

# A Multi-Dimensional, Per-Pass Empirical Study of the LLVM Optimization Pipeline

Federico Bruzzone 

Università degli Studi di Milano  
Computer Science Department, Milan, IT  
federico.bruzzone@unimi.it

Walter Cazzola 

Università degli Studi di Milano  
Computer Science Department, Milan, IT  
cazzola@di.unimi.it

**Abstract**—Quantifying the marginal impact of individual optimization passes underpins phase ordering, pass selection, optimization design, and analysis of pass/hardware interactions. In LLVM—the standard backend for C/C++, Rust, and ML stacks via MLIR—interactions among optimization passes, measurement noise, and pipeline scale make this difficult. We present a systematic, empirical study of the LLVM -O3 optimization pipeline. We decompose the pipeline into cumulative per-pass prefixes. We then measure execution time, compile time, binary size, hardware counters, and RAPL energy across 84,750 measurements covering 113 cumulative prefixes of the -O3 pipeline evaluated on 30 PolyBench/C kernels under rigorous noise mitigation. On these compute-bound affine kernels, the pipeline is non-monotone (6.6–9.7% of transitions regress) and strongly back-loaded (the median non-regressing kernel needs 84.8% of the pipeline for 80% of its speedup). Most gains are driven by a small Pareto-dominant core of passes, while the final -O3 configuration is Pareto-dominated on (size, speedup) for 29 of 30 kernels. We further show that IR instruction count is an unreliable predictor of runtime, that runtime-targeted passes are *de facto* energy-targeted (30–60% savings), and that the search-free idealized-additive upper bound on losses due to phase interference is 46.35%. These findings enable more informed pass pruning, cost-model calibration, and autotuning.

**Index Terms**—Compiler optimizations, Optimizing Compilers, LLVM, Multi-dimensional Empirical study

## I. INTRODUCTION

Since the dawn of computing, compiler optimizations have long bridged high-level abstractions and efficient machine code [1]–[6]. Optimizing compilers [7], [8] use analysis passes to guide transformations [9]–[11] that enhance performance without altering observable behavior [12]–[14]. Modern architectures make this a delicate interplay of interdependent passes [9]. LLVM [15] exemplifies this complexity, serving as the standard backend for C/C++, Rust [16], Julia [17], and—via MLIR [18]—for *machine learning* (ML) and *deep learning* (DL) stacks [19], [20] that rely on ML/DL optimizing compilers [21]–[24] (e.g., OpenXLA<sup>1</sup>, TVM [25]) and automotive pipelines.<sup>2</sup> Quantifying individual pass impact is therefore critical across these domains.

**The Phase-Ordering Problem.** The optimal sequence of optimizations is known to be program-specific [26]–[28]. However, the *phase-ordering problem* [29]–[31]—finding the

optimal pass order, which is undecidable in general [32]–[34]—and the complexity of optimization selection [35]–[37] continue to hinder the identification of such ideal configurations. *Compiler autotuning* [38]–[44] searches this large space [45]–[47], but effectiveness depends on accurately estimating pass impact—complicated by inter-pass dependencies, noisy measurements, and scale [48]. Some transformations even degrade performance [49], [50]. Despite this, both industry practice and academic research still rely on predefined, *one-size-fits-all* pipelines (e.g., -O3) due to efficiency and limited user knowledge of the optimization process [51]–[53].

**Motivation.** Quantifying the impact of individual optimization passes is critical for compiler engineers and end users [54]–[56] as it enables: (i) more effective phase ordering, (ii) improved pass-selection heuristics, (iii) informed optimization design, and (iv) analysis of optimization-pass interactions with hardware [33]. However, existing work evaluates LLVM optimizations in isolation or focuses mainly on execution time [57], overlooking trade-offs across optimization objectives: a pass that improves runtime may increase binary size or degrade cache locality, effects rarely captured in a unified framework. Quantifying pass contributions is also important because the effectiveness of modern optimization pipelines remains under scrutiny. In particular, the debate over handwritten versus compiler-optimized kernels remains open [58].<sup>3</sup>

**Our Study.** Despite LLVM’s widespread adoption, contribution of individual optimization passes to overall optimization outcomes remains poorly understood [57]. We address this gap with a systematic empirical study of the LLVM -O3 pipeline, quantifying the effect of individual passes across multiple metrics. Beyond runtime, we measure compile time, binary size, and a broad set of hardware performance counters<sup>4</sup> following a methodology similar to that of Chen *et al.* [62]. We formulate the following research questions (RQs):

**RQ1.** *What is the marginal impact of individual optimization passes within the LLVM -O3 pipeline?*

<sup>3</sup>octoml-optimizes-apache-tvm-for-apples-m1-beats-core-ml-4-by-29

<sup>4</sup>E.g., x86 MSR counters for cache misses and cycles/instruction, plus energy measurements via running average power limit (RAPL) [59]–

<sup>1</sup>openxla.org

<sup>2</sup>Tesla Full Self-Driving v14.3: stats.tessie.com/versions/2026.2.9.6[61].

**RQ<sub>2</sub>.** To what extent do optimization passes trade off execution speed, compilation overhead, and binary footprint?

**RQ<sub>3</sub>.** How do optimization passes affect hardware-level behavior as reflected by performance counters?

**RQ<sub>4</sub>.** What fraction of achievable speedup is lost to phase interference within the pipeline, and which passes contribute most to that loss?

To support this study, we built an infrastructure that decomposes an optimization pipeline into its constituent passes and builds a sequence of cumulative prefixes, each extending the previous one by a single pass. For every prefix, we measure the marginal impact of the newly added pass relative to both the preceding prefix (delta) and the -O0 baseline. This enables both a *differential* analysis of per-pass deltas and a *cumulative* analysis of the progressive evolution of the pipeline. The infrastructure is platform-independent and configuration-driven.

**Contributions.** The paper contributions are:

- a systematic empirical characterization of the LLVM -O3 pipeline, jointly quantifying the per-pass impacts of optimization passes on *intermediate representation* (IR) transformations, runtime, microarchitectural behavior, and RAPL energy, extending prior studies focused primarily on runtime [57], [62];
- a quantitative analysis of pass interactions and optimization dynamics, revealing key performance patterns and pitfalls, introducing an idealized-additive upper bound on phase-interference loss, and showing that IR instruction count is an unreliable predictor of runtime; and
- a reusable and extensible infrastructure for automatic pipeline decomposition and multi-metric measurement across optimization pipelines.

The rest of the paper is organized as follows. §II provides background, §III describes the methodology, §IV presents the results, §V discusses their implications, §VI addresses threats to validity, §VII reviews related work, and §VIII concludes.

## II. THE LLVM OPTIMIZATION PIPELINE

LLVM is a modular SSA-based compilation framework [63]–[65]. Its `opt` tool applies sequences of analysis and transformation passes that constitute the optimization pipeline. The *new pass manager*<sup>5</sup> organizes the pipeline hierarchically through nested pass managers operating at different IR granularities: **module**, **function**, **cgsc**, **loop**, and **loop-mssa**. Passes may therefore appear multiple times within the pipeline at different levels of the hierarchy. At the top level, the **module** manager coordinates module-wide optimizations (e.g., `called-value-propagation` [66], [67]) and invokes nested managers. The remaining managers apply passes that operate respectively on (i) individual functions (e.g., `mem2reg` which promotes stack allocations to virtual registers [68], [69], and `sroa` which performs *scalar replacement of aggregates* [70], [71]), (ii) strongly connected com-

ponents of the call graph (e.g., `interprocedural inline` [72]), (iii) loops (e.g., `loop-unroll` [73], [74]), and (iv) loops in memory SSA form (e.g., `licm` which performs *loop-invariant code motion* [75], [76]). Passes execute sequentially over an evolving IR. Consequently, the effect of a pass depends on the transformations that precede it, making optimization pipelines inherently *order-dependent*, *non-commutative*, and *non-additive*.

## III. EXPERIMENTAL METHODOLOGY

We develop an infrastructure that decomposes LLVM pipelines into passes and measures their marginal impacts. The four RQs introduced in §I are operationalized as follows.

**RQ<sub>1</sub>.** We quantify the marginal impact of each optimization pass by measuring the delta of its performance relative to both the preceding pipeline state and the -O0 baseline.

**RQ<sub>2</sub>.** We quantify pass-level trade-offs across execution time, compile time, and binary size.

**RQ<sub>3</sub>.** We characterize the effect of optimization passes on hardware behavior by tracking the evolution of performance counters across the pipeline.

**RQ<sub>4</sub>.** We estimate bounded phase-interference loss  $L$  from the sequence of per-pass speedup deltas, quantifying the fraction of achievable speedup lost to pipeline interference and identifying the pass positions that contribute most to the loss.

**Evaluation Infrastructure.** An *orchestrator* coordinates pipeline construction, execution, and metric collection. Benchmarks, metrics, and experimental parameters are specified in TOML, allowing new studies without modifying the core implementation. Although pipeline-agnostic, in this study the infrastructure targets LLVM’s -O3 pipeline. It first extracts and flattens the pipeline into an ordered sequence of passes  $p_0, p_1, \dots, p_n$ , then constructs cumulative prefixes  $P_i = p_0 \rightarrow \dots \rightarrow p_i$ , preserving the original execution order and pass dependencies. Each prefix is applied independently to the same unoptimized IR  $I_0$ , producing an optimized IR  $I_i$  and executable binary  $B_i$ . Comparing consecutive binaries  $B_i$  and  $B_{i-1}$  yields the marginal impact of pass  $p_i$  within its execution context.

**Benchmark Selection.** We use the 30 kernels of PolyBench/C 4.2.1<sup>6</sup> in default mode. To keep the 95% confidence interval below 1% of the mean, benchmarks must be self-contained (single translation unit, `libm` only), deterministic (single-threaded), and execution-time tractable. PolyBench satisfies these requirements, whereas larger suites (e.g., SPEC or MultiSource) would either mask single-pass effects or exceed our repetition budget. Despite its modest size, PolyBench covers a heterogeneous set of workloads [77]. Its six categories span the computation patterns most targeted by -O3, including loop nests, affine accesses, reductions, stencils, and triangular solvers. PolyBench is also the *de facto* benchmark suite for autotuning and ML-based pass-selection studies [40], facilitating comparison with prior work. We use the default dataset configuration, which maximizes level-1 cache (L1)

<sup>5</sup>[llvm.org/docs/NewPassManager.html](http://llvm.org/docs/NewPassManager.html)

<sup>6</sup>[github.com/MatthiasJReisinger/PolyBenchC-4.2.1/polybench.pdf](https://github.com/MatthiasJReisinger/PolyBenchC-4.2.1/polybench.pdf)

utilization while keeping working sets within the last-level cache (LLC), thereby exposing optimization effects without being dominated by memory bottlenecks. PolyBench is also relevant to the MLIR-based argument of §I. Modern ML stacks lower tensor operations to LLVM IR through MLIR, where code generation is ultimately driven by the same -O3 pipeline studied here. Many MLIR-generated kernels are structurally similar to PolyBench kernels: matrix multiplications and convolutions map naturally to `gemm`, `2mm`, and `3mm`; attention and reduction patterns resemble `atax`, `bicg`, `gesummv`, and `mvt`; normalization workloads resemble `correlation` and `covariance`; and iterative PDE-style computations resemble `jacobi-1d`, `jacobi-2d`, `heat-3d`, `seidel-2d`, and `fdtd-2d`. These structural correspondences suggest that our findings may transfer to MLIR-based ML pipelines, although direct validation remains future work.

**Metric Collection.** For each pipeline variant, we measure execution time, binary size, and end-to-end compilation time (wall-clock time for `clang+opt+llc`, excluding linking). Hardware behavior is characterized using `perf stat` and nine validated events [78], including L1 data-cache (D1), L1 instruction-cache (L1-I), and last-level cache (LLC) misses, cycles, instructions, branch misses, and package energy measured via RAPL (power/energy-pkg/, DRAM/GPU excluded). Cache references, peak RSS, page faults, and context switches are collected as confound checks.

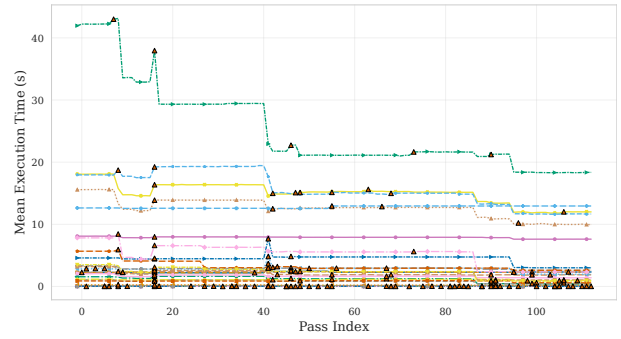
**Experimental Setup.** All experiments ran on Intel Core i9-12900KF (Alder Lake, 8P+8E cores, 5.2 GHz max), 128 GB RAM, Fedora Linux 43 (kernel 6.19), with cache hierarchy: 640 KiB D1, 768 KiB L1-I, 14 MiB L2, 30 MiB L3. We used LLVM 21.1.8 (release, no assertions, no PGO). The -O3 pass sequence was extracted via `opt -print-pipeline-passes -O3` and flattened. The pipeline has 113 distinct invocations (including repeats of `instcombine`, `simplifycfg`, `early-cse`). Our protocol entails 84,750 measurements (50,850 runtime, 33,900 hardware-counter runs). Runtime measurements use `hyperfine` (3 warmup, 15 timed runs), while hardware counters are collected with `perf stat` (2 warmup, 10 timed runs) in a separate batch. Ten runs are sufficient to keep the 95% confidence interval below 1% of the mean for all counter metrics.

**Noise Mitigation.** To reduce measurement variability, benchmarks are pinned to a dedicated P-core using `taskset`, executed under real-time FIFO scheduling (`chrt`), run with the CPU governor set to performance (`cpupower`), and evaluated with address-space layout randomization (ASLR) disabled (`setarch -R`). Hardware performance counters are collected in separate runs to avoid instrumentation overhead.

Artifacts, including the enumerated pass sequences, are available on GitHub<sup>7</sup>.

#### IV. RESULTS

We report median statistics as recommended in [79], [80], and mean statistics with a 95% confidence interval. All multi-benchmark figures use the following legend (unique color,



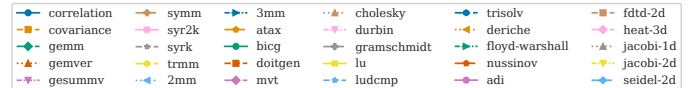
(a) runtime evolution with 95% confidence interval.



(b) cumulative speedup  $S_i$  (top) and normalized gain  $G_b(i)$  (bottom).

Fig. 1: Runtime evolution and optimization saturation.

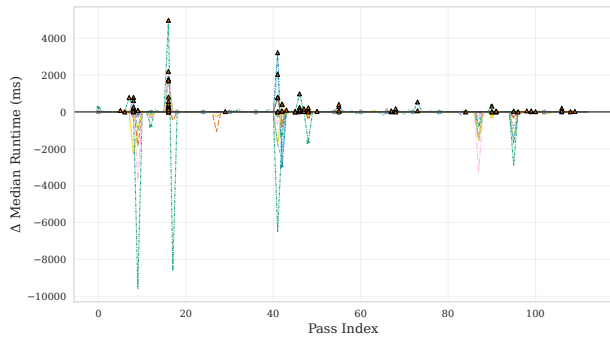
marker, line style per kernel):



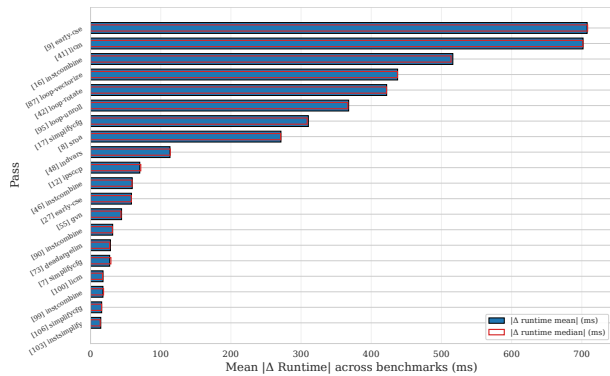
#### A. $RQ_1$ : Marginal Performance Impact

Fig. 1a plots execution time against the pipeline prefix index  $i \in [0, 112]$  ( $i = -1$  denotes unoptimized -O0). Measurement uncertainty is negligible: 95% confidence intervals remain below 1% of the mean throughout. Three findings emerge. First, the pipeline is strongly *non-monotone*: 159/1643 informative transitions (9.7%, mean) and 108/1633 (6.6%, median) degrade performance, providing direct evidence of phase-ordering interference within a single optimization pipeline. Second, speedup is strongly back-loaded. For each benchmark, the normalized gain  $G_b^i = (S_b^i - 1)/(S_b^* - 1) \in [0, 1]$  measures the fraction of the final speedup achieved at prefix  $i$ . Across the 27 non-regressing benchmarks, the median prefix required to reach 80% of the final speedup contains 84.8% of the passes (middle 50% range: 77.7%–84.8%). Thus, removing the final  $\sim 15\%$  of the pipeline sacrifices roughly 20% of the attainable speedup, contradicting the intuition that most gains occur early. Third, Fig. 1b reveals a characteristic

<sup>7</sup>[github.com/FedericoBruzzone/llvm-passview](https://github.com/FedericoBruzzone/llvm-passview)



(a) marginal-utility plot of  $\Delta T_i$  along the pipeline.



(b) top-20 passes ranked by mean of  $|\Delta T_i|$ ; median ranking overlaid.

Fig. 2: Marginal impact analysis and top-20 passes ranking.

S-shaped profile: an initial canonicalization phase, a steep growth region dominated by loop vectorization, and a final asymptote. Several curves temporarily exceed  $G_b = 1$ , indicating intermediate pipeline states that outperform the final variant. These overshoots coincide with the regressions observed in Fig. 1a, confirming that phase interference is systematic. The largest gains concentrate on dense-matrix and stencil kernels: `gemm` ( $3.86\times$ ), `jacobi-2d` ( $3.72\times$ ), `heat-3d` ( $3.37\times$ ), `fddt-2d` ( $3.02\times$ ), `syrr` ( $2.51\times$ ), `floyd-warshall` ( $2.29\times$ ), `nussinov` ( $2.18\times$ ). Three finish slower than -00: `correlation` ( $0.96\times$ ), `covariance` ( $0.98\times$ ), `seidel-2d` ( $0.98\times$ ); -03 *generates* opportunities mid-pipeline that subsequent passes *undo* (§IV-D). At  $\sim 25\%$  of the pipeline almost all benchmarks remain at  $S \approx 1$ ; by  $\sim 50\%$ , 70% have crossed  $1.2\times$ ; at the end, 80% reach  $1.5\times$  or higher.

Decomposing the cumulative trajectory into per-pass deltas  $\Delta T_i = T_{i-1} - T_i$  (Fig. 2a) reveals most lines cluster on zero, with outside spikes in two bursts (pass indices 8–18 and 41–48)—direct evidence for an “80/20” Pareto-dominated regime. Simultaneous spikes across many lines identify *universal* optimizations, whereas isolated deviations identify *program-specific* ones. Aggregating marginal contributions across benchmarks yields a cross-benchmark mean-absolute-impact ranking (Fig. 2b). The distribution is severely Pareto-skewed: the top decile accounts for the majority of impact, while the bottom half is indistinguishable from zero. The ranking is dominated by

`licm`, `instcombine`, `loop-vectorize`, `loop-rotate`, `loop-unroll`, `simplifycfg`, `sroa`, `indvars`. Top spot is occupied by `early-cse` (27 out of 30 benchmarks), which canonically eliminates common sub-expressions [81], [82]. `loop-vectorize` ranks 4th despite large impact on only  $\sim 6$  benchmarks, confirming the ranking captures both universal and program-specific passes. Some passes (`licm`, `instcombine`, `simplifycfg`, `early-cse`) appear at multiple pipeline positions, reflecting LLVM’s interleaved scheduling. The ranking is *per-occurrence*: each position of `instcombine` receives its own entry, preserving position-specific context.

We complement the impact ranking with a per-benchmark, per-pass Cohen’s  $d$  analysis (Fig. 3, clipped to  $[-3, +3]$ ; ‡:  $|d| > 3$ , \*: Wilcoxon  $p < 0.05$ ). The heatmap is dominated by ‡ cells, requiring careful reading: extreme  $d$  arises from genuinely large mean deltas *and*, more frequently, from near-zero variance that mechanically inflates the formula; ‡ signals *out of scale*, not *large absolute effect*. Cells with negligible effect ( $|d| < 0.2$ ) appear for lower-ranked passes, consistent with the long-tail picture of Fig. 2b. The *defensible core*, simultaneously large and reliable effects, is identified by joint ‡ (or high  $|d|$ ) *and* \*, whereas practical magnitude should be read jointly from Figs. 2a and 2b.

Across categories the speedup spread is substantial: `blas` and `stencils` see dramatic late-pipeline gains (medians crossing  $1.5\times$  and  $2.0\times$ ), while `datamining` remains near  $1.0\times$  (memory-bound reductions are vulnerable to ill-timed transformations). Solvers and `medley` improve near-monotonically.

### Finding 1 (RQ<sub>1</sub>)

The -03 pipeline is empirically *non-monotone*: 6.6% (median estimator) to 9.7% (mean) of pass-to-pass transitions induce runtime regressions. A small subset of passes (`early-cse`, `licm`, `instcombine`, `loop-vectorize`, and `loop-rotate`) accounts for the dominant share of realized speedup, while the long tail produces near-zero marginal impact. Contrary to the common “front-loaded” intuition, the realized gain is in fact *back-loaded*: across the 27 non-regressing kernels, 24 cross  $G_b=0.8$  at one of two specific late-pipeline transitions—immediately after `loop-vectorize` (pass 87, 10 kernels) or `loop-unroll` (pass 95, 14 kernels)—driving the median crossing to 84.8% of the pipeline.

### B. RQ<sub>2</sub>: Multi-Dimensional Trade-offs

RQ<sub>2</sub> asks whether runtime improvements are bought at the price of compile-time or footprint regressions. Fig. 4 shows compile time (near-monotonic, accelerating where loop/inlining passes fire) and binary size (*non-monotone*: early passes shrink it, late inlining/unrolling inflate it), foreshadowing dominated configurations. Fig. 5 unifies the trade-off picture. The compile-time-vs-speedup frontier is L-shaped: most speedup is attained at the pipeline tail ( $\sim$ variant 80) for

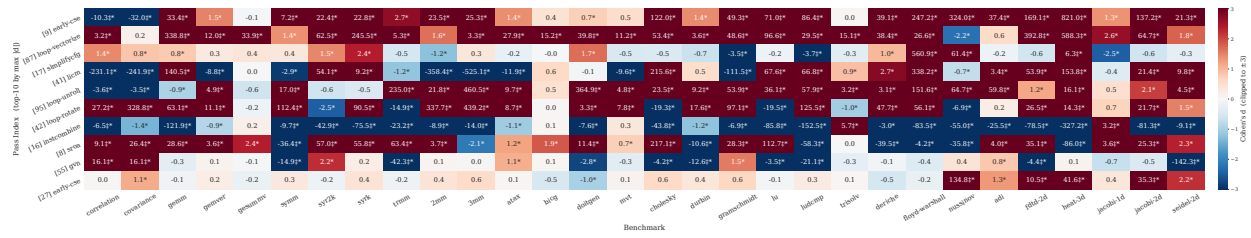


Fig. 3: Cohen’s  $d$  between consecutive variants for the top-10 passes, clipped to  $[-3, +3]$ . †:  $|d| > 3$ . \*: Wilcoxon  $p < 0.05$ .

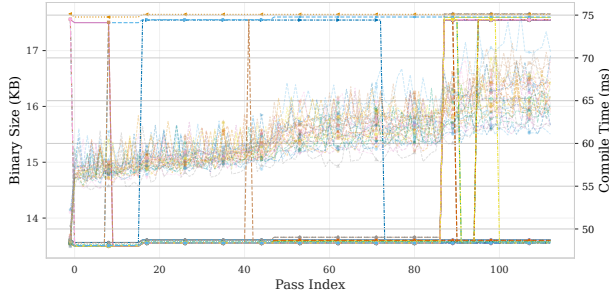


Fig. 4: Compile time and binary size trajectories.

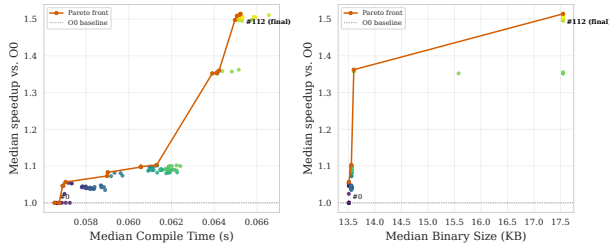


Fig. 5: Pareto views of the multi-objective trade-off, each point aggregating all benchmarks at one variant index.

modest additional compile cost. The binary-size-vs-speedup frontier is two-step: an initial wave of speedups is essentially free in size, while final jumps coincide with abrupt size increases. In 29 out of 30 benchmarks, the -03 endpoint is *dominated*: an earlier checkpoint achieves higher speedup at no larger binary size. The mechanism differs by kernel class: for BLAS/linear-algebra kernels ( $\sim 29\%$  binary expansion vs. -00), regressions arrive after size has plateaued, while for stencils, the pattern is pure speedup regression at constant size. Stopping at peak speedup yields a point that is simultaneously no larger and strictly faster—a per-benchmark signature of phase-ordering interference.

Fig. 6 shows that static IR-instruction count cannot serve as a proxy for dynamic runtime. The left panel (IR count vs. median runtime, one fitted trend per benchmark) yields a cross-benchmark Spearman  $\rho = 0.31$ , only a weak aggregate relationship. The right panel makes the within-benchmark picture explicit: 27 out of 30 benchmarks exhibit *negative*  $\rho$  (more IR implies faster runtime), reflecting that passes which expand the IR (i.e., loop unrolling, vectorization) simultaneously reduce execution time. Only *nussinov*, *bigc*, and *floyd-warshall* show positive  $\rho$ . The majority of bars fall

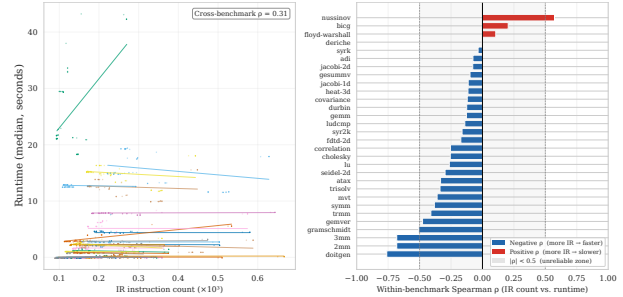


Fig. 6: *Left*: IR instruction vs. median runtime by benchmark; cross-benchmark Spearman  $\rho = 0.31$ . *Right*: within-benchmark Spearman  $\rho$  for all 30 benchmarks.

within the  $|\rho| < 0.5$  unreliable zone regardless of sign.

### Finding 2 (RQ<sub>2</sub>)

Compile time grows near-monotonically; binary size is non-monotone—early passes shrink it via dead-code elimination, late inlining and unrolling inflate it. In 29 out of 30 benchmarks the final -03 configuration is *Pareto-dominated* on (binary size, speedup) by an earlier pipeline checkpoint. The *opt* stage accounts for less than half of total compile time and its share is essentially benchmark-independent: pass-budget decisions can be made per pipeline, not per program. Static IR-instruction count is *not* a reliable proxy for dynamic runtime under pipeline-prefix variations.

### C. RQ<sub>3</sub>: Hardware-Level Insights

RQ<sub>3</sub> asks how IR-level transformations propagate to hardware behavior. Fig. 7 reports the percentage change relative to -00 across nine counters for the top-10 pass indices, faceted by benchmark. The colormap is anchored at the 95th-percentile magnitude to prevent outliers from flattening remaining cells; blue encodes improvement, and red encodes regression. The top-10 positions cluster into three groups. Scalar-canonicalization (16 *instcombine*, 17 *simplifycfg*, 8 *sroa*, 9 *early-cse*) produces the most uniform *instructions per cycle* (IPC) gain (median +11–14%, 26–28/30 benchmarks) at the cost of transient instruction inflation (+5–13%), whereas cycles nonetheless fall –3–4%, confirming net throughput gain. Propagation (12 *ipscp*) reduces instructions (–6–7%), cycles (–5%), and energy (–4–6%) consistently (19–28/30 benchmarks). The LLC cell



Fig. 7: Per-benchmark 5×6 grid of hardware-counter signatures at the top-10 most impactful pass positions.

at pass 9 is the most heterogeneous in the early pipeline (−56% to +262%), reflecting program-specific working-set changes induced by redundancy elimination. Loop-transformation/vectorization (41 licm, 42 loop-rotate, 48 indvars, 87 loop-vectorize, 95 loop-unroll) span the widest range: pass 41 is the most variable (cycles −47% to +70%; LLC +295%), while passes 42 and 48 are uniformly beneficial (instructions −7–11%, energy −8–9%, 25–28 benchmarks). Branch misses at passes 17 (simplifcfcg), 41, 42, and 48 show isolated −98% cells for several benchmarks, as loop transforms virtually eliminate loop-closing mispredictions for statically predictable loops. Pass 95 yields

the deepest suite-wide reductions (instructions −37% median, 29/30; cycles −33%; L1-I misses −11%; energy −29%) but splits the LLC column: deep blue for most benchmarks, deep red for the lu/ludcmp/cholesky triad due to working-set expansion from aggressive unrolling on dense triangular kernels—the same pass producing qualitatively different microarchitectural signatures across benchmarks.

Fig. 8 presents per-pass  $\Delta$ -trajectories for six metrics. All six panels share the same structure: for the vast majority of passes every benchmark sits on a flat baseline (median  $\Delta \approx 0$ ), and activity concentrates in a few vertical bands at passes 8–9, 16–17, 41–42, 87, 90, and 95. Most peaks are transients—

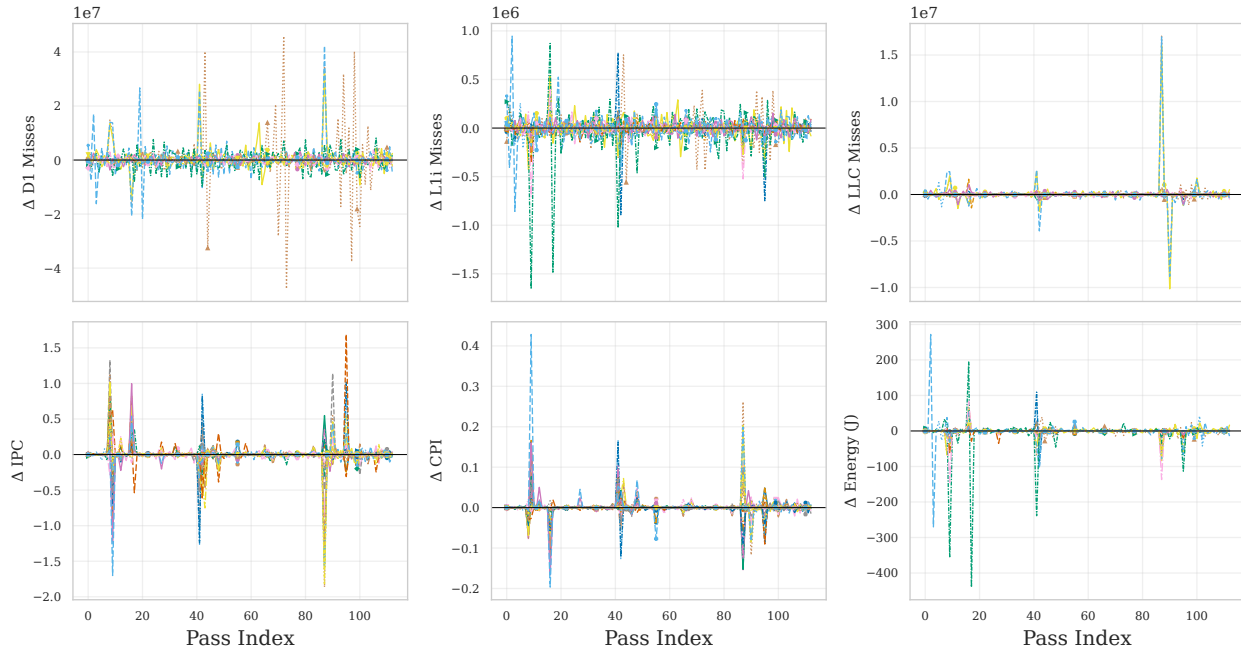


Fig. 8: Per-pass  $\Delta$ -trajectories for runtime, IPC, D1 misses, LLC misses, CPU cycles, and energy along the full -03 pipeline (median estimator).

a spike at pass  $p$  is largely undone at pass  $p+1$ —yielding near-symmetric  $\pm$  distributions ( $\approx 49/51\%$ ), while each band is driven by a handful of benchmarks. D1 is dominated by the dense linear-algebra solvers: `cholesky` spikes  $+4.6 \times 10^7$  at pass 72 and reverses ( $-4.8 \times 10^7$ ) at pass 73, while `ludcmp` adds  $+4.2 \times 10^7$  at pass 87. L1-I ( $\leq 0.17 \times 10^7$ ) is dominated by `floyd_warshall`, with deepest reductions at passes 9 and 17. LLC is almost entirely driven by `cholesky/ludcmp/lu`, peaking at pass 87 ( $+16.0$ – $17.0 \times 10^6$ ), partially reversing at pass 90 ( $-8.8$  to  $-10.1 \times 10^6$ ). IPC and *cycles per instruction* (CPI) move in anti-phase and oscillate almost suite-wide across consecutive passes; it goes up at pass 8 (28/30), down at pass 9 (27/30), up at pass 16 (27/30), down at pass 41 (25/30). The largest IPC gains are at pass 95 (`doitgen`,  $+1.69$ ) and pass 8 (`syrk`,  $+1.33$ ), and the deepest degradation at pass 87 (`cholesky`,  $-1.86$ ; `jacobi_2d`,  $-1.82$ )—pass 87 is uniquely multi-dimensionally disruptive (simultaneously degrades IPC, inflates LLC, increases D1). Energy is dominated by `floyd_warshall` ( $-440$ J at pass 17,  $-354$ J at pass 9) and `seidel_2d` (symmetric  $\pm 270$ J at passes 2–3, negligible end-to-end).

Fig. 9 presents a unified  $2 \times 3$  view of microarchitectural co-evolution. In the cache and energy panels, the cross-benchmark spread dwarfs the median—the -00 *interquartile range* (IQR) spans nearly two orders of magnitude (D1:  $0.89$ – $48.9 \times 10^6$ )—so a near-flat median hugs the floor while the solver kernels (`cholesky/lu/ludcmp`) and `floyd_warshall` float far above. The first pass (-00  $\rightarrow$  pass 0) is inert in all six panels, and each median advances as a plateau broken only by steps at the recurring hot passes 8–9, 16, 41–42, 87, 95, and 106 (`simplifycfg`) and 112 (`globaldce`). D1

median is nearly flat ( $2.78$ – $3.18 \times 10^6$ ,  $+2.6\%$  end-to-end); the spread is dominated by `cholesky/lu/ludcmp`, peaking at  $327$ – $366 \times 10^6$ . L1-I declines  $-36.8\%$  ( $7.22 \rightarrow 4.56 \times 10^5$ ) through two drops at passes 9 ( $-18\%$ ) and 87 ( $-25\%$ ). LLC rises  $+35.6\%$  ( $3.86 \rightarrow 5.23 \times 10^5$ ), with the dominant inflection at pass 87 ( $+41\%$ ); the `lu` triad forms an outlier cluster at  $\sim 2 \times 10^7$ . IPC/CPI open at  $2.96/0.338$ , peak at  $3.50/0.286$  (pass 8), collapse to  $2.55/0.393$  at pass 41, recover partially, then end at  $2.44/0.409$ —a  $-17.4\%$  IPC degradation consistent with large wall-clock speedups, since *single instruction multiple data* (SIMD) reduces instruction count far more than latency increases penalize throughput. Energy (RAPL power/energy-pkg/) decreases  $-40.9\%$  ( $92.2 \rightarrow 54.5$ J), with the dominant step at pass 87 ( $-35\%$ ,  $-30.7$ J)—to our knowledge the first *per-pass* energy profile of the LLVM pipeline. L1-I and energy are the only metrics that genuinely converge: their IQRs contract end-to-end ( $1.44 \rightarrow 0.87 \times 10^6$  and  $221 \rightarrow 113$ J), whereas the D1 spread stays essentially fixed. The cross-cutting insight: the binary-bloat trade-off (§IV-B) does *not* translate into instruction-cache pressure—L1-I misses fall with the instruction count—and its only microarchitectural cost is data-side: *late-pipeline unrolling inflates LLC misses on the dense solvers, while the suite-wide speedup comes from fewer instructions rather than from higher IPC*.

### Finding 3 (RQ<sub>3</sub>)

Runtime gains come from large reductions in instruction count and cycles—*not* from higher throughput: IPC *falls* ( $-17.4\%$ ) as vectorization trades many cheap instructions for fewer, longer-latency SIMD operations, while the D1

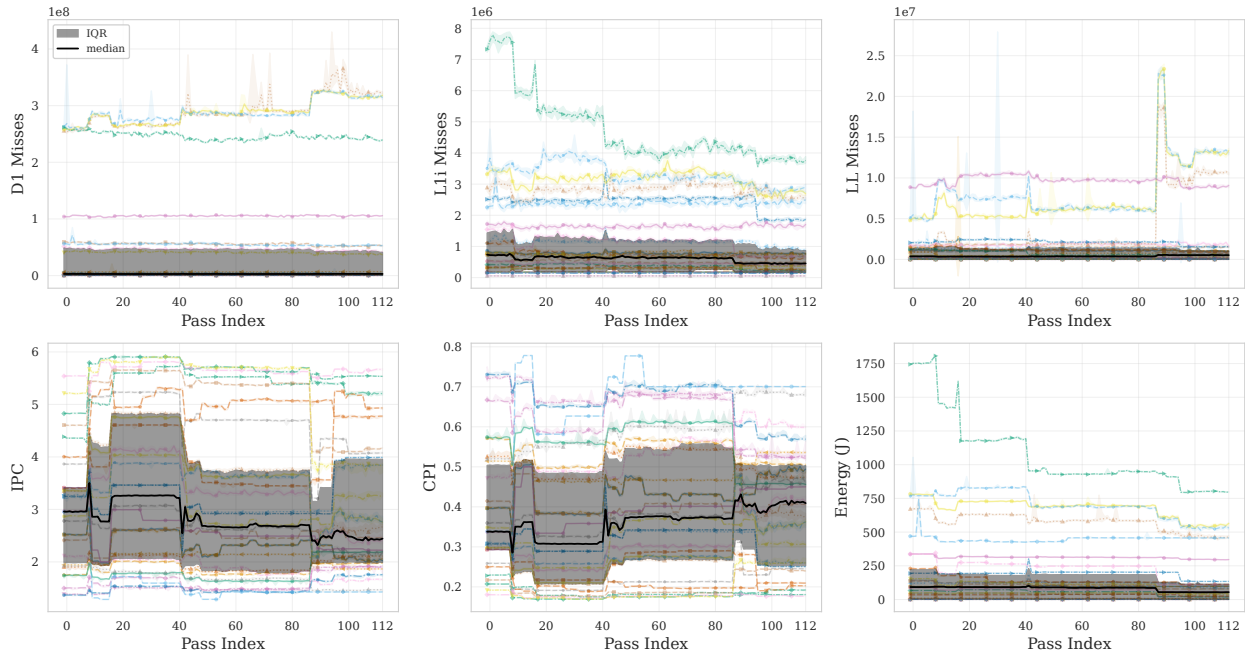


Fig. 9: Per-pass evolution of D1, LLC, and L1-I miss rates, IPC, CPI, and energy.

miss rate stays essentially flat. Energy savings track runtime savings (30–60% suite-wide median): runtime-targeted passes are *de facto* energy-targeted for compute-bound kernels. Conversely, loop-vectorize and loop-unroll *lower* L1-I misses (fewer instructions fetched); the binary-bloat trade-off of §IV-B surfaces instead as *data* working-set expansion, inflating LLC misses on the dense solvers (lu/ludcmp/cholesky). Branch-misprediction behavior is essentially flat: control-flow optimization is not on the critical path for PolyBench.

#### D. RQ<sub>4</sub>: Phase-Interference Loss

The non-monotonic trajectories raise the question of how much speedup is lost to phase interference. We define the optimistic additive bound:

$$S_{\text{opt}} = \sum_i \max(\Delta S_i, 0), \quad L = \frac{S_{\text{opt}} - [S_{\text{actual}}]_0^{S_{\text{opt}}}}{S_{\text{opt}}},$$

where  $[x]_0^{S_{\text{opt}}} = \min(\max(x, 0), S_{\text{opt}})$  and  $S_{\text{actual}} = S_{\text{final}} - 1$ . Fig. 10 reports  $L$ : mean 46.35% (median 39.3%, IQR [30.4%, 54.2%]);  $L$  measures the gap from the idealized additive ceiling  $S_{\text{opt}}$ , not from a recoverable target. Three benchmarks finish below -00: correlation (0.96×), covariance, seidel-2d (0.98×). The worst case is correlation ( $L=100\%$ ,  $S_{\text{opt}}=46.9\%$  neutered to  $S_{\text{actual}}=-3.9\%$ ), with the dominant regression pinned to pass 41 (licm,  $\Delta S=-25.98\%$ , +173,637 L1-I misses), most of which is recovered *in situ* by the immediately-following pass 42 (loop-rotate,  $\Delta S=+24.59\%$ ). Critically,  $L$  is cumulative:  $|\Delta S_i|$  of order 0.2–0.3% across 113 passes can accumulate  $S_{\text{opt}} \approx 9\%$  while erasing 6%, yielding  $L \approx 67\%$

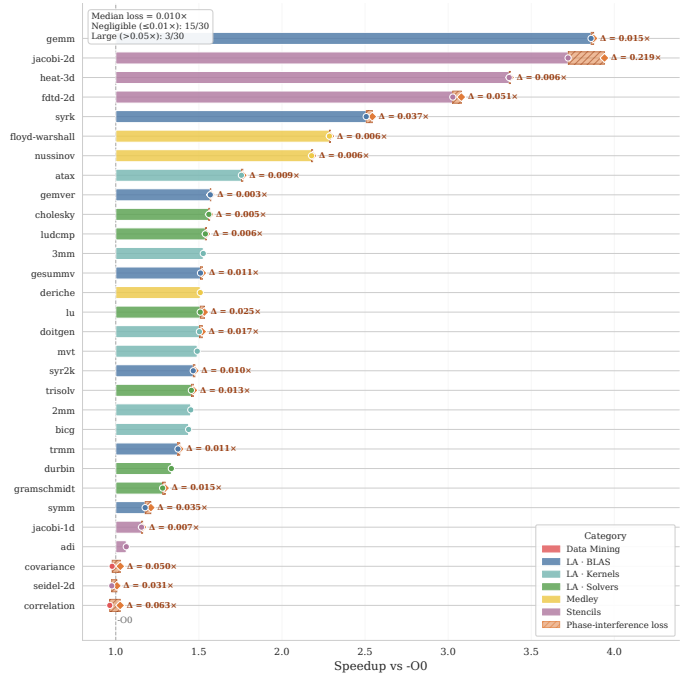


Fig. 10: Bounded phase-interference loss  $L$  per benchmark.

with no single catastrophic pass—the waterfall localizes the regression,  $L$  quantifies the discarded opportunity. Unlike prior phase-ordering work, which measures loss relative to an alternative ordering discovered by search [32], [49], [83], [84],  $L$  is—to our knowledge—the first search-free, per-pass-attributed quantification of the speedup lost to destructive phase interactions [37], [85] within the default -O3 pipeline of a production compiler.

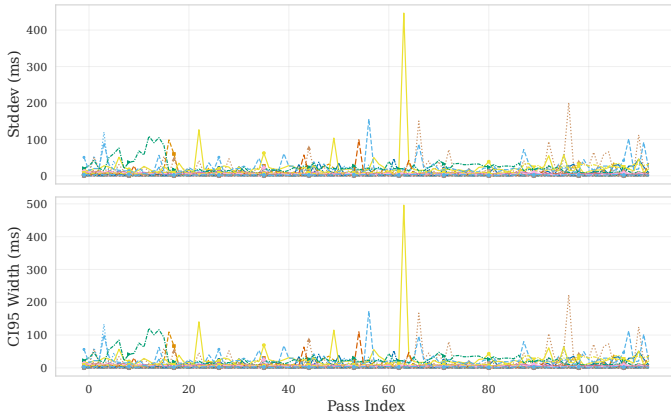


Fig. 11: Per-pass runtime standard deviation (top) and 95% confidence-interval width (bottom). Both remain below 1% of the mean and under 1 ms.

#### Finding 4 (RQ<sub>4</sub>)

The mean idealized-additive upper bound on phase-interference loss is  $L=46.35\%$  (median 39.3%, IQR [30.4%, 54.2%]);  $L$  ceilings the speedup that could be recovered from intra-pipeline interference, but the non-commutative composition of passes (§II) makes the additive ceiling itself unrealizable. Three benchmarks finish below -00 (correlation 0.96 $\times$ , covariance and seidel-2d 0.98 $\times$ ); on correlation ( $L=100\%$ ), the worst transition is pass 41 (loop-rotate,  $\Delta S=-25.98\%$ , +173,637 L1-I misses) and the immediately-following pass 42 (licm(allowspeculation),  $\Delta S=+24.59\%$ ) recovers most of it: the pipeline self-corrects *in situ* (§V).

## V. DISCUSSION & IMPLICATIONS

**Pipeline Surgery.** The bounded phase-interference loss  $L = 46.35\%$  (§IV-D) is, to our knowledge, the first *search-free*, reproducible bound on speedup lost to intra-pipeline interference, complementing prior search-based phase-ordering estimates [83], [85]. We stress that  $S_{\text{opt}}$  is an *idealized additive ceiling*: because passes are non-commutative and non-additive (§II),  $L$  *upper-bounds*, rather than estimates, the recoverable speedup. The attainable fraction is likely much smaller, and closing it is a search problem, not a free subtraction. In the worst case correlation ( $L = 100\%$ , §IV-D), the dominant regression at pass 41 is largely repaired *in situ* by the immediately following pass 42 rather than by an external guard, suggesting that the residual  $-3.9\%$  accumulates across later non-canonical interactions rather than at this single transition. Three concrete edits are suggested by the data:

- 1) the Pareto plots of Fig. 5 show that the final -03 variant is dominated by an earlier checkpoint in 29 of 30 benchmarks, while a “stop at trajectory-best” flag removes late-pipeline binary bloat with no runtime regression;
- 2) near-zero in-context marginal impact (measured in the canonical order) does not imply global dispensability: a pass that is inert at its current position may create IR

structures exploited by later passes or be critical in a reordered pipeline. Pruning decisions should therefore be validated by re-running the modified pipeline end-to-end, not by marginal impact alone;

- 3) phase-interference data (Fig. 10) localize the responsible pass at single-index resolution, enabling targeted “do not run  $p_i$  on IR matching signature  $X$ ” guards, which are strictly less invasive than full phase-ordering search.

The unsupervised top-10 ranking recovers exactly the existing hand-tuned core (early-cse, licm, instcombine, loop-vectorize, loop-rotate, ...) that any pruning effort must treat as load-bearing.

**Implications for Cost Models and MLIR.** Fig. 6 highlights the sharpest negative result of our study: IR-instruction count and dynamic runtime are essentially uncorrelated both cross-benchmark and within-benchmark. Any cost model that uses IR count as a dominant runtime proxy will systematically misrank pipeline variants and must be augmented with dynamic or microarchitectural signals (cycles, IPC, D1 misses) [40], [56]. The pass-impact ranking is robust to estimator choice at the top (mid-rank passes re-order under the median), making the median-based ranking the conservative baseline for pass-pruning heuristics. Conversely, the near-determinism of counters such as instructions-retired and cycles is an autotuning-useful signal: per-pass IR-transformation deltas on PolyBench are reproducible well below measurement noise, so Bayesian-optimization priors built over these counters can treat per-pass effects as essentially deterministic. For compute-bound numerical kernels, passes reducing runtime also reduce energy 30–60% (Fig. 9); -03 is a free-lunch energy intervention for this class, though transfer to memory-bound/batched/accelerator-hosted ML requires validation (§VI). On the dense solvers late-pipeline loop-vectorize and loop-unroll expand the *data working set*, inflating LLC misses (L1-I misses fall with instruction count); an MLIR lowering pipeline gating these behind a working-set/LLC-capacity predicate would exploit this mechanism.

**Methodology Recommendations and Limits.** Reporting mean and median jointly caught estimator-driven artifacts at negligible cost—we recommend it as a default for pass-level studies. The robustness diagnostics (Fig. 11) were specified *a priori*, making per-pass confidence-interval widths a core component of the experimental design. All recommendations are conditioned on PolyBench (default-dataset), LLVM 21.1.8, -03, single-threaded, Intel Alder Lake. The energy co-benefit and the working-set/LLC mechanism depend on the host’s RAPL domain and cache hierarchy. The pass-impact ranking is version-locked to LLVM 21.1.8, though the dominance of instcombine, licm, and early-cse is expected to be stable across versions and optimization levels, as they are the core of the hand-tuned pipeline. Re-validation on SPEC CPU, -0s/-0z, or different platforms is required before generalizing.

## VI. THREATS TO VALIDITY

The discussion follows the Wohlin *et al.*'s taxonomy [86].

**Conclusion Validity.** We keep the 95% confidence interval under 1% of the mean for every metric (Fig. 11) and report median and mean jointly [79], [80], so estimator-driven artifacts surface rather than hide. The Cohen's  $d$  heatmap (Fig. 3) is dominated by near-deterministic counters whose vanishing variance mechanically inflates  $|d|$ ; we therefore read effect size with Wilcoxon significance and the magnitude views of Figs. 2a and 2b, never promoting a pass on  $d$  alone. We apply no family-wise correction across the  $113 \times 30$  grid: our claims rest on consistent cross-benchmark patterns, not isolated  $p$ -values. Finally,  $L$  is an *idealized additive* upper bound (§IV-D), not an estimate of recoverable speedup.

**Internal Validity.** The differential design ascribes the  $B_{i-1} \rightarrow B_i$  delta to  $p_i$ , but this delta is context- and order-dependent: a pass inert in the canonical -O3 order may be load-bearing for later passes or in a reordered pipeline. We thus refrain from inferring global dispensability and require end-to-end revalidation before pruning (§V). System confounds are controlled as in §III (core pinning, RT-FIFO scheduling, performance governor, disabled ASLR), with cache references, RSS, page faults, and context switches as checks. Runtime and counters are collected separately to avoid instrumentation overhead.

**Construct Validity.** "Marginal pass impact" is operationalized as the in-context prefix delta, measuring contribution *within* the canonical order—aligned with how -O3 deploys passes, not a standalone-effect measurement. Runtime, `clang+opt+llc` compile time, and binary size are direct measures; RAPL package energy excludes the DRAM and GPU domains, under-capturing full-system energy for memory-bound workloads. IR-instruction count is treated as a candidate runtime proxy and found inadequate (Fig. 6)—a reported negative result, not an unaddressed threat. Flattening the hierarchical pass managers abstracts away nesting but preserves execution order and dependencies, so each prefix is a faithful, replayable -O3 sub-pipeline.

**External Validity.** The findings are conditioned on PolyBench/C 4.2.1 (default dataset), LLVM 21.1.8, -O3, single-threaded, on one Intel Alder Lake host. PolyBench's regular affine loop nests isolate single-pass effects but do not represent irregular, pointer-heavy, or multi-threaded code; larger suites would mask those effects or exceed our repetition budget (§III), so cross-suite replication is required before generalizing. The ranking is version-locked, though the dominance of `instcombine`, `licm`, and `early-cse`—the hand-tuned core—is expected to be stable; the energy and LLC/working-set effects depend on the host's RAPL domain and cache hierarchy. The MLIR transfer argument rests on structural isomorphism (§III) and is an unvalidated hypothesis; direct MLIR measurement, -O5/-Oz, and memory-bound or accelerator are future work.

## VII. RELATED WORK

We are not aware of any systematic study quantifying individual LLVM pass impact across multiple performance

dimensions; we review the relevant strands below.

**Search-Based and Learning-Based Optimization.** Iterative compilation [26], [27], [87] explores the optimization space by evaluating compiler configurations: small search budgets can outperform -O3 [26], [49] with program-dependent optimal configurations [45], [47], [88], while heuristic [42]–[44] and ML [36], [46], [89] approaches, collaborative filtering [90], and graph-based models [35] address the combinatorial space. Autotuning [40] adds Bayesian optimization [91], neural predictors [32], [46], and reinforcement learning (AutoPhase [84], CompilerGym [92]), with recent work reducing search cost via multi-phase learning [39], sub-sequence modeling [38], and proxy metrics [31]; a recurring limitation is reliance on expensive or inaccurate performance signals. More broadly, ML links hardware counters to optimization decisions [56] or uses deep learning for vectorization [93] and phase ordering [84], yet surveys [40] highlight the lack of fine-grained training data and difficulty isolating pass-level effects. Our work is orthogonal and complementary: we quantify each pass's marginal contribution across runtime, binary size, and counters, producing fine-grained data for supervision, surrogate targets, or priors while filling this pass-level data gap.

**Compiler Fuzzing and Testing.** Csmith [94], EMI [95], YARPGen [96], and CLsmith [97] target correctness, not performance. Our work is complementary: correctness enables reliable measurement, while performance characterization can guide fuzzing toward less understood optimizations.

**Performance Analysis of Compiler Optimizations.** De La Torre *et al.* [57] study LLVM passes but only execution time on a limited set. Chen *et al.* [62] quantify per-pass contributions in GCC/ICC, showing Pareto-skewed impact consistent with our findings—suggesting this is a property of optimization structure rather than a compiler-specific artifact. Cavazos *et al.* [56] use counters to predict benefits; Demertzi *et al.* [55] analyze reliability; Popescu and Lopes [54] study undefined behavior; Tan *et al.* [53] identify missed optimizations. None provides a systematic, cumulative, multi-dimensional evaluation of the full LLVM -O3 pipeline.

**Superoptimization and Peephole Verification.** Our ranking, dominated by `instcombine` and `early-cse`, is consistent with superoptimization [98], [99]: peephole algebraic simplifications account for a disproportionate share of achievable optimizations. Our empirical ranking provides data-driven corroboration that these rewrites are load-bearing in LLVM -O3.

**Profile-Guided Optimization.** Finding 2 (§IV-B) concludes that pass-budget decisions can be made per pipeline based on the benchmark-independence of the `opt/total` compile-time ratio. This is in tension with PGO/FDO literature [100] showing per-program profiling is essential for optimal pass selection, but the tension is not a contradiction: our finding concerns *compile-time budget allocation*, while PGO concerns *which transformations to apply*—both can simultaneously hold.

## VIII. CONCLUSION

We presented a systematic, multi-dimensional study of the LLVM -O3 pipeline, decomposing it into cumulative pass prefixes across 84,750 measurements over 113 variants of 30 PolyBench/C kernels. The pipeline is non-monotone (6.6–9.7% of transitions regress) and back-loaded, with a small Pareto-dominant core driving the gains. The final -O3 variant is Pareto-dominated on (size, speedup) in 29/30 kernels. IR-instruction count is an unreliable runtime proxy; runtime-targeted passes are *de facto* energy-targeted (30–60%). The search-free, idealized-additive upper bound on phase-interference loss is  $L = 46.35\%$  (a ceiling, not a recoverable target). With our infrastructure, these results inform pass pruning, cost-model calibration, and autotuning. Future work targets SPEC CPU, -O0s/-Oz, more platforms, and MLIR-lowered ML kernels.

## REFERENCES

- [1] A. P. Ershov, "On Programming of Arithmetic Operations," *Communications of the ACM*, vol. 1, no. 8, pp. 3–6, Aug. 1958.
- [2] J. Heller, "Sequencing Aspects of Multiprogramming," *Journal of ACM*, vol. 8, no. 3, pp. 426–439, Jul. 1961.
- [3] W. M. McKeeman, "Peephole Optimization," *Communications of the ACM*, vol. 8, no. 7, pp. 443–444, Jul. 1965.
- [4] C. W. Gear, "High Speed Compilation of Efficient Object Code," *Communications of the ACM*, vol. 8, no. 8, pp. 483–488, Aug. 1965.
- [5] E. S. Lowry and C. W. Medlock, "Object Code Optimization," *Communications of the ACM*, vol. 12, no. 1, pp. 13–22, Jan. 1969.
- [6] V. A. Busam and D. E. Englund, "Optimization of Expressions in Fortran," *Comm. of the ACM*, vol. 12, no. 12, pp. 666–674, Dec. 1969.
- [7] F. E. Allen, "Program Optimization," in *Annual Review of Automatic Programming*. Pergamon Press, 1969, vol. 5, pp. 239–307.
- [8] W. A. Wulf, *The Design of an Optimizing Compiler*. CMU, Dec. 1973.
- [9] K. D. Cooper and L. Torczon, *Engineering a Compiler*, Nov. 2022.
- [10] K. Kennedy and J. R. Allen, *Optimizing Compilers for Modern Architectures: A Dependence-Based Approach*. Morgan Kaufmann, Oct. 2001.
- [11] R. Paige, "Future Directions in Program Transformations," *ACM Computing Surveys*, vol. 28, no. 4, pp. 170–174, Dec. 1996.
- [12] D. F. Bacon, S. L. Graham, and O. J. Sharp, "Compiler Transformations for High-Performance Computing," *ACM Computing Surveys*, vol. 26, no. 4, pp. 345–420, Dec. 1994.
- [13] M. Dodds, M. Batty, and A. Gotsman, "Compositional Verification of Compiler Optimisations on Relaxed Memory," in *Proc. ESOP'18*, LNCS 10801, Thessaloniki, Greece: Springer, Apr. 2018, pp. 1027–1055.
- [14] A. V. Aho, M. S. Lam, R. Sethi, and J. D. Ullman, *Compilers: Principles, Techniques, and Tools*, 2nd ed. Addison-Wesley, 2006.
- [15] C. Lattner and V. Adve, "LLVM: A Compilation Framework for Lifelong Program Analysis and Transformation," in *Proc. CGO'04*, San José, CA, USA: IEEE, Mar. 2004, pp. 75–86.
- [16] N. D. Matsakis and F. S. Klock, "The Rust Language," *ACM SIGAda Letters*, vol. 34, no. 3, pp. 103–104, Oct. 2014.
- [17] J. Bezanson, S. Karpinski, V. B. Shah, and A. Edelman, "Julia: A Fast Dynamic Language for Technical Computing," *arXiv e-prints*, arXiv:1209.5145, pp. 1–27, Sep. 2012.
- [18] C. Lattner, M. Amini, U. Bondhugula, A. Cohen, A. Davis, J. Pienaar, R. Riddle, T. Shpeisman, N. Vasilache, and O. Zinenko, "MLIR: Scaling Compiler Infrastructure for Domain Specific Computation," in *Proc. CGO'21*. Virtual: IEEE, Feb./Mar. 2021, pp. 2–14.
- [19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Proc. NeurIPS'19*. Vancouver, Canada: Curran Associates, Inc., Dec. 2019, pp. 8026–8037.
- [20] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A System for Large-Scale Machine Learning," in *Proc. OSDI'16*. Savannah, GA, USA: USENIX, Nov. 2016, pp. 265–283.
- [21] M. Li, Y. Liu, X. Liu, Q. Sun, X. You, H. Yang, Z. Luan, L. Gan, G. Yang, and D. Qian, "The Deep Learning Compiler: A Comprehensive Survey," *IEEE Trans. Par. and Distr. Syst.*, vol. 32, no. 3, pp. 708–727, Mar. 2021.
- [22] Y. Xing, J. Weng, Y. Wang, L. Sui, Y. Shan, and Y. Wang, "An In-Depth Comparison of Compilers for Deep Neural Networks on Hardware," in *Proc. ICESST'19*. Las Vegas, NV, USA: IEEE, Jun. 2019, pp. 1–8.
- [23] R. Zhao, S. Liu, H.-C. Ng, E. Wang, J. J. Davis, X. Niu, X. Wang, H. Shi, G. A. Constantinides, P. Y. K. Cheung, and W. Luk, "Hardware Compilation of Deep Neural Networks: An Overview," in *Proc. ASAP'18*. Milan, Italy: IEEE, Jul. 2018, pp. 1–8.
- [24] H.-I. C. Liu, M. Brehler, M. Ravishankar, N. Vasilache, B. Vanik, and S. Laurenzo, "TinyIREE: An ML Execution Environment for Embedded Systems from Compilation to Deployment," *IEEE Micro*, vol. 42, no. 5, pp. 9–16, Sep. 2022.
- [25] T. Chen, T. Moreau, Z. Jiang, H. Shen, E. Q. Yan, L. Wang, Y. Hu, L. Ceze, C. Guestrin, and A. Krishnamurthy, "TVM: An Automated End-to-End Optimizing Compiler for Deep Learning," *arXiv e-prints*, arXiv:1802.04799, pp. 1–16, Oct. 2018.
- [26] K. D. Cooper, P. J. Schielke, and D. Subramanian, "Optimizing for Reduced Code Space Using Genetic Algorithms," in *Proc. LCTES'99*, Atlanta, GA, USA: ACM, May 1999, pp. 1–9.
- [27] K. D. Cooper, D. Subramanian, and L. Torczon, "Adaptive Optimizing Compilers for the 21st Century," *J. Supercomputing*, vol. 23, no. 1, pp. 7–22, Aug. 2002.
- [28] S. Triantafyllis, M. Vachharajani, N. Vachharajani, and D. I. August, "Compiler Optimization-Space Exploration," in *Proc. CGO'03*, San Francisco, CA, USA: IEEE, Mar. 2003, pp. 204–215.
- [29] S.-A.-A. Touati and D. Barthou, "On the Decidability of Phase Ordering Problem in Optimizing Compilation," in *Proc. CF'06*, Ischia, Italy: ACM, May 2006, pp. 147–156.
- [30] S. M. Freudenberger and J. C. Ruttenberg, "Phase Ordering of Register Allocation and Instruction Scheduling," in *Proc. WoCC'92*. Dagstuhl, Germany: Springer, May 1991, pp. 146–170.
- [31] J. Zhao, C. Xia, and Z. Wang, "Leveraging Compilation Statistics for Compiler Phase Ordering," in *Proc. IPDPS'25*. Milan, Italy: IEEE, Jun. 2025, pp. 533–545.
- [32] A. H. Ashouri, A. Bignoli, G. Palermo, C. Silvano, S. Kulkarni, and J. Cavazos, "MiCOMP: Mitigating the Compiler Phase-Ordering Problem Using Optimization Sub-Sequences and Machine Learning," *ACM Trans. Arch. Code Opt.*, vol. 14, no. 3, pp. 29:1–29:28, Sep. 2017.
- [33] J. W. Davidson, G. S. Tyson, D. B. Whalley, and P. A. Kulkarni, "Evaluating Heuristic Optimization Phase Order Search Algorithms," in *Proc. CGO'07*. San José, CA, USA: IEEE, Mar. 2007, pp. 157–169.
- [34] S. R. Vegdahl, "Phase Coupling and Constant Generation in an Optimizing Microcode Compiler," in *Proc. MICRO'82*, Oct. 1982, pp. 125–133.
- [35] R. Nobre, L. G. A. Martins, and J. a. M. P. Cardoso, "A Graph-Based Iterative Compiler Pass Selection and Phase Ordering Approach," in *Proc. LCTES'16*. Santa Barbara, USA: ACM, Jun. 2016, pp. 21–30.
- [36] F. Agakov, E. V. Bonilla, J. Cavazos, B. Franke, G. Fursin, M. F. P. O'Boyle, J. Thomson, M. Toussaint, and C. K. I. Williams, "Using Machine Learning to Focus Iterative Optimization," in *Proc. CGO'06*, New York, NY, USA: IEEE, Mar. 2006, pp. 305–315.
- [37] D. L. Whitfield and M. L. Soffa, "An Approach to Ordering Optimizing Transformations," in *Proc. PPOPP'90*, ACM, Mar. 1990, pp. 137–146.
- [38] H. Pan, Y. Wei, M. Xing, Y. Wu, and C. Zhao, "Towards Efficient Compiler Auto-tuning: Leveraging Synergistic Search Spaces," in *Proc. CGO'25*. Las Vegas, NV, USA: IEEE, Mar. 2025, pp. 614–627.
- [39] M. Zhu, D. Hao, and J. Chen, "Compiler Autotuning through Multiple-phase Learning," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 4, pp. 1–38, May 2024.
- [40] A. H. Ashouri, W. Killian, J. Cavazos, G. Palermo, and C. Silvano, "A Survey on Compiler Autotuning using Machine Learning," *ACM Computing Surveys*, vol. 51, no. 5, pp. 96:1–96:42, Sep. 2018.
- [41] A. Tiwari, C. Chen, J. Chame, M. Hall, and J. K. Hollingsworth, "A Scalable Auto-Tuning Framework for Compiler Optimization," in *Proc. IPDPS'09*, Rome, Italy: IEEE, May 2009, pp. 1–12.
- [42] Z. Pan and R. Eigenmann, "PEAK—A Fast and Effective Performance Tuning System via Compiler Optimization Orchestration," *ACM*

- Transactions on Programming Languages and Systems*, vol. 30, no. 3, pp. 17:1–17:43, May 2008.
- [43] —, “Fast and Effective Orchestration of Compiler Optimizations for Automatic Performance Tuning,” in *Proc. CGO’06*, 2006, pp. 319–332.
- [44] P. Kulkarni, S. Hines, J. Hiser, D. Whalley, J. Davidson, and D. Jones, “Fast Searches for Effective Optimization Phase Sequences,” in *Proc. PLDI’04*, Washington, DC, USA: ACM, Jun. 2004, pp. 171–182.
- [45] S. Purini and L. Jain, “Finding Good Optimization Sequences Covering Program Space,” *ACM Trans. Arch. Code Opt.*, vol. 9, no. 4, Jan. 2013.
- [46] S. Kulkarni and J. Cavazos, “Mitigating the Compiler Optimization Phase-Ordering Problem Using Machine Learning,” in *Proc. OOPSLA’12*. Tucson, AZ, USA: ACM, Oct. 2012, pp. 147–162.
- [47] P. Kulkarni, W. Zhao, H. Moon, K. Cho, D. Whalley, J. Davidson, M. Bailey, Y. Paek, and K. Gallivan, “Finding Effective Optimization Phase Sequences,” in *Proc. LCTES’03*, ACM, Jun. 2003, pp. 12–23.
- [48] D. L. Whitfield and M. L. Soffa, “An Approach for Exploring Code Improving Transformations,” *ACM Transactions on Programming Languages and Systems*, vol. 19, no. 6, pp. 1053–1084, Nov. 1997.
- [49] L. Almagor, K. Cooper, A. Grosul, T. Harvey, S. Reeves, D. Subramanian, L. Torczon, and T. Waterman, “Finding Effective Compilation Sequences,” in *Proc. LCTES’04*, ACM, Jun. 2004, pp. 231–239.
- [50] M. Zhao, B. R. Childers, and M. L. Soffa, “An Approach Toward Profit-Driven Optimization,” *ACM Transactions on Architecture and Code Optimization*, vol. 3, no. 3, pp. 231–262, Sep. 2006.
- [51] J. Liu, J. Fang, T. Wang, J. Xie, C. Huang, and Z. Wang, “Efficient Compiler Optimization by Modeling Passes Dependence,” *CCF Transactions on High Performance Computing*, vol. 6, no. 6, pp. 588–607, Dec. 2024.
- [52] K. Georgiou, C. Blackmore, S. Xavier-de Souza, and K. Eder, “Less is More: Exploiting the Standard Compiler Optimization Levels for Better Performance and Energy Consumption,” in *Proc. SCOPES’18*, Sankt Goar, Germany: ACM, May 2018, pp. 35–42.
- [53] J. Tan, S. Jiao, M. Chabbi, and X. Liu, “What Every Scientific Programmer Should Know about Compiler Optimizations?” in *Proc. ICS’20*. Barcelona, Spain: ACM, Jun./Jul. 2020, pp. 1–12.
- [54] L. Popescu and N. P. Lopes, “Exploiting Undefined Behavior in C/C++ Programs for Optimization: A Study on the Performance Impact,” in *Proc. PLDI’25*, Seoul, South Korea: ACM, Jun. 2025, pp. 348–371.
- [55] M. Demertzi, M. Annavamam, and M. Hall, “Analyzing the Effects of Compiler Optimizations on Application Reliability,” in *Proc. IISWC’11*, Austin, TX, USA: IEEE, Nov. 2011, pp. 184–193.
- [56] J. Cavazos, G. Fursin, F. Agakov, E. Bonilla, M. F. P. O’Boyle, and O. Temam, “Rapidly Selecting Good Compiler Optimizations using Performance Counters,” in *Proc. CGO’07*, Mar. 2007, pp. 185–197.
- [57] J. C. de la Torre, P. Ruiz, B. Dorronsoro, and P. L. Galindo, “Analyzing the Influence of LLVM Code Optimization Passes on Software Performance,” in *Proc. IPMU’18*, CCIS 855. Springer, Jun. 2018, pp. 272–283.
- [58] S. Palkar, J. Thomas, D. Narayanan, P. Thaker, R. Palamuttam, P. Negi, A. Shanbhag, M. Schwarzkopf, H. Pirk, S. Amarasinghe, S. Madden, and M. Zaharia, “Evaluating End-to-End Optimization for Data Analytics Applications in Weld,” in *Proc. VLDB’18*, May 2018, pp. 1002–1015.
- [59] N. van Kempen, H.-J. Kwon, D. T. Nguyen, and E. D. Berger, “It’s Not Easy Being Green: On the Energy Efficiency of Programming Languages,” in *Proc. ASE’25*, IEEE, Nov. 2025, pp. 1553–1565.
- [60] S. U. Lee, N. Fernando, K. Lee, and J.-G. Schneider, “A Survey of Energy Concerns for Software Engineering,” *J. Syst. Soft.*, vol. 210, Apr. 2024.
- [61] R. Pereira, M. Couto, F. Ribeiro, R. Rua, J. Cunha, J. P. Fernandes, and J. Saraiva, “Energy Efficiency Across Programming Languages,” in *Proc. SLE’17*. Vancouver, BC, Canada: ACM, Oct. 2017, pp. 256–267.
- [62] Y. Chen, S. Fang, Y. Huang, L. Eeckhout, G. Fursin, O. Temam, and C. Wu, “Deconstructing Iterative Optimization,” *ACM Trans. Arch. Code Opt.*, vol. 9, no. 3, pp. 21:1–21:30, Oct. 2012.
- [63] B. K. Rosen, M. N. Wegman, and F. K. Zadeck, “Global Value Numbers and Redundant Computations,” in *Proc. PoPL’88*, Jan. 1988, pp. 12–27.
- [64] B. Alpern, M. N. Wegman, and F. K. Zadeck, “Detecting Equality of Variables in Programs,” in *Proc. PoPL’88*, ACM, Jan. 1988, pp. 1–11.
- [65] R. Cytron, J. Ferrante, B. K. Rosen, M. N. Wegman, and F. K. Zadeck, “An Efficient Method of Computing Static Single Assignment Form,” in *Proc. POPL’89*, Austin, TX, USA: ACM, Jan. 1989, pp. 25–35.
- [66] D. Callahan, K. D. Cooper, K. Kennedy, and L. Torczon, “Interprocedural Constant Propagation,” in *Proc. SCC’86*, ACM, Jun. 1986, pp. 152–161.
- [67] M. N. Wegman and F. K. Zadeck, “Constant Propagation with Conditional Branches,” *ACM Transactions on Programming Languages and Systems*, vol. 13, no. 2, pp. 181–210, Apr. 1991.
- [68] R. Cytron, J. Ferrante, B. K. Rosen, M. N. Wegman, and F. K. Zadeck, “Efficiently Computing Static Single Assignment Form and the Control Dependence Graph,” *ACM Transactions on Programming Languages and Systems*, vol. 13, no. 4, pp. 451–490, Oct. 1991.
- [69] S. S. Muchnick, *Advanced Compiler Design and Implementation*, 1st ed. Morgan Kaufmann, Aug. 1997.
- [70] S. Carr and K. Kennedy, “Scalar Replacement in the Presence of Conditional Control Flow,” *Software—Practice and Experience*, vol. 24, no. 1, pp. 51–77, Jan. 1994.
- [71] J. Lu and K. D. Cooper, “Register Promotion in C Programs,” in *Proc. PLDI’97*, Las Vegas, NV, USA: ACM, Jun. 1997, pp. 308–319.
- [72] P. P. Chang and W.-W. Hwu, “Inline Function Expansion for Compiling C Programs,” in *Proc. PLDI’89*, Portland, USA: Jun. 1989, pp. 246–257.
- [73] J. J. Dongarra and A. R. Hinds, “Unrolling loops in Fortran,” *Software—Practice and Experience*, vol. 9, no. 3, pp. 219–226, Mar. 1979.
- [74] J. W. Davidson and S. Jinturkar, “An Aggressive Approach to Loop Unrolling,” University of Virginia, VA, USA, TR-CS-95-26, Jan. 1995.
- [75] J. Knoop, O. Rüthing, and B. Steffen, “Partial Dead Code Elimination,” in *Proc. PLDI’94*, Orlando, FL, USA: ACM, Jun. 1994, pp. 147–158.
- [76] F. E. Allen and J. Cocke, “A Catalogue of Optimizing Transformations,” in *Design and Optimization of Compilers*, Prentice-Hall, 1972, pp. 1–30.
- [77] T. Yuki, “Understanding PolyBench/C 3.2 Kernels,” in *Proc. IM-PACT’14*, Wien, Austria, Jan. 2014, pp. 1–5.
- [78] V. M. Weaver and S. A. McKee, “Can Hardware Performance Counters Be Trusted?” in *Proc. IISWC’08*, IEEE, Sep. 2008, pp. 141–150.
- [79] A. Georges, D. Buytaert, and L. Eeckhout, “Statistically Rigorous Java Performance Evaluation,” in *Proc. OOPSLA’07*, Montréal, Canada: ACM, Oct. 2007, pp. 57–76.
- [80] T. Mytkowicz, A. Diwan, M. Hauswirth, and P. F. Sweeney, “Producing Wrong Data Without Doing Anything Obviously Wrong!” in *Proc. ASPLOS’09*, Washington, USA: ACM, Mar. 2009, pp. 265–276.
- [81] J. Cocke, “Global Common Subexpression Elimination,” in *Proc. SCO’70*, Urbana-Champaign, USA: ACM, Jul. 1970, pp. 20–24.
- [82] J. D. Ullman, “Fast Algorithms for the Elimination of Common Subexpressions,” *Acta Informatica*, vol. 2, no. 3, pp. 191–213, Sep. 1973.
- [83] P. A. Kulkarni, D. B. Whalley, G. S. Tyson, and J. W. Davidson, “Exhaustive Optimization Phase Order Space Exploration,” in *Proc. CGO’06*, New York, NY, USA: IEEE, Mar. 2006, pp. 306–318.
- [84] A. Haj-Ali, Q. Huang, J. Xiang, W. Moses, K. Asanovic, J. Wawrzyniak, and I. Stoica, “AutoPhase: Juggling HLS Phase Orderings in Random Forests with Deep Reinforcement Learning,” in *Proc. MLSys’20*, Austin, TX, USA: Curran, Mar. 2020, pp. 70–81.
- [85] M. R. Jantz and P. A. Kulkarni, “Analyzing and Addressing False Interactions During Compiler Optimization Phase Ordering,” *Software—Practice and Experience*, vol. 44, no. 6, pp. 643–679, Jun. 2014.
- [86] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering*. Springer, 2012.
- [87] F. Bodin, T. Kisuki, P. Knijnenburg, M. O’Boyle, and E. Rohou, “Iterative Compilation in a Non-linear Optimisation Space,” in *Proc. PFDC’98*, Paris, France, Oct. 1998.
- [88] L. Eeckhout, H. Vandierendonck, and K. De Bosschere, “Quantifying the Impact of Input Data Sets on Program Behavior and Its Applications,” *Journal of Instruction-Level Parallelism*, vol. 5, no. 1, pp. 1–33, Feb. 2003.
- [89] G. Fursin, Y. Kashnikov, A. W. Memon, Z. Chamski, O. Temam, M. Namolaru, E. Yom-Tov, B. Mendelson, A. Zaks, E. Courtois, F. Bodin, P. Barnard, E. Ashton, E. Bonilla, J. Thomson, C. K. I. Williams, and M. O’Boyle, “Milepost GCC: Machine Learning Enabled Self-Tuning Compiler,” *International Journal of Parallel Programming*, vol. 39, no. 3, pp. 296–327, Jan. 2011.
- [90] H. Liu, J. Luo, Y. Li, and Z. Wu, “Iterative Compilation Optimization Based on Metric Learning and Collaborative Filtering,” *ACM Transactions on Architecture and Code Optimization*, vol. 19, no. 1, pp. 2:1–2:25, Mar. 2021.
- [91] A. H. Ashouri, G. Mariani, G. Palermo, E. Park, J. Cavazos, and C. Silvano, “COBAYN: Compiler Autotuning Framework Using Bayesian Networks,” *ACM Transactions on Architecture and Code Optimization*, vol. 13, no. 2, pp. 21:1–21:25, Jun. 2016.
- [92] C. Cummins, B. Wasti, J. Guo, B. Cui, J. Ansel, S. Gomez, S. Jain, J. Liu, O. Teytaud, B. Steiner, Y. Tian, and H. Leather, “CompilerGym: Robust, Performant Compiler Optimization Environments for AI Research,” in *Proc. CGO’22*, Seoul, South Korea: IEEE, Apr. 2022, pp. 92–105.

- [93] A. Haj-Ali, N. K. Ahmed, T. Willke, Y. S. Shao, K. Asanovic, and I. Stoica, "NeuroVectorizer: End-to-End Vectorization with Deep Reinforcement Learning," in *Proc. CGO'20*, San Diego, CA, USA: IEEE, Feb. 2020, pp. 242–255.
- [94] X. Yang, Y. Chen, E. Eide, and J. Regehr, "Finding and Understanding Bugs in C Compilers," in *Proc. PLDI'11*, San Josè, CA, USA: ACM, Jun. 2011, pp. 283–294.
- [95] V. Le, M. Afshari, and Z. Su, "Compiler Validation via Equivalence Modulo Inputs," in *Proc. PLDI'14*, Edinburgh, United Kingdom: ACM, Jun. 2014, pp. 216–226.
- [96] V. Livinskii, D. Babokin, and J. Regehr, "Random Testing for C and C++ Compilers with YARPGen," in *Proc. OOPSLA'20*, Chicago, IL, USA: ACM, Nov. 2020, pp. 1–25.
- [97] C. Lidbury, A. Lascu, N. Chong, and A. F. Donaldson, "Many-Core Compiler Fuzzing," in *Proc. PLDI'15*, Portland, OR, USA: ACM, Jun. 2015, pp. 65–76.
- [98] H. Massalin, "Superoptimizer: A Look at the Smallest Program," in *Proc. ASPLOS'87*, Palo Alto, CA, USA: ACM, Oct. 1987, pp. 122–126.
- [99] S. Bansal and A. Aiken, "Automatic Generation of Peephole Superoptimizers," in *Proc. ASPLOS'06*, ACM, Oct. 2006, pp. 394–403.
- [100] D. Chen, T. Moseley, and D. X. Li, "AutoFDO: Automatic Feedback-Directed Optimization for Warehouse-Scale Applications," in *Proc. CGO'16*, Barcelona, Spain: IEEE, Mar. 2016, pp. 12–23.