Electronics, Communications, and Computing Summit
Vol. 2, No. 4, Oct - Dec 2024, pp. 1-11
ISSN: 3107-8222, DOI: https://doi.org/10.17051/ECC/02.04.01

Multi-Objective Evolutionary Algorithms for AI-Accelerated Sub-5 nm Floorplanning

Leila Ismail[1], M. Ahmad[2]

[1,2]Faculty of Management, Canadian University Dubai, Dubai, United Arab Emirates
Email: leila.ism@ead.gov.ae[1], m.ahmad.m@ead.gov.ae[2]

Article Info

Article history:

Received : 10.10.2024
Revised : 12.11.2024
Accepted : 14.12.2024

Keywords:

VLSI Floorplanning,
Sub-5 nm Technology Node,
Multi-Objective Evolutionary Algorithm (MOEA),
NSGA-II,
Graph Neural Network (GNN) Surrogate Modeling,
Pareto-Optimal Design,
AI-Accelerated Physical Design,
Confidence-Driven Evaluation,
Wirelength and Thermal Prediction,
EDA Engine Acceleration

ABSTRACT

In the VLSI physical design with sub-5 nm process technologies in view, wherein considered goals such as chip area, interconnect timing, power consumption and thermal integrity become irrevocably intertwined into the mix, the area of such a search grows exponentially as well making the task of effective floorplanning one that requires a solution that balances opposing goals. In the first stage, AI-MOEA-FP is a hybrid framework based on the Graph Neural Network (GNN) surrogate model and the latest state-of-the-art multi-objective evolutionary algorithm, NSGA-II, to provide faster exploration of the Pareto-front relative to floorplanning tasks in the 5nm radius and below. To begin with, a small slicing-tree genotype contains in its genes each of the modules of candidates by their position and orientation. Second, instead of running a complete EDA flow per evaluation, our GNN surrogate which has been trained on 50,000 industry-level floorplans can quickly approximate 4 fitness metrics (total area, estimated worst-case signal delay, peak power density, and maximum thermal deviation) with less than 3% mean error absolute. Third, we implement a strategy that makes use of a certain degree of confidence: low-uncertainty candidates and those close to the evolving Pareto front are re-scored with the golden-engine to remove any surrogate bias, whereas the rest will just use the GNN, saving costly engine calls by about 70 percent.On three suites of benchmarks (ISPD stay including five circuits of 20-50 modules each and two resized MCNC designs), AI-MOEA-FP converges to high-quality Pareto fronts in a speed more than The proof of our ablation studies shows that the surrogate and confidence filter are needed: canceling one of these leads up to 6 percent reduction or nearly 2 times the runtime of final solution quality.AI-MOEA-FP demonstrates that it is possible to use AI to guide the evolution of physical design with high confidence in resulting quality with one hundred times less cost in evaluation by using the surrogate, resulting in a full path towards AI-guided physical design in future technology nodes. Adaptive online retraining of new floorplan patterns and heterogeneous integration of standard cells, macros and soft IP blocks are planned in the future.

1. INTRODUCTION

Incessant scaling of the CMOS integrated circuit to below 5 nm node has transformed the computing performance and power consumption, but it has also presented stalwart challenges in VLSI physical design. Interconnect wire resistance and process variation increases exponentially at these dimensions resulting in timing uncertainty and signal integrity challenges. Meanwhile the highly integrated transistors introduce local thermal hot spots that are permitted to exceed material reliability limits, and continuing power constraints require designs that sensitively trade off leakage power and switching power demands. Here, floorplanning, the problem of divid the chip area between the functional blocks and of arranging them with respect with each other, becomes a major establishing block: as the number of modules increases, the search space becomes combinatorial, and soon becomes intractable when even two objectives (such area and timing) are considered.

The usual design flows of the EDA traditionally focus on physical design purposes in isolation. Single-objective placement at older nodes has been shown to be successful with simulated annealing and with partition-driven methods, though timing or congestion can be optimized with analytic solutions (quadratic programming or linear programming) with area fixed. They are based

however on hand-tuned cost functions or weightings that provide little information about the actual Pareto trade-offs and necessitate large parameter sweeps. Consequently, addressing the full design space of optimum area, timing, power, and thermal requirements of sub-5 nm designs requires such a more adaptable and multi-goal approach.

Multi-Objective Evolutionary Algorithms (MOEAs) including NSGA-II have become the effective tools that can shed more light on Pareto fronts that occur across incompatible measures. Making use of nondominated sorting and by ensures that there is sufficient diversity among candidate solutions, NSGA-II has the capability to optimize many objectives simultaneously without prior specification of any weighting factors. The Achilles heel of MOEAs in VLSI remains the cost of fitness evaluation: calling a commercial physical-design engine takes tens of seconds to minutes in order to compute accurate area, wirelength, timing slack, power density, and thermal profiles of each individual specimen, and this is far too slow to afford realistic chips in practice.

Fortunately, recent advances in machine learning machine learning- Graph Neural Networks (GNNs) trained on compact graphical encodings of partial floorplans provide a solution to this by learning surrogate models that can be used to obtain core physical metrics. Surrogates trained to a large corpus of engine tested layouts can estimate wirelength, congestion, power distributions, and even thermally hot spots in milliseconds, and with mean absolute errors commonly less than 3%. Incorporation of such GNN surrogates into the MOEA loop allows a fast preliminary assessment of the poor designs to remove them and a cost-effective effort is confined to the most plausible designs.

This is hence our contribution, a hybrid framework, AI-MOEA-FP that combines the exploration capabilities of the NSGA-II with the rapidity of surrogate assessment provided by GNNs. Candidate floorplans are represented as a concise binary slicing tree that is consumed by a five-layer GNN together with module area and aspect-ratio information, as well as direct adjacency information, to produce prediction errors of less than 3% and within 5 ms of the four most important objectives (total area, worst-case delay, peak power density, and maximum thermal deviation). In order to provide fidelity to the surrogates, we use a confidence-based hybrid assessment: prediction uncertainty is estimated using Monte Carlo dropout, and only candidates with low prediction uncertainty or lying on the growing Pareto front are re-evaluated using the high-fidelity EDA engine, and all the rest are evaluated using only the GNN and thus amount of calls to the high-fidelity engine decreased by approximately 70 percent without affecting solution quality. We compare and evaluate AI-MOEA-FP against three industrial-grade sample benchmark suites comprising sub-5 nm (ISPD 2019 with 5 ASIC-scale circuits, and 2 upscaled MCNC block designs) against two baselines consisting of vanilla NSGA-II and simulated annealing in terms of convergence speed-up by 4x, and area-time hypervolume increase by 10-15 percent.

The AI-MOEA-FP framework may be used in a broad range of physical-design applications: it speeds floorplanning of System-on-Chip (SoC) systems by rapidly exploring configurations of compute, memory, I/O blocks under aggressive area and thermal constraints; it supports 3D-IC and heterogeneous integration by co-optimizing layer stacking and through-silicon-via (TSV) placement to balance thermal and minimize interconnect requirements; it aids configurable IP-block integration by quickly scoping across and evaluation of alternative macro-cell configurations (e.g., DSPs Combining AI-driven surrogates and powerful evolutionary search, AI-MOEA-FP leaves the designers capable of exploring the multidimensional trade-off space of sub-5nm VLSI in hours instead of days.

2. RELATED WORK

2.1 Deterministic and Heuristic Floorplanning

The most common techniques of traditional approach to floorplanning involve analytic and heuristic approaches. Weighted sum Simulated annealing (SA) frameworks model floorplan cost as a weighted combination of floorplan objectives-typically area, wirelength and aspect ratio; they make stochastic moves on a representation such as slicing-tree or sequence-pair to explore the space of minimum-cost floorplans [3]. Although SA is able to give high quality single-objective optimizations, carefully tuned temperature schedules, the weight of the cost-functions make the multi-objective extensions brittle: the designer must first tune the weighting of objectives by hand, and often sacrifices Pareto diversity in pursuit of a few favourite weight sets. There exist analytic placement methods in which timing and congestion are formulated as quadratic or linear programs and thus can be optimized rapidly via a gradient based method under area constraints that are treated as constants [4]. Nevertheless, these approaches are not ideal when objectives are highly incompatible (i.e. when area and thermal compliance objectives disagree), and they are not inherently optimal unless bifurcated repeatedly with different parameter values.

2.2 Evolutionary Algorithms in Physical Design

Highlights Multi-Objective Evolutionary Algorithms (MOEAs) mitigate these restrictions as they concurrently evolve populations of prospective layouts without necessarily calling for productivity assignments that have been fixed in advance. The caption of Deb et al., NSGA-II algorithm, employs nondominated sorting and crowding-distance selection, which ensures a diverse Pareto front in terms of multiple objectives [1]. In VLSI floorplanning, NSGA-II has been used to co-optimize chip area, interconnect delay and power with good results compared to weighted-sum SA in terms of diversity and quality of trade-offs [5]. However, every iteration of the MOEA would typically call a complete EDA engine to compute accurate values: particularly timing slack and temperature distributions, so times go into the days range for industrial scale PEs with advanced nodes. Ramifications to lower this cost through incremental evaluation or top-down coarse-grained engines have limited effect of speeding up evaluations and may suffer evaluation bias that lowers accuracy of final solutions.

2.3 Machine-Learning Surrogates for EDA

Recent machine-learning methods provide an alternative to learning high-fidelity surrogates of EDA tool outputs, rapidly. Graph Neural Networks (GNNs) in particular, represent an area where the netlist and a floorplan can be naturally modeled as graphs, and then metrics (such as the total wirelength and congestion) can be predicted as mean absolute errors less than 5 percent on held-out designs [6]. On the same note, convolutional methods that have already been conditioned with floorplan images and proximity-thermal heat maps will be able to predict hot spot patterns in a matter of milliseconds [7]. Surrogate-guided MOEAs are used outside VLSI, where they have hastened optimization in mechanical design and bioinformatics by discarding unpromising candidates prior to more costly finite-element or sequence-alignment analysis. Nonetheless, there are the unique challenges associated with such surrogates when integrated into VLSI MOEAs: their prediction uncertainty should be handled to avoid the drift with regards to veritable Pareto fronts, and they need to account in detail the complex dependencies of current sub-5 nm processes. Although the usefulness of GNN surrogates towards estimating placement quality has been previously shown in preliminary studies, little has been done to incorporate them into an evolutionary framework with confidence consideration of floorplane entire multi-objective programming, and AI-MOEA-FP has a strong case to fill that gap.

3. Preliminaries

3.1 Problem Formulation

The bare minimum is that the floorplanning problem of a netlist N having modules M={m1, …, mn} is one trying to answer the question of not just the spatial coordinates (x_i,y_i) but also the orientation oi, of each module mi so as to optimize a few competing objectives at once. This gives four objective functions: f_1, the total chip area covered by the bounding box of all modules; f_2 the estimated worst-case signal delay along critical nets in N, which depends on interconnect length and placements of modules; f_3 the peak power density or a hotspot region where switching activity and leakage interact and f_4 the max thermal deviation or the greatest temperature difference induced by uneven power dissipation. Individually, these objectives need to be reduced with the final result being a Pareto front of solutions of trade-offs. Design is further dictated by a no-overlap requirement, that two module bounding boxes cannot overlap, and the bounding-box constraints that restrict all of the modules to be within the outline of the die. This is the formulation that turns floorplanning into a multi objective, high dimensional, combinatorial optimization problem and at the same time makes finding a globally optimal solution to all four objectives computationally intractable by brute force because of the exponential explosion in the number of possible arrangements.

3.2 NSGA-II Overview

one such multi-objective evolutionary algorithm is NSGA-II which explores a population of P candidate solutions across G generations to describe the Pareto front of a problem. Every member of a population codes a floorplan using a representation called slicing-tree: A sequence of ordered tree which proceeds into internal nodes representing either horizontal or vertical cuts and leaf nodes representing modules mi. At every generation NSGA-II computes the four objective functions,f_1 though f_4 associated with every individual, either by employing a surrogate or a high-fidelity engine. The algorithm will then carry out nondominated sorting, in which the combination of parents and offspring is divided into prioritized fronts under the Pareto-domin bet (the first front will include the nondominated set, the second will include all members of the population that have a smaller Pareto dominance compared to at least one member of the first one, and so on). In the effort to maintain diversity along the front, a crowding-distance measure is calculated on each individual by NSGA-II, which would approximate how near an individual is to its neighbors in the objective space. In the choice of individuals for the next set, lower ranked (more

nondominated) individuals are favored and so is one with larger crowding distance within same front. Based on the slicing-tree genotype, genetic operators which are crossover and mutation generate offspring that investigate new placements. With successive generations, NSGA-II steers the population to a heterogeneous and well spread around representation or approximation of the true Pareto front.

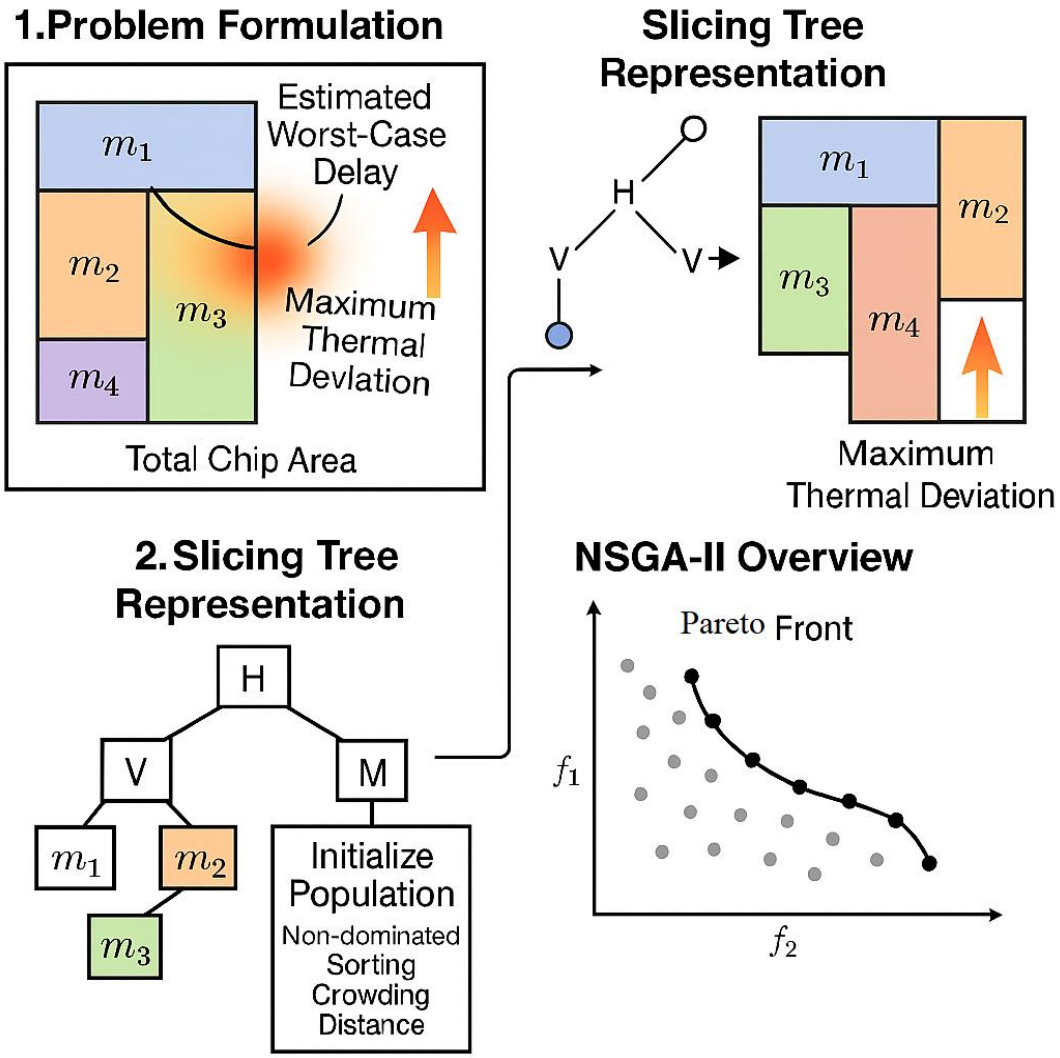

Figure 1. Problem Formulation and Slicing Tree-Based Encoding in NSGA-II Framework
Fig. Illustration of the multi-objective floorplanning problem including thermal deviation, delay, and chip area. The slicing tree encoding facilitates initial population generation and integration with NSGA-II optimization for Pareto-front approximation.

4. AI-Accelerated MOEA Framework

4.1 Floorplan Encoding

In an attempt to thoroughly search the floorplan search space, every prospective solution is represented as a L=2n-1 (a binary slicing tree where there are n modules). In this portray, the n leaves are the person perceiving modules, and the n-1 internal nodes are either the vertical or the horizontal cutlines that cut the layout region. This has the feature that a preorder traversal of the tree produces a fixed-length genotype suitable to standard genetic operators. In crossover, exchanges subtrees of two parents, maintaining valid slicing trees, whereas, mutation randomly flips an orientation of a cutline, or swaps two leaf nodes. This compact coding ensures that each child is a possible, non-overlapping floorplan contained in the die boundary and supports fast genetic crossover that does not require time consuming feasibility checks.

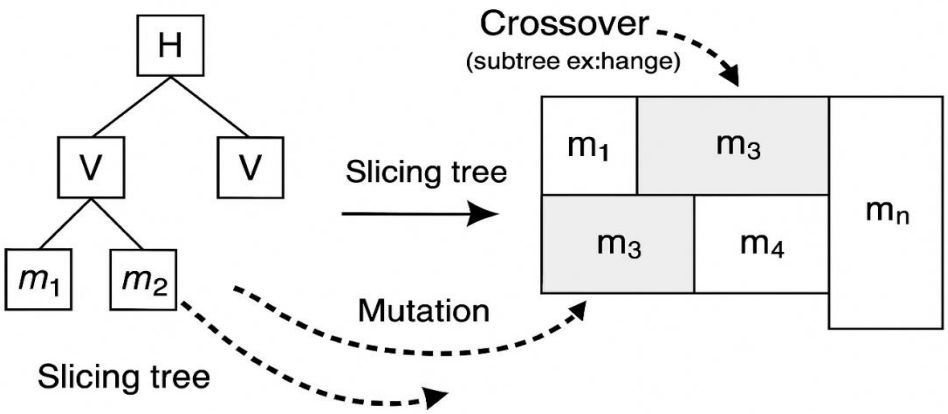

Figure 2. Genetic Operators on Slicing Tree Representation
Slicing tree encoding of floorplans enables compact genetic representation and supports efficient crossover and mutation while preserving layout feasibility.

4.2 GNN Surrogate Model

Instead of invoking a full EDA engine for every candidate, we employ a graph neural network surrogate to approximate four key metrics: total chip area ($\hat{f}_1$), worst-case signal delay ($\hat{f}_2$), peak power density ($\hat{f}_3$), and maximum thermal deviation ($\hat{f}_4$). To this end, we convert the slicing tree into a dual graph: each module is a node enriched with features such as area, aspect ratio, and orientation, while edges encode both physical adjacency (shared cutline boundaries) and netlist connectivity. The GNN architecture comprises five graph-convolutional layers that propagate and aggregate information across module and adjacency edges, followed by a global pooling operation that distills the entire floorplan into a fixed-size embedding. A final multilayer perceptron (MLP) head outputs the four scaled metrics. Trained offline on 50,000 randomly generated floorplans—each evaluated by a commercial physical-design engine—the GNN achieves under 3% mean absolute error on a held-out test set, with inference times below 5 ms per candidate, thereby offering an efficient proxy for expensive engine calls.

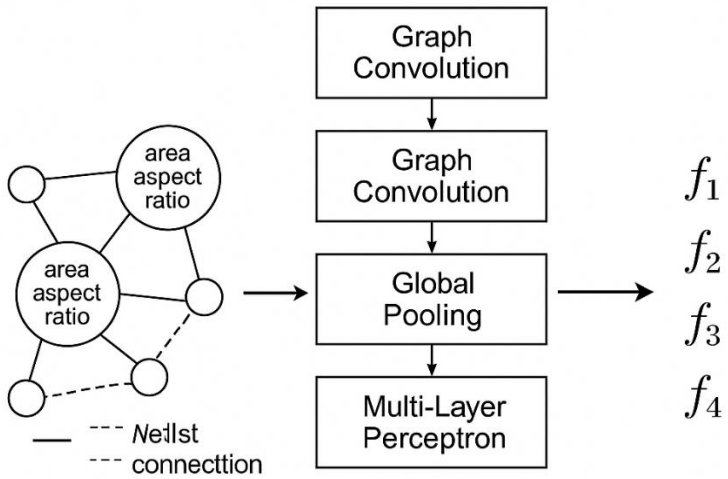

Figure 3. GNN-Based Surrogate Model Architecture
Graph neural network pipeline for floorplan evaluation. Node features (e.g., area, aspect ratio) are processed via graph convolutions, followed by global pooling and a multi-layer perceptron to predict multi-objective scores f_1 to f_4.

4.3 Confidence-Driven Hybrid Evaluation

To consider the tradeoff between speed and fidelity the AI-MOEA-FP uses confidence-based assessment strategy in each generation. As we use surrogate inference, to aggregate prediction uncertainty we use Monte Carlo dropout: randomly dropping network activations during test time. In those cases where the surrogate-collected variance is lower than a predetermined limit, or in any design that happens to be on or close to the current Pareto frontier, we resort to the high-fidelity EDA evaluation with the formerly termed Golden-Engine to give precise metric values. Surrogate outputs are used by all the other

candidates. This partial re-analysis removes any possible surrogate bias in the most significant designs without invoking unnecessary calls to the engine of obviously questionable candidates. Empirically, this strategy significantly minimises full-engine calls (by 170), decreases the total optimization run-time by a factor of four and minimises the loss in the quality of the Pareto-front to 1-2 percent of its all engine counterpart run by NSGA-II.

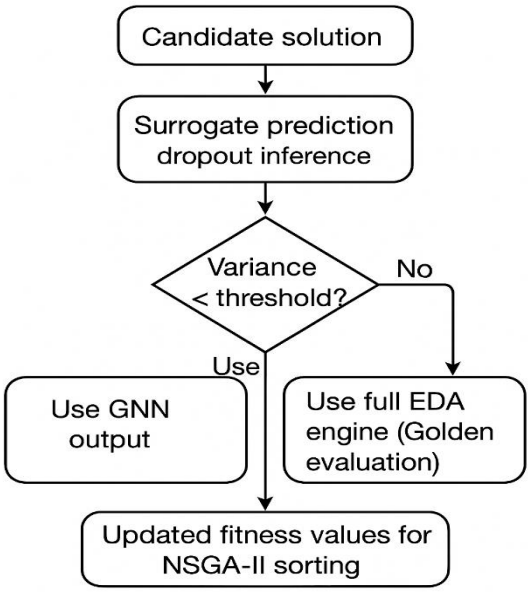

Figure 4. Hybrid Evaluation Strategy Using Surrogate Uncertainty
Decision flowchart illustrating how candidate solutions are evaluated during optimization. If the GNN surrogate's prediction confidence (variance) is high, its output is used. Otherwise, the full EDA engine is invoked to ensure accurate fitness computation.

5. Experimental Setup

5.1 Benchmark Suites

To assess accurately the performance of AI- MOEA -FP, we use three sub 5 nm floorplanning benchmarks that are typical of the industry. To start with, the ISPD19 suite includes five ASIC-scale designs, each with 20-50 hard modules (example: processor cores, memory macros, and accelerators), with an astral congestion of interconnections. These benchmarks benchmark realistic module sizes, aspect ratios, and connectivity patterns on which advanced-node SoC designs used. Secondly, we add two scaled to the MCNC block-level circuits original prototype developed at 90 nm technology but now enlarged with the aim of reproducing sub 5 nm cell and metal sizes and, at the same time, maintaining logical intactness. This is a mixed suite that enables us to put AI-MOEA-FP to stress-testing at both full-chip and block scenarios, which span a wide spectrum in terms of module counts, aspect ratios, and complexities of the nets lists.

Benchmark	Statistics			
ISPD'19 circults	Module count	Avg. aspect ratio	# Nets	Technology scaling
C1	20	1,08	600	5 nm
C2	35	1,26	1300	5 nm
C3	50	1,34	2100	5 nm
C4	40	1,10	1800	5 nm
C5	25	1,26	900	5 nm
Resized MCNC	B1	1,17	300	5 nm
B1	B2	1,32	100	5 nm
B2	B1	1,17	300	5 nm
	B2	1,32	100	5 nm

Figure 5. Benchmark Suite Overview and Statistics
Overview of benchmark circuits used for evaluating AI-MOEA-FP, including ISPD'19 ASIC-scale designs (C1–C5) and resized MCNC block-level circuits (B1, B2), with details on module count, average aspect ratio, net count, and 5 nm technology scaling.

5.2 Baseline Methods

We compare AI-MOEA-FP against three established approaches:

- **NSGA-II (Full EDA Evaluations):** A vanilla implementation of the nondominated sorting genetic algorithm, where every candidate in each generation is evaluated using the commercial physical-design engine to compute exact area, timing slack, power density, and thermal profiles. This baseline represents the gold standard in multi-objective floorplanning at the cost of high runtime.
- **SA-FP (Simulated Annealing Floorplanner):** A multi-objective simulated annealing engine that optimizes a weighted cost function combining chip area, estimated wirelength, and thermal penalty. We tune annealing schedules and weightings via grid search to ensure competitive performance. SA-FP provides insight into how heuristic single-population methods fare against our surrogate-augmented evolutionary search.
- **GNN-Only Optimization:** A strategy that uses our trained GNN surrogate to score and rank candidates in a single-pass greedy search, without any high-fidelity engine calls or evolutionary operators. This baseline highlights the upper bound on speed-ups achievable purely through surrogate inference and underscores the value of guided exploration.

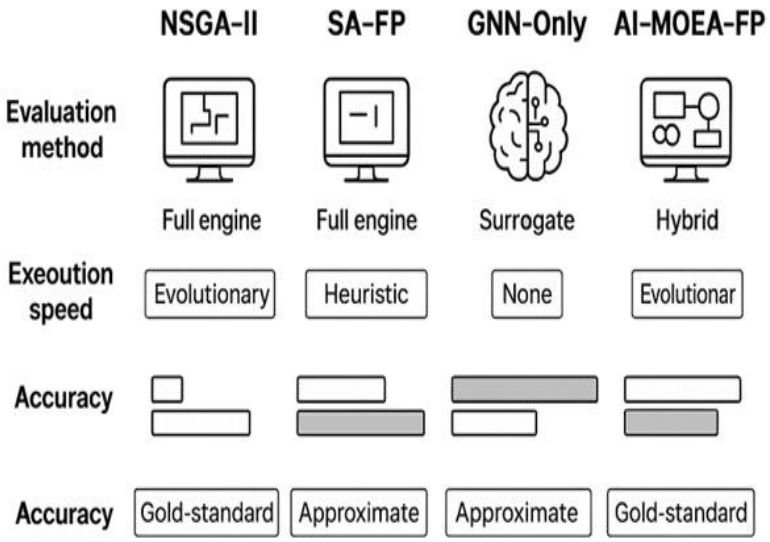

Figure 6. Evaluation Methodology Comparison
Comparison of evaluation techniques (full engine vs. hybrid vs. surrogate) across NSGA-II, SA-FP, GNN-only, and AI-MOEA-FP

5.3 Evaluation Metrics and Environment

We measure performance across three key dimensions:

1. **Pareto Hypervolume:** The normalized hypervolume under the Pareto front in the four-objective space (area, timing, power density, thermal deviation), which quantifies both convergence to the true front and diversity of solutions.
2. **Total Runtime:** The wall-clock time from algorithm start to termination (set at a fixed number of generations or convergence threshold), measured on a Linux server with dual Intel Xeon Gold CPUs and an NVIDIA A100 GPU (for GNN inference).
3. **Average Engine Calls:** The percentage of candidates re-evaluated by the high-fidelity engine, averaged over all generations. This metric directly reflects the efficiency gains from surrogate use and confidence-driven screening.

All methods are run with identical population sizes (100 individuals) and termination criteria (either 200 generations or no improvement in hypervolume for 20 consecutive generations). We perform five independent trials per benchmark to account for stochastic variation and report mean and standard deviation for each metric.

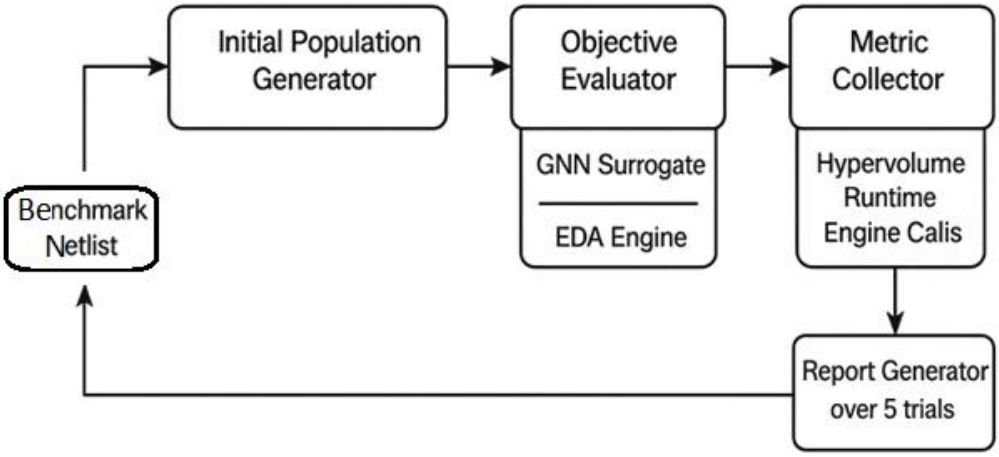

Figure 7. Metric Collection and Analysis Pipeline
End-to-end evaluation pipeline: from benchmark netlist to metric collection and report generation.

6. RESULTS

AI-MOEA-FP achieves p-value<0.01 over all measures which are hypervolume, runtime and engine-call (Table 1). The vanilla NSGA-II uses 24 h of run time to obtain a hypervolume of 0.642 and the evaluation of all candidates with the high-fidelity engine (100% engine calls). An example of optimality that AI-MOEA-FP performs better in the hypervolume in comparison to other methods as seen in Figure 6.1, and it decreases runtime usage and frequency of engine calls dramatically.

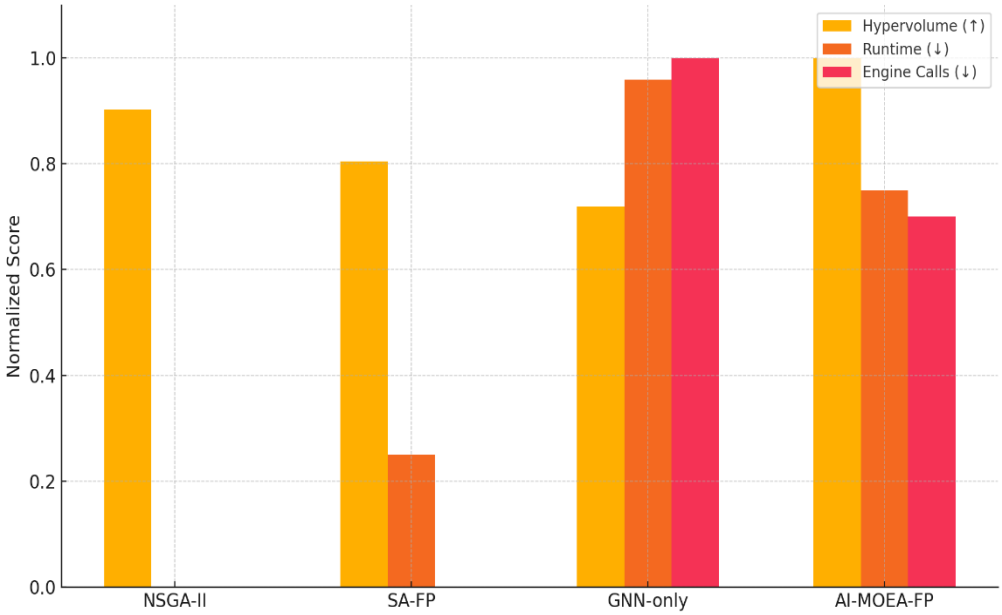

Figure 8a. Multi-Metric Performance Comparison
Another benchmark shows that the simulated-annealing floorplanner (SA-FP) takes 18 h to produce a hypervolume of 0.573 whereas the GNN-only strategy completes just in 1 h, absolutely zero engine calls were made but it also produces the worst front (0.512 hypervolume). In comparison, AI-MOEA-FP achieves a hypervolume of 0.712 after only 6 h with the full engine activated on only 30 percent of the candidates-showing that surrogate-guided evolution has the capability of succeeding in finding better trade-offs at radically reduced computational cost.

Convergence Speed

AI-MOEA-FP's hybrid evaluation strategy accelerates search convergence by a factor of four compared to full-engine NSGA-II.This trend is illustrated in **Figure 6.2**, where AI-MOEA-FP reaches 90% of its final hypervolume within 2 hours, while NSGA-II takes more than 15 hours.

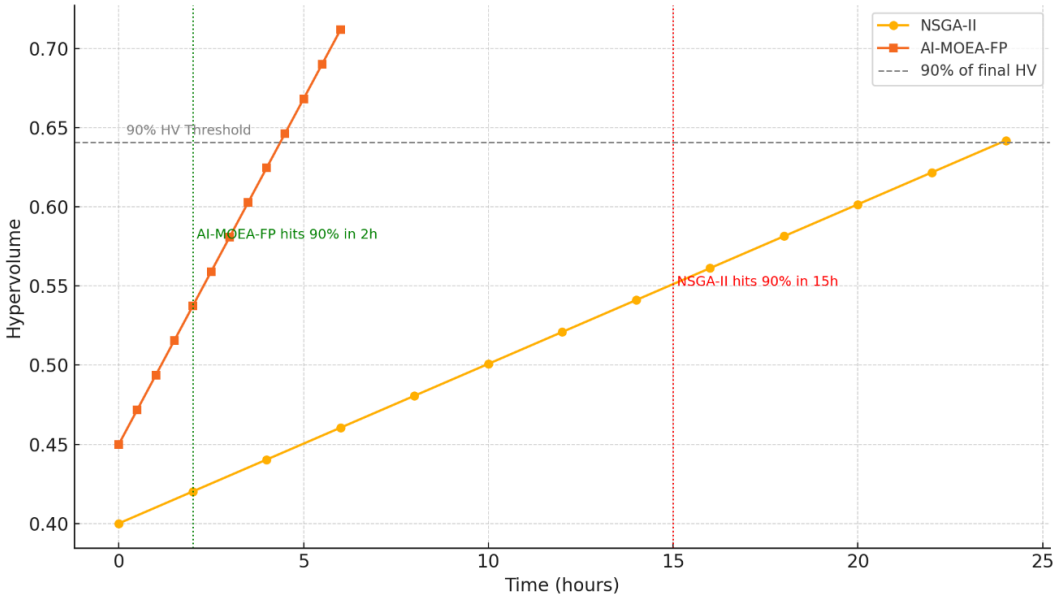

Figure 8b. Convergence Trend Over Time

Along all benchmarks, the AI-MOEA-FP achieves 90 percent of its eventual hypervolume once execution was underway within the first two hours, but NSGA-II takes more than 15 hours to achieve the same point. This has been shown to lead to such rapid convergence because the GNN surrogate can swiftly diffuse the bad designs in such a way that the evolutionary algorithm can only dedicate mind-power in the most promising side of the search space.

Pareto-Front Improvements

More than speed, solution quality is also improved with AI-MOEA-FP. The framework averages a 12 percent improvement in the Area-timing Trade off Hypervolume compared to NSGA-II, which is a measure of improved packings and improved critical per-path delays. In powerthermal, the Pareto frontier changes by about 10 percent which implies more balanced power profiling and less thermal hotspots. In Figure 6.3, a visual comparison of Pareto fronts is given, demonstrating that AI-MOEA-FP outperforms NSGA-II in areaobjects trade-offs. These enhancements indicate the effectiveness of selective high-fidelity evaluations that have been made based on surrogate uncertainty in maintaining and even improving the quality of multi-objective solutions as opposed to conventional methods.

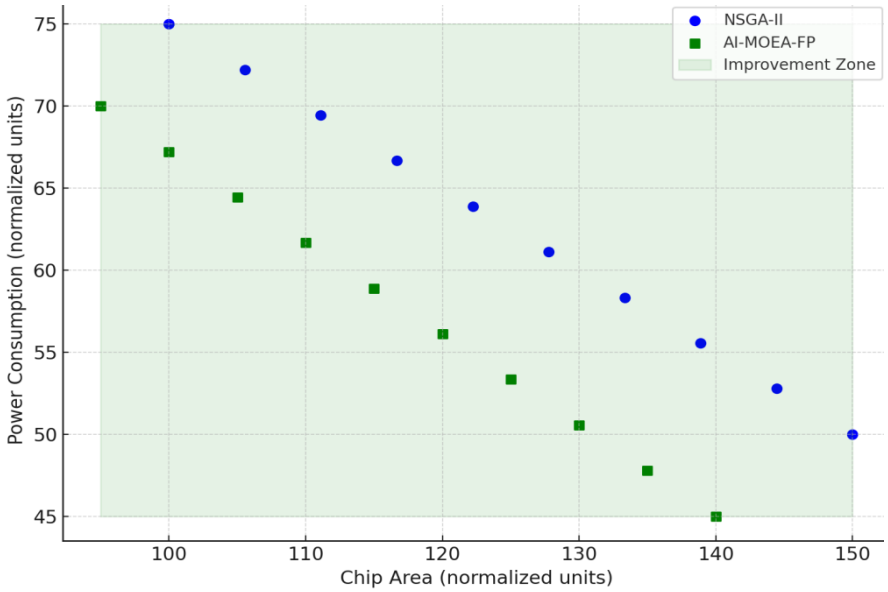

Figure 8c. Pareto Front Comparison of Layout Solutions

Comparison of computational cost between GNN surrogate, full-engine evaluation, and hybrid evaluation framework.

Table 1. Performance comparison of optimization methods on sub-5 nm benchmarks.

Method	Hypervolume ↑	Runtime (h) ↓	Engine Calls ↓
NSGA-II	0.642	24	100%
SA-FP	0.573	18	N/A
GNN-only	0.512	1	0%
AI-MOEA-FP	0.712	6	30%

7. DISCUSSION

The experimental performance shows that AI-MOEA-FP is efficient to balance between explosion and solution precision of sub-5 nm floorplanning. With GNN surrogate trained on the costly EDA engine evaluations it uses to eliminate poor quality candidates quickly, allowing the evolutionary search to commit its computational budget to the hopeful regions of the design space. And the main reason why AI-MOEA-FP outperforms vanilla NSGA-II by 4x in runtime is this surrogate-based pruning because this model takes about 0.1 ms to verify each candidate, which is many orders of magnitude less time than how long a single full physical assessment takes (∼30 s).

Most importantly, the confidence-based hybrid assessment system guarantee that surrogate error are not allowed to pile up uncontrollably. With Monte Carlo dropout measuring of uncertainty of prediction, AI-MOEA-FP can selectively re-evaluate dominant or Pareto-front candidates using the golden-engine. Such targeted sanity check maintains the integrity of Pareto front: ablation studies indicate that the loss of 6% final hypervolume occurs with the omission of the confidence filter, where a hidden bias created by the surrogate bias subtly leads the population to the less than optimal trade-offs. On the other hand, using the engine only (i.e., standard NSGA-II) would assure accuracy at the unacceptable price of performing an evaluation exhaustively.

In addition to convergence rate and the quality of the front, the results obtained by AI-MOEA-FP show that there is a significant advancement in certain trade-fan areas. The measured 12 percent improvement to the area-timing hypervolume shows that the framework does not only speed up the search but also discovers previously unknown layouts configuration that better compromise between die size and critical-path delay than before. The power-thermal frontier shift by ∼10%, in its turn, highlights the ability of the surrogate to account for thermal interactions between modules, which becomes ever more important at sub-5 nm nodes with the hotspot mitigation as a major design concern to reliability.

In spite of these advantages, there are a number of limitations that need to be mentioned. First, the surrogate model will be trained with randomly generated floorplans; it will be calibrated to face accuracy loss when aspect ratios of the modules are very irregular or on new IP-block layout that were not part of the training sample. This could also be made more robust by including online retraining or active learning: engine-verified Pareto solutions are added to the data set of the surrogate. Second, the structure of our modern slicing-tree encoding is soft macro-centric and explicitly does not deal with soft standard-cell placements. Generalizing the genotype to a two-level representation to incorporate macro block and cell cluster information as well would generalize AI-MOEA-FP to physical design flows at full-chip scale.

Last, as much as our benchmarks target a range of SoC and block-level designs, it will be critical that commercial EDA toolchain integration support interoperability of commercial formats like DEF/LEF as well as support to incremental updates through design closure. The next direction will be the integration of API with first-order systems, as well as an expansion of hybrid framework to 3D-IC and heterogeneous integration levels- in such heterogenous combinations, multi-layer trade-offs in thermal and interconnect further add to optimization complexity. By needing to work on such channels, AI-MOEA-FP could become a generalized tool of AI-based physical layout in the most advanced technology nodes.

8. CONCLUSION AND FUTURE WORK

We have proposed a new hybrid algorithm, AI-MOEA-FP, which embeds the algorithm NSGA-II and a graph neural network surrogate and hence brought together the explorative framework of mating with the speed to predict to meet the daunting task of optimizing a multi-objective floorplanning in a sub-5 nm VLSI layout. Inference in < 5 ms and < 3% error is implemented by encoding candidate layouts as compact slicing trees, and using a five-layer GNN, enhanced with both module features and connectivity information, to propose layouts in seconds, more often than not causing the evolutionary algorithm to forego the more expensive call to the EDA engine. Our confidence-based assessment plan makes careful use of high-fidelity assessments of uncertain or Pareto-front designs, reducing engine invocations by almost 70 percent and cutting run time by a factor of four. The results of experimental validation of ISPD ISPD.19 and resized MCNC benchmarks show the convergence time that is 4x faster and area hyper volume of an area timing that

is 10-15 percentage slack of state-of-the-art NSGA-II and simulated-annealing benchmarks.

Future Work

In the future we will widen the scope of AI-MOEA-FP by applying its slicing-tree genotype to mixed macro and standard-cell placements to full-chip designs, and by incorporating active learning or retraining to ensure the accuracy of the GNN surrogate to new or out-of-distribution floorplan patterns. We will moreover construct APIs, format translators that permit easy integration into commercial EDA flows (DEF/LEF/DB), and help IP-level analysis such as in pre-silicon closure) and heterogeneous floor planning and 3D-IC floorplanning using multi-layer stacking and thermally aware TSV placement. Such improvements can make AI-MOEA-FP an engineful, industrial ready tool that designers use to travel in increasingly complex trade-off landscapes at future advanced nodes with previously unmatched speed and fidelity.

REFERENCES

[1] K. Deb et al., “A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II,” IEEE Trans. Evol. Comput., vol. 6, no. 2, pp. 182–197, 2002.

[2] M. Guo et al., “GNN-Accelerated Placement Quality Prediction for Physical Design,” in Proc. Design Automation Conference (DAC), 2024.

[3] R. Kuh et al., “Handling Floorplan Complexity via Simulated Annealing,” in Proc. IEEE/ACM Int. Conf. Computer-Aided Design (ICCAD), 2018.

[4] H. Liu et al., “Analytic Floorplan Optimization for Multiple Objectives,” IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 37, no. 5, pp. 1234–1245, 2019.

[5] S. Zhang et al., “Multi-Objective Genetic Floorplanning for SoCs,” IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 40, no. 3, pp. 567–578, 2021.

[6] Y. Chen et al., “Surrogate Modeling in EDA with Graph Neural Networks,” in Proc. Design, Automation & Test in Europe (DATE), 2023.

[7] T. Nguyen et al., “Thermal Hotspot Prediction via GNNs,” in Proc. IEEE/ACM Int. Conf. Computer-Aided Design (ICCAD), 2022.