RESCALING MATTERS: BEYOND PERMUTATION EQUIVARIANCE IN GRAPH NEURAL NETWORKS FOR LINEAR PROGRAMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph neural networks (GNNs) have proven efficient for solving optimization problems, particularly linear programming (LP), largely because they respect a fundamental symmetry: LPs are equivariant to permutations of their variables and constraints, and GNNs are permutation-equivariant by design. However, permutation symmetry represents only one class of important LP symmetries. LPs also remain equivalent under positive rescaling of the objective, constraint rows, and variables-a structure largely ignored by existing GNN-based solvers. We show this omission is structural: any traditional GNN that is exactly rescaling-equivariant must entirely discard the affected LP coefficients, which carry information essential for prediction. To overcome this limitation, we introduce canonicalization for LP, a model-agnostic pipeline that maps each instance to a single canonical representative of its rescaling-equivalent formulations, applies a permutation-equivariant GNN, and transforms the prediction back to the original coordinates. The resulting model is equivariant to both permutations and rescalings, without sacrificing the expressive power of the backbone. We prove that every LP admits such a canonicalization, and compute it via a unified energy-based objective, solved by smooth, unconstrained optimization with per-iteration cost linear in the number of nonzero coefficients. Finally, experiments across LP distributions and GNN backbones confirm our symmetry guarantees, and show that canonicalization generally improves in-distribution accuracy and consistently improves robustness to coefficient ranges unseen during training, at modest overhead.

1 INTRODUCTION

Exploiting the underlying symmetries of data provides a principled way to encode inductive biases for machine learning methods, often improving generalization and sample efficiency (Elesedy & Zaidi, 2021; Wang et al., 2020). GNNs provide a prominent example: their key structural advantage is respecting the permutation symmetry of graph-structured data (Scarselli et al., 2008). Many optimization problems, including LPs and Boolean satisfiability (SAT), admit natural graph representations (Zhang et al., 2024; Chen et al., 2023a; Gasse et al., 2019) , for which permuting variables or constraints preserves the underlying problem and simply relabels the solution accordingly. This makes GNNs a natural choice for learning solution maps for such problems, where they have shown promising empirical results (Li et al., 2026; Wu et al., 2024; Chen et al., 2026).

However, permutation symmetry captures only part of the symmetry structure of optimization problems. Real-world models often admit multiple equivalent parameterizations, with LPs providing a simple but fundamental example. Multiplying the objective by a positive scalar preserves the optimal solution set, rescaling each constraint by a positive constant preserves the feasible region, and changing variables, e.g., $x = Dx'$ for positive diagonal D, induces a corresponding transformation of optimal solutions.

This rescaling has long been central to the stability and efficiency of optimization algorithms. Poor scaling can impair simplex performance, making prescaling a standard remedy; related techniques are also used in interior-point and primal-dual hybrid gradient methods (Elble & Sahinidis, 2012; Tomlin, 2009; Czyzyk et al., 1999; Pock & Chambolle, 2011; Applegate et al., 2022). How-

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

ever, rescaling-related symmetries remain comparatively underexplored in learning to optimize, where GNN-based models typically encode permutation equivariance without explicitly incorporating rescaling invariance or equivariance. This motivates the question: *Can existing GNN-based LP solvers naturally respect these rescaling symmetries for better generalization?*

To answer the question, we show that exact enforcement faces a structural obstruction. For standard LP-GNNs, enforcing these symmetries, even locally, induces **degeneracy**: predictors must ignore essential LP coefficients, such as the objective vector or the constraint system. This limitation applies both to GNNs that directly predict solutions and to those that imitate specific optimization algorithms. It also helps explain their poor generalization to rescaled instances with unseen different numerical scales, as illustrated in Figure 1. Moreover, global rescaling equivariance forces even arbitrary continuous predictors to discard relevant LP coefficients. Thus, simply designing a new continuous architecture cannot recover the full rescaling inductive bias while retaining expressive power needed to solve LPs. The challenge is therefore to exploit rescaling symmetry for better generalization while preserving the expressive power of GNNs. We address this challenge through a model-agnostic canonicalization framework, with the following contributions:

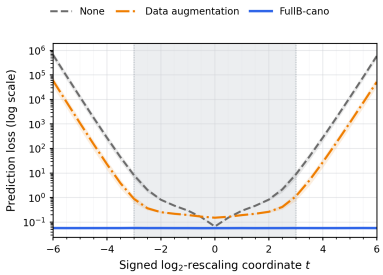

Figure 1: Prediction performance under rescaling. The shaded region marks the range of log-rescaling severities used to generate training augmentation samples.

▷ We identify rescaling symmetries as an overlooked inductive bias for learning LP solutions and establish degeneracy results that reveal a fundamental incompatibility between conventional GNNs and these symmetries.

▷ We prove the existence of a canonicalization that enforces rescaling invariance while preserving permutation equivariance. We then develop an efficient, model-agnostic construction and introduce a rescaling mask that together yield a model equivariant to both permutations and rescalings. Further, we show that the resulting canonicalization-based GNNs preserve the expressive power of their GNN backbones.

▷ Across LP distributions and GNN backbones, extended numerical experiments show accurate symmetry enforcement, better in-distribution generalization, and stronger scale-OOD robustness beyond exact rescaling changes, with modest overhead.

2 Related Work

Learning to optimize. GNN-based learning to optimize uses variable-constraint graphs to learn solver components, such as branching policies (Gasse et al., 2019; Gupta et al., 2020; Nair et al., 2020). For LP solution prediction, expressivity has been studied through WL-based representation theory (Chen et al., 2023a) and algorithm simulation or unrolling (Qian et al., 2023; Li et al., 2024a). These analyses provide complementary guarantees but do not automatically confer rescaling symmetry. Related symmetry-aware methods cover equivalent integer solutions (Chen et al., 2024a), exploit formulation automorphisms (Chen et al., 2025), or augment LPs and QPs with solution-preserving transformations (Qian & Morris, 2026b). However, almost all these works overlook rescaling symmetries despite their strong emphasis on permutation equivariance.

Geometric deep learning. Among the symmetry-enforcing methods, frame averaging symmetrizes a backbone using input-dependent transformations (Puny et al., 2021), while canonicalization selects representatives in the input space; their relationship is formalized by Ma et al. (2024). General energy minimization is an established canonicalization mechanism (Kaba et al., 2023), and Lie algebra canonicalization addresses noncompact actions and nonclosed orbits (Shumaylov et al., 2025). (Kalogeropoulos et al., 2024) represents neural networks as graphs and incorporates equivariance to neuron permutations and valid parameter rescalings into graph metanetworks. (Villar et al., 2023) enforces scaling symmetry through invariant inputs and restoration of output units . Related theories establish continuity obstructions for orbit selectors (Dym et al., 2024) and analyze the regularity and generalization of canonicalized predictors (Tahmasebi & Jegelka, 2025b;a). In contrast

to these works, we establish LP-specific failure theorems for direct symmetry enforcement, characterize compatibility between permutation and rescaling symmetries, and prove the existence of canonicalizations and the preservation of backbone universality.

Numerical prescaling for optimization. Prescaling is widely used in simplex and interior-point methods (Tomlin, 2009; Czyzyk et al., 1999), while diagonal preconditioning also supports first-order primal–dual methods (Pock & Chambolle, 2011; Applegate et al., 2022). Rather than seeking numerically favorable formulations to reduce iteration counts or stabilize algorithms, our canonicalization exploits rescaling symmetries to improve the generalization of deep-learning-based solvers.

3 PRELIMINARIES

Throughout this paper, we consider LPs in the general form

$$\min_{x\in\mathbb{R}^n} \; c^\top x \quad \text{s.t.} \quad Ax \le b, \;\; Cx = d, \;\; l \le x \le u, \tag{1}$$

and denote it by $X = (c, A, b, C, d, l, u)$. All such LPs with m inequality constraints and k equality constraints lie in the instance space $\mathbb{P}_{m,n,k} := \mathbb{R}^n \times \mathbb{R}^{m\times n} \times \mathbb{R}^m \times \mathbb{R}^{k\times n} \times \mathbb{R}^k \times [-\infty,+\infty)^n \times (-\infty,+\infty]^n$. There are two main solution mappings:

- $\Phi : \mathbb{P}_{m,n,k} \to 2^{\mathbb{R}^n}$ is the set-valued optimal solution mapping: $\Phi(X)$ is the optimal solution set of X if an optimum exists, and $\Phi(X) = \emptyset$ otherwise.
- $\Psi(X) := \arg\min_{y\in\Phi(X)} \|y\|_2$ is the minimal ℓ_2-norm optimal solution.

It is well established that LP solution mappings are *equivariant to permutations.* Let $\mathcal{K} := S_m \times S_k \times S_n$ be the permutation group, with permutations identified with their permutation matrices. Its action on $\mathbb{P}_{m,n,k}$ is

$$(P_m, P_k, P_n)\cdot(c, A, b, C, d, l, u) := (P_n c,\; P_m A P_n^\top,\; P_m b,\; P_k C P_n^\top,\; P_k d,\; P_n l,\; P_n u).$$

This action reorders the inequality constraints, equality constraints, and variables, modifying only the coordinate representation of the LP. For $S \subseteq \mathbb{R}^n$, define $P_n S := \{P_n x : x \in S\}$. Then both solution mappings are permutation equivariant: $\Phi\big((P_m, P_k, P_n)\cdot X\big) = P_n \Phi(X),\; \Psi\big((P_m, P_k, P_n)\cdot X\big) = P_n \Psi(X)$. Another important symmetry, largely overlooked in the learning to optimize literature, is ***positive diagonal rescaling***. Let $\mathcal{N} := \mathbb{R}_{++} \times \mathrm{Diag}_{m\times m}(\mathbb{R}_{++}) \times \mathrm{Diag}_{k\times k}(\mathbb{R}_{++}) \times \mathrm{Diag}_{n\times n}(\mathbb{R}_{++})$ be the rescaling group, acting on $\mathbb{P}_{m,n,k}$ by

$$(\alpha, R_{\le}, R_{=}, B)\cdot(c, A, b, C, d, l, u) := (\alpha B c,\; R_{\le} A B,\; R_{\le} b,\; R_{=} C B,\; R_{=} d,\; B^{-1} l,\; B^{-1} u).$$

Here, B corresponds to the change of variables $x = By$, $\alpha > 0$ rescales the objective, and $R_{\le}, R_{=}$ rescale the constraint rows. These transformations alter the representation but preserve the optimization problem up to the change of variables induced by B. Defining $B^{-1}S := \{B^{-1}x : x \in S\}$, the set-valued solution mapping is equivariant: $\Phi\big((\alpha, R_{\le}, R_{=}, B)\cdot X\big) = B^{-1}\Phi(X)$. For the minimal ℓ_2-norm selector Ψ, the situation is more delicate:

- Ψ is invariant to objective and constraint rescaling.
- Ψ is equivariant under scalar variable rescaling: $\Psi\big((\alpha, R_{\le}, R_{=}, \beta I)\cdot X\big) = \beta^{-1}\Psi(X)$.
- For a general diagonal B, Ψ is not equivariant, since the ℓ_2 norm is not preserved under non-uniform coordinate rescaling.

Thus, Φ respects the full rescaling symmetry, whereas Ψ respects only its objective, constraint, and scalar-variable components. Following (Berthold & Hendel, 2021), we restrict B and $R_{=}$ to diagonal matrices with strictly positive entries. This captures changes of scale while excluding sign flips: relative signs may encode semantic information (e.g., price is always positive) and therefore should not be regarded as LP symmetries. In practice, LPs are commonly represented as bipartite graphs (Gasse et al., 2019; Chen et al., 2023a), with one node per variable and one node per constraint:

- Each variable node j carries the feature (c_j, l_j, u_j).
- Each constraint node i carries the feature $(\mathbf{1}_{\le},\, \mathbf{1}_{=},\, b_i \text{ or } d_i)$, where $\mathbf{1}_{\le}$ and $\mathbf{1}_{=}$ are indicators of the constraint type.
- As $A_{ij} \neq 0$, variable node j is connected to constraint node i via an edge with weight A_{ij}.

The graphs are fed into specific GNN architectures for predicting the solution. Other graph representations for solving LPs have also been proposed (Qian et al., 2023; Li et al., 2024a); we discuss several of them in Appendix C. A more detailed preliminary is provided in Appendix A.

Table 1: Comparison of representative symmetry-enforcement mechanisms. Here $\mathcal{N}$ is the non-compact rescaling group and $\mathcal{K}$ is the permutation group already respected by the GNN backbone.

Method	Symmetry properties				Backbone evaluations	
	Exact Rescaling Symmetry	Applicable to non-compact $\mathcal{N}$	$\mathcal{K}$-equivariant	Model expressive power	Inference per instance	Training per sample
Designing continuous networks	✓	✓	✓	Degenerate[†]	$1\times$	$1\times$
Data augmentation	×	✓	✓	Universal	$1\times$	$K_{\mathrm{aug}}\times$
Group averaging	$\times^*$	$\times^*$	$\times^*$	$\times^*$	$\lvert G\rvert\times^*$	$\lvert G\rvert\times^*$
Frame averaging	△	△	△	△	$\lvert F(X)\rvert\times^*$	$\lvert F(X)\rvert\times^*$
Singleton canonicalization (ours)	✓	✓	✓	Universal	$\mathbf{1}\times$	$\mathbf{1}\times$

[†]: For traditional GNNs and even continuous architectures, exact rescaling enforcement is non-degenerately impossible by Theorem 4.1.
*: Standard group averaging based on normalized Haar measure is unavailable for the non-compact group $\mathcal{N}$, since its Haar measure cannot be normalized to have finite total mass; $|G|\times$ denotes the finite-group case. Typically, $1 < |F(X)| < |G|$. Moreover, $|F(X)|$ must be infinite for some X, see Corollary E.4.
△: Construction-dependent; the corresponding frame or symmetrization may not satisfy the required conditions.

4 Are GNNs Able to Yield Rescaling Symmetries?

GNN-based LP solvers are permutation-equivariant by construction, matching the permutation symmetry of LP solution mappings. Given the rescaling symmetries described in Section 3, a natural question is whether these models can also be invariant to objective and constraint-row rescaling and equivariant to variable rescaling. Surprisingly, we find that *exact enforcement comes at a fundamental cost: standard LP-GNNs (Chen et al., 2023a) must discard information essential for predicting optimal solutions.* For example, variable-rescaling equivariance forces the prediction to depend only on the variable bounds (l, u), regardless of the objective vector or constraint system. The model therefore produces identical predictions for LPs with the same bounds, even when their objectives or constraints yield different unique optimal solutions. Thus, enforcing rescaling symmetry within these architectures makes the model unable to distinguish not only equivalent formulations of the same problem, but also distinct problems that require different predictions. The following theorem makes this obstruction precise for ReLU-based LP-GNNs under each of the three types of rescaling.

Theorem 4.1. *Let $f : \mathbb{P}_{m,n,k} \to \mathbb{R}^n$ be any ReLU-based LP-GNN. Then:* (i) *invariance under positive objective rescaling $c \mapsto \alpha c$ forces f to be independent of c;* (ii) *invariance under positive constraint-row rescaling $(A, b; C, d) \mapsto (R_{\leq}A, R_{\leq}b, R_{=}C, R_{=}d)$ forces f to be independent of (A, b, C, d); and* (iii) *equivariance under positive diagonal variable rescaling $(c, A, C, l, u) \to (Bc, AB, CB, B^{-1}l, B^{-1}u)$ forces f to depend only on the variable bounds (l, u).* [1]

The proof of Theorem 4.1 extends to other common activations, e.g., tanh, sigmoid, GELU, and piecewise linear ones. Meanwhile, the same results hold even when the global symmetry enforcement is relaxed to local ones, as seen in Theorem C.16. For specific LP distributions (e.g., the standard form of LP), we obtain an even stronger collapse for general continuous functions (not merely GNN); see Corollary C.15. All proofs are deferred to Appendix C.

5 Canonicalization: Selecting Representatives from Orbits

Since requiring GNNs (or even general continuous functions in some cases) to satisfy exact rescaling symmetries leads to degenerate predictors, we naturally turn to *model-agnostic symmetrization methods*, summarized in Table 1. Data augmentation exposes the model to finitely many rescaled training samples and hence encourages, but does not guarantee, the desired symmetry (Chen et al., 2020). Group averaging can enforce exact symmetry by averaging predictions over the group, but normalized Haar averaging is unavailable for our non-compact rescaling group $\mathcal{N}$. Frame averaging instead uses an input-dependent subset $\mathcal{F}(X)$ of transformations (Puny et al., 2021), but its validity depends on the frame construction and it generally requires $|\mathcal{F}(X)|$ backbone evaluations.

Canonicalization avoids such averaging by selecting representatives directly from the rescaling orbit $\mathcal{N} \cdot X$, i.e., all equivalent numerical representations of the same LP under rescaling. Given a rescaling-invariant singleton representative $\mathcal{C}(X)$, let $g_X \in \mathcal{N}$ be a rescaling transformation that maps its coefficients back to those of the original LP, so that $X = g_X \cdot \mathcal{C}(X)$. Applying a backbone f to this representative and transforming its prediction back usually gives an equivariant model

[1] An analogous degeneracy result holds for scalar variable rescaling; see Corollary C.13.

$F(X) = \rho(g_X) f(\mathcal{C}(X))$, where $\rho(g_X)$ maps the predicted solution back to the original variable coordinates .[2] Thus, a singleton canonicalization requires only one backbone evaluation and avoids redundant evaluations caused by different frame elements producing the same representative (Ma et al., 2024). Since the GNN backbone already respects the permutation group $\mathcal{K}$, we further require canonicalization to preserve this symmetry: rescaling should leave the representative unchanged, whereas permuting variables or constraints should permute the representative accordingly.

Definition 5.1. *A mapping $\mathcal{C} : \mathbb{P}_{m,n,k} \to 2^{\mathbb{P}_{m,n,k}} \setminus \{\emptyset\}$ is a* permutation-equivariant rescaling orbit canonicalization *if, for every $X \in \mathbb{P}_{m,n,k}$, $g \in \mathcal{N}$, and $k \in \mathcal{K}$,*

$$\mathcal{C}(g \cdot X) = \mathcal{C}(X), \qquad \mathcal{C}(k \cdot X) = k \cdot \mathcal{C}(X), \qquad \mathcal{C}(X) \subseteq \mathcal{N} \cdot X.$$

An LP instance X is called canonicalizable *if such a canonicalization exists with $|\mathcal{C}(X)| = 1$.*

The key question is whether a single representative satisfying Definition 5.1 always exists. The difficulty is that a permutation and a rescaling can produce the same LP representation, creating a *collision*. For example, swapping $x_1 \le 1$ and $2x_1 \le 2$ is equivalent to multiplying these rows by 2 and $1/2$, respectively. Rescaling invariance and permutation equivariance then require the representative to be unchanged by this swap, so its two rows must be identical. More generally, the representative must remain unchanged under every permutation that can be reproduced by rescaling. A single representative must satisfy all such symmetry requirements simultaneously; the following theorem formally characterizes the relation between canonicalizability and this collision.

Theorem 5.2. *Let $\Sigma_X := \{k \in \mathcal{K} : k \cdot X \in \mathcal{N} \cdot X\}$. Then X is canonicalizable if and only if there exists $Y \in \mathcal{N} \cdot X$ such that $k \cdot Y = Y$ for all $k \in \Sigma_X$.*

For LPs, this condition can always be satisfied, as proved in Theorem D.4, which leads to:

Theorem 5.3. *Every linear program instance is canonicalizable. Moreover, there exists an optimal permutation-equivariant rescaling orbit canonicalization, where optimal means the least possible backbone evaluations to achieve exact equivariance are used.*[3]

6 Energy-Based Canonicalization: From Theory to Practice

Theorem 5.3 guarantees the existence of a canonical representative, but how can we find one in practice? We do so by minimizing an energy over the rescaling orbit. Let $H \le \mathcal{N}$ be a log-linear subgroup normalized by $\mathcal{K}$: $\mathrm{Log}(H)$ is a real linear subspace of $\mathbb{R}^{m+n+k+1}$ and $kHk^{-1} = H$ for every $k \in \mathcal{K}$. Log-linearity captures the usual closed, connected rescaling symmetries, which become real linear subspaces under taking component-wise logarithm; these include the rescaling group $\mathcal{N}$ and its scalar variable scaling subgroup $\mathcal{N}|_{B=\beta I}$. Normalization means that reordering an admissible scaling must leave it admissible. For a $\mathcal{K}$-equivariant feature map Θ_θ and a $\mathcal{K}$-invariant energy $\mathcal{E}_\theta$, define

$$C_\theta^H(X) := \operatorname*{arg\,min}_{Y \in H \cdot X} \mathcal{E}_\theta(\Theta_\theta(Y)) \,. \tag{2}$$

Whenever the minimum is attained at a unique representative for each X, C_θ^H is an H-invariant, $\mathcal{K}$-equivariant, single-valued orbit canonicalization. Thus, H determines the target symmetry compatible with $\mathcal{K}$, e.g., $\mathcal{N}$ or its scalar variable scaling subgroups. $(\Theta_\theta, \mathcal{E}_\theta)$ selects the representative.

FullB and scalar canonicalization. We develop the canonicalization energy based on two central design principles: (i) equilibrating the message-passing matrix (Diamond & Boyd, 2017) and (ii) stabilizing transformed coefficient magnitudes. We realize these principles through an equilibration term and quadratic log-magnitude penalties, yielding the two different canonicalization constructions below. Stack the constraint data as $\bar{A} = (A^\top, C^\top)^\top$ and $\bar{b} = (b^\top, d^\top)^\top$. Let $\Omega := \{(i,j) : \bar{A}_{ij} \neq 0\}$, $I_{\bar{A}} := \{i : \exists j \text{ such that } (i,j) \in \Omega\}$, $n_{\bar{A}} := |\Omega|$, $m_{\bar{A}} := |I_{\bar{A}}|$. For each connected component containing an edge in the bipartite support graph, let I_k, J_k be its constraint and variable sets, with $m_k = |I_k|$ and $n_k = |J_k|$. FullB-canonicalization uses $B = \mathrm{diag}(e^{\eta_1}, \dots, e^{\eta_n})$, $R = \mathrm{diag}(e^{r_1}, \dots, e^{r_{m+k}})$ and $y_{ij} = \log|\bar{A}_{ij}| + r_i + \eta_j$, its matrix energy is

$$E_A = \frac{\lambda_A}{m_{\bar{A}}} \left[\frac{1}{2} \sum_{(i,j)\in\Omega} e^{2y_{ij}} - \sum_{i \in I_{\bar{A}}} r_i - \sum_k \frac{m_k}{n_k} \sum_{j \in J_k} \eta_j \right] + \frac{\mu_A}{2n_{\bar{A}}} \sum_{(i,j)\in\Omega} y_{ij}^2. \tag{3}$$

[2] The transformation g_X need not be unique, and different valid choices can yield different restored predictions. We resolve this ambiguity by applying a rescaling mask P_X to the backbone output, giving $F(X) = \rho(g_X) P_X f(\mathcal{C}(X))$, which is equivariant for any valid choice of g_X; see Appendix E.4.

[3] These two theorems hold for scalar variable rescaling (i.e., $B = \beta I$) as well.

Ignoring the stabilizers, then the first order condition targets the uniform row norm 1 and compatible column norm $\sqrt{\frac{m_k}{n_k}}$. The β variant restricts $\eta_j = v$ for all j, giving $B = e^v I$ and reducing the component-weighted term to $m_{\bar{A}} v$.

The complete energy to be minimized is $E = E_A + E_c + E_b + E_l + E_u$. The four anchors are one half of their respective weights $\lambda_c, \lambda_b, \lambda_l, \lambda_u$ times the mean squared log-magnitude of the finite nonzero entries of the transformed vectors $c', \bar{b}', l', u'$. Primes denote coefficients after objective, row, and variable rescaling, with objective rescale parameter e^a.

For $\lambda_A, \mu_A > 0$ and positive weights on all nonempty anchor blocks, the objective equals an orbit energy satisfying Equation 2, up to a candidate-independent constant. Thus, we obtain the canonical representative by optimizing $r_i, \eta_j, a, \cdots$ and use them to rescale the LP for getting the representative, rather than directly optimizing over $H \cdot X$. This could let us get the canonicalization and the transporter g_X for restoring the prediction at the same time. Theorem E.7 establishes a unique minimizing representative $C_\theta^H(X)$ with the required symmetry and Borel measurability properties for the above constructions. Uniqueness concerns the representative, not necessarily the log-scale parameters $r_i, \eta_j, a, \cdots$. Though the transporters $r_i, \eta_j, a, \cdots$ may be non-unique, we could use the specifically designed rescaling mask P_X to ensure the final model equivariance and the Borel measurability of $\rho(g_X)P_X$ for any g_X, as proved in Appendix E.4. The rescaling mask becomes active only for LPs with nontrivial rescaling stabilizers, typically caused by degeneracy or redundancy. It thus provides a worst-case guarantee of exact equivariance, and whether it is needed can be efficiently determined from the LP's support structures, See remark E.11.

Computational cost. The energy is smooth and unconstrained in the independent log-scales. Each structured iteration requires $O(N_{\text{coef}})$ time and linear memory, where N_{coef} counts the finite nonzero coefficients in (A, b, C, d, c, l, u). This matches the per-pass complexity of conventional row/column prescaling. The total cost of computing and applying projection P_X is $O(N_{\text{coef}})$ too.

7 Canonicalization does not hurt The Expressive power of GNNs

Preserving provably approximately correct (PAC) universality under canonicalization is non-trivial because the guarantee of Chen et al. (2023a) controls approximation under a given probability measure. Canonical representatives may lie in a null set of the original measure, so the PAC bound need not control the inputs encountered after canonicalization. We overcome this using the Borel regular pushforward measure of C_θ^H and the measurable composite map $\rho_H(g_X)P_X$ in Section 6. Define $\mathcal{F}_{\text{LP}-\text{Cano}}^{\text{GNN}} = \left\{ X \mapsto \rho_H(g_X)P_X F_W(C_\theta^H(X)) \;\middle|\; F_W \in \mathcal{F}_{\text{LP}}^{\text{GNN}} \right\}$, where $X = g_X \cdot C_\theta^H(X)$, $\rho_H(\alpha, R_\leq, R_=, B)x = B^{-1}x$, and $\mathcal{F}_{\text{LP}}^{\text{GNN}}$ is the GNN in Chen et al. (2023a). Then, we have:

Theorem 7.1. *Let $\mathcal{P}$ be any Borel regular probability measure on $\mathbb{P}_{m,n,k}$ satisfying $\mathcal{P}[\Phi_{\text{obj}}(X) \in \mathbb{R}] = 1$. If $C_\theta^H(X)$ and $\rho_H(g_X)P_X$ are both Borel measurable, then, for every $\varepsilon, \delta > 0$, there exists $\widehat{F} \in \mathcal{F}_{\text{LP}-\text{Cano}}^{\text{GNN}}$ such that*

$$\mathcal{P}\left[\left\| \widehat{F}(X) - \rho_H(g_X)P_X \Psi(C_\theta^H(X)) \right\| > \delta \right] < \varepsilon.$$

Theorem E.7, E.10 provides Borel measurable C_θ^H and $\rho_H(g_X)P_X$. Moreover, as proved in Lemma F.1, $P_X\Psi(C_\theta^H(X)) = \Psi(C_\theta^H(X))$. Hence the theorem establishes PAC universality for optimal LP solutions due to the equivariance of Φ with respect to g_X . The same argument applies to other measurable canonicalizers and backbones with analogous PAC guarantees. This shows that *canonicalization preserves backbone expressive power while enforcing the prescribed symmetries.*

8 Experiments

Several numerical experiments are done to further validate our architecture's empirical ability. To be more precise, we aim to answer the following questions:

Q1: Do the proposed canonicalizations enforce their respective rescaling symmetries in finite-precision arithmetic while preserving permutation equivariance?

Q2: Do the proposed canonicalizations improve the base GNNs' in-distribution generalization on LPs, and enhance robustness to out-of-distribution numerical scale shifts beyond exact rescaling transformations?

Q3: How do the proposed canonicalizations compare with conventional solver-oriented prescaling in predictive performance and computational cost when used as preprocessing for GNN-based LP solvers?

Datasets and Experiment Design: We evaluate our canonicalization-based GNNs on nine different linear program distributions. Four of them are combinatorial optimization problems, which usually have good numerical coefficients. We adopt *set covering(SC), maximum independent set(IS), combinatorial auction(CA), and capacitated facility location(CFLP)* and get the LP dataset by using their relaxations. Meanwhile, we also leverage five real-world problems to further validate our method. We use the IEEE 57-bus, 118-bus, and 300-bus benchmark topologies as distributed in PGLib (Babaeinejadsarookolaee et al., 2019), and construct randomized DC-OPF LP instances on top of these fixed networks. Other real-world LP distributions include minimum-cost flow problems on strongly connected directed graphs (MCF) and randomly perturbed instances from the Netlib collection. We describe the dataset and parameters for dataset generation in Appendix G.2, G.3. We compare seven preprocessing/canonicalization choices: *none*, *data augmentation*, *HiGHS scaling*, *Ruiz Equilibration*, $(\alpha, R)-$*Normalization*, *fullB-cano*, and β-*cano*. The none baseline feeds the original LP representation to the GNN. The data augmentation generates four additional training samples using rescaling symmetry for each training sample, yielding five training instances in total. The *HiGHS* baseline applies HiGHS prescaling choice for traditional algorithms and then recovers the solution after GNN giving the prediction for the scaled problem. It is worth noticing that the HiGHS scaling also uses component-wise variable scaling besides rescaling rows. The $(\alpha, R)-$*Normalization* baseline applies row-wise coefficient normalization for the objective vector and constraint matrix. *Ruiz equilibration* (Ruiz, 2001) is a standard iterative diagonal-scaling procedure for improving conditioning in modern first-order LP solvers (Applegate et al., 2022). After ten Ruiz iterations, the final ℓ_∞ norm of every nonzero row and column across all instances from the nine distributions lies between 0.9917 and 1.0000. *FullB-cano* and β-*cano* are the canonicalizations using full diagonal and scalar variable scaling respectively in Section 6. For out-of-distribution evaluation, we test scale shifts on the five real-world LP distributions. The OOD instances keep the underlying structure unchanged but with coefficients sampled from a different numerical range. *Hence, it's much different to directly apply the rescaling symmetries on the original distribution.* For all the experiments, we train the neural networks with the AdamW optimizer with a default learning rate of 1e-3, and run for at most 400 epochs. LayerNorms are used to isolate the effect of numerical normalization. Following Chen et al. (2023a) and Liu et al. (2026), we use the relative solution error $\frac{\|x_{\text{GNN}} - x_{\text{solution}}\|^2}{\max\{\|x_{\text{solution}}\|^2, 1\}}$ as the prediction loss. During training, we leverage learning rate decay based on the validation relative solution distance, with a decay ratio of 0.5 and patience 50. All main experiments are averaged over three random seeds and run on two NVIDIA RTX Pro 5000 GPUs.

8.1 The ability to enforce the rescaling symmetries:

To answer **Q1**, we evaluate rescaling symmetry under scale ranges matched to data augmentation, testing objective, row, and variable rescalings both separately and jointly. For an LP instance P, let $x^\star$ denote its reference optimal solution and $\widehat{x}(P)$ the model prediction. Given a rescaling transformation $g = (\alpha, R_{\leq}, R_{=}, B)$, the transformed prediction is mapped back to the original variable space as $B\widehat{x}(g \cdot P)$. We measure both its accuracy relative to the reference solution and its consistency with the original prediction: $\mathcal{L}_{\text{orbit}}(g; P) = \frac{\|B\widehat{x}(g\cdot P) - x^\star\|_2^2}{\max\{\|x^\star\|_2^2, 1\}}$, $\mathcal{L}_{\text{cons}}(g; P) = \frac{\|B\widehat{x}(g\cdot P) - \widehat{x}(P)\|_2^2}{\max\{\|x^\star\|_2^2, 1\}}$. The label-orbit loss measures predictive accuracy under rescaling, whereas the prediction-consistency loss measures deviations from rescaling equivariance. Sampling details of g are provided in Appendix G.5.1.

Table 2: The rescale-orbit losses with respect to full rescaling group.

Backbone	Method	α	Row-R	Column-B	Joint
LP-GNN	None	0.5640,(0.4109)	1.9377,(1.7882)	11.7643,(11.6169)	14.8731,(14.7306)
	Data Augmentation	0.3089,(0.1412)	0.3246,(0.2046)	0.3270,(0.1837)	0.2797,(0.1699)
	HiGHS	0.5986,(0.4407)	7.4737,(7.3309)	9.3077,(9.1555)	534.0353,(533.8292)
	Ruiz Equilibration	1.0101,(0.8591)	0.5961,(0.4617)	0.5737,(0.4471)	0.8553,(0.7290)
	α,R-Normalization	0.1587,(**6.61e−13**)	0.1587,(**1.05e−12**)	15.2562,(15.1061)	15.2562,(15.1061)
	FullB-cano	**0.1026**,(3.55e−4)	**0.1027**,(6.68e−4)	**0.1029**,(**8.83e−4**)	**0.1028**,(**7.62e−4**)
	β-cano	0.1434,(2.86e−4)	0.1437,(7.84e−4)	2.7397,(2.6153)	2.7415,(2.6171)
Implicit Model	None	0.3637,(0.2891)	0.9596,(0.8918)	4.5280,(4.4644)	21.2507,(21.1926)
	Data Augmentation	0.2472,(0.0540)	0.2525,(0.0683)	0.6072,(0.4145)	2.1563,(1.9478)
	HiGHS	0.4046,(0.2391)	22.8638,(22.7034)	26.4553,(26.2926)	1528.5107,(1528.3389)
	Ruiz Equilibration	0.2757,(0.2127)	0.3971,(0.3375)	0.5126,(0.4553)	1.1163,(1.0592)
	α,R-Normalization	0.0845,(**2.43e−10**)	0.0845,(**3.91e−9**)	3.4322,(3.3518)	3.4322,(3.3518)
	FullB-cano	**0.0627**,(2.46e−4)	**0.0630**,(7.51e−4)	**0.0632**,(**9.90e−4**)	**0.0630**,(**7.04e−4**)
	β-cano	0.0798,(8.26e−5)	0.0799,(1.72e−4)	1.9067,(1.8421)	1.9080,(1.8434)

∘ Black values are label-orbit losses; blue parenthesized values are prediction-consistency losses.

As shown in Table 2, we compare the label-orbit and prediction-consistency losses for all seven methods, averaging over all datasets across two backbones: LP-GNNs (Chen et al., 2023a), and implicit GNNs (Liu et al., 2026). Here, α denotes pure objective rescaling, Row-R denotes pure constraint-row rescaling, Column-B denotes pure variable rescaling, and Joint applies all three

Table 3: Final training and validation-selected ID test losses for the LP-GNN backbone.

Final training loss

Dataset	None	Data Augmentation	HiGHS	Ruiz Equilibration	(α, R)-Normalization	FullB-cano	β-cano
CA	$0.2225_{\pm 0.0096}$	$0.2312_{\pm 0.0065}$	$0.2033_{\pm 0.0042}$	$0.2331_{\pm 0.0022}$	$0.1295_{\pm 0.0030}$	$\mathbf{0.1086}_{\pm 0.0013}$	$0.1209_{\pm 0.0033}$
SC	$0.1611_{\pm 0.0020}$	$0.2192_{\pm 0.0083}$	$0.1646_{\pm 0.0053}$	$0.1585_{\pm 0.0079}$	$0.1293_{\pm 0.0021}$	$\mathbf{0.0586}_{\pm 5.14e-4}$	$0.1118_{\pm 0.0050}$
OPF-57	$0.0016_{\pm 2.83e-4}$	$0.0083_{\pm 9.90e-4}$	$\mathbf{6.91e-4}_{\pm 5.63e-5}$	$8.82e-4_{\pm 2.14e-4}$	$0.0036_{\pm 0.0022}$	$0.0015_{\pm 3.81e-4}$	$0.0015_{\pm 3.60e-4}$
OPF-118	$0.0173_{\pm 0.0021}$	$0.0339_{\pm 5.10e-4}$	$0.0161_{\pm 0.0023}$	$0.0198_{\pm 2.07e-4}$	$0.0298_{\pm 8.03e-4}$	$\mathbf{0.0136}_{\pm 0.0010}$	$0.0242_{\pm 0.0033}$
OPF-300	$0.0598_{\pm 0.0080}$	$0.1014_{\pm 0.0051}$	$0.0538_{\pm 0.0037}$	$\mathbf{0.0402}_{\pm 0.0014}$	$0.0870_{\pm 0.0065}$	$0.0489_{\pm 0.0040}$	$0.0718_{\pm 0.0036}$
MCF	$0.0444_{\pm 0.0015}$	$0.0791_{\pm 0.0028}$	$0.0435_{\pm 0.0017}$	$\mathbf{0.0419}_{\pm 9.89e-4}$	$0.0504_{\pm 0.0014}$	$0.0510_{\pm 9.13e-4}$	$0.0465_{\pm 0.0023}$
CFLP	$0.1328_{\pm 0.0135}$	$0.1956_{\pm 0.0146}$	$0.0852_{\pm 0.0054}$	$\mathbf{0.0711}_{\pm 0.0058}$	$0.2595_{\pm 0.0237}$	$0.1085_{\pm 0.0132}$	$0.0968_{\pm 0.0030}$
IS	$0.4204_{\pm 5.63e-6}$	$0.2127_{\pm 0.0047}$	$0.4204_{\pm 7.32e-6}$	$0.4204_{\pm 4.44e-5}$	$0.4204_{\pm 1.55e-5}$	$\mathbf{0.1322}_{\pm 4.80e-4}$	$0.4204_{\pm 3.72e-5}$
NETLIB	$0.0339_{\pm 0.0086}$	$0.0870_{\pm 0.0048}$	$0.0237_{\pm 0.0028}$	$0.0399_{\pm 0.0045}$	$0.1165_{\pm 0.0118}$	$\mathbf{0.0192}_{\pm 0.0027}$	$0.0316_{\pm 0.0044}$

Validation-selected ID test loss

Dataset	None	Data Augmentation	HiGHS	Ruiz Equilibration	(α, R)-Normalization	FullB-cano	β-cano
CA	$0.2263_{\pm 0.0060}$	$0.2565_{\pm 0.0018}$	$0.2168_{\pm 0.0076}$	$0.2268_{\pm 0.0040}$	$0.1301_{\pm 9.49e-4}$	$\mathbf{0.1176}_{\pm 5.92e-4}$	$0.1250_{\pm 6.55e-4}$
SC	$0.1701_{\pm 0.0021}$	$0.2120_{\pm 0.0043}$	$0.1749_{\pm 0.0017}$	$0.1720_{\pm 0.0012}$	$0.1280_{\pm 0.0038}$	$\mathbf{0.0748}_{\pm 5.05e-4}$	$0.1230_{\pm 8.71e-4}$
OPF-57	$0.0202_{\pm 5.07e-4}$	$0.0238_{\pm 9.84e-4}$	$0.0200_{\pm 0.0010}$	$0.0108_{\pm 4.00e-4}$	$0.0134_{\pm 0.0024}$	$\mathbf{0.0070}_{\pm 3.51e-4}$	$0.0131_{\pm 4.24e-4}$
OPF-118	$0.0887_{\pm 0.0011}$	$0.0877_{\pm 0.0020}$	$0.0951_{\pm 0.0018}$	$0.0783_{\pm 0.0020}$	$0.0877_{\pm 2.65e-4}$	$\mathbf{0.0738}_{\pm 0.0011}$	$0.0834_{\pm 6.31e-4}$
OPF-300	$0.1645_{\pm 0.0011}$	$0.1701_{\pm 0.0012}$	$0.1824_{\pm 5.88e-4}$	$0.1409_{\pm 2.85e-4}$	$0.1571_{\pm 0.0019}$	$\mathbf{0.1262}_{\pm 0.0015}$	$0.1501_{\pm 0.0022}$
MCF	$0.0789_{\pm 2.98e-4}$	$0.0841_{\pm 3.78e-4}$	$0.0785_{\pm 3.93e-4}$	$0.0790_{\pm 4.09e-4}$	$0.0758_{\pm 6.45e-5}$	$0.0740_{\pm 4.38e-4}$	$\mathbf{0.0738}_{\pm 2.23e-4}$
CFLP	$0.2119_{\pm 0.0132}$	$0.2061_{\pm 0.0120}$	$0.1991_{\pm 0.0017}$	$\mathbf{0.1702}_{\pm 4.82e-4}$	$0.2557_{\pm 0.0196}$	$0.1771_{\pm 0.0126}$	$0.1713_{\pm 0.0032}$
IS	$0.4205_{\pm 4.96e-7}$	$0.4205_{\pm 3.91e-7}$	$0.4205_{\pm 2.86e-7}$	$0.4205_{\pm 7.57e-7}$	$0.4205_{\pm 7.49e-7}$	$\mathbf{0.1382}_{\pm 8.95e-5}$	$0.4205_{\pm 2.44e-7}$
NETLIB	$0.1680_{\pm 0.0035}$	$0.1657_{\pm 0.0025}$	$0.1827_{\pm 0.0057}$	$0.1582_{\pm 0.0027}$	$0.1596_{\pm 0.0058}$	$0.1342_{\pm 0.0023}$	$\mathbf{0.1300}_{\pm 0.0025}$

transformations together. FullB-based canonicalization almost performs best across backbones and metrics, achieving much lower label-orbit losses and nearly zero prediction-consistency loss. Both canonicalization methods usually outperform baselines that ignore part of the symmetry, such as Ruiz equilibration, HiGHS, and α, R-normalization in the label-orbit losses. Data augmentation provides some improvement in symmetry learning, but requires five times as much training data and is still outperformed by FullB-based canonicalization. Since the rescaling group is infinite and non-compact, finite augmentation cannot cover all possible transformations, highlighting the need for exact canonicalization. Meanwhile, we also test the permutation equivariance of the canonicalization-based methods, as shown in Table 7. All these results show that *canonicalization-based methods could both enforce rescaling symmetry in finite precision and preserve permutation equivariance of backbones, while other baselines fail.*

8.2 The expressive power of canonicalization-based GNNs:

To answer **Q2, Q3**, we evaluate all the methods on nine datasets across two different backbones. We further examine scale-OOD generalization on the real-world distributions, for which natural test distributions with different coefficient ranges can be generated without directly rescaling ID instances. As shown in Table 3 and the corresponding implicit-GNN results in Appendix G.5, the canonicalization methods, particularly FullB-cano, generally achieve much lower final training and validation-selected ID test losses. These results suggest that canonicalization facilitates learning by removing nuisance scaling degrees of freedom while preserving task-relevant information: once the backbone learns to predict on a canonical representative, its prediction transfers to the entire rescaling orbit by construction. HiGHS scaling can sometimes improve training, but generalizes less reliably because it is only a solver-oriented heuristic; it may improve numerical conditioning while shifting the backbone's data distribution. Ruiz equilibration is also outperformed by canonicalization, showing that the gains come from exploiting symmetry, not merely better numerical conditions. Data augmentation improves rescaling consistency, but does not consistently improve in-distribution prediction accuracy across the evaluated datasets and backbones, and sometimes weakens the model's expressive power, which aligns with our degeneracy theory. This degeneracy may also explain the seemingly surprising observation that, under data augmentation, the training loss is sometimes higher than the test loss: the GNNs struggle to fit samples across the rescaling orbit, and the training set—unlike the test set—contains such rescaled instances. $\alpha, R-$Normalization usually improves the model's performance; however, the performance improvement is much lower than that of canonicalizations. This shows that the additional change-of-variable symmetry also plays a significant role in learning. Consistently, full vector variable-scaling canonicalization almost always outperforms its scalar-scaling counterpart. Together, these results demonstrate the importance of accounting for the full rescaling group and further validate our theory.

Moreover, the canonicalization methods consistently dominate on the scale-OOD tests, even though these test distributions are generated by sampling coefficients over different numerical ranges from the original distribution rather than by directly applying rescaling transformations.

This demonstrates that our method provides not only exact orbit consistency but also numerical robustness under genuine distribution shifts. In summary, it acts as a *distributional canonicalization* for learning-based backbones, rather than merely as an *algorithmic preconditioner* designed for traditional optimization algorithms.

Table 4: ID and OOD test losses for LP-GNN.

Dataset	Method	ID test	OOD-1	OOD-2	OOD-3
OPF-57	None	$0.0202_{\pm 5.07e-4}$	$0.2310_{\pm 0.0126}$	$0.5433_{\pm 0.0124}$	$0.7552_{\pm 0.0082}$
	Data Augmentation	$0.0238_{\pm 9.84e-4}$	$0.0748_{\pm 0.0082}$	$0.0939_{\pm 0.0033}$	$0.2864_{\pm 0.0259}$
	HiGHS	$0.0200_{\pm 0.0010}$	$0.1129_{\pm 0.0061}$	$0.2584_{\pm 0.0121}$	$0.4544_{\pm 0.0114}$
	Ruiz Equilibration	$0.0108_{\pm 4.00e-4}$	$0.3226_{\pm 0.0053}$	$0.6249_{\pm 0.0045}$	$0.8056_{\pm 0.0038}$
	(α, R)-Normalization	$0.0134_{\pm 0.0024}$	$0.2547_{\pm 0.0234}$	$0.6017_{\pm 0.0264}$	$0.7914_{\pm 0.0228}$
	FullB-cano	$\mathbf{0.0070}_{\pm 3.51e-4}$	$\mathbf{0.0440}_{\pm 4.59e-4}$	$\mathbf{0.0378}_{\pm 1.84e-4}$	$\mathbf{0.0654}_{\pm 9.83e-4}$
	β-cano	$0.0131_{\pm 4.24e-4}$	$0.0497_{\pm 9.39e-4}$	$0.0447_{\pm 4.45e-4}$	$0.0704_{\pm 7.08e-4}$
OPF-118	None	$0.0887_{\pm 0.0011}$	$0.3574_{\pm 0.0045}$	$0.6153_{\pm 0.0050}$	$0.8049_{\pm 0.0052}$
	Data Augmentation	$0.0877_{\pm 0.0020}$	$0.2525_{\pm 0.0049}$	$0.2818_{\pm 0.0051}$	$0.4444_{\pm 0.0564}$
	HiGHS	$0.0951_{\pm 0.0018}$	$0.3012_{\pm 0.0040}$	$0.4494_{\pm 0.0106}$	$0.6112_{\pm 0.0168}$
	Ruiz Equilibration	$0.0783_{\pm 0.0020}$	$0.4355_{\pm 0.0065}$	$0.6678_{\pm 0.0062}$	$0.8250_{\pm 0.0041}$
	(α, R)-Normalization	$0.0877_{\pm 2.65e-4}$	$0.3412_{\pm 0.0049}$	$0.5825_{\pm 0.0029}$	$0.7822_{\pm 0.0026}$
	FullB-cano	$\mathbf{0.0738}_{\pm 0.0011}$	$\mathbf{0.1675}_{\pm 0.0028}$	$\mathbf{0.1647}_{\pm 0.0023}$	$\mathbf{0.1770}_{\pm 5.58e-4}$
	β-cano	$0.0834_{\pm 6.31e-4}$	$0.1882_{\pm 0.0036}$	$0.1876_{\pm 0.0029}$	$0.1973_{\pm 0.0037}$
OPF-300	None	$0.1645_{\pm 0.0011}$	$0.3227_{\pm 0.0049}$	$0.5726_{\pm 0.0037}$	$0.7529_{\pm 0.0038}$
	Data Augmentation	$0.1701_{\pm 0.0012}$	$0.2548_{\pm 0.0045}$	$0.3387_{\pm 0.0071}$	$0.4560_{\pm 0.0073}$
	HiGHS	$0.1824_{\pm 5.88e-4}$	$0.3069_{\pm 0.0050}$	$0.4537_{\pm 0.0137}$	$0.6114_{\pm 0.0179}$
	Ruiz Equilibration	$0.1409_{\pm 2.85e-4}$	$0.3422_{\pm 0.0025}$	$0.5935_{\pm 0.0066}$	$0.7707_{\pm 0.0086}$
	(α, R)-Normalization	$0.1571_{\pm 0.0019}$	$0.3330_{\pm 0.0028}$	$0.5794_{\pm 0.0024}$	$0.7634_{\pm 0.0013}$
	FullB-cano	$\mathbf{0.1262}_{\pm 0.0015}$	$\mathbf{0.1480}_{\pm 0.0030}$	$\mathbf{0.1718}_{\pm 0.0026}$	$\mathbf{0.1926}_{\pm 0.0041}$
	β-cano	$0.1501_{\pm 0.0022}$	$0.1763_{\pm 0.0028}$	$0.2089_{\pm 0.0034}$	$0.2383_{\pm 0.0035}$
MCF	None	$0.0789_{\pm 2.98e-4}$	$0.3037_{\pm 0.0063}$	$0.5811_{\pm 0.0171}$	$0.7697_{\pm 0.0257}$
	Data Augmentation	$0.0841_{\pm 3.78e-4}$	$0.1327_{\pm 0.0019}$	$0.1946_{\pm 0.0022}$	$0.3159_{\pm 0.0183}$
	HiGHS	$0.0785_{\pm 3.93e-4}$	$0.3158_{\pm 0.0175}$	$0.6960_{\pm 0.0120}$	$0.8769_{\pm 0.0112}$
	Ruiz Equilibration	$0.0790_{\pm 4.09e-4}$	$0.1689_{\pm 0.0042}$	$0.5396_{\pm 0.0267}$	$0.7780_{\pm 0.0234}$
	(α, R)-Normalization	$0.0758_{\pm 6.45e-5}$	$0.1481_{\pm 0.0030}$	$0.4157_{\pm 0.0304}$	$0.7429_{\pm 0.0544}$
	FullB-cano	$0.0740_{\pm 4.38e-4}$	$\mathbf{0.0996}_{\pm 0.0010}$	$\mathbf{0.1141}_{\pm 4.20e-4}$	$\mathbf{0.1276}_{\pm 6.08e-4}$
	β-cano	$\mathbf{0.0738}_{\pm 2.23e-4}$	$0.1073_{\pm 0.0014}$	$0.1257_{\pm 0.0029}$	$0.1375_{\pm 0.0045}$
NETLIB	None	$0.1680_{\pm 0.0035}$	$0.5187_{\pm 0.0501}$	$2.6592_{\pm 0.2972}$	$31.6674_{\pm 20.5834}$
	Data Augmentation	$0.1657_{\pm 0.0025}$	$0.4074_{\pm 0.0227}$	$1.4143_{\pm 0.0976}$	$4.2451_{\pm 0.3578}$
	HiGHS	$0.1827_{\pm 0.0057}$	$0.4979_{\pm 0.0337}$	$1.9004_{\pm 0.1558}$	$6.8881_{\pm 0.8570}$
	Ruiz Equilibration	$0.1582_{\pm 0.0027}$	$0.5364_{\pm 0.0215}$	$3.4833_{\pm 0.2533}$	$20.9035_{\pm 4.8904}$
	(α, R)-Normalization	$0.1596_{\pm 0.0058}$	$0.5004_{\pm 0.0608}$	$5.4590_{\pm 2.9922}$	$31.1802_{\pm 15.0058}$
	FullB-cano	$0.1342_{\pm 0.0023}$	$\mathbf{0.1496}_{\pm 0.0035}$	$\mathbf{0.1956}_{\pm 0.0077}$	$\mathbf{0.3155}_{\pm 0.0151}$
	β-cano	$\mathbf{0.1300}_{\pm 0.0025}$	$0.2327_{\pm 0.0030}$	$0.2722_{\pm 0.0063}$	$0.3415_{\pm 0.0180}$

8.3 The computational overhead comparison

Regarding the computational cost, we report preprocessing time per test instance, the average training time per epoch, and end-to-end inference time in Table 5. Canonicalization incurs only modest additional training and inference costs, arising from computing the canonical representative before the backbone forward pass, applying a rescaling mask, and restoring the prediction to the original variable space afterward. Its preprocessing cost remains on the same millisecond scale as solver-oriented prescaling methods such as HiGHS and Ruiz Equilibration, while its training overhead is substantially lower than that of data augmentation that uses a fivefold training set. These empirical results are consistent with our linear-cost analysis and with the fact that the proposed single-valued canonicalization requires only one backbone forward pass. Thus, canonicalization delivers exact rescaling symmetry and the aforementioned ID and scale-OOD generalization benefits at a practical computational cost, providing a computationally efficient, scalable and model-agnostic approach to symmetry enforcement.

Table 5: Computational cost averaged over all datasets.

Method	Preprocessing (ms/instance)	Training time per epoch (s)		Inference time (ms/instance)	
		LP-GNN	Implicit Model	LP-GNN	Implicit Model
None	**0.0000**	**0.5781**	**0.9291**	5.7824	**8.8301**
Data Augmentation	**0.0000**	1.5477	3.4570	**5.7634**	8.8424
HiGHS	2.9994	0.6325	1.0047	9.4145	12.4547
Ruiz Equilibration	3.4338	0.6338	1.0326	9.9433	13.0238
(α, R)-Normalization	1.7413	0.6194	0.9960	7.9878	11.0766
FullB-cano	4.0337	0.6231	0.9959	10.3866	13.4639
β-cano	2.8228	0.6223	1.0095	9.1171	12.2072

9 Conclusion

We identify rescaling symmetry as a missing inductive bias for learning solutions of LPs, and found that exact rescaling equivariance/invariance forces continuous predictors to discard essential coefficients, while rescaling equivariance/invariance even locally similarly degenerates standard LP-GNNs. We then characterize canonicalizability via permutation–rescaling compatibility, prove that every LP admits a single-valued orbit canonicalization respecting both symmetries, and develop an energy-based construction with a rescaling mask to ensure the final equivariance. Under mild conditions, the canonicalization is globally Borel measurable, enabling the canonicalized LP-GNNs to retain backbone PAC universality with both desired symmetries. Across several LP distributions and backbones, experiments show accurate symmetry enforcement, generally improved training and in-distribution performance, and stronger scale-OOD generalization beyond exact rescaling, with modest overhead. Canonicalization thus provides model-agnostic distributional symmetry reduction beyond solver-oriented numerical preconditioning.

Future work includes extending the framework to quadratic, conic, mixed-integer, and nonlinear optimization; jointly learning task-adaptive energies and backbones while preserving symmetry and regularity guarantees; and selecting rescaling subgroups H (Section 6) and canonicalizations for structured LP families.

REFERENCES

Erling D Andersen and Knud D Andersen. Presolving in linear programming. *Mathematical programming*, 71(2):221–245, 1995.

David Applegate, Mateo Díaz, Oliver Hinder, Haihao Lu, Miles Lubin, Brendan O'Donoghue, and Warren Schudy. Practical large-scale linear programming using primal-dual hybrid gradient, 2022. URL https://arxiv.org/abs/2106.04756.

David Applegate, Oliver Hinder, Haihao Lu, and Miles Lubin. Faster first-order primal-dual methods for linear programming using restarts and sharpness. *Mathematical Programming*, 201(1):133–184, 2023.

Sogol Babaeinejadsarookolaee, Adam Birchfield, Richard D Christie, Carleton Coffrin, Christopher DeMarco, Ruisheng Diao, Michael Ferris, Stephane Fliscounakis, Scott Greene, Renke Huang, et al. The power grid library for benchmarking ac optimal power flow algorithms. *arXiv preprint arXiv:1908.02788*, 2019.

Timo Berthold and Gregor Hendel. Learning to scale mixed-integer programs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):3661–3668, May 2021. doi: 10.1609/aaai.v35i5.16482. URL https://ojs.aaai.org/index.php/AAAI/article/view/16482.

AL Brearley, Gautam Mitra, and H Paul Williams. Analysis of mathematical programming problems prior to applying the simplex algorithm. *Mathematical programming*, 8(1):54–83, 1975.

Hao Chen, Chendi Qian, Christopher Morris, Andrea Lodi, and Can Li. Solving max-cut to global optimality via feasibility-preserving graph neural networks. *arXiv preprint arXiv:2605.07113*, 2026.

Qian Chen, Tianjian Zhang, Linxin Yang, Qingyu Han, Akang Wang, Ruoyu Sun, Xiaodong Luo, and Tsung-Hui Chang. Symilo: A symmetry-aware learning framework for integer linear optimization. *Advances in Neural Information Processing Systems*, 37:24411–24434, 2024a.

Qian Chen, Lei Li, Qian Li, Jianghua Wu, Akang Wang, Ruoyu Sun, Xiaodong Luo, Tsung-Hui Chang, and Qingjiang Shi. When gnns meet symmetry in ilps: an orbit-based feature augmentation approach. In *International Conference on Learning Representations*, volume 2025, pp. 56982–56997, 2025.

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.

Ziang Chen, Jialin Liu, Xinshang Wang, Jianfeng Lu, and Wotao Yin. On representing linear programs by graph neural networks, 2023a. URL https://arxiv.org/abs/2209.12288.

Ziang Chen, Jialin Liu, Xinshang Wang, Jianfeng Lu, and Wotao Yin. On representing mixed-integer linear programs by graph neural networks. 2023b.

Ziang Chen, Xiaohan Chen, Jialin Liu, Xinshang Wang, and Wotao Yin. Expressive power of graph neural networks for (mixed-integer) quadratic programs. *arXiv preprint arXiv:2406.05938*, 2024b.

Ziang Chen, Jialin Liu, Xiaohan Chen, Xinshang Wang, and Wotao Yin. Rethinking the capacity of graph neural networks for branching strategy. *Advances in neural information processing systems*, 37:123991–124024, 2024c.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Joseph Czyzyk, Sanjay Mehrotra, Michael Wagner, and Stephen J Wright. Pcx: An interior-point code for linear programming. *Optimization Methods and Software*, 11(1-4):397–430, 1999.

Steven Diamond and Stephen Boyd. Stochastic matrix-free equilibration. *Journal of Optimization Theory and Applications*, 172(2):436–454, 2017.

Nadav Dym, Hannah Lawrence, and Jonathan W Siegel. Equivariant frames and the impossibility of continuous canonicalization. *arXiv preprint arXiv:2402.16077*, 2024.

Joseph M Elble and Nikolaos V Sahinidis. Scaling linear optimization problems prior to application of the simplex method. *Computational Optimization and Applications*, 52(2):345–371, 2012.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *International conference on machine learning*, pp. 2959–2969. PMLR, 2021.

Maxime Gasse, Didier Chételat, Nicola Ferroni, Laurent Charlin, and Andrea Lodi. Exact combinatorial optimization with graph convolutional neural networks. *Advances in neural information processing systems*, 32, 2019.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th international conference on international conference on machine learning*, pp. 399–406, 2010.

Prateek Gupta, Maxime Gasse, Elias Khalil, Pawan Mudigonda, Andrea Lodi, and Yoshua Bengio. Hybrid models for learning to branch. *Advances in neural information processing systems*, 33: 18087–18097, 2020.

Samar Hadou and Alejandro Ribeiro. Unrolled neural networks for constrained optimization. *arXiv preprint arXiv:2601.17274*, 2026.

Qingyu Han, Linxin Yang, Qian Chen, Xiang Zhou, Dong Zhang, Akang Wang, Ruoyu Sun, and Xiaodong Luo. A gnn-guided predict-and-search framework for mixed-integer linear programming. *arXiv preprint arXiv:2302.05636*, 2023.

Yu He and Ellen Vitercik. Primal-dual neural algorithmic reasoning. *arXiv preprint arXiv:2505.24067*, 2025.

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *International Conference on Machine Learning*, pp. 15546–15566. PMLR, 2023.

Ioannis Kalogeropoulos, Giorgos Bouritsas, and Yannis Panagakis. Scale equivariant graph metanetworks, 2024. URL https://arxiv.org/abs/2406.10685.

Elias B Khalil, Christopher Morris, and Andrea Lodi. Mip-gnn: A data-driven framework for guiding combinatorial solvers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10219–10227, 2022.

Jinwoo Kim, Dat Nguyen, Ayhan Suleymanzade, Hyeokjun An, and Seunghoon Hong. Learning probabilistic symmetrization for architecture agnostic equivariance. *Advances in Neural Information Processing Systems*, 36:18582–18612, 2023.

Bingheng Li, Linxin Yang, Yupeng Chen, Senmiao Wang, Qian Chen, Haitao Mao, Yao Ma, Akang Wang, Tian Ding, Jiliang Tang, and Ruoyu Sun. Pdhg-unrolled learning-to-optimize method for large-scale linear programming, 2024a. URL https://arxiv.org/abs/2406.01908.

Qian Li, Tian Ding, Linxin Yang, Minghui Ouyang, Qingjiang Shi, and Ruoyu Sun. On the power of small-size graph neural networks for linear programming. *Advances in Neural Information Processing Systems*, 37:38695–38719, 2024b.

Qian Li, Minghui Ouyang, Tian Ding, Yuyi Wang, Qingjiang Shi, and Ruoyu Sun. Towards explaining the power of constant-depth graph neural networks for structured linear programming. In *International Conference on Learning Representations*, volume 2025, pp. 86353–86368, 2025.

Ruizhe Li, Enming Liang, and Minghua Chen. On the universality and complexity of gnn for solving second-order cone programs. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ya-Wei Eileen Lin and Ron Levie. Adaptive canonicalization with application to invariant anisotropic geometric networks. In *International Conference on Learning Representations*, volume 2026, pp. 23056–23099, 2026.

Yuchao Lin, Jacob Helwig, Shurui Gui, and Shuiwang Ji. Equivariance via minimal frame averaging for more symmetries and efficiency. *arXiv preprint arXiv:2406.07598*, 2024.

Jialin Liu, Lisang Ding, Stanley Osher, and Wotao Yin. Expressive power of implicit models: Rich equilibria and test-time scaling. In *The Fourteenth International Conference on Learning Representations*, 2026.

George Ma, Yifei Wang, Derek Lim, Stefanie Jegelka, and Yisen Wang. A canonicalization perspective on invariant and equivariant learning. *Advances in Neural Information Processing Systems*, 37:60936–60979, 2024.

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. *arXiv preprint arXiv:1812.09902*, 2018.

Boris Mityagin. The zero set of a real analytic function. *arXiv preprint arXiv:1512.07276*, 2015.

Vishal Monga, Yuelong Li, and Yonina C Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.

Vinod Nair, Sergey Bartunov, Felix Gimeno, Ingrid Von Glehn, Pawel Lichocki, Ivan Lobov, Brendan O'Donoghue, Nicolas Sonnerat, Christian Tjandraatmadja, Pengming Wang, et al. Solving mixed integer programs using neural networks. *arXiv preprint arXiv:2012.13349*, 2020.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 2006.

Alex Oshin, Rahul Vodeb Ghosh, and Evangelos A Theodorou. Scalable deep unfolding of conic optimizers. *arXiv preprint arXiv:2606.13825*, 2026.

Thomas Pock and Antonin Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *2011 International Conference on Computer Vision*, pp. 1762–1769. IEEE, 2011.

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. *arXiv preprint arXiv:2110.03336*, 2021.

Chendi Qian and Christopher Morris. Towards graph neural networks for provably solving convex optimization problems. *arXiv preprint arXiv:2502.02446*, 2025.

Chendi Qian and Christopher Morris. On the expressive power of gnns to solve linear sdps. *arXiv preprint arXiv:2604.27786*, 2026a.

Chendi Qian and Christopher Morris. Principled data augmentation for learning to solve quadratic programming problems. *Advances in Neural Information Processing Systems*, 38:132937–132969, 2026b.

Chendi Qian, Didier Chételat, and Christopher Morris. Exploring the power of graph neural networks in solving linear optimization problems, 2023. URL `https://arxiv.org/abs/2310.10603`.

Daniel Ruiz. A scaling algorithm to equilibrate both rows and columns norms in matrices. *Institut National Polytechnique de Toulouse, Toulouse,...*, 2001.

Rajiv Sambharya, Georgina Hall, Brandon Amos, and Bartolomeo Stellato. End-to-end learning to warm-start for real-time quadratic optimization. In *Learning for dynamics and control conference*, pp. 220–234. PMLR, 2023.

Rajiv Sambharya, Georgina Hall, Brandon Amos, and Bartolomeo Stellato. Learning to warm-start fixed-point optimization algorithms. *Journal of Machine Learning Research*, 25(166):1–46, 2024.

Augustinos Saravanos, Hunter Kuperman, Alex Oshin, Arshiya Taj Abdul, Vincent Pacelli, and Evangelos Theodorou. Deep distributed optimization for large-scale quadratic programming. In *International Conference on Learning Representations*, volume 2025, pp. 93580–93619, 2025.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

Zakhar Shumaylov, Peter Zaika, James Rowbottom, Ferdia Sherry, Melanie Weber, and Carola-Bibiane Schönlieb. Lie algebra canonicalization: Equivariant neural operators under arbitrary lie groups. In *International Conference on Learning Representations*, volume 2025, pp. 1809–1857, 2025.

Behrooz Tahmasebi and Stefanie Jegelka. Generalization bounds for canonicalization: A comparative study with group averaging. In *International Conference on Learning Representations*, volume 2025, pp. 72051–72070, 2025a.

Behrooz Tahmasebi and Stefanie Jegelka. Regularity in canonicalized models: A theoretical perspective. In *AISTATS*, pp. 4789–4797, 2025b.

John A Tomlin. On scaling linear programming problems. In *Computational practice in mathematical programming*, pp. 146–166. Springer, 2009.

Soledad Villar, Weichi Yao, David W Hogg, Ben Blum-Smith, and Bianca Dumitrascu. Dimensionless machine learning: Imposing exact units equivariance. *Journal of Machine Learning Research*, 24(109):1–32, 2023.

Rui Wang, Robin Walters, and Rose Yu. Incorporating symmetry into deep dynamics models for improved generalization. *arXiv preprint arXiv:2002.03061*, 2020.

Chenyang Wu, Qian Chen, Akang Wang, Tian Ding, Ruoyu Sun, Wenguo Yang, and Qingjiang Shi. On representing convex quadratically constrained quadratic programs via graph neural networks. *arXiv preprint arXiv:2411.13805*, 2024.

Jinxin Xiong, Xi Gao, Linxin Yang, Jiang Xue, Xiaodong Luo, and Akang Wang. Solving quadratic programs via deep unrolled douglas-rachford splitting. *arXiv preprint arXiv:2508.11869*, 2025.

Zhixiao Xiong, Fangyu Zong, Huigen Ye, and Hua Xu. Neuralqp: A general hypergraph-based optimization framework for large-scale qcqps. *arXiv preprint arXiv:2410.03720*, 2024.

Linxin Yang, Bingheng Li, Tian Ding, Jianghua Wu, Akang Wang, Yuyi Wang, Jiliang Tang, Ruoyu Sun, and Xiaodong Luo. An efficient unsupervised framework for convex quadratic programs via deep unrolling. *arXiv preprint arXiv:2412.01051*, 2024.

Morris Yau, Nikolaos Karalias, Eric Lu, Jessica Xu, and Stefanie Jegelka. Are graph neural networks optimal approximation algorithms? *Advances in Neural Information Processing Systems*, 37: 73124–73181, 2024.

Zhanguang Zhang, Didier Chételat, Joseph Cotnareanu, Amur Ghose, Wenyi Xiao, Hui-Ling Zhen, Yingxue Zhang, Jianye Hao, Mark Coates, and Mingxuan Yuan. Grass: Combining graph neural networks with expert knowledge for sat solver selection. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 6301–6311, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671627. URL https://doi.org/10.1145/3637528.3671627.

CONTENTS

A Preliminary **14**

A.1 Groups, group actions, and equivariance 15

A.2 Measure and probability measure 16

A.3 Group averaging, frame averaging, and canonicalization 16

A.4 Graph neural networks for solving LPs 17

A.5 The equilibration principles inside canonicalization 18

B Related Work **18**

B.1 Learning to optimize with graph neural networks 18

B.2 Geometric deep learning 19

B.3 Numerical scaling and presolve for linear programming 20

C The Incompatibility Between Graph Neural Networks and Rescaling Symmetry **21**

C.1 The degeneracy when enforcing the symmetry globally 21

C.2 The degeneracy when enforcing the symmetry locally 27

D The Existence of The Single Valued Canonicalization **29**

E The Regularity of Canonicalization **33**

E.1 Global Continuity Obstruction for Canonicalizations 33

E.2 Local Continuity Obstruction for Canonicalizations 33

E.3 Borel Measurability of Canonicalizations and Related Functions 36

E.4 Rescaling Mask for Exact Model Equivariance 40

F The Universality of Canonicalization-based GNNs **43**

G Extended experimental details **46**

G.1 Canonicalization and Prescaling Implementation 46

G.2 Dataset Background 47

G.3 Instance Generation 49

G.4 Backbone Realizations 52

G.5 Additional Experimental Results across Backbones 52

AI USE STATEMENT

We used large language models solely to aid and polish the writing of this paper, including improving its clarity, grammar, and readability. LLMs were not used to develop the research ideas, methodology, theoretical results, experiments, or conclusions. All AI-assisted text was carefully reviewed and revised by the authors, who take full responsibility for the final content.

A PRELIMINARY

Below, we briefly introduce some mathematical background related to this paper to make the paper self-contained.

A.1 GROUPS, GROUP ACTIONS, AND EQUIVARIANCE

Groups and subgroups. A binary operation on a set G is a map $\circ : G \times G \to G$ that assigns to every pair $(g_1, g_2) \in G \times G$ an element $g_1 \circ g_2 \in G$. A group $(G, \circ)$ is a set G equipped with a binary operation $\circ$ satisfying the following properties:

1. *Associativity:* $(g_1 \circ g_2) \circ g_3 = g_1 \circ (g_2 \circ g_3)$ for all $g_1, g_2, g_3 \in G$.
2. *Identity:* There exists an element $e_G \in G$, called the identity element, such that $e_G \circ g = g \circ e_G = g$ for every $g \in G$.
3. *Inverse:* For every $g \in G$, there exists an element $g^{-1} \in G$, called the inverse of g, such that $g \circ g^{-1} = g^{-1} \circ g = e_G$.

When the operation is clear from context, we simply write $g_1 g_2$ instead of $g_1 \circ g_2$. A group is called abelian if $g_1 g_2 = g_2 g_1$ for every $g_1, g_2 \in G$. A nonempty subset $H \subseteq G$ is called a subgroup of G, denoted by $H \leq G$, if it is closed under the group operation and inverse, i.e., $h_1 h_2 \in H$ and $h^{-1} \in H$ for every $h_1, h_2, h \in H$. The identity e_G then automatically belongs to H, and H itself forms a group under the operation inherited from G. Given groups $G_1, \ldots, G_r$, their direct product $G_1 \times \cdots \times G_r$ is the Cartesian product equipped with componentwise multiplication: $(g_1, \ldots, g_r)(g'_1, \ldots, g'_r) = (g_1 g'_1, \ldots, g_r g'_r)$. Its identity and inverse are respectively $(e_{G_1}, \ldots, e_{G_r})$ and $(g_1^{-1}, \ldots, g_r^{-1})$.

In our paper, the positive real numbers $\mathbb{R}_{++}$ form an abelian group under multiplication. The set $\mathrm{Diag}_{n\times n}(\mathbb{R}_{++})$ of positive diagonal matrices forms an abelian group under matrix multiplication. The symmetric group S_n is the group of all permutations of $[n] := \{1, \ldots, n\}$. Throughout this paper, we identify a permutation with its corresponding permutation matrix when convenient; under this identification, the group operation is matrix multiplication and the inverse of a permutation matrix P is $P^{-1} = P^\top$.

Group actions. A group describes abstract transformations, while a group action specifies how these transformations act on a particular space. Let G be a group and $\mathcal{X}$ be a set. A left group action of G on $\mathcal{X}$ is a map $G \times \mathcal{X} \to \mathcal{X}$, $(g, X) \mapsto g \cdot X$, satisfying $e_G \cdot X = X$ and $(g_1 g_2) \cdot X = g_1 \cdot (g_2 \cdot X)$ for every $g_1, g_2 \in G$ and $X \in \mathcal{X}$. We then say that G acts on $\mathcal{X}$. For $X \in \mathcal{X}$, the orbit of X under G is $G \cdot X := \{g \cdot X : g \in G\}$. The stabilizer of X is $G_X := \{g \in G : g \cdot X = X\}$, which contains precisely the transformations that leave X unchanged. The stabilizer is always a subgroup of G.

In this paper, the rescaling orbit $\mathcal{N} \cdot X$ therefore consists of all numerical representations of an LP instance that can be obtained from X by the admissible positive rescalings. Canonicalization amounts to selecting a representative from this orbit.

Invariance and equivariance. Suppose that a group G acts on an input space $\mathcal{X}$ and an output space $\mathcal{Y}$. Denote the two actions by $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$, respectively. A mapping $f : \mathcal{X} \to \mathcal{Y}$ is called G-equivariant if $f(\rho_{\mathcal{X}}(g)X) = \rho_{\mathcal{Y}}(g) f(X)$ for every $g \in G$ and $X \in \mathcal{X}$. Equivariance therefore means that transforming the input and then applying f gives the same result as first applying f and then transforming the output. If the action on the output space is trivial, i.e., $\rho_{\mathcal{Y}}(g)y = y$ for every $g \in G$ and $y \in \mathcal{Y}$, the above condition reduces to $f(g \cdot X) = f(X)$. In this case, f is called G-invariant. Thus, invariance is a special case of equivariance in which the output does not change under the group action.

The notion of equivariance also extends naturally to set-valued mappings. Let $\Phi : \mathcal{X} \to 2^{\mathcal{Y}}$, where $2^{\mathcal{Y}}$ denotes the collection of all subsets of $\mathcal{Y}$. For $S \subseteq \mathcal{Y}$, define $\rho_{\mathcal{Y}}(g)S := \{\rho_{\mathcal{Y}}(g)y : y \in S\}$. Then Φ is called G-equivariant if $\Phi(\rho_{\mathcal{X}}(g)X) = \rho_{\mathcal{Y}}(g)\Phi(X)$ for every $g \in G$ and $X \in \mathcal{X}$. This set-valued definition is particularly relevant for optimization because an optimization problem may admit multiple optimal solutions.

Conjugation and normalization. We frequently consider two different groups of transformations acting on the same space. Let H and K be groups whose elements act as invertible transformations of $\mathcal{X}$. For $k \in K$ and $h \in H$, the conjugate of h by k is the transformation $khk^{-1} := k \circ h \circ k^{-1}$. For a subgroup H, define $kHk^{-1} := \{khk^{-1} : h \in H\}$.

We say that K normalizes H if $kHk^{-1} = H$ for every $k \in K$. Equivalently, conjugating any admissible transformation in H by a transformation in K produces another admissible transfor-

mation in H. This condition implies $k \cdot (H \cdot X) = H \cdot (k \cdot X)$ for every $X \in \mathcal{X}$. Indeed, $k \cdot (h \cdot X) = (khk^{-1}) \cdot (k \cdot X)$, and $khk^{-1} \in H$ whenever K normalizes H.

This notion is important when two symmetries must be respected simultaneously. In particular, if K represents permutations and H represents rescalings, the normalization condition means that reordering variables or constraints cannot change which rescalings are admissible.

Groups used in this paper. Recall that the permutation group considered in this paper is $\mathcal{K} := S_m \times S_k \times S_n$. For $k = (P_m, P_k, P_n) \in \mathcal{K}$, its action on an LP instance $X = (c, A, b, C, d, l, u)$ is $k \cdot X = (P_n c, P_m A P_n^\top, P_m b, P_k C P_n^\top, P_k d, P_n l, P_n u)$. This action only changes the ordering of inequality constraints, equality constraints, and variables. The rescaling group is $\mathcal{N} := \mathbb{R}_{++} \times \mathrm{Diag}_{m\times m}(\mathbb{R}_{++}) \times \mathrm{Diag}_{k\times k}(\mathbb{R}_{++}) \times \mathrm{Diag}_{n\times n}(\mathbb{R}_{++})$. For $g = (\alpha, R_{\leq}, R_{=}, B) \in \mathcal{N}$, its action is $g \cdot X = (\alpha Bc, R_{\leq} AB, R_{\leq} b, R_{=} CB, R_{=} d, B^{-1} l, B^{-1} u)$. Here α rescales the objective, $R_{\leq}$ and $R_{=}$ rescale constraint rows, and B corresponds to the change of variables $x = By$.

The permutation group $\mathcal{K}$ normalizes the full rescaling group $\mathcal{N}$. More precisely, for $k = (P_m, P_k, P_n) \in \mathcal{K}$ and $g = (\alpha, R_{\leq}, R_{=}, B) \in \mathcal{N}$, conjugating the transformation induced by g by k gives $kgk^{-1} = (\alpha, P_m R_{\leq} P_m^\top, P_k R_{=} P_k^\top, P_n B P_n^\top)$. Since conjugating a positive diagonal matrix by a permutation matrix only reorders its diagonal entries, the resulting matrices remain positive diagonal. Hence $kgk^{-1} \in \mathcal{N}$ and therefore $k\mathcal{N}k^{-1} = \mathcal{N}$.

A.2 Measure and probability measure

Measure and probability measure. Let X be a set. A collection $\mathcal{A} \subseteq 2^X$ is a σ-algebra on X if $X \in \mathcal{A}$; $E \in \mathcal{A}$ implies $X \setminus E \in \mathcal{A}$; and $\bigcup_{i=1}^{\infty} E_i \in \mathcal{A}$ whenever $E_i \in \mathcal{A}$ for every $i \in \mathbb{N}$. The pair $(X, \mathcal{A})$ is called a measurable space.

A measure on $(X, \mathcal{A})$ is a mapping $\mu : \mathcal{A} \to [0, \infty]$ such that $\mu(\varnothing) = 0$ and, for every pairwise disjoint sequence $(E_i)_{i\in\mathbb{N}}$ in $\mathcal{A}$, $\mu(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu(E_i)$. The triple $(X, \mathcal{A}, \mu)$ is called a measure space. If $\mu(X) = 1$, then μ is called a probability measure.

Regular Borel measure. Let X be a Hausdorff topological space with topology τ, and let $\mathcal{B}(X) := \sigma(\tau) = \sigma\big(\{U \subseteq X : U \text{ is open}\}\big)$ be its Borel σ-algebra which is the smallest $\sigma-$algebra containing all open sets. A measure $\mu : \mathcal{B}(X) \to [0, \infty]$ is called a Borel measure. The measure μ is outer regular if, for every $E \in \mathcal{B}(X)$, $\mu(E) = \inf\big\{\mu(U) : E \subseteq U,\ U \subseteq X \text{ open}\big\}$, and inner regular if, for every $E \in \mathcal{B}(X)$, $\mu(E) = \sup\big\{\mu(K) : K \subseteq E,\ K \subseteq X \text{ compact}\big\}$. A Borel measure that is both outer and inner regular is called a regular Borel measure. If, in addition, $\mu(X) = 1$, it is called a Borel regular probability measure. In the LP settings, this definition is equivalent to the outer-measure based definitions.

Measurable and Borel measurable. Let $(X, \mathcal{A})$ and $(Y, \mathcal{C})$ be measurable spaces. A mapping $f : X \to Y$ is called measurable if $f^{-1}(E) \in \mathcal{A}$ for every $E \in \mathcal{C}$. If Y is a topological space equipped with its Borel σ-algebra $\mathcal{B}(Y)$, then a mapping $f : (X, \mathcal{A}) \to Y$ is called Borel measurable if $f^{-1}(B) \in \mathcal{A}$ for every $B \in \mathcal{B}(Y)$. Equivalently, it suffices that $f^{-1}(U) \in \mathcal{A}$ for every open set $U \subseteq Y$. In particular, if X is also a topological space, then $f : X \to Y$ is Borel measurable when it is measurable as a mapping from $(X, \mathcal{B}(X))$ to $(Y, \mathcal{B}(Y))$.

A.3 Group averaging, frame averaging, and canonicalization

Let a group G act on an input space $\mathcal{X}$, and let $\rho_{\mathcal{Y}}$ be a linear action on an output space $\mathcal{Y}$. Given an arbitrary backbone $\varphi : \mathcal{X} \to \mathcal{Y}$, the following model-agnostic constructions can enforce the desired symmetry.

Group averaging. For a finite group G, the group averaging defines $\mathrm{Av}_G[\varphi](X) := \frac{1}{|G|} \sum_{g\in G} \rho_{\mathcal{Y}}(g)\varphi(g^{-1} \cdot X)$. A change of variables shows that $\mathrm{Av}_G[\varphi]$ is G-equivariant; taking the output action to be trivial gives invariance. For a compact group, the sum is replaced by integration against its normalized Haar probability measure. However, the Haar measure of a locally compact group has finite total mass if and only if the group is compact. Hence, standard normalized Haar averaging is unavailable for the non-compact rescaling group $\mathcal{N}$.

Frame averaging. Frame averaging replaces the whole group by an input-dependent subset (Puny et al., 2021). A finite G-equivariant frame is a mapping $\mathcal{F} : \mathcal{X} \to 2^G \setminus \{\varnothing\}$ satisfying $\mathcal{F}(g \cdot X) = g\mathcal{F}(X), g\mathcal{F}(X) := \{gh : h \in \mathcal{F}(X)\}$. It induces the equivariant predictor

$$\mathrm{FA}_{\mathcal{F}}[\varphi](X) := \frac{1}{|\mathcal{F}(X)|} \sum_{g \in \mathcal{F}(X)} \rho_{\mathcal{Y}}(g)\varphi(g^{-1} \cdot X),$$

which requires $|\mathcal{F}(X)|$ backbone evaluations. Let $G_X := \{h \in G : h \cdot X = X\}$ be the stabilizer, also called the automorphism group, of X. Frame equivariance implies $G_X g \subseteq \mathcal{F}(X)$ for every $g \in \mathcal{F}(X)$, while all elements of $G_X g$ produce the same transformed input. Thus, a nontrivial stabilizer creates redundant evaluations, and an uncountable stabilizer forces every equivariant frame at X to be uncountable.

Canonicalization. Canonicalization removes this redundancy by operating on distinct transformed inputs (Kaba et al., 2023; Ma et al., 2024). A canonicalization is a mapping $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{X}} \setminus \{\varnothing\}$ satisfying $\mathcal{C}(g \cdot X) = \mathcal{C}(X)$. It is an orbit canonicalization if $\mathcal{C}(X) \subseteq G \cdot X$. Every equivariant frame induces the orbit canonicalization $\mathcal{C}_{\mathcal{F}}(X) := \{g^{-1} \cdot X : g \in \mathcal{F}(X)\}$. Whenever $\mathcal{F}(X)$ is finite, $|\mathcal{F}(X)| = |G_X|\,|\mathcal{C}_{\mathcal{F}}(X)|$, where $G_X := \{h \in G : h \cdot X = X\}$ is the stabilizer. Thus, $|\mathcal{C}_{\mathcal{F}}(X)|$ counts the distinct backbone inputs. For invariant targets and finite $\mathcal{C}(X)$, canonical averaging is $|\mathcal{C}(X)|^{-1} \sum_{Y \in \mathcal{C}(X)} \varphi(Y)$.

For a singleton canonicalization $\mathcal{C}(X) = \{C(X)\}$ and an equivariant target, select a transporter g_X satisfying $X = g_X \cdot C(X)$ and restore predictions by $\mathrm{CA}_{\mathcal{C}}[\varphi](X) := \rho_{\mathcal{Y}}(g_X)\varphi(C(X))$. Writing $Z = C(X)$, any two valid transporters differ by right multiplication by an element of G_Z. Consequently, exact equivariance holds if and only if the canonical prediction is fixed by this stabilizer: $\rho_{\mathcal{Y}}(s)\varphi(Z) = \varphi(Z), s \in G_Z$. Under this condition, any valid transporter yields the same prediction, so uniqueness of g_X is unnecessary. The condition holds automatically when G_Z acts trivially on the output space. A deterministic transporter selection alone does not suffice.

In our paper, $G = \mathcal{N}$, and the canonicalization must additionally preserve the permutation symmetry $\mathcal{K}$: $\mathcal{C}(g \cdot X) = \mathcal{C}(X), \mathcal{C}(k \cdot X) = k \cdot \mathcal{C}(X), \mathcal{C}(X) \subseteq \mathcal{N} \cdot X$. A singleton orbit canonicalization therefore uses the minimum possible averaging complexity—one backbone evaluation. To overcome the equivariance obstruction arising from transporter non-uniqueness, we combine energy-based canonicalization with a projection onto the stabilizer-fixed output subspace. Energy minimization provides a canonical representative and a valid transporter g_X, while the projection ensures that the resulting predictor is both rescaling- and permutation-equivariant, independently of the chosen valid transporter.

A.4 Graph neural networks for solving LPs

LP graph representation. Following the variable-constraint bipartite representation commonly used for learning on mathematical programs (Gasse et al., 2019; Chen et al., 2023a), an LP is represented as an attributed weighted bipartite graph whose two types of nodes correspond to variables and constraints.

Consider the LP instance $X = (c, A, b, C, d, l, u)$ introduced in the main text. For convenience, stack the inequality and equality constraint matrices and right-hand sides as $\bar{A} := (A^T, C^T)^T$ and $\bar{b} := (b^T, d^T)^T$, and let $M := m + k$ be the total number of constraints. The corresponding bipartite graph consists of a set of constraint nodes $V := \{v_1, \ldots, v_M\}$ and a set of variable nodes $W := \{w_1, \ldots, w_n\}$. Variable node w_j represents the variable x_j and carries the feature $s_j^W := (c_j, l_j, u_j)$. Constraint node v_i carries its right-hand side together with an indicator of its constraint type. More precisely, $s_i^V := (1, 0, b_i)$ for $i \in [m]$, while $s_{m+r}^V := (0, 1, d_r)$ for $r \in [k]$. Whenever $\bar{A}_{ij} \neq 0$, constraint node v_i and variable node w_j are connected by an edge with scalar weight $E_{ij} := \bar{A}_{ij}$. Equivalently, we may set $E_{ij} = 0$ for non-adjacent pairs and sum over all nodes.

Message-passing LP-GNNs. We next describe the LP-GNN class used throughout this paper, following the message-passing formulation of Chen et al. (2023a). This architecture is the base of (Li et al., 2026; Wu et al., 2024; Chen et al., 2024b). Each constraint and variable node first receives a hidden representation through shared encoding maps: $h_i^{0,V} = f_{\mathrm{in}}^V(s_i^V)$ for $i \in [M]$, and $h_j^{0,W} = f_{\mathrm{in}}^W(s_j^W)$ for $j \in [n]$. The network then alternates message passing between the

two node types. At layer $\ell = 1, \ldots, L$, the constraint-node embeddings are updated according to $h_i^{\ell,V} = g_\ell^V \left(h_i^{\ell-1,V}, \sum_{j=1}^n E_{ij} f_\ell^W (h_j^{\ell-1,W})\right)$, where f_ℓ^W is a learnable message map and g_ℓ^V is a learnable constraint-node update map. Similarly, variable-node embeddings are updated by $h_j^{\ell,W} = g_\ell^W \left(h_j^{\ell-1,W}, \sum_{i=1}^M E_{ij} f_\ell^V (h_i^{\ell-1,V})\right)$. The functions $f_{\text{in}}^V, f_{\text{in}}^W, f_\ell^V, f_\ell^W, g_\ell^V, g_\ell^W$ are typically parameterized by multilayer perceptrons. All nodes of the same type share the same maps.

After L message-passing layers, a primal solution is predicted by applying a shared readout to each variable node. A representative readout has the form $[F_\theta(X)]_j = f_{\text{out}}^W \left(h_j^{L,W}, \sum_{i=1}^M h_i^{L,V}, \sum_{r=1}^n h_r^{L,W}\right)$. The same readout map is used for every variable, and $F_\theta(X) \in \mathbb{R}^n$ has one output coordinate for each LP variable. It's easy to check that the GNN is permutation-equivariant under permutations of variables or constraints.

Implicit LP-GNNs. We further consider the implicit GNN architecture of Liu et al. (2026). As before, an LP instance X is represented by a bipartite graph whose constraint nodes are indexed by $i \in [M]$, whose variable nodes are indexed by $j \in [n]$, and whose edge feature E_{ij} is the corresponding constraint coefficient. The constraint and variable features are first encoded as $\bar{h}_i^V = f_{\text{in}}^V(s_i^V), \bar{h}_j^W = f_{\text{in}}^W(s_j^W)$. In addition to the static embedding $\bar{h}_j^W$, each variable node carries a dynamic state z_j. Given a current state $z = (z_1, \ldots, z_n)$, define $u_i^{0,V} = \bar{h}_i^V, u_j^{0,W} = \begin{bmatrix} \bar{h}_j^W \\ z_j \end{bmatrix}$. The implicit update operator applies t successive rounds of bipartite message passing. For $r = 1, \cdots, t$, the constraint-node representations are updated by $u_i^{r,V} = g_r^V(u_i^{r-1,V}, \sum_{j=1}^n E_{ij} f_r^W(u_j^{r-1,W}))$, and the resulting constraint representations are passed to the variable nodes: $u_j^{r,W} = g_r^W(u_j^{r-1,W}, \sum_{i=1}^M E_{ij} f_r^V(u_i^{r,V}))$ The next dynamic state is obtained from the final variable-node representation, $[G_\theta(z; X)]_j = q_\theta(u_j^{t,W})$, where the same operator G_θ is shared across implicit iterations. The implicit representation is defined as an equilibrium of this operator: $z^\star = G_\theta(z^\star; X)$. A primal solution is then predicted independently at each variable node through a shared readout: $[F_\theta(X)]_j = f_{\text{out}}^W \left(\bar{h}_j^W, z_j^\star\right)$. Practically, people always use finite unrolling of the implicit model, as in (Liu et al., 2026).

A.5 The equilibration principles inside canonicalization

Ignoring the stabilizer and feature anchors in equation 3, differentiation with respect to r_i gives $\sum_{j:(i,j)\in\Omega} (\bar{A}'_{ij})^2 = 1, i \in I_{\bar{A}}$. For $j \in J_k$, differentiation with respect to η_j gives $\sum_{i:(i,j)\in\Omega} (\bar{A}'_{ij})^2 = \frac{m_k}{n_k}$. Hence the effective row norm is one, and the effective column norm in component k is $\sqrt{m_k/n_k}$. With the stabilizer and anchors included, these identities hold only approximately.

For β-version's energy, the situation is similar. However, the shared scalar v cannot impose independent column targets; it only enforces that the sum of row masses is equal to the sum of column masses.

B Related Work

B.1 Learning to optimize with graph neural networks

The variable-constraint bipartite graph representation was first introduced by Gasse et al. (2019). Subsequent work has used the same representation to accelerate branching, diving, node selection, and primal search (Gupta et al., 2020; Nair et al., 2020; Khalil et al., 2022; Han et al., 2023). For end-to-end theoretical analyses, two complementary paradigms are particularly relevant here: Weisfeiler–Leman (WL)-based representation theory and algorithm-unrolled networks.

WL-based representation theory. This series of work encodes an optimization instance as an attributed graph and defines a specific WL test matched to the message-passing architecture. For fixed-size LPs, Chen et al. (2023a) show that WL-indistinguishable instances agree on feasibility, boundedness, optimal value, and the minimum-ℓ_2 optimal solution up to permutation. For MILPs, standard message-passing GNNs cannot distinguish some feasible and infeasible instances, although

universality holds for unfoldable instances and, with random identifiers, on finite data sets (Chen et al., 2023b). For strong-branching scores, message-passing GNNs are universal on an MP-tractable subclass, while second-order folklore GNNs are universal for all MILPs (Chen et al., 2024c).

For continuous convex linearly constrained quadratic programs, message-passing GNNs can represent feasibility, finite optimal value, and a canonical minimum-norm optimizer; for mixed-integer variants, universality holds only on restricted subclasses (Chen et al., 2024b). Similar universality results hold for convex QCQPs but not for general nonconvex QCQPs (Wu et al., 2024). For SOCPs, universality holds for feasibility and optimal-solution prediction and extends to p-order cones (Li et al., 2026). For linear SDPs, standard message-passing and some second-order architectures are insufficient, whereas a VC–2-FWL-equivalent architecture is expressive enough to capture the required matrix interactions (Qian & Morris, 2026a).

Algorithm-unrolling-based framework. Algorithm unrolling represents a finite number of optimization iterations as neural-network layers, with selected algorithmic components made learnable (Gregor & LeCun, 2010; Monga et al., 2021). For LPs, MPNNs can simulate interior-point iterations (Qian et al., 2023), while PDHG-Net unrolls primal–dual hybrid gradient iterations and approximates optimal solutions with polynomially many neurons (Li et al., 2024a). Constant-width, polylogarithmic-depth GNNs can compute $(1+\epsilon)$-approximate solutions to packing and covering LPs (Li et al., 2024b), and constant-depth GNNs suffice for almost all sparse binary LPs under an average-instance counting model (Li et al., 2025).

For convex LCQPs, GNN constructions simulate interior-point methods (Qian & Morris, 2025) and unroll QP-specialized PDHG (Yang et al., 2024), Douglas–Rachford splitting (Xiong et al., 2025; Sambharya et al., 2023), and consensus OSQP for network-structured QPs (Saravanos et al., 2025). Beyond LPs and QPs, NeuralQP combines learned QCQP prediction with McCormick-based repair and neighborhood optimization (Xiong et al., 2024). Learned fixed-point warm starts apply across QPs, SOCPs, and SDPs (Sambharya et al., 2024), while ADMM-based unrolling has been developed for SOCPs and SDPs (Oshin et al., 2026) and constrained dual unrolling for MIQPs (Hadou & Ribeiro, 2026).

For combinatorial optimization, OptGNN realizes SDP-relaxation-based approximation algorithms for Max-CSPs, with its strongest optimality guarantee conditional on the Unique Games Conjecture (Yau et al., 2024). Primal–dual-aligned bipartite GNNs similarly inherit approximation guarantees for minimum vertex cover, set cover, and hitting set (He & Vitercik, 2025).

Optimization symmetries and relation to our work. Recent work exploits optimization symmetries to improve supervision, feature construction, and data efficiency. SymILO aligns supervision over equivalent integer solutions (Chen et al., 2024a), while orbit-based feature augmentation breaks ties among formulation-symmetric ILP variables (Chen et al., 2025). For LPs and QPs, solution-preserving transformations also support supervised and contrastive data augmentation (Qian & Morris, 2026b). However, as an optimization symmetry, rescaling symmetry has always been overlooked.

B.2 Geometric deep learning

Geometric deep learning incorporates a group action either within every layer or by symmetrizing an unrestricted backbone. Group convolutions are an early example of the architectural route (Cohen & Welling, 2016); invariant and equivariant graph-network constructions extend the same principle to relational data (Maron et al., 2018). Correctly restricting the hypothesis class can also yield a strict statistical benefit under suitable assumptions (Elesedy & Zaidi, 2021). In the model-agnostic route, averaging $\rho_Y(g)^{-1}\phi(g\cdot x)$ over a finite group, or over normalized Haar measure for a compact group, makes an arbitrary backbone equivariant. Full group averaging can be expensive, and a noncompact group has no finite normalized Haar measure, so this construction cannot be transferred unchanged to the positive rescaling group considered here. Frame averaging reduces the expensive cost by replacing the whole group with an input-dependent equivariant set of transformations (Puny et al., 2021). Subsequent work constructs minimal frames for broader symmetries (Lin et al., 2024) and learns equivariant probability distributions over transformations rather than using an unweighted set (Kim et al., 2023). However, these constructions still retain multiple backbone evaluations when several frame elements map an input to the same representative, particularly in the presence of a nontrivial stabilizer, as noted in (Ma et al., 2024). Ma et al. (2024) show that an equivariant frame

induces an orbit canonicalization after quotienting redundant transformations by the stabilizer, and conversely that an orbit canonicalization induces a frame. Energy minimization is one mechanism for learning or computing such a representative (Kaba et al., 2023).

Regularity, generalization, and noncompact actions. A globally continuous orbit selector need not exist. For several common point cloud actions, Dym et al. (2024) prove continuity obstructions and show that an unweighted frame can turn a smooth backbone into a discontinuous symmetric model; appropriately weighted frames can preserve continuity. Conversely, a discontinuous selector does not automatically make the end-to-end predictor discontinuous: Tahmasebi & Jegelka (2025b) characterize regularity after canonicalization. Statistical analyses further compare the generalization benefit of canonicalization with group averaging (Tahmasebi & Jegelka, 2025a). These results separate three requirements that are often conflated—exact symmetry, regularity of a selector, and regularity or learnability of the composed model. More recent work develops adaptive canonicalization for invariant anisotropic geometric networks (Lin & Levie, 2026).

ScaleGMN incorporates neuron permutations and parameter rescalings into graph metanetworks (Kalogeropoulos et al., 2024). Dimensionless machine learning enforces units equivariance through invariant features and output restoration (Villar et al., 2023), while Lie Algebra Canonicalization uses energy minimization to handle noncompact actions and nonclosed orbits (Shumaylov et al., 2025). Transferring these ideas to LPs is highly non-trivial: representatives must remain within the original rescaling orbits to preserve LP equivalence and enable solution transport, while retaining permutation equivariance and PAC approximation guarantees. We address these obstacles by establishing failure theorems for direct symmetry enforcement, characterizing compatibility between rescaling and permutation actions, and proving that every LP admits a single-valued, permutation- and rescaling-equivariant orbit canonicalization requiring one backbone evaluation. We further develop a unified energy-based framework, and establish Borel measurability and preservation of backbone PAC universality under mild assumptions.

B.3 Numerical scaling and presolve for linear programming

Scaling in LP algorithms. For LPs, Tomlin (2009) systematically compared geometric-mean, equilibration, and optimization-based scaling rules and studied their effects on the numerical behavior of the simplex method. A later large-scale comparison by Elble & Sahinidis (2012) found that no more elaborate rule consistently outperformed simple equilibration across the tested simplex solvers and LP benchmarks. Scaling is also standard in interior-point implementations. For example, the primal–dual interior-point solver PCx applies Curtis–Reid row and column scaling to the constraint matrix before solving the Newton systems (Czyzyk et al., 1999). In first-order primal–dual methods, diagonal scaling is interpreted as a metric or step-size preconditioner (Pock & Chambolle, 2011). PDLP makes this idea LP-specific by combining Ruiz equilibration with Pock–Chambolle scaling and consistently transforming the constraint matrix, objective, right-hand sides, and variable bounds (Applegate et al., 2022). Thus, positive diagonal scaling appears across the three main families of LP algorithms, but is evaluated using solver-dependent numerical criteria: reliable simplex pivots, stable interior-point linear systems, or faster first-order convergence.

Difference from our canonicalization. Although conventional LP scaling and our method use closely related algebraic transformations, they solve different problems. Traditional scaling selects a numerically favorable formulation for a particular solver. It is not required to map every member of a complete rescaling orbit to the same representative, to commute with row and variable permutations, or to resolve the objective, row, and variable scalings jointly. Even an exact equilibration of the constraint matrix need not canonicalize the full tuple (c, A, b, l, u) or remove its remaining scaling gauges. By contrast, our map is constructed to be exactly rescaling-invariant and permutation-equivariant for the full LP action. Its purpose is therefore to enhance the generalization performance of the learned model by leveraging the symmetries, rather than improved conditioning of a numerical solver.

Presolve. LP presolve removes redundant rows and columns, fixes variables, tightens bounds, and may detect infeasibility or unboundedness before the main algorithm is called (Brearley et al., 1975; Andersen & Andersen, 1995). We do not treat presolve as a competing canonicalization baseline because it is not an invertible action on a fixed-dimensional LP space: it can change the numbers of variables and constraints, alter sparsity, and even reduce an instance to an empty problem, after

which a postsolve stack reconstructs a solution in the original coordinates. Comparing a presolved model only against unpresolved learning models would therefore confound symmetry enforcement with a change in graph size and problem difficulty.

C The Incompatibility Between Graph Neural Networks and Rescaling Symmetry

Below, we discuss the general bottlenecks faced by GNNs when leveraging them for optimization problems. Specifically, we prove that vanilla GNNs have a poor ability to enforce the rescaling symmetry. The results are presented in two parts, corresponding to global and local enforcement of symmetry.

C.1 The degeneracy when enforcing the symmetry globally

Here, we first introduce the first degeneracy theorem, which shows that general continuous functions would ignore necessary coefficients when enforcing the invariance globally.

Theorem C.1. *For any continuous function $f : \mathbb{P}_{m,n,k} \to \mathbb{R}^n$:*

- *If $f(c, A, b, C, d, l, u) = f(\alpha c, A, b, C, d, l, u)$ for all $\alpha > 0$, then $f(c, A, b, C, d, l, u) = g(A, b, C, d, l, u)$ for all $c \in \mathbb{R}^n$ and some function g.*
- *If $f(c, A, b, C, d, l, u) = f(c, R_{\leq} A, R_{\leq} b, C, d, l, u)$ for all $R_{\leq} \in Diag_{m\times m}(\mathbb{R}_{++})$, then $f(c, A, b, C, d, l, u) = h(c, C, d, l, u)$ for all $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$ and some function h.*
- *If $f(c, A, b, C, d, l, u) = f(c, A, b, R_{=} C, R_{=} d, l, u)$ for all $R_{=} \in Diag_{k\times k}(\mathbb{R}_{++})$, then $f(c, A, b, C, d, l, u) = h(c, A, b, l, u)$ for all $C \in \mathbb{R}^{k\times n}, d \in \mathbb{R}^k$ and some function h.*

Proof. For the first claim, it suffices to notice that:

$$f(c, A, b, C, d, l, u) = \lim_{n\to+\infty} f(\frac{1}{n}c, A, b, C, d, l, u) = f(0, A, b, C, d, l, u)$$

Hence, there exists a continuous function g such that: $f(c, A, b, C, d, l, u) = g(A, b, C, d, l, u)$.

For the second claim, it suffices to notice that:

$$f(c, A, b, C, d, l, u) = \lim_{n\to+\infty} f(c, \frac{1}{n}A, \frac{1}{n}b, C, d, l, u) = f(c, 0, 0, C, d, l, u)$$

Hence, there exists a continuous function h such that: $f(c, A, b, C, d, l, u) = h(c, C, d, l, u)$.

For the third claim, it suffices to notice that:

$$f(c, A, b, C, d, l, u) = \lim_{n\to+\infty} f(c, A, b, \frac{1}{n}C, \frac{1}{n}d, l, u) = \lim_{n\to+\infty} f(c, A, b, l, u)$$

Hence, there exists a continuous function h such that: $f(c, A, b, C, d, l, u) = h(c, A, b, l, u)$. □

Then, to prove the degenerate theorem for enforcing change-of-variable inductive bias globally, we first introduce some definitions and theorems that would help.

Definition C.2. *A collection of sets $\cup_{\alpha\in I}\mathcal{U}_\alpha$ is called a countable almost disjoint open partition of C if and only if*

- *I is countable.*
- *$\mathcal{U}_\alpha \subset C$ is open for any $\alpha \in I$.*
- *$\mathcal{U}_\alpha \cap \mathcal{U}_\beta = \emptyset$ for all $\alpha \neq \beta \in I$.*
- *$\mu(C \cap (\cup_{\alpha\in I}\mathcal{U}_\alpha)^c) = 0$, where μ is the Lebesgue measure.*

Definition C.3. *Let $d \in \mathbb{N}$. For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$, we write $x^\alpha := x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ and $|\alpha| := \alpha_1 + \cdots + \alpha_d$. A function $p : \mathbb{R}^d \to \mathbb{R}$ is called a real polynomial on $\mathbb{R}^d$ if there exist a finite set $F \subset \mathbb{N}^d$ and coefficients $c_\alpha \in \mathbb{R}$ such that $p(x) = \sum_{\alpha \in F} c_\alpha x^\alpha$. We say that p is nontrivial or non-constant if at least one coefficient c_α is nonzero with $\alpha \neq 0^d$.*

Definition C.4. *A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is defined to be a positive degree piecewise polynomial if and only if there exists an countable almost disjoint open partition $\cup_{\alpha \in I} \mathcal{U}_\alpha$ of $\mathbb{R}^n$ such that f is a (real) polynomial on $\mathcal{U}_\alpha$ for all $\alpha \in I$ and f is continuous. A function $f = (f_1, \cdots, f_p) : \mathbb{R}^n \mapsto \mathbb{R}^p$ is defined to be a piecewise polynomial if and only if each component f_i is a piecewise polynomial.*

Definition C.5. *A function $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is defined to be piecewise linear if and only if σ is continuous and there exists an countable almost disjoint intervals $\cup_{\alpha' \in I'} [r_{\alpha'}, l_{\alpha'}]$ such that $\cup_{\alpha' \in I'} [r_{\alpha'}, l_{\alpha'}] = \mathbb{R}$ while σ is affine on $[r_{\alpha'}, l_{\alpha'}]$ for all $\alpha' \in I'$.*

Now, we present some lemmas that are useful for the following.

Lemma C.6 (Zero set of a nonzero polynomial (Mityagin, 2015))**.** *Let $p : \mathbb{R}^d \to \mathbb{R}$ be a nonzero real polynomial. Define its zero set by $Z(p) := \{x \in \mathbb{R}^d \mid p(x) = 0\}$. Then $Z(p)$ has Lebesgue measure zero in $\mathbb{R}^d$.*

Lemma C.7. *If $f = (f_1, \cdots, f_p) : \mathbb{R}^n \mapsto \mathbb{R}^p$ is a piecewise polynomial and $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is a piecewise linear activation function. Then $\sigma(f) = (\sigma(f_1), \cdots, \sigma(f_p))$ is also a piecewise polynomial.*

Proof. By definition, it suffices to prove the situation when $p = 1$. Under this situation, first note that $\sigma(f)$ is continuous as the composition of two continuous functions. Since f is piecewise polynomial on $\mathbb{R}^n$, we can take an countable almost disjoint open partition $\cup_{\alpha \in I \cup I''} \mathcal{U}_\alpha$ of $\mathbb{R}^n$ such that f is a nontrivial polynomial on $\mathcal{U}_\alpha$ for all $\alpha \in I$ and f is a trivial polynomial on $\mathcal{U}_{\alpha''}$ for all $\alpha'' \in I''$. Similarly, since σ is piecewise linear function, we can take an countable almost disjoint open partition $\cup_{\alpha' \in I'} (r_{\alpha'}, l_{\alpha'})$ of $\mathbb{R}$, such that σ is affine on $(r_{\alpha'}, l_{\alpha'})$. Since f is continuous, we can see that $f^{-1}(r_{\alpha'}, l_{\alpha'})$ is open. Hence, note that

- On the open set $O_{\alpha,\alpha'} := f^{-1}(r_{\alpha'}, l_{\alpha'}) \cap U_\alpha$, f is polynomial for any $\alpha \in I, \alpha' \in I'$. On the open set $\mathcal{U}_{\alpha''}$ where $\alpha'' \in I''$, we have $\sigma(f)$ is a constant, hence a trivial polynomial.
- $I \times I'$ is countable and hence $(\cup_{\alpha \in I, \alpha' \in I'} O_{\alpha,\alpha'}) \cup (\cup_{\alpha'' \in I''} \mathcal{U}_{\alpha''})$ is a countable union
- Let $Q = (\cup_{\alpha \in I, \alpha' \in I'} O_{\alpha,\alpha'}) \cup (\cup_{\alpha'' \in I''} \mathcal{U}_{\alpha''}), \mathcal{U} = \cup_{\alpha \in I} U_\alpha$, we can see that

$$\begin{aligned}
\mu(Q^c) &= \mu((\cup_{\alpha \in I \cup I''} U_\alpha)^c \cup (\cup_{\alpha' \in I'} \{x \in \mathcal{U} \mid f(x) = l_{\alpha'}\}) \cup (\cup_{\alpha' \in I'} \{x \in \mathcal{U} \mid f(x) = r_{\alpha'}\})) \\
&\leq \mu((\cup_{\alpha \in I \cup I''} U_\alpha)^c) + \sum_{\alpha' \in I'} \mu(\{x \in \mathcal{U} \mid f(x) = l_{\alpha'}\}) + \sum_{\alpha' \in I'} \mu(\{x \in \mathcal{U} \mid f(x) = r_{\alpha'}\}) \\
&= \sum_{\alpha' \in I'} \mu(\{x \in \mathcal{U} \mid f(x) = l_{\alpha'}\}) + \sum_{\alpha' \in I'} \mu(\{x \in \mathcal{U} \mid f(x) = r_{\alpha'}\}) \\
&\leq \sum_{\alpha \in I} \sum_{\alpha' \in I'} \mu(\{x \in \mathcal{U}_\alpha \mid f(x) = l_{\alpha'}\}) + \sum_{\alpha \in I} \sum_{\alpha' \in I'} \mu(\{x \in \mathcal{U}_\alpha \mid f(x) = r_{\alpha'}\})
\end{aligned}$$

  due to the $\sigma-$subadditivity of measure. Since on $\mathcal{U}_\alpha$, $f(x)$ is a polynomial, we can see that: $\mu(\{x \in \mathcal{U}_\alpha \mid f(x) = r_{\alpha'}\}) = 0$ or $\mu(\{x \in \mathcal{U}_\alpha \mid f(x) = r_{\alpha'}\}) = \mu(\mathcal{U}_\alpha)$ by Lemma C.6. And $\mu(\{x \in \mathcal{U}_\alpha \mid f(x) = r_{\alpha'}\}) = \mu(\mathcal{U}_\alpha)$ holds if and only if $f(x) = r_{\alpha'}$ on $\mathcal{U}_\alpha$, i.e., f is trivial on $\mathcal{U}_\alpha$, which lead to a contradiction. Hence $\mu(\{x \in \mathcal{U}_\alpha \mid f(x) = r_{\alpha'}\}) = 0$ for all $\alpha \in I, \alpha' \in I'$. Similarly, we can show that $\mu(\{x \in \mathcal{U}_\alpha \mid f(x) = l_{\alpha'}\}) = 0$ for all $\alpha \in I, \alpha' \in I'$. Combining all above, we get $\mu(Q^c) = 0$.

It's easy to verify that the construction satisfies the disjoint property. Combining all these finish the proof. □

Corollary C.8. *The following operations still preserve the piecewise polynomial property:*

- *Concatenate: (f, g) is a piecewise polynomial if f, g is a piecewise polynomial.*
- *Linear Combination: $af + bg$ is a piecewise polynomial for all $a, b \in \mathbb{R}$ if f, g is a piecewise polynomial.*

- *Product: fg is a piecewise polynomial if f, g is a piecewise polynomial.*

Proof. The first term is just the definition of piecewise polynomial. The second term can be derived by noting that $af + bg = (a, b)(f, g)^T$ and using Lemma C.7. As for the third term, we can take $\cup_{\alpha\in I}\mathcal{U}_\alpha$ to be a countable almost disjoint open partition of the whole space such that f is polynomial on each $\mathcal{U}_\alpha$ and $\cup_{\beta\in J}\mathcal{Q}_\beta$ to be a countable almost disjoint open partition of the whole space such that g is polynomial on each $\mathcal{Q}_\beta$. On $\mathcal{U}_\alpha \cap \mathcal{Q}_\beta$, fg is a polynomial. And $\cup_{\alpha\in I,\beta\in J}\mathcal{U}_\alpha \cap \mathcal{Q}_\beta$ is a countable collection of disjoint open sets. Since $(\cup_{\alpha\in I,\beta\in J}\mathcal{U}_\alpha \cap \mathcal{Q}_\beta)^c \subset (\cup_{\alpha\in I}\mathcal{U}_\alpha)^c \cup (\cup_{\beta\in J}\mathcal{Q}_\beta)^c$, it's easy to get $\mu((\cup_{\alpha\in I,\beta\in J}\mathcal{U}_\alpha \cap \mathcal{Q}_\beta)^c) \le \mu((\cup_{\alpha\in I}\mathcal{U}_\alpha)^c) + \mu((\cup_{\beta\in J}\mathcal{Q}_\beta)^c) = 0$, which finishes the proof □

Now, we start to prove the second degenerate theorem.

Theorem C.9. *For any LP-GNN $f_{GNN} : \mathbb{P}_{m,n,k} \to \mathbb{R}^n$, if $f(c, A, b, C, d, l, u) = Bf(Bc, AB, b, CB, d, B^{-1}l, B^{-1}u))$ for all $B \in Diag_{n\times n}(\mathbb{R}_{++})$, then there exists g, such that $f(c, A, b, C, d, l, u) = g(l, u)$.*

Proof. To prove the theorem, we introduce a lemma first.

Lemma C.10. *Any LP-GNN using piecewise linear activation function is a piecewise polynomial with respect to the input c, A, b, C, d, l, u.*

Proof. We prove by induction on the three type of layers.

Embedding Layer: Note that $h^{0,\text{variable}_j} = f_{\text{Encode-MLP}}(c_j, l_j, u_j)$. Since passing through the piecewise linear activation function σ and affine transformation $Wx + b$ preserve the piecewise polynomial property, it's directly to get that $h^{0,\text{variable}_j}$ is a piecewise polynomial with respect to the input c, A, b, C, d, l, u. Similarly, $h^{0,\text{constraint}_i}$ is a piecewise polynomial with respect to the input c, A, b, C, d, l, u.

Message Passing Layer: We still proceed by induction. Assume at the t-th layer, $h^{t,\text{variable}_j}, h^{t,\text{constraint}_i}$ is a piecewise polynomial for all possible i, j. Without loss of generality, we only consider the inequality constraint here. Then, for all possible variable j, we have $\sum_i A_{ij} f_{\text{Message-MLP}}(h^{t,\text{constraint}_i})$ is still a piecewise polynomial with respect to the input c, A, b, C, d, l, u by Lemma C.7 and Corollary C.8. Apply Lemma C.7 and Corollary C.8 again, we get $(h^{t,\text{variable}_j}, \sum_i A_{ij} f_{\text{Message-MLP}}(h^{t,\text{constraint}_i}))$ is a piecewise polynomial with respect to the input c, A, b, C, d, l, u . Hence $h^{t+1,\text{variable}_j} = f_{\text{Update}}(h^{t,\text{variable}_j}, \sum_i A_{ij} f_{\text{Message-MLP}}(h^{t,\text{constraint}_i}))$ is a piecewise polynomial. Similar results can be established for $h^{t+1,\text{constraint}_i}$ for all possible constraint i as well. Hence, after the final message passing layer, the features $h^{F,\text{variable}_j}, h^{F,\text{constraint}_i}$ is still a piecewise polynomial.

Output Layer: By using the same augment, since concatenate, finite summation of piecewise polynomials is still a piecewise polynomial, we can see that the final output is a piecewise polynomial.

Combining all above finish the proof. □

We now continue the proof of Theorem C.9. By Lemma C.10, the output of the LP-GNN is a piecewise polynomial. Hence, there exists a countable almost disjoint open partition $\cup_{\alpha\in I}\mathcal{U}_\alpha$ of the input space such that on each $\mathcal{U}_\alpha$, the output coordinates f_j are polynomials.

Under the diagonal rescaling $X' = B \cdot X$ with $B = \text{Diag}(\beta_1, \ldots, \beta_n) \in \text{Diag}_{n\times n}(\mathbb{R}_{++})$, the input features transform as follows:

$$c_j \mapsto \beta_j c_j, \quad A_{ij} \mapsto \beta_j A_{ij}, \quad C_{rj} \mapsto \beta_j C_{rj}, \quad l_j \mapsto \beta_j^{-1} l_j, \quad u_j \mapsto \beta_j^{-1} u_j,$$

while b_i and d_r remain invariant. We can track this scaling by assigning a B-multidegree in $\mathbb{Z}^n$ to each input coordinate:

$$\begin{aligned}
&\deg_B(c_j) = \deg_B(A_{ij}) = \deg_B(C_{rj}) = e_j,\\
&\deg_B(l_j) = \deg_B(u_j) = -e_j,\\
&\deg_B(b_i) = \deg_B(d_r) = 0,
\end{aligned}$$

where $e_j \in \mathbb{Z}^n$ is the j-th standard basis vector.

By induction on the message-passing layers, it is straightforward to verify that on any fixed open set $\mathcal{U}_\alpha$ in the partition, every coordinate of the LP-GNN output is a linear combination of monomials of the form $M(X) = (\prod_{t=1}^{q} E_t)\, z$, where $q \geq 0$, each $E_t \in \{A_{ik}, C_{rk}\}$ is an edge coefficient, and $z \in \{1, b_i, d_r, c_k, l_k, u_k\}$ is a node feature scalar. The B-degree of such a monomial is precisely $\deg_B(M) = \sum_{t=1}^{q} e_{\kappa(E_t)} + \deg_B(z)$, where $\kappa(E_t)$ denotes the variable index k of the edge E_t. Since $e_{\kappa(E_t)} \geq 0$ componentwise, we always have $\deg_B(M) \geq \deg_B(z)$ componentwise.

The equivariance assumption $f(B \cdot X) = B^{-1} f(X)$ implies that for the j-th output coordinate, $f_j(B \cdot X) = \beta_j^{-1} f_j(X)$. Since this polynomial identity must hold locally on the open set $\mathcal{U}_\alpha$ for all positive diagonal matrices B, every non-zero monomial M in the polynomial representation of f_j on $\mathcal{U}_\alpha$ must strictly satisfy: $\deg_B(M) = -e_j$.

We now determine all admissible monomials. Because $\deg_B(M) \geq \deg_B(z)$ and $-e_j$ has a strictly negative component, the anchor feature z must have a negative degree. The only input features with negative degrees are l_k and u_k, both of which have degree $-e_k$. Thus, $z \in \{l_k, u_k\}$, and the degree constraint becomes: $\sum_{t=1}^{q} e_{\kappa(E_t)} - e_k = -e_j$.

Because $\sum_{t=1}^{q} e_{\kappa(E_t)} \geq 0$ componentwise, the only valid solution to this equation is $k = j$ and $q = 0$. This implies that the monomial contains no edge factors and must be identically l_j or u_j (up to a constant). Consequently, on every open set $\mathcal{U}_\alpha$ in the partition, f_j takes the explicit local affine form:

$$f_j(X) = \rho_{\alpha,j} l_j + \tau_{\alpha,j} u_j,$$

for some constants $\rho_{\alpha,j}, \tau_{\alpha,j} \in \mathbb{R}$.

Finally, we extend this local conclusion to a global one. Let $y := (c, A, b, C, d)$ denote all input coordinates except the bounds, and let D be the dimension of the vector y. Fix finite bounds $\xi := (l, u) \in \mathbb{R}^n \times \mathbb{R}^n$. For each output coordinate j, define $h_{j,\xi}(y) := f_j(y, \xi)$. From the degree argument above, on every activation-region interior $\mathcal{U}_\alpha$, the function f_j has the form $f_j(X) = \rho_{\alpha,j} l_j + \tau_{\alpha,j} u_j$. Therefore, $\nabla_y h_{j,\xi}(y) = 0$ for all y such that $(y, \xi) \in \cup_{\alpha \in I} \mathcal{U}_\alpha$. For the fixed slice ξ, define the corresponding open sets in $\mathbb{R}^D$ as

$$\mathcal{U}_{\alpha,\xi} := \{y \in \mathbb{R}^D : (y, \xi) \in \mathcal{U}_\alpha\}.$$

Let $\mathcal{Z}_\xi$ denote the complement of the union of these open sets in the fixed-bound slice, strictly defined as

$$\mathcal{Z}_\xi := \mathbb{R}^D \setminus \cup_{\alpha \in I} \mathcal{U}_{\alpha,\xi} = \{y \mid (y, \xi) \notin \cup_{\alpha \in I} \mathcal{U}_\alpha\}.$$

Note that $\{(y, \xi) \mid (y, \xi) \notin \cup_{\alpha \in I} \mathcal{U}_\alpha\}$ is measurable and has measure 0. (By Fubini theorem, we can see that

$$0 = \int_\xi \int_y \mathbf{1}_{(y,\xi) \notin \cup_{\alpha \in I} \mathcal{U}_\alpha} dy d\xi = \int_\xi \mu_D(\mathcal{Z}_\xi) d\xi.$$

Hence, $\mu_D(\mathcal{Z}_\xi) = 0$ for almost every ξ, where μ_D denotes the D-dimensional Lebesgue measure.)

We now show that $h_{j,\xi}$ is globally constant in y for almost every ξ. Take any ξ such that $\mu_D(\mathcal{Z}_\xi) = 0$. Let $K \subset \mathbb{R}^D$ be an arbitrary compact box. Define

$$G_K := \{(y_1, y_2) \in K \times K : \mu_1(\{t \in [0,1] : (1-t)y_1 + t y_2 \in \mathcal{Z}_\xi\}) = 0\},$$

where μ_1 is the 1-dimensional Lebesgue measure. We claim that G_K has full Lebesgue measure in $K \times K$. To prove this, consider the indicator function $\mathbf{1}_E(y_1, y_2, t)$ of the set

$$E = \{(y_1, y_2, t) \in K \times K \times [0,1] : (1-t)y_1 + t y_2 \in \mathcal{Z}_\xi\}.$$

By Tonelli's (or Fubini's) theorem, we can compute the measure of E by iterating the integrals in two different orders. First, integrating over $K \times K$ and then over t yields

$$\mu_{2D+1}(E) = \int_0^1 \Big(\int_{K \times K} \mathbf{1}_E(y_1, y_2, t)\, d\mu_{2D}(y_1, y_2) \Big) dt.$$

For any fixed $t \in (0, 1)$, the mapping $(y_1, y_2) \mapsto (1-t)y_1 + t y_2$ is a surjective linear transformation from $\mathbb{R}^{2D}$ to $\mathbb{R}^D$. Since $\mu_D(\mathcal{Z}_\xi) = 0$, the preimage of $\mathcal{Z}_\xi$ under this map has μ_{2D}-measure zero.

(This can be shown by using Fubini theorem again for fixed y_1) Thus, the inner integral is 0 for almost all t, giving

$$\mu_{2D+1}(E) = \int_0^1 0\, dt = 0.$$

Now, applying Fubini's theorem to reverse the order of integration, we have

$$\begin{aligned}\mu_{2D+1}(E) &= \int_{K\times K} \big(\int_0^1 \mathbf{1}_E(y_1, y_2, t)\, dt\big) d\mu_{2D}(y_1, y_2) \\ &= \int_{K\times K} \mu_1\big(\{t \in [0,1] : (1-t)y_1 + ty_2 \in \mathcal{Z}_\xi\}\big)\, d\mu_{2D}(y_1, y_2).\end{aligned}$$

Since the overall integral evaluates to 0 and the integrand is non-negative, the integrand must be 0 for μ_{2D}-almost all $(y_1, y_2) \in K \times K$. By definition, the set of points where this integrand is strictly positive is exactly the complement $(K \times K) \setminus G_K$. Hence, we rigorously conclude $\mu_{2D}\big((K \times K) \setminus G_K\big) = 0$.

Take any $(y_1, y_2) \in G_K$, and define the line segment $\gamma(t) := (1-t)y_1 + ty_2$ for $t \in [0,1]$. The LP-GNN output is locally Lipschitz, because it is obtained by finitely many affine maps, coordinate multiplications by edge coefficients, finite sums, and piecewise linear activations. Hence $h_{j,\xi} \circ \gamma$ is absolutely continuous on $[0,1]$. Moreover, for almost every $t \in [0,1]$, we have $\gamma(t) \notin \mathcal{Z}_\xi$, and at such points $\gamma(t)$ lies in an activation-region interior where $\nabla_y h_{j,\xi} = 0$. Therefore, for almost every t,

$$\frac{d}{dt} h_{j,\xi}(\gamma(t)) = \nabla_y h_{j,\xi}(\gamma(t)) \cdot (y_2 - y_1) = 0.$$

By the fundamental theorem of calculus for absolutely continuous functions,

$$h_{j,\xi}(y_2) - h_{j,\xi}(y_1) = \int_0^1 \frac{d}{dt} h_{j,\xi}(\gamma(t))\, dt = 0.$$

Thus $h_{j,\xi}(y_1) = h_{j,\xi}(y_2)$ for every $(y_1, y_2) \in G_K$. Since G_K has full measure in $K \times K$, it is dense in $K \times K$. For arbitrary $(y_1, y_2) \in K \times K$, choose a sequence $(y_1^{(r)}, y_2^{(r)}) \in G_K$ such that $(y_1^{(r)}, y_2^{(r)}) \to (y_1, y_2)$. For every r, $h_{j,\xi}(y_1^{(r)}) = h_{j,\xi}(y_2^{(r)})$. Taking $r \to \infty$ and using the global continuity of $h_{j,\xi}$, we obtain $h_{j,\xi}(y_1) = h_{j,\xi}(y_2)$. Since K was arbitrary, $h_{j,\xi}$ is constant on the whole coefficient space. Hence, for almost every ξ,$h_{j,\xi}$ is constant on the whole coefficient space. For any possible ξ and any possible (y_1, y_2), we can take a sequence of points $\xi^{(r)}$ with $\mu_D(\mathcal{Z}_{\xi^{(r)}}) = 0$. Then we have

$$h_{j,\xi}(y_1) = \lim_{r\to+\infty} h_{j,\xi^{(r)}}(y_1) = \lim_{r\to+\infty} h_{j,\xi^{(r)}}(y_2) = h_{j,\xi}(y_2)$$

by global continuity. So for all possible ξ,$h_{j,\xi}$ is constant on the whole coefficient space.

Therefore, for each fixed finite (l, u), the value of $f_j(c, A, b, C, d, l, u)$ is independent of c, A, b, C, d. Hence there exists a function $q_j : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ such that $f_j(c, A, b, C, d, l, u) = q_j(l, u)$. The continuity of q_j follows by evaluating f_j on a fixed continuous section, for example $q_j(l, u) := f_j(0, 0, 0, 0, 0, l, u)$. Combining all output coordinates, define $g(l, u) := (q_1(l, u), \ldots, q_n(l, u))$. Then $f(c, A, b, C, d, l, u) = g(l, u)$. This completes the proof of Theorem C.9. □

Now, we extend the activation function to general activation functions like tanh, sigmoid, and GELU. We retain the LP-GNN used above (Section A.4). Still, we introduce some concepts first.

Definition C.11. *An activation σ has a bounded homogeneous recession remainder if there exists a continuous, positive homogeneous, finite piecewise-linear map σ^∞ such that*

$$\sigma^\infty(ts) = t\sigma^\infty(s) \quad (t \geq 0), \qquad \sup_{s\in\mathbb{R}} |\sigma(s) - \sigma^\infty(s)| < \infty.$$

For the stated activations, ReLU and leaky ReLU are already positively homogeneous; sigmoid and $\tanh$ are bounded and have recession map zero. GELU, SiLU, and softplus have recession map

$s_+ = \max\{s, 0\}$. Indeed, writing $t = |s|$, then we have

$$|\operatorname{SiLU}(s) - s_+| = \frac{t}{1+e^t} \le e^{-1},$$
$$|\operatorname{softplus}(s) - s_+| = \log(1+e^{-t}) \le \log 2,$$
$$|\operatorname{GELU}(s) - s_+| = t\Phi_{\mathrm{G}}(-t) \le \int_t^\infty v\phi_{\mathrm{G}}(v)\,dv = \phi_{\mathrm{G}}(t) \le (2\pi)^{-1/2},$$

where Φ_{G} and ϕ_{G} are the standard Gaussian distribution function and density, and $\operatorname{GELU}(s) = s\Phi_{\mathrm{G}}(s)$.

Theorem C.12. *Suppose every activation has a bounded homogeneous recession remainder. If a LP-GNN f satisfies $f(B \cdot X) = B^{-1}f(X)$ for every $X \in \mathbb{P}_{m,n,k}$ and $B \in \operatorname{Diag}_{n\times n}(\mathbb{R}_{++})$, then, for each j, there exists a continuous function $q_j : \mathbb{R}^2 \to \mathbb{R}$ such that $f_j(X) = q_j(l_j, u_j)$. Thus the prediction for variable j is independent of (c, A, b, C, d) and of the bounds of all other variables. Meanwhile, q_j is positive homogeneous.*

Proof. First consider the subgroup $B = \beta I$. Evaluate the network at $(\beta I) \cdot X$, and denote each intermediate hidden state vector by h_β. We show inductively that βh_β converges as $\beta \downarrow 0$.

Every non-bound node feature has normalized limit zero, including the objective c, right-hand sides b, d, and fixed indicators. The bound features instead satisfy $\beta(\beta^{-1}l) = l, \beta(\beta^{-1}u) = u$. Affine maps preserve convergence, and their biases vanish. If $\beta h_\beta \to H$, the recession assumption gives

$$\beta\sigma(h_\beta) = \sigma^\infty(\beta h_\beta) + r_\beta \longrightarrow \sigma^\infty(H),\ |r_\beta| \le C_\sigma\beta \text{ for some constant } C_\sigma$$

For a stored edge coefficient $E \in \{A_{ij}, C_{rj}\}$, we also have $\beta[(\beta E)h_\beta] = \beta E(\beta h_\beta) \longrightarrow 0$. Finite aggregation and concatenation also preserve these limits. Consequently,

$$\lim_{\beta\downarrow 0} \beta f((\beta I)\cdot X) = Q(l, u),$$

where Q is computed by retaining only the bound inputs, discarding biases and edge products, and replacing each activation by its homogeneous recession map. Hence Q is continuous and positive homogeneous. Exact equivariance yields

$$f(X) = \lim_{\beta\downarrow 0} \beta f((\beta I)\cdot X) = Q(l, u).$$

To obtain the coordinatewise conclusion, fix j and let B_t have jth diagonal entry one and every other diagonal entry t. Diagonal equivariance gives

$$Q_j(l, u) = f_j(X) = f_j(B_t \cdot X) = Q_j(B_t^{-1}l, B_t^{-1}u).$$

Letting $t \to \infty$ and using continuity, $f_j(X) = Q_j(l_j e_j, u_j e_j)$. Define $q_j(a, b) := Q_j(ae_j, be_j)$. This proves the claimed representation, and the stated regularity and homogeneity properties follow directly from those of Q. □

Combining all the theorems above gives the following corollary.

Corollary C.13. *For any LP-GNN $f_{GNN} : \mathbb{P}_{m,n,k} \to \mathbb{R}^n$ with common activations listed above, if $f(c, A, b, C, d, l, u) = \beta f(\beta c, \beta A, b, \beta C, d, \beta^{-1}l, \beta^{-1}u)$ for all $\beta \in \mathbb{R}_{++}$, then there exists g, such that $f(c, A, b, C, d, l, u) = g(l, u)$.*

Proof. Since the LP-GNN structure is the same, it suffices to proceed using the same proof of Theorem C.9 while changing the degree with respect to B to the degree with respect to β □

Remark C.14. *Here, we only consider the finite (l, u) case since it's contained in the general case where (l, u) can has infinity values. Practically, people always use a fixed placeholder vector a to denote $+\infty$ and its opposite value to denote $-\infty$. In that situation, we adopt the convention that $h \cdot a = a$ and $h \cdot -a = -a$ for $h > 0$. This would lead to the following corollary for standard form LPs.*

Below, we only consider the standard form LPs, which have the constraint $Ax = b, x \ge 0$. Let $S_{n,m} = \{(c, A, b, 0^n, (+\infty)^n) \mid \min c^T x, s.t.\ Ax = b, x \ge 0\}$ be the space of standard form LPs. Under this case, we have a stronger result.

Corollary C.15. *For any continuous function $f : S_{n,m} \to \mathbb{R}^n$:*

- *If $f(c, A, b, 0^n, (a)^n) = f(\alpha c, A, b, 0^n, (a)^n)$ for all $\alpha > 0$, then $f(c, A, b, 0^n, (a)^n) = g(A, b)$ for all $c \in \mathbb{R}^n$ and some function g.*
- *If $f(c, A, b, 0^n, (a)^n) = f(c, RA, Rb, 0^n, (a)^n)$ for all $R \in Diag_{m\times m}(\mathbb{R}_{++})$, then $f(c, A, b, 0^n, (a)^n) = g(c)$ for all $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$ and some function g.*
- *If $f(c, A, b, 0^n, (a)^n) = \beta f(\beta c, \beta A, b, 0^n, (a)^n)$ for all $\beta \in \mathbb{R}_{++}$, then $f(c, A, b, 0^n, (a)^n) \equiv 0$.*

Proof. The first two terms' proof is the same as the Theorem C.1. For the third term, it suffices to note that

$$f(c, A, b, 0^n, (a)^n) = \lim_{\beta\to 0^+} \beta f(\beta c, \beta A, b, 0^n, (a)^n) = \lim_{\beta\to 0^+} \beta \cdot \lim_{\beta\to 0^+} f(\beta c, \beta A, b, 0^n, (a)^n) = 0$$

□

Insights for algorithm-unrolling GNNs. Beyond WL-based approaches, several GNN models solve LPs or warm-start solvers by simulating optimization algorithms, including interior-point methods (IPMs) (Nocedal & Wright, 2006; Qian et al., 2023) and primal–dual hybrid gradient (PDHG) methods (Li et al., 2024a; Applegate et al., 2022; 2023). Algorithm unrolling does not generally inherit the rescaling symmetry of the LP solution map. For example, fixed-step PDHG is not rescaling-equivariant when the same step sizes are retained after rescaling. In contrast, the IPM simulated by Qian et al. (2023) is equivariant under consistently transformed initial states. Specifically, for $g \cdot (c, A, b) = (\alpha Bc, RAB, Rb)$, its state transforms as $T_g(x, w, s, r, \mu) = (B^{-1}x, \alpha R^{-1}w, \alpha Bs, Rr, \alpha\mu)$, and each exact IPM iteration commutes with T_g; the property therefore propagates across iterations under an equivariant initialization and rescaling-invariant stopping rule.

Nevertheless, algorithmic equivariance is neither necessary nor sufficient for the desired inductive bias of an unrolled GNN. It is not necessary because an unrolled model imitates a finite solver trajectory, which numerical prescaling may deliberately change to improve conditioning and reduce iterations while preserving the underlying LP. Hence, numerical scaling may make the whole solver trajectory easy to learn. It is not sufficient because an equivariant algorithm does not imply that its GNN simulator can enforce the same symmetry. Indeed, the multidegree argument above extends to the state-augmented piecewise-linear GCN/GIN constructions of Qian et al. (2023). Below, x_j denotes the current primal iterate stored in the GNN state, rather than an LP coefficient. Under $B = \mathrm{Diag}(\beta_1, \ldots, \beta_n)$, x_j scales as $\beta_j^{-1}x_j$ to make the algorithm equivariant, whereas the numerical LP coefficients c_j and A_{ij} carry only nonnegative powers of the β_j's. Since message passing can only introduce additional coefficient factors, it cannot generate the required negative scaling degree from (c, A, b). Thus, exact equivariance forces the j-th output to take the degenerate form $\kappa_j x_j$, where κ_j is independent of the numerical LP data, ruling out any nontrivial coefficient-dependent IPM update.

Insights for data augmentation. Rescaling augmentation encourages equivariance by training on transformed instances, providing a finite-sample proxy for reducing

$$\Delta_{\mathcal{N}}(f; X, g) := \|f(g \cdot X) - \rho_Y(g)f(X)\|.$$

This is related to averaging the training objective over transformation orbits (Chen et al., 2020). However, under the assumptions of the preceding theorems, exact objective- and row-scaling invariance removes dependence on the corresponding coefficients, while exact variable-scaling equivariance restricts the specified LP-GNNs to bound-only predictions. Augmentation can therefore improve robustness over finite scale ranges, but it cannot remove the architecture's expressive limitations under exact symmetry.

C.2 The degeneracy when enforcing the symmetry locally

For LPs with bounded coefficients, some might consider enforcing rescaling symmetry only over a local numerical range. They may hope this could make the former degeneracy theory fail. Here,

local symmetry on a dataset $\mathcal{D}$ means that, for every $X \in \mathcal{D}$, the required equivalence holds for all rescalings in some open neighborhood of the identity $(1, I_m, I_k, I_n)$. Below, we call a function real analytic if it locally admits a convergent power-series representation.

Theorem C.16. *For all LP-GNNs f, the following results hold:*

- *If all activations are piecewise-linear with finitely many pieces, local diagonal variable equivariance on any dataset $\mathcal{D}$ implies*

$$f_j(X) \in \{a_{rj} l_j + b_{rj} u_j : 1 \le r \le N_j\},\ \forall X \in \mathcal{D},$$

  for finitely many fixed coefficients a_{rj}, b_{rj}. Consequently, f_j is constant on each connected subset of $\mathcal{D}$ with (l_j, u_j) fixed. Under local scalar variable equivariance, the retained formulas are linear combinations of all bounds, and f is constant on connected subsets with (l, u) fixed.

- *If all activations are real analytic on $\mathbb{R}$ and satisfy the recession assumption of Theorem C.12, local diagonal variable equivariance on an open dataset $\mathcal{D}$ implies*

$$f_j(X) = a_j l_j + b_j u_j$$

  for fixed constants a_j, b_j. Under local scalar variable equivariance on an open dataset $\mathcal{D}$,

$$f(X) = L_l l + L_u u$$

  for fixed matrices L_l, L_u.

Local objective or row invariance likewise removes dependence on the corresponding coefficient block: on connected subsets with all other fields fixed in the first case, and throughout the real coordinate space in the second.

Proof. For the first case, the final output of GNN takes values among finitely many polynomials $p_1, \dots, p_N$. As established in the proof of Theorem C.9, each monomial term in such polynomials contains stored edge coefficients from A, C and at most one node-feature factor from $(1, c, l, u, b, d)$.

Fix X and an output coordinate j, set $B(t) = \mathrm{Diag}(e^{t_1}, \dots, e^{t_n})$. The union of closed zero sets of $p_r(B(t) \cdot X) - e^{-t_j} f_j(X)$ ranging over $r \in [N]$ covers an open neighborhood of zero. At least one contains an open subset; otherwise, we can see that the complement of the zero set of $p_r(B(t) \cdot X) - e^{-t_j} f_j(X)$ is dense and open for each r. Then we could intersect these dense open sets and find that the intersection is indeed non-empty, which contradicts the covering fact. Without loss of generality, let r_1 be the index such that the zero set of $p_{r_1}(B(t) \cdot X) - e^{-t_j} f_j(X)$ contains a non-empty interior.

For a monomial M with B-multidegree $\nu = \deg_B(M)$, substituting the formula, we have: $M(B(t) \cdot X) = e^{\nu^\top t} M(X)$. Moreover, for distinct weights $\lambda_1, \dots, \lambda_R$,

$$\sum_{r=1}^{R} a_r e^{\lambda_r^\top t} = 0 \text{ on a nonempty open set } U \implies a_1 = \cdots = a_R = 0.$$

Indeed, choose $t_0 \in U$ and v such that the numbers $\lambda_r^\top v$ are pairwise distinct. Differentiating along $t = t_0 + sv$ at $s = 0$ gives $\sum_{r=1}^{R} a_r e^{\lambda_r^\top t_0} (\lambda_r^\top v)^k = 0, k = 0, \dots, R-1$. The resulting Vandermonde matrix is invertible, proving the implication. Applying this independence to the zero set of $p_{r_1}(B(t) \cdot X) - e^{-t_j} f_j(X)$ shows that $f_j(X)$ equals the weight-$-e_j$ component of p_{r_1}. By the multidegree argument of Theorem C.9, this component is a linear combination of l_j, u_j; under scalar scaling, it is a linear combination of all bounds. Fixing the corresponding bounds restricts the prediction to finitely many values. Continuity therefore forces constancy on each connected subset with those bounds fixed.

For the second case, note that f is real analytic. Let U be the nonempty open set of LP inputs on which local equivariance holds. Differentiating $f_j(B(t) \cdot X) = e^{-t_j} f_j(X)$ at $t = 0$ gives $\mathcal{L}_k f_j(X) = -\delta_{jk} f_j(X), X \in U$, where $\mathcal{L}_k := c_k \partial_{c_k} + \sum_i A_{ik} \partial_{A_{ik}} + \sum_r C_{rk} \partial_{C_{rk}} - l_k \partial_{l_k} - u_k \partial_{u_k}$. Here $t = 0$ denotes the identity rescaling. For scalar equivariance, differentiation instead gives $\mathcal{L} f_j = -f_j$ on U, where $\mathcal{L} = \sum_k \mathcal{L}_k$.

Each residual $\mathcal{L}_k f_j + \delta_{jk} f_j$, or $\mathcal{L} f_j + f_j$ in the scalar case, is analytic and zero on U. By the identity theorem for real-analytic functions, an analytic function on a connected open domain that vanishes on a nonempty open subset vanishes identically. Each residual is analytic on the entire real coefficient space and vanishes on U. Hence the differential identities hold throughout that space.

The resulting differential equations give the scaling law for arbitrary positive scales. Specifically, let $B_k(s)$ have kth diagonal entry e^s and all other entries one. For every $s \in \mathbb{R}$, note that

$$\frac{d}{ds}\left[e^{\delta_{jk} s} f_j(B_k(s) \cdot X)\right] = e^{\delta_{jk} s}(\mathcal{L}_k f_j + \delta_{jk} f_j)(B_k(s) \cdot X) = 0.$$

This proves equivariance under $B_k(s)$. Every positive diagonal matrix is a product of these matrices, so diagonal equivariance holds globally. In the scalar case,

$$\frac{d}{ds}\left[e^s f_j((e^s I) \cdot X)\right] = e^s(\mathcal{L} f_j + f_j)((e^s I) \cdot X) = 0,$$

which proves global scalar equivariance. The proof of Theorem C.12 then establishes the continuous limit $Q(l, u) = \lim_{\beta \downarrow 0} \beta f((\beta I) \cdot X)$. Global scalar equivariance makes the expression inside this limit equal to $f(X)$, hence $f(X) = Q(l, u)$. In particular, $Q(l, u) = f(0, 0, 0, 0, 0, l, u)$, so Q is an analytic restriction of f and is differentiable at zero. Writing $v = (l, u)$, positive homogeneity gives $Q(0) = 0$ and $DQ(0)v = \lim_{t \downarrow 0} \frac{Q(tv) - Q(0)}{t} = Q(v)$, where $DQ(0)$ is the Jacobian matrix at 0. Therefore Q is linear: $f(X) = L_l l + L_u u$ for fixed matrices L_l, L_u.

Since $f(X) = L_l l + L_u u$ and diagonal rescaling sends (l, u) to $(B^{-1} l, B^{-1} u)$, equivariance requires $L_l B^{-1} l + L_u B^{-1} u = B^{-1}(L_l l + L_u u)$. Taking the jth coordinate and collecting terms gives

$$\sum_k (\beta_k^{-1} - \beta_j^{-1})\big((L_l)_{jk} l_k + (L_u)_{jk} u_k\big) = 0.$$

This identity holds for all bound vectors in the real coordinate space, so the coefficient of each l_k and u_k must vanish: $(\beta_k^{-1} - \beta_j^{-1}) L_{jk} = 0, L \in \{L_l, L_u\}$. For $j \neq k$, choosing $\beta_j \neq \beta_k$ forces $L_{jk} = 0$. Thus both matrices are diagonal, yielding $f_j(X) = a_j l_j + b_j u_j$.

Finally, objective or row scaling assigns positive weight to the affected coefficients and zero weight to all others. In the finite-piece case, weight-zero polynomial components omit the affected block, and the connected-set argument applies. In the analytic case, similar arguments show that local invariance would lead to global invariance; then Theorem C.1 gives the corresponding independence as well. □

The analytic result includes sigmoid, tanh, GELU, SiLU, and softplus, while the piecewise result includes the ReLU and leaky ReLU. Hence, we successfully extend our theory to the local version.

D THE EXISTENCE OF THE SINGLE VALUED CANONICALIZATION

Theorem D.1. *Let $\Sigma_X = \{k \in \mathcal{K} \mid k \cdot X \in \mathcal{N} \cdot X\}$. Then X is canonicalizable if and only if $\exists Y \in \mathcal{N} \cdot X$, such that $\forall k' \in \Sigma_X, k' \cdot Y = Y$.*

Proof. We need some lemmas to finish the proof.

Lemma D.2. *For each element $p \in \mathcal{K}$, we have: $p \cdot \mathcal{N} \cdot X = \mathcal{N} \cdot p \cdot X$.*

Proof. Without loss of generality, assume $p = (P_1, P_2, P_3)$, where $P_1 \in S_m$, $P_2 \in S_n$, $P_3 \in S_k$. It suffices to show that for any $g = (\alpha, R_{\leq}, R_{=}, B) \in \mathcal{N}$, there exists $g' = (\alpha', R'_{\leq}, R'_{=}, B') \in \mathcal{N}$ such that $p \cdot (g \cdot X) = g' \cdot (p \cdot X)$, and conversely, that for any $g' \in \mathcal{N}$, there exists $g \in \mathcal{N}$ satisfying the same equality. Let $X = (c, A, b, C, d, l, u)$. The action of $g = (\alpha, R_{\leq}, R_{=}, B)$ yields

$$g \cdot X = (\alpha B c, R_{\leq} A B, R_{\leq} b, R_{=} C B, R_{=} d, B^{-1} l, B^{-1} u),$$

and the subsequent action by $p = (P_1, P_2, P_3)$ gives

$$p \cdot (g \cdot X) = (P_2(\alpha B c), P_1(R_{\leq} A B) P_2^T, P_1(R_{\leq} b), P_3(R_{=} C B) P_2^T, P_3(R_{=} d), P_2 B^{-1} l, P_2 B^{-1} u).$$

Since α is a scalar and P_1, P_2, P_3 are orthogonal permutation matrices (i.e., $P_i^{-1} = P_i^T$ for all i), we can rewrite these components by inserting identity matrices. Specifically, the objective vector becomes

$$P_2(\alpha Bc) = \alpha(P_2 B P_2^T)P_2 c,$$

the constraint matrices become

$$P_1(R_{\leq}AB)P_2^T = (P_1 R_{\leq} P_1^T)(P_1 A P_2^T)(P_2 B P_2^T)$$

and

$$P_3(R_{=}CB)P_2^T = (P_3 R_{=} P_3^T)(P_3 C P_2^T)(P_2 B P_2^T),$$

the constraint vectors become

$$P_1(R_{\leq}b) = (P_1 R_{\leq} P_1^T)P_1 b$$

and

$$P_3(R_{=}d) = (P_3 R_{=} P_3^T)P_3 d,$$

and the variable bounds become

$$P_2 B^{-1} l = (P_2 B P_2^T)^{-1} P_2 l \quad \text{and} \quad P_2 B^{-1} u = (P_2 B P_2^T)^{-1} P_2 u.$$

We can then construct

$$g' = (\alpha', R'_{\leq}, R'_{=}, B') = (\alpha, P_1 R_{\leq} P_1^T, P_3 R_{=} P_3^T, P_2 B P_2^T).$$

Because conjugating a diagonal matrix by a permutation matrix merely rearranges its diagonal entries, $R'_{\leq}, R'_{=}$, and B' remain valid positive diagonal matrices, ensuring $g' \in \mathcal{N}$. With this construction, our expanded expression exactly matches the action of g' on $p \cdot X$, which is

$$g' \cdot (P_2 c, P_1 A P_2^T, P_1 b, P_3 C P_2^T, P_3 d, P_2 l, P_2 u) = g' \cdot (p \cdot X).$$

This establishes that $p \cdot \mathcal{N} \cdot X \subseteq \mathcal{N} \cdot p \cdot X$.

To prove the reverse inclusion $\mathcal{N} \cdot p \cdot X \subseteq p \cdot \mathcal{N} \cdot X$, we consider an arbitrary element $g' = (\alpha', R'_{\leq}, R'_{=}, B') \in \mathcal{N}$ acting on $p \cdot X$. By naturally reversing the conjugation, we choose $g = (\alpha', P_1^T R'_{\leq} P_1, P_3^T R'_{=} P_3, P_2^T B' P_2) \in \mathcal{N}$. An identical algebraic expansion confirms that $p \cdot (g \cdot X) = g' \cdot (p \cdot X)$. Combining both sides, we get the result $p \cdot \mathcal{N} \cdot X = \mathcal{N} \cdot p \cdot X$. □

Lemma D.3. *Let $\mathcal{O}(X) := \{n_1 k_1 \cdots n_m k_m \cdot X \mid m \in \mathbb{N}; n_i \in \mathcal{N}, k_i \in \mathcal{K}, \forall i \in [m]\}$ be all the possible orbit of X under the action of $\mathcal{N} \cup \mathcal{K}$. Then for any $X \neq Y$, $\mathcal{O}(X), \mathcal{O}(Y)$ are either the same or disjoint.*

Proof. Suppose that $\mathcal{O}(X)$ and $\mathcal{O}(Y)$ are not disjoint, meaning there exists some element $Z \in \mathcal{O}(X) \cap \mathcal{O}(Y)$. Thus, there exist finite sequences of actions $g_X = n_1 k_1 \cdots n_m k_m$ and $g_Y = n'_1 k'_1 \cdots n'_l k'_l$ (where all $n_i, n'_i \in \mathcal{N}$ and $k_i, k'_i \in \mathcal{K}$) such that $Z = g_X \cdot X = g_Y \cdot Y$. Because $\mathcal{N}$ and $\mathcal{K}$ are both well-defined groups, every individual scaling matrix and permutation matrix is invertible, and their respective inverses still belong to $\mathcal{N}$ and $\mathcal{K}$. Consequently, the composed action g_X is also invertible, and its inverse $g_X^{-1} = k_m^{-1} n_m^{-1} \cdots k_1^{-1} n_1^{-1}$ remains a finite sequence of elements from $\mathcal{N}$ and $\mathcal{K}$. Applying g_X^{-1} to both sides of our equality yields $X = g_X^{-1} \cdot (g_Y \cdot Y) = (g_X^{-1} g_Y) \cdot Y$. Since the concatenated product $g_X^{-1} g_Y$ is simply another finite alternating sequence of actions from $\mathcal{N}$ and $\mathcal{K}$, this construction explicitly proves that $X \in \mathcal{O}(Y)$. Now, to prove set equality, consider any arbitrary element $W \in \mathcal{O}(X)$. By definition, $W = g_W \cdot X$ for some valid finite sequence of actions g_W. Substituting our derived expression for X, we obtain $W = g_W \cdot ((g_X^{-1} g_Y) \cdot Y) = (g_W g_X^{-1} g_Y) \cdot Y$. Because the combined operation $g_W g_X^{-1} g_Y$ is still formed entirely by elements from $\mathcal{N}$ and $\mathcal{K}$, it immediately follows that $W \in \mathcal{O}(Y)$, thereby establishing the subset relation $\mathcal{O}(X) \subseteq \mathcal{O}(Y)$. By a perfectly symmetric argument, we can similarly obtain the reverse inclusion $\mathcal{O}(Y) \subseteq \mathcal{O}(X)$. Therefore, $\mathcal{O}(X) = \mathcal{O}(Y)$. [4] □

[4] By adding identity elements in the action set, it's obvious that any finite order of elements from $\mathcal{N} \cup \mathcal{K}$ can be written in the alternating form above.

$\Leftarrow$: Now, let κ be the canonicalization defined by $\kappa(n_1k_1\cdots n_mk_m\cdot x) = k_1\cdots k_m\cdot Y$, where $m\in\mathbb{N}$, and $n_i\in\mathcal{N}, k_i\in\mathcal{K}$ for $i\in[m]$. We first show that κ is well-defined. By adding identity elements in the group action to make the sequence have the same length while having the form $nkn'k'\cdots$, it suffices to prove that: for any $m\in\mathbb{N}$, and $n_i,n_i'\in\mathcal{N}, k_i,k_i'\in\mathcal{K}$ for $i\in[m]$, $n_1k_1\cdots n_mk_m\cdot x = n_1'k_1'\cdots n_m'k_m'\cdot x$ implies $\kappa(n_1k_1\cdots n_mk_m\cdot x) = \kappa(n_1'k_1'\cdots n_m'k_m'\cdot x)$. By lemma D.2, there exists $n_m''\in\mathcal{N}$, such that

$$k_{m-2}n_{m-1}k_{m-1}n_m\cdot(k_m\cdot x) = k_{m-2}n_{m-1}n_m''k_{m-1}\cdot(k_m\cdot x).$$

Apply lemma D.2 repeatedly, we obtain:

$$n_1k_1\cdots n_mk_m\cdot x = n_ok_1\cdots k_m\cdot x = n_o'k_1'\cdots k_m'\cdot x = n_1'k_1'\cdots n_m'k_m'\cdot x$$

for some $n_o,n_o'\in\mathcal{N}$. Hence,

$$(k_1\cdots k_m)^{-1}(n_o^{-1}n_o')k_1'\cdots k_m'\cdot x = x.$$

Then, by using lemma D.2 again, we have

$$n_l(k_1\cdots k_m)^{-1}k_1'\cdots k_m'\cdot x = x$$

for some $n_l\in\mathcal{N}$. As a result, applying n_l^{-1} on both side, we get: $(k_1\cdots k_m)^{-1}k_1'\cdots k_m'\in\Sigma_x$. Thus $(k_1\cdots k_m)^{-1}k_1'\cdots k_m'\cdot Y = Y$, which implies that

$$k_1'\cdots k_m'\cdot Y = k_1\cdots k_m\cdot Y.$$

Hence κ is well defined for $\mathcal{O}(x)$. By the definition, it's easy to check that κ is invariant under $\mathcal{N}$ and equivariant under $\mathcal{K}$ over $\mathcal{O}(x)$. Moreover, since $Y\in\mathcal{N}\cdot x$, then by applying lemma D.2, we have: $\mathcal{N}\cdot n_1k_1\cdots n_mk_m\cdot x = \mathcal{N}\cdot k_1\cdots n_mk_m\cdot x = k_1\cdot\mathcal{N}n_2k_2\cdots n_mk_m\cdot x = \cdots = k_1k_2\cdots k_m\mathcal{N}\cdot x$. Therefore, $k_1\cdots k_mY\in k_1k_2\cdots k_m\mathcal{N}\cdot x = \mathcal{N}\cdot n_1k_1\cdots n_mk_m\cdot x$, which ensures that κ is a permutation-equivariant rescaling orbit canonicalization over $\mathcal{O}(x)$. By lemma D.3, for other $z\in\mathcal{O}(w)\neq\mathcal{O}(x)$, we define $\kappa(z) = \mathcal{N}\cdot z$. By the definition and lemma D.2 and D.3, we can see that: κ is also a well-defined permutation-equivariant rescaling orbit canonicalization over $\mathbb{P}_{m,n,k}\setminus\mathcal{O}(x)$. Combining all the above, we have: κ is a well-defined permutation-equivariant rescaling orbit canonicalization and $|\kappa(x)| = 1$. Hence, x is canonicalizable.

$\Rightarrow$: Now, assume x is canonicalizable. Hence, there exists a permutation-equivariant rescaling orbit canonicalization κ, such that: $\kappa(x) = Y\in\mathbb{P}_{m,n,k}$. For any $k\in\Sigma_x, \exists n_k\in\mathcal{N}, s.t. k\cdot x = n_k\cdot x$. Hence, $Y = \kappa(n_k\cdot X) = \kappa(k\cdot x) = k\cdot Y$.

Combining $\Rightarrow$ and $\Leftarrow$ would quickly finish the proof of Theorem D.1. □

So, to study the canonicalizability of the elements x, we just need to consider the "collision" group Σ_x. Below, we give an elegant existence proof to show that: every LP instance is in fact canonicalizable. To begin with, we note the following basic facts:

Fact 1: The group $\mathcal{N}$ is an abelian group.

Fact 2: The group $\mathcal{N}$ is isomorphic to the group $\mathbb{R}^{m+n+k+1}$

Proof. Define the group binary operations in $\mathbb{R}^{m+n+k+1}$ by

$$\begin{aligned}&(u,r_1,\cdots,r_m,\eta_1,\cdots,\eta_n,\zeta_1,\cdots,\zeta_k)\circ(u',r_1',\cdots,r_m',\eta_1',\cdots,\eta_n',\zeta_1',\cdots,\zeta_k')\\&=(u+u',r_1+r_1',\cdots,r_m+r_m',\eta_1+\eta_1',\cdots,\eta_n+\eta_n',\zeta_1+\zeta_1',\cdots,\zeta_k+\zeta_k')\end{aligned}$$

and define $\phi:\mathbb{R}^{m+n+k+1}\to\mathcal{N}$ by

$$\begin{aligned}&\phi(u,r_1,\cdots,r_m,\eta_1,\cdots,\eta_n,\zeta_1,\cdots,\zeta_k)\\=&(e^u,\mathrm{Diag}(e^{r_1},\cdots,e^{r_m}),\mathrm{Diag}(e^{\eta_1},\cdots,e^{\eta_n}),\mathrm{Diag}(e^{\zeta_1},\cdots,e^{\zeta_k})).\end{aligned}$$

Then, ϕ is a homomorphism from $\mathbb{R}^{m+n+k+1}$ to $\mathcal{N}$. Since ϕ is bijective, we can see that ϕ is an isomorphism. □

Now, we begin to prove the following theorems

Theorem D.4. *Every linear program instance is canonicalizable.*

Proof. Let $x = (c, A, b, C, d, l, u)$; $N_x = \{n \in \mathcal{N} \mid n \cdot x = x\}$ and $W_x = \phi^{-1}(N_x)$. Notice that $(u, r_1, \cdots, r_m, \eta_1, \cdots, \eta_n, \zeta_1, \cdots, \zeta_k) \in W_x$ if and only if :

- If $\exists j \in [n], s.t. c_j \neq 0$, then $u + \eta_j = 0$.
- If $\exists i \in [m], j \in [n], s.t. A_{ij} \neq 0$, then $r_i + \eta_j = 0$.
- If $\exists i \in [m], s.t. b_i \neq 0$, then $r_i = 0$
- If $\exists i \in [k], j \in [n], s.t. C_{ij} \neq 0$, then $\zeta_i + \eta_j = 0$.
- If $\exists i \in [k], s.t. d_i \neq 0$, then $\zeta_i = 0$
- If $\exists j \in [n], s.t. l_j \neq 0, \infty$ or $u_j \neq 0, \infty$, then $\eta_j = 0$.

Hence, W_x is a linear subspace of $\mathbb{R}^{m+n+k+1}$. Below, we use $V_x = \mathbb{R}^{m+n+k+1}/W_x$ to denote the quotient space. Since each element k in $\mathcal{K}$ is some kind of rearrangement of the orders of constraints and variables, we can naturally define the action of k on space $\mathbb{R}^{m+n+k+1}$ to be the corresponding rearrangement of the rescaling coefficients respectively. And it's not hard to verify that: $k \cdot (\phi(z) \cdot x) = \phi(k \cdot z) \cdot (k \cdot x)$ holds for any $k \in \mathcal{K}$. Next, we prove a lemma to help well-define the action of Σ_x on V_x.

Lemma D.5. *For each $k \in \Sigma_x$, we have: $k \cdot W_x = W_x$, i.e. W_x is invariant under Σ_x.*

Proof. For any $w \in W_x$, we have: $\phi(w) \in N_x$. Meanwhile, since $k \cdot x \in \mathcal{N} \cdot x$, then we can conclude that for any $n \in \mathcal{N}$: $n \cdot k \cdot x = k \cdot x \Leftrightarrow n \cdot x = x$, since $\mathcal{N}$ is commutative(abelian). Since $\phi(k \cdot w) \cdot k \cdot x = k \cdot \phi(w) \cdot x = k \cdot x$, we can see that: $\phi(k \cdot w) \cdot x = x$. Hence, $\phi(k \cdot w) \in N_x$, which implies that $k \cdot w \in W_x$. Since w is arbitrary, we have $k \cdot W_x \subseteq W_x$. Applying lemma D.2, we can see that $k^{-1} \in \Sigma_x$. (There exists $n_k \in \mathcal{N}$ such that $k \cdot x = n_k \cdot x$, which means that $n_k^{-1} \cdot k \cdot x = x$. By Lemma D.2, there exists $n'_k \in \mathcal{N}$ such that $k \cdot n'_k \cdot x = x$, i.e., $n'_k \cdot x = k^{-1} \cdot x$) Hence, $k^{-1} W_x \subseteq W_x$, which implies that $W_x \subseteq k \cdot W_x$. Combining both side would finish the proof. □

Using the lemma, we can well define the group action of Σ_x on V_x in sense that: $k \cdot [a] = k \cdot (a + W_x) = k \cdot a + k \cdot W_x = k \cdot a + W_x = [k \cdot a]$ for any $a \in \mathbb{R}^{m+n+k+1}, [a] \in V_x$. For each $k \in \Sigma_x$, there exists $x_k \in \mathbb{R}^{m+n+k+1}$, such that $k \cdot x = \phi(x_k) \cdot x$. Take $k, l \in \Sigma_x$, then we have: $\phi(x_{kl}) \cdot x = kl \cdot x = k \cdot l \cdot x = k \cdot \phi(x_l) \cdot x = \phi(k \cdot x_l) \cdot k \cdot x = \phi(k \cdot x_l)\phi(x_k) \cdot x = \phi(k \cdot x_l + x_k) \cdot x$. As a result, we have: $\phi(x_{kl} - k \cdot x_l - x_k) \cdot x = x$, which implies $x_{kl} - k \cdot x_l - x_k \in W_x$. So in the quotient space V_x, we have: $[x_{kl}] = [x_k] + [k \cdot x_l] = [x_k] + k \cdot [x_l]$.

Define $[v] = \frac{1}{|\Sigma_x|} \sum_{h \in \Sigma_x} [x_h] \in V_x$. Note that for any $k \in \Sigma_x$, $\sum_{h \in \Sigma_x} [x_{kh}] = \sum_{h \in \Sigma_x} [x_k] + k \cdot [x_h] = |\Sigma_x|[x_k] + k \cdot \sum_{h \in \Sigma_x} [x_h]$. Divide both sides by $|\Sigma_x|$, we get $[v] = [x_k] + k \cdot [v]$. Now we take $\hat{v} \in \mathbb{R}^{m+n+k+1}$ such that: $[\hat{v}] = [v]$. Thus, $\exists w_k \in W_x, s.t. \hat{v} = x_k + k \cdot \hat{v} - w_k$. Let $Y = \phi(\hat{v}) \cdot x$, then for any $k \in \Sigma_x$, it holds that

$$k \cdot Y = k \cdot \phi(\hat{v}) \cdot x = \phi(k \cdot \hat{v}) \cdot k \cdot x = \phi(k \cdot \hat{v} + x_k) \cdot x = \phi(\hat{v} + w_k) \cdot x = \phi(\hat{v}) \cdot x = Y.$$

By theorem D.1, we can conclude that any x is canonicalizable. □

Now, we can prove the main theorem fully.

Theorem D.6. *There exists a permutation-equivariant rescaling orbit canonicalization $\mathcal{C}$ such that $|\mathcal{C}(x)| = 1$ for every $x \in \mathbb{P}_{m,n,k}$.*

Proof. By Lemma D.3, choose a set S containing exactly one representative from each joint orbit. For each $x \in S$, Theorem D.4 provides a canonicalization $\mathcal{C}_x$ with $|\mathcal{C}_x(x)| = 1$. Its rescaling invariance and permutation equivariance imply $|\mathcal{C}_x(z)| = 1$ throughout $\mathcal{O}(x)$.

Define $\mathcal{C}(z) := \mathcal{C}_x(z)$ whenever $z \in \mathcal{O}(x)$ and $x \in S$. The orbit partition makes this definition unambiguous. Since $\mathcal{N}$ and $\mathcal{K}$ preserve each joint orbit, $\mathcal{C}$ inherits rescaling invariance, permutation

equivariance, and the orbit condition from the maps $\mathcal{C}_x$. Moreover, $|\mathcal{C}(z)| = 1$ for every $z \in \mathbb{P}_{m,n,k}$. □

E THE REGULARITY OF CANONICALIZATION

E.1 GLOBAL CONTINUITY OBSTRUCTION FOR CANONICALIZATIONS

Regularity is always a fundamental topic in geometric deep learning. In this section, we show some theoretical results, from continuity to measurability. Below, we briefly show that a global continuous rescaling-invariant orbit canonicalization doesn't exist.

Theorem E.1. *Assume $C : \mathbb{P}_{m,n,k} \to \mathbb{P}_{m,n,k}$ is a single-valued map such that for all $X \in \mathbb{P}_{m,n,k}$ and all $g \in N$,*

$$C(X) \in N \cdot X, \qquad C(g \cdot X) = C(X).$$

Then C cannot be continuous on all of $\mathbb{P}_{m,n,k}$.

Proof. Assume the contrary. For any $t > 0$, we have: $n_t = (t^2, t^2 I_m, t^{-1} I_n, t^2 I_k) \in \mathcal{N}$ and $n_t \cdot X = (t^2, t^2 I_m, t^{-1} I_n, t^2 I_k) \cdot (c, A, b, C, d, l, u) = (tc, tA, t^2 b, tC, t^2 d, tl, tu) \in \mathcal{N} \cdot X$. So, $C(tc, tA, t^2 b, tC, t^2 d, tl, tu) = C(c, A, b, C, d, l, u)$ for all $t > 0$. Hence, we have:

$$C(c, A, b, C, d, l, u) = \lim_{t \to 0^+} C(tc, tA, t^2 b, tC, t^2 d, tl, tu) = C(0, 0, 0, 0, 0, 0, 0) = (0, 0, 0, 0, 0, 0, 0) \tag{4}$$

The second equality holds since $C(X)$ is continuous at $(0, 0, 0, 0, 0, 0, 0)$, and the last equality holds since the orbit of $(0, 0, 0, 0, 0, 0, 0)$ only contains $(0, 0, 0, 0, 0, 0, 0)$. So, it suffices to take X which has a different support type from all zeros for a contradiction. □

E.2 LOCAL CONTINUITY OBSTRUCTION FOR CANONICALIZATIONS

Beyond global continuity, the following theorems show that it's even impossible to get local continuity outside a particular area. The proof also gives a sufficient condition for non-trivial permutation–rescaling collision to happen in Theorem D.1 while leading to the infinite frames as well. Below, we still write the logarithmic rescaling parameters as $z = (u, r_1, \ldots, r_m, \eta_1, \ldots, \eta_n, \zeta_1, \ldots, \zeta_k) \in \mathbb{R}^{1+m+n+k}$, where $u = \log \alpha$, $r_i = \log(R_{\leq})_{ii}$, $\zeta_i = \log(R_{=})_{ii}$, and $\eta_j = \log B_{jj}$. Let $e_\alpha, e_{r_1}, \ldots, e_{r_m}, e_{\eta_1}, \ldots, e_{\eta_n}, e_{\zeta_1}, \ldots, e_{\zeta_k}$ denote the corresponding standard basis vectors of $\mathbb{R}^{1+m+n+k}$. For $X = (c, A, b, C, d, l, u)$, define $J_c(X) := \{j \in [n] : c_j \neq 0\}$, $E_A(X) := \{(i, j) \in [m] \times [n] : A_{ij} \neq 0\}$, $I_b(X) := \{i \in [m] : b_i \neq 0\}$, $E_C(X) := \{(i, j) \in [k] \times [n] : C_{ij} \neq 0\}$, $I_d(X) := \{i \in [k] : d_i \neq 0\}$, and $I_{l,u}(X) := \{j \in [n] : l_j \notin \{0, -\infty\}$ or $u_j \notin \{0, +\infty\}\}$. Define the support-weight set $\mathcal{S}_B(X)$ as follows:

- for each $j \in J_c(X)$, include $e_\alpha + e_{\eta_j}$;
- for each $(i, j) \in E_A(X)$, include $e_{r_i} + e_{\eta_j}$;
- for each $(i, j) \in E_C(X)$, include $e_{\zeta_i} + e_{\eta_j}$;
- for each $i \in I_b(X)$, include e_{r_i};
- for each $i \in I_d(X)$, include e_{ζ_i};
- for each $j \in I_{l,u}(X)$, include $-e_{\eta_j}$.

Let M_X^B be the matrix whose rows are exactly the vectors in $\mathcal{S}_B(X)$. We also associate with X a graph $G_X^B = (V_X^B, E_X^B)$. Its vertex set is $V_X^B := \{\alpha\} \sqcup \{r_1, \ldots, r_m\} \sqcup \{\zeta_1, \ldots, \zeta_k\} \sqcup \{\eta_1, \ldots, \eta_n\}$. Its edge set is defined as follows: add an edge (α, η_j) whenever $c_j \neq 0$; add an edge (r_i, η_j) whenever $A_{ij} \neq 0$; and add an edge (ζ_i, η_j) whenever $C_{ij} \neq 0$.

A vertex in V_X^B is called *anchored* if its corresponding unary weight is present in $\mathcal{S}_B(X)$. Specifically, a row vertex r_i is anchored if $b_i \neq 0$, a row vertex ζ_i is anchored if $d_i \neq 0$, and a column vertex η_j is anchored if $j \in I_{l,u}(X)$. Define $U_{\max}^B$ to be the set of all $X \in \mathbb{P}_{m,n,k}$ such that every connected component of G_X^B contains at least one anchored vertex. Equivalently, $X \in U_{\max}^B$ if and only if every vertex of G_X^B is reachable from the set of anchored vertices.

Lemma E.2. *For every $X \in \mathcal{P}_{m,n,k}$, the following statements are equivalent:*

1. $X \in U^B_{\max}$.
2. $\operatorname{span} \mathcal{S}_B(X) = \mathbb{R}^{1+m+n+k}$.
3. $\operatorname{rank} M^B_X = 1 + m + n + k$.

Moreover, if $q_0(X)$ denotes the number of connected components of G^B_X that contain no anchored vertex, then $\operatorname{rank} M^B_X = 1 + m + n + k - q_0(X)$.

Proof. Let $L_X := \operatorname{span} \mathcal{S}_B(X)$. Assume first that $X \in U^B_{\max}$. Take any connected component Γ of G^B_X. By definition, Γ contains some anchored vertex v (which could be r_i, ζ_i, or η_j), so its corresponding basis vector e_v (or $-e_{\eta_j}$, which spans the same space as e_{η_j}) is in $\mathcal{S}_B(X) \subseteq L_X$. For every edge between a left vertex $p \in \{\alpha, r_1, \dots, r_m, \zeta_1, \dots, \zeta_k\}$ and a right vertex $q \in \{\eta_1, \dots, \eta_n\}$, the corresponding support weight is $e_p + e_q \in \mathcal{S}_B(X)$. Hence, if $e_p \in L_X$, then $e_q = (e_p + e_q) - e_p \in L_X$. Similarly, if $e_q \in L_X$, then $e_p = (e_p + e_q) - e_q \in L_X$. Since every connected component contains an anchored vertex, this propagation along paths shows that every standard basis vector of $\mathbb{R}^{1+m+n+k}$ lies in L_X. Therefore $L_X = \mathbb{R}^{1+m+n+k}$.

Conversely, suppose that some connected component Γ of G^B_X contains no anchored vertex. Define $v_\Gamma \in \mathbb{R}^{1+m+n+k}$ by assigning value 1 to every left vertex in $\Gamma \cap \{\alpha, r_1, \dots, r_m, \zeta_1, \dots, \zeta_k\}$, value -1 to every right vertex in $\Gamma \cap \{\eta_1, \dots, \eta_n\}$, and value 0 to every vertex outside Γ. Then $v_\Gamma \neq 0$. For every possible edge weight $e_p + e_q$ inside Γ, where p is a left vertex and q is a right vertex, we have $\langle v_\Gamma, e_p + e_q \rangle = 1 - 1 = 0$. Every edge weight outside Γ is also orthogonal to v_Γ. Finally, since Γ contains no anchored vertex, every anchor weight (whether e_{r_i}, e_{ζ_i}, or $-e_{\eta_j}$) is orthogonal to v_Γ. Thus $\langle v_\Gamma, s \rangle = 0$ for every $s \in \mathcal{S}_B(X)$. Since $v_\Gamma \neq 0$, we have $L_X \neq \mathbb{R}^{1+m+n+k}$.

The same argument also gives the rank formula. For any $n \in \Gamma$, where Γ is an unanchored connected component. Adding e_n to the support vector set S_Γ of Γ would cause dim span $S_\Gamma \cup \{e_n\} = |\Gamma|$ while the vectors in spanS_Γ can only have the corresponding vertex's space to be non-zero. If dim span $S_\Gamma = |\Gamma|$, then it should cover all the basis vectors of the vertices in Γ. However, as indicated above, there exists a non-zero vector v_Γ with all the non-zero elements lying on the corresponding space of the vertices in Γ. Hence, dim span $S_\Gamma \leq |\Gamma| - 1$. Combining together, we get: dim span $S_\Gamma = |\Gamma| - 1$.

Then, it suffices to note that dim span $\mathcal{S}_B(X) = \sum_{\text{connected component } \Gamma} \text{dim span} S_\Gamma = \sum_{\text{anchored connected component } \Gamma} |\Gamma| + \sum_{\text{unanchored connected component } \Gamma} (|\Gamma| - 1) = 1 + m + n + k - q_0(X)$. □

Remark E.3. *If $X \notin U^B_{\max}$, then* $\ker M^B_X \neq \{0\}$. *Any nonzero $z \in \ker M^B_X$ defines a non-identity rescaling $g = \phi(z) \in \mathcal{N}$ satisfying $\ell_{g \cdot X} = \ell_X + M^B_X z = \ell_X$, where ℓ_X denotes the componentwise logarithm of the nonzero, non-infinite elements of X with order aligning with M^B_X. Hence $g \cdot X = X$. If X also admits a non-identity permutation symmetry $k \in \mathcal{K}$, then*

$$k \cdot X = X = g \cdot X,$$

providing a sufficient condition for non-trivial permutation–rescaling collision to happen in Theorem D.1. The results below connect these stabilizers to obstructions to local continuous canonicalization and to infinite frames.

Below, we show that for some LP instance, the frame size is infinite, which prevent frame-averaging from being efficient.

Corollary E.4. *Let $X \in \mathbb{P}_{m,n,k}$ satisfy* $\operatorname{rank} M^B_X < 1 + m + n + k$, *or equivalently, $X \notin U^B_{\max}$. Define its rescaling automorphism group by* $\operatorname{Aut}_{\mathcal{N}}(X) := \{g \in \mathcal{N} : g \cdot X = X\}$. *Then* $\operatorname{Aut}_{\mathcal{N}}(X) = \phi\big(\ker M^B_X\big) \cong \big(\mathbb{R}^{q_0(X)}, +\big)$, *which implies* $|\operatorname{Aut}_{\mathcal{N}}(X)| = |\mathbb{R}|$. *Moreover, if $\mathcal{F} : \mathbb{P}_{m,n,k} \to 2^{\mathcal{N}} \setminus \{\varnothing\}$ is any $\mathcal{N}$-equivariant frame, i.e., $\mathcal{F}(g \cdot Y) = g\mathcal{F}(Y)$ for every $g \in \mathcal{N}$ and $Y \in \mathbb{P}_{m,n,k}$, then*

$$|\mathcal{F}(X)| = |\mathbb{R}|.$$

Hence, no finite or countable equivariant frame exists at X.

Proof. By the preceding remark, the logarithmic parametrization $\phi : (\mathbb{R}^{1+m+n+k}, +) \longrightarrow \mathcal{N}$ is a group isomorphism and maps $\ker M_X^B$ bijectively onto $\mathrm{Aut}_{\mathcal{N}}(X)$. The rank formula above gives

$$\dim \ker M_X^B = 1 + m + n + k - \operatorname{rank} M_X^B = q_0(X) \geq 1.$$

This proves $\mathrm{Aut}_{\mathcal{N}}(X) \cong (\mathbb{R}^{q_0(X)}, +)$. Now choose any $z \in \mathcal{F}(X)$. For every $h \in \mathrm{Aut}_{\mathcal{N}}(X)$, frame equivariance yields

$$\mathcal{F}(X) = \mathcal{F}(h \cdot X) = h\mathcal{F}(X).$$

Hence $hz \in \mathcal{F}(X)$. The map $h \mapsto hz$ is injective by group cancellation, and therefore

$$|\mathcal{F}(X)| \geq |\mathrm{Aut}_{\mathcal{N}}(X)| = |\mathbb{R}|.$$

Since $\mathcal{F}(X) \subseteq \mathcal{N}$, the reverse inequality also holds, completing the proof. □

Next, we apply Lemma E.2 to show that it's impossible to design continuous canonicalizations outside $U_{\max}^B$.

Theorem E.5. *Let $X_0 \in \mathbb{P}_{m,n,k} \setminus U_{\max}^B$. Then there is no open neighborhood O of X_0 and no single-valued map $C : O \to \mathbb{P}_{m,n,k}$ satisfying the following three conditions simultaneously:*

1. *C is continuous at X_0.*
2. *$C(Y) \in N \cdot Y$ for every $Y \in O$.*
3. *$C(g \cdot Y) = C(Y)$ for every $g \in N$ and every $Y \in O$ such that $g \cdot Y \in O$.*

Proof. Assume for contradiction that such O and C exist. Since $X_0 \notin U_{\max}^B$, the previous lemma gives $\operatorname{span} \mathcal{S}_B(X_0) \neq \mathbb{R}^{1+m+n+k}$. Let $L := \operatorname{span} \mathcal{S}_B(X_0)$. Let $\mathcal{W}_B$ be the support weight set of an instance in $U_{\max}^B$ where all variable bounds are $[0, +\infty)$, consisting of all vectors $e_\alpha + e_{\eta_j}$ with $j \in [n]$, all vectors $e_{r_i} + e_{\eta_j}$ with $i \in [m]$ and $j \in [n]$, all vectors $e_{\zeta_i} + e_{\eta_j}$ with $i \in [k]$ and $j \in [n]$, all vectors e_{r_i} with $i \in [m]$, all vectors e_{ζ_i} with $i \in [k]$ (without any $-e_{\eta_j}$ with $j \in [n]$). The set $\mathcal{W}_B$ spans $\mathbb{R}^{1+m+n+k}$. Since L is a proper linear subspace, there exists $w \in \mathcal{W}_B \setminus L$. Let Π_L be the Euclidean orthogonal projection onto L, and define $v := \Pi_L w - w$. Then $v \neq 0$, $v \perp L$, and $\langle v, w \rangle = -\|w - \Pi_L w\|_2^2 < 0$. Since $w \notin L$, in particular $w \notin \mathcal{S}_B(X_0)$, the coefficient corresponding to w of X_0 is zero. More explicitly, if $w = e_\alpha + e_{\eta_j}$, then $(c_0)_j = 0$; if $w = e_{r_i} + e_{\eta_j}$, then $(A_0)_{ij} = 0$; and if $w = e_{r_i}$, then $(b_0)_i = 0$. Similar results hold for other situations as well. (The chosen w isn't the η_j by construction. Then w much corresponds to 0 in X_0 instead of $+\infty$.)

Since O is open and contains X_0, choose $\delta \neq 0$ sufficiently small and define $Y \in O$ by changing only this one zero coefficient of X_0 to δ, while leaving every other coefficient unchanged.Let $\phi(z) \in N$ denote the rescaling element with $\alpha = e^u$, $(R_{\leq})_{ii} = e^{r_i}$, $(R_{=})_{ii} = e^{\zeta_i}$, and $B_{jj} = e^{\eta_j}$ for $z = (u, r_1, \ldots, r_m, \eta_1, \ldots, \eta_n, \zeta_1, \cdots, \zeta_k)$. For $t \geq 0$, define $Y_t := \phi(tv) \cdot Y$. We claim that $Y_t \to X_0$ as $t \to +\infty$. Every nonzero coefficient of X_0 has a support weight $s \in \mathcal{S}_B(X_0)$. Since $v \perp L$ and $s \in L$, we have $\langle v, s \rangle = 0$. Hence all originally nonzero coefficients of X_0 remain unchanged along the path Y_t. The perturbed coefficient has weight w, so its value along the path is $e^{t\langle v, w \rangle}\delta$. Since $\langle v, w \rangle < 0$, this coefficient converges to 0 as $t \to +\infty$. All other coefficients that are zero in X_0 and were not perturbed remain zero. Therefore $Y_t \to X_0$.

Moreover, by choosing $|\delta|$ sufficiently small, we may ensure $Y_t \in O$ for every $t \geq 0$. Indeed, along the path Y_t, the only coordinate that differs from X_0 has magnitude at most $|\delta|$. Since $Y_t = \phi(tv) \cdot Y$, the local rescaling invariance of C gives $C(Y_t) = C(Y)$ for every $t \geq 0$. By continuity of C at X_0, we get $C(X_0) = \lim_{t \to +\infty} C(Y_t) = C(Y)$. However, the action of $\mathcal{N}$ multiplies every coefficient of (c, A, b, C, d, l, u) by a strictly positive scalar. Therefore it preserves the exact zero–nonzero support pattern. The instances X_0 and Y have different support patterns, because Y has exactly one additional nonzero coefficient. Hence $N \cdot X_0 \cap N \cdot Y = \varnothing$. On the other hand, C is an orbit canonicalization on O, so $C(X_0) \in N \cdot X_0$ and $C(Y) \in N \cdot Y$. Together with $C(X_0) = C(Y)$, this implies $N \cdot X_0 \cap N \cdot Y \neq \varnothing$, a contradiction. Therefore no such O and C can exist. □

The obstruction above only uses rescaling invariance and the orbit condition. Therefore, it also applies to permutation-equivariant rescaling orbit canonicalizations.

E.3 Borel Measurability of Canonicalizations and Related Functions

Despite these continuity obstructions, our energy construction yields a Borel-measurable canonicalization and transporter. This parallels LP solution selection: on the domain where optimal solutions exist, Borel-measurable selectors are available, whereas continuous selectors (even locally) may not exist. (Chen et al., 2023a) Similar to Section 6, we stack the inequality and equality constraint matrices and their right-hand sides and continue to denote them by $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, respectively. Let $\mathcal{I}$ index the total coefficient slots. Thus, in the standard-form notation $X = (c, A, b, l, u)$ used above, $|\mathcal{I}| = d_0 := 3n + mn + m$. We define the coefficient-type pattern $\tau(X) = (\tau_p(X))_{p\in\mathcal{I}}$ by

$$\tau_p(X) = \begin{cases} 0, & x_p(X) = 0, \\ +, & 0 < x_p(X) < +\infty, \\ -, & -\infty < x_p(X) < 0, \\ +\infty, & x_p(X) = +\infty, \\ -\infty, & x_p(X) = -\infty. \end{cases}$$

Here the last two cases occur only for bounds. Let $\Omega_\tau := \{X : \tau(X) = \tau\}, S_\tau := \{p : \tau_p \in \{+,-\}\}$. There are finitely many such strata, each Ω_τ is Borel, and the nonempty strata form a partition of the LP instance space. On Ω_τ, define the active log-magnitude chart $\ell_\tau(X) := \big(\log|x_p(X)|\big)_{p\in S_\tau} \in \mathbb{R}^{|S_\tau|}$. The map ℓ_τ is continuous on the domain. Its inverse coefficient reconstruction χ_τ keeps the zero and infinite coordinates fixed and, for $p \in S_\tau$, sets $x_p\big(\chi_\tau(y)\big) = s_{\tau,p}e^{y_p}, s_{\tau,p} := \operatorname{sgn}(x_p(X)) \in \{-1, 1\}$. The value $s_{\tau,p}$ is independent of the choice of $X \in \Omega_\tau$. Thus χ_τ is continuous on its Euclidean chart. Let H be one of the positive-rescaling subgroups considered in Section 6, i.e.,

$$H = \{\varphi_H(z) : z \in \mathbb{R}^{q_H}\}, \qquad \varphi_H(z + z') = \varphi_H(z)\varphi_H(z'), \qquad \varphi_H(z)^{-1} = \varphi_H(-z).$$

Here, φ_H is defined to be the isomorphism between $\mathbb{R}^{q_H}$ and H where q_H denotes the dimension of the linear space $\mathrm{Log}(H)$.Positive rescaling preserves every coefficient-type pattern. Moreover, on each Ω_τ there is a constant matrix $M_{\tau,H} \in \mathbb{R}^{|S_\tau|\times q_H}$ such that $\ell_\tau\big(\varphi_H(z)\cdot X\big) = \ell_\tau(X) + M_{\tau,H}z$ due to the log-linearity of H. The matrix is constant on the stratum because it only depends on the support type and the sign of the coefficients, which is fixed in a stratum. Let K be the row-and-variable permutation group. We assume, as in section 6, that K normalizes H. A feature map Θ and a scalar energy E are called compatible with K if

$$\Theta(k\cdot X) = \Pi_k\Theta(X), \qquad E(\Pi_k v) = E(v), \tag{5}$$

where Π_k is the induced permutation of the feature coordinates. On a fixed stratum, write the resulting log-coordinate energy as

$$\Psi_{\tau,H}(y) := (E\circ\Theta)\big(\chi_\tau(y)\big).$$

Then we have the following result.

Theorem E.6. *Assume that K normalizes H and that Θ and E satisfy equation 5. Suppose further that, for every nonempty coefficient-type stratum Ω_τ, $\Psi_{\tau,H}$ is C^2 and there exists $\sigma_{\tau,H} > 0$ such that*

$$v^\top\nabla^2\Psi_{\tau,H}(y)v \geq \sigma_{\tau,H}\|v\|_2^2 \tag{6}$$

for every $y \in \mathbb{R}^{|S_\tau|}$ and every $v \in \mathrm{Im}(M_{\tau,H})$. For $X \in \Omega_\tau$, define

$$y^*_{\tau,H}(X) := \operatorname*{arg\,min}_{y\in\ell_\tau(X)+\mathrm{Im}(M_{\tau,H})} \Psi_{\tau,H}(y), \tag{7}$$

$$z^*_{\tau,H}(X) := M^\dagger_{\tau,H}\big(y^*_{\tau,H}(X) - \ell_\tau(X)\big), \tag{8}$$

where $M^\dagger$ denotes the Moore–Penrose pseudoinverse. Set

$$C_H(X) := \chi_\tau\big(y^*_{\tau,H}(X)\big), \qquad g^H_X := \varphi_H\big(-z^*_{\tau,H}(X)\big). \tag{9}$$

Then the minimizer in equation 7 exists and is unique. The maps C_H and $X \mapsto g^H_X$ are C^1 on every stratum and are therefore globally Borel measurable. They satisfy

$$C_H(X) \in H\cdot X, \qquad X = g^H_X\cdot C_H(X), \qquad C_H(h\cdot X) = C_H(X), \qquad C_H(k\cdot X) = k\cdot C_H(X) \tag{10}$$

for every $h \in H$ and $k \in K$. Furthermore, $z^*_{\tau,H}(X)$ is the unique minimum-ℓ_2-norm logarithmic scale vector that maps X to $C_H(X)$. Hence g^H_X is single-valued even when the H-action has a nontrivial stabilizer.

Proof. Fix a stratum Ω_τ and abbreviate $M := M_{\tau,H}$, $N := \ker(M)$, and $r := \operatorname{rank}(M)$. If $r = 0$, then the feasible set in equation 7 is the singleton $\{\ell_\tau(X)\}$. In this case $y^*_{\tau,H}(X) = \ell_\tau(X)$, $z^*_{\tau,H}(X) = 0$, and every assertion follows immediately. We therefore assume $r > 0$. Choose a matrix $U \in \mathbb{R}^{q_H \times r}$ whose columns form an orthonormal basis of $N^\perp$, and put $D := MU$. We first prove that D has full column rank. If $Dw = 0$, then $Uw \in \ker(M) = N$. By construction, $Uw \in N^\perp$ as well, so $Uw = 0$. Since U has linearly independent columns, $w = 0$. Every $z \in \mathbb{R}^{q_H}$ has a unique orthogonal decomposition $z = Uw + n$ with $n \in N$. Since $Mn = 0$, we have $\ell_\tau(X) + Mz = \ell_\tau(X) + Dw$. Conversely, every $w \in \mathbb{R}^r$ gives a point in the feasible affine space. Therefore, minimizing over that affine space is equivalent to minimizing

$$F(a, w) := \Psi_{\tau,H}(a + Dw), \qquad a \in \mathbb{R}^{|S_\tau|}, \quad w \in \mathbb{R}^r.$$

The function F is C^2. Since D has full column rank, its smallest singular value $s_{\min}(D)$ is strictly positive. For every $\xi \in \mathbb{R}^r$, the chain rule and equation 6 give

$$\begin{aligned}\xi^\top \nabla^2_{ww} F(a, w)\xi &= (D\xi)^\top \nabla^2 \Psi_{\tau,H}(a + Dw)(D\xi) \\ &\geq \sigma_{\tau,H} \|D\xi\|_2^2 \\ &\geq \sigma_{\tau,H} s_{\min}(D)^2 \|\xi\|_2^2.\end{aligned}$$

Thus $F(a, \cdot)$ is strongly convex, uniformly in a. Taylor's theorem with Lagrange remainder, applied along the segment from 0 to w, yields

$$F(a, w) \geq F(a, 0) + \langle \nabla_w F(a, 0), w\rangle + \frac{\sigma_{\tau,H} s_{\min}(D)^2}{2} \|w\|_2^2.$$

The right-hand side tends to $+\infty$ as $\|w\|_2 \to \infty$; hence $F(a, \cdot)$ is coercive. Choose a radius R such that $F(a, w) > F(a, 0)$ whenever $\|w\|_2 > R$. By continuity, $F(a, \cdot)$ attains its minimum on the compact ball $\{w : \|w\|_2 \leq R\}$, and this is also its global minimum. Strong convexity makes the minimizer unique; denote it by $w^*(a)$. We next prove that w^* is C^1. Define

$$G(a, w) := \nabla_w F(a, w) = D^\top \nabla \Psi_{\tau,H}(a + Dw).$$

The first-order condition gives $G(a, w^*(a)) = 0$. Its Jacobian with respect to w is

$$D_w G(a, w) = D^\top \nabla^2 \Psi_{\tau,H}(a + Dw) D,$$

which is positive definite by the preceding inequality and hence invertible. The implicit function theorem therefore provides, around every a_0, a C^1 solution of $G(a, w) = 0$. Any such solution is the unique global minimizer of the strongly convex function $F(a, \cdot)$, so it equals $w^*(a)$. The local solutions consequently agree on overlaps, and w^* is C^1 on all of $\mathbb{R}^{|S_\tau|}$. Note that $y^*_{\tau,H}(X) = \ell_\tau(X) + Dw^*(\ell_\tau(X))$. Thus $y^*_{\tau,H}$ is C^1 on the stratum. The vector $Uw^*(\ell_\tau(X))$ solves

$$Mz = y^*_{\tau,H}(X) - \ell_\tau(X)$$

and belongs to $N^\perp$. Every other solution is $Uw^*(\ell_\tau(X)) + n$ for some $n \in N$. Orthogonality gives

$$\|Uw^*(\ell_\tau(X)) + n\|_2^2 = \|Uw^*(\ell_\tau(X))\|_2^2 + \|n\|_2^2.$$

Hence $Uw^*(\ell_\tau(X))$ is the unique minimum-norm solution. By the defining property of the Moore–Penrose pseudoinverse, it is exactly $z^*_{\tau,H}(X)$ in equation 8. Therefore $z^*_{\tau,H}$ is also C^1 on the stratum. The reconstruction χ_τ and the homomorphism φ_H are continuous. Equations equation 9 therefore show that C_H and g^H are continuous, indeed C^1 in active log coordinates, on Ω_τ. Because there are finitely many Borel strata, for every open set O,

$$C_H^{-1}(O) = \bigcup_\tau \big(\Omega_\tau \cap (C_H|_{\Omega_\tau})^{-1}(O)\big)$$

is Borel. The same argument applies to $X \mapsto g^H_X$. This proves the two global measurability statements. It remains to establish the algebraic properties. By equation 8, the two instances below have the same coefficient-type pattern and the same active log-magnitudes. Therefore $C_H(X) = \varphi_H(z^*_{\tau,H}(X)) \cdot X$. Since $\varphi_H(-z) = \varphi_H(z)^{-1}$, this gives $X = g^H_X \cdot C_H(X)$. For $h = \varphi_H(t)$, we have

$$\ell_\tau(h \cdot X) + \operatorname{Im}(M) = \ell_\tau(X) + Mt + \operatorname{Im}(M) = \ell_\tau(X) + \operatorname{Im}(M).$$

The two orbit-minimization problems therefore have the same feasible set and the same unique minimizer, proving $C_H(h \cdot X) = C_H(X)$.

Finally, normalization of H by K gives $H \cdot (k \cdot X) = k \cdot (H \cdot X)$. By equation 5, the composite energy has the same value at Y and at $k \cdot Y$. Hence the unique minimizer over $H \cdot (k \cdot X)$ is $k \cdot C_H(X)$, which proves $C_H(k \cdot X) = k \cdot C_H(X)$ and completes the proof. □

We now verify all hypotheses of Theorem E.6 for the two energies in Section 6. This verification also explains why the scale-linear terms in our canonicalization.

Theorem E.7. *Consider the two energies in section 6. Assume that $\lambda_A > 0$, $\mu_A > 0$, and the weight of every nonempty anchor block is strictly positive. Then the FullB and β canonicalization maps, denoted by C_{FullB} and C_β, are Borel measurable. Their minimum-ℓ_2-log-scale restoration maps $X \mapsto g_X^{\mathrm{FullB}}$ and $X \mapsto g_X^{\beta}$ are also Borel measurable and satisfy*

$$X = g_X^{\mathrm{FullB}} \cdot C_{\mathrm{FullB}}(X), \qquad X = g_X^{\beta} \cdot C_\beta(X).$$

In fact, all four maps are C^1 on each coefficient-type stratum. The two canonicalizers are invariant under their respective positive-rescaling subgroups and equivariant under row and variable permutations.

Proof. Fix a stratum Ω_τ. We still stack the equality constraint and the inequality constraint to form a matrix (A, b) with m rows. Let us recall the definitions: $\Omega := \{(i,j) : A_{ij} \neq 0\}, I_A := \{i : \exists j \text{ with } (i,j) \in \Omega\}, n_A := |\Omega|, m_A := |I_A|$. Let J_c, I_b, J_l, and J_u denote, respectively, the index sets of the finite nonzero entries of c, b, l, and u on this stratum. These index sets are constant on fixed Ω_τ. For FullB, we use the logarithmic parameter $z_{\mathrm{FullB}} = (a, r_1, \dots, r_m, \eta_1, \dots, \eta_n) \in \mathbb{R}^{m+n+1}$ and the continuous homomorphism $\varphi_{\mathrm{FullB}}(z_{\mathrm{FullB}}) = (e^a, \mathrm{Diag}(e^{r_1}, \dots, e^{r_m}), \mathrm{Diag}(e^{\eta_1}, \dots, e^{\eta_n}))$ to denote the final rescaling variables. For β scaling, we use $z_\beta = (a, r_1, \dots, r_m, v) \in \mathbb{R}^{m+2}, \varphi_\beta(z_\beta) = (e^a, \mathrm{Diag}(e^{r_1}, \dots, e^{r_m}), e^v I_n)$. Ordering the active (i.e., non-zero and non-infinite) log-magnitudes as $y = (y_A, y_c, y_b, y_l, y_u)$, the matrices $M_{\tau,H}$ are specified by

$$M_{\tau,\mathrm{FullB}} = \big((e_{r_i} + e_{\eta_j})_{(i,j)\in\Omega}, (e_a + e_{\eta_j})_{j\in J_c}, (e_{r_i})_{i\in I_b}, (-e_{\eta_j})_{j\in J_l}, (-e_{\eta_j})_{j\in J_u}\big), \tag{11}$$

$$M_{\tau,\beta} = \big((e_{r_i} + e_v)_{(i,j)\in\Omega}, (e_a + e_v)_{j\in J_c}, (e_{r_i})_{i\in I_b}, (-e_v)_{j\in J_l}, (-e_v)_{j\in J_u}\big). \tag{12}$$

Here, e_i denotes the standard basis vector at place i. These identities follow directly from $c'_j = e^{a+q_j} c_j, A'_{ij} = e^{r_i+q_j} A_{ij}, b'_i = e^{r_i} b_i, l'_j = e^{-q_j} l_j, u'_j = e^{-q_j} u_j$, where $q_j = \eta_j$ for FullB and $q_j = v$ for β scaling. Suppose first that $\Omega \neq \varnothing$. Let

$$D_{\tau,\mathrm{FullB}} := (e_{r_i} + e_{\eta_j})_{(i,j)\in\Omega}, \qquad D_{\tau,\beta} := (e_{r_i} + e_v)_{(i,j)\in\Omega}.$$

For FullB, let $G_k = (I_k, J_k, \Omega_k)$ be the connected components of the bipartite support graph of A, and put $m_k := |I_k|$ and $n_k := |J_k|$.We then define:

$$\delta_{\mathrm{FullB}} := -\sum_{i\in I_A} e_{r_i} - \sum_k \frac{m_k}{n_k} \sum_{j\in J_k} e_{\eta_j}, \quad \delta_\beta := -\sum_{i\in I_A} e_{r_i} - m_A e_v.$$

We claim that

$$\delta_H \perp \ker(D_{\tau,H}), \qquad H \in \{\mathrm{FullB}, \beta\}. \tag{13}$$

For FullB, if $D_{\tau,\mathrm{FullB}}(r, \eta) = 0$, then $r_i + \eta_j = 0$ on every active edge. The bipartite property and connectivity implies that, on each G_k, there is a scalar t_k such that $r_i = t_k, i \in I_k; \eta_j = -t_k, j \in J_k$. Indeed, equality propagates along every alternating path in the connected bipartite component. Consequently,

$$-\sum_{i\in I_k} r_i - \frac{m_k}{n_k} \sum_{j\in J_k} \eta_j = -m_k t_k + \frac{m_k}{n_k} n_k t_k = 0.$$

Summing over k proves the FullB part of equation 13. For β, the equation $D_{\tau,\beta}(r, v) = 0$ gives $r_i = -v$ for every $i \in I_A$, hence

$$-\sum_{i\in I_A} r_i - m_A v = m_A v - m_A v = 0.$$

This proves equation 13 in both cases. For every real matrix D, $\mathrm{Im}(D^\top) = (\ker D)^\perp$. Therefore equation 13 implies the existence of an edge vector $\omega_{\tau,H} \in \mathbb{R}^{n_A}$ satisfying $D_{\tau,H}^\top \omega_{\tau,H} = \delta_H$. We choose it canonically as $\omega_{\tau,H} := (D_{\tau,H}^\dagger)^\top \delta_H$. Let $y_A = \ell_{A,\tau}(X) + D_{\tau,H}\bar{z}_H$ denote the rescaling-transformed log edge coefficients, where $\bar{z}_{\mathrm{FullB}} = (r, \eta)$ and $\bar{z}_\beta = (r, v)$, then

$$\delta_H^\top \bar{z}_H = \omega_{\tau,H}^\top D_{\tau,H}\bar{z}_H = \omega_{\tau,H}^\top y_A - \omega_{\tau,H}^\top \ell_{A,\tau}(X). \tag{14}$$

The last term is independent of the candidate representative. Therefore the matrix objectives in both canonicalizations have exactly the same minimizing representative as the intrinsic log-feature energy

$$\overline{E}_{A,\tau}^H(y_A) := \frac{\lambda_A}{m_A}\left(\frac{1}{2}\sum_{e\in\Omega} e^{2y_{A,e}} + \omega_{\tau,H}^\top y_A\right) + \frac{\mu_A}{2n_A}\|y_A\|_2^2. \tag{15}$$

When $\Omega = \varnothing$, both the original matrix term and equation 15 are omitted, and $\omega_{\tau,H}$ is defined to be the empty vector. Adding the four feature-anchor terms gives

$$\Psi_{\tau,H}(y) := \overline{E}_{A,\tau}^H(y_A) + \frac{\lambda_c}{2|J_c|}\|y_c\|_2^2 + \frac{\lambda_b}{2|I_b|}\|y_b\|_2^2 + \frac{\lambda_l}{2|J_l|}\|y_l\|_2^2 + \frac{\lambda_u}{2|J_u|}\|y_u\|_2^2, \tag{16}$$

where, as stated in the theorem, every empty block is omitted. By equation 14, minimizing equation 16 over the orbit gives exactly the same representative y as minimizing the original energies. We now verify the equivariant–invariant factorization. Define the feature-extracting map by

$$\Theta_H(X) := \big(\tau(X), \ell_{\tau(X)}(X), \omega_{\tau(X),H}\big)$$

The first two terms are obviously permutation-equivariant. A row or variable permutation sends $D_{\tau,H}$ to $PD_{\tau,H}Q^\top$ and δ_H to $Q\delta_H$, where P and Q are permutation matrices on the edge and scale coordinates. The Moore–Penrose identity $(PDQ^\top)^\dagger = QD^\dagger P^\top$ directly gives

$$\omega_{k\tau,H} = \big((PD_{\tau,H}Q^\top)^\dagger\big)^\top Q\delta_H = P(D_{\tau,H}^\dagger)^\top \delta_H = P\omega_{\tau,H}.$$

Thus Θ_H is permutation equivariant. (Here, note that the permutation equivariance of Θ_θ with respect to $\mathcal{K}$ is defined by $\Theta_\theta(k \cdot Y) = \Pi_k \Theta_\theta(Y)$, where Π_k denotes the corresponding permutation P, Q on the output space induced by k here.) The exponential sums, squared Euclidean norms, and inner product $\omega_{\tau,H}^\top y_A$ in equation 16 are unchanged when both arguments are permuted simultaneously. All cardinalities of index sets are also unchanged. Hence the energy is permutation invariant on the hidden space.

It remains to verify the curvature condition. On a fixed stratum, all index sets, normalizing constants, and vectors $\omega_{\tau,H}$ are fixed. Therefore $\Psi_{\tau,H}$ is C^∞. Its Hessian is diagonal. For an edge coordinate $e \in \Omega$,

$$\frac{\partial^2 \Psi_{\tau,H}}{\partial y_{A,e}^2} = \frac{2\lambda_A}{m_A} e^{2y_{A,e}} + \frac{\mu_A}{n_A} \geq \frac{\mu_A}{n_A} > 0.$$

The diagonal Hessian entries on the nonempty and finite c, b, l, and u blocks are, respectively,

$$\frac{\lambda_c}{|J_c|}, \qquad \frac{\lambda_b}{|I_b|}, \qquad \frac{\lambda_l}{|J_l|}, \qquad \frac{\lambda_u}{|J_u|},$$

and are strictly positive by assumption. If $S_\tau = \varnothing$, then $M_{\tau,H}$ has rank zero and the conclusion follows from the rank-zero case of Theorem E.6. Otherwise, taking the minimum of the finitely many constants corresponding to the nonempty blocks yields $\sigma_{\tau,H} > 0$ such that

$$\nabla^2 \Psi_{\tau,H}(y) \succeq \sigma_{\tau,H} I_{|S_\tau|} \qquad \text{for every } y \in \mathbb{R}^{|S_\tau|}.$$

This is stronger than the orbitwise condition equation 6. All hypotheses of Theorem E.6 therefore hold for $H = \mathrm{FullB}$ and $H = \beta$. Applying that theorem gives the claimed Borel measurability, stratumwise C^1 regularity, symmetry properties, and restoration identities. Hence, the proof is finished. □

E.4 Rescaling Mask for Exact Model Equivariance

As mentioned in the main content and Lemma E.2, it would be possible for an LP instance to have non-trivial stabilizers. Arbitrarily choosing canonicalizers g_X^H and then restoring would break the Borel measurability and the final equivariance of the model. So, it remains an open question how to overcome this difficulty efficiently. Hence, we introduce a specified rescaling mask to cooperate with the transporter for ensuring final equivariance, as specified in the following theorems . And it's worth noting that this mask, along with any valid transporter, could lead to the final equivariance. Before introducing this design in detail, we review some notations first for self-containment.

For $H \in \{\mathrm{FullB}, \beta\}$, let $\Lambda_H : \mathbb{R}^{q_H} \longrightarrow \mathbb{R}^n$ denote the linear map extracting the logarithmic variable scales: $\Lambda_{\mathrm{FullB}}(a, r, \eta) = \eta, \Lambda_\beta(a, r, v) = v\mathbf{1}_n$. Thus, if $g = \varphi_H(z)$ has variable-scaling matrix B_g, then $B_g = \mathrm{Diag}\big(e^{\Lambda_H z}\big), \rho_H(g)x := B_g^{-1}x = \mathrm{Diag}\big(e^{-\Lambda_H z}\big)x$. For a coefficient-type stratum Ω_τ, define $\mathcal{N}_{\tau,H} := \ker(M_{\tau,H}), \mathcal{S}_{\tau,H} := \varphi_H(\mathcal{N}_{\tau,H})$. Since the active log-coordinates determine every instance in Ω_τ, one has

$$\mathcal{S}_{\tau,H} = \mathrm{Stab}_H(Y) \qquad \text{for every } Y \in \Omega_\tau.$$

Let $N_{\tau,H} \in \mathbb{R}^{q_H \times d_{\tau,H}}$ have orthonormal columns spanning $\mathcal{N}_{\tau,H}$, and set

$$R_{\tau,H} := \Lambda_H N_{\tau,H} \in \mathbb{R}^{n \times d_{\tau,H}}.$$

When $d_{\tau,H} = 0$, $R_{\tau,H}$ is understood as the empty matrix. Define

$$m_{\tau,H,j} := \mathbf{1}\big\{\|(R_{\tau,H})_{j,:}\|_2 = 0\big\}, \qquad j = 1, \dots, n, \tag{17}$$

and the diagonal projection (the rescaling mask):

$$\mathcal{P}_{\tau,H} := \mathrm{Diag}\big(m_{\tau,H,1}, \dots, m_{\tau,H,n}\big).$$

Now, we introduce some basic properties of the rescaling mask using the following lemmas.

Lemma E.8. *For every nonempty stratum Ω_τ,*

$$\mathrm{Im}(\mathcal{P}_{\tau,H}) = \mathrm{Fix}_{\rho_H}(\mathcal{S}_{\tau,H}) := \{x \in \mathbb{R}^n : \rho_H(s)x = x \textit{ for every } s \in \mathcal{S}_{\tau,H}\}.$$

Consequently,

$$\rho_H(s)\mathcal{P}_{\tau,H} = \mathcal{P}_{\tau,H} \qquad \textit{for every } s \in \mathcal{S}_{\tau,H}.$$

Moreover, $\mathcal{P}_{\tau,H}$ is the Euclidean orthogonal projection onto this fixed subspace.

Proof. Let $x \in \mathbb{R}^n$ and $s = \varphi_H(z)$ with $z \in \mathcal{N}_{\tau,H}$. Then

$$\big(\rho_H(s)x\big)_j = e^{-(\Lambda_H z)_j} x_j.$$

If $(R_{\tau,H})_{j,:} = 0$, then $(\Lambda_H z)_j = 0$ for every $z \in \mathcal{N}_{\tau,H}$, so the stabilizer acts trivially on the jth output coordinate. Conversely, if $(R_{\tau,H})_{j,:} \neq 0$, there exists $z \in \mathcal{N}_{\tau,H}$ such that $(\Lambda_H z)_j \neq 0$. The identity

$$e^{-(\Lambda_H z)_j} x_j = x_j$$

then forces $x_j = 0$. Hence, the stabilizer-fixed space consists exactly of the vectors supported on the coordinates for which $m_{\tau,H,j} = 1$. The diagonal map $\mathcal{P}_{\tau,H}$ is therefore its Euclidean orthogonal projection. □

Lemma E.9 (Permutation covariance of the rescaling mask)**.** *Let $k = (P_m, P_k, P_n) \in K$. Then*

$$\mathcal{P}_{k\tau,H} = P_n \mathcal{P}_{\tau,H} P_n^\top.$$

Thus, the stratumwise projection is permutation covariant. Here, $k\tau$ denotes the new support stratum using the same indexing order after permuting by k.

Proof. Let $Q_{k,H}$ denote the permutation matrix induced by k on the logarithmic rescale coordinates, and $P_{k,\tau}$ is the permutation matrix induced by k on the active support type using the same indexing order. For FullB and β canonicalization,

$$M_{k\tau,H} = P_{k,\tau} M_{\tau,H} Q_{k,H}^\top,$$

and therefore

$$\ker(M_{k\tau,H}) = Q_{k,H} \ker(M_{\tau,H}).$$

The variable-scale maps satisfy

$$\Lambda_H Q_{k,H} = P_n \Lambda_H.$$

Consequently, the rows of $\Lambda_H \ker(M_{\tau,H})$ are merely permuted by P_n. The zero-row mask in equation 17 therefore transforms as

$$m_{k\tau,H} = P_n m_{\tau,H},$$

which proves the result. □

Theorem E.10 (Exact equivariance after rescaling mask)**.** *Let $H \in \{\mathrm{FullB}, \beta\}$ be the corresponding rescaling group, and let $f : \mathbb{P}_{m,n,k} \to \mathbb{R}^n$ be a permutation-equivariant backbone:*

$$f(k \cdot Y) = P_n f(Y), \qquad k = (P_m, P_k, P_n) \in K.$$

For every X, let $g_X^H \in H$ be any valid transporter satisfying $X = g_X^H \cdot C_H(X)$. Define the projected canonicalization-based predictor by

$$F_H^{\mathrm{proj}}(X) := \rho_H(g_X^H)\, \mathcal{P}_{\tau(X),H}\, f(C_H(X)). \tag{18}$$

Then $F_H^{\mathrm{proj}}(X)$ is independent of the choice of the valid transporter g_X^H. And the composite mapping $\rho_H(g_X^H)\, \mathcal{P}_{\tau(X),H}$ is Borel measurable. Moreover, F_H^{proj} is exactly equivariant under both positive rescalings and permutations:

$$F_H^{\mathrm{proj}}(h \cdot X) = \rho_H(h) F_H^{\mathrm{proj}}(X), \qquad h \in H, \tag{19}$$

$$F_H^{\mathrm{proj}}(k \cdot X) = P_n F_H^{\mathrm{proj}}(X), \qquad k = (P_m, P_k, P_n) \in K. \tag{20}$$

If f is Borel measurable, then F_H^{proj} is also Borel measurable.

Proof. Fix $X \in \Omega_\tau$ and write

$$v_H(X) := \mathcal{P}_{\tau,H} f(C_H(X)).$$

By Lemma E.8,

$$\rho_H(s) v_H(X) = v_H(X) \qquad \text{for every } s \in \mathrm{Stab}_H(C_H(X)). \tag{21}$$

We first prove independence from the transporter. Let g_X^H and $\widetilde{g}_X^H$ be any two valid transporters. Then

$$g_X^H \cdot C_H(X) = \widetilde{g}_X^H \cdot C_H(X) = X.$$

Consequently,

$$s_X := (g_X^H)^{-1} \widetilde{g}_X^H \in \mathrm{Stab}_H(C_H(X)),$$

and $\widetilde{g}_X^H = g_X^H s_X$. Therefore, using equation 21,

$$\begin{aligned} \rho_H(\widetilde{g}_X^H) v_H(X) &= \rho_H(g_X^H s_X) v_H(X) \\ &= \rho_H(g_X^H) \rho_H(s_X) v_H(X) \\ &= \rho_H(g_X^H) v_H(X). \end{aligned}$$

Thus equation 18 has the same value for every valid transporter. Now let $h \in H$. Since the canonicalizer is H-invariant,

$$C_H(h \cdot X) = C_H(X),$$

and positive rescaling preserves the coefficient-type stratum. Hence

$$v_H(h \cdot X) = v_H(X).$$

Moreover, $h g_X^H$ is a valid transporter for $h \cdot X$, because

$$(h g_X^H) \cdot C_H(h \cdot X) = h \cdot \big(g_X^H \cdot C_H(X)\big) = h \cdot X.$$

By transporter independence, the transporter selected at $h \cdot X$ may be replaced by hg_X^H. Therefore,

$$\begin{aligned} F_H^{\mathrm{proj}}(h \cdot X) &= \rho_H(hg_X^H) v_H(X) \\ &= \rho_H(h)\rho_H(g_X^H) v_H(X) \\ &= \rho_H(h) F_H^{\mathrm{proj}}(X). \end{aligned}$$

This proves equation 19 without requiring the selected transporter map to be rescaling equivariant.

Next, let $k = (P_m, P_k, P_n) \in K$. By permutation equivariance of the canonicalizer, the backbone, and the rescaling mask,

$$\begin{aligned} C_H(k \cdot X) &= k \cdot C_H(X), \\ f(k \cdot C_H(X)) &= P_n f(C_H(X)), \end{aligned}$$

and

$$\mathcal{P}_{k\tau,H} = P_n \mathcal{P}_{\tau,H} P_n^\top.$$

It follows that

$$\begin{aligned} v_H(k \cdot X) &= \mathcal{P}_{k\tau,H} f(C_H(k \cdot X)) \\ &= P_n \mathcal{P}_{\tau,H} P_n^\top P_n f(C_H(X)) \\ &= P_n v_H(X). \end{aligned} \tag{22}$$

Since K normalizes H, the element $kg_X^H k^{-1}$ belongs to H. It is a valid transporter for $k \cdot X$, because

$$\begin{aligned} (kg_X^H k^{-1}) \cdot C_H(k \cdot X) &= (kg_X^H k^{-1}) \cdot (k \cdot C_H(X)) \\ &= k \cdot \big(g_X^H \cdot C_H(X)\big) \\ &= k \cdot X. \end{aligned}$$

Transporter independence therefore allows us to replace the selected transporter at $k \cdot X$ by $kg_X^H k^{-1}$. Using

$$\rho_H(kgk^{-1}) = P_n \rho_H(g) P_n^\top, \qquad g \in H,$$

and equation 22, we obtain

$$\begin{aligned} F_H^{\mathrm{proj}}(k \cdot X) &= \rho_H(kg_X^H k^{-1}) v_H(k \cdot X) \\ &= P_n \rho_H(g_X^H) P_n^\top P_n v_H(X) \\ &= P_n F_H^{\mathrm{proj}}(X). \end{aligned}$$

This proves equation 20 without requiring the selected transporter map to be permutation equivariant.

Finally, Theorem E.7 provides a Borel measurable valid transporter g_X^H for $C_H(X)$, namely the minimum-ℓ_2-norm transporter. Using this transporter in equation 18, the resulting map is Borel measurable because C_H and g_X^H are Borel measurable, $\mathcal{P}_{\tau,H}$ is constant on each of the finitely many Borel strata, f is Borel measurable, and ρ_H is continuous. Transporter independence shows that every other valid transporter selection defines exactly the same predictor. Hence F_H^{proj} is Borel measurable independently of the transporter used to evaluate it. Moreover, $\rho_H(g_X^H)\mathcal{P}_{\tau(X),H}$ is measurable since $\rho_H(g_X^H)\mathcal{P}_{\tau(X),H}$ is still C^1 for g_X^H in Theorem E.7 on each strata Ω_τ and H. Note the transporter independence finishes the proof. □

Implementation for FullB. Recall what we introduced to obtain the local continuity obstruction. For a fixed stratum, construct the scale-support graph $\mathcal{G}_{\tau,\mathrm{FullB}} = (V_\tau, E_\tau)$ with vertices

$$V_\tau = \{\alpha\} \sqcup \{r_1, \dots, r_{m+k}\} \sqcup \{\eta_1, \dots, \eta_n\}.$$

Add an edge (r_i, η_j) whenever $\bar{A}_{ij} \neq 0$, and add an edge (α, η_j) whenever $c_j \neq 0$. Mark r_i as anchored whenever $\bar{b}_i \neq 0$, and mark η_j as anchored whenever l_j or u_j is finite and nonzero. Here, $\bar{A}$ denotes the stacked constraint matrix of inequality and equality constraints.

Lemma E.2 implies that every unanchored connected component contributes one stabilizer direction which is non-zero on each vertex of this component, whereas the rescale parameter of a vector in

$\mathcal{N}_{\tau,H}$ on an anchored component must vanish. Then the rescaling mask can be computed efficiently without forming a null-space basis :

$$m_{\tau,\mathrm{FullB},j} = \mathbf{1}\left\{\text{the connected component containing } \eta_j \text{ contains an anchored vertex}\right\}. \tag{23}$$

The connected components and their anchor flags can be computed by depth-first search, breadth-first search, or union–find in $O\big(\mathrm{NNZ}(\bar{A}) + m + k + n\big)$ time and memory. Applying the resulting projection to a backbone output costs $O(n)$ time.

Implementation for β canonicalization. For β canonicalization, the corresponding variable-scale component of a null vector in $\mathcal{N}_{\tau,H}$ is $e^{-v}\mathbf{1}_n$. Hence the fixed output space is either $\mathbb{R}^n$ or $\{0\}$. Let

$$I_{\bar{A}} := \{i : \exists j \text{ such that } \bar{A}_{ij} \neq 0\}; J_l, J_u \text{ are the indexes of non-infinite,non-zero bounds}$$

, and $I_{\bar{b}}$ is the constraint that has a nonzero $\bar{b}_i$. Similar arguments show that every null vector has $v = 0$ if and only if

$$J_l \cup J_u \neq \varnothing \qquad \text{or} \qquad I_{\bar{A}} \cap I_{\bar{b}} \neq \varnothing.$$

Consequently,

$$\mathcal{P}_{\tau,\beta} = \begin{cases} I_n, & J_l \cup J_u \neq \varnothing \text{ or } I_{\bar{A}} \cap I_{\bar{b}} \neq \varnothing, \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

This condition can be checked in $O(\mathrm{NNZ}(\bar{A}) + m + n + k)$ time, and applying the projection costs $O(n)$ time.

Remark E.11. *For FullB canonicalization, the rescaling mask satisfies $P_X = I_n$ whenever every support-graph component containing a variable is anchored. An anchor is provided by a constraint with $b_i \neq 0$ or by a variable with a finite, nonzero bound. Thus, P_X is unnecessary, for example, when every variable is connected to a nonzero demand or capacity constraint, or has a finite, nonzero lower or upper bound. For scalar β-canonicalization, the condition is even weaker: one finite nonzero bound, or one nonempty constraint row with a nonzero right-hand side, already implies $P_X = I_n$.*

There are still some degenerate LPs that P_X isn't identity. A practical example arises in a fixed-size multi-commodity transportation model:

$$\begin{aligned} \min_{z,x\geq 0} \quad & c^\top z \\ \text{s.t.} \quad & Nz = d, d \neq 0, \\ & Nx = 0, \end{aligned}$$

where N is a network-incidence matrix. The variables z describe an active commodity with nonzero supplies and demands d, whereas x describes an inactive commodity whose demand and objective weight are both zero but whose variables are retained to preserve a fixed model structure. Assuming that the two commodities have separate balance constraints and no shared capacity constraints, the z-component is anchored by $d \neq 0$, while the x-component contains only homogeneous constraints and zero lower bounds. The latter therefore admits a nontrivial stabilizer. FullB removes the resulting ambiguity through

$$P_{X,\mathrm{FullB}} = \mathrm{Diag}\big(I_{|z|}, 0_{|x|}\big).$$

For scalar β-canonicalization, the nonzero vector d fixes the common variable scale, and hence $P_{X,\beta} = I$.

F The Universality of Canonicalization-based GNNs

In this section, we establish a PAC universality guarantee for canonicalization-based GNNs with rescaling mask. Fix $H \in \{\mathrm{FullB}, \beta\}$, and let $C_H : \mathbb{P}_{m,n,k} \longrightarrow \mathbb{P}_{m,n,k}$ be the Borel measurable, H-invariant, and K-equivariant canonicalization constructed in Theorem E.7. For each X, choose any valid transporter $g_X^H \in H$ satisfying $X = g_X^H \cdot C_H(X)$. Let $\mathcal{P}_{\tau,H}$ be the rescaling mask in Lemma E.8. We distinguish the group-valued transporter g_X^H from the masked output operator

$$\mathcal{R}_H(X) := \rho_H(g_X^H)\mathcal{P}_{\tau(X),H}. \tag{25}$$

In particular, $\mathcal{R}_H(X)$ acts on output vectors and is not, in general, a group element acting on LP instances. Let $\mathcal{F}_{\mathrm{LP}}^{\mathrm{GNN}}$ be a class of Borel measurable, permutation-equivariant LP-GNNs satisfying the PAC approximation property stated below. Define the masked canonicalization-based model class by

$$\mathcal{F}_{\mathrm{LP\text{-}Cano},H}^{\mathrm{GNN}} := \left\{ X \longmapsto \mathcal{R}_H(X) F_W(C_H(X)) \;\middle|\; F_W \in \mathcal{F}_{\mathrm{LP}}^{\mathrm{GNN}} \right\}. \tag{26}$$

The operator $\mathcal{R}_H(X)$ is independent of the chosen valid transporter as well. Indeed, let $\bar{g}_X^H$ be another Borel measurable reference transporter supplied by Theorem E.7. Then

$$s_X := (\bar{g}_X^H)^{-1} g_X^H \in \mathrm{Stab}_H(C_H(X)),$$

and hence

$$\rho_H(g_X^H)\mathcal{P}_{\tau(X),H} = \rho_H(\bar{g}_X^H)\rho_H(s_X)\mathcal{P}_{\tau(X),H} = \rho_H(\bar{g}_X^H)\mathcal{P}_{\tau(X),H}.$$

Consequently, $\mathcal{R}_H$ is Borel measurable, even if the initially chosen transporter assignment is not. (Since if we choose g_X^H to be the minimal ℓ_2 norm transporter, then the composite map is Borel measurable, as shown in Theorem E.10)

Write $\Phi(Y)$ for the optimal-solution set of an LP instance Y. Whenever $\Phi_{\mathrm{obj}}(Y) \in \mathbb{R}$, define

$$\Psi(Y) := \operatorname*{arg\,min}_{y \in \Phi(Y)} \|y\|_2.$$

This minimizer exists and is unique because $\Phi(Y)$ is a nonempty closed convex set. We use the usual Borel measurable minimum-norm solution map and, for definiteness, extend it by zero to instances without a finite optimal value. Uniqueness and invariance of the Euclidean norm under permutations give

$$\Psi(k \cdot Y) = P_n \Psi(Y), \qquad k = (P_m, P_k, P_n) \in K.$$

Lemma F.1 (Rescaling mask preserves optimality)**.** *Let $H \in \{\mathrm{FullB}, \beta\}$ and $Y \in \Omega_\tau$ satisfy $\Phi_{\mathrm{obj}}(Y) \in \mathbb{R}$. Then*

$$\mathcal{P}_{\tau,H}\Phi(Y) \subseteq \Phi(Y), \qquad \mathcal{P}_{\tau,H}\Psi(Y) = \Psi(Y).$$

Proof. We first show that, for these two rescaling groups, there exists a family $s_t \in \mathrm{Stab}_H(Y)$, $t \geq 0$, such that

$$\rho_H(s_t) = \mathcal{P}_{\tau,H} + e^{-t}(I - \mathcal{P}_{\tau,H}). \tag{27}$$

For $H = \mathrm{FullB}$, write the logarithmic scale coordinates as $z = (a, r, \eta)$, where a is the objective-scale coordinate, r collects both types of row-scale coordinates, and $\eta \in \mathbb{R}^n$ contains the variable-scale coordinates. The output action is $\rho_H(\varphi_H(z)) = \mathrm{diag}(e^{-\eta_1}, \ldots, e^{-\eta_n})$. Let $J := \{j : \mathcal{P}_{\tau,H} e_j = 0\}$. By the characterization of the stabilizer-fixed subspace,

$$j \notin J \quad \Longleftrightarrow \quad \eta_j = 0 \text{ for every } z \in \ker M_{\tau,H}.$$

Theorem E.2 implies that all variables appearing with nonzero coefficients in the same row are in the same connected component in G_X^B. Hence, they have the same status in $\mathcal{P}_{\tau,H}$. The same holds for all variables with nonzero objective coefficients. Thus, the support of each row, and the support of the objective, is contained either in J or in its complement. Moreover, a row with a nonzero right-hand side can involve only variables outside J, and a variable with a finite nonzero bound must lie outside J.

Define $z^0 = (a^0, r^0, \eta^0)$ by $\eta_j^0 = \mathbf{1}_{\{j \in J\}}$, setting $r_i^0 = -1$ when row i contains a nonzero coefficient of a variable in J, and $r_i^0 = 0$ otherwise. Similarly, set $a^0 = -1$ when the objective contains a nonzero coefficient of a variable in J, and $a^0 = 0$ otherwise. These coordinates satisfy every stabilizer equation:

$$a^0 + \eta_j^0 = 0 \quad \text{for nonzero objective coefficients,} \qquad r_i^0 + \eta_j^0 = 0 \quad \text{for nonzero row coefficients,}$$

together with the equations imposed by nonzero right-hand sides and finite nonzero bounds. Hence $z^0 \in \ker M_{\tau,H}$. Taking $s_t = \varphi_H(tz^0)$ proves equation 27.

For $H = \beta$, every output coordinate has the same scale. Therefore, $\mathcal{P}_{\tau,H}$ is either I or 0. If it is I, take s_t to be the identity. Otherwise, the stabilizer kernel contains a vector with a nonzero shared

variable-scale coordinate. Normalize that coordinate to one and exponentiate t times the resulting kernel vector. This gives $\rho_H(s_t) = e^{-t}I$, again proving equation 27.

Now let $y \in \Phi(Y)$. Equivariance of the optimal-solution set and $s_t \cdot Y = Y$ imply

$$\rho_H(s_t)y \in \Phi(Y) \qquad \text{for every } t \geq 0.$$

By equation 27,

$$\rho_H(s_t)y \longrightarrow \mathcal{P}_{\tau,H}y \qquad \text{as } t \longrightarrow \infty.$$

Closedness of $\Phi(Y)$ therefore gives $\mathcal{P}_{\tau,H}y \in \Phi(Y)$. In particular, $\mathcal{P}_{\tau,H}\Psi(Y)$ is optimal and $\|\mathcal{P}_{\tau,H}\Psi(Y)\|_2 \leq \|\Psi(Y)\|_2$. Uniqueness of the minimum-norm optimal solution yields $\mathcal{P}_{\tau,H}\Psi(Y) = \Psi(Y)$. □

We next recall some useful lemmas for the final proof.

Lemma F.2 ((Chen et al., 2023a)). *Let P be a Borel regular probability measure on $\mathbb{P}_{m,n,k}$ such that $P[\Phi_{\text{obj}}(X) \in \mathbb{R}] = 1$. Then, for every $\varepsilon, \delta > 0$, there exists $F_W \in \mathcal{F}_{\text{LP}}^{\text{GNN}}$ such that*

$$P[\|F_W(X) - \Psi(X)\|_2 > \delta] < \varepsilon.$$

Lemma F.3. *Let $\mathcal{P}$ be a Borel regular probability measure on $\mathbb{P}_{m,n,k}$ and $C : \mathbb{P}_{m,n,k} \to \mathbb{P}_{m,n,k}$ be a Borel measurable function. Then the pushforward measure $\mathcal{M}$ defined by $\mathcal{M}(A) = \mathcal{P}(C^{-1}(A))$ is still a Borel regular probability measure.*

Proof. We first show that $(\mathbb{P}_{m,n,k}, \mathcal{B}(\mathbb{P}_{m,n,k},), \mathcal{M})$ is indeed a Borel measurable space. To be more precise, consider

- $\mathcal{M}(\mathbb{P}_{m,n,k}) = \mathcal{P}(C^{-1}(\mathbb{P}_{m,n,k})) = \mathcal{P}(\mathbb{P}_{m,n,k}) = 1$.
- Since C is Borel measurable, for any $A \in \mathcal{B}(\mathbb{P}_{m,n,k})$, $C^{-1}(A) \in \mathcal{B}(\mathbb{P}_{m,n,k})$. Thus, we can well define the measure of any $A \in \mathcal{B}(\mathbb{P}_{m,n,k})$ to be $\mathcal{P}(C^{-1}(A))$ as $C^{-1}(A)$ is $\mathcal{P}$-measurable.
- If $(B_i)_{i\geq 1}$ are pairwise disjoint Borel sets, then the sets $C^{-1}(B_i)$ are pairwise disjoint Borel sets, hence we have:

$$\mathcal{M}(\cup_i B_i) = \mathcal{P}(C^{-1}(\cup_i B_i)) = \mathcal{P}(\cup_i C^{-1}(B_i)) = \sum_i \mathcal{P}(C^{-1}(B_i)) = \sum_i \mathcal{M}(B_i)$$

Combining all of the above, we can see that $\mathcal{M}$ is indeed a Borel probability measure. Next, we show the regularity. Since C is Borel measurable, we can find a compact set K such that $\mathcal{P}(\mathbb{P}_{m,n,k} \setminus K) \leq \frac{\epsilon}{2}$ such that $C|_K$ is continuous by the Lusin theorem. For any $A \in \mathcal{B}(\mathbb{P}_{m,n,k})$, we have:

$$\begin{aligned}\mathcal{M}(A) &= \mathcal{P}(C^{-1}(A)) = \mathcal{P}(C^{-1}(A) \cap K) + \mathcal{P}(C^{-1}(A) \setminus K) \\ &\leq \mathcal{P}(C^{-1}(A) \cap K) + \mathcal{P}(\mathbb{P}_{m,n,k} \setminus K) \leq \mathcal{P}(C^{-1}(A) \cap K) + \frac{\epsilon}{2}\end{aligned}$$

Since $C^{-1}(A), K$ are both Borel set, so as their intersection $C^{-1}(A) \cap K$. By the regularity of $\mathcal{P}$, we can see that $\exists$ compact $F \subseteq C^{-1}(A) \cap K$ such that: $\mathcal{P}(C^{-1}(A) \cap K \setminus F) \leq \frac{\epsilon}{2}$. Hence, we have:

$$\mathcal{M}(A) \leq \mathcal{P}(C^{-1}(A) \cap K) + \frac{\epsilon}{2} \leq \mathcal{P}(C^{-1}(A) \cap K \setminus F) + \mathcal{P}(F) + \frac{\epsilon}{2} \leq \mathcal{P}(F) + \epsilon$$

Moreover, since $F \subseteq C^{-1}(A) \cap K$, $C(F) \subseteq A$. Note that C is continuous on K, hence on F. So $C(F)$ is compact, hence Borel. So, $\mathcal{M}(A) \leq \mathcal{M}(C(F)) + \epsilon$, which implies that: $\mathcal{M}(A) \leq \sup\{\mathcal{M}(V) \mid V \subseteq A, V \text{ is compact}\} + \epsilon$. Since ϵ is arbitrary, we get $\mathcal{M}(A) \leq \sup\{\mathcal{M}(V) \mid V \subseteq A, V \text{ is compact}\}$ by letting $\epsilon \to 0$. By the monotonicity of measure, we have: $\mathcal{M}(A) \geq \sup\{\mathcal{M}(V) \mid V \subseteq A, V \text{ is compact}\}$. Combining both sides, we get: $\mathcal{M}(A) = \sup\{\mathcal{M}(V) \mid V \subseteq A, V \text{ is compact}\}$

For the other side, notice that A^c is still Borel as all Borel sets form a $\sigma-$algebra. Applying what we just proved before, we see that: for any $\epsilon > 0$, there exists a compact set $F' \subseteq A^c$ such that

$\mathcal{M}(A^c) \leq M(F') + \epsilon$, which implies that $1 - \mathcal{M}(A^c) \geq 1 - M(F') - \epsilon$. That is $\mathcal{M}(A) + \epsilon \geq M(F'^c) \geq \inf\{\mathcal{M}(O) \mid A \subset O, \text{O is open}\}$. The last inequality holds since $A \subseteq F'^c$. Similarly, by letting $\epsilon \to 0$, we can see that: $\mathcal{M}(A) \geq \inf\{\mathcal{M}(O) \mid A \subset O, \text{O is open}\}$. Combining with the monocity of measure would give $\mathcal{M}(A) = \inf\{\mathcal{M}(O) \mid A \subset O, \text{O is open}\}$.

Combining all above finish the lemma's proof. □

Theorem F.4 (PAC universality with rescaling mask)**.** *Under the assumptions above, let P be a Borel regular probability measure on $\mathbb{P}_{m,n,k}$ satisfying $P[\Phi_{\text{obj}}(X) \in \mathbb{R}] = 1$. Define the canonicalization-induced target by $T_H(X) := \mathcal{R}_H(X)\Psi(C_H(X))$. For every instance with a finite optimal value, $T_H(X) = \rho_H(g_X^H)\Psi(C_H(X)) \in \Phi(X)$. Then, for every $\varepsilon, \delta > 0$, there exists $f \in \mathcal{F}^{\text{GNN}}_{\text{LP-Cano},H}$ such that*

$$P[\|f(X) - T_H(X)\|_2 > \delta] < \varepsilon. \tag{28}$$

Proof. First, positive rescaling preserves the coefficient-type stratum and the existence of a finite optimal value. Thus, whenever X has a finite optimal value, Lemma F.1 gives

$$\mathcal{P}_{\tau(X),H}\Psi(C_H(X)) = \Psi(C_H(X)).$$

Together with $X = g_X^H \cdot C_H(X)$ and equivariance of the optimal-solution set, this implies

$$T_H(X) = \rho_H(g_X^H)\Psi(C_H(X)) \in \rho_H(g_X^H)\Phi(C_H(X)) = \Phi(X).$$

To prove the approximation statement, let $Q := (C_H)_{\#}P$. By Lemma F.3, Q is a Borel regular probability measure. Since $C_H(X)$ belongs to the same positive-rescaling orbit as X, $Q[\Phi_{\text{obj}}(Y) \in \mathbb{R}] = 1$. Hence Lemma F.2 applies to Q. Define $a_H(X) := \|\mathcal{R}_H(X)\|_2$. This function is Borel measurable and finite everywhere. Consequently,

$$P[a_H(X) > M] \longrightarrow 0 \qquad \text{as } M \longrightarrow \infty.$$

Choose $M \geq 1$ such that $P[a_H(X) > M] < \frac{\varepsilon}{2}$. Applying Lemma F.2 with accuracy δ/M and failure probability $\varepsilon/2$ yields $F_W \in \mathcal{F}^{\text{GNN}}_{\text{LP}}$ such that

$$Q\left[\|F_W(Y) - \Psi(Y)\|_2 > \frac{\delta}{M}\right] < \frac{\varepsilon}{2}.$$

Equivalently, $P\big[\|F_W(C_H(X)) - \Psi(C_H(X))\|_2 > \frac{\delta}{M}\big] < \frac{\varepsilon}{2}$. Set $f(X) := \mathcal{R}_H(X)F_W(C_H(X))$. On the event $a_H(X) \leq M$, $\|F_W(C_H(X)) - \Psi(C_H(X))\|_2 \leq \frac{\delta}{M}$, we have

$$\begin{aligned}\|f(X) - T_H(X)\|_2 &= \big\|\mathcal{R}_H(X)\big(F_W(C_H(X)) - \Psi(C_H(X))\big)\big\|_2 \\ &\leq a_H(X)\|F_W(C_H(X)) - \Psi(C_H(X))\|_2 \\ &\leq M\frac{\delta}{M} = \delta.\end{aligned}$$

Therefore, the union bound gives

$$\begin{aligned}P[\|f(X) - T_H(X)\|_2 > \delta] &\leq P[a_H(X) > M] \\ &\quad + P\left[\|F_W(C_H(X)) - \Psi(C_H(X))\|_2 > \frac{\delta}{M}\right] \\ &< \varepsilon.\end{aligned}$$

This proves equation 28. □

G Extended experimental details

G.1 Canonicalization and Prescaling Implementation

For all proposed canonicalizers, we set

$$\lambda_A = 1, \qquad \lambda_c = 0.1, \qquad \lambda_b = \lambda_l = \lambda_u = 0.5, \qquad \mu_A = 10^{-4}.$$

These values are fixed across all datasets and backbones, without dataset-specific tuning. We minimize the resulting sparse energy using analytic gradients and SciPy L-BFGS-B, initialized at the identity scaling. Its limited-memory quasi-Newton updates avoid forming or storing a dense Hessian, making it well-suited to LPs of different sizes and sparsity patterns. We use `maxiter=50`, `gtol=1e-8`, and `maxls=20`; the solver terminates before reaching the iteration limit whenever the gradient criterion is satisfied. *Note that canonicalization can be further accelerated by warm-starting from distribution-informed initial points or by precomputing and caching $M_{\tau,H}$, together with its factorization when applicable, for LPs with a fixed support pattern.* These procedures would not affect equivariance, since the rescaling mask could cooperate with any transporter g_X. For each LP instance, canonicalization is computed once, cached during training, and excluded from backpropagation. We use the same canonicalization procedure and solver configuration without cache during dataset preprocessing and at inference time. The projection P_X is computed without explicitly constructing the stabilizer nullspace. We build an augmented support graph containing one node for each constraint and variable, together with a shared objective-scaling node. An edge connects constraint i and variable j whenever $\bar{A}_{ij} \neq 0$, and connects the objective node to variable j whenever $c_j \neq 0$. A component is anchored if it contains a constraint with $\bar{b}_i \neq 0$ or a variable with a finite nonzero bound. (Same as the construction of G_X^B, $\bar{A}$ denote the stacked matrix of both inequality constraint and equality constraint.) We then set

$$(P_X)_{jj} = \begin{cases} 1, & \text{if variable } j \text{ belongs to an anchored component,} \\ 0, & \text{otherwise.} \end{cases}$$

This requires only a sparse connected-components computation.

For the HiGHS baseline, we use HiGHS 1.14.0, commit `7df0786`. A minimal read-only pybind wrapper included with our code calls the internal HiGHS `scaleLp` routine. Given a source LP, the wrapper returns a scaled copy, from which we read the row and column scaling factors. We apply the returned factors to the original graph rather than reconstructing the graph from the scaled LP. This preserves all source metadata and uses the same source-coordinate restoration convention as the proposed canonicalizers.

G.2 Dataset Background

We evaluate all backbones on LP relaxations from nine distributions: combinatorial auction (CA), set cover (SC), maximum independent set (IS), capacitated facility location (CFLP), minimum-cost flow (MCF), DC optimal power flow on the IEEE 57-, 118-, and 300-bus topologies (OPF57, OPF118, and OPF300), and the netlib instances (NETLIB). The binary restrictions in the four discrete-origin problem families—CA, SC, IS, and CFLP—are relaxed to the interval $[0, 1]$.

Set cover. Given a universe $\mathcal{U}$, candidate subsets $\{S_j\}_{j=1}^n$, and nonnegative costs c_j, set cover chooses a minimum-cost collection whose union covers every element:

$$\min_{x\in[0,1]^n} \sum_{j=1}^{n} c_j x_j \quad \text{s.t.} \quad \sum_{j:u\in S_j} x_j \geq 1, \qquad u \in \mathcal{U}. \tag{29}$$

Combinatorial auction. Give a universe $\mathcal{M}$ of items, each bid j requests a bundle S_j of items and is willing to offer value v_j for the bundle; the winner-determination LP accepts mutually compatible bids:

$$\max_{x\in[0,1]^n} \sum_{j=1}^{n} v_j x_j \quad \text{s.t.} \quad \sum_{j:k\in S_j} x_j \leq 1, \qquad k \in \mathcal{M}. \tag{30}$$

Maximum independent set. For an undirected graph $G = (V, E)$, IS selects a maximum set of pairwise nonadjacent vertices:

$$\max_{x\in[0,1]^{|V|}} \sum_{v\in V} x_v \quad \text{s.t.} \quad x_u + x_v \leq 1, \qquad (u, v) \in E. \tag{31}$$

Capacitated facility location. Let y_j denote whether facility j is open, x_{ij} the fraction of customer i assigned to facility j, d_i the demand, v_j the capacity, f_j the opening cost, and c_{ij} the transportation cost. The stored LP is

$$\begin{aligned}
\min_{x,y} \quad & \sum_{j\in\mathcal{F}} f_j y_j + \sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{F}} c_{ij}x_{ij} && && (32)\\
\text{s.t.} \quad & \sum_{j\in\mathcal{F}} x_{ij} \geq 1, && i\in\mathcal{C}, && (33)\\
& \sum_{i\in\mathcal{C}} d_i x_{ij} \leq v_j y_j, && j\in\mathcal{F}, && (34)\\
& \sum_{j\in\mathcal{F}} v_j y_j \geq \sum_{i\in\mathcal{C}} d_i, && && (35)\\
& x_{ij} \leq y_j, && i\in\mathcal{C},\ j\in\mathcal{F}, && (36)\\
& 0 \leq x_{ij} \leq 1, \quad 0 \leq y_j \leq 1. && && (37)
\end{aligned}$$

Because all transportation costs are nonnegative, the first inequality is tight at an optimum; it is stored as an inequality to keep the graph representation uniform.

Minimum-cost flow. For a directed graph with node–arc incidence matrix N, let f_a denote the amount of flow routed through arc a, c_a its per-unit transportation cost, and u_a its maximum admissible capacity. The vector s specifies the prescribed net supply or demand at each node, with the interpretation of its sign determined by the orientation convention used in N. The minimum-cost flow problem is

$$\min_{0\leq f\leq u} \ c^\top f \quad \text{s.t.} \quad Nf = s. \tag{38}$$

The objective is the total cost of transporting flow through the network. The box constraint $0 \leq f_a \leq u_a$ requires every arc flow to be nonnegative and prevents it from exceeding the physical or operational capacity of that arc. For each node i, the equality $(Nf)_i = s_i$ is a flow-conservation constraint: the net flow entering or leaving the node must match its prescribed supply or demand. In particular, a transshipment node has $s_i = 0$, so its total inflow equals its total outflow. Every column of an incidence matrix contains one positive and one negative entry, corresponding to the two endpoints of an arc. Consequently, $\mathbf{1}^\top N = 0$, and the flow-conservation equations are linearly dependent. Since a feasible instance necessarily satisfies $\mathbf{1}^\top s = 0$, one conservation row can be removed without changing the set of feasible flows; this only eliminates a redundant equality.

DC optimal power flow. We use the cycle-flow formulation of DC optimal power flow on a fixed transmission-network topology. Let p_g denote the active-power output of generator g, and let f_ℓ denote the active-power flow on transmission line ℓ. The matrix S maps generators to the buses at which they are located, while the oriented bus–branch incidence matrix N maps line flows to their corresponding nodal net flows. The vector d contains the active power demand at each bus. The resulting LP is

$$\begin{aligned}
\min_{p,f} \quad & c^\top p\\
\text{s.t.} \quad & Sp - Nf = d,\\
& K\,\mathrm{diag}(b)^{-1} f = -K\phi,\\
& p^{\min} \leq p \leq p^{\max}, \qquad -F \leq f \leq F.
\end{aligned}$$

The objective vector is the total generation cost, and the objective selects the least-cost generation dispatch that satisfies network and operating constraints. The nodal balance equation $Sp - Nf = d$ enforces conservation of active power at every bus. The component $(Sp)_i$ is the total generation located at bus i, while $(Nf)_i$ represents the net branch flow associated with that bus under the adopted line-orientation convention. Thus, the equation states that local generation together with net power imported from or exported to neighboring buses must exactly meet the demand d_i. This constraint is the power-system analogue of the flow-conservation equation in minimum-cost flow. The cycle equation $K\,\mathrm{diag}(b)^{-1} f = -K\phi$ enforces Kirchhoff's voltage law. Here, b_ℓ is the susceptance of

Table 6: ID dataset sizes used by every backbone.

Dataset	Instances	Variables	Constraints	Mean $\mathrm{nnz}(A)$	Train/Val/Test
CA	1000	600	557–598	3313.581	800/100/100
SC	1000	700	700	4900.000	800/100/100
IS	1000	1000	1996	3992.000	800/100/100
CFLP	1000	1025	1066	4050.000	800/100/100
MCF	1000	500	99	989.925	800/100/100
OPF57	1000	84	81	410.000	800/100/100
OPF118	1000	205	187	836.000	800/100/100
OPF300	1000	468	412	1733.000	800/100/100
NETLIB	1000	41–143	43–185	500.400	800/100/100

line ℓ, ϕ_ℓ is its fixed phase shift, and the rows of K describe a basis of independent cycles in the transmission network. Under the DC power flow approximation, f_ℓ/b_ℓ corresponds to the voltage-angle difference across line ℓ, adjusted by the fixed phase shift. The cycle constraint requires these implied angle differences to sum consistently around every independent network cycle. It therefore guarantees that the predicted line flows can be generated by a globally consistent set of bus voltage angles. On a connected fixed topology, this formulation is equivalent to the usual voltage-angle formulation of DC optimal power flow. The bounds correspond to the physical limit of the line or generator.

NETLIB. NETLIB is a standard collection of linear-programming test problems. We use five templates: ADLITTLE, SCAGR7, SC50A, KB2, and BOEING2. They have heterogeneous sparse constraint structures and can be written as

$$\min_{\ell^x \le x \le u^x} c^\top x \qquad \text{s.t.} \qquad \ell^r \le Ax \le u^r.$$

Equality rows satisfy $\ell_i^r = u_i^r$. During graph construction, equalities are retained, while each finite side of a one-sided or ranged row is represented as a deterministic inequality. The five templates define five distinct matrix supports and are pooled into a single NETLIB-derived distribution.

G.3 Instance Generation

Every retained instance is solved to optimality with HiGHS, and one optimal primal solution is stored as the supervised label. The data split seed is 2026 for every backbone. Each distribution contains 1000 instances and uses an $800/100/100$ train/validation/test split. The three model training seeds are 2027, 2028, and 2029. Table 6 reports the number of non-bound constraints; box bounds are stored separately as variable features, which is the most common choice in (Chen et al., 2023a; Qian et al., 2023; Liu et al., 2026). We define three OOD regimes, denoted OOD-1, OOD-2, and OOD-3, with progressively larger shifts in the data-generation parameters. The datasets at different levels are sampled independently and are not nested; therefore, a higher OOD index does not necessarily imply a monotonically larger test loss. Below, we briefly introduce how we generate the datasets.

SC generation. Each set-covering instance is

$$\min_{0 \le x \le 1} c^\top x \quad \text{s.t.} \quad Mx \ge \mathbf{1},$$

where $M \in \{0,1\}^{700\times 700}$ and $c_j \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[0,1)$. We allocate exactly $700^2 \times 0.01 = 4900$ binary nonzero positions: every column is assigned at least two rows, every row appears at least once, and row indices within each column are sampled without replacement.

CA generation. Each combinatorial-auction instance contains $n = 500$ real items and terminates after generating 600 bids. Each item i has a public baseline value $v_i \sim \mathrm{Unif}[0.01, 0.1]$, shared by all bidders. For every bidder k, we independently sample item interests $h_{ki} \sim \mathrm{Unif}[0,1]$ and define the bidder-specific value $v_{ki} = v_i + 0.05(2h_{ki} - 1)$. Thus, h_{ki} controls both the bidder's preference for

item i and the deviation of its private value from the public baseline. To model complementarities, we sample $\widetilde{Q}_{ij} \sim \mathrm{Unif}[0,1]$ for $i < j$, set $\widetilde{Q}_{ji} = \widetilde{Q}_{ij}$ and $\widetilde{Q}_{ii} = 0$, and normalize each row: $Q_{ij} = \frac{\widetilde{Q}_{ij}}{\sum_{q \neq i} \widetilde{Q}_{iq}}$. For a nonempty partial bundle S, the compatibility of a candidate item $j \notin S$ with the current bundle is $\kappa_j(S) = \frac{1}{|S|} \sum_{i \in S} Q_{ij}$. For bidder k, the first item $I_k^{(0)}$ is sampled according to

$$\Pr\Big(I_k^{(0)} = i\Big) = \frac{h_{ki}}{\sum_{q=1}^{n} h_{kq}}, \qquad S_k^{(0)} = \{I_k^{(0)}\}.$$

At each subsequent step t, an independent Bernoulli trial with success probability 0.7 is performed. Upon success, an unselected item is appended according to

$$\Pr\Big(J_k^{(t)} = j \mid S_k^{(t)}\Big) = \frac{\mathbf{1}\{j \notin S_k^{(t)}\} h_{kj} \kappa_j(S_k^{(t)})}{\sum_{q \notin S_k^{(t)}} h_{kq} \kappa_q(S_k^{(t)})},$$

and $S_k^{(t+1)} = S_k^{(t)} \cup \{J_k^{(t)}\}$. The process stops at the first failed trial, and its terminal set is denoted by S_k^{init}. Bidder k assigns every bundle S the price $p_k(S) = \sum_{i \in S} v_{ki} + |S|^{1.2}$. Alternative bundles are generated by taking each $a \in S_k^{\mathrm{init}}$ as an anchor, initializing a new bundle with $\{a\}$, and applying the same compatibility-based sampling rule until its cardinality equals $|S_k^{\mathrm{init}}|$. At most five additional distinct alternatives are retained, subject to

$$0 \leq p_k(S) \leq 1.5 p_k(S_k^{\mathrm{init}}) \quad \text{and} \quad \sum_{i \in S} v_i \geq 0.5 \sum_{i \in S_k^{\mathrm{init}}} v_i.$$

If bidder k contributes more than two bids, a bidder-specific dummy item is included in all of them, enforcing that at most one alternative from that bidder can be selected. Let $\mathcal{B}(i)$ be the set of bids containing real or dummy item i. The resulting LP relaxation is

$$\min_{0 \leq x \leq 1} -\sum_b p_b x_b \qquad \text{s.t.} \qquad \sum_{b \in \mathcal{B}(i)} x_b \leq 1 \quad \forall i.$$

Here, items receiving no bid induce no constraint row.

IS generation. Each unweighted independent-set instance is an independently sampled Barabasi–Albert graph with 1000 vertices and attachment parameter two. It has 1996 edges, one constraint $x_u + x_v \leq 1$ per edge, and objective $\min -\mathbf{1}^\top x$ over $0 \leq x \leq 1$.

CFLP generation. Each capacitated facility-location instance has 40 customers and 25 facilities, with locations $r_i, s_j \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([0,1]^2)$. We first sample $d_i^0 \sim \mathrm{Unif}\{5, \ldots, 35\}$, $q_j^0 \sim \mathrm{Unif}\{10, \ldots, 160\}$, and $f_j^0 = a_j \sqrt{q_j^0} + e_j$, where $a_j \sim \mathrm{Unif}\{100, \ldots, 111\}$ and $e_j \sim \mathrm{Unif}\{0, \ldots, 90\}$. The ID distribution then uses

$$d_i = \max\{1, \mathrm{round}\left(d_i^0\, 10^{\xi_i}\right)\}, \qquad \xi_i \sim \mathrm{Unif}[-1,1],$$
$$\widetilde{v}_j = q_j^0\, 10^{\rho_j}, \qquad \rho_j \sim \mathrm{Unif}[-1,1].$$

The capacities $\widetilde{v}_j$ are globally rescaled and integer-rounded so that $\sum_j \widetilde{v}_j \approx 2 \sum_i d_i$. Opening and assignment costs are

$$f_j = f_j^0\, 10^{\phi_j}, \quad c_{ij} = 10 d_i \|r_i - s_j\|_2\, 10^{\omega_{ij}}, \qquad \phi_j \sim \mathrm{Unif}[-2,2], \quad \omega_{ij} \sim \mathrm{Unif}[-1,1].$$

MCF generation. Each minimum-cost-flow instance independently samples a strongly connected directed graph with 100 nodes and 500 arcs by inserting a Hamiltonian cycle and then drawing 400 additional arcs without replacement. For each arc,

$$\bar{f}_a = 2^{g_f + h_a}, \qquad c_a = 2^{g_c + k_a}, \qquad g_f, g_c, h_a, k_a \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[-1,1]. \tag{39}$$

The feasible witness sets the node-balance vector s (i.e., $N\bar{f} = s$), and capacities satisfy $u_a = \bar{f}_a(1 + 0.15\, 2^{\epsilon_a})$ with $\epsilon_a \sim \mathrm{Unif}[-0.15, 0.15]$. One redundant conservation row is dropped.

OPF generation. We use the fixed IEEE 57-, 118-, and 300-bus network layouts and nominal equipment data from PGLib-OPF (Babaeinejadsarookolaee et al., 2019).

To add the randomness, for each instance, all the demand d is multiplied by a global factor $s = 2^g$, $g \sim \mathrm{Unif}[-0.6, 0.6]$, and redistributed across 4, 6, or 8 connected network regions for the three respective topologies. Regional and bus-level log-demand perturbations are Gaussian with zero mean in log base two. Their standard deviations are $(0.28, 0.10)$ for the three topologies. Generator and line parameters are perturbed around their nominal values by clipped lognormal device factors,

$$\log_2 s_g \sim \mathrm{clip}\big(\mathcal{N}(0, 0.30^2), [-0.85, 0.85]\big), \qquad \log_2 s_\ell \sim \mathrm{clip}\big(\mathcal{N}(0, 0.35^2), [-1, 1]\big), \tag{40}$$

together with regional and device-specific residual factors. The residual standard deviation is 0.18. Total generation capacity is calibrated so that its ratio to total demand is uniform on $[1.10, 1.38]$. Generator minima are perturbed nominal fractions of their maxima and clipped to $[0, 0.35]$.

Each line's transfer coefficient and safe-flow limit receive coherent regional and device factors, with an additional line-limit multiplier sampled from $[0.60, 1.00]$ for OPF57/118 and $[0.80, 1.35]$ for OPF300. Linear generation costs are perturbed nominal PGLib marginal costs: the global fuel multiplier has log-two exponent $\mathrm{Unif}[-0.5, 0.5]$, the device noise is Gaussian with standard deviation 0.38, and a weak inverse capacity exponent 0.78 makes larger generators moderately cheaper per unit. Infeasible or numerically invalid samples are rejected.

NETLIB generation. For each template, we generate 200 numerical instances while keeping its matrix support, coefficient signs, dimensions, and variable-bound pattern fixed. Every nonzero entry of A and c is independently multiplied by $\exp(\mathrm{Unif}[-0.5, 0.5])$. Starting from an optimal feasible solution of the source template, finite row bounds are reconstructed around its activity under the perturbed matrix, with the original row slacks independently multiplied by the same range of random factors. Equality bounds are set to the perturbed activity, and variable bounds remain unchanged. Every generated LP is re-solved with HiGHS, and infeasible, unbounded, or numerically invalid samples are rejected. The resulting 1,000 instances use a stratified $800/100/100$ train/validation/test split, with $160/20/20$ instances from each template.

OOD generation. Let $q \in \{1, 2, 3\}$ denote the OOD level, and define

$$\mathcal{I}_1 = [1.5, 2.0], \qquad \mathcal{I}_2 = [2.5, 3.0], \qquad \mathcal{I}_3 = [3.5, 4.0].$$

OPF. We replace the ID global-demand exponent $g \sim \mathrm{Unif}[-0.6, 0.6]$ by $g \sim \mathrm{Unif}(\mathcal{I}_q)$, so that the total demand is multiplied by 2^g. The regional and bus-level log-two demand perturbations have standard deviations 0.26 and 0.05, respectively. We also introduce generator injection efficiencies by replacing the bus-generator incidence matrix S with $S\,\mathrm{diag}(\eta)$. We sample

$$\log_2 \eta_j \sim \mathrm{clip}\big(\mathcal{N}(0, (w\bar{\sigma}_q)^2), [-3w\bar{\sigma}_q, 3w\bar{\sigma}_q]\big), \qquad (\bar{\sigma}_1, \bar{\sigma}_2, \bar{\sigma}_3) = (0.55, 0.75, 0.85),$$

where the topology weights for OPF57, OPF118, and OPF300 are $w = 0, 15/53$, and 1, respectively. This size-dependent weight avoids over-perturbing small systems with only a few generators, while allowing the efficiency dispersion to reach its full strength on larger systems. Finally, η is normalized so that $\sum_j p_j^{\max} \eta_j = \sum_j p_j^{\max}$.

MCF. Using the notation $\bar{f}_a = 2^{g_f + h_a}$ and $c_a = 2^{g_c + k_a}$, we replace $g_f \sim \mathrm{Unif}[-1, 1]$ by $g_f \sim \mathrm{Unif}(\mathcal{I}_q)$. The local exponent h_a contains three terms: an arc-specific draw from $\mathrm{Unif}[-0.25, 0.25]$, the average regional shock of the origin and destination nodes of arc a, and an independent $\mathcal{N}(0, 0.05^2)$ residual. Each regional shock follows $\mathrm{clip}(\mathcal{N}(0, 0.26^2), [-1, 1])$. For transportation costs, we set $g_c = 0$ and sample $k_a \sim \mathrm{Unif}[-0.20, 0.20]$. Each incidence column is also multiplied by an arc-efficiency factor η_a, with $\log_2 \eta_a$ sampled using standard deviation $\bar{\sigma}_q$ with $w = 1$ and clipped at three standard deviations. The efficiencies are normalized so that $n_A^{-1} \sum_a \eta_a = 1$, where n_A is the number of arcs.

NETLIB. We keep A unchanged and divide the rows and variables into four fixed sectors. Let z_r, $r = 1, \ldots, 4$, be a random permutation of $(-1, -1/3, 1/3, 1)$ multiplied by a draw from $\mathrm{Unif}[0.9, 1.1]$. At the three OOD levels, we use

$$(\mu_q, \delta_q, \gamma_q) = (0.75, 0.30, 0.20), \quad (1.50, 0.50, 0.40), \quad (2.25, 0.75, 0.60),$$

respectively. Sector r receives the log-two quantity factor $-\mu_q + \delta_q z_r$ and the log-two price factor $\gamma_q z_r$. The quantity factors modify equality right-hand sides, finite inequality resource limits, and finite nonzero variable bounds, whereas the price factors modify the nonzero objective coefficients.

G.4 Backbone Realizations

For the LP-GNN backbone, we use four message-passing layers with a hidden dimension of 64. All encoding, message, update, and readout MLPs contain two linear layers with ReLU activations. Layer normalization is applied after each message-passing layer.

For the implicit GNN backbone, we use a hidden dimension of 64 and assign one scalar dynamic state to each variable node. The implicit operator is applied six times with shared parameters, as in (Liu et al., 2026). Each application contains two message-passing rounds, giving twelve complete bipartite rounds in total. The message and update maps use two-layer MLPs with ReLU activations. Layer normalization is applied after every message-passing update.

G.5 Additional Experimental Results across Backbones

G.5.1 Data Augmentation and Rescaling Evaluation Protocol

Below, we denote $R = (R_{\leq}, R_{=})$ for notational convenience, and the logarithm is with basis 2.

Data augmentation. The extra samples are generated independently by transforming the original LP and its labels by the group $g = (\alpha, R, B)$'s action, with $\log \alpha, \log R_i,$ and $\log B_j$ sampled independently from $\mathrm{Uniform}[-3, 3]$. For each LP in the training set, we sample four additional training samples using the above protocol.

Transformation sampling. For each original test instance $P_{d,k}$ from distribution d, we independently sample five transformation triples

$$g_{d,k,\ell} = (\alpha_{d,k,\ell}, R_{d,k,\ell}, B_{d,k,\ell}), \qquad \ell = 1, \ldots, 5.$$

The logarithms of $\alpha_{d,k,\ell}$ and of all diagonal entries of $R_{d,k,\ell}$ and $B_{d,k,\ell}$ are sampled independently from $\mathrm{Uniform}([-3, 3])$, It uses the same log-scale range as training-time augmentation.

Evaluation actions. Each sampled triple defines four evaluation actions:

$$\begin{aligned} g^{(\alpha)}_{d,k,\ell} &= (\alpha_{d,k,\ell}, I, I), \\ g^{(R)}_{d,k,\ell} &= (1, R_{d,k,\ell}, I), \\ g^{(B)}_{d,k,\ell} &= (1, I, B_{d,k,\ell}), \\ g^{(\mathrm{joint})}_{d,k,\ell} &= (\alpha_{d,k,\ell}, R_{d,k,\ell}, B_{d,k,\ell}). \end{aligned}$$

We use $N_d = 100$ original test instances for each of the nine distributions. For each checkpoint and scale regime, the five probes per instance yield $5N_d$ transformed instances per action and $20N_d$ across all four actions. Let $\widehat{x}^{(s)}$ denote the predictor associated with the checkpoint trained using seed s, and let $\mathcal{M}^{(s)}$ denote either loss evaluated using $\widehat{x}^{(s)}$. For each action $a \in \{\alpha, R, B, \mathrm{joint}\}$, we report

$$\overline{\mathcal{M}}^{(a)} = \frac{1}{9}\sum_{d=1}^{9}\frac{1}{3}\sum_{s=1}^{3}\frac{1}{5N_d}\sum_{k=1}^{N_d}\sum_{\ell=1}^{5}\mathcal{M}^{(s)}\left(g^{(a)}_{d,k,\ell}; P_{d,k}\right), \qquad \mathcal{M} \in \{\mathcal{L}_{\mathrm{orbit}}, \mathcal{L}_{\mathrm{cons}}\}.$$

Thus, each reported value averages over transformation probes, test instances, and training seeds, with equal weight assigned to each distribution. The results are shown in Table 2. We also test the permutation equivariance of the canonicalization-based GNNs, results are shown in Table 7.

G.5.2 Training and Test Loss

Besides LP-GNNs, we test our canonicalization on the implicit models based on GNNs (Liu et al., 2026). These two choices are the most common choices to directly predict the solution mapping of LPs. The results are shown in Table 8 to Table 9. On the implicit GNN backbone, FullBcano achieves the lowest validation-selected ID test loss on all nine datasets, while one of the two canonicalized variants attains the lowest final training loss on six of them. Its advantage is even more pronounced under distribution shift: a canonicalized model achieves the lowest mean loss in

Table 7: Joint row/column permutation prediction-consistency loss.

Dataset	LP-GNN		Implicit Model	
	FullB-cano	β-cano	FullB-cano	β-cano
CA	$2.51\text{e}{-13}_{\pm 3.88\text{e}-14}$	$\mathbf{2.32\text{e}{-13}}_{\pm 7.11\text{e}-14}$	$\mathbf{6.45\text{e}{-13}}_{\pm 9.44\text{e}-14}$	$1.31\text{e}{-12}_{\pm 1.80\text{e}-13}$
SC	$\mathbf{1.12\text{e}{-13}}_{\pm 4.38\text{e}-15}$	$3.77\text{e}{-13}_{\pm 1.58\text{e}-14}$	$1.54\text{e}{-13}_{\pm 1.27\text{e}-15}$	$\mathbf{1.35\text{e}{-13}}_{\pm 7.81\text{e}-15}$
IS	$2.05\text{e}{-13}_{\pm 8.86\text{e}-14}$	$\mathbf{4.89\text{e}{-15}}_{\pm 2.81\text{e}-15}$	$\mathbf{1.58\text{e}{-14}}_{\pm 2.06\text{e}-15}$	$1.60\text{e}{-14}_{\pm 1.25\text{e}-15}$
CFLP	$\mathbf{5.25\text{e}{-13}}_{\pm 3.68\text{e}-14}$	$9.30\text{e}{-13}_{\pm 8.25\text{e}-14}$	$\mathbf{3.20\text{e}{-12}}_{\pm 4.64\text{e}-13}$	$3.94\text{e}{-12}_{\pm 9.67\text{e}-13}$
MCF	$1.59\text{e}{-13}_{\pm 7.89\text{e}-15}$	$\mathbf{1.36\text{e}{-13}}_{\pm 1.29\text{e}-14}$	$\mathbf{4.05\text{e}{-13}}_{\pm 1.41\text{e}-13}$	$5.56\text{e}{-13}_{\pm 2.78\text{e}-13}$
OPF57	$\mathbf{4.51\text{e}{-14}}_{\pm 3.36\text{e}-15}$	$1.43\text{e}{-13}_{\pm 9.35\text{e}-15}$	$9.79\text{e}{-14}_{\pm 5.00\text{e}-14}$	$\mathbf{6.45\text{e}{-14}}_{\pm 3.97\text{e}-15}$
OPF118	$\mathbf{1.01\text{e}{-13}}_{\pm 6.74\text{e}-16}$	$1.39\text{e}{-13}_{\pm 4.93\text{e}-15}$	$\mathbf{2.47\text{e}{-13}}_{\pm 7.25\text{e}-14}$	$3.28\text{e}{-13}_{\pm 5.46\text{e}-14}$
OPF300	$\mathbf{2.03\text{e}{-13}}_{\pm 2.87\text{e}-14}$	$1.63\text{e}{-7}_{\pm 7.98\text{e}-9}$	$\mathbf{6.69\text{e}{-13}}_{\pm 1.24\text{e}-13}$	$4.13\text{e}{-7}_{\pm 4.61\text{e}-8}$
NETLIB	$\mathbf{1.13\text{e}{-13}}_{\pm 5.21\text{e}-15}$	$9.53\text{e}{-8}_{\pm 1.02\text{e}-8}$	$\mathbf{5.19\text{e}{-12}}_{\pm 4.47\text{e}-12}$	$4.26\text{e}{-8}_{\pm 1.39\text{e}-8}$
Unweighted mean	$\mathbf{1.90\text{e}{-13}}$	$2.87\text{e}{-8}$	$\mathbf{1.18\text{e}{-12}}$	$5.06\text{e}{-8}$

Note. Dataset rows report three-seed mean $\pm$ standard error after averaging five shared joint row/column permutation probes per test instance. Each dataset uses 100 original test instances.

Table 8: Final training and validation-selected ID test losses for the Implicit backbone.

Final training loss

Dataset	None	Data Augmentation	HiGHS	Ruiz Equilibration	(α, R)-Normalization	FullB-cano	β-cano
CA	$0.0531_{\pm 0.0011}$	$0.0847_{\pm 6.20\text{e}-4}$	$0.0520_{\pm 0.0011}$	$0.0505_{\pm 3.46\text{e}-4}$	$0.0604_{\pm 0.0010}$	$\mathbf{0.0482}_{\pm 6.66\text{e}-4}$	$0.0527_{\pm 0.0014}$
SC	$0.0204_{\pm 6.70\text{e}-4}$	$0.0502_{\pm 0.0024}$	$0.0217_{\pm 9.88\text{e}-4}$	$0.0193_{\pm 6.44\text{e}-4}$	$0.0292_{\pm 0.0010}$	$\mathbf{0.0188}_{\pm 4.15\text{e}-4}$	$0.0197_{\pm 2.51\text{e}-4}$
OPF-57	$0.0013_{\pm 2.05\text{e}-4}$	$0.0750_{\pm 0.0106}$	$0.0290_{\pm 0.0079}$	$0.0010_{\pm 1.90\text{e}-4}$	$0.0011_{\pm 6.85\text{e}-5}$	$0.0015_{\pm 0.0010}$	$\mathbf{9.82\text{e}{-4}}_{\pm 2.15\text{e}-4}$
OPF-118	$0.0233_{\pm 0.0013}$	$0.2023_{\pm 0.0372}$	$0.1964_{\pm 0.1050}$	$\mathbf{0.0065}_{\pm 8.08\text{e}-4}$	$0.0105_{\pm 0.0015}$	$0.0146_{\pm 0.0030}$	$0.0129_{\pm 0.0034}$
OPF-300	$0.0751_{\pm 0.0028}$	$0.3095_{\pm 0.0111}$	$0.3935_{\pm 0.1381}$	$0.0283_{\pm 0.0028}$	$0.0596_{\pm 0.0067}$	$0.0334_{\pm 0.0141}$	$\mathbf{0.0239}_{\pm 0.0020}$
MCF	$0.0236_{\pm 3.64\text{e}-4}$	$0.1451_{\pm 0.0165}$	$0.0219_{\pm 0.0015}$	$0.0230_{\pm 6.19\text{e}-4}$	$0.0252_{\pm 0.0022}$	$\mathbf{0.0201}_{\pm 0.0010}$	$0.0225_{\pm 0.0011}$
CFLP	$0.1247_{\pm 0.0222}$	$1.3522_{\pm 0.2471}$	$\mathbf{0.0209}_{\pm 0.0026}$	$0.0307_{\pm 0.0060}$	$0.0623_{\pm 0.0116}$	$0.0294_{\pm 0.0030}$	$0.0272_{\pm 8.09\text{e}-4}$
IS	$0.1242_{\pm 4.12\text{e}-4}$	$0.1278_{\pm 1.54\text{e}-4}$	$\mathbf{0.1242}_{\pm 1.77\text{e}-4}$	$0.1247_{\pm 3.18\text{e}-4}$	$0.1243_{\pm 2.59\text{e}-4}$	$0.1244_{\pm 7.14\text{e}-4}$	$0.1245_{\pm 2.62\text{e}-4}$
NETLIB	$0.2495_{\pm 0.0722}$	$0.7136_{\pm 0.1267}$	$0.7018_{\pm 0.0889}$	$0.1604_{\pm 0.0644}$	$0.4760_{\pm 0.2321}$	$\mathbf{0.0243}_{\pm 0.0095}$	$0.3418_{\pm 0.0850}$

Validation-selected ID test loss

Dataset	None	Data Augmentation	HiGHS	Ruiz Equilibration	(α, R)-Normalization	FullB-cano	β-cano
CA	$0.0559_{\pm 2.72\text{e}-4}$	$0.0667_{\pm 1.65\text{e}-4}$	$0.0551_{\pm 4.45\text{e}-4}$	$0.0550_{\pm 2.60\text{e}-4}$	$0.0611_{\pm 6.31\text{e}-4}$	$\mathbf{0.0543}_{\pm 1.47\text{e}-4}$	$0.0558_{\pm 7.33\text{e}-4}$
SC	$0.0273_{\pm 3.61\text{e}-4}$	$0.0380_{\pm 2.63\text{e}-4}$	$0.0271_{\pm 1.83\text{e}-4}$	$0.0272_{\pm 2.17\text{e}-4}$	$0.0294_{\pm 1.52\text{e}-4}$	$\mathbf{0.0260}_{\pm 3.40\text{e}-4}$	$0.0269_{\pm 6.76\text{e}-4}$
OPF-57	$0.0055_{\pm 3.29\text{e}-4}$	$0.0323_{\pm 0.0020}$	$0.0293_{\pm 0.0088}$	$0.0043_{\pm 3.01\text{e}-4}$	$0.0040_{\pm 4.09\text{e}-4}$	$\mathbf{0.0031}_{\pm 0.0010}$	$0.0034_{\pm 2.88\text{e}-5}$
OPF-118	$0.0728_{\pm 0.0011}$	$0.1252_{\pm 0.0070}$	$0.1859_{\pm 0.0909}$	$0.0690_{\pm 0.0010}$	$0.0672_{\pm 6.37\text{e}-4}$	$\mathbf{0.0589}_{\pm 9.02\text{e}-4}$	$0.0651_{\pm 3.59\text{e}-4}$
OPF-300	$0.1139_{\pm 0.0030}$	$0.2175_{\pm 0.0115}$	$0.3648_{\pm 0.1220}$	$0.1094_{\pm 0.0019}$	$0.1179_{\pm 0.0027}$	$\mathbf{0.0924}_{\pm 0.0048}$	$0.1125_{\pm 0.0036}$
MCF	$0.0376_{\pm 7.89\text{e}-5}$	$0.0951_{\pm 0.0099}$	$0.0381_{\pm 2.82\text{e}-4}$	$0.0376_{\pm 2.69\text{e}-4}$	$0.0361_{\pm 5.78\text{e}-4}$	$\mathbf{0.0348}_{\pm 2.34\text{e}-4}$	$0.0361_{\pm 0.0016}$
CFLP	$0.0907_{\pm 0.0060}$	$0.5092_{\pm 0.0372}$	$0.0776_{\pm 0.0019}$	$0.0689_{\pm 0.0018}$	$0.0696_{\pm 0.0121}$	$\mathbf{0.0582}_{\pm 0.0022}$	$0.0659_{\pm 0.0032}$
IS	$0.1269_{\pm 1.36\text{e}-4}$	$0.1281_{\pm 3.71\text{e}-4}$	$0.1264_{\pm 1.45\text{e}-4}$	$0.1267_{\pm 3.36\text{e}-4}$	$0.1269_{\pm 2.16\text{e}-4}$	$\mathbf{0.1263}_{\pm 6.68\text{e}-5}$	$0.1266_{\pm 2.31\text{e}-4}$
NETLIB	$0.2201_{\pm 0.0417}$	$0.5445_{\pm 0.1392}$	$0.6281_{\pm 0.0575}$	$0.1719_{\pm 0.0440}$	$0.2486_{\pm 0.0250}$	$\mathbf{0.1098}_{\pm 0.0083}$	$0.2261_{\pm 0.0222}$

Note. Entries are three-seed mean $\pm$ standard error.

all 15 dataset–OOD-level combinations. FullB-cano ranks first in 14 cases, with β-cano performing best only on OPF-118 under OOD-3. These results demonstrate that the benefits of canonicalization are not specific to LP-GNNs, but transfer consistently to the implicit GNN backbone.

This behavior is also informative for understanding the role of numerical conditioning. Implicit models repeatedly apply the same message-passing operator for many times, so better-conditioned inputs can help stabilize the iterative computation. However, the comparison with equilibration-based preprocessing shows that conditioning alone is not sufficient: although equilibration also improves the numerical scale of the LP coefficients, it does not consistently match the performance of FullB-cano. Similarly, β-cano improves over the baseline in many cases, but is generally weaker than FullB-cano because its scalar variable scaling does not provide enough degrees of freedom to fully balance the problem or enforce the full variable-scaling symmetry. These results suggest that FullB-cano benefits from both numerical equilibration and exact canonicalization under the full diagonal variable-scaling group. In the shallow GNN regimes commonly used for LP solution prediction, the symmetry-enforcement effect appears to play the leading role, with improved conditioning providing additional stability.

Table 9: ID and OOD test losses for Implicit Model.

Dataset	Method	ID test	OOD-1	OOD-2	OOD-3
OPF-57	None	$0.0055_{\pm 3.29e-4}$	$0.1293_{\pm 0.0116}$	$0.2815_{\pm 0.0172}$	$0.3467_{\pm 0.0164}$
	Data Augmentation	$0.0323_{\pm 0.0020}$	$0.0792_{\pm 0.0040}$	$0.0935_{\pm 0.0074}$	$0.1480_{\pm 0.0107}$
	HiGHS	$0.0293_{\pm 0.0088}$	$0.0994_{\pm 0.0156}$	$0.1576_{\pm 0.0148}$	$0.2298_{\pm 0.0148}$
	Ruiz Equilibration	$0.0043_{\pm 3.01e-4}$	$0.1982_{\pm 0.0057}$	$0.3227_{\pm 0.0134}$	$0.3954_{\pm 0.0209}$
	(α, R)-Normalization	$0.0040_{\pm 4.09e-4}$	$0.1587_{\pm 0.0186}$	$0.3016_{\pm 3.41e-4}$	$0.3755_{\pm 0.0153}$
	FullB-cano	$\mathbf{0.0031}_{\pm 0.0010}$	$\mathbf{0.0458}_{\pm 0.0022}$	$\mathbf{0.0382}_{\pm 0.0010}$	$\mathbf{0.0676}_{\pm 0.0031}$
	β-cano	$0.0034_{\pm 2.88e-5}$	$0.0468_{\pm 8.61e-4}$	$0.0394_{\pm 3.04e-4}$	$0.0677_{\pm 7.86e-4}$
OPF-118	None	$0.0728_{\pm 0.0011}$	$0.2675_{\pm 0.0079}$	$0.4004_{\pm 0.0151}$	$0.5037_{\pm 0.0131}$
	Data Augmentation	$0.1252_{\pm 0.0070}$	$0.2640_{\pm 0.0127}$	$0.2966_{\pm 0.0164}$	$0.3792_{\pm 0.0182}$
	HiGHS	$0.1859_{\pm 0.0909}$	$0.3502_{\pm 0.0950}$	$0.4133_{\pm 0.0765}$	$0.4767_{\pm 0.0507}$
	Ruiz Equilibration	$0.0690_{\pm 0.0010}$	$0.3212_{\pm 0.0097}$	$0.4755_{\pm 0.0132}$	$0.5764_{\pm 0.0084}$
	(α, R)-Normalization	$0.0672_{\pm 6.37e-4}$	$0.2654_{\pm 0.0050}$	$0.4282_{\pm 0.0049}$	$0.5622_{\pm 0.0053}$
	FullB-cano	$\mathbf{0.0589}_{\pm 9.02e-4}$	$\mathbf{0.1422}_{\pm 0.0047}$	$\mathbf{0.1377}_{\pm 0.0038}$	$0.1557_{\pm 0.0043}$
	β-cano	$0.0651_{\pm 3.59e-4}$	$0.1469_{\pm 0.0041}$	$0.1442_{\pm 0.0054}$	$\mathbf{0.1495}_{\pm 0.0032}$
OPF-300	None	$0.1139_{\pm 0.0030}$	$0.2340_{\pm 0.0045}$	$0.4371_{\pm 0.0066}$	$0.5879_{\pm 0.0094}$
	Data Augmentation	$0.2175_{\pm 0.0115}$	$0.2810_{\pm 0.0171}$	$0.3657_{\pm 0.0232}$	$0.4725_{\pm 0.0212}$
	HiGHS	$0.3648_{\pm 0.1220}$	$0.4682_{\pm 0.1138}$	$0.5622_{\pm 0.0881}$	$0.6356_{\pm 0.0527}$
	Ruiz Equilibration	$0.1094_{\pm 0.0019}$	$0.2956_{\pm 0.0026}$	$0.5050_{\pm 0.0113}$	$0.6381_{\pm 0.0111}$
	(α, R)-Normalization	$0.1179_{\pm 0.0027}$	$0.2935_{\pm 0.0061}$	$0.5313_{\pm 0.0123}$	$0.6688_{\pm 0.0106}$
	FullB-cano	$\mathbf{0.0924}_{\pm 0.0048}$	$\mathbf{0.1113}_{\pm 0.0012}$	$\mathbf{0.1211}_{\pm 0.0055}$	$\mathbf{0.1416}_{\pm 0.0039}$
	β-cano	$0.1125_{\pm 0.0036}$	$0.1164_{\pm 0.0011}$	$0.1273_{\pm 3.54e-5}$	$0.1517_{\pm 0.0031}$
MCF	None	$0.0376_{\pm 7.89e-5}$	$0.1479_{\pm 0.0081}$	$0.3080_{\pm 0.0217}$	$0.5567_{\pm 0.0419}$
	Data Augmentation	$0.0951_{\pm 0.0099}$	$0.1261_{\pm 0.0121}$	$0.1593_{\pm 0.0174}$	$0.2327_{\pm 0.0400}$
	HiGHS	$0.0381_{\pm 2.82e-4}$	$0.2026_{\pm 0.0137}$	$0.4780_{\pm 0.0103}$	$0.6648_{\pm 0.0236}$
	Ruiz Equilibration	$0.0376_{\pm 2.69e-4}$	$0.0837_{\pm 0.0025}$	$0.3434_{\pm 0.0176}$	$0.5477_{\pm 0.0108}$
	(α, R)-Normalization	$0.0361_{\pm 5.78e-4}$	$0.0913_{\pm 9.93e-4}$	$0.2684_{\pm 0.0100}$	$0.5435_{\pm 0.0186}$
	FullB-cano	$\mathbf{0.0348}_{\pm 2.34e-4}$	$\mathbf{0.0514}_{\pm 4.37e-4}$	$\mathbf{0.0554}_{\pm 7.85e-4}$	$\mathbf{0.0581}_{\pm 5.25e-4}$
	β-cano	$0.0361_{\pm 0.0016}$	$0.0631_{\pm 2.60e-4}$	$0.0826_{\pm 0.0030}$	$0.0941_{\pm 0.0043}$
NETLIB	None	$0.2201_{\pm 0.0417}$	$0.4479_{\pm 0.0502}$	$1.8761_{\pm 0.1626}$	$13.7716_{\pm 8.1460}$
	Data Augmentation	$0.5445_{\pm 0.1392}$	$0.6828_{\pm 0.1572}$	$1.3202_{\pm 0.1808}$	$3.2754_{\pm 0.3374}$
	HiGHS	$0.6281_{\pm 0.0575}$	$0.7366_{\pm 0.0680}$	$1.2691_{\pm 0.1404}$	$3.6016_{\pm 0.4813}$
	Ruiz Equilibration	$0.1719_{\pm 0.0440}$	$0.4049_{\pm 0.0409}$	$1.9000_{\pm 0.2668}$	$7.5683_{\pm 1.3013}$
	(α, R)-Normalization	$0.2486_{\pm 0.0250}$	$1.2684_{\pm 0.3824}$	$14.3785_{\pm 2.8367}$	$68.8093_{\pm 6.9315}$
	FullB-cano	$\mathbf{0.1098}_{\pm 0.0083}$	$\mathbf{0.1085}_{\pm 0.0053}$	$\mathbf{0.1287}_{\pm 0.0084}$	$\mathbf{0.1935}_{\pm 0.0163}$
	β-cano	$0.2261_{\pm 0.0222}$	$0.3059_{\pm 0.0119}$	$0.3349_{\pm 0.0155}$	$0.3902_{\pm 0.0153}$