slack bus. This yields a state vector with $2N-1$ unknowns, related to each other by a set of measurements $z \in \mathbb{R}^m$ through the AC power-flow equations:

$$z = h(x) + e, \qquad e \sim \mathcal{N}(0, R), \quad R = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2), \tag{16}$$

The measurement set includes real and reactive nodal injections P_i, Q_i, branch flows P_{ij}, Q_{ij}, and voltage magnitudes V_i; e.g. the injection functions are

$$P_i = V_i \sum_j V_j \big(G_{ij}\cos\theta_{ij} + B_{ij}\sin\theta_{ij}\big), \quad Q_i = V_i \sum_j V_j \big(G_{ij}\sin\theta_{ij} - B_{ij}\cos\theta_{ij}\big), \tag{17}$$

with $\theta_{ij} = \theta_i - \theta_j$ and $Y = G + jB$ the bus admittance matrix.

A.5 Distribution System State Estimation

DSSE is analogous to SE, but targeting three-phase unbalanced distribution feeders under partial observation. Where SE is typically overdetermined, DSSE is underdetermined, and the task is to recover voltage magnitudes and angles at all buses from limited, noisy measurements. Distribution networks are radial, three-phase unbalanced, and severely under-observed. The weighted least-squares form is

$$\min_{x} \sum_{m \in \mathcal{M}} \frac{(z_m - h_m(x))^2}{\sigma_m^2} \tag{18}$$

with state x, measurement z_m, measurement function $h_m(\cdot)$ and variance σ_m^2. For three-phase systems the state expands per phase, $x = [v_1^a, v_1^b, v_1^c, \theta_1^a, \theta_1^b, \theta_1^c, \ldots]^T$. The problem is under-determined as there are more states than measurements. Pseudo-measurements are thus typically required.

A.6 Distribution Optimal Power Flow

Given a three-phase feeder with loads fixed at the sampled operating point, DOPF selects photovoltaic real-power curtailment $c_i \geq 0$ and inverter reactive setpoints q_i to minimize total curtailment with a weighted penalty on reactive effort,

$$\min_{c,\,q} \;\; \sum_i c_i \;+\; \lambda \sum_i |q_i|,$$

subject to the LinDist3Flow linearized unbalanced branch-flow equations, per-phase voltage bounds, inverter reactive limits declared in the source models, and inverter apparent-power limits enforced by an inscribed polygon (conservative with respect to the true circular limit). Capacitor and regulator-tap positions are frozen at their power-flow values, which keeps the problem a linear program. The released label is this LP *surrogate optimum*, with every solution validated by an exact OpenDSS AC power-flow round trip (Appendix D.7); learned models are scored by per-unit state RMSE against the surrogate optimum, since a cost gap against a surrogate is not a meaningful oracle (Table 4).

A.7 Reliability Assessment

RA quantifies the ability to meet demand under uncertain component availability. It is a three-stage construction: an outage process generates system states, each state is resolved by a cost-minimizing dispatch, and reliability indices are estimated from the resulting shed variables.

Outage process. Each component $c \in \mathcal{C}$ (generator, branch or storage unit) is modelled as a two-state Markov process with failure rate λ_c and repair rate μ_c, so its steady-state unavailability — the forced outage rate — is

$$q_c = \frac{\lambda_c}{\lambda_c + \mu_c}, \qquad \mathrm{MTTF}_c = \lambda_c^{-1}, \quad \mathrm{MTTR}_c = \mu_c^{-1}. \tag{19}$$

Sequential sampling over the horizon yields availability masks $u_{c,t}^{(s)} \in \{0, 1\}$ for scenario s and hour $t \in \mathcal{T}$, preserving the temporal correlation of an outage that persists for its repair duration — which a state-sampling scheme drawing each hour independently would destroy.

Scenario dispatch. Given a mask, the scenario is a 24-hour dispatch that prices unserved energy at the value of lost load VOLL, so shedding is admissible but expensive:

$$\min_{p,e,s} \quad \sum_{t \in \mathcal{T}} \Big[\sum_{g \in \mathcal{G}} c_g(p_{g,t}) + \mathrm{VOLL} \cdot s_t \Big] \tag{20}$$

$$\text{s.t.} \quad \sum_g p_{g,t} + \sum_{b \in \mathcal{B}} \big(p_{b,t}^{\mathrm{dis}} - p_{b,t}^{\mathrm{ch}} \big) + s_t = D_t \tag{21}$$

$$u_{g,t}^{(s)} p_g^{\min} \leq p_{g,t} \leq u_{g,t}^{(s)} p_g^{\max} \tag{22}$$

$$e_{b,t} = e_{b,t-1} + \eta_b^{\mathrm{ch}} p_{b,t}^{\mathrm{ch}} \Delta t - \big(\eta_b^{\mathrm{dis}}\big)^{-1} p_{b,t}^{\mathrm{dis}} \Delta t \tag{23}$$

$$0 \leq e_{b,t} \leq \bar{e}_b, \qquad 0 \leq s_t \leq D_t \tag{24}$$

where s_t is load shed, $e_{b,t}$ storage state of charge, and the mask $u_{g,t}^{(s)}$ gates each generator's capability. Outaged branches are removed from the network before the solve, so a mask can render load unservable through topology as well as through capacity.

Index estimation. The indices are estimated from the solved shed variables, not from a capacity comparison, which is what makes transmission limits and storage dynamics enter the answer:

$$\mathrm{LOLE} = \frac{1}{N}\sum_{s=1}^{N}\sum_{t\in\mathcal{T}} \mathbf{1}\Big[s_t^{(s)} > 0\Big]\,\Delta t, \tag{25}$$

$$\mathrm{EENS} = \frac{1}{N}\sum_{s=1}^{N}\sum_{t\in\mathcal{T}} s_t^{(s)}\Delta t, \qquad \mathrm{EIR} = 1 - \frac{\mathrm{EENS}}{\mathbb{E}\big[\sum_t D_t \Delta t\big]}. \tag{26}$$

LOLP is the hourly rate $\mathrm{LOLE}/|\mathcal{T}|$.

Sampling cost. Because shedding is rare, the Monte Carlo estimator is governed by the coefficient of variation of the per-scenario energy not served, $\mathrm{ENS}^{(s)} = \sum_t s_t^{(s)}\Delta t$. Its relative standard error is

$$\beta = \frac{\sigma\big[\mathrm{ENS}\big]}{\sqrt{N}\,\mathbb{E}\big[\mathrm{ENS}\big]}, \tag{27}$$

and for a rare event of probability π the variance is dominated by the $\mathrm{Bernoulli}(\pi)$ factor, giving $\beta \approx \sqrt{(1-\pi)/(N\pi)}$, i.e. $N \approx (1-\pi)/(\pi\beta^2)$. At the observed $\pi \approx 0.026$ this requires $N \approx (1-0.026)/(0.026\times 0.1^2) \approx 3.75\times 10^3$ scenarios for $\beta = 0.1$, which is why the export is sized in thousands of risk scenarios (the released export carries ≈59,000 solved risk scenarios, comfortably above this floor) and why any learned surrogate must be scored against a positive-class baseline rather than on aggregate error.

B EXTENDED TABLES

This appendix collects the full-detail tables summarized in the main text: the resource comparison (Table 1), the per-task I/O contract (Table 14), the task tier split is given in (Table 9), and the enumerated per-case ACOPF, DSSE and SE inventories are given in task-specific tables below.

Table 9: Dataset size tiers across tasks (nested: each tier includes the smaller ones; cumulative case/feeder counts and sample totals). Transmission tasks tier by bus count (NANO is the single case14 network by specification; SMALL ≤ 300, MEDIUM $\leq 6{,}470$, FULL $\leq 13{,}659$ buses); distribution tasks tier analogously on feeder size. Cells are cases / samples. RA is a single system-level export and is not tiered.

	Task	NANO	SMALL	MEDIUM	FULL
Transmission	ACOPF	1 / 1,000	16 / 40,000	43 / 274,000	50 / 343,000
	SCUC	1 / 2,000	14 / 27,500	30 / 44,627	30 / 44,627
	SE	1 / 35,040	14 / 471,010	30 / 639,790	30 / 639,790
	N-k (3 obj.)[a]	1 / 6,300	16 / 764,700	42 / 34,225,500	50 / 41,442,192
Distrib.	DSSE, QSTS	4 / 140,160	12 / 420,480	32 / 1,121,280	44 / 1,541,760
	DSSE, IID[b]	4 / 1,000	12 / 6,000	32 / 14,200	44 / 19,000
	DOPF[c]	1 / 500	8 / 4,000	28 / 14,000	40 / 20,000

[a] Task instances summed over the three co-registered objectives (cost/loss/shed); solve labels are being regenerated (v2) and instance counts are the invariant quantity.

Table 10: SE corpus details. Transmission state estimation over the 30-network core topology at 15-minute resolution with hourly ACOPF redispatch. Small/mid networks carry the full calendar year (35,040 states)

Case	Buses	States	Conv.
case3_lmbd	3	35,040	35,040
case5_pjm	5	35,040	35,040
case14_ieee	14	35,040	35,040
case30_as	30	35,040	35,040
case30_ieee	30	35,040	35,040
case39_epri	39	35,040	35,040
case57_ieee	57	35,040	35,040
case60_c	60	35,040	35,040
case73_ieee_rts	73	35,040	35,040
case89_pegase	89	35,040	35,040
case118_ieee	118	35,040	35,040
case162_ieee_dtc	162	35,040	35,040
case197_snem	197	35,040	35,040
case300_ieee	300	15,434	15,434
case500_goc	500	15,817	15,817
case588_sdet	588	35,040	35,040
case1951_rte	1,951	2,754	2,753
case2000_goc	2,000	4,918	4,918
case2383wp_k	2,383	15,646	15,646
case2736sp_k	2,736	15,072	15,072
case2742_goc	2,742	6,506	6,506
case2746wop_k	2,746	12,011	12,011
case2746wp_k	2,746	14,249	14,249
case2848_rte	2,848	3,454	3,454
case2868_rte	2,868	2,954	2,952
case3970_goc	3,970	6,232	6,232
case4020_goc	4,020	6,232	6,232
case4601_goc	4,601	9,933	9,933
case4619_goc	4,619	11,451	11,451
case4837_goc	4,837	6,567	6,517
Total	**30 cases**	**639,790**	**639,737**

Table 11: DSSE corpus. One network description per feeder carries two sampling regimes: IID draws independent load scalings with no time axis, sampled from the feasible operating envelope and split randomly, while QSTS is a 15-minute calendar year whose states are temporally correlated and must be split temporally. "Nodes" counts energised bus-phases.

Feeder	Buses	Nodes	Loads	IID states	QSTS steps	Synth.
ieee13_qsts	16	41	15	—	35,040	
ieee34_qsts	37	95	68	—	35,040	
ieee37_qsts	39	117	30	—	35,040	
p4uhs0_4–p4udt4	71	156	36	1,000	35,040	
p6uhs12_1247–p6udt2478	174	353	91	1,000	35,040	
p13uhs19_1247–p13udt22083	260	638	118	600	35,040	
p13uhs3_1247–p13udt8678	424	1,191	167	600	35,040	
p9uhs11_1247–p9udt12047	460	1,380	164	600	35,040	
p6uhs7_1247–p6udt8181	615	1,181	394	600	35,040	
p10uhs3_1247–p10udt4601	691	1,464	445	600	35,040	
p13uhs11_1247–p13udt16436	806	2,038	390	600	35,040	
p9uhs13_1247–p9udt8152	864	1,485	567	400	35,040	
p6uhs8_1247–p6udt370	872	1,637	635	400	35,040	
p6uhs9_1247–p6udt2485	940	1,719	673	400	35,040	
p9uhs7_1247–p9udt14468	1,103	1,909	748	400	35,040	
p6uhs8_1247–p6udt6574	1,118	2,158	812	400	35,040	
p9uhs5_1247–p9udt1073	1,382	2,814	1,257	400	35,040	
p1uhs0_1247–p1udt1470	1,566	3,403	1,428	400	35,040	
p6uhs5_1247–p6udt4837	1,596	2,705	1,005	400	35,040	
p6uhs0_1247–p6udt1718	1,656	3,000	1,180	400	35,040	
p9uhs2_1247–p9udt97	1,789	3,108	1,267	400	35,040	
p13uhs24_1247–p13udt28487	2,177	4,880	1,627	400	35,040	
p6uhs0_1247–p6udt580	2,289	4,668	1,810	400	35,040	
p13uhs29_1247–p13udt9061	2,397	4,745	2,073	400	35,040	
p13uhs14_1247–p13udt25270	2,402	5,069	1,941	400	35,040	
p13uhs11_1247–p13udt7624	2,551	6,326	1,337	400	35,040	
p9uhs24_1247–p9udt21969	2,706	5,030	2,210	400	35,040	
p13uhs22_1247–p13udt26466	2,828	5,895	2,556	400	35,040	
p13uhs0_1247–p13udt15681	2,901	6,094	2,253	400	35,040	
p13uhs18_1247–p13udt23375	3,004	6,767	2,290	400	35,040	
p9uhs13_1247–p9udt10219	3,033	5,591	2,765	400	35,040	
p13uhs17_1247–p13udt7025	3,559	7,553	2,942	400	35,040	
p13uhs26_1247–p13udt28949	3,632	7,790	3,161	400	35,040	
p9uhs8_1247–p9udt6585	3,749	6,941	2,902	400	35,040	
p13uhs4_1247–p13udt13674	3,883	8,036	3,318	400	35,040	
p13uhs29_1247–p13udt19738	3,929	7,795	3,684	400	35,040	
p13uhs23_1247–p13udt28279	4,178	8,134	3,856	400	35,040	
p9uhs0_1247–p9udt20385	4,462	8,559	4,240	400	35,040	
p13uhs30_1247–p13udt33140	4,515	8,724	4,382	400	35,040	
p9uhs3_1247–p9udt19049	5,219	10,024	4,930	400	35,040	
Total	**40 feeders**			**17,200**	**1,401,600**	

Table 12: N-k contingency data over the core set of topologies. Each core network carries three objectives (cost, loss, shed) with equal sample count, so *total* is three times *samples/obj*.

Case	Tier	Contingencies	Scenarios	Samples/obj	Total samples
case4619_goc	medium	8,150	100	815,100	2,445,300
case4837_goc	medium	7,765	100	776,600	2,329,800
case4601_goc	medium	7,199	100	720,000	2,160,000
case4020_goc	medium	6,988	100	698,900	2,096,700
case3970_goc	medium	6,641	100	664,200	1,992,600
case2000_goc	medium	3,639	100	364,000	1,092,000
case2736sp_k	medium	3,504	100	350,500	1,051,500
case2383wp_k	medium	2,896	100	289,700	869,100
case1951_rte	medium	2,596	100	259,700	779,100
case500_goc	medium	733	100	73,400	220,200
case588_sdet	medium	686	100	68,700	206,100
case300_ieee	small	411	100	41,200	123,600
case197_snem	small	286	100	28,700	86,100
case162_ieee_dtc	small	284	100	28,500	85,500
case89_pegase	small	210	100	21,100	63,300
case118_ieee	small	186	100	18,700	56,100
case73_ieee_rts	small	120	100	12,100	36,300
case60_c	small	88	100	8,900	26,700
case57_ieee	small	80	100	8,100	24,300
case39_epri	small	46	100	4,700	14,100
case30_ieee	small	41	100	4,200	12,600
case30_as	small	41	100	4,200	12,600
case14_ieee	nano	20	100	2,100	6,300
case5_pjm	small	6	100	700	2,100
case3_lmbd	small	3	100	400	1,200
Total, 25 networks					**15,793,200**

C Schema and Representation Details

C.1 IEEE Common Information Model

We borrow from the Common Information Model (International Electrotechnical Commission, 2020) its separation of *conducting equipment* (physical devices), *terminals* (the points at which a device connects), and *connectivity* or *topological nodes* (the electrical nodes): our **bus** corresponds to a CIM topological node, so the standard transmission bus-branch abstraction is preserved rather than re-invented; our **generator**, **load** and **shunt** nodes are conducting equipment; and our **link** edges are terminals recording the incidence of a device on a bus. We also borrow the CGMES equipment/topology/state separation (ENTSO-E, 2020), with static equipment, a topology, and a solved state kept as distinct layers (Section D.8). For distribution we align to the CIM distribution profile (CDPSM) (International Electrotechnical Commission, 2013), and we adopt the PowerModelsDistribution (Fobes et al., 2020) distinction between an *engineering* model (components as an engineer describes them such as a transformer with a winding configuration, a load with a connection type) and the *mathematical* model that a solver flattens them into. Due to missing detail in some IEEE standard topologies, we only claim to align our vocabulary to these standards and cannot provide uniform compliance guarantees across all target topologies. This last is left to future work.

C.2 Node and Edge Classes

Six node classes are defined: **bus** (base voltage, value of lost load, temporal per-phase load, distributed generation, voltage measurements); **generator** (capacity limits, ramp rates, minimum up/down times, marginal and startup costs, piecewise curves, temporal states); **storage** (energy capacity, charge/discharge rates, efficiency, state-of-charge); **price-sensitive load** (revenue parameters, curtailable demand); **reserve** (ancillary requirements, shortfall penalties); and **shunt** (conductance, susceptance). Five edge classes are defined: **transmission line** (impedance r, x, b in per-unit,

Table 13: ACOPF corpus, enumerated in full. Every sample carries the primal solution and the dual variables. "Core" denotes whether a topology is in the multi-task core set.

Case	Buses	Samples	Core
case3_lmbd	3	1,000	✓
case5_pjm	5	1,000	✓
case14_ieee	14	1,000	✓
case30_as	30	1,000	✓
case30_ieee	30	1,000	✓
case39_epri	39	1,000	✓
case57_ieee	57	1,000	✓
case60_c	60	1,000	✓
case73_ieee_rts	73	1,000	✓
case89_pegase	89	1,000	✓
case118_ieee	118	5,000	✓
case162_ieee_dtc	162	5,000	✓
case179_goc	179	5,000	—
case197_snem	197	5,000	✓
case240_pserc	240	5,000	—
case300_ieee	300	5,000	✓
case500_goc	500	5,000	✓
case588_sdet	588	5,000	✓
case793_goc	793	5,000	—
case1951_rte	1,951	9,500	✓
case2000_goc	2,000	10,000	✓
case2312_goc	2,312	9,500	—
case2383wp_k	2,383	9,500	✓
case2736sp_k	2,736	3,500	✓
case2742_goc	2,742	10,000	✓
case2746wop_k	2,746	10,000	✓
case2746wp_k	2,746	9,500	✓
case2848_rte	2,848	10,000	✓
case2853_sdet	2,853	9,500	—
case2868_rte	2,868	10,000	✓
case2869_pegase	2,869	10,000	—
case3022_goc	3,022	10,000	—
case3120sp_k	3,120	1,000	—
case3375wp_k	3,375	10,000	—
case3970_goc	3,970	10,000	✓
case4020_goc	4,020	10,000	✓
case4601_goc	4,601	10,000	✓
case4619_goc	4,619	7,000	✓
case4661_sdet	4,661	10,000	—
case4837_goc	4,837	10,000	✓
case4917_goc	4,917	10,000	—
case6468_rte	6,468	10,000	—
case6470_rte	6,470	10,000	—
case6495_rte	6,495	10,000	—
case6515_rte	6,515	10,000	—
case7336_epigrids	7,336	10,000	—
case8387_pegase	8,387	9,000	—
case9241_pegase	9,241	10,000	—
case10000_goc	10,000	10,000	—
case13659_pegase	13,659	10,000	—
Total	**50 cases**	**343,000**	

Table 14: Per-task executable I/O contract (also provided in the code supplement). "Prohibited at inference" lists solver outputs / targets a submission must not consume as inputs. N buses, G generators, T=36 SCUC hours, $P \leq 3$ phases; p.u. is per-unit on the system MVA base. (Section 2.1).

Task	Inputs (dims)	Targets (dims)	Angle ref.	Prohibited at inference
ACOPF	`bus.x`$[N,4]$, `gen.x`$[G,11]$, `load.x`$[L,2]$, branch feats	`bus.y`$[N,2]$=$[va, vm]$, `gen.y`$[G,2]$=$[p_g, q_g]$; duals$[N,2]$	ref. bus, va=0	`bus.y`, `gen.y`, duals, objective, status
SCUC	`bus.x`$[N,2]$, `gen.x`$[G,2]$, `bus_load`$[N,T]$	`is_on`$[G,T]$, `pg`$[G,T]$, `LMP`$[N,T]$	DC slack	`is_on`, `pg`, `LMP`, objective
Cont.	`bus.x`$[N,4]$ $[p_d, q_d, w_p, w_q]$, post-outage topology	`bus`$[N,2]$, `gen`$[G,2]$, `insecure`, `severity`	ref. bus	solution, `insecure`, `severity`, status
DSSE	`bus.x`$[NP,2]$=$[P,Q]$ (noisy/pseudo), branch	`bus.y`$[NP,2]$=$[vm, va]$	per-phase source	`bus.y`, converged
SE	`bus.x`$[N,2]$=$[P,Q]$ injections, topology	`bus.y`$[N,2]$=$[vm, va]$	ref. bus	`bus.y`
DOPF	`bus.x`$[NP, \geq 2]$ load+PV avail., branch	`pv_curtail`, `pv_q`, `bus_vm`$[NP]$, active set	per-phase source	setpoints, `bus_vm`, objective, active set
RA	`X`$[n,$ feat$]$ demand+outage mask+storage	`shed`$[T]$, `ENS`, indices (EUE/LOLE)	n/a (1 bus)	`shed`, `ENS`, indices, dispatch

thermal ratings, angle limits, contingency status); **transformer** (line fields plus tap ratio and phase shift); and **link** edges (load, generator, shunt) connecting component nodes to parent buses.

C.2.1 Entity Versus Role

The profile is *identity first, role second*. A device's identity is stable across the corpus; whether it is in service, committed, dispatched, providing reserve, curtailable, or measured is a role or state it carries under a given scenario and formulation, not a change of identity. This is why equipment identifiers survive the projection and why the same component recurs across tasks that share a network. It also forces one distinction we keep explicit: the *electrical basis* (positive-sequence single-phase, or phase-conductor three-phase) is separate from the *formulation* (AC, a DC approximation, or a linearized unbalanced model). We do not encode a formulation by silently reinterpreting a feature position, for instance by writing a zero into a reactive-power slot to signal "DC", because *not represented by this formulation* is not the same as *measured to be zero*, *unknown*, or *unobserved*. Each stored quantity carries its representation status, so a consumer can tell a modelling omission from a physical value.

C.2.2 Schema layers

The canonical model is organized in five layers, making explicit what the release ships as `base`, `variations` and `samples`. *Network* N is the physical asset consisting of the topology, equipment identities and ratings. This is stable across scenarios. *Scenario* S is an operating condition drawn over the network: loads, availabilities, renewable levels, a topology variation. *Formulation* F_τ is the mathematical problem posed on (N, S) for task τ, including its electrical basis, objective and constraint set. This is a necessary distinction as the same (N, S) supports (e.g.) both an ACOPF and a DC-OPF formulation, so the formulation is not a property of the network. *Observation* O_τ is what is measured and how; here we adopt, conceptually, the SOSA/SSN (Haller et al., 2019) separation of entity, observed property, procedure (the instrument model), result and time, which keeps the DSSE noise model an evaluation-time *procedure* rather than a value baked into the artifact (Section D.8). This is necessary in order to allow the variation of observability from the same stored dataset for state estimation tasks. *Result* R is the output of solving a formulation (dispatch, voltages, commitment, a feasibility label, duals). We adopt PROV-O-style provenance (Lebo et al., 2013)

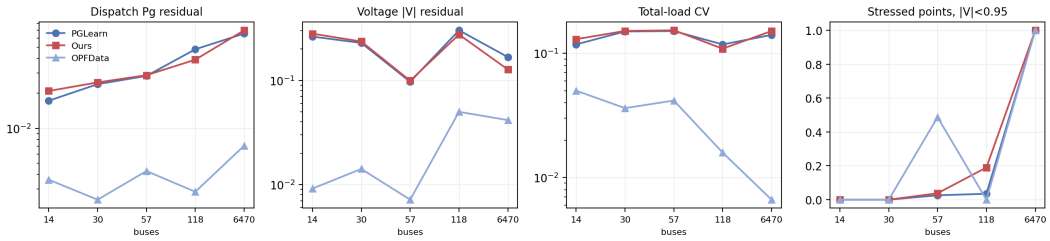


Figure 3: Operating point diversity measures across OPFData, PGLearn, and our dataset. We match PGLearn on a majority of diversity metrics while providing greater coverage of cases at or near stressed voltage levels, and a substantially broader set of operating conditions than OPFData, across a broader range of tasks.

wherein a result is produced by a specific execution of a specific formulation with specific inputs and tolerances.

C.2.3 Units and Normalization

The profile carries enough normalization metadata to interpret a stored array unambiguously, with power and voltage bases (system MVA base, base kV), the angle reference, whether a quantity is per-phase or a total, and the sign convention.

C.2.4 Scope

We adopt the CIM vocabulary and its equipment/terminal/node and equipment/topology/state distinctions as an application profile, and the SOSA/SSN and PROV-O separations conceptually. We do *not* ship a full CIM/CGMES object model, an RDF or JSON-LD serialization, and we do not adopt OEO, IEC 61850 or SAREF. Those are steps towards a fully standards-native footing, but certified standards compliance is left to future work as explained above.

D Corpus Construction

D.1 AC Optimal Power Flow

ACOPF instances are generated over PGLib-OPF with PowerModels.jl and an interior-point solver. Operating points come from a two-level sampler: a global loading factor $b \sim \mathcal{U}(b_{\text{lo}}, b_{\text{hi}})$ scales the demand vector, and an independent multiplier $\varepsilon_i \sim \mathcal{U}(0.8, 1.2)$ is applied per load. The interval $[b_{\text{lo}}, b_{\text{hi}}]$ is *calibrated per case* by bisection toward infeasibility, so each network is sampled across its own usable loading range. This matters for a foundation-model corpus: a range imposed uniformly across networks can leave large systems in a narrow, nearly constant active set.

Sample depth follows network size: 1,000 samples up to 100 buses, 5,000 up to 1,000 buses, 10,000 above. Every solved instance stores, alongside the primal solution, the *dual variables* of the nodal power-balance constraints and branch flow limits. Three conventions must be carried to use them. The locational marginal price is $\text{LMP}_{\$/\text{MWh}} = -\lambda_{\text{kcl}}/\text{baseMVA}$ (the raw PowerModels active-balance dual is negative; on case14 this yields 7.9–60.6 $\$/\text{MWh}$, physically sane). Branch multipliers are stored as `NaN` off the binding set, denoting a feature marking a non-binding limit. And because the ACOPF is a slack-relaxed formulation (per-bus p/q and per-branch thermal slacks penalized at 3×10^5), the balance duals at buses where slack activates are dominated by the penalty and are *not* true prices; they are filtered by `total_power_slack`>0 before use, and a small fraction of solved scenarios ($\approx$3.4% on case14) carry no retrievable duals (`NaN`) and are filtered likewise. Duals are converged to the solver's default tolerance (Ipopt/MA27). These are what make the corpus usable for Lagrangian and constraint-aware objectives. The corpus covers 50 networks and 343,000 solved instances, of which 30 networks and 169,000 instances form the core set shared with SCUC. Per-case achieved counts, including cases short of their band target, are given in the inventory.

Although operating-point coverage is valuable for task generalization more generally, variation in operating conditions does not necessarily guarantee the diversity of the *binding-constraint set*. We count unique *convergent* active-constraint sets by a primal-at-bound criterion (a constraint is active iff its variable sits on a box bound within 10^{-4} of a converged AC-OPF solve: generator p_g on $[p^{\min}, p^{\max}]$ and q_g on $[q^{\min}, q^{\max}]$, bus $|V|$ on $[v^{\min}, v^{\max}]$) and compare directly against OPFData (Lovett et al., 2024) at matched sample budgets. On case14, case30 and case118 our corpus spans 13, 15 and 2,080 distinct active sets against OPFData's 5, 2 and 1,231. Ours exceeds OPFData on every case, and the case118 ordering is robust to the tolerance (ours 1,969–2,210 vs OPFData 1,231–1,243 over 10^{-5}–10^{-3}) (Figure 4a).

The gap is primarily due to the fact that the correlated-b sampler moves system-total load (cv ≈11% vs OPFData's ≈0.4%), so more distinct constraints bind across samples. As a control, the generator bound columns of OPFData's arrays were validated against the MATPOWER case file before counting, and the criterion, budget and convergence filter are identical across corpora. These values are distinct from single-corpus active-set counts computed under a different definition, and should just be understood as a controlled cross-dataset comparison measure. OPF-Learn's (Joswig-Jones et al., 2021) published counts instead use a *non-converged* nonzero-dual criterion whose interior-point barrier leaves duals never exactly zero, so nearly every sample registers a distinct set. These are not comparable to a primal-at-bound count and are not included to avoid an apples-to-oranges comparison.

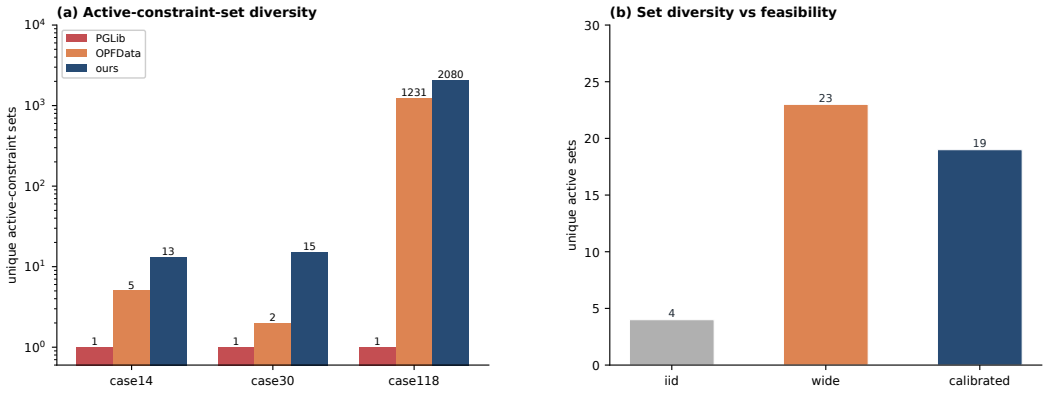


Figure 4: Active-constraint-set diversity. **(a)** Unique *convergent* active-constraint sets (primal-at-bound, TOL 10^{-4}: generator p_g/q_g and bus $|V|$ on their box bounds). Ours (13/15/2,080) exceeds OPFData (5/2/1,231); the case118 ordering holds across TOL 10^{-5}–10^{-3}. The correlated-b sampler moves total load (cv ≈11% vs OPFData's ≈0.4%), producing more distinct binding regimes; OPFData's generator bound columns were validated against the MATPOWER case file; the count axis is logarithmic. OPF-Learn's counts (thousands per case, e.g. 6,938/9,931/9,980) use a non-comparable non-converged nonzero-dual criterion and are not plotted. **(b)** Diversity is worthless without feasibility, on case30 over 2,000 samples: an uncalibrated iid window ($\mathcal{U}[0.8, 1.2]$, 19% slack-relaxed) yields few sets, a wide window ($[0.9, 1.5]$) inflates the count to 23 but 77% of those sets are slack-relaxed (not AC-OPF solutions), and the calibrated window ($[0.6, 1.0]$) reaches nearly the same diversity (19) at 1.1% relaxation.

The correlated global loading factor is what makes dispatch a learnable target, and its effect is reported here as a measured quality property of the corpus. Under independent per-load perturbation the system total averages to a near-constant across samples so dispatch does not move, a per-unit constant predictor of generator active power scores $R^2 = 0.9997$, and physics accounts for under 1% of the target variance, leaving model comparisons inside the noise floor. Under the two-level sampler the total-load coefficient of variation is 11–15% (≈ 21× the independent-perturbation value at the corpus median, and ≈ 25× on the 2,000-bus case), and the constant-predictor R^2 is 0.947 for active power and 0.71 for voltage magnitude. On small networks the active-power constant-R^2 remains high (0.95–0.98) because merit order pins the majority of generators at their bounds, so generator active power on small grids is substantially cross-sectional and is the wrong target on

which to argue a dispatch-modeling win. Voltage and reactive targets do not share this and remain learnable throughout.

D.2 Security-Constrained Unit Commitment

SCUC instances are produced in Julia with UnitCommitment.jl and the HiGHS MILP solver. Load scenarios on the base topology across 30 networks, extending commitment data to network families for which none previously existed publicly. On large systems the majority of units are baseload and only a minority carry a scenario-dependent commitment. Mid-size networks (e.g. case118, case197, case1354) vary a majority of units across scenarios and are the credible commitment benchmark. Dispatch is a learnable target across all scales, and commitment uniqueness is high on non-trivial networks.

D.3 Contingency Analysis

The contingency corpus pairs operating-point variation with *topology* variation. For each of 25 PGLib-OPF networks, load scenarios are drawn and, for every scenario, the full non-islanding N-k set is solved: the intact base case, the outage of each line and transformer whose removal does not disconnect the network, and the outage of each generator. The contingency count per network therefore equals its branch count, capped at 3,000 (Table 12).

Load scenarios. Each network carries $S = 100$ load scenarios, and contingency data varies primarily in topology variations.

Each scenario s rescales every load independently,

$$p_i^d \leftarrow \alpha_i \, p_i^d, \qquad q_i^d \leftarrow \beta_i \, q_i^d, \qquad \alpha_i, \beta_i \overset{\text{iid}}{\sim} \mathcal{U}[0.9,\, 1.1],$$

with the random stream seeded by the scenario index alone. Seeding by s makes the perturbation reproducible and *shared*: scenario s applies the identical load vector across all contingencies of a network and across the three objective variants, so the corpus is co-registered along both axes. This is the OPFData-style independent per-load sampler; because independent noise averages out across loads, total demand stays within a few percent of nominal, and the stress in this corpus comes from the outages rather than from the loading level.

Task structure. The unit of generation is the (scenario, contingency) pair. For each network the task queue is the cross product of S scenarios with the contingency set, plus one base-case task (no outage) per scenario:

$$N_{\text{tasks}} \;=\; S \cdot (C + 1),$$

where C is the contingency count. Each task yields one HDF5 record; the same queue is solved once per objective, so a network contributes $3\,S\,(C+1)$ records in total.

Per-task solves. Each task applies the scenario's load perturbation and the outage to the base network, then runs two AC solves in polar form:

1. a flat-start AC power flow at the network's reference dispatch, recording whether the pre-contingency setpoints remain AC-feasible on the post-outage topology (security of the uncorrected dispatch);
2. a full AC-OPF re-dispatch on the post-outage topology. The strict OPF is attempted first; if it fails to converge, the solve is repeated with a relaxed formulation that adds per-bus active/reactive power-balance slacks and per-branch thermal slacks, each penalized at 3×10^5 in the objective, so that stressed instances still return an operating point together with the slack magnitudes needed to reach it.

Solves use Ipopt with tolerance 10^{-8}. Records store the perturbed load vector, both solver statuses, and, for converged solves, bus voltages (va, vm), generator dispatch (pg, qg), and per-load served power; for the relaxed OPF the total power and line slack are stored explicitly. Following the label-fidelity audit, security labels are defined from the power-flow convergence and from the OPF slack magnitudes, not from the OPF termination status, which the penalized formulation renders uninformative.

Objective variants. The same task queue is labeled under three OPF objectives, giving three co-registered label sets per network:

- **cost**: polynomial generation cost from the case data, the standard PGLib objective, plus the slack penalties when relaxed;
- **loss**: total active generation $\sum_g p^g$, which at fixed served load is equivalent to minimizing network losses and requires no cost data;
- **shed**: total power-balance slack (unserved power), solved with the slack formulation always active and a small $10^{-3} \sum_g p^g$ tie-break; this is the min-shed / max-load-served objective, and is non-trivial precisely on the stressed post-outage instances.

Contingency sets are enumerated per scenario, so set size varies by a few elements between scenarios of the same case; totals are summed per scenario, not multiplied. The total set yields 5,264,400 solved states, i.e. 15,793,200 samples across three objectives (cost, loss, shed).

D.4 Distribution System State Estimation

Unlike SE, DSSE is generated on unbalanced three-phase feeders. A distribution feeder is unbalanced by construction, so a positive-sequence benchmark measures a different problem than the one operators face, and results against it are not comparable to three-phase estimators.

The corpus is a single set of feeders carrying two load sampling regimes over one network description. Both regimes are generated from the same OpenDSS models through the same code path and share an identical `base/` group; what differs is how the operating points are drawn, and that difference is confined to a regime group in the file (Section D.8). The IID regime draws independent load scalings with no time axis and supplies observability and load-range diversity; the QSTS regime is a quasi-static time series in which load follows real loadshapes, giving one calendar year at 15-minute resolution, 35,040 timesteps per feeder. Each regime declares its own split policy, which is the one place the distinction genuinely matters: IID samples are exchangeable and may be split at random, whereas QSTS states must be split temporally.

The regimes are complementary. The IID load range of 0.40–1.10 of nominal is deliberately wider than a real year ever spans, so it visits operating points no trajectory reaches; QSTS supplies the realism and temporal structure that independent draws cannot. Neither subsumes the other, and Table 11 reports both per feeder.

D.4.1 Scenario-sampled regime (IID)

Instances are produced with OpenDSS over the 40 SMART-DS feeders of Table 11, covering 17,200 solved states. The feeder set is drawn from SMART-DS, chosen for coverage of the axes along which distribution networks differ: voltage class from 240 V secondary to 35 kV primary, radial and lightly meshed configurations, overhead and underground construction, and scale from tens to several thousand buses. Scenario depth is weighted by network size because per-scenario solve cost grows sharply superlinearly in bus count, and a flat per-feeder count concentrates nearly all compute in the few largest networks for far fewer distinct states. Every quantity is per phase, with phase presence carried as feature width: bus targets are six columns (magnitude and angle on each phase) with absent phases masked, and the graph shape matches the transmission tasks.

Each solved state is paired with 48 measurement placements (three observability densities, sixteen samples each), giving 825,600 training samples. Placements vary the observed-bus mask only, not the underlying grid state.

Observed loads carry measurement noise, and a fraction of loads is unknown. Without both, a weighted least-squares estimator saturates the task and the benchmark measures nothing. On the released corpus a per-component constant predictor scores a voltage-magnitude constant-R^2 of 0.85–0.87.

D.4.2 Yearly quasi-static regime (QSTS)

The corpus covers 40 feeder-years each solved at all 35,040 timesteps of the year with no non-convergence, occupying approximately 36 GB before screening. The layout stores state *per ener-*

gised node, and load as a shape library plus a per-load index rather than per load per step; with compression this measures 6.2 bytes per node-state.

Splits over this corpus must be *temporal*, for instance months 1–9 for training and 10–12 for test. A random split over 15-minute states leaks almost perfectly, because the neighbouring state is very nearly the same state.

The Tier A IEEE and EPRI feeders are static test cases and ship no yearly loadshapes, so SMART-DS load profiles are attached to them by customer class: single-phase loads at or below 25 kW are treated as residential and all others as commercial. The profiles themselves are real and only the mapping of profile to load is synthetic.

D.5 State Estimation

DSSE poses state estimation only on distribution feeders, so nothing in the corpus lets a model be trained on one voltage level and evaluated on the other. SE is the transmission counterpart, specified so that the two are the *same task*: one calendar year per network at 15-minute resolution, the same HDF5 layout, and the same temporal split policy. Transmission is represented in positive sequence, so there is exactly one node per bus and the per-phase node index degenerates to phase zero; that single convention is what allows one reader and one encoder to consume both corpora.

Shared load process. Bus loads are driven by the same 15-minute load profile pool that drives DSSE-QSTS. This is the substantive choice: a transfer claim between voltage levels is only meaningful if the underlying demand process is common, and pairing a transmission corpus with balancing-authority demand while distribution runs on SMART-DS would leave the two with unrelated temporal structure. Two adjustments are required. Aggregation over many customers smooths, so each bus draws a mixture over a random subset of profiles to preserve the coincident system peak while supplying per-bus diversity. The leading principal component of the annual state is then 0.618, against 0.942 for a distribution feeder-year. And the pool carries residential and commercial classes only, with no industrial profiles, which would overstate the diurnal swing at transmission buses; a synthetic low-variance industrial component is mixed in. Both the profile-to-bus pairing and the industrial component are synthetic and are declared as such.

Dispatch and its cadence. Load alone does not determine a solvable power flow so the operating point is set by ACOPF at a fixed cadence, economic redispatch between ACOPF solves, and an AC power flow at every step to record the state. The cadence is not merely a cost parameter. With generator voltage setpoints held across a day, bus voltage magnitudes are pinned and carry almost no temporal variation; re-optimising hourly raises the mean per-bus temporal standard deviation of $|V|$ by a factor of 4.8 and is what makes the magnitude half of the state recoverable at all. Angles vary strongly regardless, since they follow the dispatch directly.

Released scope. The released SE corpus is complete over the full 30-network ACOPF+SCUC core: the small and mid networks carry the full calendar year ($35{,}040$ 15-minute states), and the larger networks carry the largest contiguous window that fit the compute budget, floored so every case clears $\geq$ 2,000 converged states for 639,790 solved states in total. The SE *targets are AC power-flow states*, under the PowerModels generator a full AC power flow is solved at *every* 15-minute step, and the operating point is re-anchored by a genuine ACOPF every fourth step, with an economic DC redispatch and AGC-like participation tracking on the intervening steps. These states are learnable: per-component constant-R^2 on voltage magnitude runs 0.65–0.94, and the periodic re-optimization keeps $|V|$ from collapsing to a constant. What separates cases is the *dispatch* that sets the operating point, not the physics of the state: on `case162_ieee_dtc`, for example, all $35{,}040$ steps carry a converged AC power-flow state, but its dispatch is set by ACOPF on 25% of steps, DC economic redispatch on 58%, and AGC fallback on the remaining 17%; the per-step `dispatch_kind` is stored so a consumer can filter on economic-dispatch realism; all 30 networks are AC-grounded (every stored step carries a converged AC power-flow state). The corpus is single-solver (Ipopt/MUMPS via the PowerModels port, `format_version` 4) and schema-consistent with the exagrid ACOPF corpus, with the per-seed shards retained as provenance behind the clean per-case release files. $|V|$ is dynamic on all 30 networks and the mean-state baseline does not win, with the angle metric taken within-bus/slack-referenced. It is released as the transmission counterpart of DSSE, so state estimation is posed on a common representation across the transmission–distribution boundary.

D.6 Reliability Assessment

RA instances are exported from the A-LEAF framework. Each record is one solved risk scenario over a 24-hour horizon carrying hourly demand, the component outage mask, dispatch, load shed and energy not served, storage state, branch flows where available, and solver metadata; the released conversion covers a 30-day, 5,000-risk export.

Two properties bound the task. The export uses an aggregated regional representation with a single bus and no branches, so it exercises temporal and outage structure but not graph structure; a network-flow variant with explicit topology exists in the raw export but is not part of the released conversion. And positive energy-not-served events are rare — approximately 2.6% of the natural training split — so any reported metric requires a positive-class baseline, for which positive-only and oversampled training variants are provided. All three standard reliability indices — expected unserved energy (EUE), loss-of-load expectation (LOLE), and a loss-of-load / risk index — are reported from the one scenario ensemble; EUE is the primary index for the aggregate score because it measures magnitude and not merely occurrence, with LOLE as the fallback where an export stores only a loss-of-load flag.

D.7 Distribution Optimal Power Flow

DOPF is released on the 35 SMART-DS feeders that carry DSSE, at 500 operating points each for 17,500 instances, so that the distribution side of the corpus is multi-task on a shared feeder set in the same way ACOPF and SCUC share the transmission core. Each sample carries the photovoltaic real-curtailment and inverter reactive-power setpoints, the resulting bus voltage state, the objective, a convergence flag, and the active set.

Formulation. Instances are posed as LinDist3Flow, a linear unbalanced branch-flow model, and solved as a linear program. Controls are photovoltaic real curtailment and inverter reactive power; the objective minimises curtailment with a weighted penalty on reactive effort. Inverter apparent-power limits are enforced by an inscribed polygon, which is conservative with respect to the true circular limit, and inverter reactive limits declared in the source models are enforced as bounds in the optimisation. Capacitors and regulator taps are held at their power-flow positions, since they are discrete and including them makes the problem a mixed-integer program. Local droop controllers present in the source models are not active during generation, so that the recorded state corresponds to the computed setpoints.

Surrogate optimum with AC-validated states. The label is a LinDist3Flow LP *surrogate optimum* whose states are validated by an exact OpenDSS AC power-flow round trip, and not an "exact feasibility" label. Every solution is round-tripped: the optimised setpoints are applied, the full AC power flow is solved, and the resulting state and violations are recorded. Four quantities that a single "feasible" flag would conflate are stored and reported separately, because they *dissociate* across feeders. (i) *AC convergence* of the round trip ranges from 1.8% to 100% (mean 84.2%); a few feeders barely converge. (ii) The *LP-vs-AC voltage MAE* `lindist_vm_mae` is small (median 3.1×10^{-2} p.u.), and samples above the 0.05 p.u. tolerance are discarded. (iii) The *AC voltage-bound violation rate* is high: on most feeders 70–100% of AC-validated states violate the $[0.95, 1.05]$ band (over- voltages to 1.37 p.u.), so a small surrogate MAE does *not* imply an AC-feasible optimum. (iv) *Inverter saturation* (`inverter_at_limit`) holds on $\approx$$71\%$ of samples on nearly every feeder. The stored `bus_vm_pu` state is the exact OpenDSS AC state regardless; it is the LP *setpoints* that carry the surrogate disagreement, and a few hard split-phase feeders retain only 25–30% of their samples at the MAE gate.

Controllable devices. The distribution models underlying DSSE carry no distributed generation, so DOPF is generated on the SMART-DS solar and battery scenarios, which ship photovoltaic systems, storage and capacitors over the same feeders alongside identical load and line definitions.

Difficulty metric. DOPF is scored by the per-unit RMSE of the recovered controls and voltage state (curtailment, inverter reactive power, and bus voltage magnitude) against a mean state predictor, consistent with the other state recovery tasks. The active set is additionally stored per sample in order to allow for degeneracy screening.

D.8 Storage Format

Data are stored in HDF5 with an indexed schema that minimizes redundancy. Static network parameters are held once in a `base` group; scenario-varying inputs (load profiles, costs) and topology variations (line service status vectors) in a `variations` group; and each solved sample in a `samples` group carrying integer indices into the scenario and topology tables, solver metrics, and solution arrays. Numerical data use 32-bit precision with chunked, compressed storage.

Where one network supports more than one way of drawing operating points, those are stored as *regimes* under a shared `base` group. DSSE uses this: a feeder carries `regimes/iid` and `regimes/qsts` beside one network description, and each regime declares its own `split_policy` (`random` and `temporal` respectively) so that a loader cannot silently apply a random split to correlated states.

E Benchmark Protocol Details

Per-task I/O contract. The per-task input/output contract is fixed executably: with the supplied schema files in the code supplement declaring, for every task, the input fields, target fields, tensor dimensions, masks, units, the angle reference, and the fields prohibited at inference (Table 14). This is the projection contract of Section 2.1: each field is a named projection choice and an absent field is typed ("not represented" $\neq$ "measured zero"). Because every transmission task presents the same typed node interface, a single encoder consumes an ACOPF, a SCUC and an SE instance unchanged, and the only required difference to train across tasks is that the head that maps node embeddings to targets differs (ACOPF regresses $[va, vm]$ and $[p_g, q_g]$; SCUC swaps in a commitment-plus-dispatch head over the same generator nodes with an LMP head on buses; SE estimates $[vm, va]$ from bus injections).

Tiers. Four case-size tiers scale the compute demand while keeping the task contracts fixed: a single-network NANO tier, a SMALL tier of networks up to 300 buses; a MEDIUM tier up to $\sim$6,470 buses; and a FULL tier spanning the whole library to 13,659 buses. The distribution tasks are tiered analogously on the feeder size.

Splits and leakage control. Every submission reports one in-distribution split and four generalization splits, shipped as fixed published index files (`splits/<task>/<split>.json`) at a frozen seed. *In-distribution* holds out disjoint samples of the training networks. *Topology transfer* holds out whole networks. *Scale transfer* trains on networks $\leq$ 300 buses and tests on networks $\geq$ 2,000. *Regime shift* partitions within each network by loading factor (below-median b trains, above-median tests). *N-k perturbation* trains on the base topology and tests on post-outage topologies. Partitions that could leak an operating point are grouped by the underlying object, not the row: DSSE splits by *solved state* before the 48-measurement-placement fan-out, and contingency by the shared *load scenario*; a sample with no group is its own group. For cross-task transfer, a network held out to test the target is removed from every source task's training split, so pretraining cannot leak it. A task$\times$split applicability matrix is published; a non-applicable (task, split) cell is blank, not zero.

Reference-model training. Reference models are heterogeneous GNNs (per-node-type projection to a 64-dim hidden space; three `HeteroConv` layers, sum aggregation, ReLU; GraphSAGE default, with GCN/GIN/GAT/HGT variants) plus a stock MLP. Each cell is trained with Adam (lr $= 10^{-3}$), batch 64, up to 200 epochs under a 25-minute per-cell wall-clock budget, on up to 2,000 scenarios, with a 20% validation fraction and MSE loss (optional physics/cost objectives for ACOPF/DOPF), repeated over three seeds and three label budgets (100/10/1%). Per-cell wall/device seconds and train/val sizes are recorded in every scorecard and released in the results file.

E.1 Transfer study

This section details generalization results across tasks and across topologies. All results use the shared heterogeneous-GNN encoder, pretrained on one or more *source* tasks (or, for topology transfer, leave-one-out on the same task's *other* networks), then adapted to the target by loading the encoder and training a fresh task head on the target's labeled split under a fixed compute budget,

scored on the identical held-out split as every other method. We compare *scratch* (single-task, no pretraining), *frozen* (encoder fixed) and *finetune* (encoder updated); finetuning did not systematically beat frozen (median difference ≈ 0), so *transfer* below pools the two.

Cross-task transfer. The full source×target census is Table 6 in the body: on ACOPF no pretraining source approaches from-scratch training (0.84–1.0 vs 0.59), while on SCUC every source modestly improves over the weak from-scratch encoder (0.19–0.29 vs 0.31) yet still trails the classical merit-order baseline (0.12). A random-init encoder anchors how much is learned: training reduces SCUC commitment Hamming ($0.51 \rightarrow 0.31$) but barely improves ACOPF over random init ($1.0 \rightarrow 0.59$).

F AI USE DISCLOSURE

In this work, we used generative AI tools for writing assistance, copyediting, and software engineering support. Synthetic data generation processes where an AI tool may have impacted related softare were validated extensively against existing public datasets, dataset comparisons are reported extensively throughout this work and were validated by mutiple members of the research team. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.