MOBO-Driven Advanced Sub-3-nm Device Optimization for Enhanced PDP Performance

HyunJoon Jeong, *Graduate Student Member, IEEE*, JinYoung Choi, *Student Member, IEEE*, HyungMin Cho, *Student Member, IEEE*, SangMin Woo, Yohan Kim, *Member, IEEE*, Jeong-Taek Kong, and SoYoung Kim, *Senior Member, IEEE*

***Abstract*— Optimizing the nonlinear electrical characteristics of sub-3-nm devices requires considerable trial and error. However, due to the complexity of physics and secondary effects, technology computer-aided design (TCAD) simulations are time-consuming. Even with a combination of TCAD and a suitable design of experiments (DOEs), comprehensive exploration of the design space using TCAD is a challenging task. In this study, we propose a device optimization framework that can dramatically reduce the number of TCAD simulations while identifying the optimal device structure. The framework we propose consists of an artificial neural network (ANN)-based objective function derived from a dataset generated by weighted Sobol sampling, a multiobjective Bayesian optimization (MOBO) model for device optimization, and an ANN-based compact model for circuit simulation. The framework produced a device structure that showed a 51.5% performance improvement compared to the best device performance found from individual TCAD simulations of 128 structures. In contrast, the manual determination of a device achieving similar results required more than 2048 TCAD simulations.**

***Index Terms*— Machine learning (ML), optimization, sampling, SPICE, sub-3-nm device, technology computer-aided design (TCAD).**

Manuscript received 8 January 2024; revised 3 March 2024; accepted 12 March 2024. This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by Korea Government, Ministry of Science and ICT (MSIT) through Software Systems for AI Semiconductor Design under Grant 2021-0-00754; and in part by the National Research Foundation of Korea (NRF) Grant funded by Korean Government (MSIT) under Grant 2020R1A5A1019649. The review of this article was arranged by Editor M. A. Pavanello. *(Corresponding authors: Jeong-Taek Kong; SoYoung Kim.)*

HyunJoon Jeong is with the Department of Electrical and Computer Engineering, College of Information and Communication Engineering, Sungkyunkwan University, Suwon 16419, Republic of Korea (e-mail: hyeonjoon.jeong@gmail.com).

JinYoung Choi and HyungMin Cho are with the Department of Semiconductor and Display Engineering, College of Information and Communication Engineering, Sungkyunkwan University, Suwon 16419, Republic of Korea.

SangMin Woo is with Circuit Design and TCAD Solutions Team, Synopsys Korea, Seongnam 13494, Republic of Korea.

Yohan Kim is with the Computational Science and Engineering Team, Innovation Center, Samsung Electronics, Suwon 16677, Republic of Korea.

Jeong-Taek Kong and SoYoung Kim are with the Department of Semiconductor Systems Engineering, College of Information and Communication Engineering, Sungkyunkwan University, Suwon 16419, Republic of Korea (e-mail: jtkong@skku.edu; ksyoung@skku.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2024.3378224.

Digital Object Identifier 10.1109/TED.2024.3378224

I. Introduction

As TRANSISTORS become smaller, structures become more complex, circuits become more highly integrated, and optimization of nonlinear device performance becomes more difficult [1], [2], [3], [4], [5]. In our previous study, an artificial neural network (ANN)-based compact model was proposed to overcome model parameter extraction (MPE). The model became complicated due to the nonlinearity of sub-3-nm devices [6], [7]. However, TCAD simulations of state-of-the-art devices are time-consuming, and incur high computational cost to deliver optimal design parameters prior to MPE. Therefore, various works of research have been conducted to reduce the time required to explore the design space and to optimize parameters through machine learning (ML) [8], [9], [10], [11], [12], [13], [14]. However, previous research did not consider the time for training data preparation before optimization and largely depended on formulas or simulators for performance evaluation. Circuit-level verification of the optimized device has often been neglected. To overcome these limitations, we propose a highly cost-effective framework that rapidly enhances the digital performance of nanosheet FET (NSFET) devices in sub-3-nm process nodes and facilitates circuit-level verification. The flow of the proposed framework is shown in Fig. 1, with the main contributions highlighted by dashed boxes. Our proposed contributions are as follows.

1) Comprehensive exploration of the device design space requires extensive training data and is time-consuming and costly. Mehta and Wong [15] achieved R2 scores of 0.97 to 0.99 with an autoencoder using less data (25∼200). However, the size of the autoencoder model was large and a third-order polynomial model was needed to reflect the structural changes in the device. On the other hand, our model utilizes data efficiently with weighted Sobol sampling and achieves an R2 score of over 0.999 using an effective ANN architecture.
2) A newly introduced objective function model (ANN) captures the nonlinear electrical characteristics of sub-3-nm devices and serves as an efficient measure for the subsequent optimizer in device optimization, where accuracy is critical for performance evaluation. Previous studies [12], [13], [14] have relied on mathematical models or simulator connections for this function, but the former provides high speed but low accuracy, and the latter is more accurate but more time-consuming.

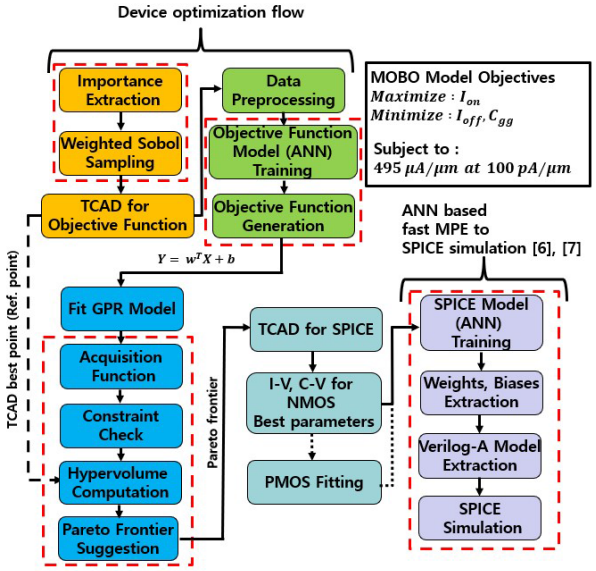

Fig. 1. Flow of the proposed ML-based device optimization technique for circuit simulation.

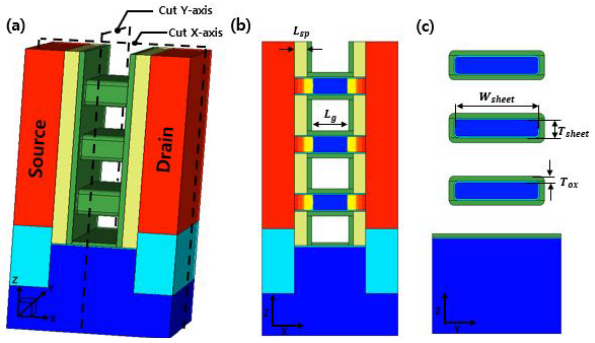

Fig. 2. (a) NSFET device structure, (b) *Y*-axis section, and (c) *X*-axis section.

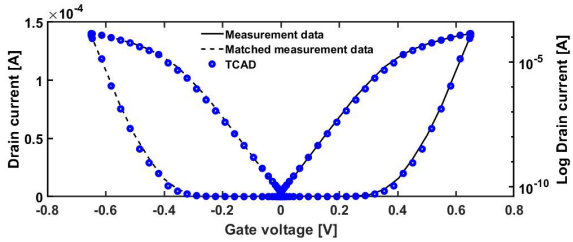

Fig. 3. Calibrated *I*–*V* curves with measurement data [17].

3) Our new framework accelerates verification of circuit-level performance by extracting I–V and C–V curves under various bias conditions only for optimized devices and automatically generating model parameters required for SPICE simulation through a SPICE model (ANN) [6], [7]. This circuit-level verification is essential to evaluate the energy efficiency and performance of the system. Nevertheless, prior works [8], [9], [10], [11], [12], [13], [14] have not extended their verification to the SPICE simulation.
4) The proposed framework, including the above contributions, searches for the optimal device and verifies it down to the circuit level more than 15 times faster than conventional TCAD-based optimization.

This study proceeds as follows according to each step of the proposed framework. Section II describes the generation of datasets of devices that are used and compared in this article through weighted Sobol sampling. Section III describes the generation of the objective function through the objective function model (ANN) using the produced training datasets. Section IV describes the multiobjective Bayesian optimization (MOBO) process for minimizing the power delay product (PDP) of the devices. In Section V, we run TCAD simulations under different voltage bias conditions based on the optimal structure and perform MPE with the SPICE model (ANN). We also perform SPICE simulations using the generated model parameters and analyze the results. Section VI presents the conclusions of the article.

II. Device Structure and Sampling Method

A. Simulation Conditions

The NSFETs were designed and simulated using Synopsys Sentaurus TCAD [16]. The simulation conditions for the NSFETs were similar to those in our previous papers [6], [7]. The structural parameters used to optimize the NSFETs are shown in Fig. 2. First, NSFET devices were designed and their electrical characteristics were compared with IBM's NSFET measurement data [17]. The structural parameters used for calibration are shown in Table I. Additionally, the channel doping concentrations are 10^{16} and 10^{15} cm^{-3} for nMOS and pMOS, respectively, and the source/drain doping concentrations are 6.5×10^{20} and 10^{21} cm^{-3} [6]. The quantum-potential model and the Fermi model were applied to study nanoscale quantum effects. The movement of carriers in a low electric field along a short channel length was captured using the quasi-ballistic mobility model. For remote phonon and Coulomb scattering effects, the Lombardi model was used, while the inversion and accumulation layer model was used to describe the surface roughness due to impurities and phonons in thin layers. To describe the bandgap changes due to doping, the Slotboom bandgap narrowing model was applied to all regions of the semiconductor [18]. Shockley-Read-Hall (SRH), Auger, and SurfaceSRH models were used as recombination models, and the band-to-band tunneling (BTBT) model was used to consider tunneling and quantum confinement in small devices [19], [20]. We calibrated the measurement data using IBM's NSFETs to achieve an accuracy greater than 99%. The resulting calibrated I–V curves and the structure parameters used for calibration are shown in Fig. 3 and Table I.

The pMOS was developed to produce electrical characteristics (i.e., matched measurement data in Fig. 3) as symmetrical as possible for stable operation of logic circuits after calibration with real-world data for nMOS. Since hole mobility is generally slower than electron mobility, processes are added to accelerate hole mobility. The stress-induced hole mobility in pMOS was studied using the h-multivalley model. A subband model among piezo models was used to apply the doping-dependent quasi-Fermi energy level, and the channel orientation was that of silicon ⟨110⟩ [21]. The change in each carrier valley due to stress was calculated using the deformation potential model [22].

TABLE I
CALIBRATION VALUES, VARIATION RANGES [17], AND ABSOLUTE IMPORTANCE VALUES OF NSFET STRUCTURAL PARAMETERS

Parameters	L_g	W_{sheet}	T_{sheet}	L_{sp}	T_{ox}
Calibration values [nm]	12	50	5	5	1.5
Value ranges [nm]	11∼ 16	20∼ 50	4∼ 6	3∼ 5	1∼ 2
Importance	0.221	0.140	0.927	0.018	0.038

Algorithm 1 Weighted Sobol Sampling

$W_{seq_{i,j}} = S_{seq_{i,j}} \times \mathrm{I}_j \times (\mathrm{B}_{j,max} - B_{j,min}) + B_{j,min}$
$\Rightarrow$ cumulative weights$_{i,j} = \sum_{i=1}^{k} W_{seq_{i,j}}$
$\Rightarrow$ probabilities$_{i,j} = \frac{cumulative\ weights_{i,j}}{cumulative\ weights_{N,j}}$
$\Rightarrow \mathrm{y}_{i,j} = F_i^{-1}(u_{i,j})$

B. Weighted Sobol Sampling

To generate the datasets, Spearman correlation analysis was used to extract the importance between device variables and the PDP. Table I shows the ranges of NSFET structural parameters along with their absolute importance, calculated using the absolute correlation coefficient. These ranges are defined based on measured device structure parameters in calibration [17] and sub-3-nm device structure parameters from IRDS [25]. The analysis used data from 1000 devices generated through Monte Carlo simulations at the device level using BSIM-CMG 111.1 model parameters [6], [7]. The importance according to Spearman correlation analysis is expressed as follows [26]:

$$\text{Importance} = 1 - \frac{6\sum d_i^2}{n\left(n^2-1\right)}. \tag{1}$$

In (1), n represents the number of data pairs used for Spearman correlation analysis, and d_i is the rank difference between the ith device structure and PDP among the 1000 BSIM devices.

Sobol sampling is an effective method to cover a design space with a small number of samples. In particular, a new weighted Sobol sampling method that weights the sampling distribution during Sobol sampling is implemented [23], [24]. In Algorithm 1, $i = 1, 2, \ldots, N$, and $j = 1, 2, \ldots, d$, where N is the number of sample data points, d is the input dimension, and $S_{\mathrm{seq}_{i,j}}$ is a random sample generated through the Sobol sequence. I_j represents the extracted importance. $B_{j,\max}$ and $B_{j,\min}$ are the maximum and minimum input values, respectively, which specify the range of each input through two variables. F_i^{-1} is the inverse cumulative distribution function, and $u_{i,j}$ is a uniform random number between 0 and 1. The probability of each sequence is calculated using the cumulative sum of the weights, and samples are extracted by applying the inverse cumulative distribution function to each probability. That is, more samples are assigned to more important variables. For example, the density of T_{sheet} is more than four times higher than that of L_g.

III. ANN MODEL FOR MULTIOBJECTIVE FUNCTIONS

The electrical characteristics of the nMOS devices were generated by inputting 128 samples, extracted through weighted Sobol sampling, in TCAD simulation. The ANN model was trained using the 128 TCAD samples. A multiobjective function was created through the trained ANN model, which is expressed as follows:

$$Y_k = W^T X + b \quad \left(k = I_{\mathrm{ON}}, I_{\mathrm{OFF}}, C_{\mathrm{gg}}\right). \tag{2}$$

The objective function model (ANN) has a multiinput multioutput structure that consists of one input layer, two hidden layers with ten hidden neurons, and one output layer, as shown in Fig. 4(a). The inputs consist of L_g, W_{sheet}, T_{sheet}, T_{ox}, and L_{sp}, which represent the structure of the NSFET, and the outputs consist of I_{ON}, I_{OFF}, and C_{gg}, which are the electrical characteristics of the device to be optimized. In our previous research [6], [7], we verified that the five structural parameters of NSFET represent more than 99% of the current and capacitance characteristics of the device. In determining the model size, we considered the amount of training data because the ANN model size increases excessively when the number of training data is too large. When the model size becomes larger, the computational cost and the risk of model overfitting increase. Therefore, we chose the minimum model size to achieve our 99% target accuracy. During data preprocessing, inputs were scaled to the MinMax size, and outputs were scaled logarithmically for learning efficiency. The hyperbolic tangent was used for the activation function. Adam was used as an optimizer during model training. The mean square error loss function was used to evaluate model accuracy. In addition, early stopping was used to prevent overfitting due to larger epochs.

In the study, we evaluated test accuracy as the training data increased from 16 to 128 samples with weighted Sobol sampling. As shown in Fig. 4(b), we achieved 99% accuracy with 128 samples. Finally, the training and test datasets consisted of 128 weighted Sobol samples and 30 random samples, respectively, with 20 000 iterations in the experiment. We also used an NVIDIA Titan XP GPU to accelerate training and found that the objective function model (ANN) completed training in around 1.5 min. The test accuracies for Sobol sampling and weighted Sobol sampling were analyzed by repeating the training process ten times. Fig. 4(c) shows the comparison between Sobol sampling and weighted Sobol sampling, with weighted Sobol sampling showing a 3.26% higher accuracy improvement. Moreover, Fig. 4(b) and (c)

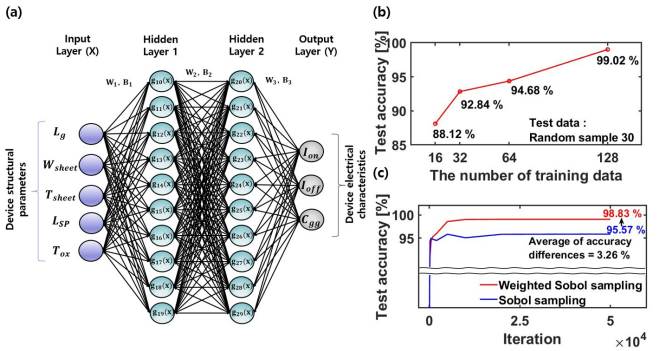

Fig. 4. (a) Objective function model (ANN) architecture, (b) test accuracy according to the number of training data with weighted Sobol sampling, and (c) test accuracies of the objective function model (ANN) according to weighted Sobol sampling and Sobol sampling.

show that the accuracy obtained with 128 Sobol samples matches that of 64 weighted Sobol samples, which means that weighted Sobol sampling reduces the sample size by half. This validates the use of weighted Sobol data to optimize design objectives with fewer sample data points. The MOBO model finds Pareto optimal points that simultaneously reduce intrinsic delay ($C_{gg}V_{dd}/I_{ON}$) and power ($P_{static} + P_{dynamic} = V_{dd}I_{OFF} + C_{gg}V_{dd}^2 f/2$). The PDP of the device is the product of power and intrinsic delay. This objective can be expressed as

$$\max\left(Y_{I_{ON}}\right) \quad \& \quad \max\left(-Y_{I_{OFF}}\right) \quad \& \quad \max\left(-Y_{C_{gg}}\right). \tag{3}$$

IV. Device Optimization With the MOBO Model

Our MOBO model uses Gaussian process regression (GPR) to express the relationship between device structure and performance while accounting for uncertainty in the objective function as represented by the variance of new data predicted by the GPR model. The initial training data obtained through weighted Sobol sampling and the current explored dataset are modeled with uncertainty using GPR. New samples are then selected, taking uncertainty into account, and the acquisition function is used to identify device structures that maximize the hypervolume. Optimal device structures that do not meet the sub-3-nm node specifications are then filtered out through constraint functions.

A. GPR Model

The surrogate model is based on a GPR model consisting of a mean function ($m(X)$) and a covariance matrix ($K(X)$) that includes a kernel function. The kernel function plays an important role in predicting GPR models, as it identifies the correlation of each input over a given distance of values [10]. We used the Matern kernel function as follows [27], [28]:

$$k_{\text{matern}}(x_i, x_j) = \theta^2 \exp\left(-\sqrt{5}r\right)\left(1 + \sqrt{5}r + \frac{5}{3}r^2\right) \tag{4}$$

where θ is a hyperparameter that includes the length scale and amplitude, and r is the distance between x_i and x_j. The Matern kernel function performs better than the radial basis function because it smooths the function for subsequent samples [28]. In GPR modeling, the maximum likelihood estimation is used to optimize θ of (4). A covariance matrix containing a kernel function with the derived θ is used to compute the posterior distribution for the new input, x_{new} [29]. The equations for mean ($\mu(X)$) and variance ($\sigma^2(X)$) of the posterior for the new input x_{new} are as follows [10], [27]:

$$\mu(x_{\text{new}}) = m_0(x_{1:n}) + k(x_{\text{new}}, x_{1:n})^T \times \left(K(x_{1:n}) + \sigma_n^2 I\right)^{-1}(y - m_0(x_{1:n})) \tag{5}$$

$$\sigma^2(x_{\text{new}}) = k(x_{\text{new}}, x_{\text{new}}) - k(x_{\text{new}}, x_{1:n})^T \times \left(K(x_{1:n}) + \sigma_n^2 I\right)^{-1} k(x_{\text{new}}, x_{1:n}). \tag{6}$$

B. Acquisition and Constraint Functions

Given the posterior distribution, the next candidate is determined using qNEHVI (parallel noise estimation hypervolume improvement), an acquisition function in the Botorch package for parallel processing of multiobjective optimization [30]. qNEHVI optimizes noisy objectives in a MOBO process to select the next data point that provides a hypervolume improvement over the reference point. Here, it is set to the optimal point among 128 TCAD samples. Fig. 5(a) shows the hypervolume for the acquisition function, which indicates convergence for 80 iterations. The hypervolume serves as a performance metric for Bayesian optimization, allowing a larger region to be explored in the objective function space. Constraints are applied to the ON- and OFF-currents based on the effective width of the 3-nm node technology in the IRDS device roadmap [25]. Gate capacitances that are not in the specification of IRDS are directly minimized without filtering by constraint functions. Fig. 5(b) illustrates the distribution of initial and new data across iterations, with the red dashed lines indicating the space where constraints are met. Data that does not meet the constraints is excluded from candidate groups for iteration. Each objective is optimized by assigning weights to the objective function and incorporating the PDP importance measured from the training data for each iteration. The MOBO experiment required 30 min for 80 iterations of optimization on a computer with an Intel Core i9-7900X CPU.

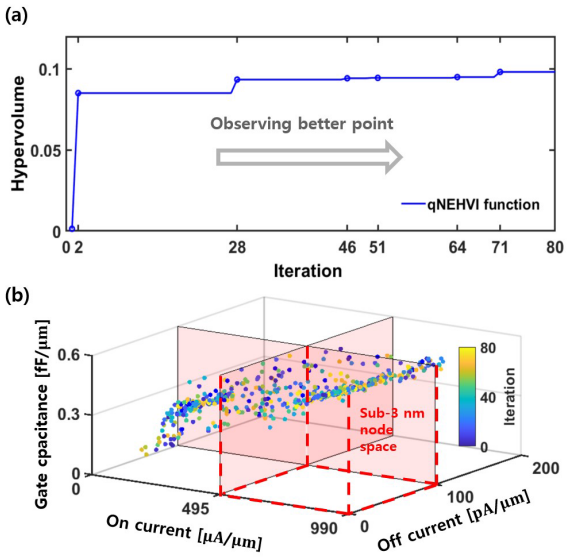

Fig. 5. (a) Hypervolume of the MOBO model according to number of iterations and (b) MOBO model 3-D scatter plot using the qNEHVI function.

C. Optimum Device Search Results

We first evaluated the accuracy between the performance of the device optimized by the MOBO model in Section IV-B and the performance verified by TCAD, as shown in Fig. 6. The accuracy between the performance of the optimized device and the performance of the TCAD device is more than 98.7% for I_{ON}, I_{OFF}, and C_{gg}, and the R2 score of the model is more than 0.98. Next, we compared the device optimized with the MOBO model to the TCAD device optimized based on the weighted Sobol sampling mentioned in Section II-B.

Fig. 7(a) shows the optimal points of the structure and the distribution of points satisfying the constraints found in each sample. When performing conventional TCAD-based

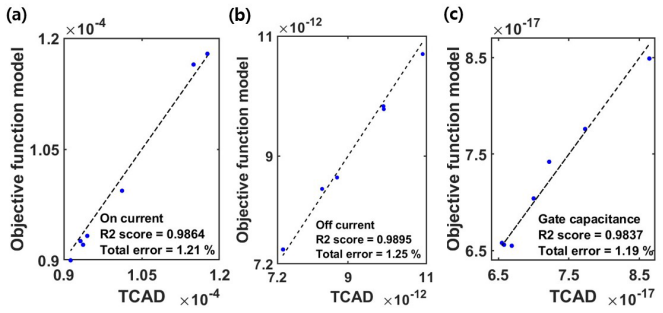

Fig. 6. R2 scores and total accuracies between the objective function model (ANN) and TCAD according to Pareto frontier (a) ON current, (b) OFF current, and (c) gate capacitance.

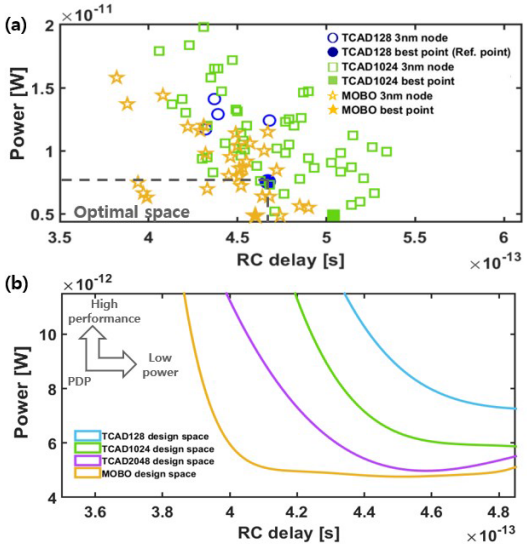

Fig. 7. (a) Points that satisfy the 3-nm node specifications and optimal points according to TCAD sample numbers and the MOBO model. (b) Explorable design space through different TCAD samples and the MOBO model.

optimization, 1024 TCAD samples are required to be located inside the optimal space. However, the MOBO model explored seven devices located in the optimal space using 128 TCAD samples. Fig. 7(b) shows the explorable design space for the MOBO model and when the number of TCAD samples is increased. Within the explored design space, devices can be selected to suit the application (high performance, low power, low PDP), and the proposed MOBO model can recommend logic devices with better performance than the TCAD samples used in the experiment. Additionally, as the number of TCAD samples gradually increases, the design space that can be explored also increases, which implies that exploring the design space of the corresponding MOBO model requires a very large number of TCAD samples. Table II and Fig. 8 show the optimal device performance explored through different TCAD sample numbers and the MOBO model, and the number of samples required for optimization. The performance of the optimal device selected by the MOBO model can achieve a 51.5% improvement compared to 128 TCAD samples. We also found that more than 2048 TCAD samples were needed to find devices similar to those optimized by the MOBO model. This means that the proposed MOBO model can optimize devices

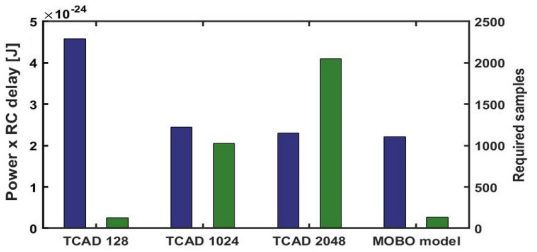

Fig. 8. PDP performance and required samples of TCAD and the MOBO model.

TABLE II
COMPARISON OF THE OPTIMAL STRUCTURE, PDP, AND SAMPLES REQUIRED FOR OPTIMIZATION ACCORDING TO THE TCAD AND MOBO MODELS FOR THE SAME DOE

Approach	Optimal structural parameters [nm]					PDP [J]	Required samples
	Lsp	Tox	Tsheet	Wsheet	Lg		
TCAD-based optimization	4.3	1.5	5.4	24.9	14.6	4.58e-24	128
	4.4	1.2	5.6	21.1	15.8	2.45e-24	1024
	4.4	1.2	6	21.1	15.7	2.30e-24	2048
MOBO model (128 TCAD + 7 Optimal candidates)	4.4	1.3	5.7	20.9	15.8	2.22e-24	135

with approximately 15 times fewer samples than conventional TCAD-based optimization.

V. CIRCUIT PERFORMANCE VALIDATION

In this section, we used a MOBO-optimized device and a conventional TCAD-optimized devices to confirm the performance improvement at the circuit level.

A. SPICE Model (ANN) for Circuit Simulation

In Section IV, the optimal structure for each nMOS device was extracted. After pMOS fitting according to the conditions mentioned in Section II-A, $I_{ds} - V_{gs}$ (V_{ds} = 0.05, 0.65 V), $I_{ds} - V_{ds}$ (V_{gs} = 0.05, 0.25, 0.5, 0.65 V), and C_{gg}, C_{gs}, $C_{gd} - V_{gs}$ (V_{ds} = 0.05, 0.65 V) curves of the optimized device were extracted through TCAD. Our previous work has demonstrated that accurate circuit simulation is possible by training only V_{ds} = 0.05 V and V_{ds} = 0.65 V [6]. As shown in Fig. 9, a SPICE model (ANN) was developed to calculate the current and capacitance values as a function of voltage biases and was divided into an I–V model and a C–V model. Unlike the objective function model (ANN) mentioned in Section III, the SPICE model (ANN) predicts the overall I–V and C–V curves under voltage bias conditions.

The I–V model has two hidden layers of five hidden neurons, and the C–V model has one hidden layer of five hidden neurons [7]. Training epochs were set to 100 000. The total accuracy of the I–V and C–V models for the SPICE model (ANN) is shown in Fig. 10. We found that both the I–V and C–V models completed training in around 5 min. With the extracted weights, biases, structural parameters, and voltage biases, a Verilog-A language-based compact model was automatically generated in the Python environment. Details of the ANN-based MPE can be found in our previous studies [6], [7]. Figs. 11 and 12 show that I–V, C–V, $g_m - V_{gs}$ (V_{ds} = 0.65 V), and $g_{ds} - V_{ds}$ (V_{gs} = 0.65 V) for the optimized device all meet the requirements with less than 1% error.

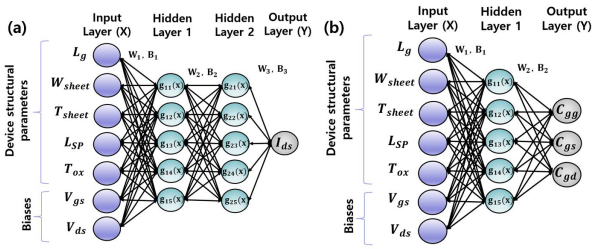

Fig. 9. SPICE model (ANN) architectures of (a) *I–V* model and (b) *C–V* model.

TABLE III
CIRCUIT PERFORMANCE OF THE MOBO MODEL, 128 TCAD SAMPLES, AND 1024 TCAD SAMPLES

Circuit	Performance	Approach		
		TCAD128	TCAD1024	MOBO model
17-stage RO	PDP [J]	2.05e-15	1.87e-15	1.72e-15
	EDP [J·s]	1.36e-25	1.23e-25	1.11e-25
TGFF	PDP [J]	4.07e-18	3.81e-18	3.31e-18
	EDP [J·s]	6.77e-29	6.20e-29	5.32e-29

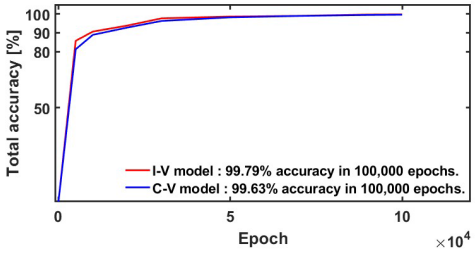

Fig. 10. Total accuracies of *I–V* and *C–V* models for the SPICE model (ANN).

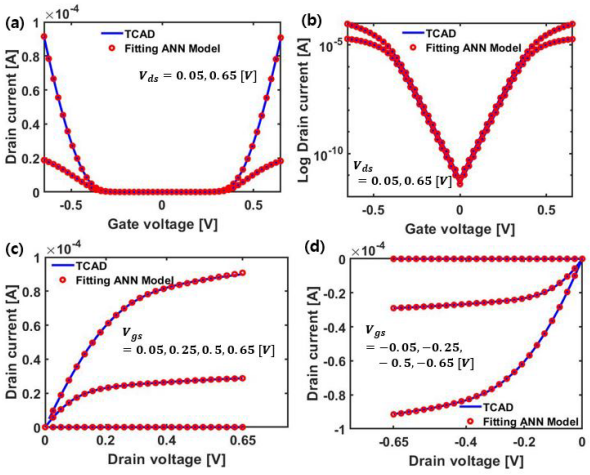

Fig. 11. SPICE model (ANN) fitting results according to the optimal device extracted by the MOBO model. (a) $I_{ds} - V_{gs}$ ($V_{ds} = 0.05$, 0.65 V), (b) log scale $I_{ds} - V_{gs}$ ($V_{ds} = 0.05$, 0.65 V), (c) nMOS $I_{ds} - V_{ds}$ ($V_{gs} = 0.05$, 0.25, 0.5, 0.65 V), and (d) pMOS $I_{ds} - V_{ds}$ ($V_{gs} = -0.05$, -0.25, -0.5, -0.65 V).

B. Circuit Simulation

Transient simulations were performed on a 17-stage ring oscillator (RO) and a transmission gate flip-flop (TGFF) [31] using the SPICE model (ANN) according to the optimized device. Table III shows the performance of PDP and energy-delay product (EDP) for each circuit using 128 or 1024 TCAD samples, and the optimal devices found with the MOBO model. The PDP performance of the 17-stage RO of the optimal device found with the MOBO model improved by 16.1% and 8.0% compared with 128 and 1024 TCAD samples, respectively. The EDP performance of the 17-stage RO found with the MOBO model improved by 18.4% and 9.8% compared with 128 and 1024 TCAD samples, respectively. Also, the PDP performance of the TGFF of the optimal device found with the MOBO model improved by 18.7% and 13.1% compared with 128 and 1024 TCAD samples, respectively. The EDP performance of the TGFF found with the MOBO model improved by 21.4% and 14.2% compared with 128 and 1024 TCAD samples, respectively.

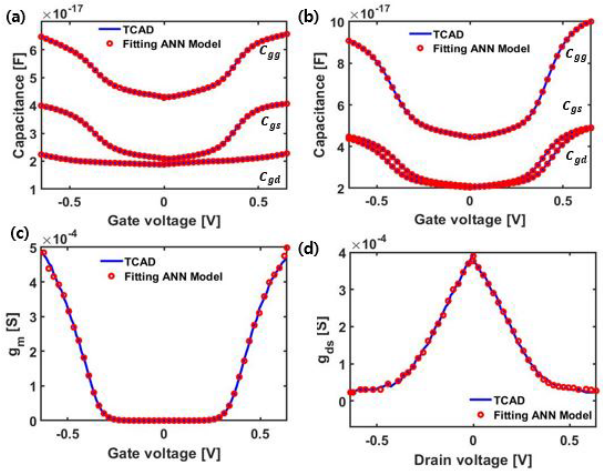

Fig. 12. SPICE model (ANN) fitting results according to the optimal device extracted by the MOBO model. (a) C_{gg}, C_{gs}, $C_{gd} - V_{gs}$ ($V_{ds} = 0.65$ V), (b) C_{gg}, C_{gs}, $C_{gd} - V_{gs}$ ($V_{ds} = 0.05$ V), (c) $g_m - V_{gs}$ ($V_{ds} = 0.65$ V), and (d) $g_{ds} - V_{ds}$ ($V_{gs} = 0.65$ V).

VI. CONCLUSION

Herein, we propose a novel framework to accelerate design space exploration and performance optimization of state-of-the-art devices. For fast performance optimization, an objective function was created to calculate the key performance of the device (I_{ON}, I_{OFF}, C_{gg}) through weighted Sobol sampling and an objective function model (ANN). Using this objective function, the MOBO model combines the GPR model and the acquisition function to achieve optimization of device performance, while meeting the stringent requirements of sub-3-nm specifications through the constraint function. We also proposed a SPICE model (ANN) coupled flow to verify the power and energy performance improvement of the circuits. The new framework enables optimal device targeting for circuit performance 15 times faster than traditional TCAD methods, resulting in performance improvements of up to 51.5%. Furthermore, analysis of PDP and EDP performance at a 17-stage RO and a TGFF circuits showed a 21.4% improvement in performance compared to TCAD-optimized devices. Therefore, the proposed framework can be further extended to new architectures to accelerate device optimization and circuit performance evaluation in sub-3-nm device development. This accelerated modeling and simulation technique is a valuable tool for pathfinding activities in sub-3-nm devices with new architectures.

ACKNOWLEDGMENT

The EDA tool was supported by the IC Design Education Center (IDEC), South Korea.

REFERENCES

[1] S. Barraud et al., "7-levels-stacked nanosheet GAA transistors for high performance computing," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2020, pp. 1–2, doi: 10.1109/VLSITechnology18217.2020.9265025.

[2] V. Jegadheesan, K. Sivasankaran, and A. Konar, "Impact of geometrical parameters and substrate on analog/RF performance of stacked nanosheet field effect transistor," *Mater. Sci. Semiconductor Process.*, vol. 93, pp. 188–195, Apr. 2019, doi: 10.1016/j.mssp.2019.01.003.

[3] J. Ajayan et al., "Nanosheet field effect transistors—A next generation device to keep Moore's law alive: An intensive study," *Microelectron. J.*, vol. 114, Aug. 2021, Art. no. 105141, doi: 10.1016/j.mejo.2021.105141.

[4] A. Veloso et al., "Nanowire & nanosheet FETs for ultra-scaled, high-density logic and memory applications," *Solid-State Electron.*, vol. 168, Jun. 2020, Art. no. 107736, doi: 10.1016/j.sse.2019.107736.

[5] S.-D. Kim, M. Guillorn, I. Lauer, P. Oldiges, T. Hook, and M.-H. Na, "Performance trade-offs in FinFET and gate-all-around device architectures for 7 nm-node and beyond," in *Proc. IEEE SOI-3D-Subthreshold Microelectron. Technol. Unified Conf. (S3S)*, Rohnert Park, CA, USA, 2015, pp. 1–3, doi: 10.1109/S3S.2015.7333521.

[6] H. Jeong et al., "Fast and expandable ANN-based compact model and parameter extraction for emerging transistors," *IEEE J. Electron Devices Soc.*, vol. 11, pp. 153–160, 2023, doi: 10.1109/JEDS.2023.3246477.

[7] S. Woo, H. Jeong, J. Choi, H. Cho, J.-T. Kong, and S. Kim, "Machine-learning-based compact modeling for sub-3-nm-node emerging transistors," *Electronics*, vol. 11, no. 17, p. 2761, Sep. 2022, doi: 10.3390/electronics11172761.

[8] H. Yamano et al., "Efficient optimization approach for designing power device structure using machine learning," *Jpn. J. Appl. Phys.*, vol. 62, Apr. 2023, Art. no. SC1050, doi: 10.35848/1347-4065/acb061.

[9] H. Xu et al., "A machine learning approach for optimization of channel geometry and source/drain doping profile of stacked nanosheet transistors," *IEEE Trans. Electron Devices*, vol. 69, no. 7, pp. 3568–3574, Jul. 2022, doi: 10.1109/TED.2022.3175708.

[10] B. Kim and M. Shin, "Bayesian optimization of MOSFET devices using effective stopping condition," *IEEE Access*, vol. 9, pp. 108480–108494, 2021, doi: 10.1109/ACCESS.2021.3101812.

[11] J. Yoon, S. Lee, H. Yun, and R. Baek, "Digital/analog performance optimization of vertical nanowire FETs using machine learning," *IEEE Access*, vol. 9, pp. 29071–29077, 2021, doi: 10.1109/ACCESS.2021.3059475.

[12] J. Chen et al., "Automatic selection of structure parameters of silicon on insulator lateral power device using Bayesian optimization," *IEEE Electron Device Lett.*, vol. 41, no. 9, pp. 1288–1291, Sep. 2020, doi: 10.1109/LED.2020.3013571.

[13] H. Gangi et al., "Design optimization of multiple stepped oxide field plate trench MOSFETs with machine learning for ultralow on-resistance," in *Proc. 33rd Int. Symp. Power Semiconductor Devices ICs (ISPSD)*, May 2021, pp. 151–154, doi: 10.23919/ISPSD50666.2021.9452194.

[14] T. Wu and J. Guo, "Multiobjective design of 2-D-material-based field-effect transistors with machine learning methods," *IEEE Trans. Electron Devices*, vol. 68, no. 11, pp. 5476–5482, Nov. 2021, doi: 10.1109/TED.2021.3085701.

[15] K. Mehta and H.-Y. Wong, "Prediction of FinFET current-voltage and capacitance-voltage curves using machine learning with autoencoder," *IEEE Electron Device Lett.*, vol. 42, no. 2, pp. 136–139, Feb. 2021, doi: 10.1109/LED.2020.3045064.

[16] *Sentaurus Device User Guide*, document Version P-2019.03, Synopsys, Sunnyvale, CA, USA, Mar. 2019.

[17] N. Loubet et al., "Stacked nanosheet gate-all-around transistor to enable scaling beyond FinFET," in *Proc. Symp. VLSI Technol.*, Jun. 2017, pp. T230–T231, doi: 10.23919/VLSIT.2017.7998183.

[18] J.-S. Yoon, J. Jeong, S. Lee, and R.-H. Baek, "Optimization of nanosheet number and width of multi-stacked nanosheet FETs for sub-7-nm node system on chip applications," *Jpn. J. Appl. Phys.*, vol. 58, no. SB, Apr. 2019, Art. no. SBBA12, doi: 10.7567/1347-4065/ab0277.

[19] D. Ryu, M. Kim, J. Yu, S. Kim, J.-H. Lee, and B.-G. Park, "Investigation of sidewall high-*k* interfacial layer effect in gate-all-around structure," *IEEE Trans. Electron Devices*, vol. 67, no. 4, pp. 1859–1863, Apr. 2020, doi: 10.1109/TED.2020.2975255.

[20] J.-S. Yoon, J. Jeong, S. Lee, and R.-H. Baek, "Systematic DC/AC performance benchmarking of sub-7-nm node FinFETs and nanosheet FETs," *IEEE J. Electron Devices Soc.*, vol. 6, pp. 942–947, 2018, doi: 10.1109/JEDS.2018.2866026.

[21] Y. Sun, S. E. Thompson, and T. Nishida, "Physics of strain effects in semiconductors and metal-oxide-semiconductor field-effect transistors," *J. Appl. Phys.*, vol. 101, no. 10, May 2007, Art. no. 104503, doi: 10.1063/1.2730561.

[22] S. Reboh et al., "An analysis of stress evolution in stacked GAA transistors," in *Proc. IEEE Silicon Nanoelectronics Workshop (SNW)*, Jun. 2016, pp. 206–207, doi: 10.1109/SNW.2016.7578053.

[23] I. M. Sobol', "On the distribution of points in a cube and the approximate evaluation of integrals," *USSR Comput. Math. Math. Phys.*, vol. 7, no. 4, pp. 86–112, Jan. 1967, doi: 10.1016/0041-5553(67)90144-9.

[24] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. New York, NY, USA: Springer, 2004.

[25] *International Roadmap for Devices and Systems (IRDS)*, 2021 ed. IEEE, Piscataway, NJ, USA, 2021.

[26] J. Hauke and T. Kossowski, "Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data," *Quaestiones Geographicae*, vol. 30, no. 2, pp. 87–93, Jun. 2011.

[27] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: A review of Bayesian optimization," *Proc. IEEE*, vol. 104, no. 1, pp. 148–175, Jan. 2016, doi: 10.1109/JPROC.2015.2494218.

[28] A. Deshwal and J. Doppa, "Combining latent space and structured kernels for Bayesian optimization over combinatorial spaces," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 8185–8200.

[29] C. Xanthopoulos, "Gaussian process-based wafer-level correlation modeling and its applications," in *Machine Learning in VLSI Computer-Aided Design*, I. M. Elfadel, D. S. Boning, and X. Li, Eds., 1st ed. Cham, Switzerland: Springer, 2019, ch. 5, pp. 119–173.

[30] S. Daulton, M. Balandat, and E. Bakshy, "Parallel Bayesian optimization of multiple noisy objectives with expected hypervolume improvement," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 2187–2200.

[31] G. Gerosa et al., "A 2.2 W, 80 MHz superscalar RISC microprocessor," *IEEE J. Solid-State Circuits*, vol. 29, no. 12, pp. 1440–1454, Dec. 1994, doi: 10.1109/4.340417.