High Dimensional Yield Estimation using Shrinkage Deep Features and Maximization of Integral Entropy Reduction

Shuo Yin, Guohao Dai, Wei W. Xing†

January 13, 2023

School of Integrated Circuit Science and Engineering,
Beihang University

Introduction

Background

- As semiconductor fabrication technology improves by shrinking down its scale to nanometer, the negative effect of the process variance will cause yield reduction.
- For the SRAM array, the failure rate of each bitcell should be lower than 10^{-6} in order to ensure the quality of the SRAM array.
- Monte Carlo (MC) analysis is generally considered the gold standard for yield estimation in industry and academia. However, MC requires a large number (usually millions) of SPICE simulations, which will be time-consuming.

Importance Sampling Method

Importance Sampling Method

Importance sampling (IS) based approaches draw samples according to a constructed distribution shifted to the likely-to-fail regions.

Importance Sampling Method

- [TCAD'10][1]
- [ISPD'16][2]
- [DAC'18][3]

---

[1] A. A. Bayrakci, A. Demir, and S. Tasiran, "Fast monte carlo estimation of timing yield with importance sampling and transistor-level circuit simulation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 29, no. 9, pp. 1328–1341, 2010.

[2] W. Wu, S. Bodapati, and L. He, "Hyperspherical clustering and sampling for rare event analysis with multiple failure region coverage," in *Proceedings of the 2016 on International Symposium on Physical Design*, 2016, pp. 153–160.

[3] X. Shi, F. Liu, J. Yang, and L. He, "A fast and robust failure analysis of memory circuits using adaptive importance sampling method," in *2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*, IEEE, 2018, pp. 1–6.

Surrogate Model Method

Surrogate Model Method

The main idea of surrogate model method is to use a data-driven model to approximate the behavior of simulator the and provide a quick circuit metric estimation for any corner process.

Surrogate Model Method

- RBF Neural Network[VLSI'14][4]
- Polynomial Chaos Expansion[DAC'19][5]
- Bayesian Method[ASPDAC'20][6]

---

[4] J. Yao, Z. Ye, and Y. Wang, "An efficient sram yield analysis and optimization method with adaptive online surrogate modeling," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 23, no. 7, pp. 1245–1253, 2014.

[5] X. Shi, H. Yan, Q. Huang, J. Zhang, L. Shi, and L. He, "Meta-model based high-dimensional yield analysis using low-rank tensor approximation," in *Proceedings of the 56th Annual Design Automation Conference 2019*, 2019, pp. 1–6.

[6] S. Zhang, F. Yang, D. Zhou, and X. Zeng, "Bayesian methods for the yield optimization of analog and sram circuits," in *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, IEEE, 2020, pp. 440–445.

Our Work

1. We use HSIC-Lasso feature selection algorithm to reduce the dimension of the process variation inputs.
2. We use deep kernel learning gaussian process as our surrogate model to capture the simulator behavior.
3. We proposed a scalable parallel acquisition strategy to enable massive parallel model updates based on entropy reduction.

The open-source code is available at github[7]

[7] https://github.com/SawyDust1228/HSIC-DKL-Yield-Estimation

Background

Parameter Definition

Parameter Definition

- Define $\boldsymbol{x} = [x^{(1)}, x^{(2)}, \cdots, x^{(d)}]^T \in \boldsymbol{X}$ denotes the variational parameters, such as threshold voltage, channel length modulation effect, and bulk effect.
- Define $\boldsymbol{z}_k = [z^{(1)}, z^{(2)}, \cdots, z^{(k)}]^T \in \mathbb{R}^k$ as circuit performance metric, such as amplifier gain and memory read/write delay.
- Define $\boldsymbol{z}^0$ as the circuit metric threshold.
- Define P_f as the circuit yield failure probability.

Background

The SPICE simulation process can be seen as a black-box function $\boldsymbol{f}_k$,

$$\boldsymbol{z}_k = \boldsymbol{f}_k(\boldsymbol{x}) \tag{1}$$

Without loss of generality, $\boldsymbol{x}$ is assumed **independent** Gaussian distributed after normalization,

$$p(\boldsymbol{x}) = \prod_i^d \exp\left(-(x^{(i)})^2/2\right) / \sqrt{2\pi} \tag{2}$$

Background

Define $I : \mathbb{R}^k \to \{0, 1\}$ as a indication function of whether a performance metric fails the predefined threshold.

$$I(\boldsymbol{z}_k) \triangleq \begin{cases} 0 & \forall i \; z^{(i)} < z_i^0 \\ 1 & \exists i \; z^{(i)} > z_i^0 \end{cases} \tag{3}$$

We can compute p_f by equation (4).

$$P_f \triangleq \int_{\mathcal{X}} I(\boldsymbol{f}_k(\boldsymbol{x}))p(\boldsymbol{x})d\boldsymbol{x} \tag{4}$$

Problem Defination

Suppose D as the currently available data observed from the simulator. We can using a model $\boldsymbol{g}(\boldsymbol{x})$ to replace the simulator and approximate P_f by equation (5).

$$\hat{P}_f = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} I(\boldsymbol{g}(\boldsymbol{x}_i)) \tag{5}$$

In order to make the $\boldsymbol{g}(\boldsymbol{x})$ best represent the simulator, We need to define a strategy to find the best candidates $\{\boldsymbol{x}_*, \boldsymbol{f}_k(\boldsymbol{x}_*)\}$ based on the currently available data D to update model $\boldsymbol{g}(\boldsymbol{x})$.

⋆ Bayesian Optimization

Surrogate Model Based Yield Estimation

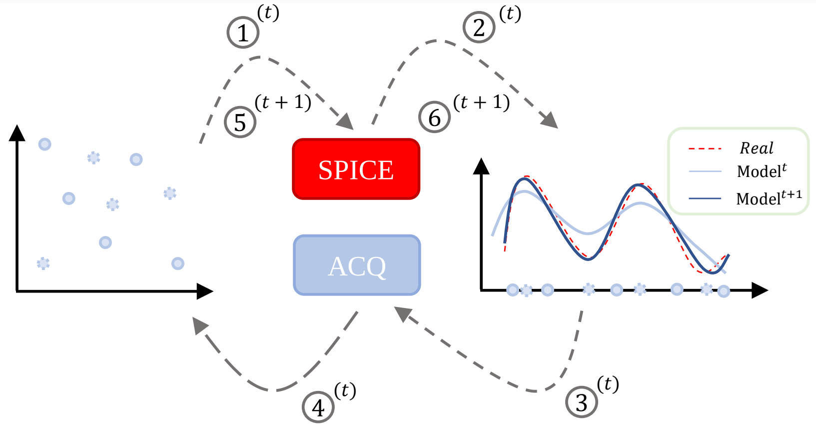

Figure 1: Illustration of the surrogate model based yield estimation: ①⑤. Conduct the SPICE simulator to get performance metrics; ②⑥. Update the surrogate model; ③④. Compute the acquisition function and find the observation candidates.

Method

Gaussian Process

Gaussian Process

$$f(\mathbf{x})|\theta \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'|\theta)) \tag{6}$$

μ is the mean function and k is the kernel function parameterized by θ.

$\star$ curse of dimensionality

Deep Kernel Learning[8]

Spectral Mixture Base Kernel

$$k_\theta(\mathbf{x}_i, \mathbf{x}_j) \to k_{\mathbf{w},\theta}(\phi(\mathbf{x}_i, \mathbf{w}), \phi(\mathbf{x}_j, \mathbf{w})) \tag{7}$$

Where $\phi(\mathbf{x}, w)$ is a non-linear mapping parameterized by weights $\mathbf{w}$, given by a deep neural network, such as a multi-layer perception (MLP) with multiple hidden layers.

1. Reduce training cost.
2. Further feature extraction.

[8] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing, "Deep kernel learning," in *Artificial intelligence and statistics*, PMLR, 2016, pp. 370–378.

Deep Kernel Learning

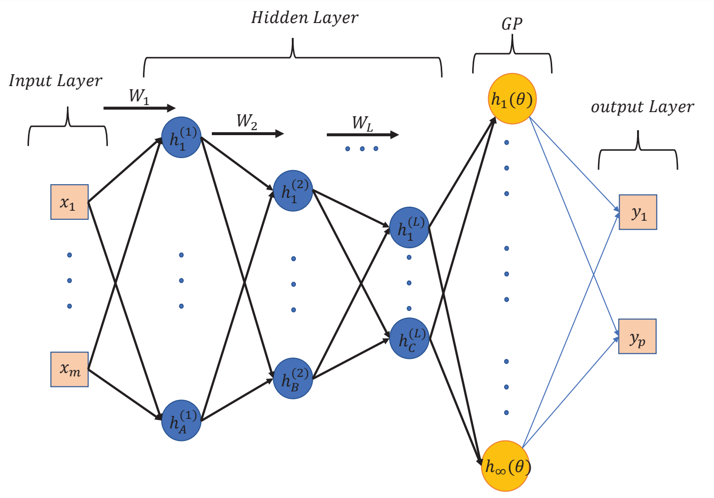

Figure 2: Illustration of the Structure of Deep kernel learning

Feature Selection

Motivation

- Previous research [DAC'18][9] noticed that not all variational parameters are equally important.

This finding makes "dimension reduction" possible to reduce the input dimension such that only the key parameters are preserved.

Because the inputs are fully independent. Thus, no dimension reduction techniques, e.g., PCA and KPCA, should be directly applied.

---

[9] J. Zhai, C. Yan, S.-G. Wang, and D. Zhou, "An efficient bayesian yield estimation method for high dimensional and high sigma sram circuits," in *2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*, IEEE, 2018, pp. 1–6.

HSIC Lasso

Hilbert-**S**chmidt **I**ndependence **C**riterion Lasso[10][11] is a nonlinear feature selection algorithm using kernel transfermation.

$$\underset{\alpha}{\operatorname{argmin}} \frac{1}{2}||\widetilde{\boldsymbol{L}} - \sum_{d=1}^{D} \boldsymbol{K}^{(d)}\boldsymbol{\alpha}^{(d)}||_2 + \lambda||\boldsymbol{\alpha}||_1 \tag{8}$$

[10] M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama, "High-dimensional feature selection by feature-wise kernelized lasso," *Neural computation*, vol. 26, no. 1, pp. 185–207, 2014.

[11] https://github.com/riken-aip/pyHSICLasso

Acquisition Function

Motivation

Since the deep kernel learning model can provide us with the uncertainty of the predicted value. So we can compute the probability that a simulation output may be failed by comparing to the threshold $\boldsymbol{z}^o$.

Advantages:

- Avoid observation in the region that the simulation performance z will "absolutely" pass or fail the threshold.
- Find multiple candidates at one iteration, which can make full use of the parallel mechanism of the SPICE simulator.

Parameter Definition

- Define $\boldsymbol{f}(\boldsymbol{x}) = [f_0(\boldsymbol{x}), f_1(\boldsymbol{x}) \dots f_k(\boldsymbol{x})]^T$ as a GP model.
- Define $l(\boldsymbol{x}) \triangleq p(\widetilde{I}(\boldsymbol{x}) = 1)$ as an approximate probability that certain variation parameter x can pass threshold, where $\widetilde{I}(\boldsymbol{x}) = I(\boldsymbol{f}(\boldsymbol{x}))$.

We can compute $l(\boldsymbol{x})$ by equation (9):

$$l(\boldsymbol{x}) = \prod_{k=1}^{K} p\left(\widetilde{f}_k(\boldsymbol{x}) \geq \boldsymbol{z}_k\right) = \prod_{k=1}^{K} \Phi\left(\frac{\mu_k(\boldsymbol{x}) - \boldsymbol{z}_k^0}{v_k(\boldsymbol{x})}\right) \tag{9}$$

Where $\Phi(\cdot)$ is the cumulative density function (CDF) of a normal distribution.

Probability Information Entropy

- Then we can compute probability information entropy $H(\boldsymbol{x})$ for the yield posterior of $\tilde{l}(\boldsymbol{x})$, which is the entropy of a Bernoulli distribution,

$$H(\boldsymbol{x}) = -l(\boldsymbol{x}) \log \left(l(\boldsymbol{x})\right) - \left(1 - l(\boldsymbol{x})\right) \log \left(1 - l(\boldsymbol{x})\right) \tag{10}$$

- We then define the total integral entropy as

$$IH = \int_{\mathcal{X}} H(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \tag{11}$$

IH (11) indicates the uncertainty of surrogate model $g(\boldsymbol{x})$ based on the current observations D.

Maximum Integral Entropy Reduction

In order to reduce the uncertainty of P_f, we can propose a candidate base on maximizing the expected integral entropy reduction.

$$
\begin{aligned}
\boldsymbol{x}^* &= \underset{\boldsymbol{x}\in\mathcal{X}}{\operatorname{argmax}}\, \mathbb{E}_f \left[(IH(\boldsymbol{x}|D) - IH(D \cup \boldsymbol{x})) \right] \\
&= \underset{\boldsymbol{x}\in\mathcal{X}}{\operatorname{argmin}}\, \mathbb{E}_f \left[IH(D \cup \boldsymbol{x}) \right]
\end{aligned} \tag{12}
$$

More specifically, we use **multi-start points** strategy to propose multiple candidates at each iteration.

$$
\boldsymbol{X}^* = \underset{\boldsymbol{X}\in\mathcal{X}}{\operatorname{argmin}}\, \mathbb{E}_f \left[IH(D \cup \boldsymbol{x}_1 \cup \cdots, \cup \boldsymbol{x}_Q) \right] \tag{13}
$$

Experiment

Baseline

- LRTA[DAC'19][12]
- HSCS[ISPD'16][13]
- HDBO[DAC'19][14]

[12] X. Shi, H. Yan, Q. Huang, J. Zhang, L. Shi, and L. He, "Meta-model based high-dimensional yield analysis using low-rank tensor approximation," in *Proceedings of the 56th Annual Design Automation Conference 2019*, 2019, pp. 1–6.
[13] W. Wu, S. Bodapati, and L. He, "Hyperspherical clustering and sampling for rare event analysis with multiple failure region coverage," in *Proceedings of the 2016 on International Symposium on Physical Design*, 2016, pp. 153–160.
[14] H. Hu, P. Li, and J. Z. Huang, "Enabling high-dimensional bayesian optimization for efficient failure detection of analog and mixed-signal circuits," in *2019 56th ACM/IEEE Design Automation Conference (DAC)*, IEEE, 2019, pp. 1–6.

Experiment Setting

To determine when to stop the yield estimation process, we follow the widely used[15][16] figure of Merit (FOM) ρ in the yield estimation literature as the stopping criteria.

$$\rho = \frac{\sqrt{\sigma_{p_f}^2}}{p_f} \tag{14}$$

Where P_f denotes the mean failure probability estimation, and σ_{p_f} the standard deviation of P_f.

[15] X. Shi, H. Yan, Q. Huang, J. Zhang, L. Shi, and L. He, "Meta-model based high-dimensional yield analysis using low-rank tensor approximation," in *Proceedings of the 56th Annual Design Automation Conference 2019*, 2019, pp. 1–6.

[16] W. Wu, S. Bodapati, and L. He, "Hyperspherical clustering and sampling for rare event analysis with multiple failure region coverage," in *Proceedings of the 2016 on International Symposium on Physical Design*, 2016, pp. 153–160.

Experiment On 6T SRAM Bitcell

Table 1: Final P_f estimation on 18-dimensional 6T SRAM

	MC	HSCS	HDBO	LRTA	Proposed
Failure prob.	4.83e-4	5.15e-4	6.25e-4	6.40e-4	4.60e-4
Relative error	Golden	6.62%	29.40%	19.46%	4.14%
# of Sim.	265000	8100	3500	2200	1350
Sim. speedup	1x	32.72x	75.71x	120.45x	196.30x
Training time	N/A	5.28s	401.62s	53.50s	1537.73s

Experiment On 6T SRAM Bitcell

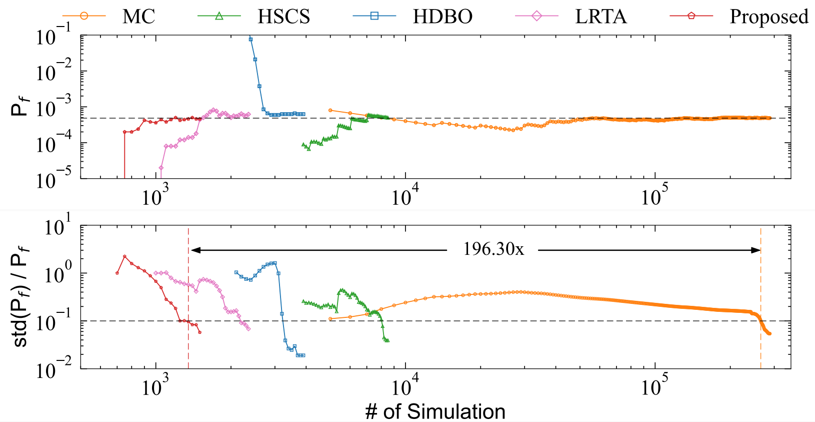

Figure 3: P_f and FOM on 18-dimensional 6T SRAM

Experiment On 6T SRAM Bitcell Array

Table 2: Final P_f on 569-dimensional SRAM column

	MC	HSCS	HDBO	LRTA	Proposed
Failure prob.	4.70e-4	5.82e-4	3.87e-4	5.60e-4	4.39e-4
Relative error	Golden	23.83%	17.66%	19.14%	6.60%
# of Sim	928500	44400	6100	5400	4000
Sim. speedup	1x	20.91x	152.21x	171.94x	232.13x
Training time	N/A	112.53s	1001.73s	12403.21s	5546.56s

Experiment On 6T SRAM Bitcell Array

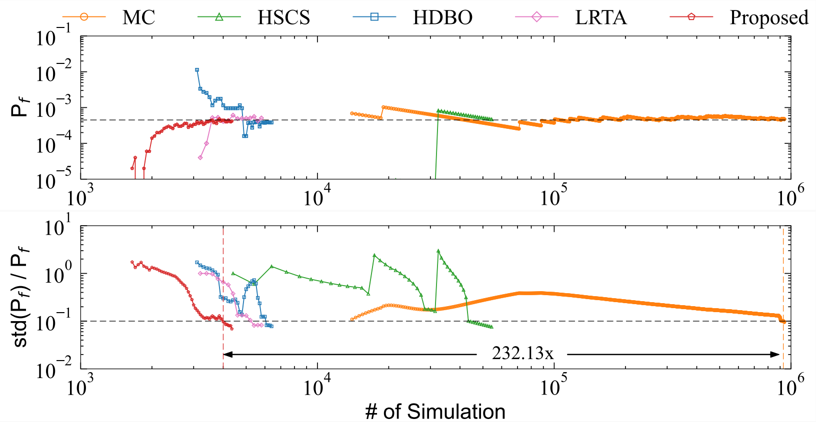

Figure 4: P_f and FOM on 569-dimensional SRAM column

Parallel Batch Update Convergence Validation

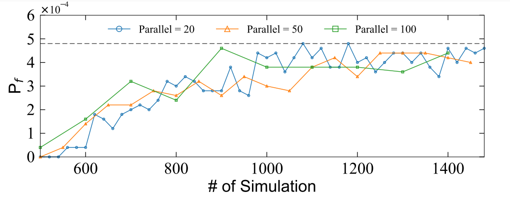

Figure 5: P_f estimation with different batch size (6T SRAM Bitcell)

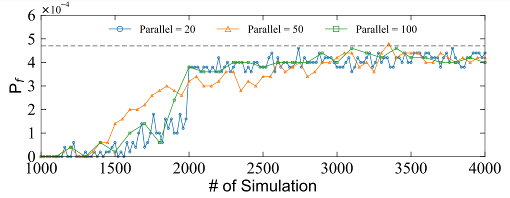

Figure 6: P_f estimation with different batch size (6T SRAM Bitcell Array)

Ablation Study

Maximum Integral Entropy Infill Validation

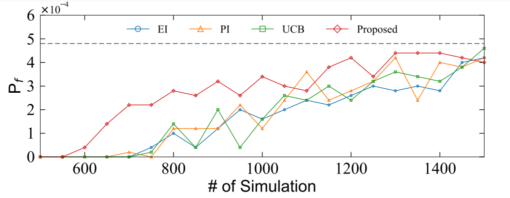

Figure 7: Acquisition function experiment (6T SRAM Bitcell)

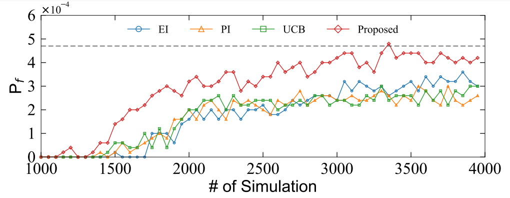

Figure 8: Acquisition function experiment (6T SRAM Bitcell Array)

Feature selection Validation.

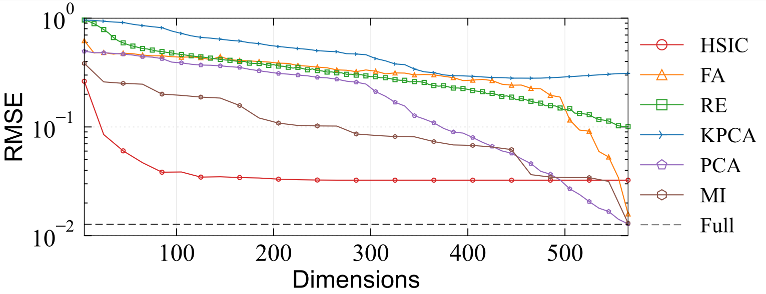

Figure 9: Feature reduction comparing to Factor Analysis (FA), Princi- pal Component Analysis (PCA), Mutual Information (MI) [DAC'18], and Random Embedding (RE) [DAC'19].

Reference I

A. A. Bayrakci, A. Demir, and S. Tasiran, “Fast monte carlo estimation of timing yield with importance sampling and transistor-level circuit simulation,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 29, no. 9, pp. 1328–1341, 2010.

W. Wu, S. Bodapati, and L. He, “Hyperspherical clustering and sampling for rare event analysis with multiple failure region coverage,” in *Proceedings of the 2016 on International Symposium on Physical Design*, 2016, pp. 153–160.

Reference II

- X. Shi, F. Liu, J. Yang, and L. He, "A fast and robust failure analysis of memory circuits using adaptive importance sampling method," in *2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*, IEEE, 2018, pp. 1–6.
- J. Yao, Z. Ye, and Y. Wang, "An efficient sram yield analysis and optimization method with adaptive online surrogate modeling," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 23, no. 7, pp. 1245–1253, 2014.

Reference III

X. Shi, H. Yan, Q. Huang, J. Zhang, L. Shi, and L. He, "Meta-model based high-dimensional yield analysis using low-rank tensor approximation," in *Proceedings of the 56th Annual Design Automation Conference 2019*, 2019, pp. 1–6.

S. Zhang, F. Yang, D. Zhou, and X. Zeng, "Bayesian methods for the yield optimization of analog and sram circuits," in *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, IEEE, 2020, pp. 440–445.

A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing, "Deep kernel learning," in *Artificial intelligence and statistics*, PMLR, 2016, pp. 370–378.

Reference IV

- J. Zhai, C. Yan, S.-G. Wang, and D. Zhou, "An efficient bayesian yield estimation method for high dimensional and high sigma sram circuits," in *2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*, IEEE, 2018, pp. 1–6.
- M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama, "High-dimensional feature selection by feature-wise kernelized lasso," *Neural computation*, vol. 26, no. 1, pp. 185–207, 2014.
- H. Hu, P. Li, and J. Z. Huang, "Enabling high-dimensional bayesian optimization for efficient failure detection of analog and mixed-signal circuits," in *2019 56th ACM/IEEE Design Automation Conference (DAC)*, IEEE, 2019, pp. 1–6.

THANK YOU!