1865 THE UNIVERSITY OF MAINE

SIEGE

Secure and Intelligent Edge Research

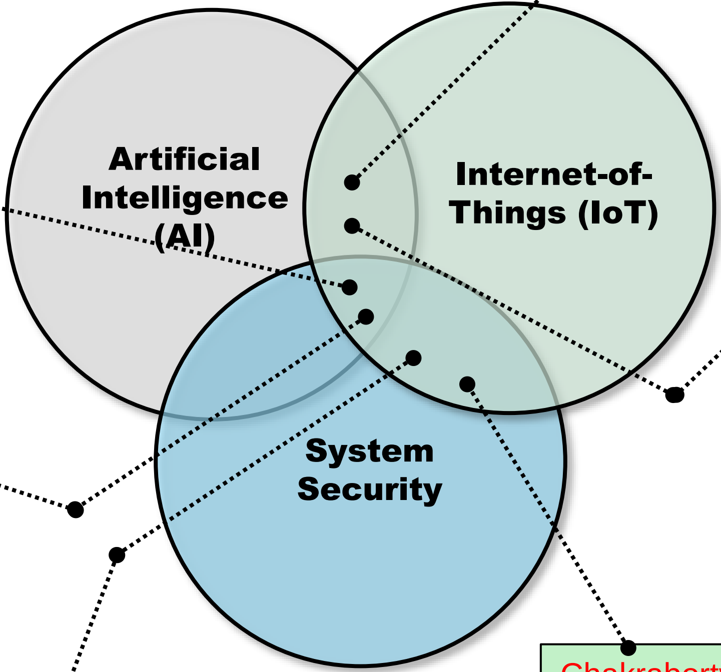
Artificial Intelligence for Microelectronics Security and Trust

Prabuddha Chakraborty

Electrical and Computer Engineering
Advanced Structures and Composites Center
University of Maine

About Myself

- PhD, University of Florida, 2022
- Assistant Professor
- University of Maine (ECE, ASCC)
- **SIEGE Research**: 7 PhD Student



Bhunia et. al., US Patent, 2020
Alaql et. al., US Patent, 2020
Bhunia el. al., Copyright, 2018

*Top Picks in Hardware and Embedded Security 2021 (IEEE HSTTC)

Chakraborty et. al., IEEE TIFS, 2021
*Chakraborty et. al., IEEE AHOST, 2018
Bhattacharyay et. al., DAC, 2022
Alaql et. al., IEEE TCAD, 2021
Yang et. al., IEEE ITC-India, 2021
Yang et. al., IEEE ISQED, 2021
^Hoque et. al., IEEE ITC, 2018

^Best Hardware Demo Award (1st) – IEEE HOST

Chakraborty et. al., Nature Scientific Reports, 2022
Chakraborty et. al., IEEE ESL, 2022
Chakraborty et. al., NCAA, 2021
Chakraborty et. al., IEEE IoT-J, 2020
Chakraborty et. al., VTC-2020 Spring
Dizon et. al., IEEE CEM, 2022

Chakraborty et. al., US Patent, 2021
Chakraborty et. al., US Patent, 2021
Chakraborty et. al., US Patent, 2020
Bhunia et. al., US Patent, 2020
Wang et. al., US Patent, 2021
Bhunia et. al., US Patent, 2021
Bhunia et. al., US Patent, 2021
Chakraborty et. al., US Patent, 2021
Bhunia et. al., US Patent, 2021

Bhattacharyay et. al., US Patent, 2020
Bhunia et. al., US Patent, 2020
Bhunia el. al., Copyright, 2020

Chakraborty et. al., IEEE ESL, 2021
Chakraborty et. al., IEEE HOST, 2019
Bhattacharyay et. al., IEEE TCAD, 2022
Yang et. al., IEEE TVLSI, 2022

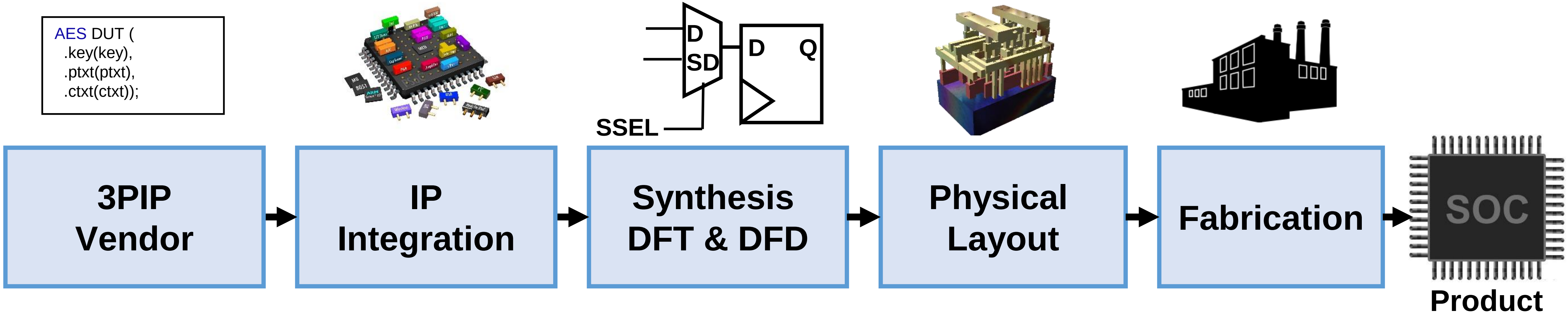
1865 THE UNIVERSITY OF MAINE

SIEGE
Secure and
Intelligent Edge
Research

Outline

- Background – Hardware Design Intellectual Property (IP) Protection
- SAIL: **S**tructural **A**nalysis using Mach**I**ne **L**earning
- SURF: Joint **S**truct**UR**al **F**unctional Attack on Logic Locking
- LeGO: **Le**arning-**G**uided Logic L**o**cking
- Background – Hardware Trojans
- MIMIC: **M**achine **I**ntelligence based **M**alicious **I**mplant **C**reation
- VIPR: **V**erification of **IP** T**R**ust
- Summary

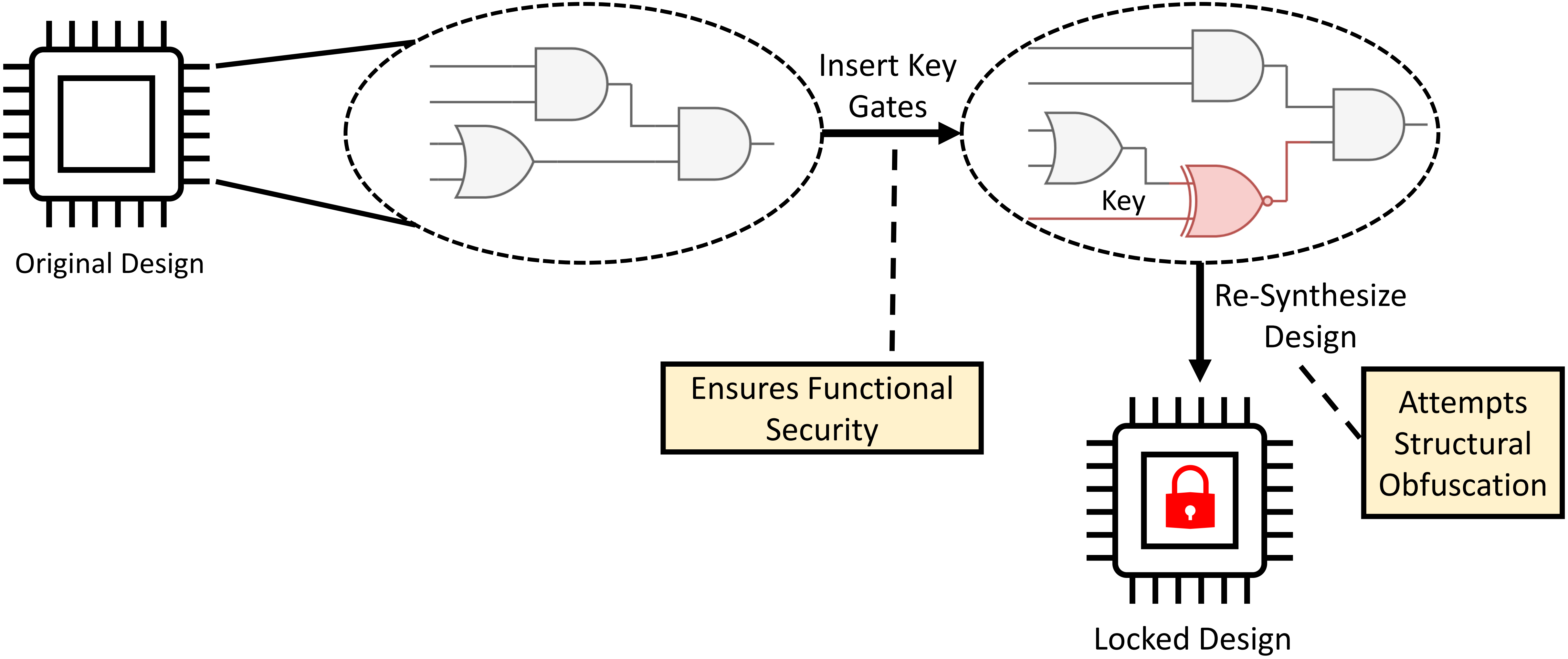
Hardware IP/IC Threats



```
AES DUT (
  .key(key),
  .ptxt(ptxt),
  .ctxt(ctxt));
```

- Security is an important design parameter
- Horizontal supply-chain brings diverse threats: IP Theft, Reverse Engineering, Trojans
- One solution is to build-in security measures in the hardware IP itself

Logic Locking: A Potential Solution

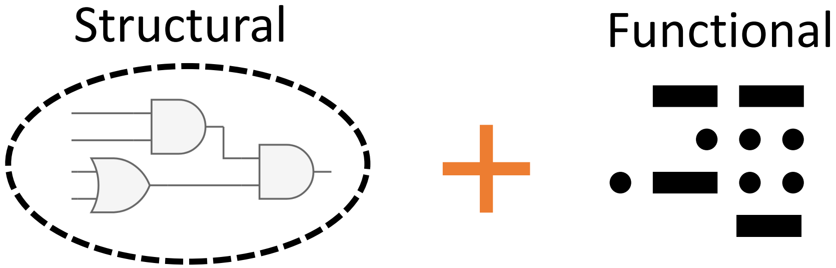


Ensures Functional Security

Attempts Structural Obfuscation

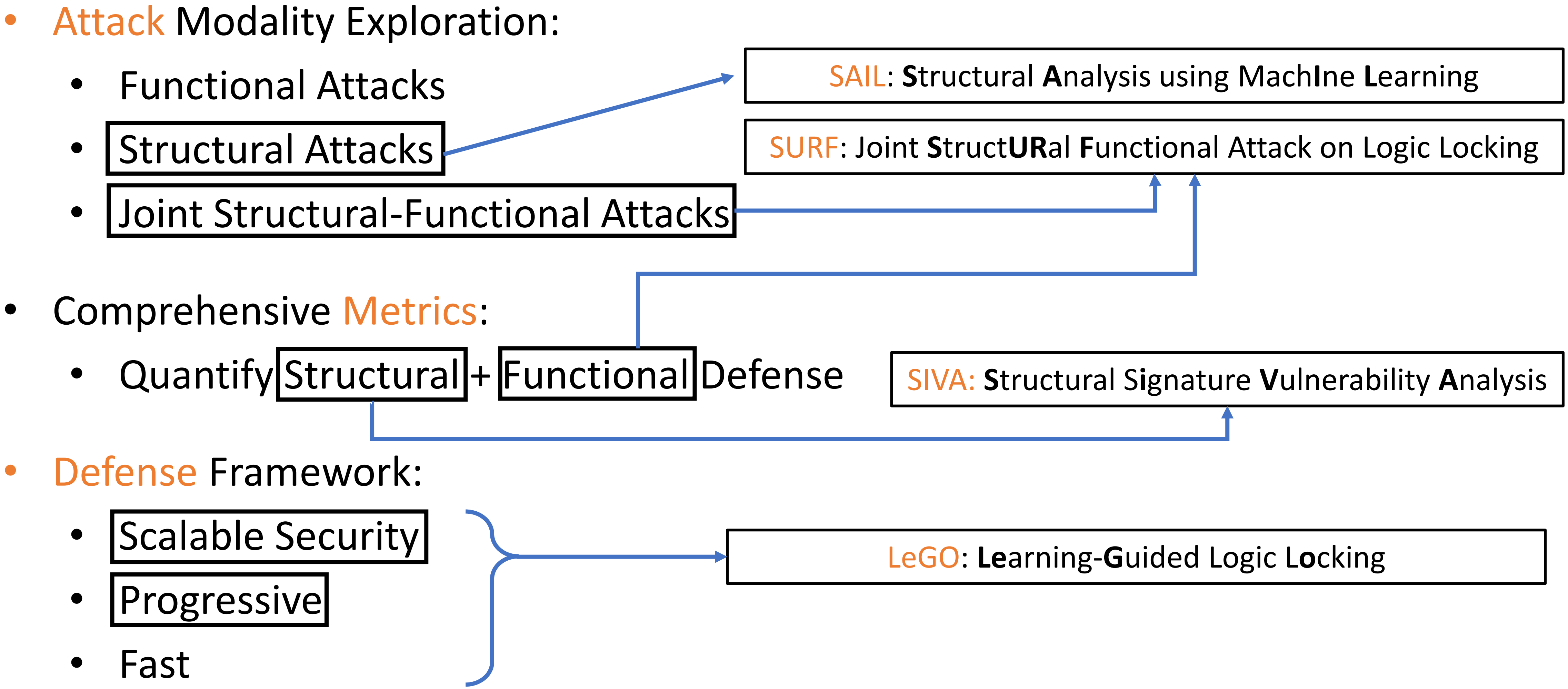
Ingredients of Good Logic Locking

- Attack Modality Exploration:
  - Functional Attacks
  - Structural Attacks
  - Joint Structural-Functional Attacks



- Comprehensive Metrics:
  - Quantify Structural + Functional Defense
- Defense Framework:
  - Scalable Security
  - Progressive
  - Fast

Summary



- Attack Modality Exploration:
  - Functional Attacks
  - Structural Attacks
  - Joint Structural-Functional Attacks
- Comprehensive Metrics:
  - Quantify Structural + Functional Defense
- Defense Framework:
  - Scalable Security
  - Progressive
  - Fast

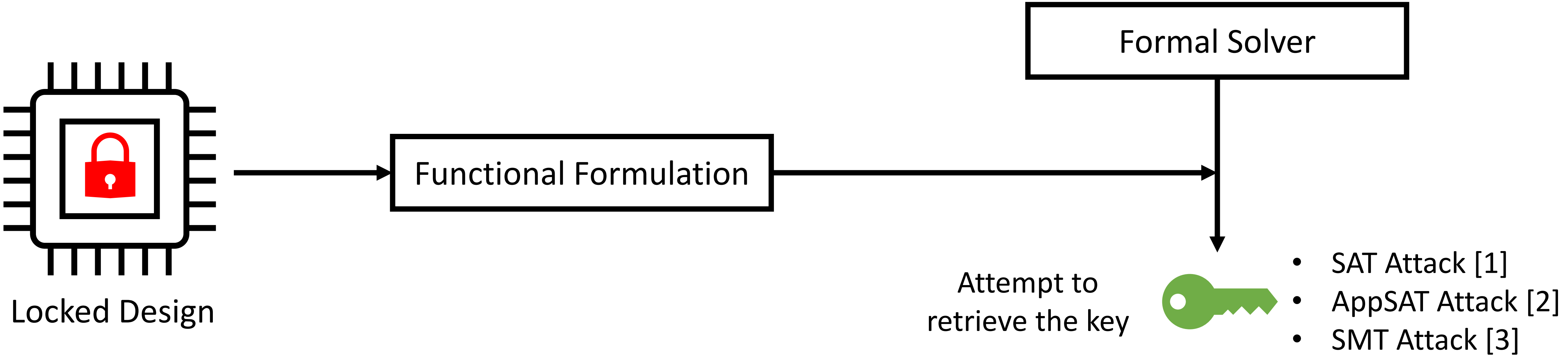
SAIL: **S**tructural **A**nalysis using Mach**I**ne **L**earning

SURF: Joint **S**truct**UR**al **F**unctional Attack on Logic Locking

SIVA: **S**tructural S**i**gnature **V**ulnerability **A**nalysis

LeGO: **Le**arning-**G**uided Logic L**o**cking

Verifying Strength of Logic Locking (How it was)



Locked Design

Functional Formulation

Formal Solver

Attempt to retrieve the key

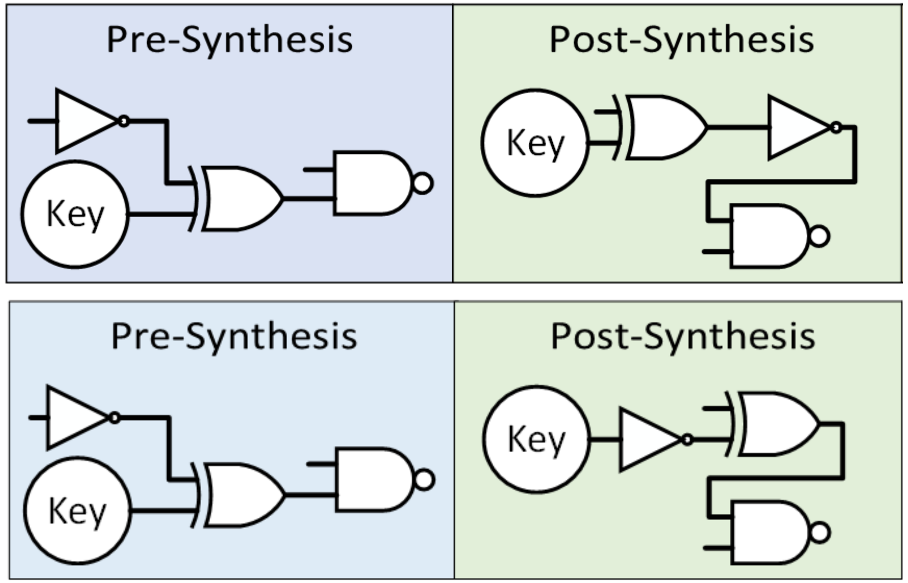
- SAT Attack [1]
- AppSAT Attack [2]
- SMT Attack [3]

- Can verify functional security of a locked design.
- Can the attacker analyze the design structurally?

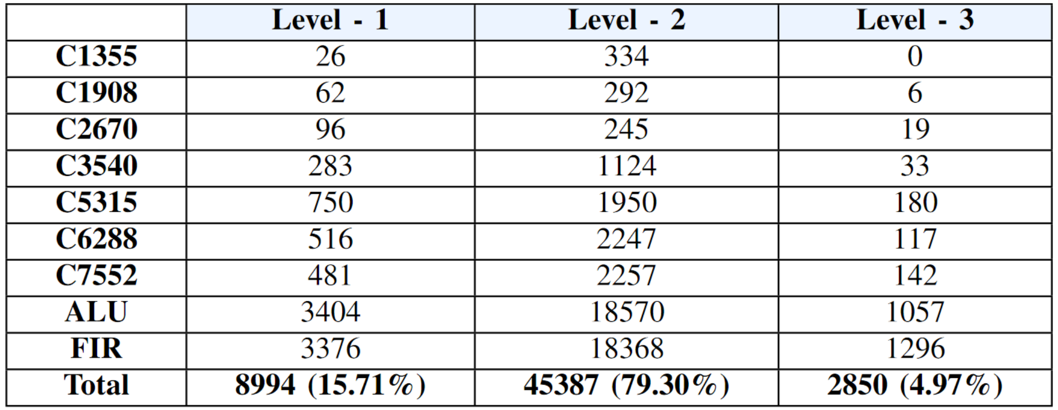
Reverse Engineer

Open a pathway for functional unlocking

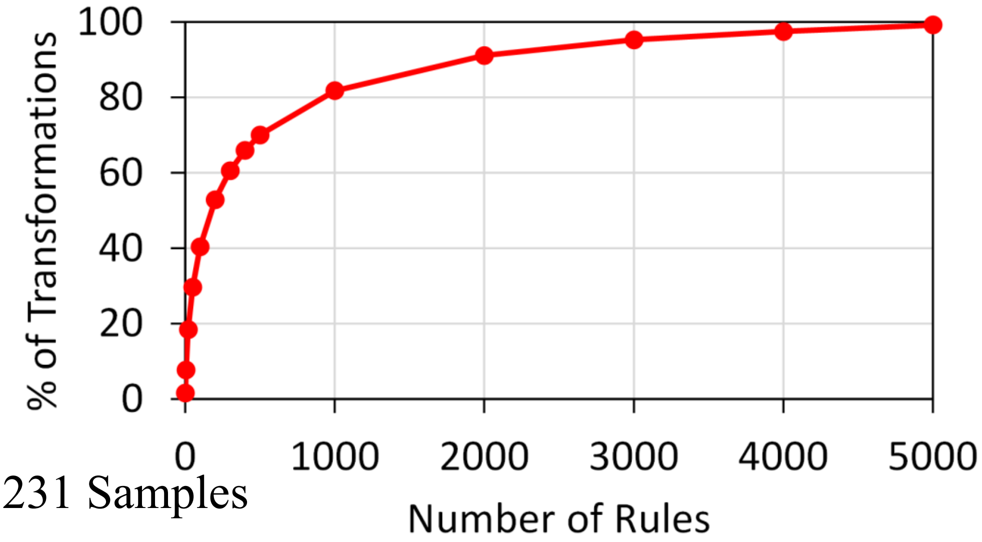
Vulnerability in the Structure: A Novel Attack Vector



- Structural changes due to logic locking is **local**.
- The **diversity** of transformation is **limited**.
- **Heavy Bias**… Can we statistically model this?



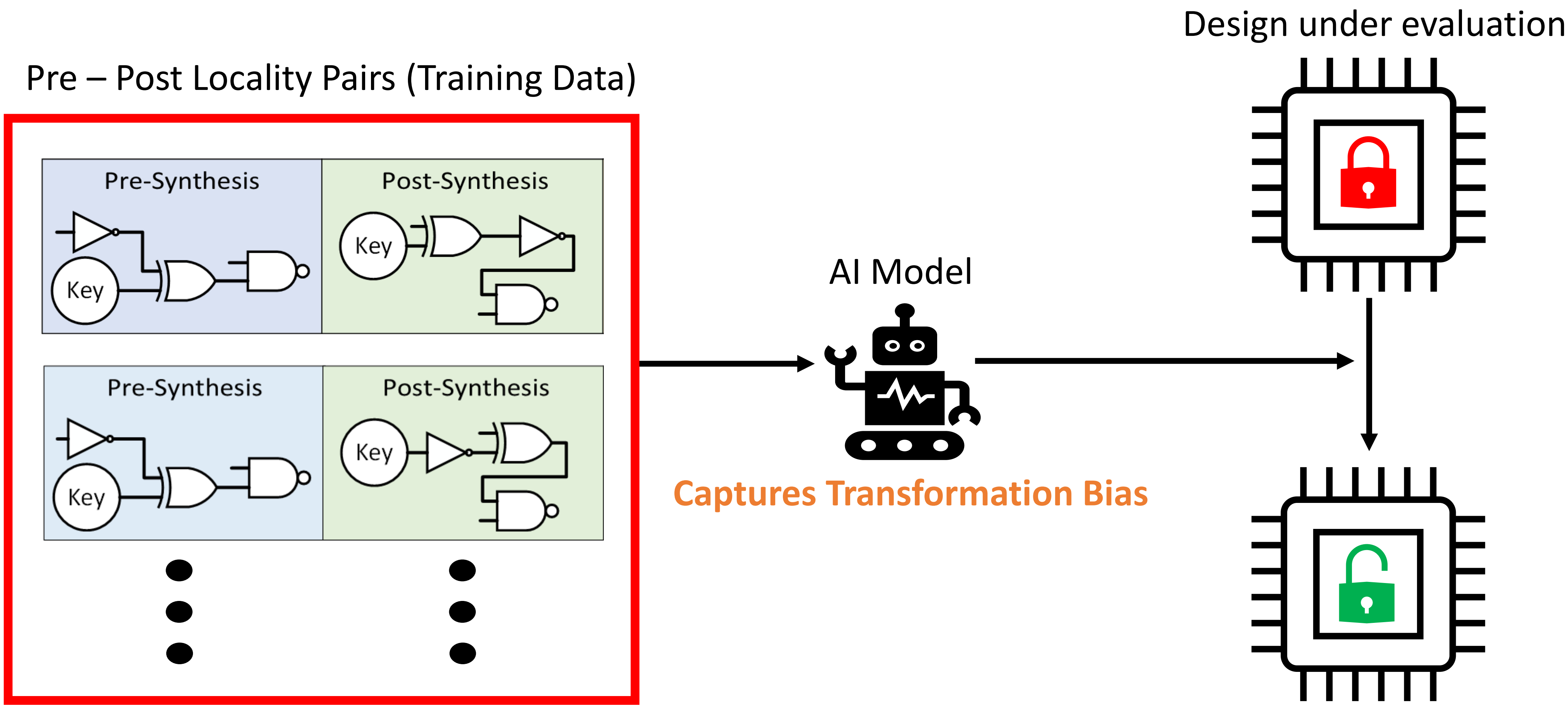
	Level - 1	Level - 2	Level - 3
C1355	26	334	0
C1908	62	292	6
C2670	96	245	19
C3540	283	1124	33
C5315	750	1950	180
C6288	516	2247	117
C7552	481	2257	142
ALU	3404	18570	1057
FIR	3376	18368	1296
Total	**8994 (15.71%)**	**45387 (79.30%)**	**2850 (4.97%)**



*57,231 Samples

Chakraborty, Prabuddha, et al. "SAIL: Analyzing structural artifacts of logic locking using machine learning." IEEE Transactions on Information Forensics and Security 16 (2021): 3828-3842.

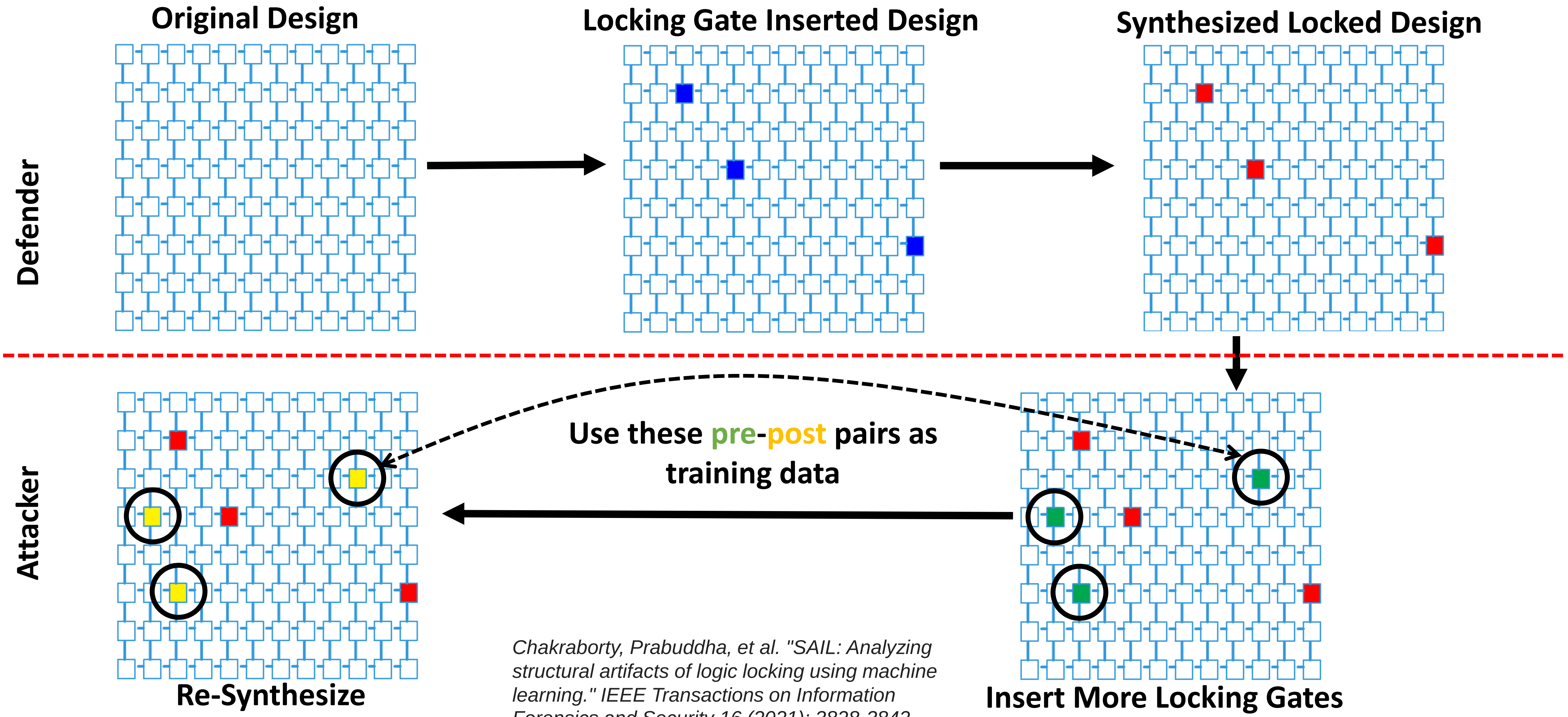
Learning the Predictable & Limited Transformations



- Structural **Unlocking**
- **Quantify** Structural Defense

Chakraborty, Prabuddha, et al. "SAIL: Analyzing structural artifacts of logic locking using machine learning." IEEE Transactions on Information Forensics and Security 16 (2021): 3828-3842.

Pseudo Self Referencing: A Golden-Free Analysis



Defender

Original Design

Locking Gate Inserted Design

Synthesized Locked Design

Attacker

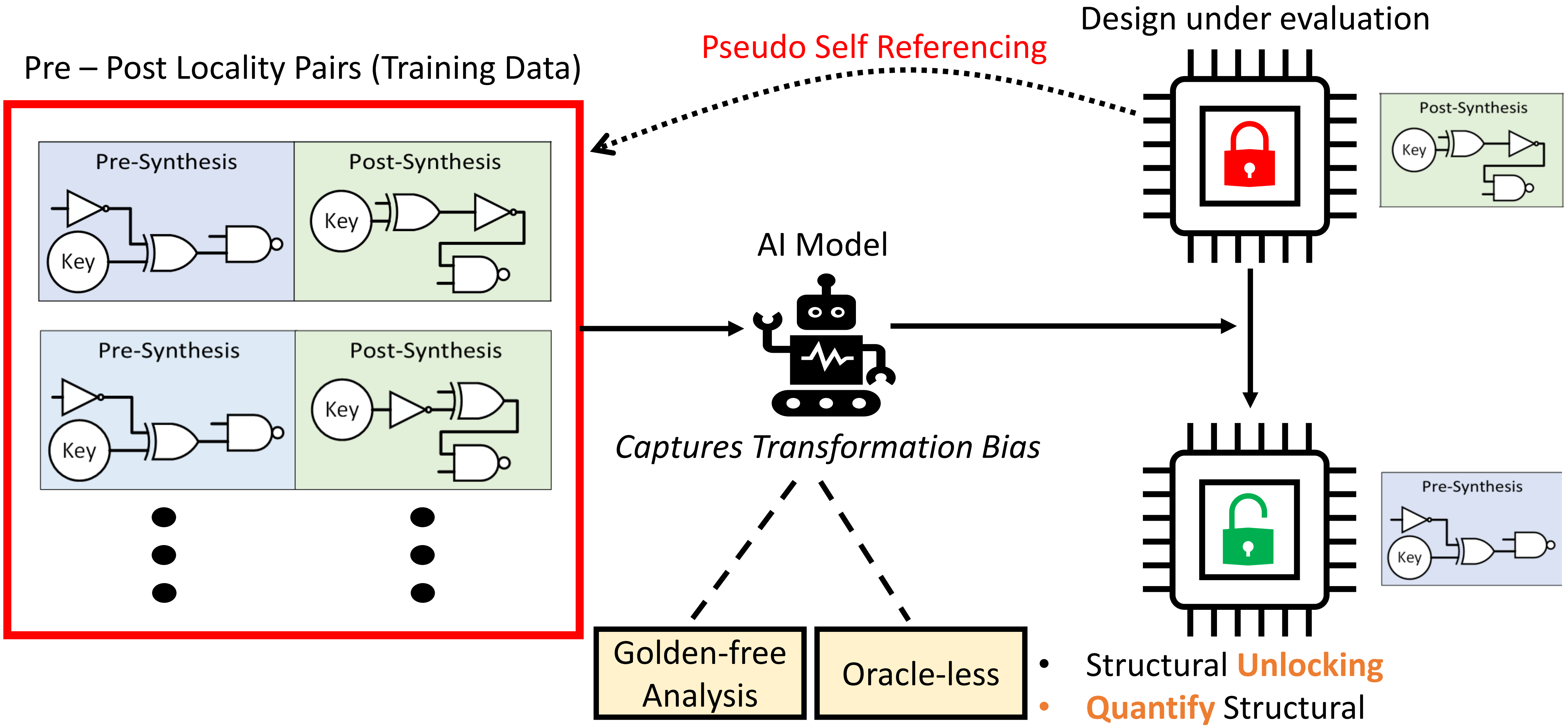
Re-Synthesize

Use these pre-post pairs as training data

Insert More Locking Gates

Chakraborty, Prabuddha, et al. "SAIL: Analyzing structural artifacts of logic locking using machine learning." IEEE Transactions on Information Forensics and Security 16 (2021): 3828-3842.

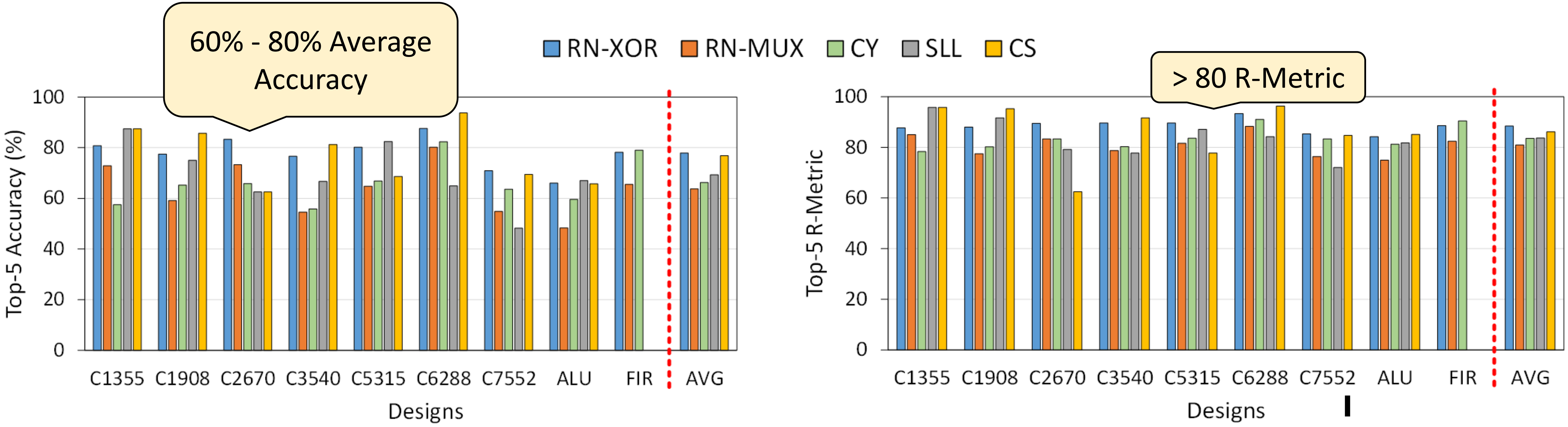
Learning the Predictable & Limited Transformations



- Structural **Unlocking**
- **Quantify** Structural Defense

Chakraborty, Prabuddha, et al. "SAIL: Analyzing structural artifacts of logic locking using machine learning." IEEE Transactions on Information Forensics and Security 16 (2021): 3828-3842.

Quantitative Analysis with SAIL



60% - 80% Average Accuracy

> 80 R-Metric

Standard Logic Locking Schemes are Structurally Vulnerable

$$R = \sum_{i=0}^{L_y} \frac{GE[i] \times 100}{T} \times \frac{L_y - i}{L_y}$$

Measure of Structural Recovery

T = Number of localities predicted during an experiment

GE[i] = The number of predicted localities with Gate Error = i and Link Error = 0

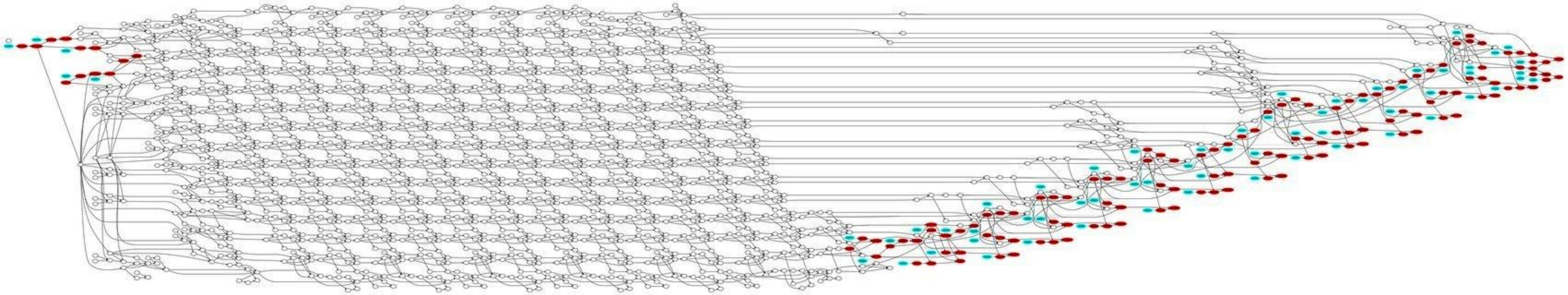
Ly	3	4	5	6
Exploration Space	5.12E+5	6.55E+5	3.35E+12	6.87E+16
SAIL-RD Avg. Top-5 Acc. (%)	77.91	60.82	41.38	29.02

Large Exploration Space

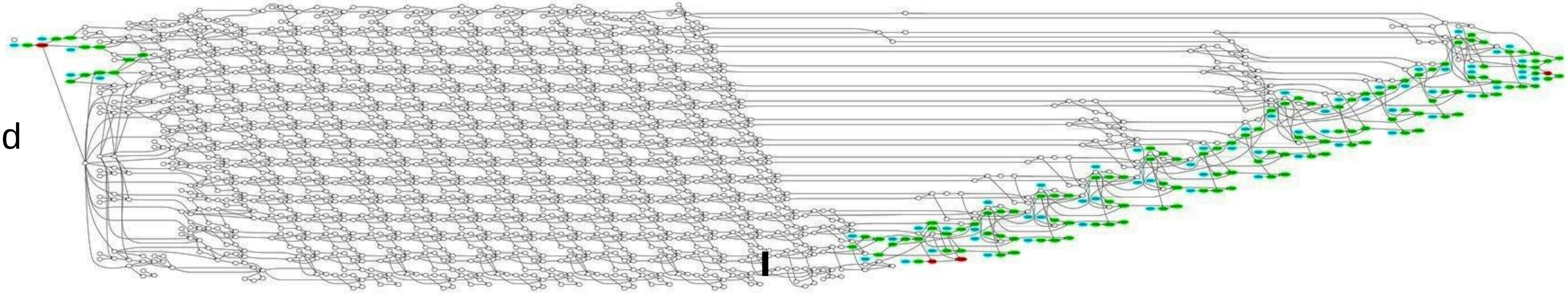
Chakraborty, Prabuddha, et al. "SAIL: Analyzing structural artifacts of logic locking using machine learning." IEEE Transactions on Information Forensics and Security 16 (2021): 3828-3842.

Quantitative Analysis with SAIL

Locked Design



Structurally Unlocked
(using SAIL)



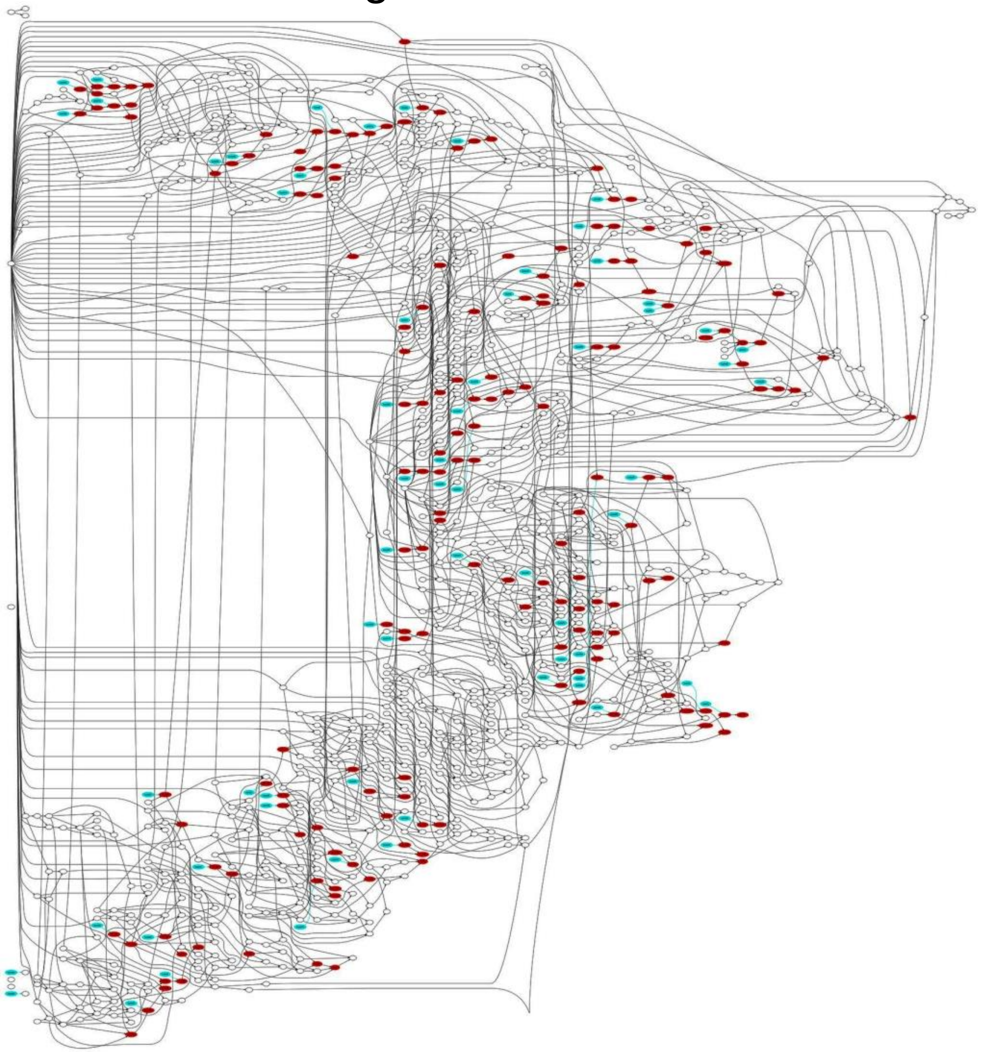
C6288:
- SAT Resistant Design
- Logic Cone Size based locking

Not SAIL Resistant!

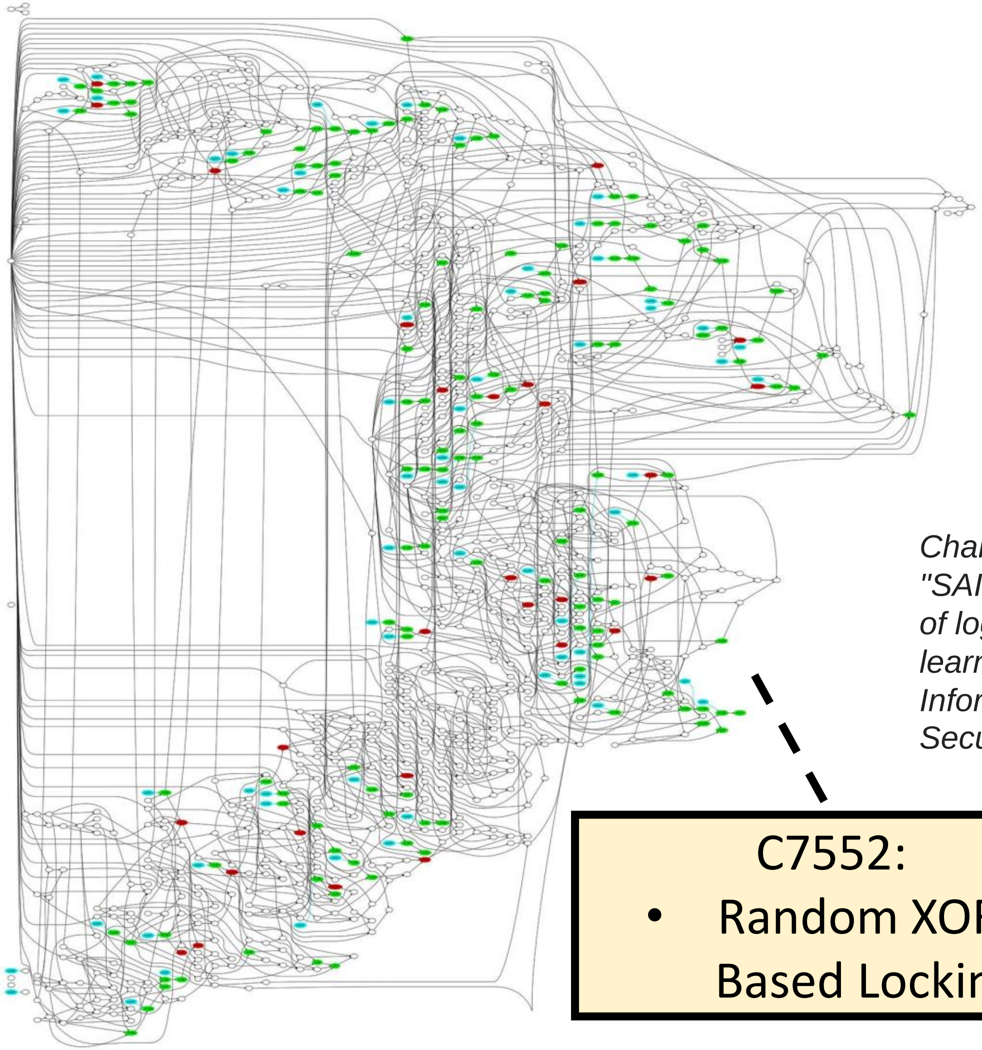
Chakraborty, Prabuddha, et al. "SAIL: Analyzing structural artifacts of logic locking using machine learning." IEEE Transactions on Information Forensics and Security 16 (2021): 3828-3842.

Quantitative Analysis with SAIL

Locked Design



Structurally Unlocked (using SAIL)



Chakraborty, Prabuddha, et al. "SAIL: Analyzing structural artifacts of logic locking using machine learning." IEEE Transactions on Information Forensics and Security 16 (2021): 3828-3842.

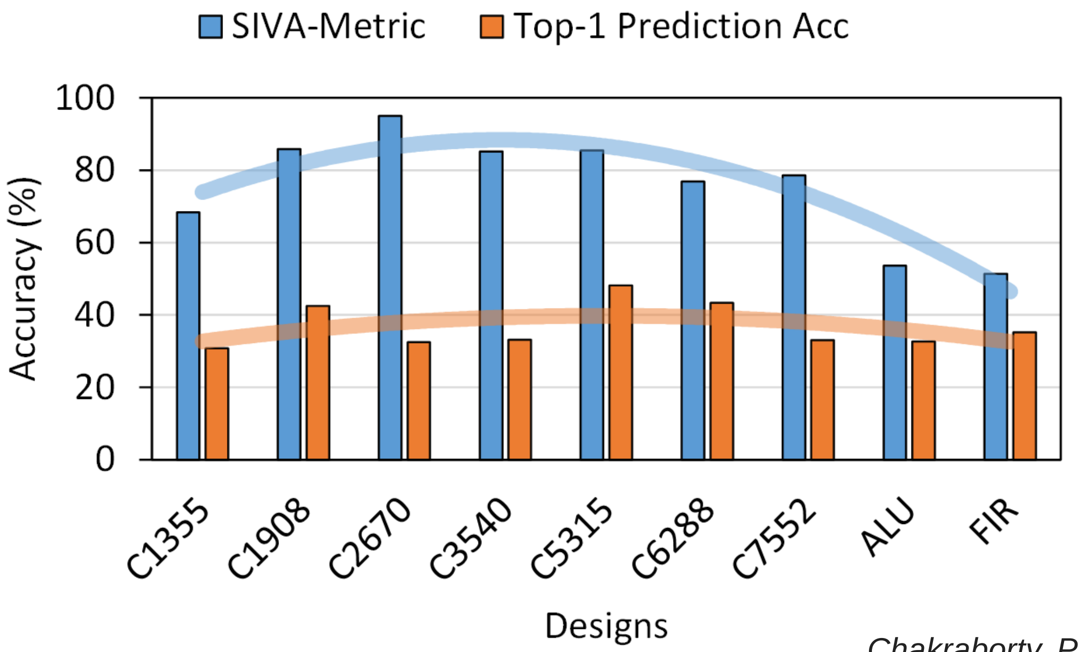
C7552:

- Random XOR-Based Locking

SIVA (**S**tructural S**i**gnature **V**ulnerability **A**nalysis) Metric

Theorem 6.1: SIVA-Metric $= (\sum_{i=1}^{n} F_i) \times \frac{100}{S}$ *implies that the SIVA-Metric is the upper bound of SAIL Accuracy*

F_i : Maximum locality recovery success for $i^{th}\ transformation$
S: Total number of localities

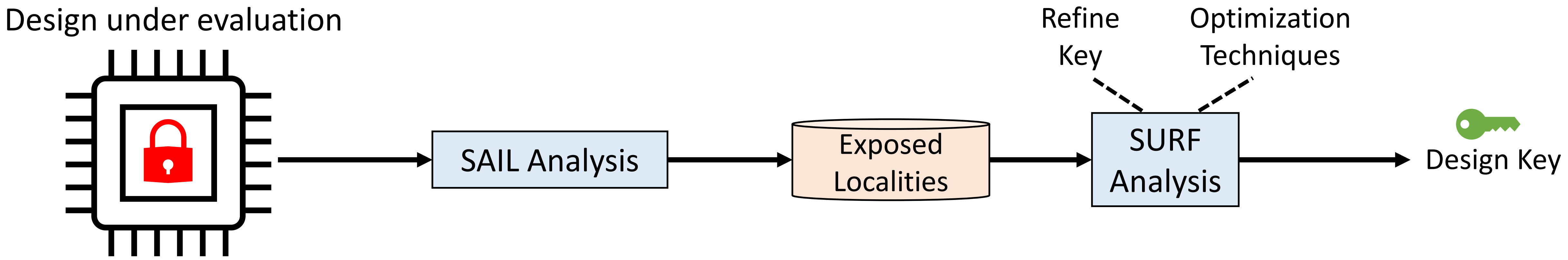


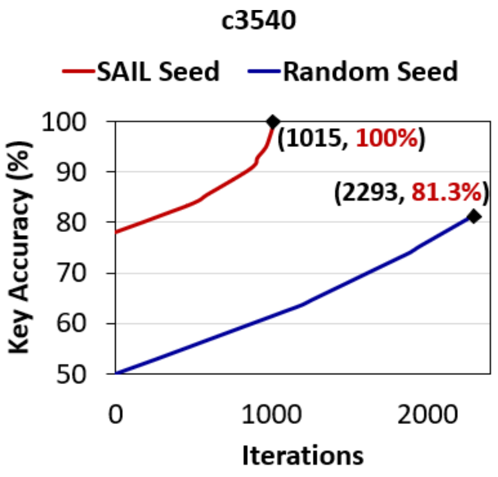
A Metric to Quantify Structural Integrity of Logic Locking

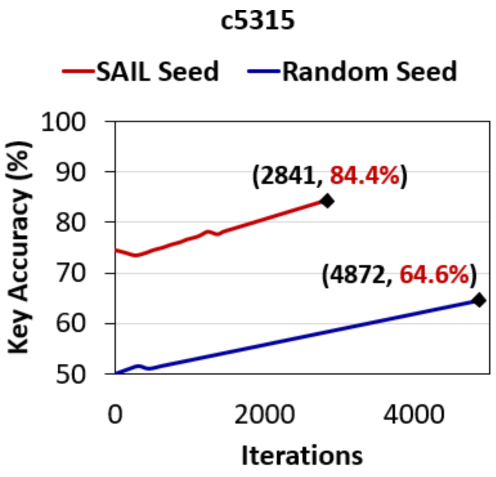
Theoretical Upper Bound for SAIL-RD

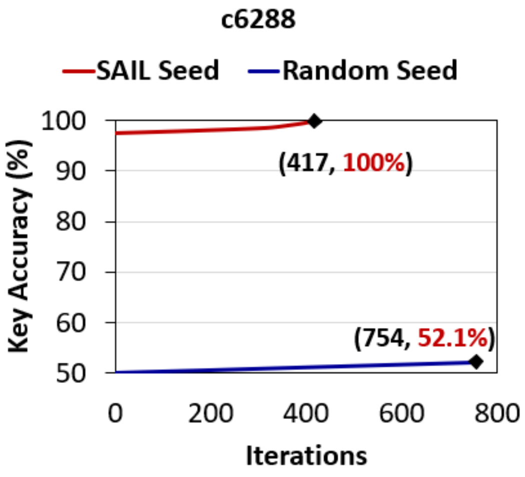
Chakraborty, Prabuddha, et al. "SAIL: Analyzing structural artifacts of logic locking using machine learning." IEEE Transactions on Information Forensics and Security 16 (2021): 3828-3842.

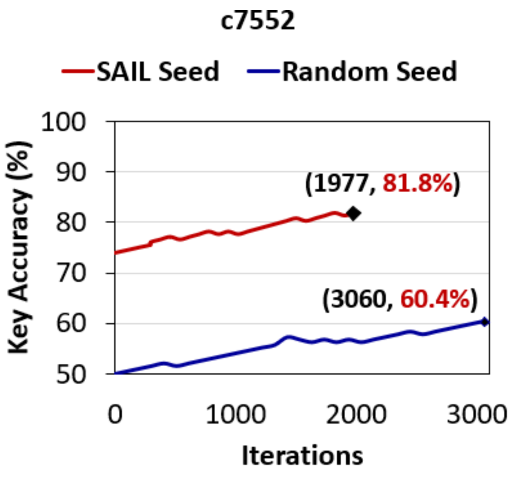
SURF: Leveraging SAIL







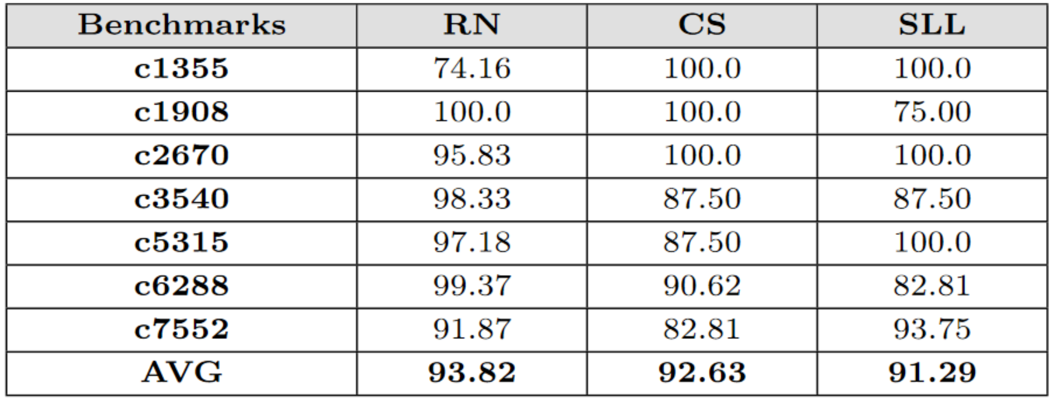




Prabuddha Chakraborty, Jonathan Cruz, and Swarup Bhunia. "SURF: Joint structural functional attack on logic locking." 2019 IEEE International Symposium on Hardware Oriented Security and Trust (HOST). IEEE, 2019.

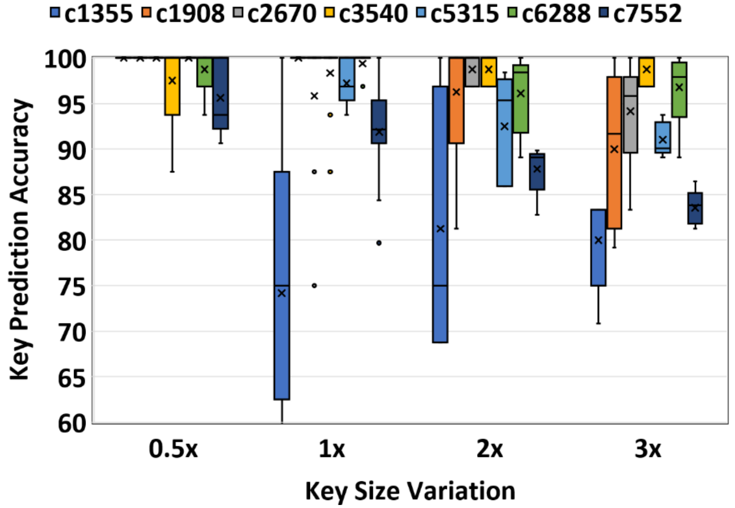
SURF: Leveraging SAIL

SURF Key Recovery Accuracy (on Average)

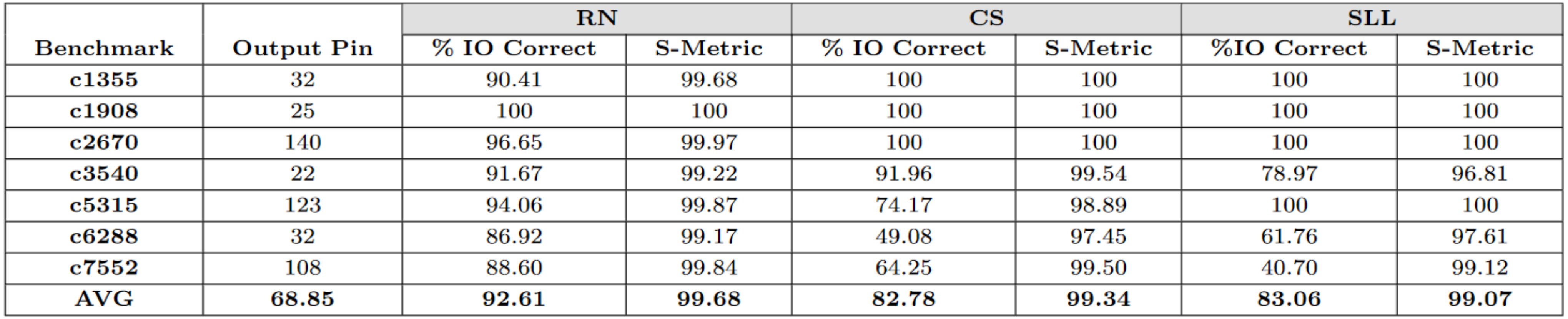


Benchmarks	RN	CS	SLL
c1355	74.16	100.0	100.0
c1908	100.0	100.0	75.00
c2670	95.83	100.0	100.0
c3540	98.33	87.50	87.50
c5315	97.18	87.50	100.0
c6288	99.37	90.62	82.81
c7552	91.87	82.81	93.75
AVG	**93.82**	**92.63**	**91.29**

SURF Key Recovery Accuracy Distribution



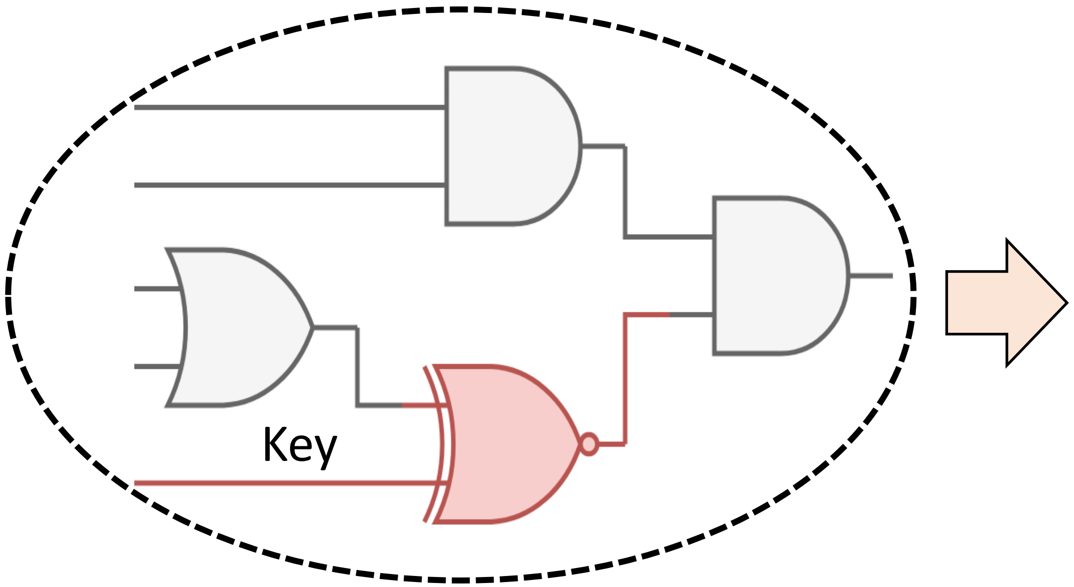
Usefulness of Partial Unlocking



Benchmark	Output Pin	RN		CS		SLL	
		% IO Correct	S-Metric	% IO Correct	S-Metric	%IO Correct	S-Metric
c1355	32	90.41	99.68	100	100	100	100
c1908	25	100	100	100	100	100	100
c2670	140	96.65	99.97	100	100	100	100
c3540	22	91.67	99.22	91.96	99.54	78.97	96.81
c5315	123	94.06	99.87	74.17	98.89	100	100
c6288	32	86.92	99.17	49.08	97.45	61.76	97.61
c7552	108	88.60	99.84	64.25	99.50	40.70	99.12
AVG	**68.85**	**92.61**	**99.68**	**82.78**	**99.34**	**83.06**	**99.07**

Prabuddha Chakraborty, Jonathan Cruz, and Swarup Bhunia. "SURF: Joint structural functional attack on logic locking." 2019 IEEE International Symposium on Hardware Oriented Security and Trust (HOST). IEEE, 2019.

Metrics of Logic Locking



SAIL → Structural Defense Metric

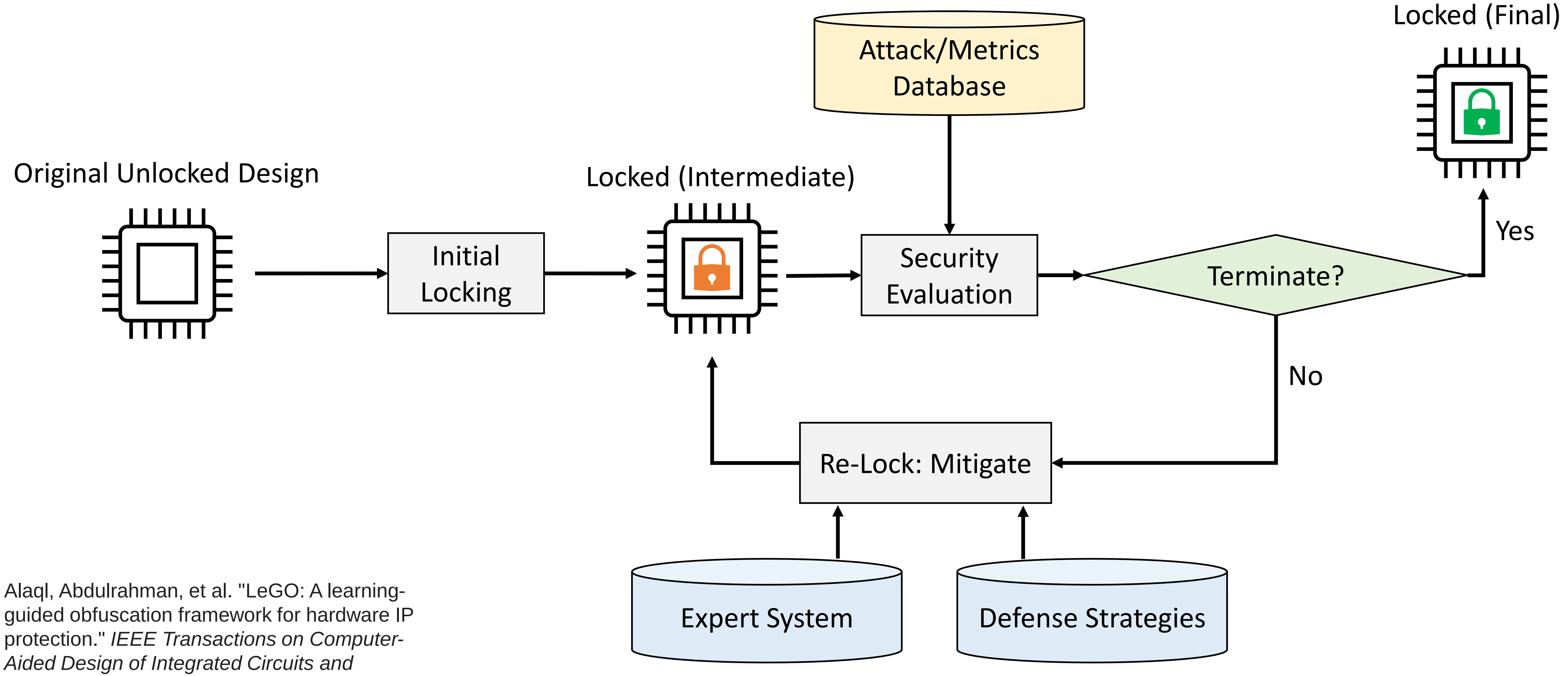
SURF → Structural + Functional Defense Metric

SAT-Attack → Functional Defense Metric

SWEEP → Structural + Functional Defense Metric

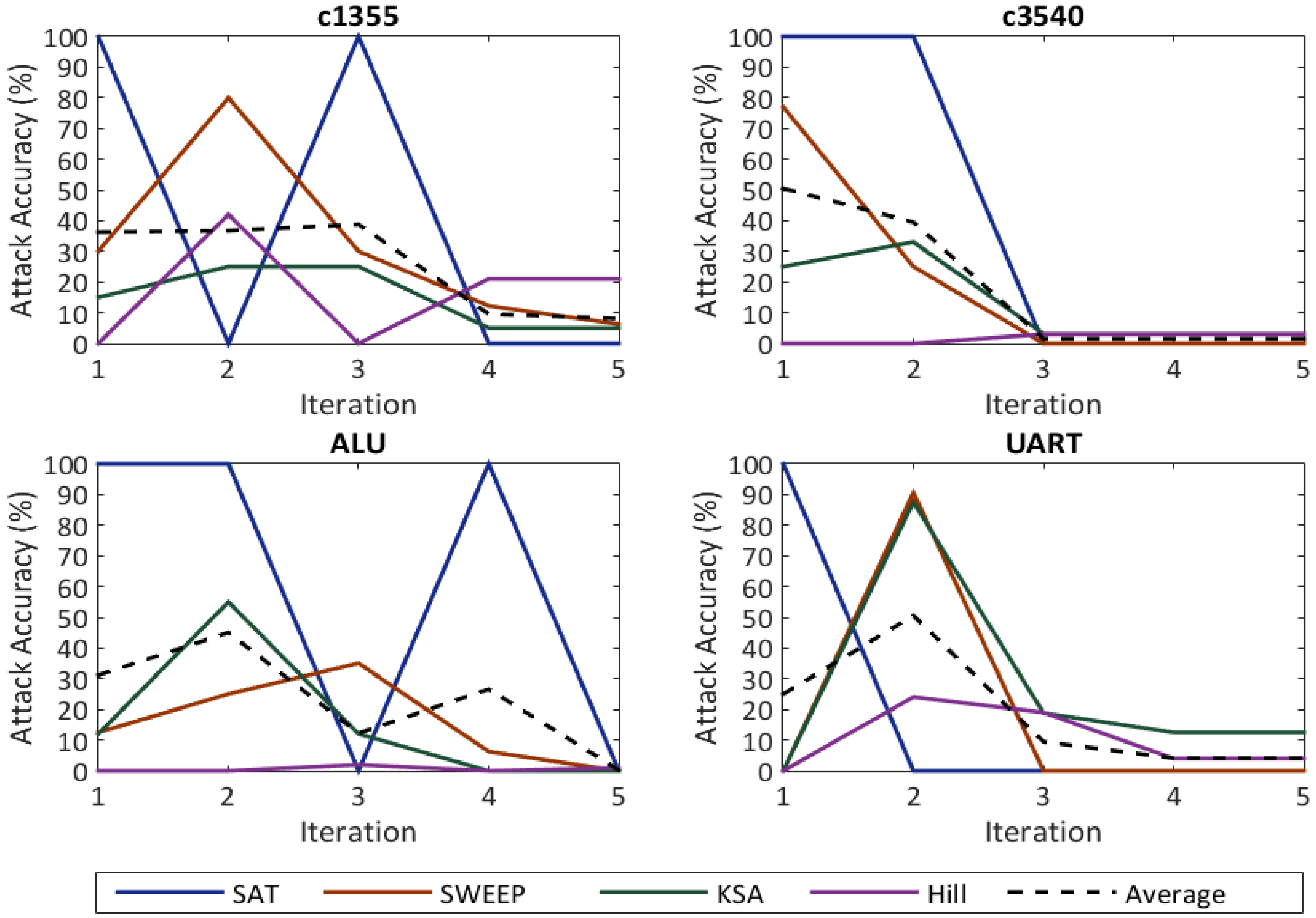
SIVA → Structural Defense Metric

LeGO: **Le**arning-**G**uided Logic L**O**cking



Alaql, Abdulrahman, et al. "LeGO: A learning-guided obfuscation framework for hardware IP protection." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 41.4 (2021): 854-867.

LeGO: Results



Fast: Rapid Convergence

Scalable: Incorporate New Attacks

Progressive: Requirement-Based

Alaql, Abdulrahman, et al. "LeGO: A learning-guided obfuscation framework for hardware IP protection." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 41.4 (2021): 854-867.

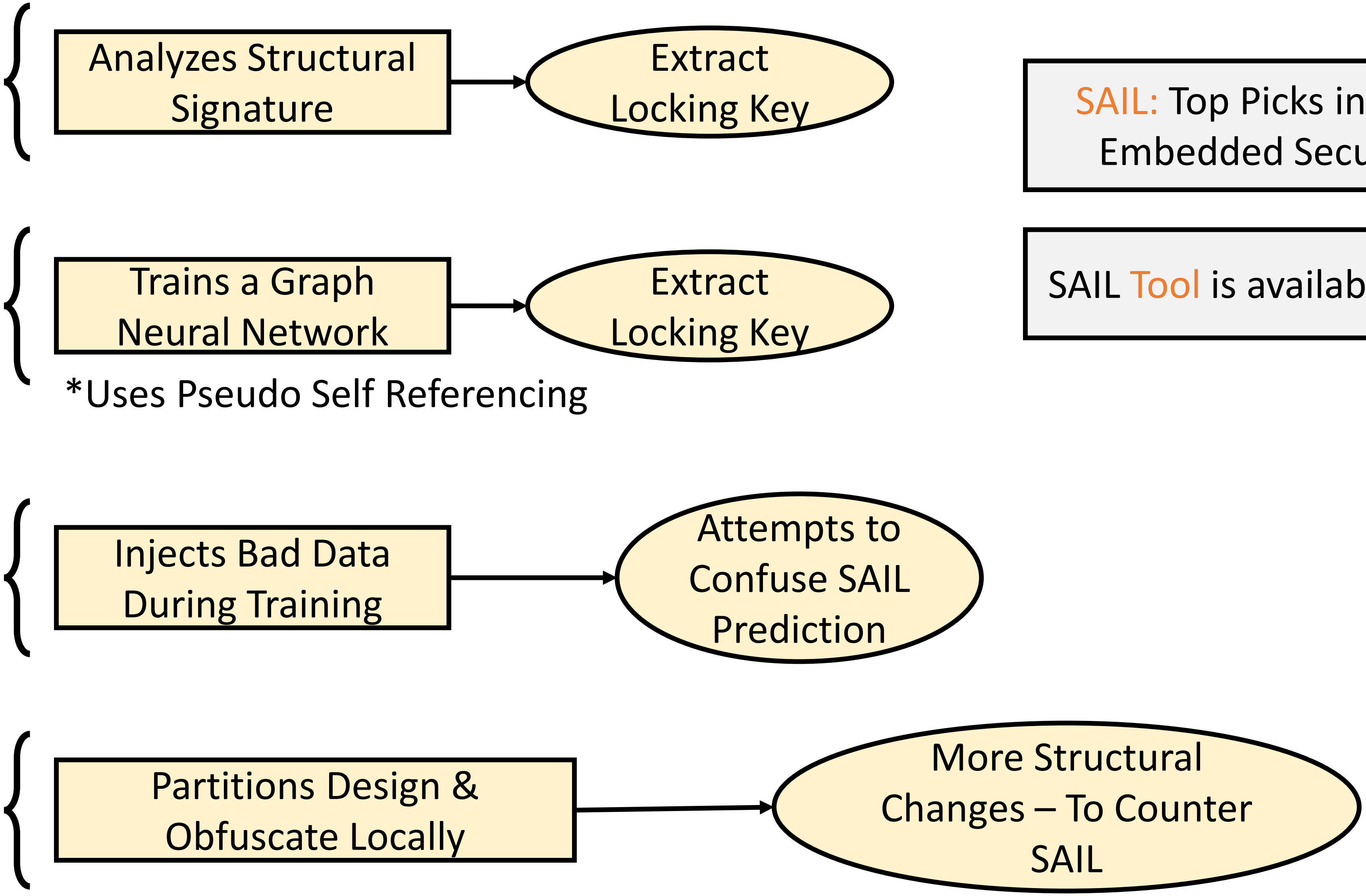
A Novel Attack Vector: Inspired Follow-up Works

- Snapshot [5]
- SWEEP [6]

- OMLA [7]

- UNSAIL [8]

- LeGO [9]
- SARO [10]

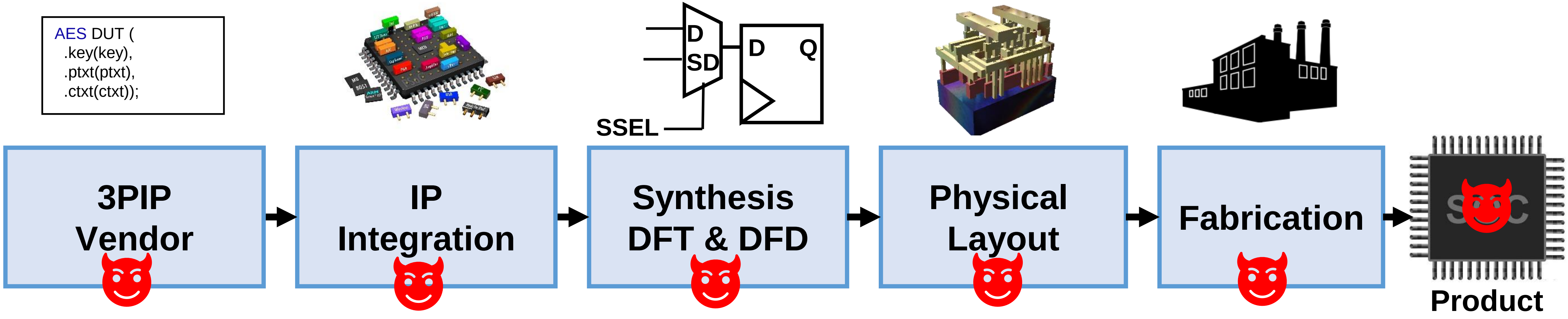


Analyzes Structural Signature
Extract Locking Key
Trains a Graph Neural Network
Extract Locking Key
*Uses Pseudo Self Referencing
Injects Bad Data During Training
Attempts to Confuse SAIL Prediction
Partitions Design & Obfuscate Locally
More Structural Changes – To Counter SAIL

SAIL: Top Picks in Hardware and Embedded Security (Winner)

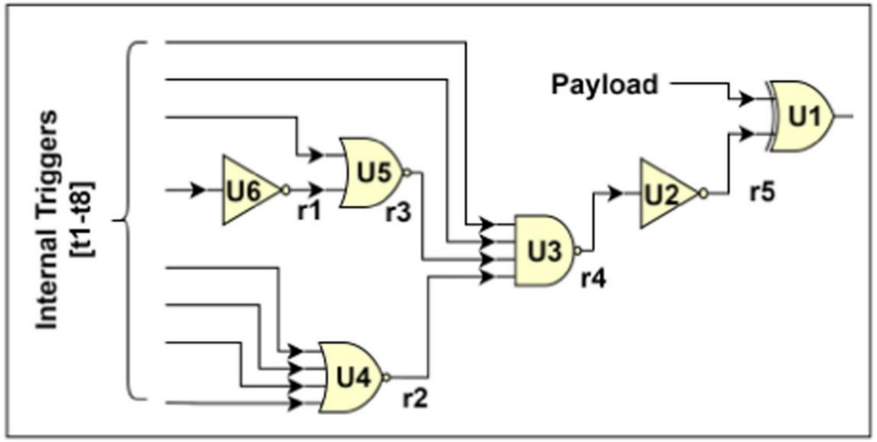
SAIL Tool is available upon Request

Hardware IP/IC Threats



```
AES DUT (
  .key(key),
  .ptxt(ptxt),
  .ctxt(ctxt));
```

Hardware Trojans can get inserted throughout the supply chain

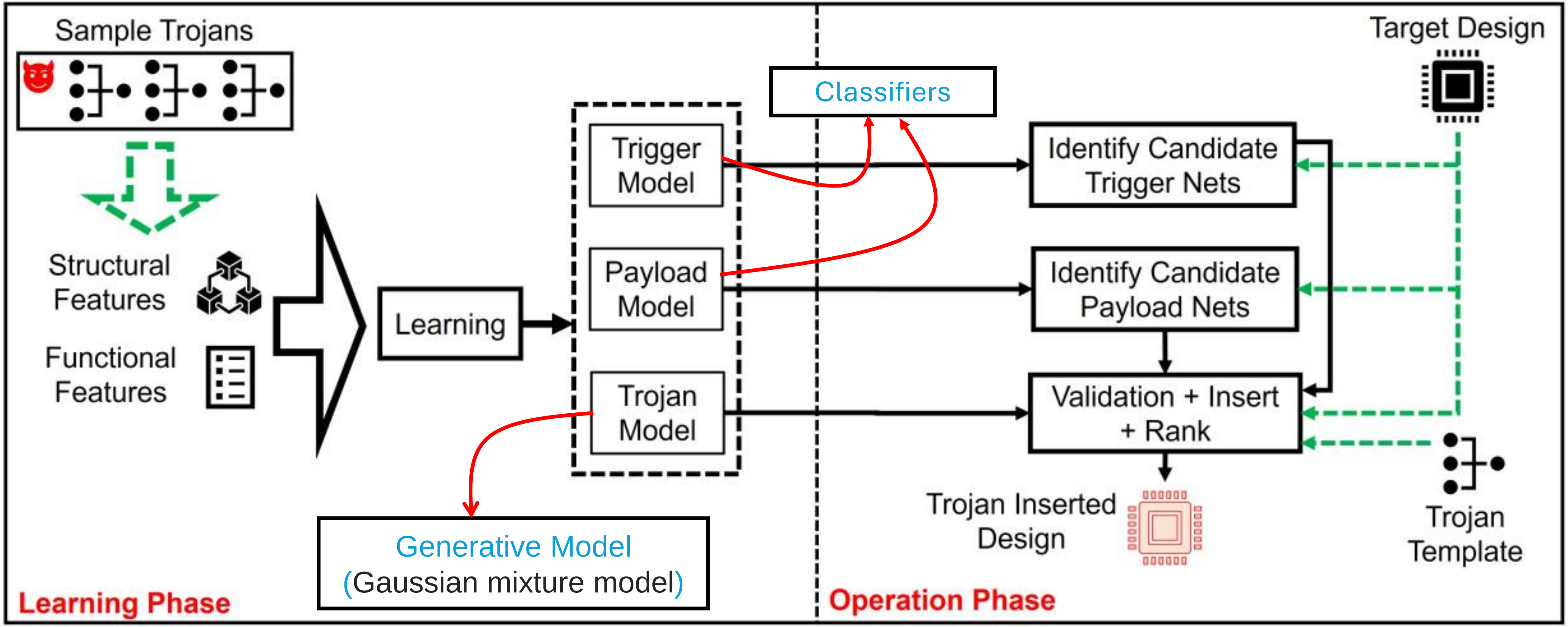


(a) 8-triggered combinational Trojan in RS232 design

- **Hardware Trojans:** Malicious modifications made in the hardware design/IC
- **Challenges** with Detecting Hardware Trojans:
  1. Lack of datasets → Limited understanding
  2. Reliance on static defense → Easy to bypass

Cruz, Jonathan, et al. "A machine learning based automatic hardware trojan attack space exploration and benchmarking framework." *2022 Asian Hardware Oriented Security and Trust Symposium (AsianHOST)*. IEEE, 2022.

MIMIC Flow



Cruz, Jonathan, et al. "A machine learning based automatic hardware trojan attack space exploration and benchmarking framework." *2022 Asian Hardware Oriented Security and Trust Symposium (AsianHOST)*. IEEE, 2022.

MIMIC Results

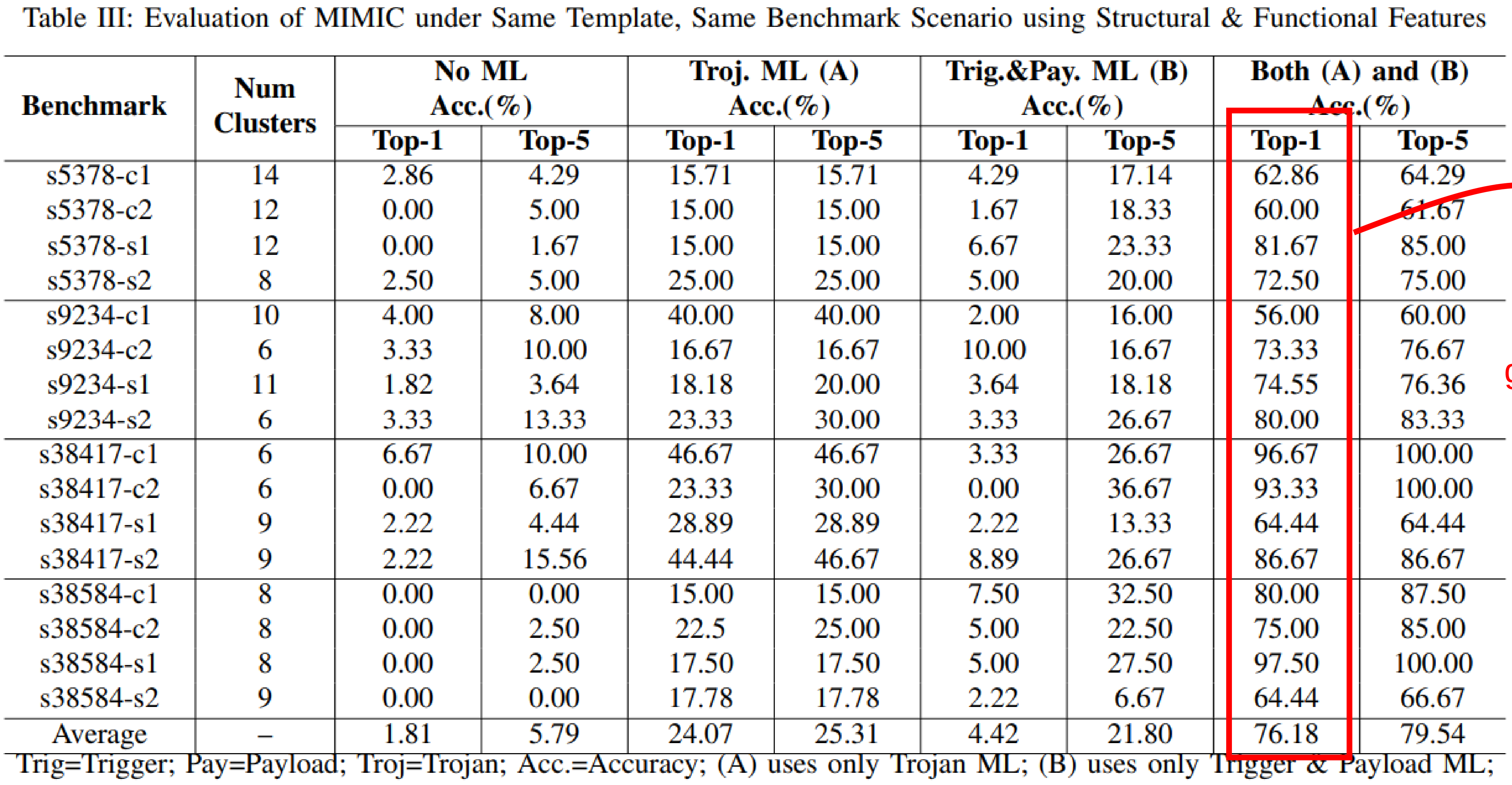


Table III: Evaluation of MIMIC under Same Template, Same Benchmark Scenario using Structural & Functional Features

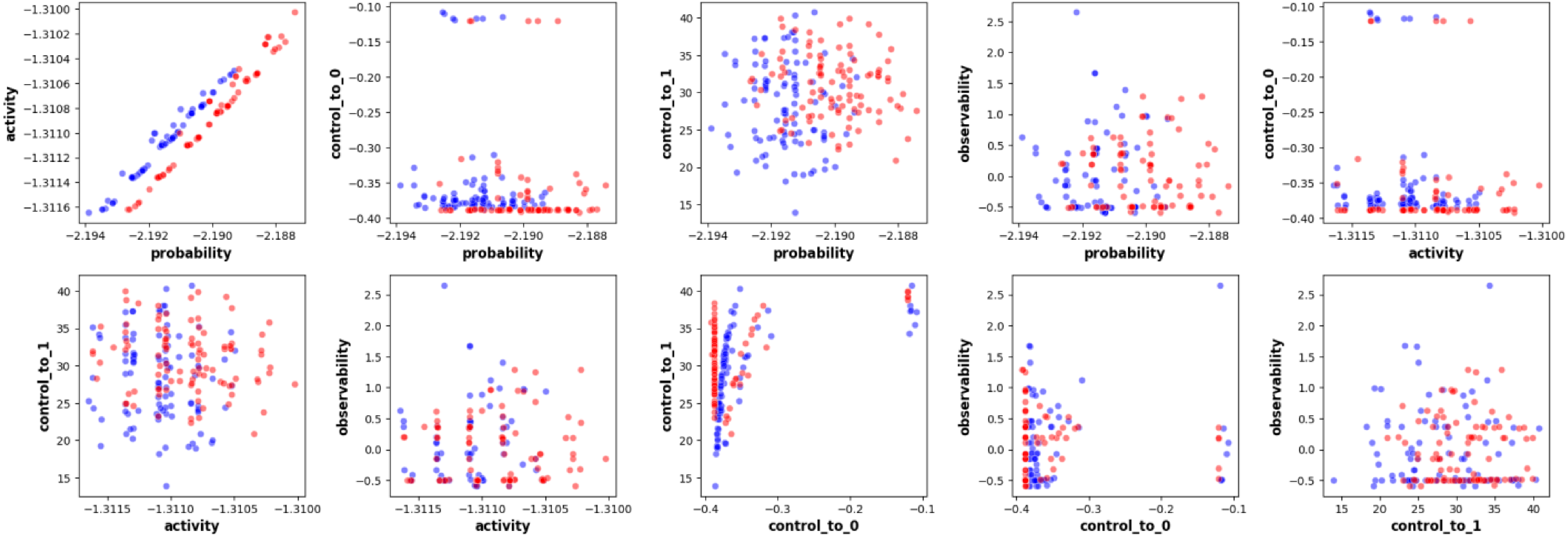
Benchmark	Num Clusters	No ML Acc.(%)		Troj. ML (A) Acc.(%)		Trig.&Pay. ML (B) Acc.(%)		Both (A) and (B) Acc.(%)	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
s5378-c1	14	2.86	4.29	15.71	15.71	4.29	17.14	62.86	64.29
s5378-c2	12	0.00	5.00	15.00	15.00	1.67	18.33	60.00	61.67
s5378-s1	12	0.00	1.67	15.00	15.00	6.67	23.33	81.67	85.00
s5378-s2	8	2.50	5.00	25.00	25.00	5.00	20.00	72.50	75.00
s9234-c1	10	4.00	8.00	40.00	40.00	2.00	16.00	56.00	60.00
s9234-c2	6	3.33	10.00	16.67	16.67	10.00	16.67	73.33	76.67
s9234-s1	11	1.82	3.64	18.18	20.00	3.64	18.18	74.55	76.36
s9234-s2	6	3.33	13.33	23.33	30.00	3.33	26.67	80.00	83.33
s38417-c1	6	6.67	10.00	46.67	46.67	3.33	26.67	96.67	100.00
s38417-c2	6	0.00	6.67	23.33	30.00	0.00	36.67	93.33	100.00
s38417-s1	9	2.22	4.44	28.89	28.89	2.22	13.33	64.44	64.44
s38417-s2	9	2.22	15.56	44.44	46.67	8.89	26.67	86.67	86.67
s38584-c1	8	0.00	0.00	15.00	15.00	7.50	32.50	80.00	87.50
s38584-c2	8	0.00	2.50	22.5	25.00	5.00	22.50	75.00	85.00
s38584-s1	8	0.00	2.50	17.50	17.50	5.00	27.50	97.50	100.00
s38584-s2	9	0.00	0.00	17.78	17.78	2.22	6.67	64.44	66.67
Average	–	1.81	5.79	24.07	25.31	4.42	21.80	76.18	79.54

Trig=Trigger; Pay=Payload; Troj=Trojan; Acc.=Accuracy; (A) uses only Trojan ML; (B) uses only Trigger & Payload ML;

Accurately generate valid & potent Trojans

Cruz, Jonathan, et al. "A machine learning based automatic hardware trojan attack space exploration and benchmarking framework." *2022 Asian Hardware Oriented Security and Trust Symposium (AsianHOST)*. IEEE, 2022.

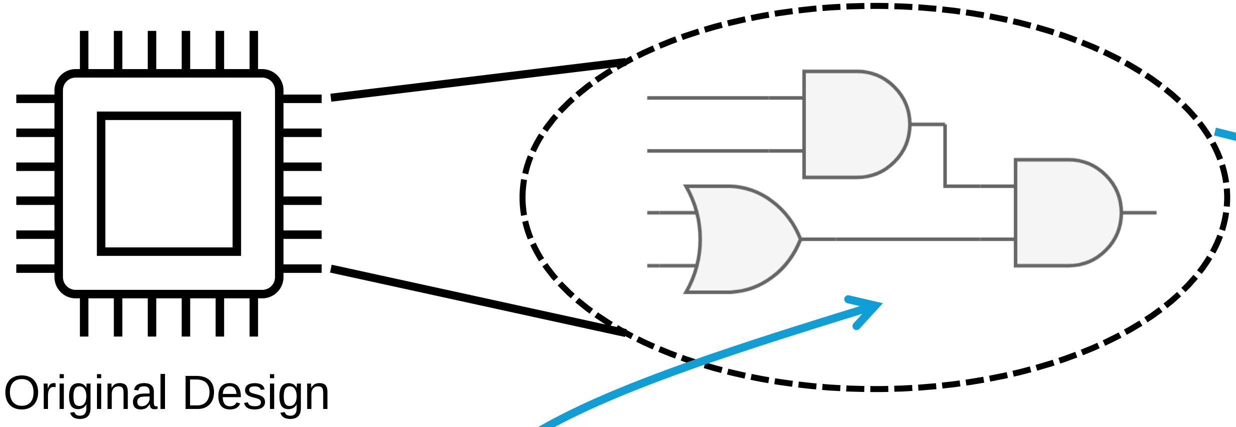
MIMIC Results



- Trojans are similar (to the training/potent Trojan population). Yet different!

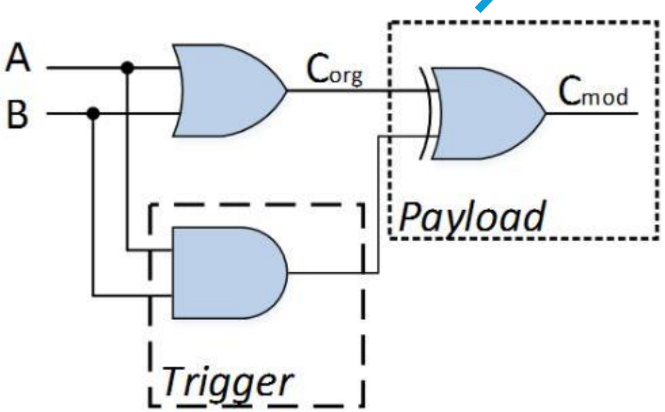
Cruz, Jonathan, et al. "A machine learning based automatic hardware trojan attack space exploration and benchmarking framework." 2022 Asian Hardware Oriented Security and Trust Symposium (AsianHOST). IEEE, 2022.

VIPR: Joint Structural-Functional Learning to Detect Trojans



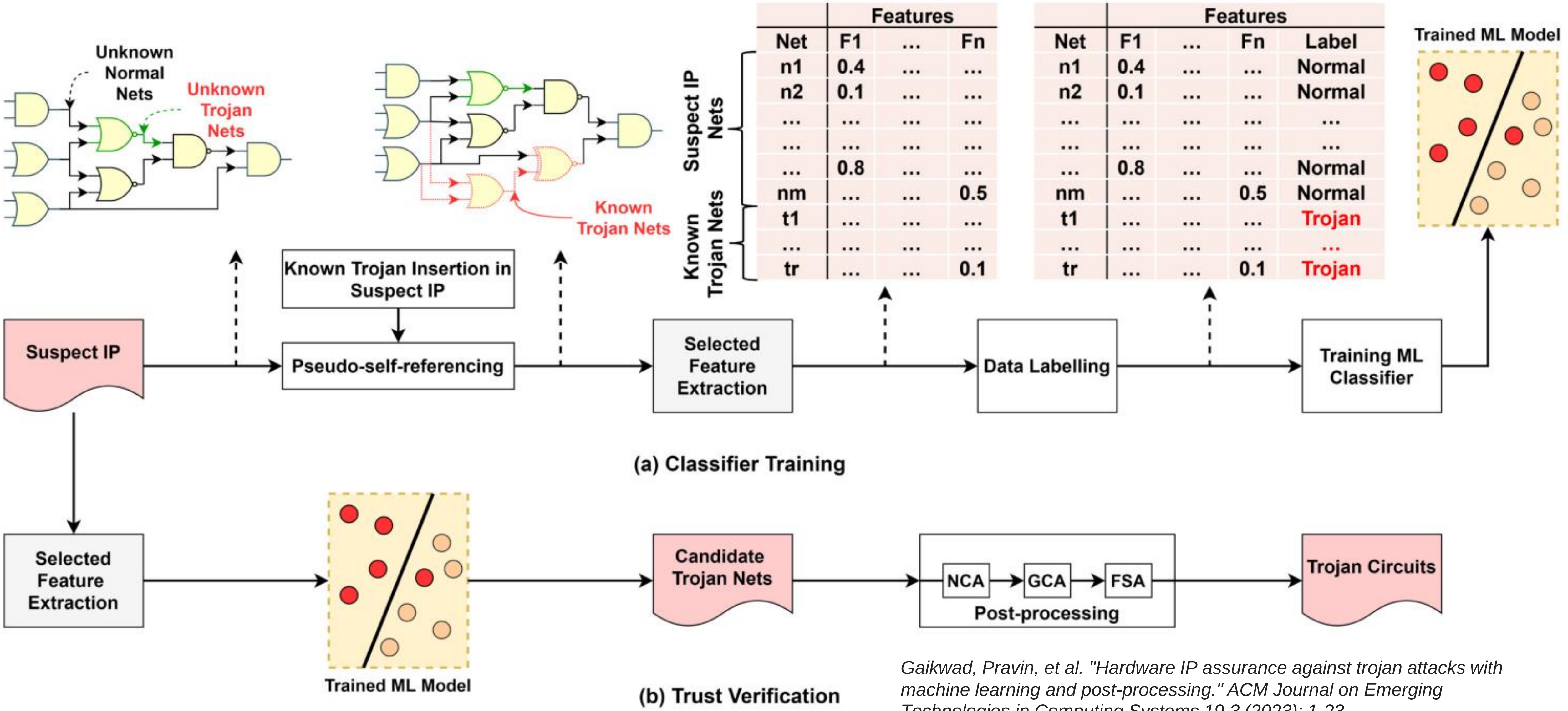
- Extract features of every wire/net
- Classify

Gaikwad, Pravin, et al. "Hardware IP assurance against trojan attacks with machine learning and post-processing." ACM Journal on Emerging Technologies in Computing Systems 19.3 (2023): 1-23.



#	Functional Feature	Description
1	Static Probability	Static probability of the net.
2	Transition Probability	Activity from 0 to 1.
3	Controllability	Controllability of the net.
4	Observability	Observability of the net.
5	Fanin Level 1	# of connected inputs at level 1
6	Fanout Level 1	# of connected outputs at level 1
7	Fanin Level 2	# of connected inputs at level 2
8	Fanout Level 2	# of connected outputs at level 2
9	Nearest_FF_D	Distance of the nearest flip-flop input
10	Nearest_FF_Q	Distance of the nearest flip-flop output
11	Min. PI Distance	Min. distance from nearest primary input
12	Min. PO Distance	Min. distance from nearest primary output

VIPR Flow



		Features		
	Net	F1	...	Fn
Suspect IP Nets	n1	0.4	...	...
	n2	0.1	...	...
	...	...	...	...
	...	...	...	...
	...	0.8	...	...
	nm	...	...	0.5
Known Trojan Nets	t1	...	...	...
	...	...	...	...
	tr	...	...	0.1

	Features			
Net	F1	...	Fn	Label
n1	0.4	...	...	Normal
n2	0.1	...	...	Normal
...	...	...	...	...
...	...	...	...	...
...	0.8	...	...	Normal
nm	...	...	0.5	Normal
t1	...	...	...	Trojan
...	...	...	...	...
tr	...	...	0.1	Trojan

(a) Classifier Training

(b) Trust Verification

Gaikwad, Pravin, et al. "Hardware IP assurance against trojan attacks with machine learning and post-processing." ACM Journal on Emerging Technologies in Computing Systems 19.3 (2023): 1-23.

VIPR Post-Processing Algorithms

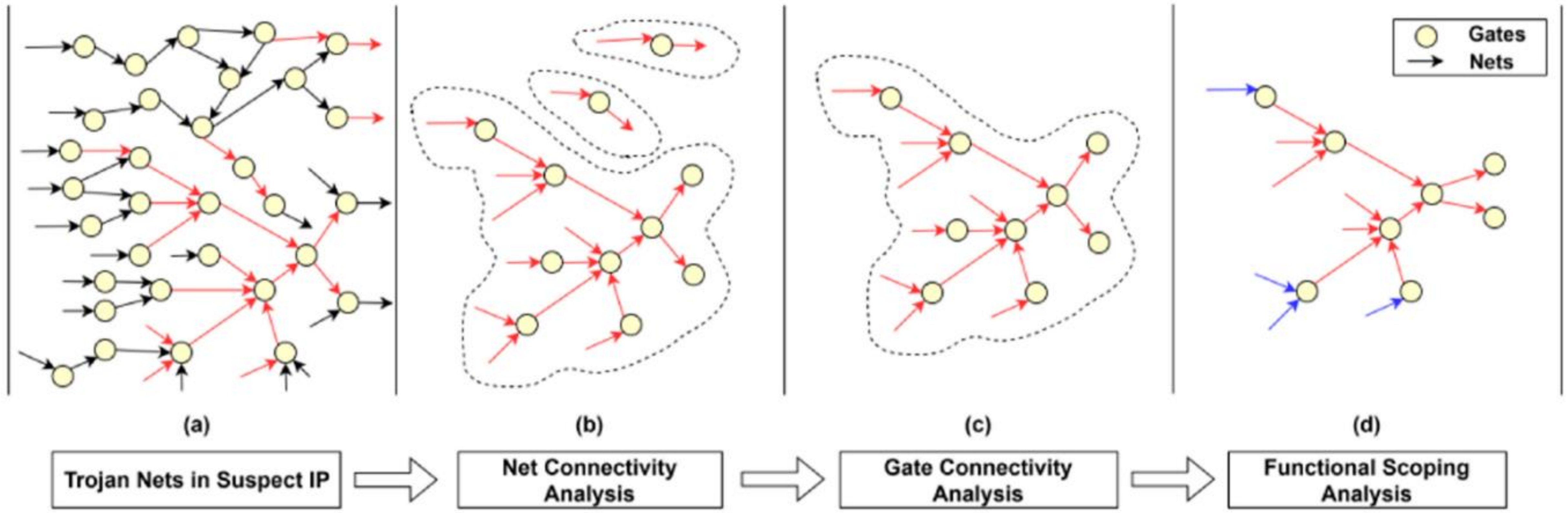
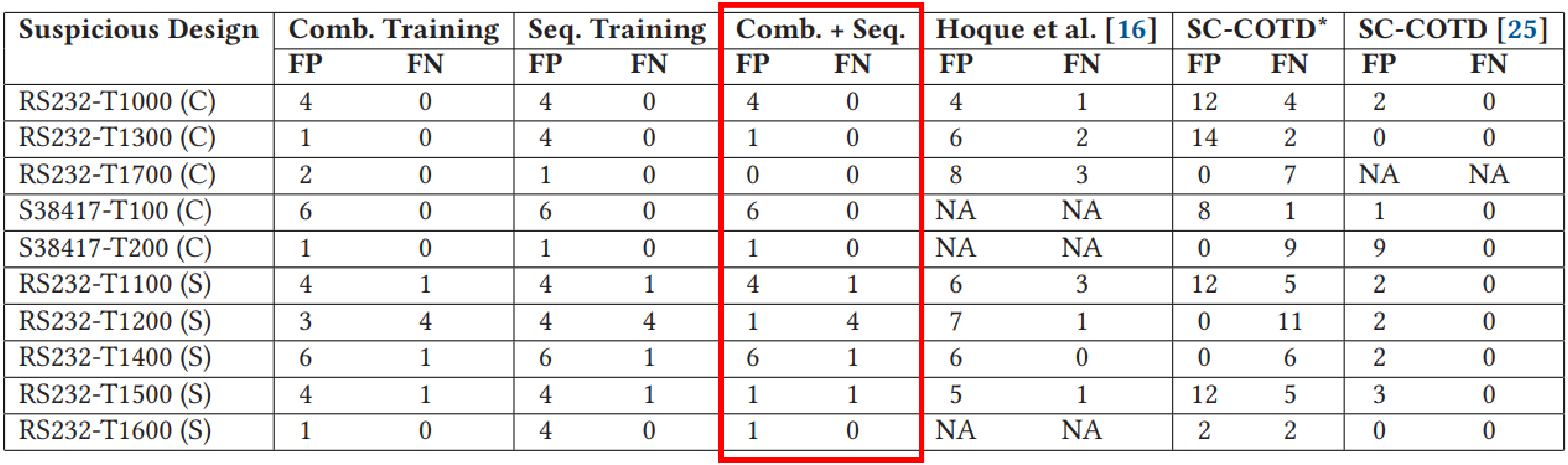

Gates
Nets
(a)
(b)
(c)
(d)
Trojan Nets in Suspect IP
Net Connectivity Analysis
Gate Connectivity Analysis
Functional Scoping Analysis

Fig. 6. Circuit reconstruction with the proposed post-processing algorithms. Nets highlighted in red color represent predictions from the ML model. Specifically for the last section, nets highlighted in blue are false-positive nets, and those highlighted in red are true-positive nets.

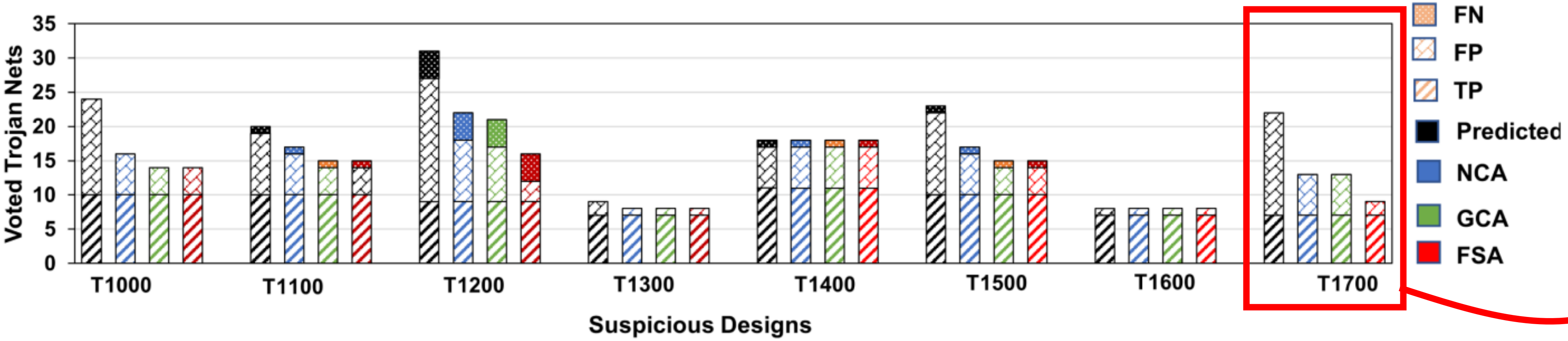
Gaikwad, Pravin, et al. "Hardware IP assurance against trojan attacks with machine learning and post-processing." ACM Journal on Emerging Technologies in Computing Systems 19.3 (2023): 1-23.

VIPR Results



Suspicious Design	Comb. Training		Seq. Training		Comb. + Seq.		Hoque et al. [16]		SC-COTD*		SC-COTD [25]	
	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
RS232-T1000 (C)	4	0	4	0	4	0	4	1	12	4	2	0
RS232-T1300 (C)	1	0	4	0	1	0	6	2	14	2	0	0
RS232-T1700 (C)	2	0	1	0	0	0	8	3	0	7	NA	NA
S38417-T100 (C)	6	0	6	0	6	0	NA	NA	8	1	1	0
S38417-T200 (C)	1	0	1	0	1	0	NA	NA	0	9	9	0
RS232-T1100 (S)	4	1	4	1	4	1	6	3	12	5	2	0
RS232-T1200 (S)	3	4	4	4	1	4	7	1	0	11	2	0
RS232-T1400 (S)	6	1	6	1	6	1	6	0	0	6	2	0
RS232-T1500 (S)	4	1	4	1	1	1	5	1	12	5	3	0
RS232-T1600 (S)	1	0	4	0	1	0	NA	NA	2	2	0	0

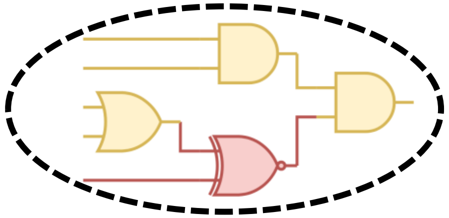
Low FP and Low FN

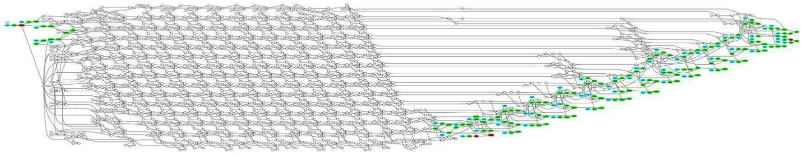


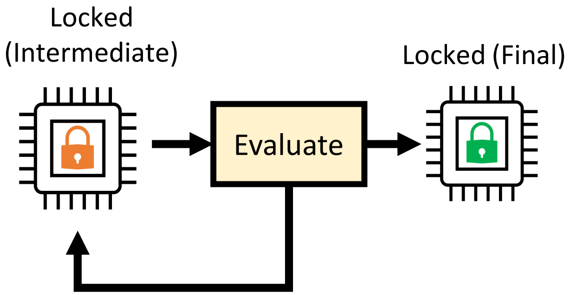
Progressive decrease in False Positives (FP)

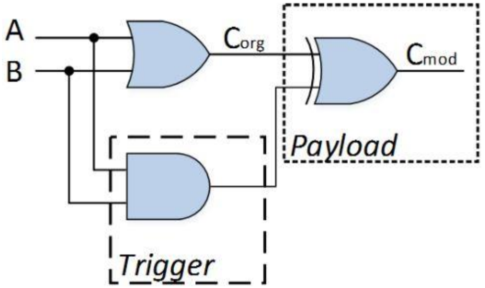
Gaikwad, Pravin, et al. "Hardware IP assurance against trojan attacks with machine learning and post-processing." ACM Journal on Emerging Technologies in Computing Systems 19.3 (2023): 1-23.

Summary & Future Works









- Designing secure hardware is challenging
- Logic locking can be a solution but has major pitfalls
  - **SAIL**: Structural attack on logic locking
  - **SURF**: Leveraging recovered structural artifacts to find key
  - **LeGO**: Learning-guided iterative locking scheme
- Hardware Trojans can have devastating impact
  - **MIMIC**: AI-guided hardware Trojan exploration
  - **VIPR**: AI-guided hardware Trojan Detection
- Significant future research possible building on these work

1865 THE UNIVERSITY OF MAINE

SIEGE Secure and Intelligent Edge Research

Thank You!

Your

Thoughts?

Prabuddha Chakraborty

(prabuddha@maine.edu)