
EVA-MILP: Towards Standardized Evaluation of MILP Instance Generation

Yidong Luo, Chenguang Wang, Jiahao Yang, Fanzeng Xia, Tianshu Yu*

School of Data Science,

The Chinese University of Hong Kong, Shenzhen

yidongluo@link.cuhk.edu.cn yutianshu@cuhk.edu.cn

Abstract

Mixed-Integer Linear Programming (MILP) is fundamental to solving complex decision-making problems. The proliferation of MILP instance generation methods, driven by machine learning's demand for diverse optimization datasets and the limitations of static benchmarks, has significantly outpaced standardized evaluation techniques. Consequently, assessing the fidelity and utility of synthetic MILP instances remains a critical, multifaceted challenge. This paper introduces a comprehensive benchmark framework designed for the systematic and objective evaluation of MILP instance generation methods. Our framework provides a unified and extensible methodology, assessing instance quality across crucial dimensions: mathematical validity, structural similarity, computational hardness, and utility in downstream machine learning tasks. A key innovation is its in-depth analysis of solver-internal features –particularly by comparing distributions of key solver outputs including root node gap, heuristic success rates, and cut plane usage – leveraging the solver's dynamic solution behavior as an ‘expert assessment’ to reveal nuanced computational resemblances. By offering a structured approach with clearly defined solver-independent and solver-dependent metrics, our benchmark aims to facilitate robust comparisons among diverse generation techniques, spur the development of higher-quality instance generators, and ultimately enhance the reliability of research reliant on synthetic MILP data. The framework’s effectiveness in systematically comparing the fidelity of instance sets is demonstrated using contemporary generative models. The code is available in <https://github.com/anonymous-neurips-submission-2025/EVA-MILP>.

1 Introduction

Mixed-Integer Linear Programming (MILP) is a fundamental optimization framework extending standard Linear Programming (LP) by incorporating integer variables, enabling the modeling of discrete decisions and logical conditions often intractable with purely continuous models [1, 2]. Its versatility in capturing complex relationships like fixed costs, mutual exclusivity, and indivisible entities makes MILP essential for practical decision-making in fields such as supply chain optimization, scheduling, financial modeling, and network design [3, 4, 5, 6].

Traditionally, MILP solver advancement relied on benchmark libraries like MIPLIB [7] and other real-world instance collections. However, these static benchmarks often lack the scale, diversity, or controllable characteristics vital for modern research. This limitation is particularly acute with the rise of machine learning (ML) in optimization, where learning-based approaches for tasks like branching strategy selection or algorithm configuration demand large, varied datasets that often exceed available collections [8]. Consequently, a significant shift towards proactive generation of synthetic MILP instances has occurred [9, 10, 11, 12, 13, 14, 15, 16]. This trend towards instance generation is motivated by multiple advantages: fulfilling the need for extensive datasets in ML

applications; achieving greater instance diversity than found in existing libraries [7]; enabling control over computational difficulty for rigorous algorithm testing; facilitating the simulation of specific problem structures [13]; aiding solver testing and debugging [17]; and creating privacy-preserving surrogates for confidential real-world data.

Methodologies for MILP instance generation have evolved in tandem. Early approaches featured parameterized generators for specific problem classes such as the traveling salesman problems [18, 19], set covering [20], or quadratic assignment problems [21]. Later techniques emphasized more generalized feature-based descriptions and sampling [9, 22]. Most recently, deep learning based frameworks like G2MILP [10], MILP-StuDio [13], and MILP-Evolve [23] have gained prominence. These often represent MILP instances as graphs (e.g., bipartite graphs) and use generative models like variational autoencoders [24] to learn and replicate specific structures from training data, marking a significant potential advancement but also introducing new evaluation challenges.

Despite this progress in generation, a critical and multifaceted challenge remains: the fair, objective, and comprehensive evaluation of synthetic MILP instance “quality.” Assessing generator outputs effectively requires considering several dimensions: **(I)** Ensuring fundamental solvability (feasibility and boundedness) is crucial yet non-trivial [11, 10, 13]. **(II)** Generated instances should ideally mirror real-world structural patterns [7], but mere resemblance may not capture true computational behavior. **(III)** Instances must exhibit appropriate and controllable computational hardness, typically evaluated via solver performance [17], though achieving desired hardness levels consistently is a significant difficulty. Furthermore, current evaluation protocols [10, 9, 16] often lack standardization, hindering robust comparisons and impeding progress.

To address the aforementioned challenges and lack of standardization, this research introduces a novel benchmark framework for the systematic and comprehensive evaluation of MILP instance generation techniques. Our framework provides a unified, extensible methodology to assess instance quality along key dimensions – mathematical validity, structural similarity, computational hardness, solver interaction patterns, and utility for downstream ML tasks – aiming to establish an objective standard. This, in turn, will facilitate fairer comparisons, guide the development of higher-quality generators, and improve the reliability of research utilizing synthetic MILP data. Specifically, our contributions are: **(I)** We define and integrate key evaluation metrics into a unified framework, assessing instances based on both Solver-Independent Metrics (covering fundamental properties and structural resemblance) and Solver-Dependent Metrics (including computational characteristics and downstream task utility), thereby providing a more holistic assessment than existing evaluations. **(II)** Our framework uniquely incorporates detailed analysis of solver-internal features (e.g., root node gap, heuristic success counts, cut plane profiles analyzed via PCA and Wasserstein distance), treating the solver’s behavior as an expert assessment to reveal deeper computational similarities beyond surface structure. **(III)** The proposed benchmark is intentionally designed with extendability in mind; its modular structure facilitates the future integration of additional evaluation metrics, new datasets representing diverse MILP problem types, and novel generation techniques as the field progresses, ensuring its long-term relevance. **(IV)** It enables objective and fair comparisons between diverse MILP instance generators (including rule-based, statistical, and machine learning-based approaches) by offering clear, operational metrics and quantitative criteria, thereby providing guidance for the development of higher-quality generation methods for the community.

Furthermore, this work highlights a limitation in current DL-based generation approaches. While representing MILP instances as bipartite graphs has enabled progress, particularly in achieving structural similarity, this representation may struggle to fully capture the intricate constraint relationships and inherent mathematical properties crucial for MILP feasibility and hardness. We posit that future advancements in MILP instance generation might benefit from shifting focus beyond direct graph structure manipulation (like node/constraint operations) towards methods that more directly model or preserve the underlying mathematical structure, or solution space characteristics. This proposed benchmark provides the necessary tools to evaluate such novel approaches rigorously.

2 Preliminaries

MILP and its Bipartite Graph Representation. We consider the standard Mixed-Integer Linear Programming (MILP) problem:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \tilde{\mathbf{c}}^\top \mathbf{x}, \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}, \quad x_i \in \mathbb{Z}, \quad \forall i \in \mathcal{I}.$$

Here, \mathbf{x} is the vector of decision variables, $\tilde{\mathbf{c}}$ contains the objective coefficients, \mathbf{A} is the constraint matrix, \mathbf{b} is the vector of constraint bounds, and \mathcal{I} identifies the integer variables.

Representing MILP instances as weighted bipartite graphs is an established practice in the relevant literature [25, 26, 27]. In this common representation, each MILP instance corresponds to a graph \mathcal{G} defined by three sets: constraint vertices \mathcal{C} , variable vertices \mathcal{V} , and edges \mathcal{E} connecting them. The constraint partition $\mathcal{C} = \{c_1, \dots, c_m\}$ includes one vertex c_i for each of the m constraints, where the vertex feature c_i typically represents the bias term, i.e., $c_i = (b_i)$ (the i -th element of \mathbf{b} from the MILP formulation). The variable partition $\mathcal{V} = \{v_1, \dots, v_n\}$ contains one vertex v_j for each of the n variables, with the corresponding 9-dimensional feature vector \mathbf{v}_j encoding information such as the objective coefficient c_j (from the MILP vector \mathbf{c}), the variable type, and the bounds l_j, u_j (from the MILP vectors \mathbf{l} and \mathbf{u}). Edges in \mathcal{E} only exist between vertices from different partitions ($c_i \in \mathcal{C}$ and $v_j \in \mathcal{V}$). An edge $e_{i,j}$ connects c_i and v_j if the corresponding element $A_{i,j}$ in the constraint matrix \mathbf{A} is non-zero; its associated feature $e_{i,j}$ is described by this coefficient, i.e., $e_{i,j} = (A_{i,j})$. If $A_{i,j}$ is zero, the edge $e_{i,j}$ is considered absent.

Linear Programming Relaxation. Given the MILP problem:

$$z_{\text{MILP}} = \min_{\mathbf{x}} \{ \tilde{\mathbf{c}}^\top \mathbf{x} \mid \mathbf{Ax} \leq \mathbf{b}, x_i \in \mathbb{Z} \forall i \in \mathcal{I} \subseteq \{1, \dots, n\} \}.$$

The *LP relaxation* is obtained by omitting the integrality constraints:

$$z_{\text{LP}} = \min_{\mathbf{x}} \{ \tilde{\mathbf{c}}^\top \mathbf{x} \mid \mathbf{Ax} \leq \mathbf{b} \}.$$

Let $\mathcal{P} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{Ax} \leq \mathbf{b}\}$ be the feasible region of the LP relaxation. The optimal objective value of the LP relaxation provides a lower bound (for minimization problems) to the optimal objective value of the original MILP, i.e., $z_{\text{LP}} \leq z_{\text{MILP}}$.

Duality Gap. In the context of solving the MILP $\min \tilde{\mathbf{c}}^\top \mathbf{x}$, the *duality gap* measures the difference between the objective value of the best known feasible integer solution, $z_{\text{incumbent}}$ (primal bound or upper bound, UB), and the best proven objective bound, z_{bound} (dual bound or lower bound, LB), typically derived from LP relaxations ($z_{\text{bound}} \geq z_{\text{LP}}$). The absolute gap is $G_{\text{abs}} = z_{\text{incumbent}} - z_{\text{bound}}$ (assuming $z_{\text{incumbent}} \geq z_{\text{bound}}$). The relative gap, often used as a termination criterion, quantifies the remaining gap relative to one of the bounds, e.g.:

$$G_{\text{rel}} = \frac{z_{\text{incumbent}} - z_{\text{bound}}}{\max(1, |z_{\text{incumbent}}|)}.$$

Optimality is proven when $z_{\text{incumbent}} = z_{\text{bound}}$, or G_{abs} (or G_{rel}) is within a predefined tolerance $\epsilon \geq 0$.

Additional preliminaries can be found in Appendix E.

3 Methodology

This section outlines the methodology employed to evaluate the quality and characteristics of generated MILP instances. We compare instances generated using reproductions of three open-source methods – G2MILP [10], ACM-MILP [12], and DIG-MILP [11] – against a baseline set of ‘original’ instances. This baseline set, comprising Set Cover, Combinatorial Auction (CA), Capacitated Facility Location (CFL), and Independent Set (IS) problems, was synthesized using the Ecole library [28] (the specific parameter settings used for this synthesis are detailed in the Appendix A), while IP and LB come from two challenging real-world problem families used in ML4CO 2021 competition [29]. To capture different facets of similarity and fidelity between the generated and original sets, we utilize a suite of metrics. These metrics are organized into two primary categories based on their reliance on the underlying optimization solver: **(I) Solver-Independent Metrics**, which assess inherent properties of the MILP instances themselves (e.g., feasibility, structure) without regard to any specific solution process; and **(II) Solver-Dependent Metrics**, which evaluate aspects intrinsically linked to the behavior and performance of a particular solver (primarily Gurobi) when applied to the instances. This dual approach allows for a robust assessment, considering both fundamental instance properties and their practical implications for optimization algorithms (See the whole framework in Figure 1).

3.1 Dataset

This study evaluates MILP instances from two sources: original instances (baselines and validation) and those from three generative models (G2MILP, ACM-MILP, DIG-MILP). Baselines include four

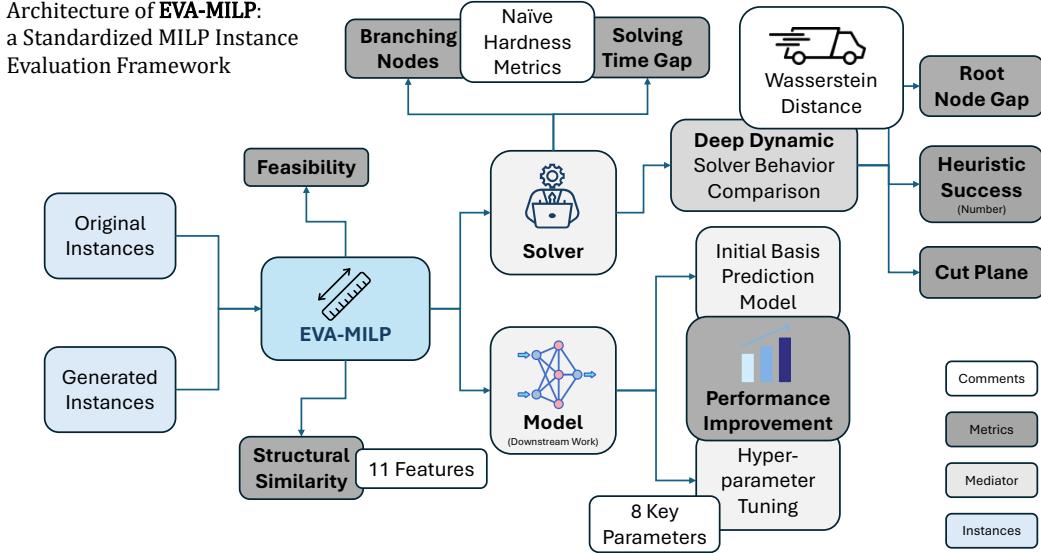


Figure 1: Architecture of EVA-MILP

synthetic datasets (SC, CA, CFL, IS) from Ecole [28] (details in Appendix A) and public benchmarks from the ML4CO Competition 2024 [29] for pipeline validation. Due to the inherent specialization of generative models to particular problem types, our comparative analysis, which forms the core of our evaluation, specifically examines SC instances from G2MILP, CA instances from ACM-MILP and DIG-MILP, and IS instances from ACM-MILP and G2MILP to demonstrate our metrics. The methodology's applicability to other datasets and models has been confirmed.

3.2 Solver-Independent Metrics

3.2.1 Feasibility Ratio

This metric measures the percentage of instances within a given dataset that are both feasible and bounded.

Results The feasibility and boundedness of generated instance sets for IS, CA, and SC problem domains were evaluated. Most generative models successfully maintained 100% feasibility and boundedness, mirroring the original datasets. A minor exception was G2MILP (MIS, $\eta = 0.10$), which had a 93.40% feasibility ratio. This suggests that while generally robust, some model configurations, particularly with specific problem structures and higher mask ratios, might occasionally produce invalid instances. Results could be found in Table 10, 11, and 12 in the Appendix.

3.2.2 Structural Similarity

To quantitatively assess the structural similarity between two sets of MILP instances, following [10], we implemented a feature-based comparison method. Each MILP instance is first converted into a graph representation. From this representation, a vector of 11 predefined structural features is computed for each instance. For each of the 11 features, we compute the Jensen-Shannon Divergence (JSD) between the distributions observed in the two sets. The JSD value for each feature i is transformed into a similarity score S_i using the formula $S_i = 1 - \frac{JSD_i}{\log(2)}$, where a score of 1 indicates identical distributions and 0 indicates maximal divergence. The overall structural similarity score between the two sets is then calculated as the arithmetic mean of the individual feature similarity scores (S_i).

Results The JSD-based metric (Table 1) shows that models vary in their ability to replicate structural characteristics across problem domains. For CA instances, ACM-MILP and DIG-MILP achieved high structural similarity to reference data (scores 0.860-0.892). In contrast, for IS problems, ACM-

Table 1: Structural Similarity Across Problem Types, Models, and Mask Ratios (η)

Problem	Model	Overall Similarity at Masked Ratio (η)			
		0.01	0.05	0.10	0.20
CA	ACM-MILP	—	0.889	0.867	0.892
	DIG-MILP	0.861	0.860	0.879	—
IS	ACM-MILP	—	0.734	0.731	0.699
	G2MILP	0.486	0.473	0.466	—
SC	G2MILP	0.990	0.950	0.921	—

MILP showed moderate similarity (0.699-0.734), while G2MILP’s scores were lower (0.466-0.486). G2MILP, however, excelled with Set Cover problems, yielding very high similarity (0.921-0.990), particularly at lower mask ratios.

3.3 Solver-Dependent Metrics

3.3.1 Branching nodes

Computational hardness was assessed using the number of branch-and-bound nodes [30] explored by the Gurobi solver during the optimization process. The formula used is:

$$\text{Relative Error} = \left| \frac{\sum \text{Nodes}_{\text{generated}} - \sum \text{Nodes}_{\text{training}}}{\sum \text{Nodes}_{\text{training}}} \right| \times 100\%$$

Results The Gurobi branching node counts (Appendix. F Table 13) reveal that computational hardness is highly sensitive to both the generative model and problem type. For IS instances, G2MILP drastically increased complexity (e.g., RE > 50,000% for $\eta = 0.10$), leading to frequent time-outs, while ACM-MILP instances remained close to the baseline’s simplicity. A similar divergence occurred with CA instances: ACM-MILP produced substantially harder problems than the baseline (e.g., RE > 26,000% for $\eta = 0.20$), whereas DIG-MILP only modestly increased difficulty. For SC problems, G2MILP’s generated instances were moderately harder but remained tractable.

3.3.2 Solving Time Gap

This metric aims to quantify the solver-conditional hardness similarity between the two sets of instances by measuring the percentage difference in their average solving times. A smaller difference indicates higher similarity in hardness for the specific solver used. Prior work, such as [10], has also compared generated and original instances based on relative differences in solving times.

Results The Solving Time Gap (Table 17) reveals varied impacts of generative models on instance hardness. For Set Cover, G2MILP instances showed modest time differences (11–22%) from originals. In Combinatorial Auctions, ACM-MILP dramatically increased solution times (gaps >2400%), while DIG-MILP instances were closer (20–29% gaps). For Independent Set problems, ACM-MILP generated easier instances (times 47–49% shorter). Overall, instance difficulty is highly sensitive to the model and problem type, with mask ratio η playing a secondary role. G2MILP (for SC) and DIG-MILP (for CA) better preserved original difficulty levels.

3.3.3 Hyperparameter Tuning

Automatically optimizing solver hyperparameters is known to be crucial for achieving peak performance on complex algorithms like MILP solvers [31]. Therefore, we employed the Sequential Model-based Algorithm Configuration (SMAC3) framework [32], which utilizes Bayesian optimization. While this optimization approach is general and flexibly extendable to tune various MILP solvers, our current work focused specifically on Gurobi. Following the approach in [13], the tuning process targeted 8 key Gurobi parameters. These parameters were selected as they govern diverse and fundamental components of the MILP solution process, including primal heuristics, search strategy focus, branching decisions, presolving routines, cutting plane generation, and node LP solution

methods. The significant impact of these core solver components, and thus the parameters controlling them, on overall performance is well-documented in the MILP literature [e.g., 33, 34, 35]. The objective function for SMAC3 was the minimization of the mean wall-clock solve time across the instances within a designated tuning set. The primary goal of this hyperparameter tuning experiment was to evaluate the generalization capability of the optimized Gurobi configurations using generated MILP instances.

Results Hyperparameter tuning with SMAC3 (Table 2) yielded varied performance gains. For example, tuning significantly benefited the ACM-MILP (IS, $\eta = 0.1$) dataset, reducing solve time by 14.15%. In contrast, improvements for DIG-MILP (CA, $\eta = 0.05$) and G2MILP (SC, $\eta = 0.1$) were more modest, at 1.15% and 1.59% respectively. These results suggest that the potential for performance enhancement through tuning these specific Gurobi parameters is notably dependent on the dataset and generative model in question.

Table 2: Solving time improvement after Gurobi hyperparameter tuning. Compares the average wall-clock time (seconds) on the test set for the default configuration versus the best configuration found by SMAC3

Dataset	Source/Model	η	Default Time (s)	Best Time (s)	Improvement (%)
IS	ACM-MILP	0.05	0.339	0.370	-9.17
		0.1	0.351	0.280	20.25
		0.2	0.339	0.377	-11.13
	Original Dataset	—	0.336	0.366	8.98
CA	DIG-MILP	0.01	0.107	0.041	61.35
		0.05	0.109	0.042	61.63
		0.1	0.108	0.042	61.38
	Original Dataset	—	0.107	0.042	61.26
SC	G2MILP	0.01	0.211	0.118	43.97
		0.05	0.211	0.118	44.06
	Original Dataset	—	0.229	0.221	3.31

3.3.4 Initial Basis Prediction

Following the methodology of [36], we investigated predicting an initial basis for MILP relaxations using a GNN. While adhering to the principles outlined by Fan et al., our implementation was developed independently from scratch. For this purpose, a GNN model was trained specifically on generated MILP instances. The model was trained to predict the basis status (basic or non-basic at lower/upper bounds) for variables and slack variables associated with the MILP relaxation. The GNN’s predictions were subsequently refined into a numerically stable basis using established techniques. To evaluate the effectiveness of training on generated data versus original data for this task, we measured the impact of using the initial basis predicted by the GNN on Gurobi’s performance metrics (solving time, runtime). This was compared against Gurobi’s default setting and potentially a model trained on original data, using unseen test instances. The validity and effectiveness of our re-implemented approach are substantiated by the experimental outcomes.

Results Table 19 in Appendix.J compares solving performance with GNN-predicted initial basis across different datasets. The effectiveness varied by problem type: For CA problems, both ACM-MILP and DIG-MILP generated instances showed modest improvements (2.6%-4.1%); for IS problems, GNN predictions on generated instances slightly decreased performance (-0.2% to -0.6%); while for SC problems, G2MILP instances showed substantial improvements (9.1%-12.4%), comparable to those on original data. These results indicate that the utility of initial basis prediction is domain-dependent, with Set Cover problems benefiting most significantly.

3.3.5 Solver-Internal Features

We introduce a novel, computationally efficient methodology to evaluate MILP instance fidelity using solver-internal features: metrics from Gurobi’s deterministic solving process. This approach rapidly assesses and effectively discerns instance set similarities. Its efficacy was validated by a split-half

experiment (see Appendix K.1). Figure 2 presents one result of the validation experiments, showing low 1-Wasserstein distances (0.130 in heuristic; 0.188 in root node gap; 0.232 in cut plane) between random original dataset subsets, confirming its capability for authentic inter-set similarity evaluation.

With fixed Gurobi settings, we extracted: Root Node Gap, Heuristic Success Count, and Cut Plane Usage vectors. Distributions for Root Node Gap and Heuristic Success Count were compared using the 1-Wasserstein distance (W_1). For Cut Plane Usage, per-instance vectors were preprocessed, underwent PCA, and distributions of scores on dominant principal components were then compared using the W_1 distance.

Results Table 3 presents W_1 distances for root gap distributions (lower W_1 indicates higher similarity). G2MILP highly replicated SC root gaps (e.g., $W_1 = 0.2294$ at $\eta = 0.05$). In contrast, DIG-MILP and ACM-MILP significantly diverged for CA. ACM-MILP showed moderate success for IS at lower η , diminishing as η increased (e.g., $W_1 \approx 0.87$ for $\eta = 0.10$ vs. $W_1 = 1.0388$ for $\eta = 0.20$).

Table 3: Comparison of Root Node Gap

Benchmark	Generated Dataset		Mean Gap (%)		Std Dev Gap (%)		W_1 Dist.
	Problem	Model	Ratio (η)	Gen	Bench	Gen	Bench
CA	DIG-MILP	0.01	1.11	2.47	1.22	2.60	1.3001
		0.05	1.09	2.47	1.17	2.60	1.3787
		0.10	1.18	2.47	1.26	2.60	1.2982
	ACM-MILP	0.05	7.93	2.47	3.85	2.60	5.4638
		0.10	7.49	2.47	3.60	2.60	5.0187
		0.20	9.54	2.47	4.40	2.60	7.0666
IS	ACM-MILP	0.05	2.65	3.42	1.82	1.34	0.8818
		0.10	2.66	3.42	1.81	1.34	0.8646
		0.20	2.45	3.42	1.80	1.34	1.0388
SC	G2MILP	0.01	8.25	7.61	4.19	4.12	0.6690
		0.05	7.75	7.61	4.27	4.12	0.2294
		0.10	8.05	7.61	4.12	4.12	0.5103

For primal heuristic success distributions (Table 4, W_1 distance), DIG-MILP (e.g., $W_1 \approx 0.14$) and ACM-MILP (e.g., $W_1 \approx 0.34 - 0.44$) achieved good similarity for CA problems. G2MILP (SC) demonstrated remarkable similarity ($W_1 \approx 0.08 - 0.10$). Conversely, ACM-MILP (IS) instances diverged significantly ($W_1 \approx 17.0 - 17.2$). Fidelity appears model and problem-dependent, with η having a lesser role.

Table 4: Comparison of Heuristic Success Frequency Distributions between Generated (Gen.) and corresponding Benchmark (Base) MILP Instances

Problem	Model	Generated Dataset Config		Gen. Stats		Base Stats		W_1
		η	Mean	Std Dev	Mean	Std Dev		
CA	ACM-MILP	0.05	2.549	0.500	2.115	0.379	0.434	
		0.10	2.552	0.501	2.115	0.379	0.437	
		0.20	2.457	0.500	2.115	0.379	0.342	
	DIG-MILP	0.01	2.238	0.451	2.115	0.379	0.135	
		0.05	2.241	0.439	2.115	0.379	0.126	
		0.10	2.258	0.440	2.115	0.379	0.143	
IS	ACM-MILP	0.05	1.662	0.473	18.883	12.112	17.221	
		0.10	1.746	0.435	18.883	12.112	17.137	
		0.20	1.849	0.358	18.883	12.112	17.034	
SC	G2MILP	0.01	2.619	0.533	2.702	0.491	0.083	
		0.05	2.600	0.548	2.702	0.491	0.102	
		0.10	2.598	0.541	2.702	0.491	0.104	

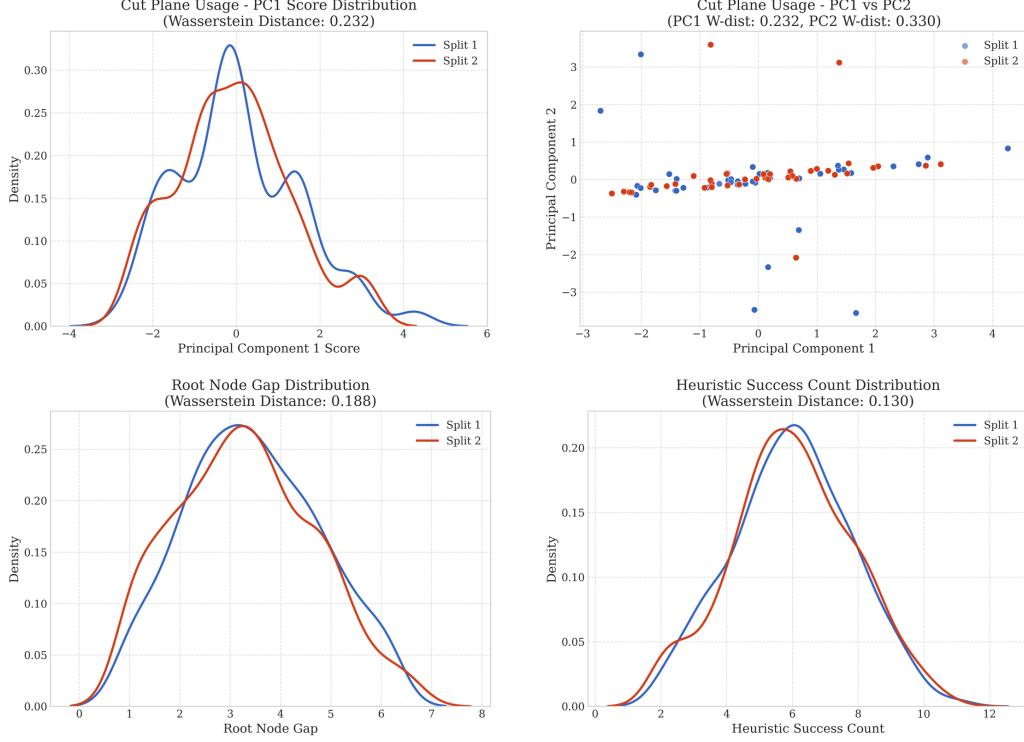


Figure 2: Internal Cut-plane Comparison of IS Dataset (Demonstration of Identical Distribution)

For cutting plane usage (Table 5, PCA then W_1 distance on PC scores), G2MILP (SC) achieved the highest similarity (e.g., PC1 $W_1 \approx 0.07 - 0.17$). In contrast, ACM-MILP and DIG-MILP for CA instances showed significant divergences (DIG-MILP PC1 $W_1 \approx 1.47 - 1.69$). ACM-MILP (IS) demonstrated intermediate similarity (PC1 $W_1 \approx 0.89 - 1.01$). Masking ratio (η) impact was inconsistent; similarity primarily depends on model-problem combinations.

Table 5: Comparison of Cutting Plane Usage Profiles (1-Wasserstein Distance on Principal Component Scores)

Problem	Model	Eta (η)	Wasserstein Distance		
			PC1	PC2	PC3
CA	ACM-MILP	0.05	1.5570	0.8352	0.5349
		0.10	1.4837	0.7155	0.3570
		0.20	1.6889	0.3166	0.3731
	DIG-MILP	0.01	1.4711	0.6673	0.3601
		0.05	1.4868	0.6826	0.3047
		0.10	1.5392	0.4394	0.5988
IS	ACM-MILP	0.05	0.8944	0.4986	0.2517
		0.10	0.9354	0.3942	0.0963
		0.20	1.0097	0.5735	0.1199
SC	G2MILP	0.01	0.0707	0.1271	0.1124
		0.05	0.1714	0.1425	0.0677
		0.10	0.1374	0.1028	0.1375

4 Discussion

- (i). **Superficial structural similarity is an unreliable predictor of computational behavior and difficulty in generated MILP instances.** ACM-MILP (CA) instances showed high structural similarity but were exceptionally difficult to solve, with Root Node Gap discrepancies suggesting that structural metrics miss key complexity drivers or that models introduce subtle, difficulty-enhancing variations.
- (ii). **Simple outcome metrics, such as solving time or branching nodes, do not adequately reveal the source of problem difficulty or underlying changes to instance structure.** This is exemplified by ACM-MILP (IS) instances, which, despite some similar outcome metrics (e.g., solving time), exhibited vastly different heuristic behavior, indicating altered internal structures that impact solvability in ways not captured by these surface-level performance measures.
- (iii). **The effectiveness of generative models is problem-dependent, linked to how well their modification strategies align with the core structural determinants of hardness for specific problem types.** G2MILP’s difficulty in generating realistic IS instances is likely due to IS problem hardness being rooted in global graph topology [37, 38], which its local edge constraint modifications struggle to preserve. G2MILP’s better performance on SC problems suggests its direct manipulation of constraint configurations [20] more effectively captures their structural hardness.

While we advocate for the EVA-MILP framework and its emphasis on solver-internal features as a significant advancement, potential limitations warrant acknowledgment. The initial selection of “useful metrics”, though comprehensive, may necessitate ongoing refinement to include emerging or problem-specific indicators and avoid inherent biases. Furthermore, extensive evaluation using rich solver-internal features can be computationally intensive, particularly for large instance sets. Solver-dependency of some internal features also requires careful consideration when generalizing findings. Finally, while EVA-MILP aims to simplify evaluation, comprehensively interpreting multi-faceted data into actionable insights may still demand considerable expertise and remains an area for future development. More limitations are in Appendix. M

5 Conclusion

The challenges observed in evaluating synthetic MILP instances, particularly the unreliability of traditional structural similarity metrics and the limitations of simple outcome measures, necessitate a paradigm shift in assessment strategies. Our findings consistently demonstrate that generating instances merely possessing superficial structural resemblance does not guarantee they will exhibit comparable computational challenges or elicit solver behaviors akin to real-world problems. We strongly advocate for leveraging solver-internal features as a primary means for a more robust and insightful assessment of instance similarity. In contrast to static features describing only surface ‘appearance’ or outcome metrics offering an aggregate ‘grade’, solver-internal features illuminate the dynamic interplay between an instance’s intrinsic structure and the solver’s algorithmic components. By examining metrics such as cutting plane usage profiles (e.g., the prevalence of Gomory cuts indicating LP relaxation quality [39]) or the success rates of various primal heuristics (suggesting amenability to finding feasible solutions [40]), a more nuanced understanding of the computational challenges inherent in an instance can be achieved.

Focusing on these internal solver dynamics provides a more reliable and fine-grained benchmark. Matching the patterns and distributions of such internal metrics facilitates a comparison based on solver-perceived structure and the actual difficulties encountered during optimization. Our proposed EVA-MILP evaluation framework, which integrates a comprehensive set of useful metrics and supports extensibility, is designed to enable researchers to conveniently assess both generated and non-generated instances. Ultimately, the adoption of such rigorous evaluation methodologies, centered on solver-internal behavior and facilitated by frameworks like EVA-MILP, will be instrumental in developing more representative benchmark instances, thereby fostering more impactful research and development within the field of mathematical optimization.

References

- [1] Tobias Achterberg and Roland Wunderling. *Mixed Integer Programming: Analyzing 12 Years of Progress*. Springer, 2013. doi: 10.1007/978-3-642-38189-8_18.
- [2] Laurence A. Wolsey. *Integer Programming*. John Wiley & Sons, 2 edition, 2020. ISBN 978-1-119-60653-6.
- [3] Michael H. Hugos. *Essentials of Supply Chain Management*. John Wiley & Sons, 2018. ISBN 978-1-119-46110-4.
- [4] Juergen Branke, Su Nguyen, Christoph W. Pickardt, and Mengie Zhang. Automated design of production scheduling heuristics: A review. *IEEE Transactions on Evolutionary Computation*, 20(1):110–124, 2015. doi: 10.1109/TEVC.2015.2429314.
- [5] Renata Mansini, Włodzimierz Ogryczak, and M. Grazia Speranza. *Linear and Mixed Integer Programming for Portfolio Optimization*, volume 21. Springer, 2015. EURO: The Association of European Operational Research Societies.
- [6] Naser Al-Falahy and Omar Y Alani. Technologies for 5g networks: Challenges and opportunities. *IT Professional*, 19(1):12–20, 2017.
- [7] Ambros Gleixner, Gregor Hendel, Gerald Gamrath, Tobias Achterberg, Michael Bastubbe, Timo Berthold, Philipp Christopel, Kati Jarck, Thorsten Koch, Jeff Linderoth, et al. Miplib 2017: Data-driven compilation of the 6th mixed-integer programming library. *Mathematical Programming Computation*, 13(3):443–490, 2021.
- [8] Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine learning for combinatorial optimization: A methodological tour d’horizon. *European Journal of Operational Research*, 290(2):405–421, 2021.
- [9] Simon Bowly et al. Generation techniques for linear programming instances with controllable properties. *Mathematical Programming Computation*, 12(3):389–415, 2020.
- [10] Zijie Geng et al. A deep instance generative framework for milp solvers under limited data availability. In *Advances in Neural Information Processing Systems*, volume 36, pages 26025–26047, 2023.
- [11] Haoyu Wang et al. Dig-milp: A deep instance generator for mixed-integer linear programming with feasibility guarantee. *arXiv preprint arXiv:2310.13261*, 2023.
- [12] Ziao Guo et al. Acm-milp: Adaptive constraint modification via grouping and selection for hardness-preserving milp instance generation. In *Forty-first International Conference on Machine Learning*, 2024.
- [13] Haoyang Liu et al. MILP-StuDio: MILP instance generation via block structure decomposition. *arXiv preprint arXiv:2410.22806*, 2024.
- [14] Tianxing Yang, Huigen Ye, and Hua Xu. Learning to generate scalable milp instances. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 2024.
- [15] Hao Zeng et al. Effective generation of feasible solutions for integer programming via guided diffusion. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024.
- [16] Yahong Zhang et al. Milp-fbgen: Lp/milp instance generation with feasibility/boundedness. In *Forty-first International Conference on Machine Learning*, 2024.
- [17] Gurobi Optimization LLC. Gurobi optimizer. URL <http://www.gurobi.com>, 2021. Accessed: DATE.
- [18] Martha G Pilcher and Ronald L Rardin. Partial polyhedral description and generation of discrete optimization problems with known optima. *Naval Research Logistics (NRL)*, 39(6):839–858, 1992.

- [19] Russ J Vander Wiel and Nikolaos V Sahinidis. Heuristic bounds and test problem generation for the time-dependent traveling salesman problem. *Transportation Science*, 29(2):167–183, 1995.
- [20] Egon Balas and Andrew Ho. *Set Covering Algorithms Using Cutting Planes, Heuristics, and Subgradient Optimization: A Computational Study*. Springer, 1980.
- [21] J Culberson. A graph generator for various classes of k-colorable graphs. URL <http://webdocs.cs.ualberta.ca/~joe/Coloring/Generators/generate.html>, 2002.
- [22] Kate Smith-Miles and Simon Bowly. Generating new test instances by evolving in instance space. *Computers & Operations Research*, 63:102–113, 2015.
- [23] Sirui Li et al. Towards foundation models for mixed integer linear programming. *arXiv preprint arXiv:2410.08288*, 2024.
- [24] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.
- [25] J. Zhang, C. Liu, X. Li, H.-L. Zhen, M. Yuan, Y. Li, and J. Yan. A survey for solving mixed integer programming via machine learning. *Neurocomputing*, 519:205–217, 2023.
- [26] Maxime Gasse, Didier Chételat, Nicola Ferroni, Laurent Charlin, and Andrea Lodi. Exact Combinatorial Optimization with Graph Convolutional Neural Networks, October 2019. URL <http://arxiv.org/abs/1906.01629>. arXiv:1906.01629 [cs, math, stat].
- [27] V. Nair, S. Bartunov, F. Gimeno, I. von Glehn, P. Lichocki, I. Lobov, B. O’Donoghue, N. Sonnerat, C. Tjandraatmadja, P. Wang, R. Addanki, T. Hapuarachchi, T. Keck, J. Keeling, P. Kohli, I. Ktena, Y. Li, O. Vinyals, and Y. Zwols. Solving mixed integer programs using neural networks. *arXiv preprint*, 2021.
- [28] Antoine Prouvost, Justin Dumouchelle, Lara Scavuzzo, Maxime Gasse, Didier Chételat, and Andrea Lodi. Ecole: A gym-like library for machine learning in combinatorial optimization solvers. In *Learning Meets Combinatorial Algorithms at NeurIPS2020*, 2020. URL <https://openreview.net/forum?id=IVc9hqgibyB>.
- [29] Maxime Gasse, Simon Bowly, Quentin Cappart, Jonas Charfreitag, Laurent Charlin, Didier Chételat, Antonia Chmiela, Justin Dumouchelle, Ambros Gleixner, Aleksandr M. Kazachkov, Elias Khalil, Paweł Lichocki, Andrea Lodi, Miles Lubin, Chris J. Maddison, Morris Christopher, Dimitri J. Papageorgiou, Augustin Parjadis, Sebastian Pokutta, Antoine Prouvost, Lara Scavuzzo, Giulia Zarpellon, Linxin Yang, Sha Lai, Akang Wang, Xiaodong Luo, Xiang Zhou, Haohan Huang, Shengcheng Shao, Yuanming Zhu, Dong Zhang, Tao Quan, Zixuan Cao, Yang Xu, Zhewei Huang, Shuchang Zhou, Chen Binbin, He Minggui, Hao Hao, Zhang Zhiyu, An Zhiwu, and Mao Kun. The machine learning for combinatorial optimization competition (ml4co): Results and insights. In Douwe Kiela, Marco Ciccone, and Barbara Caputo, editors, *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pages 220–231. PMLR, Dec 2022. URL <https://proceedings.mlr.press/v176/gasse22a.html>.
- [30] Yang Li, Xinyan Chen, Wenxuan Guo, Xijun Li, Wanqian Luo, Junhua Huang, Hui-Ling Zhen, Mingxuan Yuan, and Junchi Yan. Hardsatgen: Understanding the difficulty of hard sat formula generation and a strong structure-hardness-aware baseline. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1234–1244. ACM, 2023.
- [31] Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *International Conference on Learning and Intelligent Optimization*, pages 507–523. Springer, 2011.
- [32] Marius Lindauer, Katharina Eggensperger, Matthias Feurer, André Biedenkapp, Difan Deng, Carolin Benjamins, Tim Ruhkopf, René Sass, and Frank Hutter. Smac3: A versatile bayesian optimization package for hyperparameter optimization. *Journal of Machine Learning Research*, 23(54):1–9, 2022. URL <http://jmlr.org/papers/v23/21-0888.html>.

- [33] Andrea Lodi and Giulia Zarpellon. On the algorithmic effectiveness of modern MILP solvers. *Operations Research Perspectives*, 4:14–23, 2017.
- [34] Tobias Achterberg. SCIP: solving constraint integer programs. *Mathematical Programming Computation*, 1(1):1–41, 2009.
- [35] Timo Berthold. Primal heuristics for mixed integer programming. In *Mixed Integer Nonlinear Programming*, pages 29–62. Springer, 2012.
- [36] Zhenan Fan et al. Smart initial basis selection for linear programs. In *International Conference on Machine Learning*. PMLR, 2023.
- [37] Robert E Tarjan and Anthony E Trojanowski. Finding a maximum independent set. *SIAM Journal on Computing*, 6(3):537–546, 1977.
- [38] Douglas B West. *Introduction to Graph Theory*. Prentice Hall, 2001.
- [39] Merve Bodur, Alberto Del Pia, Santanu S. Dey, Marco Molinaro, and Sebastian Pokutta. Aggregation-based cutting-planes for packing and covering integer programs. *Mathematical Programming*, 171(1):331–359, 2018.
- [40] Gerald Gamrath, Thorsten Koch, Stephen J. Maher, Daniel Rehfeldt, and Yuji Shinano. *Structure-based primal heuristics for mixed integer programming*. Springer Japan, 2016.
- [41] Tianxing Yang, Huigen Ye, Hua Xu, and Hongyan Wang. MIPGen: Learning to Generate Scalable MIP Instances. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*, April 2024. URL <https://openreview.net/forum?id=YhwDw31DGI>. Under review at ICLR 2024.
- [42] George B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, 1963.
- [43] Michele Conforti, Gérard Cornuéjols, and Giacomo Zambelli. *Integer Programming*. Springer, 2014.
- [44] HiGHS. HiGHS - high performance software for linear optimization documentation. Available at <https://highs.dev/top>, 2020. Accessed: June 12, 2025.

Table of Contents for Appendix

A Datasets	14
B Related Work	14
C Parameter Summary	15
D Experiment Environment	16
E More Preliminaries	17
E.1 Fundamental Optimization Concepts	17
E.2 Core Solution Algorithms and Techniques	18
F Feasibility Ratio	18
G Branching nodes	20
H Structural Similarity	21
I Solving Time Gap	23
I.1 Results	24
J Hyperparameter Tuning on Gurobi	24
K Initial Basis Prediction	26
L Solver-internal features	28
L.1 Visualization	31
L.2 Validation Experiment	31
L.3 Some Insights	31
M Limitations	35
N Broader Impact	35

A Datasets

Ecole Synthesized Instances

To facilitate various experiments, including model training and evaluation across different metrics, 4 synthetic datasets of MILP instances was generated. This process utilized the **Ecole** library [28], a platform designed for machine learning research in combinatorial optimization.

- **Configuration File:** Instance generation is controlled via a YAML configuration file. This file defines:
 - **Global Settings:** A random seed for reproducibility, the base output directory, and the total number of instances to generate per problem type.
 - **Size Ranges:** Default minimum and maximum values for the number of constraints, number of variables, and density, applicable across all problem types unless overridden.
 - **Problem-Specific Settings:** Sections dedicated to each problem type allow for enabling/disabling generation and overriding global size ranges. They also accommodate problem-specific parameters, such as the `ratio` of facilities to customers in CFL or `n_items` in CA.
- **Parameter Randomization:** Parameters like the number of constraints, variables, and density are randomly sampled for each instance within the ranges specified in the configuration file. This ensures diversity in the generated dataset. The configuration allows setting both global ranges and more specific ranges per problem type. The details of synthesizing parameters are in 6.

Compatibility Testing with Public Datasets (NIPS ML4CO)

In addition to utilizing the synthetically generated dataset described above, the compatibility and robustness of the various analysis pipelines and tools developed (e.g., for feature extraction, solver log parsing, model evaluation) were verified using publicly available MILP benchmark instances. Specifically, datasets from the **ML4CO Competition** [29] were employed for testing purposes. Successfully processing and running analyses on these standard benchmarks demonstrated the broader applicability and compatibility of the developed experimental framework beyond the specific synthetically generated instances.

B Related Work

The generation of high-quality MILP instances is crucial for developing, testing, and tuning both traditional solvers and modern ML approaches for combinatorial optimization. However, the scarcity of diverse, representative real-world instances, often due to proprietary constraints or collection difficulties, presents a significant bottleneck. This has spurred research into synthetic instance generation techniques. Recently, deep learning (DL) has emerged as a promising direction for MILP instance generation, aiming to automatically learn complex data distributions and structural features from existing instances without requiring explicit, expert-designed formulations. Several DL-based frameworks have been proposed:

- **VAE-based Approaches:** G2MILP [10] introduced the first DL-based framework, using a masked Variational Autoencoder (VAE) paradigm. It iteratively corrupts and replaces constraint nodes in the bipartite graph representation. While capable of generating instances structurally similar to the training data, G2MILP does not explicitly guarantee the feasibility or boundedness of the generated instances, potentially leading to unusable samples for certain downstream tasks. DIG-MILP [11], also VAE-based, addresses the feasibility and boundedness guarantee by leveraging MILP duality theories and sampling from a space including feasible solutions for both primal and dual formats. However, this can come at the cost of structural similarity compared to the original data. ACM-MILP [12] refines the VAE approach by incorporating adaptive constraint modification. It uses probability estimation in the latent space to select instance-specific constraints for modification, preserving core characteristics. It also groups strongly related constraints via community detection for

Table 6: Parameter Settings for MILP Instance Synthesis by Problem Type

Problem Type	Parameter	Value / Range
Set Cover (SC)	Constraints Range	[200, 800]
	Variables Range	[400, 1600]
	Density Range	[0.05, 0.2]
Capacitated Facility Location (CFL)*	Ratio (Facilities/Customers)	0.5
	Constraints Range	[50, 150]
	Variables Range	[500, 5000]
	Density Range	[0.01, 0.3]
Combinatorial Auction (CA)**	Number of Items (n_items)	Auto-calculated (null)
	Value Range (min/max)	[1, 100]
	Constraints Range	[50, 200]
	Variables Range	[80, 600]
	Density Range	[0.02, 0.1]
Independent Set (IS)***	Variables (Nodes) Range	[480, 520]
	Density (Edge Prob.) Range	[0.01, 0.015]
	Constraints Range	N/A (Derived)

*For CFL, the input ‘constraints’ and ‘variables’ ranges influence the number of customers and facilities, which in turn determine the actual model size. The ‘density’ range specified might not be directly used by the ‘ecole’ generator for CFL.

**For CA, the input ‘constraints’, ‘variables’, and ‘density’ ranges influence the number of bids and potentially items (if ‘n_items’ is null), which determine the actual model size.

***For IS, the ‘variables’ range directly corresponds to the number of nodes. The ‘constraints’ range is implicitly determined by the number of nodes and the edge probability (density) and is not directly set via ‘min/max_constraints’ in the config.

collective modification, aiming to maintain constraint interrelations and improve hardness preservation.

- **Diffusion-based Approaches:** Recognizing the power of diffusion models in generation tasks, MILP-FBGen [16] proposed a diffusion-based framework. It uses a structure-preserving module to modify constraint/variable vectors and a feasibility/boundedness-constrained sampling module for the right-hand side and objective coefficients, aiming to ensure both structural similarity and feasibility/boundedness. Another diffusion-based approach focuses specifically on generating *feasible solutions* (rather than full instances) using guided diffusion. It employs contrastive learning to embed instances and solutions and uses a diffusion model conditioned on instance embeddings to generate solution embeddings, guided by constraints and objectives during sampling.
- **Structure-Focused Approaches:** Some recent works explicitly target specific structural properties. MILP-StuDio [13] focuses on preserving and manipulating *block structures* commonly found in CCMs of real-world MILPs. It decomposes instances into block units, builds a library of these units, and uses operators (reduction, mix-up, expansion) to construct new, potentially scalable instances while maintaining structural integrity and properties like hardness and feasibility. MILPGen [41] also aims at scalability, using node splitting/merging on the bipartite graph representation to decompose instances into tree structures, which are then concatenated to build larger problems.

These learning-based methods offer the potential to generate diverse and realistic MILP instances automatically, addressing the data scarcity issue. The generated instances have shown promise in downstream tasks like solver hyperparameter tuning, data augmentation for ML models predicting optimal values or guiding solvers, and potentially constructing harder benchmarks.

C Parameter Summary

See the parameters of problems in Table 7, 8, 9.

Table 7: Summary of Key Experimental Parameters (Part 1 of 3).

Category	Parameter Name	Value	Description
Feasibility Ratio			
	Gurobi time_limit	300 s	Maximum Gurobi solve time per instance for feasibility/boundedness check.
	Gurobi threads	64	Number of CPU threads used per Gurobi solve.
	parallel	true	Flag enabling parallel processing of instances.
	n_jobs	null	Number of parallel workers (default: CPU cores - 1).
Structural Similarity			
	Feature Set	11 features	Set of graph-based structural features extracted from each MILP instance (e.g., density, degree stats, clustering).
	num_workers	64	Number of parallel processes for feature extraction.
	num_samples	1000	Number of instances sampled (with replacement) from each dataset for JSD calculation.
	Comparison Metric	JSD-based	Jensen-Shannon Divergence between feature distributions, converted to similarity ($S_i = 1 - JSD_i / \log(2)$) and averaged.
Branching Nodes (Hardness Comparison)			
	Gurobi time_limit	300 s	Maximum Gurobi solve time per instance.
	Gurobi threads	1	Threads per Gurobi solve (to enhance determinism for node count comparison).
	Gurobi Heuristics	0.5	Gurobi heuristic effort parameter setting.
	Gurobi MIPFocus	1	Gurobi parameter to focus on finding feasible solutions.
	Metric	Relative Error	Relative difference in total B&B nodes: $ (\sum N_{gen} - \sum N_{train}) / \sum N_{train} \times 100\%$.
Solving Time Gap (Hardness Comparison)			
	Gurobi time_limit	100 s	Maximum Gurobi solve time per instance.
	Gurobi threads	256	Threads per Gurobi solve (high count used to expedite experiments).
	Metric	Relative Difference	Relative difference in mean solve times: $ (mean_t_{gen} - mean_t_{train}) / mean_t_{train} \times 100\%$.

D Experiment Environment

All experiments were conducted on a Linux-based system equipped with an AMD EPYC 9754 128-Core Processor with a typical operating frequency around 3.1 GHz and supports 256 threads (128 cores with 2 threads per core). The system has 256 MiB of L3 cache and 251 GiB of system RAM.

For GPU-accelerated computations, a NVIDIA GeForce RTX 3090 graphics card with 24 GiB of VRAM was utilized. The NVIDIA driver version was 535.171.04, supporting CUDA up to version 12.2. The CUDA Toolkit version used for development and compilation was 11.8 (V11.8.89).

The operating system was Ubuntu 20.04.6 LTS. Key software includes Python 3.8.20, Gurobi Optimizer version 11.0.1.

Table 8: Summary of Key Experimental Parameters (Part 2 of 3).

Category	Parameter Name	Value	Description
Hyperparameter Tuning (SMAC3 for Gurobi)			
Tuning Framework	SMAC3		Automated algorithm configuration tool used.
n_trials	50		SMAC3 evaluation budget (number of configurations tested).
random_seed	42		Seed for SMAC reproducibility.
Objective	Mean Solve Time		Minimize mean wall-clock time on the tuning instance set.
Gurobi time_limit	60 s		Gurobi time limit per solve during tuning and final evaluation.
Gurobi threads	256		Gurobi threads per solve during tuning and final evaluation.
Tuned Parameters (8)	Gurobi Controls		Heuristics, MIPFocus, VarBranch, BranchDir, Presolve, PrePasses, Cuts, Method. (Ranges/values specified in config).
Initial Basis Prediction (GNN)			
<i>GNN Model Parameters</i>			
num_layers	3		Number of GNN message passing layers.
hidden_dim	128		Dimensionality of hidden layers in the GNN.
dropout	0.1		Dropout rate used during training.
<i>GNN Training Parameters</i>			
epochs	800 (Max)		Maximum number of training epochs.
batch_size	32		Number of instances per training batch.
learning_rate	1e-3		Initial learning rate for the Adam optimizer.
weight_decay	1e-4		L2 regularization coefficient.
early_stopping	50		Patience (epochs) for early stopping.
seed	42		Random seed for training reproducibility.
Loss Function	Weighted CE		Weighted Cross-Entropy + Label Smoothing (0.1).
<i>Evaluation Parameters</i>			
Gurobi time_limit	600 s		Gurobi time limit per solve during evaluation.
Gurobi threads	256		Gurobi threads per solve during evaluation.
Gurobi mip_gap	0.01		Gurobi target MIP gap during evaluation.
Basis Repair Threshold	1e-12		Pivot threshold for basis repair stability check.
Basis Repair Iterations	50		Max iterations for basis repair algorithm.

E More Preliminaries

E.1 Fundamental Optimization Concepts

Feasibility Consider the MILP problem defined by constraints $\mathbf{Ax} \leq \mathbf{b}$ and $x_i \in \mathbb{Z}, \forall i \in \mathcal{I}$. The feasible region is the set $\mathcal{F} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{Ax} \leq \mathbf{b}, x_i \in \mathbb{Z} \forall i \in \mathcal{I}\}$. The MILP problem is *feasible* if its feasible region \mathcal{F} is non-empty, i.e., $\mathcal{F} \neq \emptyset$. A specific point $\hat{\mathbf{x}} \in \mathbb{R}^n$ is considered *feasible* if it satisfies all constraints, i.e., $\hat{\mathbf{x}} \in \mathcal{F}$.

Boundedness Given the MILP objective $\min_{\mathbf{x} \in \mathcal{F}} \tilde{\mathbf{c}}^\top \mathbf{x}$, the problem is *bounded* if the optimal objective value z^* is finite. This means $z^* = \inf_{\mathbf{x} \in \mathcal{F}} \tilde{\mathbf{c}}^\top \mathbf{x} > -\infty$. (For a maximization problem $\max_{\mathbf{x} \in \mathcal{F}} \tilde{\mathbf{c}}^\top \mathbf{x}$, boundedness would mean $z^* = \sup_{\mathbf{x} \in \mathcal{F}} \tilde{\mathbf{c}}^\top \mathbf{x} < +\infty$). If a problem is both feasible ($\mathcal{F} \neq \emptyset$) and bounded, a finite optimal objective value z_{MILP} exists.

Table 9: Summary of Key Experimental Parameters (Part 3 of 3).

Category	Parameter Name	Value	Description
Solver-Internal Features			
<i>Data Generation Phase (Gurobi Run)</i>			
Gurobi time_limit	100 s		Gurobi time limit per instance for log generation.
Gurobi threads	128		Gurobi threads per instance for log generation.
Gurobi seed	42		Gurobi seed for log generation reproducibility.
Gurobi mip_gap	1e-4		Gurobi MIP gap for log generation runs.
Callback Usage	Enabled		Gurobi callback potentially used for detailed heuristic data.
<i>Comparison Phase</i>			
Root Gap / Heuristic Metric	1-Wasserstein Dist.		Metric used to compare scalar distributions.
Cut Plane Metric	PCA + 1-Wasserstein		Method for comparing high-dim cut vectors.
PCA Components	3		Number of components retained for cut analysis.
PCA Scaling	Standard		Data scaling method (Z-score) applied before PCA.

E.2 Core Solution Algorithms and Techniques

Simplex Method Basics The *Simplex Method* [42] is an algorithm primarily designed for solving Linear Programming (LP) problems, such as the LP relaxation defined above. While it can handle various forms, it often operates by converting the problem to a standard form like $\min\{\mathbf{c}'^\top \mathbf{y} \mid \mathbf{A}'\mathbf{y} = \mathbf{b}', \mathbf{y} \geq 0\}$ (where \mathbf{y} might include original and slack/surplus variables). It explores the vertices (Basic Feasible Solutions, BFS) of the feasible polyhedron $\mathcal{P}' = \{\mathbf{y} \mid \mathbf{A}'\mathbf{y} = \mathbf{b}', \mathbf{y} \geq 0\}$. It iteratively moves between adjacent BFS by exchanging one variable in the current basis (a set of variables determining the BFS) with a non-basic variable, guided by criteria like reduced costs to ensure monotonic improvement of the objective function value, until an optimal BFS is found or unboundedness is detected.

Heuristics Within MILP solvers, *heuristics* [35] are algorithms designed to rapidly find a feasible integer solution $\hat{\mathbf{x}} \in \mathcal{F}_{MILP} = \{\mathbf{x} \mid \mathbf{Ax} \leq \mathbf{b}, x_i \in \mathbb{Z} \forall i \in \mathcal{I}\}$, whose objective value $\tilde{\mathbf{c}}^\top \hat{\mathbf{x}}$ is hopefully close to the true optimum z_{MILP} . They do not guarantee optimality. Their main purpose is to quickly establish strong primal bounds (updating the incumbent $z_{incumbent} = \min(z_{incumbent}, \tilde{\mathbf{c}}^\top \hat{\mathbf{x}})$). In branch-and-bound, these incumbents enable pruning of search nodes j where the local lower bound $z_{bound}^{(j)}$ satisfies $z_{bound}^{(j)} \geq z_{incumbent}$.

Cutting Planes (Cuts) *Cutting planes* [43] are linear inequalities, $\alpha^T \mathbf{x} \leq \beta$, that are valid for all feasible integer solutions, i.e., they hold for all $\mathbf{x} \in Conv(\mathcal{F}_{MILP})$ where $\mathcal{F}_{MILP} = \{\mathbf{x} \mid \mathbf{Ax} \leq \mathbf{b}, x_i \in \mathbb{Z} \forall i \in \mathcal{I}\}$. However, they are chosen such that they are violated by the optimal solution \mathbf{x}_{LP}^* of the current LP relaxation (i.e., $\alpha^T \mathbf{x}_{LP}^* > \beta$). Adding such cuts \mathcal{C} to the LP relaxation feasible set $\mathcal{P} = \{\mathbf{x} \mid \mathbf{Ax} \leq \mathbf{b}\}$ yields a tighter relaxation $\mathcal{P}' = \mathcal{P} \cap \{\mathbf{x} \mid \alpha^T \mathbf{x} \leq \beta \text{ for all } \alpha^T \mathbf{x} \leq \beta \in \mathcal{C}\}$. The goal is to improve the lower bound derived from the relaxation, $\min_{\mathbf{x} \in \mathcal{P}'} \tilde{\mathbf{c}}^\top \mathbf{x} \geq z_{LP}$, thereby bringing it closer to the true MILP optimum z_{MILP} and potentially reducing the search space.

F Feasibility Ratio

The primary goal of this experiment is to calculate the **Feasibility Ratio** for a given dataset of Mixed-Integer Linear Programming (MILP) instances. The Feasibility Ratio is defined as the percentage of instances within the dataset that are determined to be both **feasible** (possessing at least one valid solution satisfying all constraints) and **bounded** (having a finite optimal objective value, not tending towards infinity or negative infinity). Instances that are infeasible, unbounded, or whose status cannot be determined within the given time limit are *not* counted towards this ratio.

Methodology & Operations

1. Key parameters loaded include:

- `time_limit`: 300 seconds (Maximum solving time per instance).
- `threads`: 64 (Number of CPU threads Gurobi can use per solve).

2. **Feasibility and Boundedness Check:** The instance is considered *Feasible and Bounded* for the metric *only if* the status is GRB.OPTIMAL (2) or GRB.SUBOPTIMAL (12). It is considered *not* meeting the criteria if the status is GRB.INFEASIBLE (3), GRB.INF_OR_UNBD (4), GRB.UNBOUNDED (5), or GRB.TIME_LIMIT (9). Any other status or error during the process also results in the instance being considered not feasible/bounded.

3. Result Aggregation:

- After all instances have been checked (either sequentially or in parallel), the results (a boolean indicating feasible/bounded status for each instance) are collected.
- The total number of instances processed (`total_count`) and the number of instances deemed feasible and bounded (`feasible_count`) are tallied.
- The Feasibility Ratio is calculated as:

$$\text{Feasibility Ratio} = \left(\frac{\text{feasible_count}}{\text{total_count}} \right) \times 100$$

4. Output Generation:

- A results report file is automatically generated within this directory.
- This report contains:
 - The input instance directory path.
 - The total number of instances found.
 - The number of instances found to be feasible and bounded.
 - The calculated Feasibility Ratio (%).
 - A detailed list of each instance file name and its corresponding status ("Feasible and bounded" or "Infeasible or unbounded").
- A summary of the results (directory, counts, ratio, and output file path) is also printed to the console upon completion.

Summary of Parameters

- **Solver Time Limit per Instance:** 300 seconds
- **Solver Threads per Instance:** 64
- **Execution Mode:** Parallel
- **Number of Parallel Workers:** System CPU core count minus 1
- **Metric Definition:** Percentage of instances with Gurobi status OPTIMAL or SUBOPTIMAL within the time limit.
- **Output:** Text report and console summary.

Table 10: Feasible and Bounded Ratio for IS Instance Sets

Dataset / Model	Masked Ratio	Total Instances	Feasible & Bounded Ratio (%)
Original Training Set (raw)	N/A	1000	100.00%
G2MILP (g2milp_mis)	0.01	1000	100.00%
	0.05	1000	100.00%
	0.1	1000	93.40%
ACM-MILP (acmmilp_mis)	0.05	1000	100.00%
	0.1	1000	100.00%
	0.2	1000	100.00%

Table 11: Feasible and Bounded Ratio for Combinatorial Auction (CA) Instance Sets

Dataset / Model	Parameter (η)	Total Instances	Feasible & Bounded Ratio (%)
Original Training Set (raw)	N/A	1000	100.00%
ACM-MILP (acm-milp_ca)	0.05	1000	100.00%
	0.1	1000	100.00%
	0.2	1000	100.00%
DIG-MILP (fixed_digmilp_ca)	0.01	1000	100.00%
	0.05	1000	100.00%
	0.1	1000	100.00%

Table 12: Feasible and Bounded Ratio for Set Cover Instance Sets

Dataset / Model	Parameter (η)	Total Instances	Feasible & Bounded Ratio (%)
Original Training Set (raw/set_cover)	N/A	1000	100.00%
G2MILP (g2milp_setcover)	0.01	1000	100.00%
	0.05	1000	100.00%
	0.1	1000	100.00%

G Branching nodes

Experiment Objective:

The goal is to quantify the difference in computational hardness between two sets of MILP instances: a ‘training’ set and a ‘generated’ set. This is achieved by measuring the total number of branch-and-bound (B&B) nodes explored by the Gurobi solver for each set and calculating the relative error between these totals.

Methodology & Operations:

1. Individual Instance Solving and Node Counting:

- **Gurobi Solver Configuration:** For every instance processed, the solver operates under specific settings defined in the configuration:
 - **Time Limit:** A strict time limit of 300 seconds (`solve.time_limit`) is imposed on the solver for each instance. If the solver reaches this limit before completion, it stops, and the node count up to that point is used.
 - **Threading:** Each Gurobi solve process is restricted to using 1 CPU thread (`solve.threads`).
 - **Solver Strategy Tuning:** Additional parameters are set to potentially influence the B&B search and node count: `MIPFocus` is set to 1 (directing the solver to prioritize finding feasible solutions) and `Heuristics` is set to 0.5 (adjusting the effort spent on heuristic methods).
- **Node Count Retrieval:** Upon completion or termination (e.g., hitting the time limit) of the Gurobi process for an instance, the total number of branch-and-bound nodes explored (`Nodes`) is recorded. If an instance fails to process correctly or results in zero nodes (which is noted as potentially abnormal), its node count is typically excluded from the aggregation.

2. Total Node Aggregation:

- The recorded node counts from all successfully processed instances in the *training* set are summed together to get a single value: $\sum \text{Nodes}_{\text{training}}$.
- Likewise, the node counts from all successfully processed instances in the *generated* set are summed to get $\sum \text{Nodes}_{\text{generated}}$.

3. Relative Error Calculation:

- The final metric, representing the relative difference in the total B&B effort between the two sets, is calculated using the specified formula:

$$\text{Relative Error} = \left| \frac{\sum \text{Nodes}_{\text{generated}} - \sum \text{Nodes}_{\text{training}}}{\sum \text{Nodes}_{\text{training}}} \right| \times 100\%$$

- A check is performed to handle the case where $\sum \text{Nodes}_{\text{training}}$ is zero; in this scenario, the relative error cannot be meaningfully computed (often represented as infinity).

Key Parameters Influencing the Metric:

- Gurobi Time Limit per Instance:** 300 seconds.
- Gurobi Threads per Instance:** 1.
- Gurobi MIPFocus Parameter:** 1.
- Gurobi Heuristics Parameter:** 0.5.

Outcome:

This process yields a percentage value indicating how closely the total computational effort (measured in B&B nodes under specific solver settings and time limits) for the generated instances matches that of the training instances.

Table 13: Branching Node Statistics and Relative Error across Problem Types

Problem	Model	Mask Ratio (η)	Mean Nodes	Median Nodes	Std Dev	Max Nodes	Time Limit Hits	Relative Error (%)
IS	Baseline (IS)	N/A	14.3	1.0	49.1	1,056	0	—
	G2MILP	0.01	26.2	1.0	60.1	724	0	84.1
	G2MILP	0.05	765.7	175.0	1,941.6	16,314	5	5,271.6
	G2MILP	0.10	8,386.4	8,330.0	1,219.5	11,789	982	58,730.0
	ACM-MILP	0.05	8.8	1.0	36.8	572	0	38.3
	ACM-MILP	0.10	10.9	1.0	63.8	1,344	0	23.4
	ACM-MILP	0.20	12.5	1.0	78.0	1,977	0	12.3
CA	Baseline (CA)	N/A	18.5	1.0	48.1	514	0	—
	ACM-MILP	0.05	6,556.0	1,661.5	9,926.3	51,913	138	35,360.0
	ACM-MILP	0.10	6,504.9	1,547.0	9,962.8	54,385	120	35,084.3
	ACM-MILP	0.20	4,873.6	962.5	8,021.7	43,036	75	26,259.8
	DIG-MILP	0.01	34.1	1.0	76.2	799	0	84.4
	DIG-MILP	0.05	34.3	1.0	76.4	812	0	85.6
	DIG-MILP	0.10	37.8	1.0	83.5	902	0	104.3
SC	Baseline (SC)	N/A	46.4	1.0	298.6	8,135	0	—
	G2MILP	0.01	85.4	1.0	536.5	14,592	0	84.0
	G2MILP	0.05	79.6	1.0	430.9	10,643	0	71.6
	G2MILP	0.10	63.7	1.0	269.8	4,713	0	37.3

Note: Statistics calculated over 1000 instances per model/ratio. Baselines refer to original training sets (Total Nodes: IS=14,255; CA=18,488; SC=46,408). 'Time Limit Hits' indicates premature termination. Relative Error (RE) compares generated total nodes to baseline total nodes. η denotes mask ratio.

H Structural Similarity

Experiment Objective:

The experiment aims to quantitatively compare the structural similarity between two distinct sets of MILP instances (referred to as Set 1 and Set 2). The comparison is based on extracting structural features from a graph representation of each instance and then evaluating the divergence between the feature distributions of the two sets using Jensen-Shannon Divergence (JSD).

Methodology & Operations:

1. Feature Extraction:

- **Graph Conversion & Feature Calculation (per instance):** For each individual MILP instance file:
 - The instance is loaded and converted into a bipartite graph representation.
 - From this graph, a vector containing 11 predefined structural features is computed. These features are:
 - (a) `coef_dens`: Coefficient density
 - (b) `var_degree_mean`: Mean degree of variable nodes
 - (c) `var_degree_std`: Standard deviation of variable node degrees
 - (d) `cons_degree_mean`: Mean degree of constraint nodes
 - (e) `cons_degree_std`: Standard deviation of constraint node degrees
 - (f) `lhs_mean`: Mean of Left-Hand-Side coefficients (constraint matrix values)
 - (g) `lhs_std`: Standard deviation of LHS coefficients
 - (h) `rhs_mean`: Mean of Right-Hand-Side values (constraint bounds)
 - (i) `rhs_std`: Standard deviation of RHS values
 - (j) `clustering`: Graph clustering coefficient
 - (k) `modularity`: Graph modularity
- **Result:** This step produces two lists of feature vectors (NumPy arrays), one list corresponding to all instances in Set 1, and the other to all instances in Set 2.

2. Per-Feature Divergence & Similarity Calculation:

- Before comparing the feature distributions, a sampling step is performed. However, in our experiment, the sampling number is set to 1000, which is also the total number of instances of both the original and generated sets.
- The process iterates through each of the 11 feature dimensions (columns 0 through 10 in the sampled arrays `f1` and `f2`).
- For each feature i :
 - The script takes the 1000 values for this feature from Set 1 (`f1[:, i]`) and the 1000 sampled values from Set 2 (`f2[:, i]`).
 - The Jensen-Shannon Divergence (JSD) between these two empirical distributions is calculated (`js_div` function). This involves creating histograms (using 5 bins) for each set's samples and for the combined samples, then applying the standard JSD formula based on Kullback-Leibler divergence.
 - The calculated JSD value (JSD_i) is transformed into a similarity score S_i using the specified formula:

$$S_i = 1 - \frac{JSD_i}{\log(2)}$$

- This score S_i ranges from 0 (maximal divergence) to 1 (identical distributions).

3. Overall Similarity Calculation:

- After calculating the individual similarity scores (S_1, \dots, S_{11}) for all 11 features, the overall structural similarity score between Set 1 and Set 2 is computed.
- This overall score is the simple arithmetic mean of the 11 individual feature similarity scores:

$$\text{OverallSimilarity} = \frac{1}{11} \sum_{i=1}^{11} S_i$$

Key Computational Parameters:

- **Number of Parallel Workers for Feature Extraction:** 64 (`compute.num_workers`)
- **Sample Size for JSD Calculation:** 1000 (`compute.num_samples`) per dataset per feature.

Below should be the detailed results of all features.

Table 14: Overall structural similarity between generated Combinatorial Auction (CA) instances and the training dataset for different models and η values.

Generative Model	Masked Ratio (η)	Overall Similarity
ACM-MILP	0.05	0.889
	0.10	0.867
	0.20	0.892
DIG-MILP	0.01	0.861
	0.05	0.860
	0.10	0.879

Table 15: Overall structural similarity between generated IS instances and the reference dataset for different models and masked ratios (η).

Generative Model	Masked Ratio (η)	Overall Similarity
ACM-MILP	0.05	0.734
	0.10	0.731
	0.20	0.699
G2MILP	0.01	0.466
	0.05	0.473
	0.10	0.486

I Solving Time Gap

The purpose of this experiment is to assess the similarity in computational hardness between two sets of MILP instances (a ‘training’ set and a ‘generated’ set) specifically concerning the time taken to solve them with the Gurobi optimizer under configured settings. The metric calculates the percentage difference between the average solving times of the two sets.

Methodology & Operations

1. Individual Instance Solving and Time Measurement:

- Every MILP instance file within both the training and generated sets is individually subjected to the Gurobi optimization process.
- **Gurobi Solver Configuration:**
 - **Time Limit:** A maximum solving duration of 300 seconds (`solve.time_limit`) is applied per instance. If an instance’s solve time reaches this limit, the recorded time will reflect this cap.
 - **Threading:** The Gurobi solver is configured to utilize 256 CPU threads (`solve.threads`) for processing each instance. The configuration uses a high thread count primarily to accelerate the overall experimental runtime. While absolute solve times are influenced by the number of threads, the focus of this metric is the relative difference in average solve times between the two instance sets. It is assumed that using the same high thread count consistently for both sets allows for a meaningful comparison of their relative hardness under these specific multi-threaded conditions.

Table 16: Overall structural similarity between generated Set Cover instances and the reference dataset for the g2milp model and different η values.

Generative Model	η Value	Overall Similarity
G2MILP	0.01	0.990
	0.05	0.950
	0.10	0.921

- **Time Recording:** The actual wall-clock time consumed by the core Gurobi optimization call for each instance is precisely measured and recorded. Instances that encounter errors during solving or do not yield a valid time measurement are excluded from subsequent calculations.

2. Relative Time Gap Calculation:

- The "Solving Time Gap" is computed as the percentage difference between the two average times. The specific calculation used is:

$$\text{Relative Time Gap} = \frac{|\text{generated_mean_time} - \text{training_mean_time}|}{\max(\text{training_mean_time}, 10^{-10})} \times 100\%$$

Key Parameters Influencing the Metric

- **Gurobi Time Limit per Instance:** 300 seconds.
- **Gurobi Threads per Instance:** 256.

This procedure results in a percentage value that indicates how similar the two sets of instances are in terms of their average solving time when using the specified Gurobi configuration (time limit and threads). A lower percentage signifies greater similarity in solver-conditional hardness.

I.1 Results

We have omitted the G2MILP-generated Independent Set (IS) instances from this presentation. The rationale for this decision lies in the substantial surge in their computational complexity compared to the baseline instances, leading to a prohibitive increase in solution times by a factor of several thousands.

Table 17: Solving Time Gap Comparison. The table shows the percentage difference in average solving time between instances generated by different models (with varying mask ratios η) and the original training set instances.

Training Set (Original Avg Time)	Model	η	Avg Time (s)	Solving Time Gap (%)
<i>Set Cover (0.2644s)</i>	G2MILP	0.10	0.3002	13.54
		0.05	0.2945	11.38
		0.01	0.3237	22.43
<i>Combinatorial Auction (0.1020s)</i>	ACM-MILP	0.20	2.6449	2493.04
		0.10	3.1654	3003.33
		0.05	3.4461	3278.53
	DIG-MILP	0.10	0.1312	28.63
		0.05	0.1225	20.10
		0.01	0.1237	21.27
<i>Independent Set (0.3374s)</i>	ACM-MILP	0.20	0.1722	48.96
		0.10	0.1774	47.42
		0.05	0.1788	47.01

J Hyperparameter Tuning on Gurobi

The core goal is to automatically optimize a select set of Gurobi solver hyperparameters using the SMAC3 framework. The optimization aims to find a parameter configuration that minimizes the average wall-clock solving time for a specific collection of MILP instances designated as the "tuning set". Subsequently, the effectiveness of this optimized configuration is assessed by comparing its performance (average solve time) against Gurobi's default settings on a separate, unseen "test set" of instances, thereby evaluating the generalization capability of the tuned parameters.

Methodology & Operations

1. Inputs & Configuration:

- **Tuning Instance Set:** A collection of MILP instances used by SMAC3 to guide the optimization process.
- **Test Instance Set:** An independent collection of MILP instances used for final performance evaluation.
- **Tuning Framework:** SMAC3 (via the SMAC4HPO facade).
- **Target Solver:** Gurobi.

2. Hyperparameter Space Definition:

- The tuning process focuses on optimizing 8 key Gurobi parameters. The search space for each parameter is defined in `gurobi_train.yaml` under `param_space`:
 - `Heuristics`: Continuous float range [0.0, 1.0].
 - `MIPFocus`: Categorical integer choices {0, 1, 2, 3}.
 - `VarBranch`: Categorical integer choices {-1, 0, 1, 2, 3}.
 - `BranchDir`: Categorical integer choices {-1, 0, 1}.
 - `Presolve`: Categorical integer choices {-1, 0, 1, 2}.
 - `PrePasses`: Categorical integer choices {-1, 0, 1, ..., 20}.
 - `Cuts`: Categorical integer choices {-1, 0, 1, 2, 3}.
 - `Method`: Categorical integer choices {-1, 0, 1, 2, 3, 4, 5}.

3. Automated Tuning via SMAC3:

- **Optimization Loop:** SMAC executes an iterative Bayesian optimization process for a budget of 50 trials (`tuning.n_trials`). In each trial:
 - SMAC suggests a candidate Gurobi hyperparameter configuration (a specific set of values for the 8 parameters within their defined space).
 - **Objective Function Evaluation:** This suggested configuration is evaluated as follows:
 - * Every MILP instance within the `tuning set` (`paths.custom_instances_dir`) is solved using Gurobi configured with the candidate parameters.
 - * Each solve is subject to a strict time limit of 60 seconds (`solve.time_limit`) and utilizes 256 threads (`solve.threads`).
 - * The wall-clock time taken for each solve is recorded.
 - * The *arithmetic mean* of these solve times across all instances in the tuning set is calculated.
 - **SMAC Feedback:** This mean solve time is returned to SMAC as the performance "cost" of the evaluated configuration. SMAC uses this feedback (and results from previous trials) to update its internal models and decide which configuration to try next, aiming to minimize this mean solve time.
- **Result:** After 50 trials, SMAC reports the configuration that achieved the lowest mean solve time on the tuning set. This is designated as the "best configuration".

4. Performance Evaluation on Test Set:

- To assess how well the tuned parameters generalize, the performance of the "best configuration" found by SMAC is compared against Gurobi's default settings. This comparison is performed on the independent `test set` (`paths.test_instances_dir`).
- **Procedure:**
 - Each instance in the test set is solved once (`evaluation.repeat: 1`) using Gurobi with the **default parameter settings** (approximated for the 8 tuned parameters). The solve time is recorded, subject to the same 60s time limit and 256 thread count.
 - Each instance in the test set is then solved once using Gurobi with the "**best configuration**" found by SMAC, again subject to the 60s time limit and 256 threads. The solve time is recorded.
- **Calculation:** The average solve time across all test set instances is calculated separately for the default configuration and the "best configuration".

5. Output & Metric:

- The primary output is the comparison between the average solve time achieved by the default Gurobi settings and the SMAC-tuned "best configuration" on the test set.
- This difference is often reported as a percentage improvement of the tuned configuration over the default.
- The specific "best configuration" (the set of 8 hyperparameter values) found by SMAC is also a key output.

Key Parameters

- **SMAC Trials:** 50 (`tuning.n_trials`).
- **Gurobi Time Limit (per solve):** 60 seconds (`solve.time_limit`).
- **Gurobi Threads (per solve):** 256 (`solve.threads`).
- **Evaluation Repetitions (per instance on test set):** 1 (`evaluation.repeat`).
- **Tuning Instance Set Path:** Value of `paths.custom_instances_dir`.
- **Test Instance Set Path:** Value of `paths.test_instances_dir`.

Table 18 shows the best Gurobi hyperparameter values found by SMAC3 for each experiment. The default Gurobi parameter values are: `Heuristics=0.05`, `MIPFocus=0`, `VarBranch=-1`, `BranchDir=0`, `Presolve=-1`, `PrePasses=-1`, `Cuts=-1`, `Method=-1`.

Table 18: Best Gurobi hyperparameter values found for each experiment.

Experiment Name	Heur.	Focus	VarBr.	BrDir.	Pres.	PreP.	Cuts	Meth.
acmmilp_mis_0.1	0.189	2	1	-1	-1	10	0	5
acmmilp_mis_0.2	0.474	0	1	0	-1	17	1	2
acmmilp_mis_0.05	0.500	0	-1	-1	-1	-1	-1	-1
ca	0.186	1	1	1	0	1	-1	4
digmilp_ca_0.1	0.186	1	1	1	0	1	-1	4
digmilp_ca_0.05	0.186	1	1	1	0	1	-1	4
digmilp_ca_0.01	0.186	1	1	1	0	1	-1	4
mis	0.498	0	1	-1	1	-1	1	3
setcover	0.189	2	1	-1	-1	10	0	5
g2milp_setcover_0.01	0.189	2	1	-1	-1	10	0	5
g2milp_setcover_0.05	0.189	2	1	-1	-1	10	0	5

Column Abbreviations: Heur.: Heuristics, Focus: MIPFocus, VarBr.: VarBranch, BrDir.: BranchDir, Pres.: Presolve, PreP.: PrePasses, Cuts: Cuts, Meth.: Method.

K Initial Basis Prediction

Following the approach of [36], this experiment aims to leverage Graph Neural Networks (GNNs) to predict a high-quality initial basis for the LP relaxation of MILP instances. The primary goal is to train a GNN model, typically on a specified dataset (e.g., generated instances), and evaluate whether using the initial basis predicted by this trained model can accelerate the MILP solving process in Gurobi compared to Gurobi's default initialization. The evaluation measures performance improvements in terms of Gurobi runtime, node count, and iteration count on an unseen test set.

Methodology & Operations

The experiment follows a structured process encompassing data representation, feature engineering, GNN model training, basis generation/repair, and performance evaluation.

1. Problem Description and Data Representation

- An MILP instance is represented by its constraint matrix $A \in \mathbb{R}^{m \times n}$, right-hand side $b \in \mathbb{R}^m$, objective coefficients $c \in \mathbb{R}^n$, and variable bounds $l^x, u^x \in \mathbb{R}^n$.

- The task focuses on predicting an initial basis for the LP relaxation, defined by a set of m basic variables/slacks ($B_x \subset \{1..n\}$, $B_s \subset \{1..m\}$, $|B_x| + |B_s| = m$) such that the corresponding basis matrix is non-singular.
- Each MILP instance is transformed into a bipartite graph $G = (V, E)$, where V consists of n variable nodes and m constraint nodes, and an edge $(v_i, w_j) \in E$ exists if $A_{ji} \neq 0$, with the edge weight being A_{ji} .

2. Feature Engineering

- **Variable Node v_i Features:**
 1. Objective coefficient c_i .
 2. Variable density: $\text{nnz}(A_{:,i})/m$.
 3. Similarity to slack lower bounds: $\langle A_{:,i}, l^s \rangle / (\|A_{:,i}\| \|l^s\|)$.
 4. Similarity to slack upper bounds: $\langle A_{:,i}, u^s \rangle / (\|A_{:,i}\| \|u^s\|)$.
 5. Variable lower bound l_i^x (0 if infinite).
 6. Variable lower bound flag (-1 if $-\infty$, 0 if finite).
 7. Variable upper bound u_i^x (0 if infinite).
 8. Variable upper bound flag (1 if $+\infty$, 0 if finite).
- **Constraint Node w_j Features:**
 1. Similarity to objective coefficients: $\langle A_{:,j}, c \rangle / (\|A_{:,j}\| \|c\|)$.
 2. Constraint density: $\text{nnz}(A_{:,j})/n$.
 3. Similarity to variable lower bounds: $\langle A_{:,j}, l^x \rangle / (\|A_{:,j}\| \|l^x\|)$.
 4. Similarity to variable upper bounds: $\langle A_{:,j}, u^x \rangle / (\|A_{:,j}\| \|u^x\|)$.
 5. Constraint (slack) lower bound l_j^s (0 if infinite).
 6. Constraint lower bound flag (-1 if $-\infty$, 0 if finite).
 7. Constraint (slack) upper bound u_j^s (0 if infinite).
 8. Constraint upper bound flag (1 if $+\infty$, 0 if finite).
- **Normalization:** Continuous features (excluding flags) are standardized (z-score normalization) before being input to the model.

3. Graph Neural Network Model

- **Architecture:** An `InitialBasisGNN` model with a bipartite graph structure is used. It consists of:
 - Initial MLP layers projecting input features (8 dims) to a hidden dimension (128, `model.hidden_dim`).
 - $L = 3$ (`model.num_layers`) `BipartiteMessagePassing` layers that iteratively update variable and constraint node embeddings. Residual connections are used.
 - Separate output MLPs predicting 3-dimensional logits for variables and constraints.
 - Dropout (0.1, `model.dropout`) is applied for regularization.
- **Output and Knowledge Masking:**
 - The output logits correspond to the three basis states: [NonbasicAtLower, Basic, NonbasicAtUpper].
 - Knowledge Masks are applied to the logits to ensure the model does not predict physically impossible basis statuses.
 - Softmax is applied to the masked logits to obtain probability distributions p_{x_i}, p_{s_j} over the three states for each variable and slack.

4. Model Training

- **Dataset:** A single GNN model is trained using instances from a specified training directory (`datasets.train_dir`).
- **Training Loop:**

- The `GNNTrainer` orchestrates the training for a maximum of 800 epochs (`training.epochs`) using batches of size 32 (`training.batch_size`).
- **Loss Function:** A `WeightedCELoss` is employed, incorporating:
 - * Label smoothing (0.1) to improve generalization.
 - * Class weights, computed per batch, to mitigate class imbalance.
- **Optimization:** The Adam optimizer is used with an initial learning rate of 1e-3 (`training.learning_rate`) and L2 regularization (weight decay 1e-4, `training.weight_decay`).
- **Early Stopping:** Training terminates early if the validation loss does not improve for 50 consecutive epochs (`training.early_stopping`).
- **Output:** The model weights achieving the best performance on the validation set are saved to a file.

5. Initial Basis Generation and Repair

- **Candidate Selection:** The m variables/slacks with the highest predicted probability of being Basic are selected.
- **Basis Repair:** An iterative process ensures the basis matrix is numerically stable by performing LU decomposition and replacing unstable columns.
- **Nonbasic State Encoding:** Nonbasic variables/slacks are set to `NonbasicAtLower` or `NonbasicAtUpper` based on the GNN’s predictions.

6. Performance Evaluation

- **Setup:** The evaluation uses an independent test set and the trained GNN model.
- **Procedure:** Each test instance is solved twice using Gurobi:
 1. **Baseline:** Gurobi’s default initialization.
 2. **Predicted:** Gurobi is initialized using the repaired basis predicted by the GNN.
- **Metrics:** Performance improvements in runtime, node count, and iteration count are calculated.
- **Output:** Results are aggregated and saved to a JSON file, with optional visualizations.

Key Parameters

- **GNN Architecture:** 3 layers, 128 hidden dim, 0.1 dropout.
- **Training:** Max 800 epochs, Batch 32, LR 1e-3, Adam, Weighted CE Loss, Early Stopping 50, Seed 42.
- **Evaluation Gurobi:** 600s Time Limit, 256 Threads, 0.01 MIP Gap.
- **Basis Repair:** 1e-12 Pivot Threshold, 50 Max Iterations.
- **Datasets:** Training (`datasets.train_dir`), Test (`datasets.test_dir`).

L Solver-internal features

The Experiment

The experiment consists of two main phases:

Phase 1: Data Generation - Solving Instances and Extracting Metrics

1. **Objective:** To solve each MILP instance in both the original and generated sets using Gurobi under identical, fixed settings, and to record detailed information about the solver’s internal operations from the execution logs.
2. **Instance Processing (per instance in a set):**

Table 19: GNN prediction initial basis: Runtime and performance improvement

Dataset	Source/Model	η	Default Time (s)	Best Time (s)	Improvement (%)
CA	ACM-MILP	0.05	0.078	0.077	2.6%
		0.10	0.079	0.076	2.6%
		0.20	0.078	0.075	4.0%
	DIG-MILP	0.05	0.079	0.076	3.5%
		0.10	0.080	0.076	4.1%
	Original Dataset	—	0.078	0.077	2.5%
IS	ACM-MILP	0.10	0.236	0.239	-0.2%
	G2MILP	0.05	0.237	0.240	-0.6%
	Original Dataset	—	0.236	0.237	0.2%
SC	G2MILP	0.01	0.236	0.221	10.6%
		0.05	0.240	0.229	9.1%
		0.10	0.242	0.224	12.4%
	Original Dataset	—	0.236	0.226	9.1%

- **Fixed Solver Settings:** The Gurobi solve is performed with consistent parameters specified in `solver_info.yaml` to ensure comparability:
 - Time Limit (`solve.time_limit`): 100 seconds.
 - Threads (`solve.threads`): 128.
 - MIP Gap Tolerance (`solve.mip_gap`): 1e-4.
 - Random Seed (`solve.seed`): 42 (for reproducibility).
 - Presolve (`gurobi.presolve`): -1 (Automatic).
 - MIPFocus (`gurobi.mip_focus`): 0 (Balanced).
 - Heuristics (`gurobi.heuristics`): 0.5 (Default).
 - Cuts (`gurobi.cuts`): -1 (Automatic).
 - Logging (`gurobi.output_flag=1, gurobi.log_to_console=0`): Enabled to generate detailed log files for each instance solve, but console output is suppressed. Additional logging parameters (`log_file_append`, `diagnostics`, `display_interval`) might be used for more detailed logs.
- **Heuristic Data Collection:** The configuration (`gurobi.use_callback: true`) indicates that a Gurobi callback (`HeuristicCallback`, code in `metrics.gurobi_callbacks`) might be used during the solve. This callback aims to capture more detailed information about heuristic successes than might be available solely through standard log parsing.

3. Log Parsing & Metric Extraction:

- After Gurobi finishes processing an instance (reaches optimality, time limit, etc.), its generated log file is parsed.
- The following key metrics are extracted:
 - **Root Node Gap:** The relative duality gap calculated after solving the initial LP relaxation at the root node. Extracted by parsing specific lines in the log (e.g., `Root relaxation: objective ...` and `Best objective ...`).
 - **Heuristic Success Count:** Tallies the occurrences of various heuristic methods finding improved integer solutions. This uses regular expressions to find keywords (e.g., RINS, Feasibility Pump, Diving, FoundHeuristic, Improved solution) in the log. If the Gurobi callback was used successfully (`heuristic_data_source="callback"`), the counts collected by the callback might be prioritized; otherwise, counts derived from log parsing are used (`heuristic_data_source="log_parsing"`). Counts are stored per heuristic type (e.g., `heur_RINS`, `heur_FoundHeuristic`).
 - **Cut Plane Usage:** Parses the `Cutting planes`: section of the log file to count the number of times each specific type of cut (e.g., Gomory, Cover, MIR, FlowCover,

ZeroHalf) was applied during the solve. These are stored as separate counts (e.g., `cut_Gomory`, `cut_Cover`).

- Other potentially useful information like final status, solve time, node count, and iteration count are also extracted.
4. **Output:** For each dataset (original and generated), the extracted metrics for all its instances are aggregated and saved into a structured CSV file (e.g., `solver_info.csv`).

Phase 2: Comparative Analysis

1. **Objective:** To quantify the similarity between the original and generated instance sets based on the distributions of the solver metrics collected in Phase 1.
2. **Input:** The two CSV files generated in Phase 1 (one for original data, one for generated data). Paths are specified in the respective comparison config files (e.g., `compare_rootgap.yaml`).
3. **Comparison Methodology (Metric-Specific):**
 - **Root Node Gap** (`compare_solver_rootgap.py`):
 - The `root_gap` values are extracted from both CSV files.
 - The empirical distributions of these scalar values are compared using the **1-Wasserstein distance**. A smaller distance indicates more similar distributions.
 - Statistical summaries and visualizations (e.g., histograms, KDE plots, CDFs, box plots) are generated.
 - **Heuristic Success Count** (`compare_solver_heur.py`):
 - The total count of success of heuristic (`heur_FoundHeuristic`) is selected for comparison.
 - The distributions of counts for this chosen heuristic are compared between the two datasets using the **1-Wasserstein distance**.
 - Similar statistical summaries and visualizations are generated. (This process can be repeated for different heuristic types by changing the config. However, Gurobi may not provide detailed information about which specific heuristics were used. We recommend considering solvers that offer more detailed heuristic information for future experiments.)
 - **Cut Plane Usage** (`compare_solver_cutplane.py`):
 - The vectors of cut counts are extracted for each instance from both CSV files.
 - **Preprocessing:**
 - * *Normalization*: For each instance, the raw cut counts are converted into proportions (count of cut type / total cuts for that instance).
 - * *Scaling*: These proportions are then standardized (using `StandardScaler` by default, as per `compare_cutplane.yaml`) across both datasets combined to give features zero mean and unit variance.
 - **Dimensionality Reduction (PCA)**: Principal Component Analysis is applied to the combined, scaled proportion data. The top 3 principal components (`pca.n_components`: 3) capturing the most variance are retained.
 - **Comparison**: The distributions of the instance scores along each of these top 3 principal components are compared between the two datasets using the **1-Wasserstein distance**.
 - Statistical summaries, PCA variance explained, and visualizations (e.g., PCA scatter plots, score distributions, loading plots) are generated. The main focus is on the first principal component (PC1).
4. **Output:** The primary outputs are the Wasserstein distance values comparing the distributions for Root Node Gap, the selected Heuristic Success Count, and the dominant principal components of Cut Plane Usage. These distances serve as quantitative measures of similarity between the original and generated datasets in terms of these solver behaviors.

Key Parameters:

- **Gurobi Settings (Phase 1):** Time Limit=100s, Threads=128, Seed=42, MIPGap=1e-4, plus specific Presolve/MIPFocus/Heuristics/Cuts settings.

- **Comparison Metric:** 1-Wasserstein Distance.
- **Cut Plane Analysis:** PCA with 3 components, Standard Scaling.
- **Heuristic Analysis:** Comparison performed on a specific heuristic count column (e.g., `heur_FoundHeuristic`).

This detailed two-phase experiment first gathers data on how Gurobi interacts with each instance under fixed conditions, and then uses statistical techniques (e.g., Wasserstein distance, PCA) to compare the overall behavioral profiles of the original and generated instance sets.

Table 20: Root Node Gap Statistics Comparison (Dataset Halves)

Statistic	part1	part2
count	500	500
mean	3.4637	3.4194
std	1.4386	1.4514
min	0.9722	1.0417
25%	2.2222	2.4094
50%	3.4483	3.2076
75%	4.6289	4.5045
max	6.1364	6.5909
Wasserstein_Distance	0.1884	0.1884

Table 21: Heuristic Success Count Statistics Comparison (Dataset Halves)

Statistic	part1	part2
count	500	500
mean	6.0380	6.0400
std	1.7879	1.9541
min	2.0000	2.0000
25%	5.0000	4.0000
50%	6.0000	6.0000
75%	7.0000	7.0000
max	11.0000	11.0000
Wasserstein_Distance	0.1300	0.1300

Table 22: Cut Plane Usage PCA 1-Wasserstein Distances (Dataset Halves)

Principal_Component	Wasserstein_Distance
PC1	0.2318
PC2	0.3295
PC3	0.2280

L.1 Visualization

L.2 Validation Experiment

More validation experiments for solver-internal features are in the supplementary files.

L.3 Some Insights

We noted a stark divergence in the performance of the same G2MILP model on IS and SC (Set Covering) problems. For SC problems, G2MILP largely replicated the performance characteristics of the original instance set, whereas for IS problems, the difficulty of the generated instances increased dramatically. This could be attributed to the following reasons:

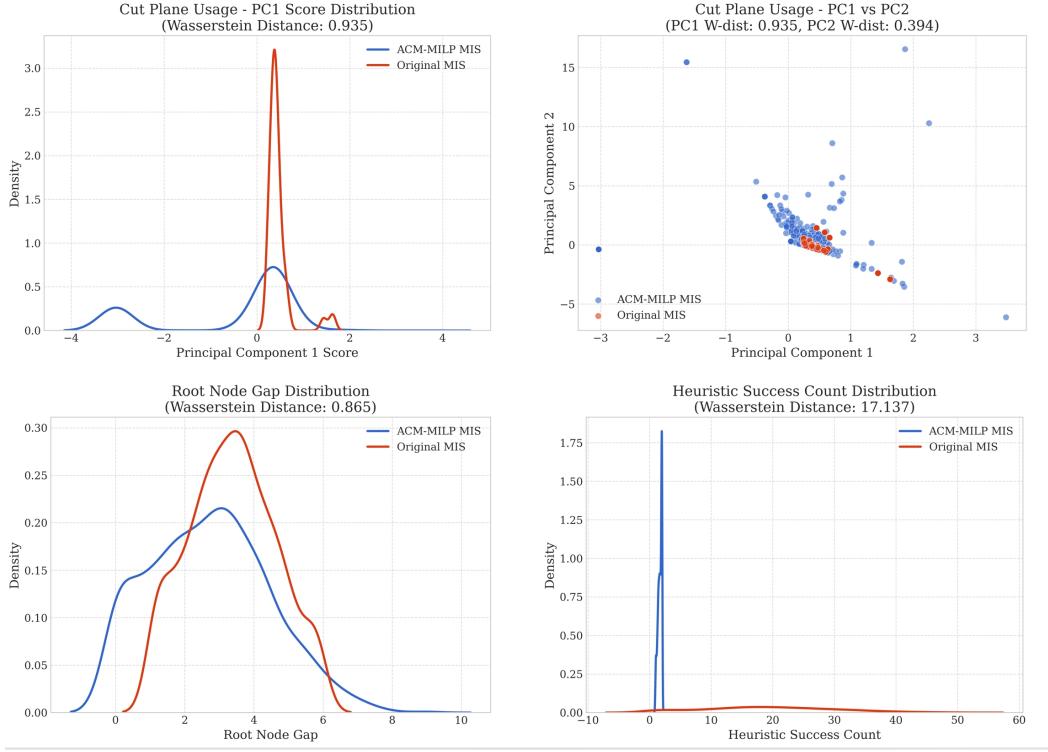


Figure 3: ACM-MILP IS at $\eta = 0.1$ compared with Original Dataset

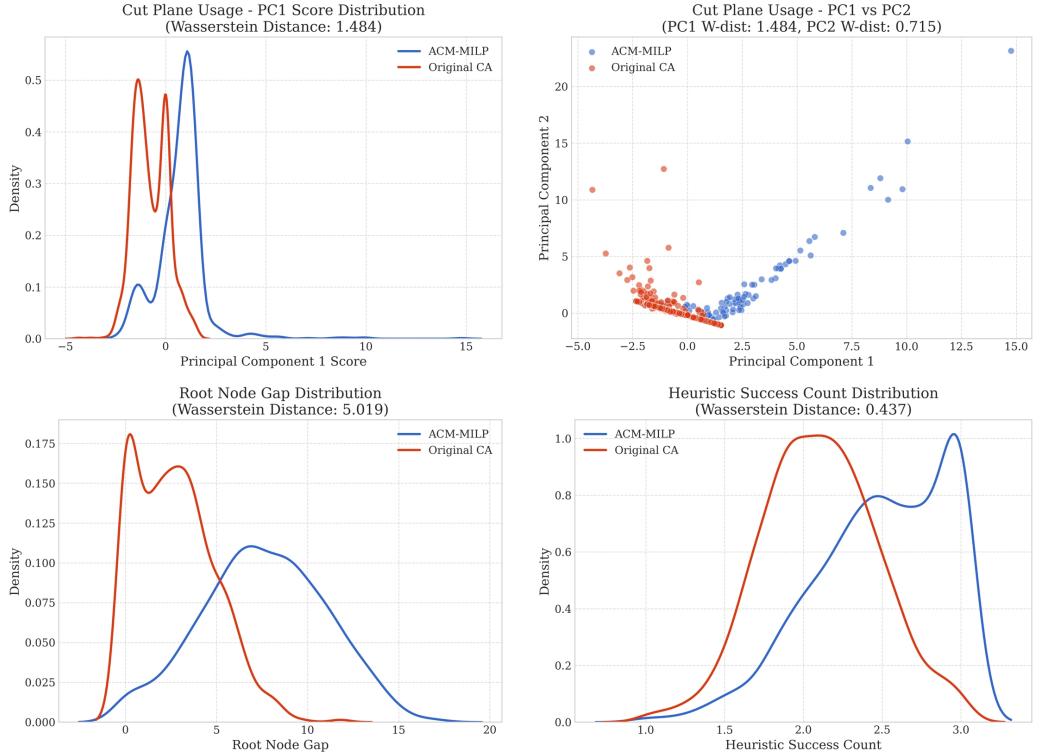


Figure 4: ACM-MILP CA at $\eta = 0.1$ compared with Original Dataset

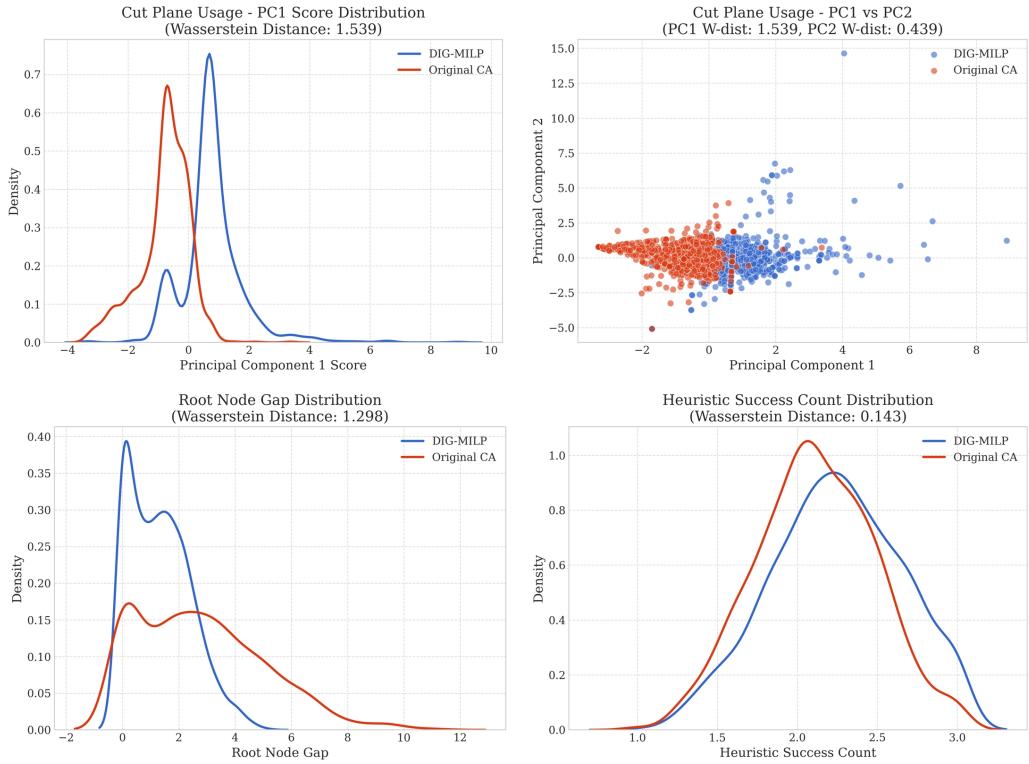


Figure 5: DIG-MILP CA at $\eta = 0.1$ compared with Original Dataset

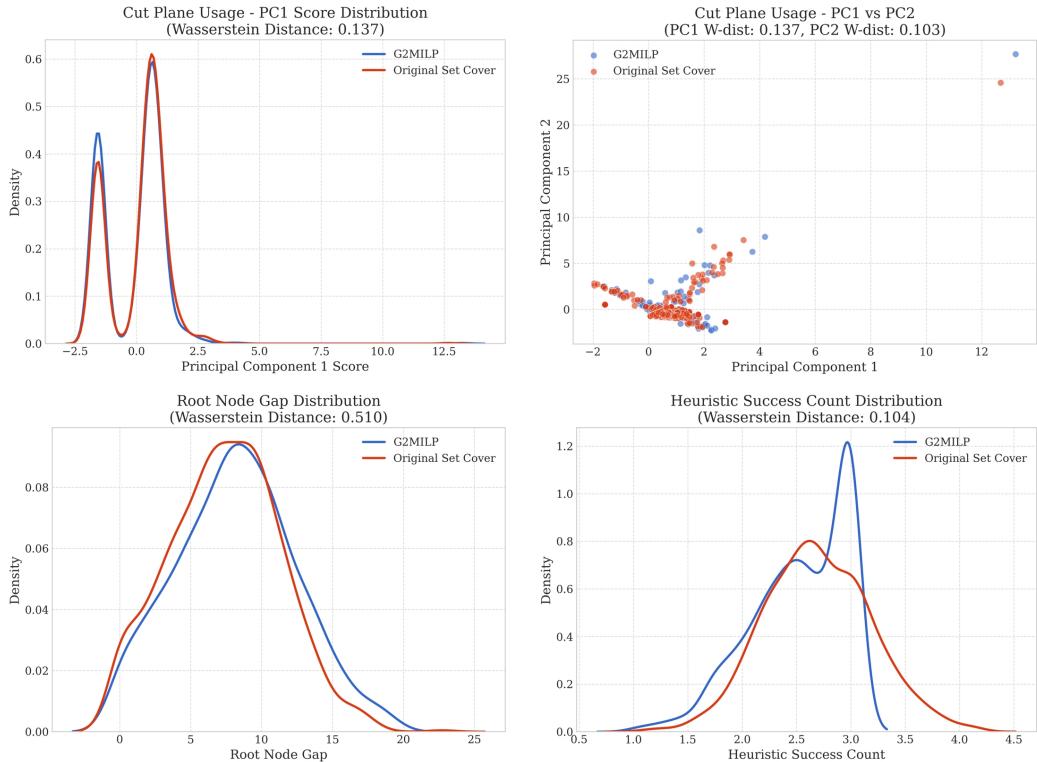


Figure 6: G2MILP SC at $\eta = 0.1$ compared with Original Dataset

Table 23: Overall Gurobi Solver Metrics Across Datasets

Dataset	Avg. Root Gap (%)	Heuristic Successes
IS (Raw)	3.42	18883
Combinatorial Auction (Raw)	2.47	2115
Set Cover (Raw)	7.61	2702
ACM-MILP CA ($\eta = 0.1$)	7.49	2552
ACM-MILP IS ($\eta = 0.1$)	2.66	1746
DIG-MILP CA ($\eta = 0.05$)	1.09	2241
G2MILP SetCover ($\eta = 0.05$)	7.75	2600

Table 24: Aggregated Cut Plane Counts (Part 1: Sum over 1000 instances)

Dataset	Gomory	ZeroHalf	Clique	MIR	RLT	FlowCover
IS (Raw)	1305	61590	89	58	43763	0
Combinatorial Auction (Raw)	11183	7014	8941	98	81	0
Set Cover (Raw)	1350	3473	80	5601	240	0
ACM-MILP CA ($\eta = 0.1$)	6524	3132	26779	25	2	229
ACM-MILP IS ($\eta = 0.1$)	1087	34032	33	13	22932	0
DIG-MILP CA ($\eta = 0.05$)	6024	5629	462	531	114	0
G2MILP SetCover ($\eta = 0.05$)	1330	3228	31	5454	209	0

For the IS problem, the objective is to find the largest subset of nodes in a graph such that no two nodes within the subset are adjacent [37, 38]. In the MILP formulation, a binary variable x_i is typically assigned to each node i (indicating whether node i is in the independent set), and a constraint $x_i + x_j \leq 1$ is added for each edge (i, j) in the graph. In G2MILP’s bipartite graph representation, “constraint nodes” correspond to these edge constraints. Thus, when G2MILP modifies a “constraint node”, it is altering a constraint associated with a specific edge in the graph. The “structure and hardness” of IS problems are predominantly derived from the global topological properties of the original graph, rather than merely the presence or absence of individual edge constraints or their coefficients (which are typically 1). G2MILP, by modifying individual or a few edge constraints, may struggle to effectively learn or preserve these critical global graph attributes. It might only be performing local “poking” or “connecting” operations, the global structural impact of which can be drastic and unpredictable. The paper mentions that G2MILP “iteratively corrupts and replaces parts of the original graphs”. For a problem like IS, which is sensitive to global structure, such local, iterative replacements might more readily lead to structural deviations.

For the SC problem, constraints directly define the covering requirements [20]. Modifications to constraints by G2MILP directly manipulate the core semantic units of the problem. The “structure and hardness” of SC instances in the training data are primarily manifested in the configuration of these constraints (i.e., which constraints exist and which variables they involve) [20]. Therefore, G2MILP’s Variational Autoencoder (VAE) has a higher probability of learning effective patterns within these configurations.

This further underscores that structural similarity, in itself, is not conclusively informative and proves useful only for certain problems. This motivates the need for our proposed, more general evaluation metric.

Table 25: Aggregated Cut Plane Counts (Part 2: Sum over 1000 instances)

Dataset	Cover	ModK	RelaxLift	InfProof	StrongCG	ImplBound
IS (Raw)	0	0	0	0	0	0
Combinatorial Auction (Raw)	3	24	0	0	6	0
Set Cover (Raw)	0	8	0	0	0	1
ACM-MILP CA ($\eta = 0.1$)	86	3	4	3	0	0
ACM-MILP IS ($\eta = 0.1$)	0	1	0	0	0	0
DIG-MILP CA ($\eta = 0.05$)	14	34	0	1	14	0
G2MILP SetCover ($\eta = 0.05$)	0	8	0	0	0	0

M Limitations

While we advocate for the EVA-MILP framework and its emphasis on solver-internal features as a significant advancement, potential limitations warrant acknowledgment.

- **Metric Selection and Evolution:** The initial selection of 'useful metrics,' though comprehensive, may necessitate ongoing refinement. Future work should consider incorporating emerging or problem-specific indicators to ensure the framework remains current and to avoid potential inherent biases in the chosen metrics.
- **Solver-Specific Feature Dependency:** Some internal features are inherently dependent on the specific solver used for evaluation. This requires careful consideration when generalizing findings across different solvers or when comparing instances evaluated with disparate solver technologies. The characteristics observed might, in part, reflect the solver's architecture rather than solely the instance's intrinsic properties.
- **Information Granularity from Solvers:** A challenge encountered during our work was the level of detail disclosed by some commercial solvers. For instance, when logging the behavior of Gurobi, specific details about the types of heuristics employed were not always available. This lack of transparency can hinder a deeper analysis of instance properties and solver interactions. To mitigate this, we suggest that future research could benefit from utilizing open-source solvers like SCIP [34] or HiGHS [44], which may offer greater access to, and control over, internal algorithmic information. The more comprehensive the information provided by a solver, the more effectively researchers can analyze instance features and hardness.
- **Interpretation Expertise:** Although EVA-MILP aims to simplify and standardize the evaluation process, comprehensively interpreting the multi-faceted data generated (spanning mathematical validity, structural similarity, computational hardness, and solver behavior) may still demand considerable expertise in optimization and solver mechanisms. Developing tools or guidelines for more automated or guided interpretation remains an area for future development."

N Broader Impact

The introduction of the EVA-MILP framework has several potential broader impacts on the fields of mathematical optimization and machine learning:

Positive Impacts:

- **Advancing Research in Combinatorial Optimization:** By providing a standardized and comprehensive methodology for evaluating MILP instance generation, EVA-MILP can foster more rigorous and comparable research. This can accelerate the development of higher-quality synthetic instances, which are crucial for testing, benchmarking, and improving MILP solvers.
- **Enhancing Machine Learning for Optimization:** The availability of diverse, well-characterized, and challenging MILP instances is vital for training and validating machine learning models aimed at improving optimization algorithms (e.g., for tasks like branching, node selection, or parameter tuning). EVA-MILP can help ensure that the synthetic data used for these purposes more accurately reflects the complexities of real-world problems, leading to more robust and effective ML-driven optimization techniques.
- **Development of More Realistic Benchmarks:** The framework encourages a deeper understanding of instance features beyond superficial structural similarity, pushing for the creation of synthetic benchmarks that better capture the computational hardness and nuanced characteristics of operational problems. This can lead to solvers that are better equipped to handle real-world challenges.
- **Facilitating Fairer Comparisons:** EVA-MILP offers a more level playing field for comparing different instance generation techniques, moving beyond limited or inconsistent evaluation practices. This transparency can guide researchers and practitioners in selecting or developing generators best suited for their specific needs.

- **Educational Tool:** The framework and its associated metrics can serve as an educational resource for students and researchers new to MILP, providing insights into what constitutes a 'good' or 'hard' instance and how various features influence solver performance.

Potential Negative Impacts and Mitigations:

- **Over-Emphasis on Standardized Metrics:** While standardization is a goal, there's a potential risk that an over-emphasis on the specific metrics within EVA-MILP could inadvertently narrow the focus of instance generation research, discouraging exploration of novel instance characteristics not yet captured by the framework.
 - **Mitigation:** The EVA-MILP framework is designed to be extensible, and we encourage the community to propose and integrate new metrics as the field evolves.
- **Increased Computational Burden for Evaluation:** thorough evaluation using EVA-MILP can be computationally intensive. This might create a barrier for researchers with limited computational resources.
 - **Mitigation:** Future work could explore methods for more lightweight yet informative evaluation protocols, or develop shared platforms/datasets that reduce the individual burden of running extensive evaluations.

Overall, the EVA-MILP framework is intended to be a positive contribution, aiming to improve the quality, diversity, and understanding of MILP instances used in research and development. Its primary societal impact is expected to be the indirect advancement of optimization technology, leading to better solutions for complex decision-making problems across various sectors."