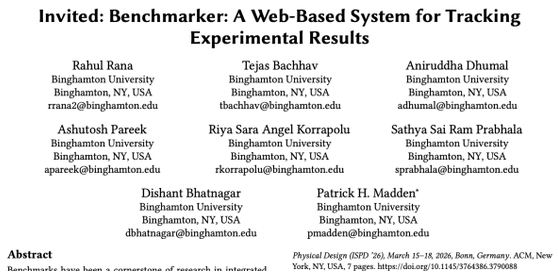
Invited: Benchmarker: A Web-Based System for Tracking Experimental Results

Rahul Rana, Tejas Bacchav, Aniruddha Dumal,
Ashutosh Pareek, Riya Sara Angel Korrapolu,
Sathya Sai Ram Prabhala, Dishant Bhatnagar,
Patrick H. Madden
Binghamton University School of Computing



Invited: Benchmarker: A Web-Based System for Tracking Experimental Results

Rahul Rana
Binghamton University
Binghamton, NY, USA
rrana2@binghamton.edu

Tejas Bachhav
Binghamton University
Binghamton, NY, USA
tbachhav@binghamton.edu

Aniruddha Dhumal
Binghamton University
Binghamton, NY, USA
adhumal@binghamton.edu

Ashutosh Pareek
Binghamton University
Binghamton, NY, USA
apareek@binghamton.edu

Riya Sara Angel Korrapolu
Binghamton University
Binghamton, NY, USA
rkorrapolu@binghamton.edu

Sathya Sai Ram Prabhala
Binghamton University
Binghamton, NY, USA
sprabhala@binghamton.edu

Dishant Bhatnagar
Binghamton University
Binghamton, NY, USA
dbhatnagar@binghamton.edu

Patrick H. Madden*
Binghamton University
Binghamton, NY, USA
pmadden@binghamton.edu

Abstract

Physical Design (ISPD '26), March 15–18, 2026, Bonn, Germany. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3764386.3790088

Background

More than twenty-five years ago, I started working on circuit placement (after having spent entirely too much time working on global routing).

To have an idea of what sort of results I'd need to get a paper published, I started gathering results from recent papers into an Excel spreadsheet. And nothing made sense.

2001 "Reporting of Standard Cell Placement Results"

Wide variation in results

	[13] Gordian	[9] PRC	[22] HALO	[14] POPINS	[19] TW7.0	[10] QUAD	[18] NRG
Fract				45225		337	25602
Struct	558+362					3780	287631
Primary1	841+553	999.5	1.439	1128245	0.83	8972	894545
Primary2	4761+3153	3665.6	6.73	4128324	3.53	36824	3412195
Biomed	5232+3123			4128324	3.22	23765	
Industry1							
Industry2					13.30	332318	
Industry3					41.53	938682	
Avqsmall				23848188	5.08	62890	
Avqlarge				28323022	5.65	65906	
Golem3					88.98		
Units	X1000	X1000	Meter	Micron	Meter	X100	Micron
Spacing	Routed				Routed	None	

	[5] FD98	[7] ARP	[21] Dragon	[24] SPADE	[11] Mongrel	[23] Feng Shui	[12] iTools
Fract		0.034		0.024		0.032	
Struct	0.338	0.34		0.291	0.266	0.380	0.272
Primary1	0.87	0.79		0.74	0.83	1.018	0.799
Primary2	3.72	3.61		3.13	2.94	3.684	3.37
Biomed	1.78	1.83		1.43		1.689	2.90
Industry1		1.50				1.606	
Industry2	14.6		12.88	11.90	11.89	15.408	11.4
Industry3	45.1	48.12	42.33	35.37	34.53	44.729	39.6
Avqsmall	4.91	6.06	5.17	5.08	4.4	5.960	4.48
Avqlarge	5.38	6.54	5.25	6.16	4.87	6.301	4.78
Golem3			77.56	19.84		21.882	79.9
Units	Meter	Meter	Meter	Meter	Meter	Meter	Meter
Spacing	Row		Row	None	None	Row	Routed

What Was Going On?

- Some placements have routing channels, others don't.
- Different numbers of rows in placements.
- Pad locations often moved substantially.
- *Huge impact on HPWL for any of these.*
- File format differences (YAL, VPNR, TW) could result in a 4X change in reported results.

I emailed a dozen different groups to ask how they computed their results. Everyone replied with a variation of **"we use the standard method."**

Surprise. There was no "standard method."

How Did This Happen?

This was the late 1990's. The web had been around for only a few years. Circuit benchmarks were often distributed on magnetic tapes.

If you wanted to submit a paper to DAC, you had to snail mail physical printed copies. Close to the submission deadline? Drive to the LAX airport (*nobody walks in LA*), where the post office stayed open late.

Many people knew that the situation wasn't ideal. I said *the quiet part, out loud. Experimental results in many papers couldn't be taken at face value.* Some feathers ruffled, some folks a bit uncomfortable, but....

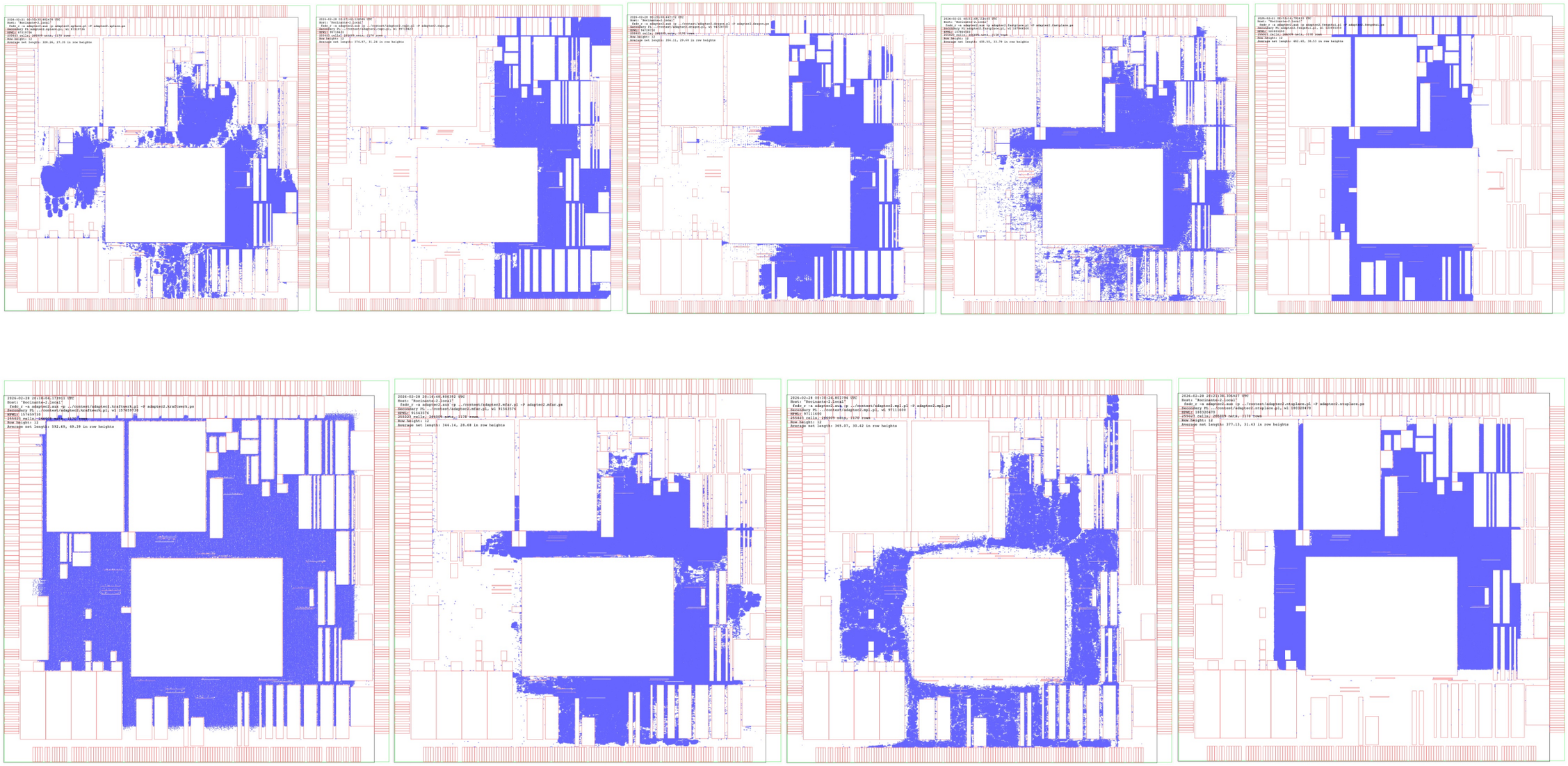
There was broad agreement that we could and should do things better.

How Things Got Better

ISPD contests started in 2005. Contest benchmark suites are created each year, and many groups participate. The tools are tested by independent experts, with all results being carefully checked. (*IBM Austin! Prof. Kahng's GSRC crew! Thank You for getting things off on the right foot!*)

The result: rapid progress, new ideas, experimental results that are the gold standard.

My 2001 paper ruffled some feathers. In 2021, there was a 20th anniversary retrospective, co-authored with many contest organizers. **Still benchmarking, after all these years!**



ISPD 2005 contest placements for Adaptec2 from Aplace, Capo, Dragon, FastPlace, feng shui, Kraftwerk, mFar, mPl, and NTUPlace. Placement files are available on the ISPD web site, wire lengths and legality checked with the *perl* scripts. Standard cells are placed around fixed macro blocks.

Benchmarker Web Site

Now, the uncomfortable part of the talk....

Artificial Intelligence is taking over everything

There's a lot of interest in using AI techniques for integrated circuit design, but it's not clear how effective it actually is.

Many of the miscommunications and differences in measurements that happened a quarter century ago have reappeared.

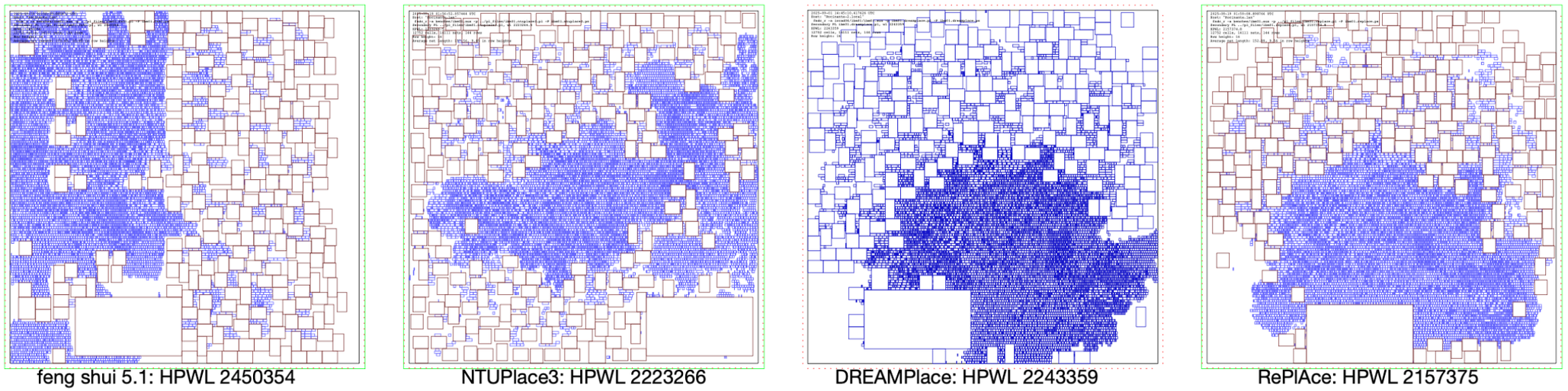
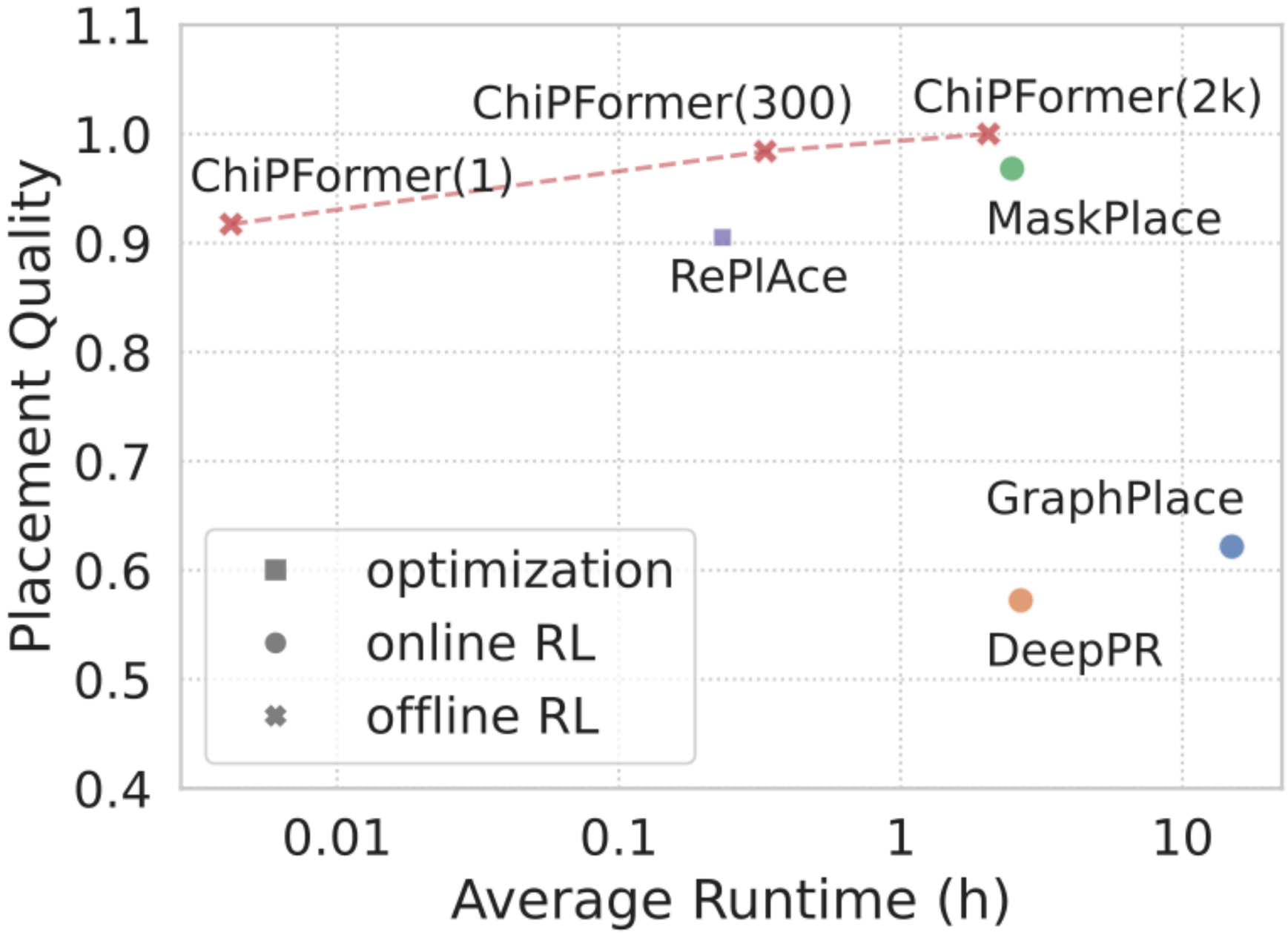


Figure 5: Placements for the ICCAD04 IBM01 benchmark, using conventional techniques. feng shui relies on recursive bisection, while NTUPlace, DREAMPlace, and RePlAce are analytic. All placements are legal, with the analytic methods obtaining the best half-perimeter wire length results.



ChiPFormer ICML 2023, https://icml.cc/virtual/2023/poster/25027

Table 8: **Comparisons of HPWL ($\times 10^7$) for mixed-size placement in *ICCAD04* benchmark.** HPWL is the smaller the better. SA (simulated annealing) method is implemented as in Mirhoseini et al. (2021). ChiPFormer(workflow) can achieve the best placement quality.

circuit	SA	RePlAce	GraphPlace	MaskPlace	ChiPFormer(workflow)
ibm01	25.85	22.82	31.71	24.18	**16.70** (-26.82%)
ibm02	54.87	47.59	55.11	47.45	**37.87** (-20.19%)
ibm03	80.68	64.36	80.00	71.37	**57.63** (-10.46%)
ibm04	83.32	72.61	86.86	78.76	**65.27** (-10.11%)
ibm06	69.09	58.07	63.48	55.70	**52.57** (-5.62%)
ibm07	111.03	98.57	117.70	95.27	**86.20** (-9.52%)
ibm08	131.07	114.67	134.77	120.64	**102.26** (-10.82%)
ibm09	135.45	120.01	148.74	122.91	**105.61** (-12.00%)
ibm10	423.14	274.29	440.78	367.55	**230.39** (-16.00%)
ibm11	210.12	169.98	218.73	202.23	**160.60** (-5.52%)
ibm12	410.05	306.33	438.57	397.25	**273.14** (-10.83%)
ibm13	259.89	220.14	278.92	246.49	**197.20** (-10.42%)
ibm14	405.80	341.80	455.32	302.67	**301.28** (-0.46%)
ibm15	510.06	451.36	520.06	457.86	**429.71** (-4.80%)
ibm16	614.54	516.05	642.08	584.67	**463.32** (-10.22%)
ibm17	720.40	635.93	814.37	643.75	**569.13** (-10.50%)
ibm18	442.00	399.43	450.67	398.83	**370.36** (-7.14%)

ChiPFormer ICML 2023, https://icml.cc/virtual/2023/poster/25027

Table 6. Comparison of HPWL (10^6) averaged over 5 seeds, using various techniques for mixed-size placement, on the IBM benchmark.

Circuit	MaskPlace + DP	WireMask-BBO + DP	ChiPFormer + DP	DREAMPlace	Diffusion (Ours)
ibm01	3.33	2.84	3.35	2.23	2.09
ibm02	7.30	6.87	6.24	5.79	4.43
ibm03	10.1	9.81	10.9	10.4	7.30
ibm04	10.4	9.65	10.1	9.13	8.00
ibm05	7.67	7.67	7.67	7.60	7.79
ibm06	7.62	8.41	7.76	6.15	8.31
ibm07	13.3	13.0	13.4	11.1	9.60
ibm08	15.5	15.9	15.7	12.3	13.3
ibm09	16.2	15.4	16.9	12.8	12.6
ibm10	46.8	45.2	45.4	44.8	30.2
ibm11	23.5	24.6	23.6	16.6	17.3
ibm12	46.1	Failed	48.8	31.0	34.0
ibm13	28.2	28.0	28.4	23.2	23.0
ibm14	45.4	48.2	46.5	31.3	34.5
ibm15	53.4	Failed	55.8	51.3	45.0
ibm16	65.9	63.2	67.3	53.0	52.7
ibm17	72.9	69.7	71.4	57.9	60.4
ibm18	42.2	41.6	41.1	37.6	38.6
Average	28.7	27.0	28.9	23.6	22.7

Diffusion, ICML 2025, https://arxiv.org/pdf/2407.12282

Which is it?

ICML 2023 reports ChiPFormer as 1.67
ICML 2025 reports ChiPFormer as 3.35

- Confirmed RePlAce results: around 2.15
- Either ChiPFormer is **27% better than RePlAce, or 54% worse**. Kind of a big gap.

ChiPFormer placement results are not available.
And there's one more gotcha...

Table 8: **Comparisons of HPWL** ($\times 10^7$) **for mixed-size placement in *ICCAD04* benchmark.** HPWL is the smaller the better. SA (simulated annealing) method is implemented as in Mirhoseini et al. (2021). ChiPFormer(workflow) can achieve the best placement quality.

circuit	SA	RePlAce	GraphPlace	MaskPlace	ChiPFormer(workflow)
ibm01	25.85	22.82	31.71	24.18	**16.70** (-26.82%)
ibm02	54.87	47.59	55.11	47.45	**37.87** (-20.19%)
ibm03	80.68	64.36	80.00	71.37	**57.63** (-10.46%)
ibm04	83.32	72.61	86.86	78.76	**65.27** (-10.11%)
ibm06	69.09	58.07	63.48	55.70	**52.57** (-5.62%)
ibm07	111.03	98.57	117.70	95.27	**86.20** (-9.52%)
ibm08	131.07	114.67	134.77	120.64	**102.26** (-10.82%)
ibm09	135.45	120.01	148.74	122.91	**105.61** (-12.00%)
ibm10	423.14	274.29	440.78	367.55	**230.39** (-16.00%)
ibm11	210.12	169.98	218.73	202.23	**160.60** (-5.52%)
ibm12	410.05	306.33	438.57	397.25	**273.14** (-10.83%)
ibm13	259.89	220.14	278.92	246.49	**197.20** (-10.42%)
ibm14	405.80	341.80	455.32	302.67	**301.28** (-0.46%)
ibm15	510.06	451.36	520.06	457.86	**429.71** (-4.80%)
ibm16	614.54	516.05	642.08	584.67	**463.32** (-10.22%)
ibm17	720.40	635.93	814.37	643.75	**569.13** (-10.50%)
ibm18	442.00	399.43	450.67	398.83	**370.36** (-7.14%)

- Note the exponent. It should be **10e5**, not **10e7**, if the RePlAce result is to make any sense. The authors confirmed that the exponent is a typo.
- Anyone who reads the ChiPFormer paper, and then implements a placer... If the result is less than 100x worse than the results in the table, they might mistake it as a massive advance.
- The ChiPFormer paper has been cited 90+ times in the past couple of years.

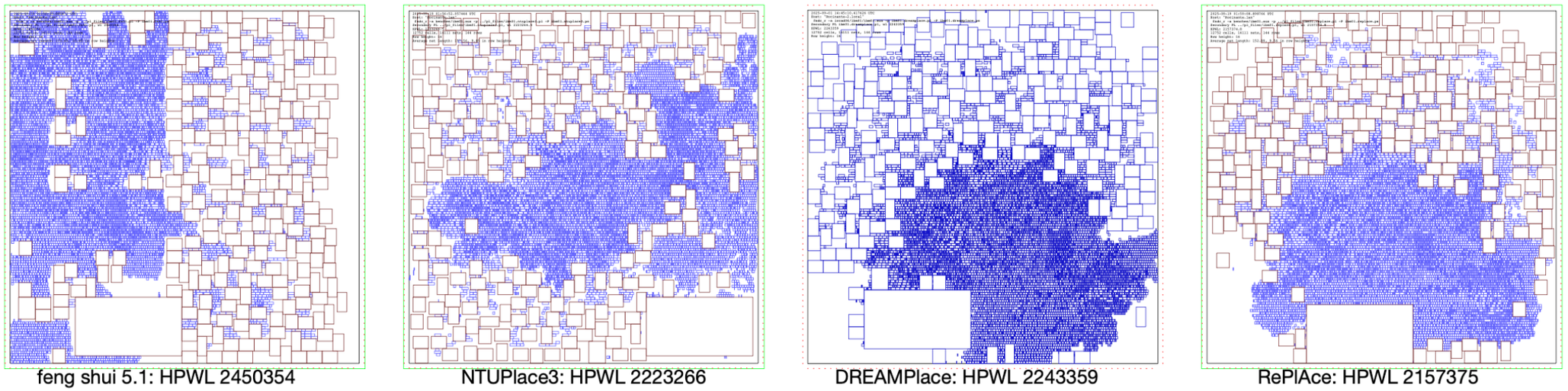


Figure 5: Placements for the ICCAD04 IBM01 benchmark, using conventional techniques. feng shui relies on recursive bisection, while NTUPlace, DREAMPlace, and RePlAce are analytic. All placements are legal, with the analytic methods obtaining the best half-perimeter wire length results.

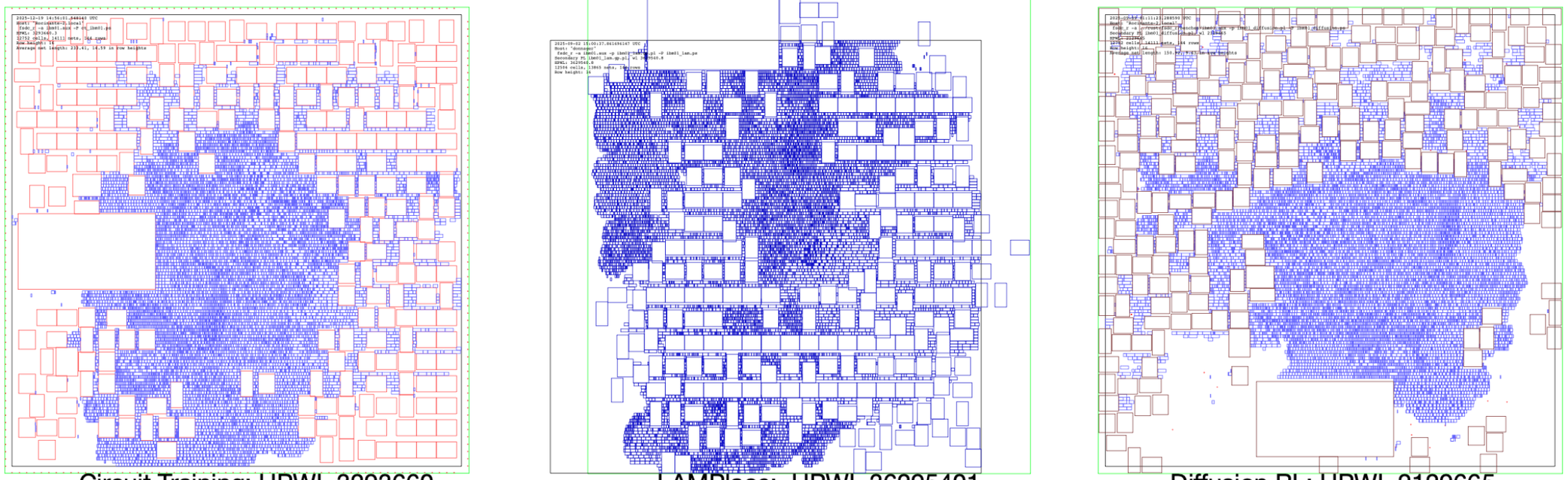


Figure 6: Placements for the ICCAD04 IBM01 benchmark using AI techniques. The circuit training placement was created using an independent implementation of the Google reinforcement learning approach; the placement is legal, but has significantly higher half-perimeter wire lengths compared to conventional approaches. The LAMPlace and Diffusion placements are illegal, having macro blocks overlapping, and macros out of the core placement region. Pads seem to be removed entirely from LAMPlace results, while pads are moved into the core by the Diffusion approach.

It's A Mess

- There are many papers from the AI community using physical design benchmarks, but experimental results seem to have gone completely unchecked. The errors are missed entirely by both authors and peer-review.
- Modified benchmarks and broken software frameworks seem to be spreading like wildfire.
- Macro blocks and cells overlap. Pads are moved (or removed). Wire lengths sometimes only consider macro blocks. Wire lengths might only consider *a subset* of macro blocks.
- Many authors do not respond to email queries. I'm tempted to call Missing Persons.

What Can Be Done?

Now, more than ever, published results need to be verified. This is the *extraordinary evidence* for *extraordinary claims*.

This verification should be timely. The Benchmarker site provides a friction free platform where papers can be tracked, errors can be flagged, results can be shared.

We've made the platform generic; it should be easy to add different benchmarks and metrics. It's not just for physical design.

Pull requests, edits, experimental results, comments, contributions, are all very much welcome! Please help!

https://github.com/profmadden/benchmarker

Benchmarker site at Binghamton, GitHub, Google Drive of benchmark suites, Benchmarker paper.