

---

# Metamorphic testing of learned AC optimal power flow surrogates reveals representation dependence

---

Stephen Peng\* Vineet Jagadeesan Nair\*

AI Science Institute, Boston Consulting Group

Peng.Stephen@bcg.com, JagadeesanNair.Vineet@bcg.com

## Abstract

Changing bus order or an angle reference leaves the AC optimal power flow (AC-OPF) problem unchanged, but can change a learned surrogate’s prediction. We test this dependence with thirteen metamorphic relations, each specifying when a transformation is valid and how to compare its outputs. Our screen covers 17 configurations: two released trained checkpoints and 15 random-weight architecture or mechanism controls. Every configuration fails at least six relations, including at tolerances ten times tighter or looser. For GridSFM-Open v1.1, source inspection and intermediate-tensor tests point to an indexing bug and three preprocessing choices; a repaired checkpoint remains untested. In a separate case500 experiment, augmentation reduces transformed-input error in three of four models and median discrepancy and constraint-violation P95 in all four. Most of the large Laplacian-model accuracy gain comes from a GPU/CPU eigensolver mismatch. Some relations worsen, and every evaluated prediction remains AC infeasible. Evaluating multiple encodings exposes failures that accuracy and feasibility scores on one encoding can miss.

## 1 Introduction

AC-OPF chooses generator dispatch to minimize cost while satisfying network equations and operating limits. Learned surrogates approximate this calculation and are now released as pretrained checkpoints for reuse across grids. Most evaluations measure accuracy, feasibility, and speed on held-out scenarios in a single serialization [1–3] (Sec. B.1). They leave untested whether another encoding of the same problem would give the same prediction.

A bus can move to a different row without changing the grid. Eligible branches can be reversed, and each island can use a different angle reference. Once predictions are returned to their physical devices and a common reference, a consistent deterministic surrogate should give the same answer. Graph layers may be permutation equivariant while positional encodings, derived features, directed edges, or index-based output heads break this property [4, 5]. We test the complete prediction pipeline using metamorphic relations: known relationships between the outputs for linked inputs [6, 7]. Each admitted transformation preserves the AC-OPF problem under stated conditions, which we check algebraically and with sampled solver controls (Sec. D.2). This differs from testing a load change or an outage, which changes the problem itself (Sec. B.2).

## 2 Metamorphic testing for AC-OPF

For a deterministic surrogate  $f$ , relation  $r$  transforms serialized case  $d$  into an equivalent input  $T_r d$ . Its output map  $P_r$  returns predictions to the original physical devices, aligns angles within each island,

---

\*Equal contribution.

and sums split-device outputs when needed. We measure departure from  $P_r f(T_r d) = f(d)$  by

$$\Delta_r(d) = \max_{c \in \mathcal{C}} \frac{\text{MAE}_c(P_r f(T_r d), f(d))}{\tau_c}, \quad \mathcal{C} = \{V_m, V_a, P_g, Q_g\}. \quad (1)$$

Here  $V_m, V_a$  denote voltage magnitude and angle, and  $P_g, Q_g$  denote active and reactive generation. Each channel’s mean absolute error (MAE) averages over its elements. We set  $\tau_c = 10^{-6}$  in native units: per unit for  $V_m, P_g, Q_g$  and radians for  $V_a$ . Angles are re-anchored without further modulo wrapping. A value of  $\Delta_r = 1$  reaches the reporting threshold; it is neither a percentage error nor an AC-feasibility limit.

**Screen verdicts.** A cell in the 17-configuration screen passes when every executed, applicable draw has  $\Delta_r \leq 1$ . We record partial coverage where controls are unavailable and check maximum component errors separately at  $10^{-5}$  (Sec. D.2).

**Relations.** Eight representation-equivalence (REP) relations change labels, orientation, angle reference, or a field used by a particular pipeline. Five electrical-equivalence (ELC) relations change component or graph bookkeeping while preserving the aggregate problem. Sec. A specifies each transformation, its validity conditions, and its output map, with a worked bus-relabel example. Non-uniqueness matters here: different predictions for equivalent inputs reveal representation dependence, but either could be a valid solution of the nonconvex problem (Secs. C.1 and C.2).

**Example.** If bus-relabel stores old buses as (3, 1, 2), restore voltage predictions as  $(\hat{V}'_2, \hat{V}'_3, \hat{V}'_1)$  and match generator outputs by device identity. If the reference bus also changes, subtract the restored reference angle within each island before comparing.

Table 1: The thirteen relations. The tiers distinguish how transformed outputs are mapped back; some relations require aggregation and are not bijective group actions.

| Tier | Relations                                 | What is held fixed / restored                               |
|------|-------------------------------------------|-------------------------------------------------------------|
| REP  | component-relabel, bus-relabel, row-order | Physical identities; inverse permutations                   |
| REP  | branch-reversal, tap-side-flip            | Branch two-port equations; endpoint/attribute transport     |
| REP  | ref-bus-move, island-ref-mix              | Per-island angle gauge after mapping back                   |
| REP  | basekv-scale                              | Exported per-unit target; conditional on schema semantics   |
| ELC  | component-split, parallel-split           | Aggregate injection, costs or admittance; sum split outputs |
| ELC  | load-as-gen, null-pad                     | Net bus injection; undo device recoding/padding             |
| ELC  | island-union                              | Original disconnected subproblem; restrict to its outputs   |

We test the adapter, preprocessing, network, and output map together, after checking that the relation applies. For example, basekv-scale changes model-visible metadata after targets have been exported in per unit; it does not rescale physical voltages. If a valid transformation falls outside a model’s documented input requirements, a failure shows an interface incompatibility rather than necessarily a defect in the intended use.

A random-weight counterexample rules out a guarantee of invariance for all weights; a pass only means that the sampled tests found none. We also compare trained and random-weight versions of GridSFM-Open v1.1 and LUMINA-2M v0.1.3 to see whether training changes the aggregate threshold classifications [8–12]. Scenarios come from the N-1 topology-perturbed OPFData release; solver controls use PGLib networks and PowerModels.jl [13–15]. Model access, sampling, and admission rules are given in Secs. D.1 and D.2.

### 3 Results

**Screen results.** Every configuration fails at least six of the thirteen relations (Fig. 1). Six failures are shared by all rows: basekv-scale, ref-bus-move, island-ref-mix, component-split, load-as-gen, and null-pad. The smallest discrepancy among these cells is  $37.5\times$  the tolerance, after rounding. Excluding basekv-scale leaves 140 of 204 cells failed at  $10^{-6}$  and at least five failures per configuration. Even at  $10^{-3}$ , every row fails at least four relations, or three without basekv-scale. These counts come from reaggregating saved results; they test sensitivity to the threshold, not the numerical noise floor (Sec. D.3).

Table 2: Case500 augmentation outcomes (baseline  $\rightarrow$  augmented; three seeds). Accuracy is equal-channel native-unit MAE; median  $\Delta_r$  uses each model’s trained relation set. Constraint entries are mean  $\pm$  sample SD of seed-level P95: first take each source’s worst violation over 53 views, then P95 across 50 sources. Equality is the larger maximum nodal active or reactive balance violation, in p.u. For pooled inequality, divide each family’s violation by its  $10^{-6}$  native-unit tolerance and report the maximum in millions of tolerances. Every arm has 0% AC-feasible sources. The LapPE row uses CPU spectral features and reflects the backend mismatch discussed below. Full definitions and family-level values appear in Secs. F.1 and F.3.

| Model      | Canonical MAE                 | Transformed MAE               | Median $\Delta_r$         |
|------------|-------------------------------|-------------------------------|---------------------------|
| MPNN       | 0.01295 $\rightarrow$ 0.01345 | 0.01312 $\rightarrow$ 0.01353 | 852 $\rightarrow$ 595     |
| LapPE+MPNN | 0.09983 $\rightarrow$ 0.01309 | 0.10211 $\rightarrow$ 0.01318 | 1,000 $\rightarrow$ 805   |
| RWSE+MPNN  | 0.01255 $\rightarrow$ 0.01250 | 0.01271 $\rightarrow$ 0.01257 | 883 $\rightarrow$ 563     |
| LUMINA     | 0.01834 $\rightarrow$ 0.01889 | 0.03594 $\rightarrow$ 0.02089 | 4,630 $\rightarrow$ 1,260 |

| Model      | Equality P95 (p.u.)                           | Pooled inequality P95 ( $10^6 \times$ tol.) |
|------------|-----------------------------------------------|---------------------------------------------|
| MPNN       | 58.1 $\pm$ 18.7 $\rightarrow$ 42.2 $\pm$ 2.3  | 19.3 $\pm$ 4.2 $\rightarrow$ 11.9 $\pm$ 1.6 |
| LapPE+MPNN | 113.9 $\pm$ 42.1 $\rightarrow$ 39.5 $\pm$ 5.2 | 42.1 $\pm$ 4.6 $\rightarrow$ 12.7 $\pm$ 3.6 |
| RWSE+MPNN  | 51.8 $\pm$ 10.9 $\rightarrow$ 33.5 $\pm$ 2.4  | 18.8 $\pm$ 2.8 $\rightarrow$ 10.2 $\pm$ 0.8 |
| LUMINA     | 871 $\pm$ 342 $\rightarrow$ 164 $\pm$ 11      | 276 $\pm$ 91 $\rightarrow$ 20.1 $\pm$ 2.1   |

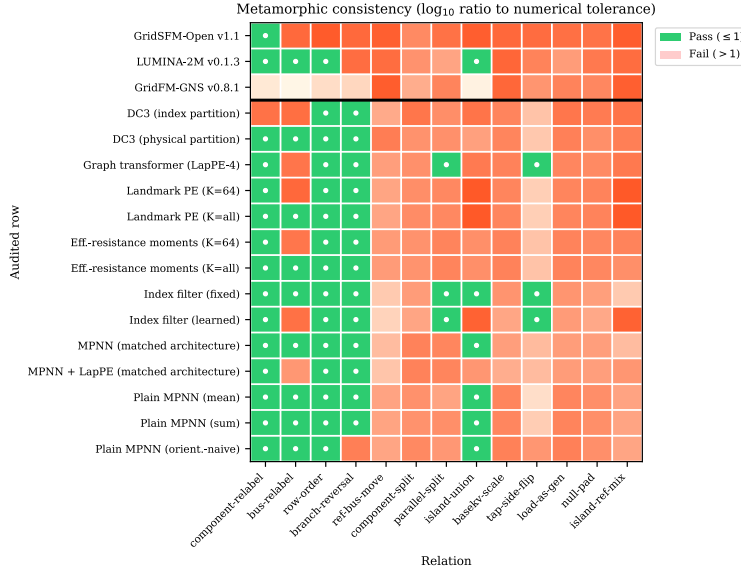


Figure 1: Metamorphic consistency for 17 model and pipeline configurations (two trained checkpoints and 15 random-weight architecture or mechanism controls) across 13 relations on case14/500. Each case uses 3 source scenarios, 3 sampled transformations per admitted case-relation pair, and 3 random initializations where applicable. Green denotes a pass at the prespecified  $10^{-6}$  tolerance.

Several matched comparisons identify specific design choices. Adding raw Laplacian positional encodings to a message passing neural network (MPNN) makes bus-relabel and island-union fail. Index-selected landmarks change physical buses when rows are reordered; the matched all-bus control removes this dependence, without demonstrating a repaired trained checkpoint. Orientation-naive edge processing breaks branch-reversal. Aggregation keyed to physical components and symmetric branch features pass the corresponding tests. The six shared failures show sensitivity to angle references, target-irrelevant inputs, and the division of a bus-level injection among device records (Sec. E.1).

**Trained and random weights.** Across five grids, both released backbones have the same 13 aggregate threshold classifications with trained and random weights, for each of three random seeds. This comparison takes the source median and then the worst grid; it does not require every draw to pass

(Sec. D.2). GridSFM is below threshold only on `component-relabel`. LUMINA is also below threshold on `bus-relabel`, `row-order`, and `island-union`. Training changes the discrepancy magnitudes but removes none of these aggregate failures. The ratios in Sec. E.2 measure this change in sensitivity, not model quality.

In GridSFM, the first detected differences occur in preprocessing. Source inspection points to index-selected landmarks, an incorrect transformer edge offset, unstable signs of near-zero DC flows, and a cycle basis that depends on edge-row order. Substituting aligned intermediate tensors removes much of the discrepancy in one diagnostic case, although a network-only relabeling test retains a nonzero residual. These tests identify candidate repairs; they do not validate a repaired checkpoint or establish network equivariance (Sec. E.3; see Sec. E.4 for cost sensitivity). LUMINA’s `branch-reversal` failure is consistent with directed message passing that does not compensate for reversed branch attributes.

**Augmentation.** We train an MPNN, the same MPNN with raw Laplacian positional encodings (LapPE+MPNN) [16, 17], an MPNN with random-walk structural encodings (RWSE+MPNN) [17, 18], and LUMINA. RWSE has no eigenvector sign or basis ambiguity. All use equal-channel MSE against the solver solution plus AC constraint penalties. The augmented arm transforms each training instance with probability 0.5, using a fixed model-specific relation set selected from earlier trained-baseline measurements, separately from the random-weight screen. Evaluation uses 50 held-out scenarios per seed, in canonical form and under all thirteen relations. Accuracy averages native-unit MAEs equally across  $V_m, V_a, P_g, Q_g$  (Sec. F.1).

On transformed inputs, augmentation reduces MAE by 87.0% for LapPE+MPNN, 41.9% for LUMINA, and 1.1% for RWSE+MPNN, but increases it by 3.2% for MPNN (Table 2). Most of the LapPE gain is explained by using a CPU eigensolver for this audit after training with a GPU eigensolver; the same mismatch may inflate its baseline violation P95 (Sec. F.4). Canonical MAE improves by 0.4% for RWSE+MPNN and worsens by 3.9% for MPNN and 3.0% for LUMINA.

Median  $\Delta_r$  falls for all four models, but remains 563 to 1,260 times the threshold. Some relations worsen, and the gains do not transfer consistently to other grids. Equality and pooled inequality P95 also fall for all four models under the same training loss, but every evaluated prediction remains AC infeasible (Secs. F.2 and F.3).

We have not isolated why augmentation reduces constraint violations, or whether preserving the physical problem during transformation causes that reduction. Lower discrepancies and violation P95 values do not by themselves show better physical understanding or fitness for operation. Code and documentation are available online.

## 4 Implications and limitations

Some failures can be prevented by design: aggregation keyed to physical components, permutation-safe features, symmetric handling of branch attributes, and a common angle reference can enforce specific relations (Sec. G.1). Augmentation offers an approximate remedy, but its effects vary by relation, and group-averaging theory applies to only part of the catalog [19–21] (Sec. C.3). Our catalog is incomplete, and the trained-checkpoint audit covers only two compatible releases. The conclusions are limited to the tested checkpoints and procedures.

The full catalog matters because models fail different relations: LUMINA’s largest baseline discrepancies are on `branch-reversal` and `basekv-scale`, while LapPE+MPNN’s are on `bus-relabel`, `island-union`, `island-ref-mix`, and `parallel-split` (Sec. F.2).

Developers can use these tests before release; operators and planners can apply them through documented inference access. For each held-out scenario, check which relations apply, generate  $K$  valid encodings, map predictions back, and inspect the worst discrepancy alongside channel accuracy, feasibility, optimality, and dispatch cost. Keep failed admissions and infeasible predictions, and treat the views as repeated measurements of one source (Sec. G.2 and Table 13). Failures guide debugging; a finite pass does not certify operational use or architectural invariance. The latter still needs an analytical argument, with random-weight screens used to find counterexamples (Sec. G.3).

## References

- [1] Luis Piloto, Sofia Liguori, Sephora Madjiheurem, Miha Zgubic, Sean Lovett, Hamish Tomlinson, Sophie Elster, Chris Apps, and Sims Witherspoon. CANOS: A Fast and Scalable Neural AC-OPF Solver Robust To N-1 Perturbations, March 2024. URL <http://arxiv.org/abs/2403.17660>. arXiv:2403.17660 [cs.LG].
- [2] Shaohui Liu, Chengyang Wu, and Hao Zhu. Topology-Aware Graph Neural Networks for Learning Feasible and Adaptive AC-OPF Solutions. *IEEE Transactions on Power Systems*, 38(6):5660–5670, November 2023. ISSN 1558-0679. doi: 10.1109/TPWRS.2022.3230555. URL <https://ieeexplore.ieee.org/document/9992121>.
- [3] Olayiwola Arowolo and Jochen L. Cremer. Towards generalization of graph neural networks for AC optimal power flow. *Energy and AI*, 25:100842, September 2026. ISSN 2666-5468. doi: 10.1016/j.egyai.2026.100842. URL <https://www.sciencedirect.com/science/article/pii/S2666546826001680>.
- [4] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks, October 2018. URL <http://arxiv.org/abs/1806.01261>. arXiv:1806.01261 [cs.LG].
- [5] Haggai Maron, Heli Ben-Hamu, Nadav Shmair, and Yaron Lipman. Invariant and Equivariant Graph Networks. September 2018. URL <https://openreview.net/forum?id=Syx72jC9tm>.
- [6] Tsong Yueh Chen, Fei-Ching Kuo, Huai Liu, Pak-Lok Poon, Dave Towey, T. H. Tse, and Zhi Quan Zhou. Metamorphic Testing. *ACM Computing Surveys*, 51(1):1–27, January 2018. doi: 10.1145/3143561. URL <https://dl.acm.org/doi/abs/10.1145/3143561>.
- [7] Sergio Segura, Gordon Fraser, Ana B. Sanchez, and Antonio Ruiz-Cortés. A Survey on Metamorphic Testing. *IEEE Transactions on Software Engineering*, 42(9):805–824, September 2016. ISSN 1939-3520. doi: 10.1109/TSE.2016.2532875. URL <https://ieeexplore.ieee.org/abstract/document/7422146>.
- [8] Weiwei Yang, Andrea Britto Mattos Lima, Thiago V. Spina, Spencer Fowers, Baosen Zhang, and Chris White. GridSFM: A Foundation Model for AC Optimal Power Flow, May 2026. URL [https://www.microsoft.com/en-us/research/wp-content/uploads/2026/05/GridFM\\_white\\_paper.pdf](https://www.microsoft.com/en-us/research/wp-content/uploads/2026/05/GridFM_white_paper.pdf).
- [9] Microsoft Corporation. GridSFM-Open model card, 2026. URL [https://huggingface.co/microsoft/GridSFM\\_Open/tree/v1.1](https://huggingface.co/microsoft/GridSFM_Open/tree/v1.1). Hugging Face, checkpoint gridsfm\_open\_v1.1.pt.
- [10] Microsoft Corporation. GridSFM software repository, 2026. URL <https://github.com/microsoft/GridSFM>. GitHub, audited at commit 1ca775f.
- [11] Yijiang Li, Zeeshan Memon, Hongwei Jin, Stefano Fenu, Keunju Song, Sunash B. Sharma, Parfait Gasana, Hongseok Kim, Liang Zhao, and Kibaek Kim. LUMINA: Foundation models for topology transferable ACOPF. In *ICLR 2026 workshop on foundation models for science: Real-world impact and science-first design*, 2026. doi: 10.48550/arXiv.2603.04300. URL <https://doi.org/10.48550/arXiv.2603.04300>.
- [12] Argonne National Laboratory. LUMINA-2M model card, 2026. URL <https://huggingface.co/argonne/LUMINA-2M/tree/v0.1.3>. Hugging Face, version v0.1.3.
- [13] Sean Lovett, Miha Zgubic, Sofia Liguori, Sephora Madjiheurem, Hamish Tomlinson, Sophie Elster, Chris Apps, Sims Witherspoon, and Luis Piloto. OPFData: Large-scale datasets for AC optimal power flow with topological perturbations, June 2024. URL <https://arxiv.org/abs/2406.07234>. arXiv:2406.07234.

- [14] Sogol Babaeinejadsarookolae, Adam Birchfield, Richard D. Christie, Carleton Coffrin, Christopher DeMarco, Ruisheng Diao, Michael Ferris, Stephane Fliscounakis, Scott Greene, Renke Huang, Cedric Jozs, Roman Korab, Bernard Lesieutre, Jean Maeght, Terrence W. K. Mak, Daniel K. Molzahn, Thomas J. Overbye, Patrick Panciatici, Byungkwon Park, Jonathan Snodgrass, Ahmad Tbaileh, Pascal Van Hentenryck, and Ray Zimmerman. The Power Grid Library for Benchmarking AC Optimal Power Flow Algorithms, January 2021. URL <http://arxiv.org/abs/1908.02788>. arXiv:1908.02788 [math.OC].
- [15] Carleton Coffrin, Russell Bent, Kaarthik Sundar, Yeesian Ng, and Miles Lubin. PowerModels.JL: An Open-Source Framework for Exploring Power Flow Formulations. pages 1–8. IEEE, June 2018. doi: 10.23919/PSCC.2018.8442948. URL <https://doi.org/10.23919/pssc.2018.8442948>.
- [16] Vijay Prakash Dwivedi, Chaitanya K. Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking Graph Neural Networks. *Journal of Machine Learning Research*, 24(43):1–48, 2023. URL <http://jmlr.org/papers/v24/22-0567.html>.
- [17] Ladislav Rampásek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a General, Powerful, Scalable Graph Transformer. In *Advances in Neural Information Processing Systems*, volume 35, pages 14501–14515. Curran Associates, Inc., 2022. doi: 10.52202/068431-1054. URL [https://proceedings.neurips.cc/paper\\_files/paper/2022/hash/5d4834a159f1547b267a05a4e2b7cf5e-Abstract-Conference.html](https://proceedings.neurips.cc/paper_files/paper/2022/hash/5d4834a159f1547b267a05a4e2b7cf5e-Abstract-Conference.html).
- [18] Vijay Prakash Dwivedi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Graph Neural Networks with Learnable Structural and Positional Representations. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=wTTjnvGphYj>.
- [19] Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A Group-Theoretic Framework for Data Augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020. ISSN 1533-7928. URL <http://jmlr.org/papers/v21/20-163.html>.
- [20] Bryn Elesedy and Sheheryar Zaidi. Provably Strict Generalisation Benefit for Equivariant Models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2959–2969. PMLR, July 2021. URL <https://proceedings.mlr.press/v139/elesedy21a.html>.
- [21] Oskar Nordenfors, Fredrik Ohlsson, and Axel Flinth. Optimization Dynamics of Equivariant and Augmented Neural Networks. *Transactions on Machine Learning Research*, September 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=PTTa3U29NR>.
- [22] Rahul Nellikkath and Spyros Chatzivasileiadis. Physics-Informed Neural Networks for AC Optimal Power Flow. *Electric Power Systems Research*, 212:108412, November 2022. ISSN 0378-7796. doi: 10.1016/j.epsr.2022.108412. URL <https://www.sciencedirect.com/science/article/pii/S0378779622005636>.
- [23] Priya L. Donti, David Rolnick, and J. Zico Kolter. DC3: a learning method for optimization with hard constraints. In *International conference on learning representations*, 2021. URL <https://arxiv.org/abs/2104.12225>.
- [24] Damian Owerko, Fernando Gama, and Alejandro Ribeiro. Optimal Power Flow Using Graph Neural Networks. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5930–5934, May 2020. doi: 10.1109/ICASSP40776.2020.9053140. URL <https://ieeexplore.ieee.org/abstract/document/9053140>. ISSN: 2379-190X.
- [25] Salah Ghamizi, Aoxiang Ma, Jun Cao, and Pedro Rodriguez Cortes. OPF-HGNN: Generalizable Heterogeneous Graph Neural Networks for AC Optimal Power Flow. In *2024 IEEE Power & Energy Society General Meeting (PESGM)*, pages 1–5, July 2024. doi: 10.1109/PESGM51994.2024.10688560. URL <https://ieeexplore.ieee.org/abstract/document/10688560>. ISSN: 1944-9933.

- [26] Aihui Wen, Bao Wen, Jining Li, and Jin Xu. Heterogeneous Graph Neural Network with Local and Global Message Passing for AC-Optimal Power Flow Solutions. *Applied System Innovation*, 9(1):18, January 2026. ISSN 2571-5577. doi: 10.3390/asi9010018. URL <https://www.mdpi.com/2571-5577/9/1/18>.
- [27] Helge Spieker, Nassim Belmecheri, Arnaud Gotlieb, and Nadjib Lazaar. Evaluating human trajectory prediction with metamorphic testing. In *Proceedings of the 9th ACM international workshop on metamorphic testing*, pages 34–40, 2024. doi: 10.1145/3679006.3685071.
- [28] Andreas Venzke, Guannan Qu, Steven Low, and Spyros Chatzivasileiadis. Learning optimal power flow: Worst-case guarantees for neural networks, 2020. URL <https://arxiv.org/abs/2006.11029>.
- [29] Derek Lim, Joshua Robinson, Lingxiao Zhao, Tess Smidt, Suvrit Sra, Haggai Maron, and Stefanie Jegelka. Sign and basis invariant networks for spectral graph representation learning. In *International Conference on Learning Representations*, 2023. URL <https://arxiv.org/abs/2202.13013>.
- [30] Meng Li, Xiaohua Yang, Jie Liu, and Shiyu Yan. Domain-validity-gated metamorphic testing of scientific ML surrogates, 2026. URL <https://arxiv.org/abs/2606.17529>.
- [31] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models’ sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The twelfth international conference on learning representations*, 2024. URL <https://openreview.net/forum?id=RIu5lyNXjT>.
- [32] Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. doi: 10.1162/tac1\_a\_00681.
- [33] Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 10776–10787, 2023. doi: 10.18653/v1/2023.findings-emnlp.722.
- [34] Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel Kochenderfer. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *Advances in neural information processing systems 37*, pages 21763–21813, 2024. doi: 10.52202/079017-0685.
- [35] Maria Eriksson, Erasmo Purificato, Arman Noroozian, João Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. Can we trust AI benchmarks? An interdisciplinary review of current issues in AI evaluation. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society*, volume 8, pages 850–864, 2025. doi: 10.1609/aies.v8i1.36595.
- [36] Daniel Bienstock and Abhinav Verma. Strong NP-hardness of AC power flows feasibility. *Operations Research Letters*, 47(6):494–501, November 2019. ISSN 0167-6377. doi: 10.1016/j.orl.2019.08.009. URL <https://doi.org/10.1016/j.orl.2019.08.009>.
- [37] Waqqas A. Bukhsh, Andreas Grothey, Ken I. M. McKinnon, and Paul A. Trodden. Local Solutions of the Optimal Power Flow Problem. *IEEE Transactions on Power Systems*, 28(4): 4780–4788, November 2013. ISSN 1558-0679. doi: 10.1109/TPWRS.2013.2274577. URL <https://ieeexplore.ieee.org/abstract/document/6581918>.
- [38] Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks. *arXiv preprint arXiv:2005.00178*, 2020. URL <https://arxiv.org/abs/2005.00178>.
- [39] GridFM Project. gridfm-graphkit Release v0.8.1, 2026. URL <https://github.com/gridfm/gridfm-graphkit/releases/tag/v0.8.1>. GitHub release, tag v0.8.1, commit b3d663b.
- [40] Andreas Wächter and Lorenz T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, April 2005. ISSN 0025-5610. doi: 10.1007/s10107-004-0559-y. URL <https://doi.org/10.1007/s10107-004-0559-y>.

## A Relation catalog and output maps

Table 3 lists the transformations, the conditions under which they preserve AC-OPF, and the maps used to compare predictions. Figs. 2 and 3 give examples. REP relations change labels or reference conventions; ELC relations change graph or device bookkeeping and may require summing outputs or undoing a recoding.

**Example: bus relabeling.** Suppose the transformed bus array stores old buses in the order  $(3, 1, 2)$ . To restore voltage predictions to the original order, take  $(\hat{V}'_2, \hat{V}'_3, \hat{V}'_1)$ . Match generator outputs by physical device identity. If the reference bus also changes, subtract the restored reference angle from every angle in that island. Any remaining difference is representation dependence. It need not imply infeasibility: nonconvex AC-OPF can have several physically valid solutions.

Table 3: Catalog of the 13 metamorphic relations. For each relation,  $T_r$  constructs an equivalent input and  $P_r$  aligns or aggregates the transformed prediction with the original physical quantities; the validity condition states when that equivalence holds.

| Relation          | Tier | Transformation $T_r$                                                                                                                                                                                                                                                                                                                                        | Validity condition                                                                                                        | Output map $P_r$                                                                                                |
|-------------------|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| component-relabel | REP  | Reorder the generator array (or load, or shunt) by a permutation $\sigma$ , moving its bus-link rows with it; buses and branches untouched.                                                                                                                                                                                                                 | always (every $\sigma$ )                                                                                                  | apply $\sigma^{-1}$ to the per-component outputs                                                                |
| bus-relabel       | REP  | Reorder the bus array by a permutation $\pi$ ; rewrite every stored bus index through $\pi$ (branch from/to endpoints and all component-to-bus links).                                                                                                                                                                                                      | always (every $\pi$ )                                                                                                     | apply $\pi^{-1}$ to the per-bus outputs                                                                         |
| row-order         | REP  | Reorder the rows of one edge family ( <code>ac_line</code> , transformer, or a component-link table); every field value and endpoint unchanged (a pure serialisation reshuffle).                                                                                                                                                                            | always                                                                                                                    | identity on physical quantities; re-sort through <code>canonical_link_order</code> before positional conversion |
| branch-reversal   | REP  | Reverse a branch: swap its from/to endpoints, swap the shunt halves $b_{fr} \leftrightarrow b_{to}$ , and replace $(\text{angmin}, \text{angmax})$ with $(-\text{angmax}, -\text{angmin})$ .                                                                                                                                                                | per row: $\tau=1$ , shift $=0$ , $g_{fr}=g_{to}=0$ (all <code>ac_line</code> rows; only untapped, unshifted transformers) | reverse that branch's directed-flow channels, $(P_f, Q_f, P_t, Q_t) \rightarrow (P_t, Q_t, P_f, Q_f)$           |
| ref-bus-move      | REP  | Move the angle reference: set a chosen generator bus's type to reference and demote the former reference to PV; no other field changes.                                                                                                                                                                                                                     | the chosen bus is a generator bus in the same island (one reference per island)                                           | identity on components; re-anchor both solutions' angles to a common physical bus                               |
| basekv-scale      | REP  | Multiply the model-visible <code>base_kv</code> field at every bus by a common factor $\alpha > 0$ after OPFData export.                                                                                                                                                                                                                                    | pipeline-specific: the standard per-unit AC-OPF target is unchanged                                                       | identity (solver quantities are unchanged)                                                                      |
| tap-side-flip     | REP  | Swap a transformer's from/to endpoints. For complex tap $\tau = te^{j\phi}$ , set $\tau' = \tau^{-1}$ , $z' = t^2 z$ , $(b'_{fr}, b'_{to}) = (b_{to}/t^2, b_{fr}/t^2)$ , and $(\theta'_{\min}, \theta'_{\max}) = (-\theta_{\max}, -\theta_{\min})$ .                                                                                                        | $t > 0$ ; $Y' = \Pi Y \Pi$ for the endpoint-swap matrix $\Pi$                                                             | reverse directed-flow channels, $(P_f, Q_f, P_t, Q_t) \rightarrow (P_t, Q_t, P_f, Q_f)$                         |
| island-ref-mix    | REP  | Multi-island only: permute which island serialises first and choose each island's reference independently (bus relabeling on island order composed with per-island ref-bus-move).                                                                                                                                                                           | multi-island instance; one valid generator reference per island                                                           | re-anchor each island's angles, then undo the island/bus permutation                                            |
| component-split   | ELC  | Split load $\ell$ into co-located loads $\ell_1, \ell_2$ satisfying $(P_{\ell_1}^d + P_{\ell_2}^d, Q_{\ell_1}^d + Q_{\ell_2}^d) = (P_{\ell}^d, Q_{\ell}^d)$ ; or split generator $g$ into two with halved $(P_g^{\min}, P_g^{\max}, Q_g^{\min}, Q_g^{\max})$ bounds and cost $(c_{g,2}, c_{g,1}, c_{g,0}) \rightarrow (2c_{g,2}, c_{g,1}, c_{g,0}/2)$ each. | loads: always; generators: convex quadratic cost and symmetric halved bounds                                              | sum the two parts' injections back (aggregation, not a permutation)                                             |
| parallel-split    | ELC  | Replace one branch by two identical parallel branches: $r, x \rightarrow 2r, 2x$ (parallel admittance preserved), $b_{fr}, b_{to} \rightarrow b_{fr}/2, b_{to}/2$ , rate $\rightarrow \text{rate}/2$ ; angle limits kept.                                                                                                                                   | always (admittances recombine; flow splits evenly)                                                                        | sum the two branches' flows                                                                                     |

Continued on next page

Table 3 (continued)

| Relation     | Tier | Transformation $T_r$                                                                                                                                                                                                                             | Validity condition                                                     | Output map $P_r$                                                                           |
|--------------|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|
| island-union | ELC  | Concatenate two solved scenarios into one instance with two disconnected islands, renumbering the second scenario's buses and components after the first.                                                                                        | always (no coupling; optimum and cost are the direct sum)              | split the joint prediction into the two island blocks                                      |
| load-as-gen  | ELC  | Recast a fixed load $\ell$ with demand $(P_\ell^d, Q_\ell^d)$ as a zero-cost generator $g'$ at the same bus with $P_{g'}^{\min}=P_{g'}^{\max}=-P_\ell^d$ and $Q_{g'}^{\min}=Q_{g'}^{\max}=-Q_\ell^d$ , then remove $\ell$ from the load table.   | always (a clamped zero-cost generator imposes the identical injection) | read $(P_{g'}, Q_{g'})$ and map it back to the load (must equal $(-P_\ell^d, -Q_\ell^d)$ ) |
| null-pad     | ELC  | Add an inert element: a load $\ell'$ with $P_{\ell'}^d=Q_{\ell'}^d=0$ , a shunt with $G^{\text{sh}}=B^{\text{sh}}=0$ , or a zero-cost generator $g'$ with $P_{g'}^{\min}=P_{g'}^{\max}=Q_{g'}^{\min}=Q_{g'}^{\max}=0$ ; bus types are unchanged. | always (the element injects nothing; only counts change)               | drop the padded element from the prediction                                                |

## Representation and gauge relations: the physical grid is unchanged

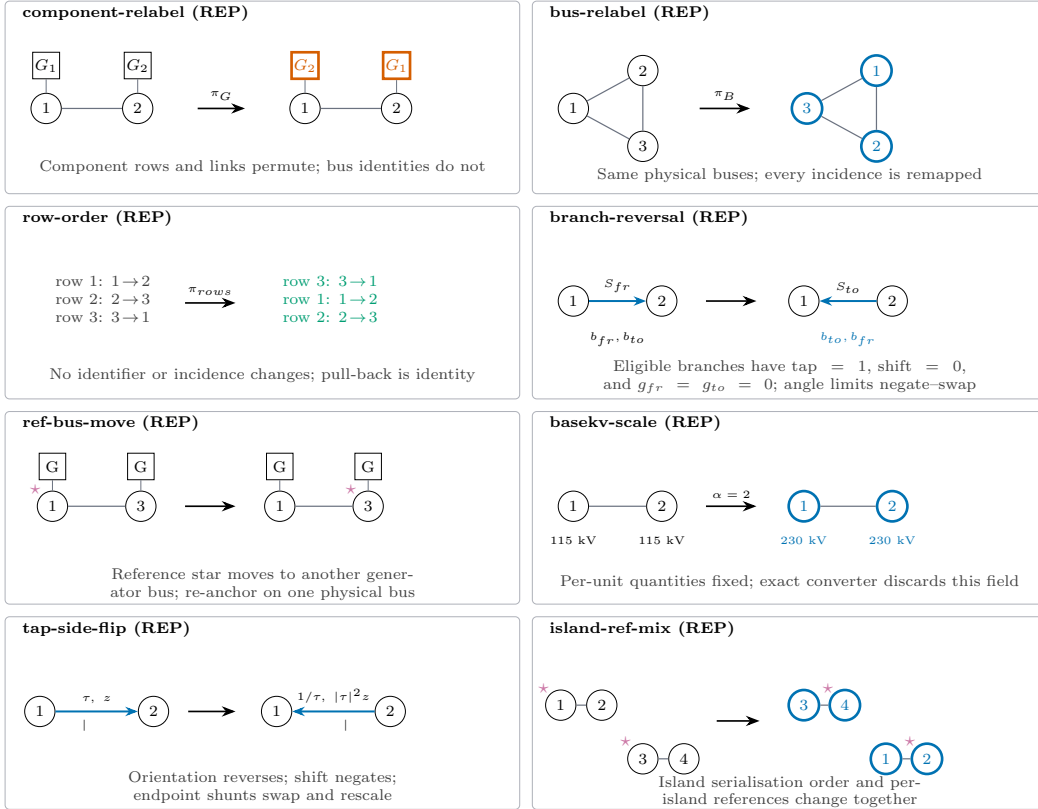


Figure 2: Examples of REP relations. They change labels, row order, branch orientation, or angle reference without changing the underlying AC-OPF problem, and we compare predictions after undoing the declared map. The pipeline-specific `basekv-scale` relation changes an exported model input that is absent from the standard per-unit AC-OPF target.

## B Related work

### B.1 Learned AC-OPF surrogates

Learned AC-OPF models map demands, generator limits, costs, and network parameters to voltages and dispatch. They may replace an optimizer, provide its starting point, or work with a correction

### Electrical-equivalence relations: graph structure changes, but the OPF does not

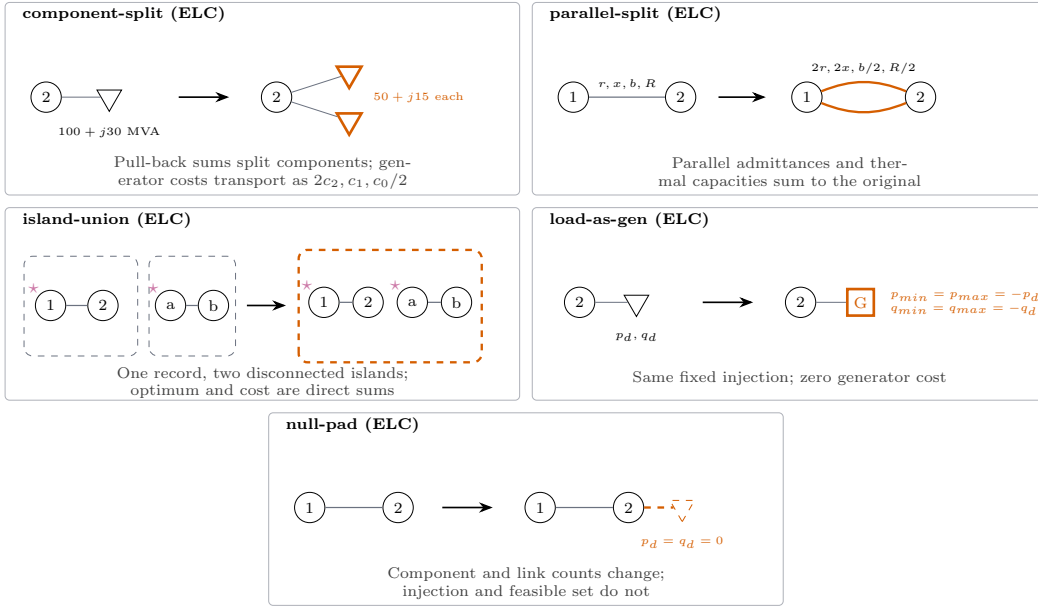


Figure 3: Examples of ELC relations. They change graph size or component-family bookkeeping while preserving the AC-OPF feasible set and objective under the stated validity conditions. Their output maps sum split quantities, separate joined islands, reassign a fixed injection, or remove inert padding, depending on the relation.

step. Fast predictions with low label error can still violate power balance or operating limits and can have high dispatch cost. Physics-informed losses and differentiable completion or correction layers address some of these problems [22, 23].

Early surrogates often used fully connected networks with a fixed topology and variable order. Graph neural networks use bus-branch connectivity and shared message-passing rules, allowing permutation-equivariant computation [4, 5]. Homogeneous and heterogeneous AC-OPF models use this structure [1, 2, 24, 25]. Distinguishing buses, generators, loads, shunts, lines, and transformers preserves useful physical information, but also requires consistent handling of each family’s order, identity, and aggregation.

Local message passing has difficulty representing long-range electrical coupling. Global attention, graph-transformer blocks, and positional encodings provide ways to extend its reach [3, 17, 26]. Laplacian eigenvectors, effective resistance, random-walk statistics, and landmark summaries supply global information, but some also depend on arbitrary indices, eigenvector signs or bases, reference nodes, or row order. These features can break the equivariance of an otherwise equivariant network.

Our screen includes mean- and sum-aggregating MPNNs, DC3-inspired output partitions, graph transformers with spectral or effective-resistance features, index-selected landmarks, learned index filters, and public models that accept varying grid topologies. When no checkpoint is available, random-weight tests can expose failures of built-in equivariance but cannot tell us what training would learn. GridSFM-Open v1.1 and LUMINA-2M v0.1.3 provide both trained weights and runnable inference, so we also test their released checkpoints [8–12].

## B.2 Metamorphic testing and equivariance

Metamorphic testing checks known relationships between outputs for related inputs when the correct output is unavailable or costly to obtain [6, 7]. A violated necessary relationship can reveal a defect without a unique reference label. Scientific software and machine-learning applications often obtain such relationships from physical symmetries [27].

The relations here have different mathematical forms. A bus permutation requires the same permutation of bus outputs; after mapping them back, predictions should be unchanged. Changing the reference angle shifts absolute angles while preserving angle differences, flows, and dispatch. Component and branch splits may change output dimension and require many-to-one aggregation. We reserve *built-in equivariance* for a property that holds at every parameter setting. Consistency observed in a trained checkpoint is an empirical result over the tested inputs.

Learned-OPF evaluations usually measure accuracy, feasibility, speed, and generalization to changed loads or contingencies [1–3]. Those tests change the physical problem; our admitted relations preserve it after a known output map. In the literature we reviewed, we found no AC-OPF study that combined a relation catalog with solver controls, trained-checkpoint tests, failure localization, and an augmentation experiment. Our contribution is this AC-OPF construction and audit, building on established metamorphic-testing and graph-equivariance methods.

**Relation to formal guarantees and invariant spectral features.** Worst-case verification asks whether a predictor meets a constraint or quality bound throughout an input domain; Venzke et al. [28], for example, study neural DC-OPF guarantees. Our tests find counterexamples on sampled equivalent AC-OPF inputs and cannot certify a whole domain. The raw Laplacian coordinates in our baseline have known sign and eigenspace-basis ambiguities. SignNet and BasisNet address these symmetries explicitly [29]; RWSE avoids these ambiguities through a different construction. We did not train SignNet or BasisNet, so our LapPE results do not apply to spectral graph models in general.

**Related scientific-surrogate metamorphic testing.** Li et al. [30] use metamorphic tests for scientific ML surrogates with explicit validity conditions and numerical thresholds. We share their requirement to check that a relation applies before interpreting a failure. Here we develop the AC-OPF relations and output maps, distinguish trained-checkpoint tests from random-weight controls, and test complete grid-model pipelines. Our experiments cover power systems only.

### B.3 Evaluation validity beyond power systems

Evaluation choices can affect conclusions in other domains as well. Meaning-preserving prompt-format changes can shift measured accuracy by tens of points [31]; model rankings from one prompt may not hold across prompts [32]; and undetected data contamination can inflate scores [33]. Reviews also find that many benchmarks leave construct validity and reporting practices unexamined [34, 35].

A heuristic perturbation may change the task it is meant to test. AC-OPF gives us explicit equations with which to check that a transformation preserves the problem, and sampled solver controls check the implementation. The same approach could be used in another domain where task-preserving transformations can be verified.

## C Technical foundation

### C.1 AC optimal power flow and non-uniqueness

Let  $\mathcal{N}$ ,  $\mathcal{E}$ , and  $\mathcal{G}$  denote buses, branches, and generators. For bus  $i$ , the AC-OPF variables include voltage magnitude  $V_i$ , voltage angle  $\theta_i$ , and active and reactive generation  $(P_g, Q_g)$ ; fixed demands are  $(P_i^d, Q_i^d)$ . With commitment fixed, a common economic-dispatch objective is

$$\min_{V, \theta, P, Q} \sum_{g \in \mathcal{G}} c_{g,2} P_g^2 + c_{g,1} P_g + c_{g,0}. \quad (2)$$

Writing  $Y = G + jB$  and  $\theta_{ij} = \theta_i - \theta_j$ , active and reactive nodal balance require

$$\sum_{g \in \mathcal{G}_i} P_g - P_i^d = V_i \sum_{j \in \mathcal{N}} V_j (G_{ij} \cos \theta_{ij} + B_{ij} \sin \theta_{ij}), \quad (3)$$

$$\sum_{g \in \mathcal{G}_i} Q_g - Q_i^d = V_i \sum_{j \in \mathcal{N}} V_j (G_{ij} \sin \theta_{ij} - B_{ij} \cos \theta_{ij}). \quad (4)$$

Generator outputs and voltage magnitudes obey bounds; branch flows obey apparent-power and angle-difference limits. One reference angle per connected island removes the additive angle gauge.

Voltage products and trigonometric angle differences make the feasible set nonconvex. AC feasibility is strongly NP-hard in general, and multiple local optima can occur [36, 37]. Production cost may also leave reactive dispatch and voltage magnitude weakly identified. Our test checks whether one deterministic surrogate is consistent across equivalent inputs and does not assume a unique AC-OPF optimum.

## C.2 Fibers of the data-to-problem map

Serialized cases include choices that the physical optimization does not depend on, such as bus names, array order, stored branch direction, and angle reference. The problem itself is specified by aggregate bus injections, series and shunt admittances, transformer ratios, connectivity, bounds, and cost coefficients. Let  $s : \mathcal{D} \rightarrow \mathcal{Z}$  map serialized data to these sufficient quantities. A *fiber* of  $s$  contains the encodings with the same sufficient quantities. Valid relations stay within such a fiber, allowing for label permutations or changes of angle reference. After mapping outputs back, consistency requires  $P_r f(T_r d) = f(d)$ , or  $\Delta_r(d) = 0$  in (1).

Table 4: Schema distinctions discarded by the data-to-problem map  $s$ , with the relations that test each resulting many-to-one collapse.

| Collapse in $s$                                                                    | Relations                                                                                                           |
|------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|
| Incidence is label-free                                                            | <code>component-relabel</code> , <code>bus-relabel</code> ,<br><code>row-order</code> , <code>island-ref-mix</code> |
| Only the aggregate branch model enters the equations                               | <code>branch-reversal</code> , <code>tap-side-flip</code>                                                           |
| Bus balance sees aggregate injection instead of per-device                         | <code>component-split</code> , <code>load-as-gen</code> ,<br><code>null-pad</code>                                  |
| Parallel branches collapse to one effective admittance                             | <code>parallel-split</code>                                                                                         |
| Angle reference is a gauge convention                                              | <code>ref-bus-move</code> , <code>island-ref-mix</code>                                                             |
| The post-export per-unit target does not use model-visible<br><code>base_kv</code> | <code>basekv-scale</code>                                                                                           |
| Disconnected islands cannot interact                                               | <code>island-union</code>                                                                                           |

Table 4 lists serialization choices that leave the optimization problem unchanged. Our catalog tests one transformation for each listed fiber, with the coverage limits described in Sec. C.2.

Unequal or  $N$ -way component splits and arbitrary shunt redistribution would extend the discrete tests to continuous families. A sufficiently small relaxation of a strictly inactive bound could also preserve a locally stable solution, although it need not preserve the global OPF problem.

Most fibers are continuous, so one discrete test covers only part of each. For example, `branch-reversal` swaps endpoints without testing arbitrary redistribution of  $b_{fr} + b_{to}$ , and `component-split` uses equal halves rather than an arbitrary  $N$ -way split. We have not proved that composing the 13 relations reaches every equivalent encoding.

## C.3 When augmentation theory applies

For squared loss, orbit averaging projects a predictor onto the equivariant function class when the transformations form a compact group, the target is equivariant, and the input distribution is invariant or balanced across each orbit. Under these assumptions, the non-equivariant part contributes avoidable risk [19, 20, 38]. Sampling from the normalized Haar measure gives an unbiased stochastic estimate of this objective.

These assumptions hold for only part of our catalog. Finite relabelings and eligible involutions fit the group setting; `basekv-scale` is unbounded and pipeline-specific, reference changes are conditional gauge actions, and ELC relations can change input and output dimensions. Our model-specific 0.5 augmentation mixture does not sample a single group’s Haar measure. The theorem also concerns orbit-balanced risk, not error on one canonical encoding, and nonconvex training need not reach the projection [21]. We therefore measure canonical error, transformed error, discrepancy, and feasibility separately.

## D Audit design and model inventory

Fig. 4 summarizes the audit workflow.

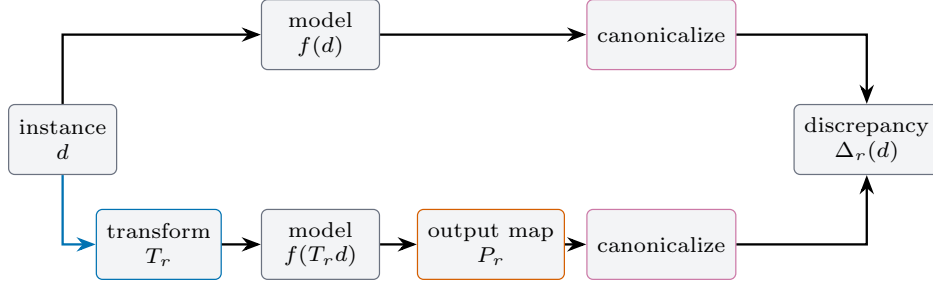


Figure 4: Metamorphic audit workflow. The audit compares canonical and transformed predictions after applying the declared output map and gauge canonicalization. A numerical solver-control suite samples both paths to check the implementation of valid relations.

### D.1 Model access and scope of the tests

The 17 rows in Table 5 comprise two trained checkpoints, one public AC-OPF architecture at random initialization, and fourteen randomly initialized reimplementations or controls.

We selected GridSFM-Open and LUMINA-2M because each accepts varying grid topologies and provides released weights and runnable inference [8–12]. We could also run the GridFM-GNS architecture, but found no compatible public AC-OPF checkpoint for it as of August 25, 2026 [39].

The reimplementations test five design choices: index-based versus electrical-role-based output partitions; raw Laplacian coordinates; summaries over index-selected landmarks versus all buses; learned per-index versus graph-fixed filters; and directed versus orientation-symmetric branch processing. Where practical, matched pairs use the same surrounding network and random seed, changing only the component of interest. They test that component’s effect, not the accuracy of a published trained model.

We also include two MPNN rows that connect the architecture screen to the augmentation study. They are randomly initialized instances of the study architectures, each with four message-passing layers, width 128, sum aggregation, orientation-invariant branch features, and masked solution inputs; the second row adds 16 raw Laplacian eigenvectors. The three plain MPNNs provide smaller family controls for mean versus sum aggregation and for branch orientation.

A random-weight failure rules out guaranteed invariance for the tested pipeline. It tells us nothing about its eventual trained accuracy, feasibility, or discrepancy magnitude. The DC3 rows test the partition-and-completion choice with a stand-in; they do not run DC3’s Newton power-flow completion or correction loop.

Table 5: Inventory of the 17 model, reimplementation, and control rows in the built-in consistency screen. “Scope” names the released pipeline or design mechanism behind each row.

| Row                         | Class                   | Scope                                                                                                                                                          |
|-----------------------------|-------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| GridSFM-Open v1.1           | Open weight             | Released 15,148,227-parameter GridSFM graph-transformer checkpoint and inference pipeline [8–10].                                                              |
| LUMINA-2M v0.1.3            | Open weight             | Released 2,375,548-parameter, six-layer heterogeneous model and inference pipeline [11, 12].                                                                   |
| GridFM-GNS v0.8.1           | Public architecture     | Released heterogeneous preprocessing, model, and decoder at random initialization; we found no compatible public AC-OPF checkpoint as of August 25, 2026 [39]. |
| DC3 (index partition)       | Design reimplementation | Random index masks for partial $V_m$ , $V_a$ , $P_g$ outputs followed by co-variant mean completion [23].                                                      |
| DC3 (physical partition)    | Matched control         | Same completion stand-in with masks keyed by slack/PV/PQ roles rather than array position.                                                                     |
| Graph transformer (LapPE-4) | Design reimplementation | Global bus self-attention with four raw, uncanonicalized Laplacian eigenvectors [17].                                                                          |

Continued on next page

Table 5 (continued)

| Row                              | Class                        | Scope                                                                                                                                                                                                                    |
|----------------------------------|------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Landmark PE ( $K=64$ )           | Mechanism control            | Four shortest-path-distance moments to the first 64 bus indices, appended to an MPNN.                                                                                                                                    |
| Landmark PE ( $K=all$ )          | Matched control              | Same moments over all buses, removing the index-selected subset.                                                                                                                                                         |
| Effective resistance ( $K=64$ )  | Design reimplementa-<br>tion | Five unit-weight effective-resistance moments over the first 64 bus indices, appended to an MPNN [8, 10].                                                                                                                |
| Effective resistance ( $K=all$ ) | Matched control              | Same moments over all buses [3]. Both effective-resistance rows use unit edge weights rather than the susceptance weighting of the cited designs, isolating index selection from electrical weighting.                   |
| Index filter (fixed)             | Matched control              | Three bilinear graph-filter layers using degree-normalized adjacency directly.                                                                                                                                           |
| Index filter (learned)           | Design reimplementa-<br>tion | Same layers with learned per-grid $N \times N$ values masked to graph support [2].                                                                                                                                       |
| MPNN (matched)                   | Matched control              | Four-layer, width-128 heterogeneous MPNN with sum aggregation, symmetric branch features, and masked solution inputs.                                                                                                    |
| MPNN + LapPE (matched)           | Matched control              | Same MPNN with 16 raw, uncanonicalized Laplacian eigenvectors.                                                                                                                                                           |
| Plain MPNN (mean)                | Family control               | Two-layer heterogeneous MPNN with shared encoders and heads, mean aggregation, and symmetric branch encoding; this control is inspired by OPF-HGNN but does not reproduce its exact aggregation or branch encoding [25]. |
| Plain MPNN (sum)                 | Family control               | Same two-layer MPNN with sum aggregation.                                                                                                                                                                                |
| Plain MPNN (orientation naive)   | Family control               | Same sum MPNN with raw endpoint fields and only the stored branch direction.                                                                                                                                             |

## D.2 Data, sampling, and solver controls

Scenarios come from the N-1 topology-perturbed OPFData release, a perturbation type that appears in the training data of both released checkpoints [9, 12, 13]. LUMINA metadata includes all five audited grids. GridSFM documents a training-size range of 500 to 4,661 buses, which places case14/118 outside that range and the larger grids inside it. These scenarios do not form a common unseen-topology benchmark for the two released checkpoints. The architecture screen uses case14/500, three source scenarios per grid, three draws per applicable case–relation pair, and three initializations where applicable. Fourteen component-split cells retain only case500 evidence because the case14 control was unavailable. The independent evaluation units are source scenarios; transformed views of a source do not count as separate units.

**Solver controls.** PowerModels.jl and IPOPT check the numerical implementation of the relations [15, 40]. The algebraic conditions still determine validity, and a converged solve need not be globally optimal. The archived expanded-relation configuration uses the polar AC formulation (ACPPowerModel), a flat start in each view, an IPOPT tolerance of  $10^{-8}$ , and a 300-second limit. The wrapper accepts LOCALLY\_SOLVED, OPTIMAL, and their ALMOST variants. These status labels do not independently verify residual bounds or global optimality. The saved summaries do not identify the linear solver.

**Expanded-relation admission.** The historical control implementation admits a grid–relation pair only if all its sampled solves converge, the worst relative objective difference is at most  $10^{-9}$ , and the worst aligned primal difference meets its relation-specific limit. Relative objective error is  $|C' - C_{\text{ref}}| / \max(|C_{\text{ref}}|, 10^{-12})$ ; disconnected-union controls use the sum of component reference costs. The primal limits are 0 for basekv-scale,  $10^{-9}$  for ref-bus-move/load-as-gen,  $10^{-7}$  for null-pad,  $10^{-6}$  for component-split/parallel-split,  $2.2 \times 10^{-6}$  for tap-side-flip, and  $10^{-5}$  for island-union/island-ref-mix. The archived case14/500 table contains 18 grid–relation pairs with nine control records each. All 162 pairs of solves converge, and 17 grid–relation pairs pass admission. Case14 component-split is unavailable because its objective difference,  $7.26 \times 10^{-9}$ , exceeds  $10^{-9}$ , even though IPOPT converged. The model screen reuses these controls and does not run a fresh solve for every model draw. Although the configuration names  $10^{-6}$  equality/inequality targets, the expanded control CSV does not store independently audited residuals, so we do not claim that those targets were separately verified for every admitted view.

**Original relabeling and orientation controls.** For the four original relations, we additionally check transported labels and residual preservation over 300 records per grid–relation on case14/118/500 and 90 on case2000/4661, and we check bus relabeling for agreement with an independent converter. Transported labels match and residual changes meet  $10^{-12}$  in all 20 grid–relation summaries. A separate re-solve comparison uses  $10^{-6}$  cost and primal tolerances. Case4661 row-order and branch-reversal controls converge but exceed that primal threshold. We keep them as solver-solution

sensitivity and do not reclassify them as surrogate failures. The distinction matters when solutions are non-unique or weakly identified.

**Screen and five-grid summaries.** The screen takes the largest channel-mean discrepancy over all executed draws and compares it with  $10^{-6}$ ; the separate maximum-component check uses  $10^{-5}$ . The five-grid comparison first averages draws within a source, then takes the median across sources and the maximum across grids, retaining a value for each of three random seeds. A checkmark in that comparison means the aggregate is below threshold, even if some individual draws exceed it. Neither statistic can prove invariance beyond the sampled tests.

### D.3 Threshold sensitivity and component tails

We changed only the reporting threshold for the 221 saved model–relation summaries, keeping the samples and applicability mask fixed. No model or solver was rerun. Every configuration still fails at least six relations at  $10^{-7}$ ,  $10^{-6}$ , and  $10^{-5}$  in native channel units. At  $10^{-6}$ , all 64 passing cells also have maximum-component discrepancies at or below  $10^{-5}$ . Thus the component check reveals no additional failure among the screen passes. This result does not apply to individual draws in the separate five-grid aggregate comparison.

Table 6: Sensitivity of the saved architecture-screen verdicts to the channel-mean reporting threshold. Each row uses the same 221 cells and partial-coverage mask.

| Threshold | Failed cells / 221 | Failed relations per model | Models with a failure |
|-----------|--------------------|----------------------------|-----------------------|
| $10^{-7}$ | 160                | 6–13                       | 17/17                 |
| $10^{-6}$ | 157                | 6–13                       | 17/17                 |
| $10^{-5}$ | 154                | 6–12                       | 17/17                 |
| $10^{-4}$ | 144                | 4–12                       | 17/17                 |
| $10^{-3}$ | 134                | 4–12                       | 17/17                 |

Excluding `basekv-scale` leaves 204 model–relation cells. At  $10^{-6}$ , 140 of these cells fail, and every configuration fails between five and twelve of the twelve remaining relations. At  $10^{-4}$  and  $10^{-3}$ , respectively, 127 and 118 cells fail; every configuration still fails between three and eleven of the twelve remaining relations. The tolerance was set for reporting; it was not calibrated to each pipeline’s numerical noise. We lack a comprehensive set of repeated-forward, round-trip, CPU/GPU, and float32/float64 comparisons. A near-threshold result therefore cannot identify a numerical cause. Changing the reporting threshold leaves the broad failure-count result intact, but changing arithmetic precision might not.

## E Detailed audit evidence

### E.1 Architecture-level failure patterns

Relabeling failures separate designs that use physical identities from those that use array positions. Component-keyed designs pass `component-relabel`, `bus-relabel`, and `row-order`. Only GridSFM-Open and GridFM fail `row-order`; their preprocessing uses row positions directly. Eight of 17 rows fail `bus-relabel`, and each uses index-derived or spectral coordinates through positional encodings, output partitions, or filters. In the matched MPNN pair, adding raw Laplacian eigenvectors changes `bus-relabel` and `island-union` from passes to failures because the eigenvector signs and bases depend on order. Appending a disconnected island leaves local, permutation-safe aggregation unchanged, but affects global positional summaries and spectral coordinates. Likewise, `branch-reversal` fails only when directed edge features are used without symmetrization, consistent with LUMINA’s trained/random result below.

The six shared failures fall into two groups. Changing the angle reference (`ref-bus-move` and `island-ref-mix`) or target-irrelevant metadata (`basekv-scale` after OPFData export) changes predictions. The pipelines therefore do not consistently remove these arbitrary choices before inference. The other three failures, `component-split`, `load-as-gen`, and `null-pad`, show that predictions depend on device count or type even when the bus-level injection is unchanged. Canoni-

calizing references and irrelevant fields may address the first group. The second may require common representations for equivalent devices and additive aggregation of their contributions at each bus.

## E.2 Trained versus random-weight consistency

Table 7 uses the worst-grid statistic defined in Sec. D.2. The trained column uses the released checkpoint; the random column gives the range across three weight seeds. Checkmarks apply to the aggregate, not to every draw.

Table 7: Trained and random-weight worst-grid discrepancies by relation. Entries are unnormalized worst-channel mean absolute discrepancies; random entries span three weight seeds. A checkmark marks an aggregate value at or below the prespecified  $10^{-6}$  channel tolerance; it does not mean that every draw passes. Because values inherit the worst channel’s unit (p.u. for  $V_m, P_g, Q_g$  or radians for  $V_a$ ), they are comparable with the tolerance but not across channels.

| Relation          | GridSFM-Open v1.1                |                                                         | LUMINA-2M v0.1.3                 |                                                          |
|-------------------|----------------------------------|---------------------------------------------------------|----------------------------------|----------------------------------------------------------|
|                   | Trained                          | Random seed range                                       | Trained                          | Random seed range                                        |
| component-relabel | $3.80 \times 10^{-7} \checkmark$ | $[3.00 \times 10^{-8}, 3.53 \times 10^{-8}] \checkmark$ | $2.18 \times 10^{-8} \checkmark$ | $[4.07 \times 10^{-9}, 1.03 \times 10^{-8}] \checkmark$  |
| bus-relabel       | 0.279                            | $[0.00957, 0.0143]$                                     | $5.01 \times 10^{-9} \checkmark$ | $[5.60 \times 10^{-10}, 3.37 \times 10^{-9}] \checkmark$ |
| row-order         | 0.780                            | $[0.574, 0.578]$                                        | $4.45 \times 10^{-8} \checkmark$ | $[5.55 \times 10^{-9}, 1.17 \times 10^{-8}] \checkmark$  |
| branch-reversal   | 0.326                            | $[0.0133, 0.0365]$                                      | 0.299                            | $[0.00528, 0.00903]$                                     |
| ref-bus-move      | 0.438                            | $[0.00755, 0.0113]$                                     | 0.182                            | $[2.40 \times 10^{-4}, 0.00789]$                         |
| component-split   | 0.00940                          | $[3.45 \times 10^{-4}, 0.00232]$                        | 0.00325                          | $[2.45 \times 10^{-5}, 7.04 \times 10^{-5}]$             |
| parallel-split    | 0.0253                           | $[0.00171, 0.00206]$                                    | $1.15 \times 10^{-4}$            | $[2.86 \times 10^{-5}, 6.86 \times 10^{-5}]$             |
| island-union      | 0.680                            | $[0.0515, 0.160]$                                       | $3.30 \times 10^{-8} \checkmark$ | $[7.51 \times 10^{-9}, 2.23 \times 10^{-8}] \checkmark$  |
| basekv-scale      | 0.397                            | $[0.0168, 0.0232]$                                      | 0.513                            | $[0.0547, 0.0782]$                                       |
| tap-side-flip     | 0.104                            | $[0.00569, 0.0141]$                                     | 0.0333                           | $[1.47 \times 10^{-4}, 2.02 \times 10^{-4}]$             |
| load-as-gen       | 0.0322                           | $[0.00782, 0.0175]$                                     | 0.00157                          | $[3.12 \times 10^{-5}, 2.06 \times 10^{-4}]$             |
| null-pad          | 0.104                            | $[0.00248, 0.00767]$                                    | 0.0231                           | $[8.96 \times 10^{-4}, 0.00226]$                         |
| island-ref-mix    | 0.675                            | $[0.244, 0.268]$                                        | 0.124                            | $[2.40 \times 10^{-4}, 0.00789]$                         |

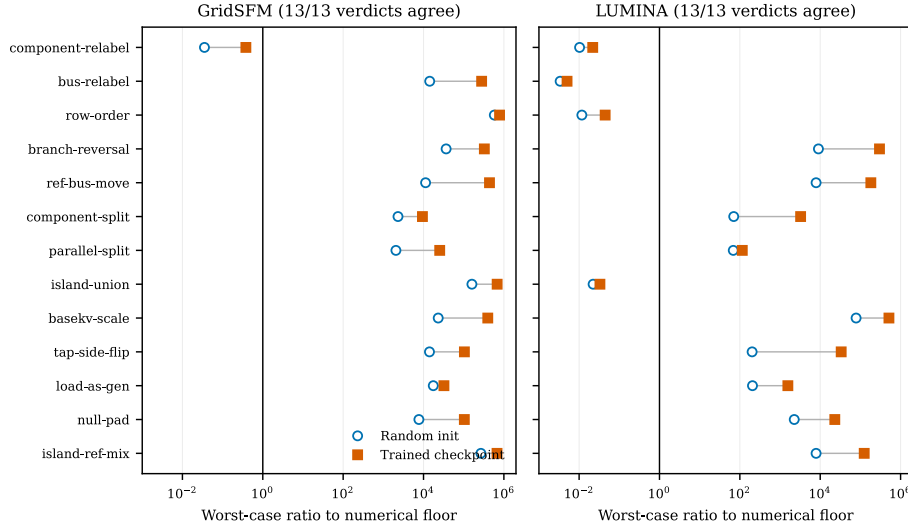


Figure 5: Trained and random-weight models agree on all 13 threshold classifications of the displayed aggregate statistic. For each relation, the horizontal coordinate is the worst-channel discrepancy, maximized over five grids, divided by the  $10^{-6}$  tolerance; random-weight markers also take the worst of three seeds. The vertical line at one marks the aggregate reporting threshold and does not serve as an all-draw pass criterion. Connecting segments show the change in magnitude after training.

GridSFM is below the aggregate threshold only on component-relabel; LUMINA is also below it on bus-relabel, row-order, and island-union, for both trained and random weights (Table 7 and Figs. 5 to 7). Excluding component-relabel, training increases discrepancy in 55 of 60 finite GridSFM case-relation cells and 50 of 58 LUMINA cells. The median ratios are 7.50 and 6.23.

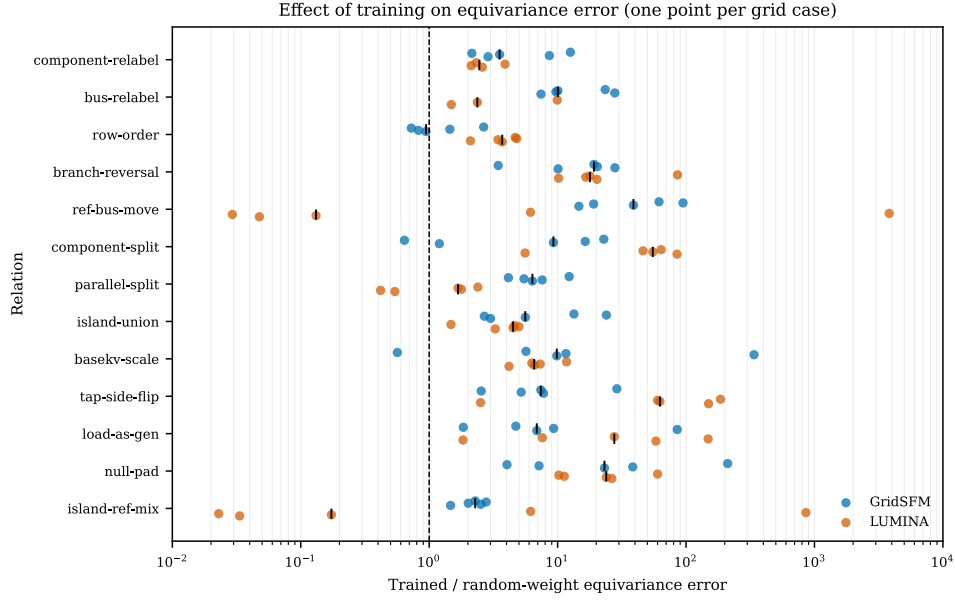


Figure 6: Per-grid effect of training on metamorphic discrepancy. Each point is the trained discrepancy divided by its random-weight control for one grid and relation; values above one indicate a larger mismatch after training. Short black bars mark relation-level medians. Ratios for `component-relabel` compare values that are already below the  $10^{-6}$  tolerance, so they carry little information.

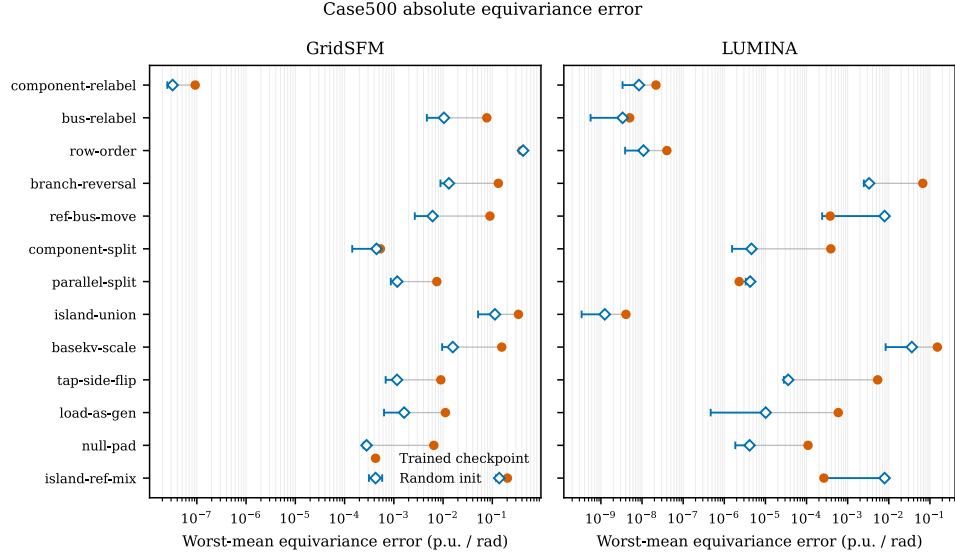


Figure 7: Absolute case500 discrepancies underlying the ratio comparison. Circles show released checkpoints; diamonds and intervals show the central value and range across three random seeds. The plot distinguishes a large relative change near zero from a discrepancy far above  $10^{-6}$ . Each point uses its worst channel, so values should be read against the tolerance and not compared across channel units.

Ratios need to be read with absolute values: a near-zero random denominator can give a large ratio, and a ratio below one can still leave a failure above tolerance. Random networks may also be nearly constant. These comparisons measure sensitivity to representation, not predictive quality; all trained and random arms fail the AC-feasibility criterion.

### E.3 Localization in released pipelines

We substituted aligned intermediate tensors to locate the source of discrepancies, without repairing or retraining a checkpoint. For case500 source 9245 under bus relabeling (draw 0), the cumulative substitutions change the legacy score from 30.4769 to 30.4712 after replacing node/cycle features, then to  $5.2535 \times 10^{-5}$  after replacing cycle incidence. This score averages channel errors divided by  $10^{-3}$ ; it is not the maximum-channel  $\Delta_r$  reported at  $10^{-6}$  elsewhere in the paper. The final-to-initial ratio is  $1.724 \times 10^{-6}$ , a relative reduction rather than an absolute pass threshold. Other relations retain substantial residuals. In a separate network-only test, edge-row shuffling leaves the output unchanged, but branch-node relabeling gives a discrepancy of  $3.51 \times 10^{-5}$  on the legacy scale. The substitutions locate much of the sensitivity in preprocessing, while the remaining network residual prevents a claim of complete network invariance.

For GridSFM, the bus-to-bus edge index is permutation-equivariant, and shuffling edge rows after feature construction leaves network predictions unchanged. Source inspection identifies four candidate causes. Effective-resistance features choose up to 64 landmarks partly by bus index. The offset used to join line and transformer edges counts appended arrays instead of edges. Taking `sign` of an almost-zero DC flow turns numerical noise into a categorical feature. The cycle basis comes from a spanning tree weighted by edge-row order. We reproduced these observations at commit 1ca775f [10].

Candidate fixes follow from these observations: choose landmarks without arbitrary indices or summarize all buses; use cumulative edge counts when joining families; replace the near-zero flow sign with a physically justified dead band or smooth mapping; and construct cycle information independently of row order. Each fix needs a patch-and-retest experiment. Retraining or fine-tuning may also be needed because the checkpoint was fitted to the existing features.

For LUMINA, both trained and random weights remain below the  $10^{-6}$  aggregate threshold on component and bus relabeling, row ordering, and disconnected-island composition. Its branch-reversal failure instead points to orientation-sensitive directed processing. Comparing weight states and intermediate computations helps distinguish preprocessing effects, architectural choices, and changes caused by training.

### E.4 Auxiliary dispatch-cost sensitivity

Let  $\hat{P}_g(d) = [f(d)]_{P_g}$  and  $\hat{P}_{g,r}(d) = [P_r f(T_r d)]_{P_g}$  be the canonical and aligned transformed active-power predictions. With the cost  $C$  in Eq. (2), define

$$S_r(d) = \frac{\left| C\left(\hat{P}_{g,r}(d)\right) - C\left(\hat{P}_g(d)\right) \right|}{\max\left(\left| C\left(P_g^*(d)\right) \right|, \varepsilon\right)}, \quad (5)$$

where  $P_g^*(d)$  is the solver dispatch and  $\varepsilon$  is machine epsilon. Table 8 reports the median numerator in cost units and the median  $S_r(d)$ . A value of 0.513, for example, means that an encoding change alters the implied cost by 51.3% of the solver’s reference cost. Because the predictions are AC infeasible,  $S_r(d)$  measures representation sensitivity and does not measure a realized dispatch loss or optimality gap. Random-weight values likewise do not measure economic quality.

Table 8: Median absolute difference between the costs implied by canonical and aligned transformed GridSFM predictions. Absolute values are in cost units; relative values divide that difference by the solver’s reference dispatch cost. Because the predictions are infeasible, these values measure sensitivity and are not optimality gaps.

| Relation          | Trained weights      |                       | Random weights        |                       |
|-------------------|----------------------|-----------------------|-----------------------|-----------------------|
|                   | Absolute             | Relative              | Absolute              | Relative              |
| component-relabel | $7.8 \times 10^{-3}$ | $1.57 \times 10^{-8}$ | $2.15 \times 10^{-3}$ | $8.18 \times 10^{-9}$ |
| bus-relabel       | $1.25 \times 10^4$   | 0.022                 | $1.65 \times 10^3$    | $3.94 \times 10^{-3}$ |
| row-order         | $6.99 \times 10^4$   | 0.234                 | $1.34 \times 10^3$    | $4.86 \times 10^{-3}$ |
| branch-reversal   | $1.98 \times 10^4$   | 0.0404                | $1.15 \times 10^3$    | $4.23 \times 10^{-3}$ |
| ref-bus-move      | $2.60 \times 10^4$   | 0.0737                | 281                   | $1.28 \times 10^{-3}$ |
| component-split   | 57.5                 | $1.40 \times 10^{-4}$ | 2.04                  | $4.92 \times 10^{-5}$ |
| parallel-split    | 180                  | $3.76 \times 10^{-4}$ | 45.2                  | $3.65 \times 10^{-4}$ |
| island-union      | $4.63 \times 10^4$   | 0.502                 | $1.98 \times 10^3$    | $5.27 \times 10^{-3}$ |
| basekv-scale      | $3.21 \times 10^4$   | 0.0350                | $1.03 \times 10^3$    | $1.81 \times 10^{-3}$ |
| tap-side-flip     | 366                  | $7.69 \times 10^{-4}$ | 105                   | $7.16 \times 10^{-4}$ |
| load-as-gen       | $2.03 \times 10^3$   | $2.46 \times 10^{-3}$ | 66.8                  | $3.32 \times 10^{-4}$ |
| null-pad          | 401                  | $1.08 \times 10^{-3}$ | 9.43                  | $9.71 \times 10^{-5}$ |
| island-ref-mix    | $9.40 \times 10^4$   | 0.513                 | $4.22 \times 10^3$    | 0.0120                |

## F Augmentation protocol and extended results

### F.1 Training and evaluation protocol

For each of four architectures, we train a baseline on canonical inputs and a second model with transformed inputs. The architectures are a four-layer, width-128 heterogeneous MPNN; that MPNN with 16 raw Laplacian eigenvectors (LapPE+MPNN); that MPNN with 16-step random-walk return probabilities (RWSE+MPNN); and LUMINA. RWSE avoids eigenvector sign and basis ambiguities [16–18]. Each uses three seeds (42, 31415, 271828), 2,000 case500 training instances, 89 epochs, 22,250 optimizer steps, batch size 8, and weight decay  $10^{-4}$ . Learning rates are  $10^{-3}$  for the MPNN variants and  $3 \times 10^{-4}$  for LUMINA. Each instance in the augmented arm is transformed by one eligible relation with probability 0.5.

All arms minimize the same equal-channel MSE plus AC constraint penalties,

$$\mathcal{L} = \frac{1}{|C|} \sum_{c \in C} \text{MSE}_c + \lambda_{\text{eq}} \log \left( 1 + \frac{1}{B} \sum_b (r_{P,b}^2 + r_{Q,b}^2) \right) + \lambda_{\text{ineq}} \frac{1}{K} \sum_k \max(0, g_k)^2. \quad (6)$$

Here  $r_{P,b}, r_{Q,b}$  are the real and reactive parts of  $S^{\text{inj}} - V \odot \overline{Y_{\text{bus}} V}$ ;  $g_k$  includes bounds, branch angle differences, and thermal limits at both ends. All 24 arms use  $\lambda_{\text{eq}} = 3.606662 \times 10^{-4}$  and  $\lambda_{\text{ineq}} = 5.929293 \times 10^{-3}$ . These weights were set offline using the median final validation-MSE-to-penalty ratio from nine earlier baseline arms on 500 validation scenarios. The recovered calibration record includes all nine arms and reproduces both weights after rounding. It describes no one-epoch tuning step. We select checkpoints using the validation objective. Within this experiment, the loss is fixed and augmentation changes; a comparison with the earlier pure-MSE experiments would change both loss and relation sets.

**Relation selection.** We froze the training sets from measured movement in earlier trained, unaugmented baseline arms and from architecture-specific exclusions. The random-initialization screen did not supply these sets. The seven MPNN and RWSE relations are ref-bus-move, component-split, parallel-split, tap-side-flip, load-as-gen, null-pad, and island-ref-mix. LapPE+MPNN adds bus-relabel and island-union. LUMINA uses the common seven plus basekv-scale and branch-reversal. The configuration records below- $10^{-6}$  baseline movement for basekv-scale in MPNN/LapPE and island-union in MPNN. Because case500 features saturate, this small movement does not imply invariance for all inputs. Dashes in Tables 9 and 10 mark relations excluded from training, and evaluation still covers all thirteen relations.

**Evaluation samples.** The main evaluation uses 50 held-out case500 source scenarios per seed. Each has 53 views: one canonical input and four draws for each of thirteen relations. The 2,650 views are repeated measurements of 50 independent sources. A separate batched evaluation uses 200 canonical scenarios from each of case14, case118, case500, case2000, and case4661, without further training. None of its case500 scenario IDs occurs in the main sample. Sec. F.4 explains how the different samples and evaluation paths affect the comparison.

Accuracy averages the native-unit MAEs of  $V_m, V_a, P_g, Q_g$  with equal weights. This convention combines per unit and radians; Table 9 also reports each channel separately. For the main tables, improvement is the mean of three paired seed-level percentages,  $100(e_{\text{base},s} - e_{\text{aug},s})/e_{\text{base},s}$ , so positive values mean lower error. Table 12 uses the opposite sign. The spreads in Sec. F.2 are population standard deviations over the three percentage changes, not confidence intervals.

To summarize consistency, we average draws within each source, take the median across sources, then the median across the model’s trained relations, and finally the mean over seeds. Per-relation entries instead use the seed with the largest source median. Since models train on different relation sets, their aggregate medians can compare baseline with augmentation within a model but cannot rank models. We recompute discrepancies from native-unit errors at the paper’s  $10^{-6}$  scale; older artifacts used  $10^{-3}$ .

For feasibility, we first take the largest violation within each constraint family for each prediction. Within a seed, we then take the worst of each source’s 53 views and compute the median, linearly interpolated P95, and maximum over 50 sources. Tables give the mean and sample standard deviation of the resulting three seed-level statistics. Equality violation is the larger maximum nodal active or reactive balance error, in p.u. Pooled inequality violation is the largest family violation divided by

that family’s  $10^{-6}$  native-unit tolerance, and is dimensionless. A source is feasible only if every view meets both the equality and all inequality thresholds. No evaluated prediction does.

Table 9: Full augmentation results under the AC constraint-penalty objective. Positive improvement (Imp.) denotes the mean seed-paired reduction under augmentation. Metamorphic values report the dimensionless normalized discrepancy  $\Delta_r(d)$  defined in (1); 0 denotes exact invariance and dashes mark excluded relations. Constraint rows use the scenario-clustered seed-level P95 defined in the text and report mean  $\pm$  sample standard deviation across three seeds. Equality is in per unit. The pooled inequality metric is dimensionless and displayed in millions of family-specific  $10^{-6}$  tolerances. Absolute family-level results appear in Table 11.

| Outcome                       | Presentation             | Metric                | MPNN               |                     |                 | LapPE+MPNN                      |                    |                    | RWSE+MPNN          |                    |                                  | LUMINA                          |                    |               |                |      |
|-------------------------------|--------------------------|-----------------------|--------------------|---------------------|-----------------|---------------------------------|--------------------|--------------------|--------------------|--------------------|----------------------------------|---------------------------------|--------------------|---------------|----------------|------|
|                               |                          |                       | Base               | Aug.                | Imp. (%)        | Base                            | Aug.               | Imp. (%)           | Base               | Aug.               | Imp. (%)                         | Base                            | Aug.               | Imp. (%)      |                |      |
| Mean absolute error (MAE)     | Canonical                | Mean                  | 0.01295            | 0.01345             | −3.9            | 0.09983                         | 0.01309            | 86.8               | 0.01255            | 0.01250            | 0.4                              | 0.01834                         | 0.01889            | −3.0          |                |      |
|                               |                          | $V_m$                 | 0.00425            | 0.00471             | −12.7           | 0.01453                         | 0.00447            | 69.1               | 0.00407            | 0.00424            | −4.2                             | 0.00674                         | 0.00758            | −12.4         |                |      |
|                               |                          | $V_e$                 | 0.01289            | 0.01290             | −0.1            | 0.12493                         | 0.01320            | 89.2               | 0.01274            | 0.01263            | 0.9                              | 0.03500                         | 0.03606            | −3.0          |                |      |
|                               |                          | $P_\beta$             | 0.01198            | 0.01268             | −5.9            | 0.15529                         | 0.01176            | 92.3               | 0.01103            | 0.01088            | 1.4                              | 0.00952                         | 0.00961            | −0.9          |                |      |
|                               |                          | $Q_\beta$             | 0.02270            | 0.02352             | −3.6            | 0.10459                         | 0.02292            | 77.9               | 0.02236            | 0.02226            | 0.4                              | 0.02211                         | 0.02233            | −1.0          |                |      |
|                               |                          | $Q_\beta$             | 0.01312            | 0.01353             | −3.2            | 0.10211                         | 0.01318            | 87.0               | 0.01271            | 0.01257            | 1.1                              | 0.03594                         | 0.02089            | 41.9          |                |      |
|                               | Transformed              | Mean                  | 0.00427            | 0.00472             | −12.5           | 0.01468                         | 0.00447            | 69.5               | 0.00409            | 0.00425            | −4.0                             | 0.00845                         | 0.00826            | 2.2           |                |      |
|                               |                          | $V_m$                 | 0.01308            | 0.01300             | 0.6             | 0.13365                         | 0.01339            | 89.9               | 0.01293            | 0.01271            | 1.6                              | 0.06418                         | 0.04219            | 34.3          |                |      |
|                               |                          | $V_e$                 | 0.01227            | 0.01279             | −4.2            | 0.15438                         | 0.01187            | 92.2               | 0.01133            | 0.01097            | 3.1                              | 0.03263                         | 0.00977            | 70.0          |                |      |
|                               |                          | $P_\beta$             | 0.02285            | 0.02361             | −3.4            | 0.10573                         | 0.02299            | 78.1               | 0.02251            | 0.02234            | 0.7                              | 0.03850                         | 0.02333            | 39.3          |                |      |
|                               |                          | $Q_\beta$             | 0.01312            | 0.01353             | −3.2            | 0.10211                         | 0.01318            | 87.0               | 0.01271            | 0.01257            | 1.1                              | 0.03594                         | 0.02089            | 41.9          |                |      |
|                               |                          | $Q_\beta$             | 0.00427            | 0.00472             | −12.5           | 0.01468                         | 0.00447            | 69.5               | 0.00409            | 0.00425            | −4.0                             | 0.00845                         | 0.00826            | 2.2           |                |      |
| Out-of-distribution (OOD) MAE | case14                   | Mean                  | 0.29138            | 0.27818             | 3.8             | 0.34591                         | 0.26954            | 21.8               | 0.29786            | 0.28374            | 4.3                              | 0.15830                         | 0.14110            | 11.0          |                |      |
|                               | case118                  | Mean                  | 0.47254            | 0.53065             | −11.3           | 0.48121                         | 0.46921            | 2.8                | 0.44662            | 0.51278            | −13.5                            | 0.29558                         | 0.33423            | −14.2         |                |      |
|                               | case2000                 | Mean                  | 0.25068            | 0.24089             | 3.9             | 0.21860                         | 0.23550            | −7.7               | 0.25316            | 0.24444            | 3.5                              | 0.22905                         | 0.21569            | 5.7           |                |      |
|                               | case4661                 | Mean                  | 0.49780            | 0.49983             | −0.2            | 0.41962                         | 0.47545            | −13.4              | 0.47101            | 0.49675            | −6.0                             | 0.48663                         | 0.44790            | 8.0           |                |      |
|                               | Across trained relations | Median $\Delta_r$     | 852                | 595                 | 30.1            | $1.00 \times 10^3$              | 805                | 16.2               | 883                | 563                | $36.1 \text{ } 4.63 \times 10^3$ | $1.26 \times 10^3$              | 70.8               |               |                |      |
| Metamorphic                   | Per relation             | bus-relabel           | 0                  | 0                   | —               | $1.95 \times 10^5$              | $5.32 \times 10^3$ | 97.3               | 0                  | 0                  | —                                | 0                               | 0                  | —             |                |      |
|                               |                          | branch-reversal       | —                  | —                   | —               | —                               | —                  | —                  | —                  | —                  | —                                | $2.16 \times 10^5$              | $5.47 \times 10^4$ | 74.7          |                |      |
|                               |                          | ref-bus-move          | $1.47 \times 10^3$ | 923                 | 37.1            | $1.50 \times 10^3$              | 949                | 36.9               | 988                | $1.02 \times 10^3$ | —                                | $3.3 \text{ } 8.02 \times 10^4$ | $6.29 \times 10^4$ | 21.5          |                |      |
|                               |                          | component-split       | 898                | 674                 | 24.9            | 577                             | 668                | −15.7              | $1.00 \times 10^3$ | 703                | 29.9                             | 190                             | 139                | 26.5          |                |      |
|                               |                          | parallel-split        | $1.42 \times 10^3$ | 975                 | 31.5            | $1.05 \times 10^3$              | $3.89 \times 10^3$ | 96.3               | $1.25 \times 10^3$ | 999                | 20.1                             | 23.8                            | 17.0               | 28.6          |                |      |
|                               |                          | island-union          | 0                  | 0                   | —               | $1.62 \times 10^5$              | $5.81 \times 10^3$ | 96.4               | 0                  | 0                  | —                                | —                               | —                  | —             |                |      |
|                               |                          | basekv-scale          | —                  | —                   | —               | —                               | —                  | —                  | —                  | —                  | —                                | $2.07 \times 10^5$              | $1.57 \times 10^4$ | 92.5          |                |      |
|                               |                          | tap-side-flip         | 167                | 90.9                | 45.4            | 61.9                            | 31.0               | 49.9               | 41.3               | 50.0               | −21.2                            | $6.33 \times 10^3$              | $1.56 \times 10^3$ | 75.3          |                |      |
|                               |                          | load-as-gen           | $1.01 \times 10^3$ | 644                 | 36.1            | 658                             | 747                | −13.4              | $1.19 \times 10^3$ | 594                | 50.0                             | $1.98 \times 10^3$              | 478                | 75.9          |                |      |
|                               |                          | null-pad              | $1.32 \times 10^3$ | 736                 | 44.4            | $803 \text{ } 1.04 \times 10^3$ | −29.4              | $1.20 \times 10^3$ | $1.02 \times 10^3$ | 15.6               | 797                              | 316                             | 60.4               | 21.7          |                |      |
|                               |                          | island-ref-mix        | $1.10 \times 10^3$ | 695                 | 37.0            | $1.59 \times 10^3$              | $5.74 \times 10^3$ | 96.4               | 744                | 766                | −2.9                             | $6.01 \times 10^4$              | $4.70 \times 10^4$ | 21.7          |                |      |
|                               |                          | Constraint violations | All views          | Equality P95 (p.u.) | 58.1 $\pm$ 18.7 | 42.2 $\pm$ 2.3                  | 22.1               | 113.9 $\pm$ 42.1   | 39.5 $\pm$ 5.2     | 61.7               | 51.8 $\pm$ 10.9                  | 33.5 $\pm$ 2.4                  | 33.4               | 871 $\pm$ 342 | 164 $\pm$ 11   | 78.7 |
|                               |                          |                       |                    | Ineq. P95 (tol.)    | 19.3 $\pm$ 4.2  | 11.9 $\pm$ 1.6                  | 35.2               | 42.1 $\pm$ 4.6     | 12.7 $\pm$ 3.6     | 70.1               | 18.8 $\pm$ 2.8                   | 10.2 $\pm$ 0.8                  | 45.4               | 276 $\pm$ 91  | 20.1 $\pm$ 2.1 | 92.3 |

Panel A of Table 10 takes the median over the relations used to train the models. Per-channel prediction MAEs appear in Table 9.

## F.2 Accuracy and relation-level effects

On transformed scenarios, mean MAE improves by +87.0% for LapPE+MPNN, +41.9% for LUMINA, and +1.1% for RWSE+MPNN. MPNN has an improvement of −3.2%, meaning its error increases. Population standard deviations across the seed-level changes are 1.2, 1.5, 0.7, and 3.3 percentage points, respectively. The first three models improve in all three seeds; MPNN improves in one. Most of the LapPE gain is explained by the eigensolver mismatch (Sec. F.4). On canonical inputs, improvements are +86.8% for LapPE+MPNN, +0.4% for RWSE+MPNN, −3.9% for MPNN, and −3.0% for LUMINA. MPNN and LUMINA worsen in every seed.

Median  $\Delta_r$  falls by 30.1%, 16.2%, 36.1%, and 70.8% for MPNN, LapPE+MPNN, RWSE+MPNN, and LUMINA, respectively. It decreases in 11 of 12 model–seed pairs; one LapPE+MPNN seed is almost unchanged at −0.1%. The relations responsible differ. LUMINA’s largest baseline discrepancies are on branch-reversal and basekv-scale. LapPE+MPNN’s are on bus-relabel, island-union, island-ref-mix, and parallel-split, which expose the sign, basis, and ordering ambiguities of raw eigenvectors. Augmentation reduces those four discrepancies by 96 to 97%.

These gains coexist with regressions. LapPE+MPNN worsens on null-pad (−29.4% improvement), component-split (−15.7%), and load-as-gen (−13.4%). RWSE+MPNN worsens on tap-side-flip (−21.2%), ref-bus-move (−3.3%), and island-ref-mix (−2.9%). MPNN improves on all seven trained relations and LUMINA on all nine. The median alone hides these differences and cannot establish that each relation has been repaired.

## F.3 Feasibility and out-of-distribution behavior

No prediction meets both  $10^{-6}$  tolerances, so no source has all 53 views feasible. The feasible-source rate is 0% for every arm, over 50 sources per seed. Equality and pooled inequality P95 fall with augmentation for all four models, but the reductions are insufficient to reach feasibility.

Table 10: Expanded augmentation results under the AC constraint-penalty objective. Panel A compares three-seed mean accuracy, median  $\Delta_r$  over each model’s trained relations, and scenario-clustered constraint P95; violation entries are mean  $\pm$  sample standard deviation across seeds. Equality is in per unit, while pooled inequality is dimensionless and displayed in millions of family-specific  $10^{-6}$  tolerances. Feasible is the percentage of 50 source scenarios per seed whose 53 views all satisfy both  $10^{-6}$  tolerances. Panel B gives the percentage reduction in  $\Delta_r$  by relation; positive values denote improvement under augmentation, and dashes mark excluded relations.

| A. Aggregate outcomes |      |               |                 |                    |                     |                                      |              |
|-----------------------|------|---------------|-----------------|--------------------|---------------------|--------------------------------------|--------------|
| Model                 | Arm  | Canonical MAE | Transformed MAE | Median $\Delta_r$  | Equality P95 (p.u.) | Inequality P95 ( $10^6 \times$ tol.) | Feasible (%) |
| MPNN                  | Base | 0.01295       | 0.01312         | 852                | $58.1 \pm 18.7$     | $19.3 \pm 4.2$                       | 0            |
|                       | Aug. | 0.01345       | 0.01353         | 595                | $42.2 \pm 2.3$      | $11.9 \pm 1.6$                       | 0            |
| LapPE+MPNN            | Base | 0.09983       | 0.10211         | $1.00 \times 10^3$ | $113.9 \pm 42.1$    | $42.1 \pm 4.6$                       | 0            |
|                       | Aug. | 0.01309       | 0.01318         | 805                | $39.5 \pm 5.2$      | $12.7 \pm 3.6$                       | 0            |
| RWSE+MPNN             | Base | 0.01255       | 0.01271         | 883                | $51.8 \pm 10.9$     | $18.8 \pm 2.8$                       | 0            |
|                       | Aug. | 0.01250       | 0.01257         | 563                | $33.5 \pm 2.4$      | $10.2 \pm 0.8$                       | 0            |
| LUMINA                | Base | 0.01834       | 0.03594         | $4.63 \times 10^3$ | $871 \pm 342$       | $276 \pm 91$                         | 0            |
|                       | Aug. | 0.01889       | 0.02089         | $1.26 \times 10^3$ | $164 \pm 11$        | $20.1 \pm 2.1$                       | 0            |

| B. Metamorphic-discrepancy change by relation (%) |      |            |           |        |
|---------------------------------------------------|------|------------|-----------|--------|
| Relation                                          | MPNN | LapPE+MPNN | RWSE+MPNN | LUMINA |
| bus-relabel                                       | —    | 97.3       | —         | —      |
| branch-reversal                                   | —    | —          | —         | 74.7   |
| ref-bus-move                                      | 37.1 | 36.9       | −3.3      | 21.5   |
| component-split                                   | 24.9 | −15.7      | 29.9      | 26.5   |
| parallel-split                                    | 31.5 | 96.3       | 20.1      | 28.6   |
| island-union                                      | —    | 96.4       | —         | —      |
| basekv-scale                                      | —    | —          | —         | 92.5   |
| tap-side-flip                                     | 45.4 | 49.9       | −21.2     | 75.3   |
| load-as-gen                                       | 36.1 | −13.4      | 50.0      | 75.9   |
| null-pad                                          | 44.4 | −29.4      | 15.6      | 60.4   |
| island-ref-mix                                    | 37.0 | 96.4       | −2.9      | 21.7   |

Panel A of Table 11 reports each constraint family separately. A source contributes its worst violation over all views; the table gives means and sample standard deviations across the seed-level medians, P95 values, and maxima. Active and reactive power, generator bounds, and thermal limits are in per unit on the 100-MVA base. Voltage is in per unit and angle-difference excess in radians. Keeping families separate avoids taking absolute maxima across incompatible units. The LUMINA baseline violates angle-difference limits even though all other arms have zero violations in that family.

Panel B expresses violations relative to physical scales. In B1, the nodal complex-power residual is  $\sqrt{r_{P,b}^2 + r_{Q,b}^2}$ . Its maximum is divided by the source’s total apparent demand,  $\sum_b \sqrt{P_{d,b}^2 + Q_{d,b}^2}$ , using the same denominator for every transformed view. The canonical metric averages these values within each seed; percentage improvement is computed before averaging over seeds. In B2, positive inequality excess is divided by the relevant element’s nominal voltage, generator capacity range, branch rating, or violated angle limit. We then take the worst view per source before computing P95 and maxima. These percentages give scale to the violations; absolute tolerances still determine feasibility. All evaluated denominators were finite, nonzero, present, and active, so no element was excluded.

Augmentation’s effect on other grids varies by model and seed (Table 12). All three seeds agree on improvement for LapPE+MPNN on case14 (21.8%) and RWSE+MPNN on case2000 (3.5%). They agree on worse LapPE+MPNN error on case2000 and case4661 (7.7% and 13.4%). LUMINA improves on case14, case2000, and case4661 (11.0%, 5.7%, and 8.0%), while worsening on the separate in-distribution case500 sample (3.4%). Every other model–grid combination has seed-

Table 11: AC constraint violations from the canonical Julia audit. Panel A reports absolute family-level violations as baseline  $\rightarrow$  augmented; each side is the mean  $\pm$  sample standard deviation across three seed-level statistics after clustering 53 views under each of 50 source scenarios. P95 uses linear interpolation. Panel B reports normalized context using scenario- or element-specific denominators. The AC-feasible source-scenario rate is 0% for every arm at the  $10^{-6}$  tolerance.

| A. Absolute violations by constraint family |                                 |      |                                               |                                               |                                               |                                               |                                               |                             |
|---------------------------------------------|---------------------------------|------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------|
| Model                                       | Constraint family               | Unit | Median                                        |                                               | P95                                           |                                               | Maximum                                       |                             |
|                                             |                                 |      | Base                                          | Aug.                                          | Base                                          | Aug.                                          | Base                                          | Aug.                        |
| MPNN                                        | Active-power balance            | p.u. | 21.1 $\pm$ 2.39                               | 13.2 $\pm$ 1.59                               | 58.0 $\pm$ 18.7                               | 42.2 $\pm$ 2.26                               | 64.1 $\pm$ 16.5                               | 50.8 $\pm$ 6.4              |
|                                             | Reactive-power balance          | p.u. | 8.47 $\pm$ 0.476                              | 6.5 $\pm$ 0.746                               | 24.2 $\pm$ 9.59                               | 14.9 $\pm$ 1.68                               | 36.9 $\pm$ 10.2                               | 22.3 $\pm$ 1.14             |
|                                             | Voltage bounds                  | p.u. | 0.00756 $\pm$ 0.00303                         | 0.00617 $\pm$ 0.00361                         | 0.0288 $\pm$ 0.0149                           | 0.0225 $\pm$ 0.0133                           | 0.0406 $\pm$ 0.0186                           | 0.0491 $\pm$ 0.0351         |
|                                             | Generator active-power bounds   | p.u. | 0.206 $\pm$ 0.116                             | 0.0970 $\pm$ 0.0385                           | 2.68 $\pm$ 0.482                              | 0.647 $\pm$ 0.215                             | 10.2 $\pm$ 2.16                               | 1.29 $\pm$ 0.468            |
|                                             | Generator reactive-power bounds | p.u. | 0.0333 $\pm$ 0.00659                          | 0.0272 $\pm$ 0.00723                          | 0.138 $\pm$ 0.041                             | 0.109 $\pm$ 0.0848                            | 0.418 $\pm$ 0.173                             | 0.163 $\pm$ 0.0549          |
|                                             | Branch thermal limits           | p.u. | 4.92 $\pm$ 1.82                               | 2.5 $\pm$ 0.253                               | 19.3 $\pm$ 4.16                               | 11.9 $\pm$ 1.57                               | 28.8 $\pm$ 11.4                               | 20.2 $\pm$ 3.97             |
| LapPE+MPNN                                  | Angle-difference limits         | rad  | 0 $\pm$ 0                                     | 0 $\pm$ 0                                     | 0 $\pm$ 0                                     | 0 $\pm$ 0                                     | 0 $\pm$ 0                                     | 0 $\pm$ 0                   |
|                                             | Active-power balance            | p.u. | 89.2 $\pm$ 29.3                               | 12.6 $\pm$ 2.86                               | 114 $\pm$ 42.1                                | 39.5 $\pm$ 5.17                               | 128 $\pm$ 45.1                                | 52.0 $\pm$ 17.5             |
|                                             | Reactive-power balance          | p.u. | 38.2 $\pm$ 5.89                               | 6.18 $\pm$ 1.14                               | 51.3 $\pm$ 2.59                               | 12.2 $\pm$ 1.76                               | 61.5 $\pm$ 6.8                                | 14.7 $\pm$ 2                |
|                                             | Voltage bounds                  | p.u. | 0.0447 $\pm$ 0.0123                           | 0.00566 $\pm$ 6.30 $\times 10^{-4}$           | 0.0583 $\pm$ 0.0147                           | 0.0162 $\pm$ 0.0048                           | 0.0777 $\pm$ 0.0302                           | 0.0464 $\pm$ 0.0117         |
|                                             | Generator active-power bounds   | p.u. | 1.32 $\pm$ 0.712                              | 0.0828 $\pm$ 0.0272                           | 2.29 $\pm$ 0.733                              | 0.565 $\pm$ 0.0711                            | 9.16 $\pm$ 4.15                               | 1.04 $\pm$ 0.208            |
|                                             | Generator reactive-power bounds | p.u. | 0.23 $\pm$ 0.0212                             | 0.0252 $\pm$ 0.0166                           | 0.404 $\pm$ 0.0964                            | 0.0709 $\pm$ 0.014                            | 0.581 $\pm$ 0.186                             | 0.103 $\pm$ 0.0321          |
| RWSE+MPNN                                   | Branch thermal limits           | p.u. | 32.9 $\pm$ 5.78                               | 2.48 $\pm$ 0.685                              | 42.1 $\pm$ 4.6                                | 12.7 $\pm$ 3.59                               | 53.1 $\pm$ 5.37                               | 17.9 $\pm$ 7.29             |
|                                             | Angle-difference limits         | rad  | 0 $\pm$ 0                                     | 0 $\pm$ 0                                     | 0 $\pm$ 0                                     | 0 $\pm$ 0                                     | 0 $\pm$ 0                                     | 0 $\pm$ 0                   |
|                                             | Active-power balance            | p.u. | 19.2 $\pm$ 3.05                               | 10.6 $\pm$ 2.18                               | 51.8 $\pm$ 10.9                               | 33.5 $\pm$ 2.38                               | 74.2 $\pm$ 13.3                               | 46.6 $\pm$ 4.53             |
|                                             | Reactive-power balance          | p.u. | 8.39 $\pm$ 1.49                               | 5.18 $\pm$ 0.414                              | 19.3 $\pm$ 4.62                               | 12.1 $\pm$ 0.743                              | 30.2 $\pm$ 5.67                               | 13.5 $\pm$ 1.31             |
|                                             | Voltage bounds                  | p.u. | 0.00707 $\pm$ 0.00302                         | 0.00629 $\pm$ 0.00287                         | 0.0206 $\pm$ 0.00408                          | 0.0196 $\pm$ 0.00329                          | 0.0347 $\pm$ 0.00747                          | 0.0333 $\pm$ 0.012          |
|                                             | Generator active-power bounds   | p.u. | 0.162 $\pm$ 0.0675                            | 0.0709 $\pm$ 0.0325                           | 2.21 $\pm$ 0.666                              | 0.525 $\pm$ 0.112                             | 10.6 $\pm$ 1.41                               | 1.02 $\pm$ 0.211            |
| LUMINA                                      | Generator reactive-power bounds | p.u. | 0.0268 $\pm$ 0.00296                          | 0.0219 $\pm$ 0.00326                          | 0.186 $\pm$ 0.128                             | 0.0868 $\pm$ 0.0609                           | 0.289 $\pm$ 0.103                             | 0.115 $\pm$ 0.0765          |
|                                             | Branch thermal limits           | p.u. | 4.33 $\pm$ 0.489                              | 2.72 $\pm$ 0.468                              | 18.8 $\pm$ 2.82                               | 10.2 $\pm$ 0.843                              | 25.8 $\pm$ 3.6                                | 13.6 $\pm$ 3.63             |
|                                             | Angle-difference limits         | rad  | 0 $\pm$ 0                                     | 0 $\pm$ 0                                     | 0 $\pm$ 0                                     | 0 $\pm$ 0                                     | 0 $\pm$ 0                                     | 0 $\pm$ 0                   |
|                                             | Active-power balance            | p.u. | 736 $\pm$ 411                                 | 91.2 $\pm$ 9.83                               | 871 $\pm$ 342                                 | 164 $\pm$ 10.6                                | 989 $\pm$ 401                                 | 188 $\pm$ 8.73              |
|                                             | Reactive-power balance          | p.u. | 568 $\pm$ 512                                 | 24.2 $\pm$ 2.75                               | 693 $\pm$ 405                                 | 40.3 $\pm$ 5.05                               | 1473 $\pm$ 1007                               | 54.5 $\pm$ 2.02             |
|                                             | Voltage bounds                  | p.u. | 2.38 $\times 10^{-8} \pm 0$                   | 0 $\pm$ 0                                     | 2.38 $\times 10^{-8} \pm 0$                   | 0 $\pm$ 0                                     | 2.38 $\times 10^{-8} \pm 0$                   | 0 $\pm$ 0                   |
| LUMINA                                      | Generator active-power bounds   | p.u. | 7.12 $\times 10^{-8} \pm 4.58 \times 10^{-8}$ | 3.21 $\times 10^{-8} \pm 1.71 \times 10^{-8}$ | 1.06 $\times 10^{-7} \pm 3.03 \times 10^{-8}$ | 3.21 $\times 10^{-8} \pm 1.71 \times 10^{-8}$ | 1.24 $\times 10^{-7} \pm 0$                   | 4.20 $\times 10^{-8} \pm 0$ |
|                                             | Generator reactive-power bounds | p.u. | 6.44 $\times 10^{-9} \pm 7.46 \times 10^{-9}$ | 0 $\pm$ 0                                     | 1.86 $\times 10^{-8} \pm 2.83 \times 10^{-8}$ | 0 $\pm$ 0                                     | 2.86 $\times 10^{-8} \pm 4.56 \times 10^{-8}$ | 0 $\pm$ 0                   |
|                                             | Branch thermal limits           | p.u. | 238 $\pm$ 119                                 | 10.8 $\pm$ 0.727                              | 276 $\pm$ 91.0                                | 20.1 $\pm$ 2.13                               | 502 $\pm$ 386                                 | 32.0 $\pm$ 3.52             |
|                                             | Angle-difference limits         | rad  | 10.9 $\pm$ 1.85                               | 0 $\pm$ 0                                     | 11.9 $\pm$ 0.445                              | 0 $\pm$ 0                                     | 13.2 $\pm$ 1.9                                | 0 $\pm$ 0                   |

**Table 11 (continued).** AC constraint violations from the canonical Julia audit. Panel B1 gives system-scale equality context and the voltage global maximum. Panel B2 reports relative inequality overruns by family. Each entry is the mean  $\pm$  sample standard deviation across three seeds.

| B. Normalized interpretability metrics        |                                                                              |                                               |                                                |                                               |                                                                                          |                  |      |          |
|-----------------------------------------------|------------------------------------------------------------------------------|-----------------------------------------------|------------------------------------------------|-----------------------------------------------|------------------------------------------------------------------------------------------|------------------|------|----------|
| B1. System-scale equality and voltage context |                                                                              |                                               |                                                |                                               |                                                                                          |                  |      |          |
| Model                                         | Global-max nodal complex-power balance residual<br>(% total apparent demand) |                                               | Global-max voltage-bound excess<br>(% nominal) |                                               | Canonical mean maximum nodal complex-power balance residual<br>(% total apparent demand) |                  |      | Imp. (%) |
|                                               | Base                                                                         | Aug.                                          | Base                                           | Aug.                                          | Base                                                                                     | Aug.             |      |          |
| MPNN                                          | 36.9 $\pm$ 7.29                                                              | 28.8 $\pm$ 2.42                               | 4.06 $\pm$ 1.86                                | 4.91 $\pm$ 3.51                               | 4.32 $\pm$ 0.484                                                                         | 3.94 $\pm$ 0.393 |      | 8.3      |
| LapPE+MPNN                                    | 72.6 $\pm$ 25.1                                                              | 28.6 $\pm$ 9.52                               | 7.77 $\pm$ 3.02                                | 4.64 $\pm$ 1.17                               | 29.6 $\pm$ 6.21                                                                          | 3.34 $\pm$ 0.385 |      | 88.4     |
| RWSE+MPNN                                     | 40.4 $\pm$ 7.42                                                              | 25.7 $\pm$ 2.71                               | 3.47 $\pm$ 0.747                               | 3.33 $\pm$ 1.2                                | 3.78 $\pm$ 0.28                                                                          | 3.31 $\pm$ 0.157 |      | 12.2     |
| LUMINA                                        | 893 $\pm$ 499                                                                | 103 $\pm$ 5.4                                 | 2.38 $\times 10^{-6} \pm 0$                    | 0 $\pm$ 0                                     | 10.0 $\pm$ 0.493                                                                         | 8.85 $\pm$ 0.481 |      | 11.8     |
| B2. Relative inequality overruns (%)          |                                                                              |                                               |                                                |                                               |                                                                                          |                  |      |          |
| Model                                         | Constraint family                                                            | P95                                           |                                                | Maximum                                       |                                                                                          |                  |      |          |
|                                               |                                                                              | Base                                          | Aug.                                           | Base                                          | Aug.                                                                                     | Base             | Aug. |          |
| MPNN                                          | Voltage bounds                                                               | 2.88 $\pm$ 1.49                               | 2.25 $\pm$ 1.33                                | 4.06 $\pm$ 1.86                               | 4.91 $\pm$ 3.51                                                                          |                  |      |          |
|                                               | Generator active-power bounds                                                | 157 $\pm$ 13.4                                | 135 $\pm$ 47.5                                 | 270 $\pm$ 77.9                                | 202 $\pm$ 129                                                                            |                  |      |          |
|                                               | Generator reactive-power bounds                                              | 838 $\pm$ 434                                 | 628 $\pm$ 282                                  | 3212 $\pm$ 2238                               | 2570 $\pm$ 1685                                                                          |                  |      |          |
|                                               | Branch thermal limits                                                        | 892 $\pm$ 197                                 | 375 $\pm$ 108                                  | 1418 $\pm$ 675                                | 708 $\pm$ 158                                                                            |                  |      |          |
|                                               | Angle-difference limits                                                      | 0 $\pm$ 0                                     | 0 $\pm$ 0                                      | 0 $\pm$ 0                                     | 0 $\pm$ 0                                                                                |                  |      |          |
| LapPE+MPNN                                    | Voltage bounds                                                               | 5.83 $\pm$ 1.47                               | 1.62 $\pm$ 0.480                               | 7.77 $\pm$ 3.02                               | 4.64 $\pm$ 1.17                                                                          |                  |      |          |
|                                               | Generator active-power bounds                                                | 616 $\pm$ 122                                 | 165 $\pm$ 104                                  | 942 $\pm$ 431                                 | 409 $\pm$ 293                                                                            |                  |      |          |
|                                               | Generator reactive-power bounds                                              | 1203 $\pm$ 408                                | 494 $\pm$ 249                                  | 2687 $\pm$ 1645                               | 1934 $\pm$ 198                                                                           |                  |      |          |
|                                               | Branch thermal limits                                                        | 2098 $\pm$ 288                                | 541 $\pm$ 222                                  | 2309 $\pm$ 306                                | 649 $\pm$ 185                                                                            |                  |      |          |
|                                               | Angle-difference limits                                                      | 0 $\pm$ 0                                     | 0 $\pm$ 0                                      | 0 $\pm$ 0                                     | 0 $\pm$ 0                                                                                |                  |      |          |
| RWSE+MPNN                                     | Voltage bounds                                                               | 2.06 $\pm$ 0.408                              | 1.96 $\pm$ 0.329                               | 3.47 $\pm$ 0.747                              | 3.33 $\pm$ 1.20                                                                          |                  |      |          |
|                                               | Generator active-power bounds                                                | 180 $\pm$ 84.6                                | 216 $\pm$ 33.6                                 | 499 $\pm$ 405                                 | 679 $\pm$ 485                                                                            |                  |      |          |
|                                               | Generator reactive-power bounds                                              | 688 $\pm$ 548                                 | 438 $\pm$ 299                                  | 1663 $\pm$ 1174                               | 1534 $\pm$ 939                                                                           |                  |      |          |
|                                               | Branch thermal limits                                                        | 803 $\pm$ 73.9                                | 458 $\pm$ 79.9                                 | 1141 $\pm$ 229                                | 601 $\pm$ 127                                                                            |                  |      |          |
|                                               | Angle-difference limits                                                      | 0 $\pm$ 0                                     | 0 $\pm$ 0                                      | 0 $\pm$ 0                                     | 0 $\pm$ 0                                                                                |                  |      |          |
| LUMINA                                        | Voltage bounds                                                               | 2.38 $\times 10^{-6} \pm 0$                   | 0 $\pm$ 0                                      | 2.38 $\times 10^{-6} \pm 0$                   | 0 $\pm$ 0                                                                                |                  |      |          |
|                                               | Generator active-power bounds                                                | 1.19 $\times 10^{-5} \pm 1.23 \times 10^{-6}$ | 4.76 $\times 10^{-6} \pm 2.62 \times 10^{-6}$  | 1.19 $\times 10^{-5} \pm 1.23 \times 10^{-6}$ | 4.85 $\times 10^{-6} \pm 2.53 \times 10^{-6}$                                            |                  |      |          |
|                                               | Generator reactive-power bounds                                              | 7.01 $\times 10^{-7} \pm 7.38 \times 10^{-7}$ | 0 $\pm$ 0                                      | 8.63 $\times 10^{-7} \pm 9.99 \times 10^{-7}$ | 0 $\pm$ 0                                                                                |                  |      |          |
|                                               | Branch thermal limits                                                        | 9400 $\pm$ 5026                               | 859 $\pm$ 27.6                                 | 9975 $\pm$ 5084                               | 1537 $\pm$ 233                                                                           |                  |      |          |
|                                               | Angle-difference limits                                                      | 2274 $\pm$ 85.0                               | 0 $\pm$ 0                                      | 2530 $\pm$ 363                                | 0 $\pm$ 0                                                                                |                  |      |          |

dependent signs. LUMINA has the most consistent improvements across the tested transfer grids, but these results do not establish a general transfer benefit from augmentation.

Table 12: Effect of transformation augmentation on other grids. Entries are average per-seed percent changes in equal-channel MAE, and **negative** values denote improvement under augmentation. An asterisk marks entries where all three seeds agree in sign. Case500 is a separate 200-source clean cohort with no scenario overlap with the main 50-source audit; evaluation paths also differ (Sec. F.4). The other grids receive no additional training.

| Grid     | MPNN  | LapPE+MPNN | RWSE+MPNN | LUMINA |
|----------|-------|------------|-----------|--------|
| case14   | −3.8  | −21.8*     | −4.3      | −11.0* |
| case118  | +11.3 | −2.8       | +13.5     | +14.2  |
| case500  | +4.8  | −0.4       | +0.4      | +3.4*  |
| case2000 | −3.9  | +7.7*      | −3.5*     | −5.7*  |
| case4661 | +0.2  | +13.4*     | +6.0      | −8.0*  |

#### F.4 Why the case500 evaluations differ

The main accuracy table uses the 50-source canonical subset of the transformed-input audit. The case500 row in Table 12 uses a separate 200-source sample, with no shared scenario IDs. Reaggregating the saved native-unit MAEs reproduces both tables. The cross-grid run records 24 checkpoint hashes and 24,000 model–scenario rows: four models, two arms, three seeds, five grids, and 200 scenarios. Its MPNN-family evaluation is batched; the transformed-input audit uses a different prediction path.

The LapPE baseline has mean canonical MAE 0.09983 in the main audit and 0.01253 in the cross-grid case500 sample. Augmented MAEs are 0.01309 and 0.01246. A matched follow-up identifies the eigensolver backend as the main cause. Training and cross-grid evaluation use batched single-precision CUDA `torch.linalg.eigh`; the main audit computes eigenvectors separately for each scenario using double-precision NumPy (LAPACK) `eigh` on CPU. Neither fixes eigenvector signs or bases in repeated eigenspaces.

We evaluated the same six checkpoints, with hashes matching the cross-grid run, on the same 200 case500 scenarios after changing only the eigensolver backend to CPU. Baseline MAE rises from 0.01253 to 0.08853 ( $7.1\times$ ; seed values 0.0797, 0.0992, and 0.0867), close to the main-audit value. Augmented MAE changes from 0.01246 to 0.01283 (+3.0%). Thus most of the large LapPE gain measures tolerance to CPU/LAPACK spectral coordinates after CUDA training. It does not demonstrate the same accuracy gain with the training backend, or a general repair of spectral encodings. The mismatch may also inflate the main audit’s baseline violation P95 and discrepancies, but this follow-up does not quantify those effects. The remaining cross-grid results retain their stated samples and sign convention.

## G Model design and evaluation

### G.1 Designing consistency by construction

After the output map is applied, an admitted relation compares two encodings of the same dispatch problem. A substantial prediction difference therefore comes from the encoding rather than a change in the physical problem. Low label error on one serialization can miss this dependence.

A relation enforced by construction holds for any learned weights, whether or not it appeared in training. The screen points to several useful design choices: positional summaries that do not depend on bus order, output partitions based on electrical roles, and symmetric handling of directed branch attributes. Preprocessing can standardize units, angle references, and row conventions. Changes in device count or type may require shared device encodings and additive aggregation at physical buses. Each proposed change must be checked against the relation it is meant to enforce.

Where an exact construction is unavailable, augmentation can reduce sensitivity, including for ELC relations that change graph size. Its effect depends on training. In our experiment, the aggregate discrepancy falls, but some relations worsen, canonical error can increase, and every prediction remains infeasible. Models therefore need separate checks of consistency, channel-level accuracy, AC feasibility, bounds, and dispatch cost.

### G.2 Testing multiple encodings of each scenario

Most learned-OPF benchmarks sample physical scenarios and evaluate one encoding of each. We propose evaluating  $K$  valid serializations per case and reporting both accuracy and operational metrics for those views, along with the discrepancy after mapping predictions back. The views share a physical scenario and must be treated as repeated measurements, not additional independent data points.

Table 13: Minimum reproducibility fields for a metamorphic learned-OPF benchmark.

| Record           | Minimum contents                                                                                                                                     |
|------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|
| Data and labels  | Release, split, topology membership, scenario identifier, units, angle-reference convention, and label-generation pipeline.                          |
| Relation         | Name, tier, parameters, applicability decision, transformed-case hash, and output map.                                                               |
| Solver control   | Status, objective difference, equality and inequality constraint violations, and admission decision for both views.                                  |
| Model            | Code and checkpoint version, checksums, preprocessing contract, feature order, normalization, and output semantics.                                  |
| Sampling         | Source-scenario count, sampled transformations, weight seeds, and the clustering unit used for uncertainty.                                          |
| Metrics          | Per-channel median, tail, and worst discrepancy plus accuracy, AC constraint violations, bounds, feasibility, and cost.                              |
| Negative records | Unsupported relations, failed solver admission checks, missing joins, infeasible predictions, and threshold saturation retained rather than dropped. |
| Artifact         | Case–relation–seed records, configurations, environment, and one command or script that regenerates tables and figures.                              |

At evaluation time, these transformations test whether the predictions and conclusions depend on the chosen encoding. At training time, they change the objective and may teach consistency. A pass by an augmented checkpoint supports a claim about that checkpoint; it cannot establish equivariance for all weights. Architectural claims still need analytical justification, with random-weight screens used to search for counterexamples.

### G.3 What an independent audit needs

Source code and a training recipe let an auditor test the architecture at random initialization. To test learned behavior, the auditor also needs the trained model, its preprocessing and normalization, the feature-column order, and a runnable path from a case file to outputs with known units and physical

identities. GridSFM-Open and LUMINA-2M provided these, making it possible to compare trained and random weights and locate specific failures [8–12].

An auditable release should also document training-topology membership and claimed invariances and include a worked numerical example. A versioned hosted endpoint can serve the same purpose if it accepts the documented inputs and exposes the required outputs; releasing the parameters is not essential. A benchmark split and preprocessing code alone are insufficient to test learned consistency.

Groups that cannot provide external inference access could run the public transformations internally and publish signed or hashed records of applicability, solver-control status, per-channel discrepancies, and failure tails. A common report format would make these results comparable without disclosing proprietary weights.

The evaluation record should retain unsupported relations, failed solver checks, infeasible predictions, and saturated feasibility thresholds. Omitting them can make a selectively populated table look more consistent than the tested model is.

## H Catalog coverage and experimental limits

The thirteen relations sample known families of equivalent AC-OPF inputs. We have not proved that they generate every equivalent representation. One valid failure refutes invariance to its relation; passing the catalog only rules out the counterexamples actually tested.

The random-weight screen can refute a guarantee for all weights with a single valid, above-tolerance counterexample. Finite passes cannot prove such a guarantee. Our reimplementations test specific model designs or components at random initialization, so they do not measure what the original implementation could learn or how accurately it would predict after training. A trained checkpoint's pass has the same finite scope (Sec. G.2).

The augmentation study uses one training grid family, three seeds, one compute budget, and model-specific relation sets. Some relations worsen, transfer to other grid sizes is mixed, and none of the predictions is fully AC feasible. These results do not justify operational use. PowerModels.jl and IPOPT supply high-accuracy numerical controls, not global-optimality certificates for nonconvex AC-OPF. Multiple local optima may differ in weakly identified quantities such as reactive dispatch. A discrepancy between equivalent inputs therefore establishes representation dependence but does not, on its own, show that both predictions are invalid. We measure consistency, label error, feasibility, bounds, and cost separately.

We found only two recent models accepting varying grid topologies with compatible trained checkpoints and executable inference. The other rows are public architectures, reconstructed designs, or mechanism controls at random initialization. Their results concern the tested code or component, not an unavailable trained original. Broader comparisons will need more runnable releases or comparable records from privately run models.

### H.1 Interpreting failures and passes

The 17-row screen gives at least six above-tolerance counterexamples per configuration. Each valid counterexample rules out all-input consistency for that relation. For the two released backbones, trained and random weights have the same thirteen threshold classifications of the source-median, worst-grid aggregate. These checkpoints change the magnitudes without eliminating the aggregate failures; the experiment does not rule out learning those relations through other training.

A pass means that no discrepancy was detected in the sampled scenarios, transformations, and seeds, unless a separate argument establishes invariance from the implementation. Consistency also says little about operational quality on its own: every prediction in the feasibility audit fails, despite reductions in representation sensitivity.

### H.2 Future work

The catalog could be extended with unequal component splits, inactive-bound changes, cost rescalings that preserve optimal dispatch, and bus-branch schema conversions. Each extension needs explicit validity conditions. Studying generators for these families and sampling their continuous parameters would help determine how much a finite catalog covers. Tolerances should also account for solver conditioning rather than relying only on one scalar threshold.

The next model comparisons should test augmentation against direct changes to reference and unit preprocessing, spectral features, physical-role output heads, branch handling, and co-located-device aggregation. These are candidate remedies, not validated repairs. Feasibility completion and correction should be tested alongside them because those steps may themselves break consistency. Larger training studies should vary grid families, compute budgets, and initializations, report uncertainty over source scenarios, and retain failures for individual relations.

Metamorphic consistency asks whether an encoding choice changes the prediction. Accuracy and feasibility ask whether the prediction is useful, while dispatch cost asks how economical it is. A learned-OPF evaluation needs all of these measurements.