Received 9 April 2024; revised 11 May 2024; accepted 29 May 2024. Date of publication 4 June 2024; date of current version 23 July 2024.
The review of this article was arranged by Editor P. Pavan.

*Digital Object Identifier 10.1109/JEDS.2024.3409572*

# Enhancing Interpretability of Neural Compact Models: Toward Reliable Device Modeling

**CHANWOO PARK [1,2], HYUNBO CHO [2], AND JUNGWOO LEE [1] (Senior Member, IEEE)**

1 Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea
2 Research and Development Center, Alsemy Inc., Seoul 06154, South Korea

CORRESPONDING AUTHOR: J. LEE (e-mail: junglee@snu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) under Grant 2021R1A2C2014504 (40%) and Grant 2021M3F3A2A02037893 (10%)), and in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) Grant funded by the Ministry of Science and ICT (MSIT), INMAC, and BK21 FOUR Program under Grant 2021-0-00106 (40%) and Grant 2021-0-02068 (10%)).

**ABSTRACT** Neural Compact Models (NCMs) have emerged as a crucial tool to meet the stringent demands of Design-Technology Co-Optimization (DTCO) and to overcome the complexities and prolonged development cycles encountered in traditional compact model creation. Despite their efficiency in simulating electronic devices, a significant barrier to the widespread adoption of NCMs in the industry remains: the lack of interpretability. In the semiconductor sector, where inaccuracies or failures can lead to considerable financial consequences, it is critical to ensure that the model's predictions are both understandable and reliable. This study aims to enhance the interpretability of NCMs used for I-V and C-V characterizations by clarifying the physical significance of latent vectors. A regularization technique is employed to disentangle features within the latent space, and interpolation is used to visualize and elucidate each dimension's physical impact. Our approach, which offers interpretable insights into the model's functionality, seeks to encourage broader implementation of NCMs in the industry, thus accelerating advancements in DTCO.

**INDEX TERMS** Neural compact models, interpretability, latent vector interpolation, design-technology co-optimization.

## I. INTRODUCTION

As semiconductor technology advances toward further miniaturization, the need for advanced and rapid compact modeling becomes increasingly pronounced. Conventional methods, dependent on analytical equations and extensive parameter extraction, are challenged to meet the expedited development cycles required by emerging technology nodes [1], [2], [3], [4]. The increased complexity added by device downscaling complicates the creation of precise compact models, often extending the turnaround time (TAT) and posing an obstacle to efficient Design-Technology Co-Optimization (DTCO).

In response to these challenges, Neural Compact Models (NCMs), which utilize Artificial Neural Networks (ANNs) for device modeling, have been recognized as a viable alternative to streamline the modeling process [5], [6], [7], [8], [9], [10], [11]. NCMs, by incorporating domain knowledge into ANN architectures and preprocessing, can quickly simulate novel devices with a high degree of accuracy and efficiency that outperforms traditional methods. Despite these advances, the inherent lack of interpretability in these NCMs is a significant hurdle that impedes their widespread adoption in the industry.

The challenge of understanding black-box ANNs is pivotal, particularly for applications in sectors where safety is critical, such as autonomous driving [12] and medical diagnostics [13]. In semiconductor device modeling, where inaccuracies can lead to significant time and financial costs, enhancing both the interpretability and reliability of NCMs is essential. Researchers have introduced various approaches, including physics-informed networks [5] and hybrid models [14], to improve precision and physical plausibility.

Physics-informed networks successfully integrate device physics to circumvent unphysical behaviors of NCMs but can restrict ANN flexibility, potentially limiting predictive performance. Conversely, while model-based approaches combine basic analytical equations with ANNs, they still retain the "black-box" nature of ANNs, achieving only partial model interpretability. This work seeks to bridge this gap by introducing a framework that not only capitalizes on the speed and accuracy of NCMs but also makes their interpretative features more accessible.

We investigate the physical relationships between latent vectors and generated device characteristics that emerge during the end-to-end training of NCMs. Employing a transformer-based encoder, our approach extracts the physical representation of the given technology and regularizes the latent vector to ensure a sparse and disentangled representation. This technique enables a clearer separation of the underlying factors in latent representations with distinct semantic meanings. With our approach, which examines and validates the model's internal processes, we aim to enhance the reliability and interpretability of NCMs, promoting their wider use in DTCO and accelerating the advancement of semiconductor device research.

Our contributions are summarized as follows:

- We introduce an encoder-decoder framework for neural compact modeling that incorporates a regularized latent space to ensure clear interpretability and distinct separation of device characteristics.
- To the best of our knowledge, this is the first work to integrate explainable AI principles into device modeling, aiming for more transparent and understandable NCMs.
- We clarify the role of each latent dimension in defining device characteristics and present a heatmap-based analysis to elucidate the internal workings of ANN-based compact models, thus making the model's predictive capabilities more transparent.

## II. RELATED WORK

### A. MODEL INTERPRETABILITY

Extensive research focuses on enhancing the interpretability of Machine Learning (ML) models, aiming to clarify their "black-box" nature and their internal representations. A technique providing visual explanations for the decisions of CNN-based models has been proposed, offering a insight into their decision-making process [15]. In autonomous driving, a framework designed to effectively process and fuse multi-modal and multi-view sensor information to achieve comprehensive scene understanding and detect adversarial events has been introduced [12]. For medical diagnosis, models that create a direct, multimodal connection between medical images and diagnostic reports have been explored, providing meaningful and visually interpretative diagnostic processes [13]. Moreover, in language modeling, attempts have been made to synergize reasoning and action, enabling models to generate reasoning traces and task-specific actions in a mutually informative manner [16]. Across these domains, interpretability stands out as a primary research objective to enhance model utility and trustworthiness.

### B. NEURAL COMPACT MODELS

Recent progress in NCMs has incorporated domain knowledge into ANNs [6], [8], [9]. This integration has led to the development of specialized architectures and loss functions that mitigate non-physical behaviors [5], [17], and the use of Autoencoders (AE) to align physical parameters with I-V characteristics [7]. Studies to address data scarcity have included the deployment of hierarchical networks for MOSFETs [10] and meta-learning techniques for prior knowledge construction [11]. Furthermore, ANNs have been shown to enhance the convergence of circuit simulation, outperforming conventional analytical models [18].

On the other hand, hybrid models that combine basic analytic models and ANNs have been proposed for emerging devices [14], [19], [20]. While these models effectively learn I-V trends by integrating both physics-based and learning-based approaches in a complementary manner, they remain partially interpretable due to the inclusion of ANNs, or are less accurate due to reliance on simple equations. Although the above strategies provide fast and accurate simulations of emerging devices, the industrial adoption of NCMs has been impeded primarily by concerns regarding their interpretability and reliability. In this work, we refine our approach by expanding the dimensionality of our model to enhance predictive performance, while applying L1 regularization for an interpretable, sparse representation. This strategy ensures a balanced solution that optimizes both interpretability and performance, fostering reliability and a deeper understanding in practical applications, promoting broader industry adoption.

## III. PROPOSED METHOD

### A. OVERVIEW OF FRAMEWORK

The workflow of our framework is depicted in Figure 1, outlining two main stages: training and latent analysis. In the training stage, each individual input point is mapped into a $d_{\text{latent}}$-dimensional embedding. This embedding is then processed through Multihead Attention (MHA) and aggregated using max pooling to create a latent vector. The decoder, augmented with additional query inputs such as $V_{DS}$ and $V_{GS}$, utilizes this latent vector to predict target values including $I_D$ and $C_{GG}$. We apply L1 regularization to the latent vectors, aiming for a disentangled representation to facilitate clear interpretation. This encoder-decoder configuration is trained end-to-end across a variety of devices to optimize performance.

In the latent analysis phase, the trained encoder extracts essential information from input measurements. For each dimension $z_i$ of the latent vector, we conduct two analysis. First, we interpolate $z_i$ between $-1$ and $1$ while keeping other dimensions constant to observe variations in decoder I-V, C-V outputs and infer the role of each latent dimension. Second, we compute the gradient of $z_i$ with respect to the input measurements, creating a heatmap to identify input data points that significantly influence $z_i$. This dual approach clarifies the causes and consequences of each latent

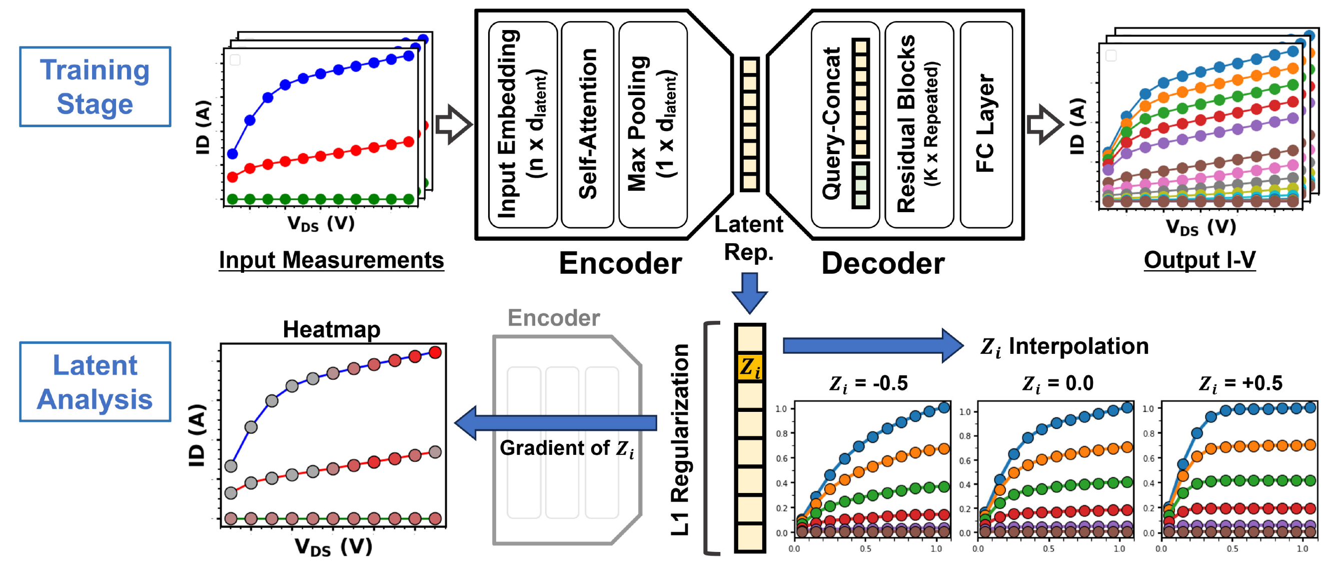

**FIGURE 1.** Overview of our framework: In the training stage, input measurements are encoded into a latent vector, which the decoder then utilizes to predict target I-V and C-V curves. The latent analysis stage focuses on interpreting each latent dimension $z_i$ through interpolation and heatmap generation, to clarify its role and identify the influential input data points.

dimension, enhancing our understanding of how the ANN functions and what it has learned in device modeling.

### B. METHOD DETAILS

In our methodology, we employ an encoder-decoder structure, designed to address the complexities of predicting I-V and C-V curves at unseen biases by utilizing crucial information extracted from the input data. We utilize a dataset generated using the BSIM4 model in SPICE, calibrated to align with the 32nm technology standards specified in the ITRS Roadmap [21]. This dataset includes variations in Width (W), Length (L), and Temperature (T), and for each device, we have organized the data into input (assumed provided measurements) and target (values for prediction) categories. The encoder is trained to extract essential information from the available input data. Subsequently, the decoder, comprising fully-connected layers with skip connections, uses a latent vector, combined with unseen biases, to predict the behavior of the target region of a device. Further details are outlined in Algorithm 1.

Incorporating a self-attention mechanism within the encoder is crucial for grasping the underlying physics in the provided data by identifying relationships among input points. This mechanism is designed to capture physical relationships between various positions in the input measurements and to understand connections between physical phenomena, such as the connection between Drain-Induced Barrier Lowering (DIBL) and output resistance ($R_{out}$) in I-V curves. This feature enables the model to discern complex patterns and comprehend the physics embedded in the device characteristics, thereby not only enhancing its predictive performance but also distilling interpretable factors into latent representations.

**Algorithm 1** Encoder-Decoder Training

```
Notations: $k$: # input data, $d$: embed size, $m$: input dimension
Input: Device data $(x_i, y_i)$ for $i = 1, \ldots, N$;
Parameter: Regularization weight $\lambda$;
Split the device data into input set $\mathcal{T}$ and target set $\mathcal{U}$;
Encoder:
Form input matrix: $\mathbf{X}_{\mathcal{T}} \in \mathbb{R}^{k \times d}$;
Apply self-attention on $\mathbf{X}_{\mathcal{T}}$: $\mathbf{S}_{\mathcal{T}} = \text{softmax}\left(\frac{\mathbf{X}_{\mathcal{T}}\mathbf{X}_{\mathcal{T}}^T}{\sqrt{d}}\right)\mathbf{X}_{\mathcal{T}}$;
Aggregate: $z = \text{maxpool}(\mathbf{S}_{\mathcal{T}}) \in \mathbb{R}^{1 \times d}$;
Decoder:
Initialize total loss: $\mathcal{L}_{\text{total}} = 0$;
for $(x_j, y_j) \in \mathcal{U}$ do
    Form decoder input: $\mathbf{Z}_j = \text{concat}(z, x_j) \in \mathbb{R}^{d+m}$;
    Predict output: $\hat{y}_j = f_{\text{dec}}(\mathbf{Z}_j)$;
    Compute loss: $\mathcal{L}_j = \lambda \cdot \|z\|_1 + \text{MSE}(\hat{y}_j, y_j)$;
    Update total loss: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{total}} + \mathcal{L}_j$;
end for
Compute gradients: $\nabla_{\text{enc}}, \nabla_{\text{dec}} = \nabla\mathcal{L}_{\text{total}}$;
Update encoder and decoder: $f_{\text{enc}} \leftarrow f_{\text{enc}} - \alpha\nabla_{\text{enc}}, f_{\text{dec}} \leftarrow f_{\text{dec}} - \alpha\nabla_{\text{dec}}$;
```

We employ L1 regularization on the latent vector to enhance interpretability by encouraging sparsity in the latent representations. This method effectively isolates the data's generative factors into separate, clear dimensions in the latent space, thereby allowing for a straightforward and easily understandable connection between changes in the device characteristics and shifts in the latent variables.

Identifying the optimal latent dimensionality is key because dimensions that are too large can hinder interpretability, whereas those too small can cause an information bottleneck, potentially harming the learning process. Preliminary experiments with Variational Autoencoders (VAEs) revealed that a latent space of 8 dimensions was sufficient to reconstruct the I-V and C-V curves. However, to achieve sparse and factored representations for improved interpretability, we chose a latent space of 32 dimensions with L1 regularization. By utilizing a larger latent space and enforcing stringent regularization, the model can learn a sparse, defined representation, in which each dimension isolates different independent characteristic factors. This technique allows us to achieve a balance between maintaining an interpretable latent space and ensuring practical computational and representational efficiency.

### C. EXPERIMENTAL SETUP

To develop a robust model, we created a comprehensive dataset using BSIM4 models calibrated to 32nm technology. The dataset includes various device dimensions and temperatures to capture a wide range of properties within semiconductor devices. Gate widths span from 50nm to 100um, and gate lengths range from 32nm to 10um. We covered operating temperatures from 0 to 125 °C.

We partitioned the devices into two sets: one for training and one for testing. This approach allowed our model to learn from the training devices and then have its performance evaluated on the unseen test devices. Further, we divided the data for each device into input and target sets, simulating a scenario where the input data are user-provided measurements used to generate latent code, which is utilized by the neural compact model (decoder).

For the input data, we conducted $V_{DS}$ sweeps at fixed $V_{GS}$ values of 0.0, 0.8, and 1.1 V, as well as $V_{GS}$ sweeps at fixed $V_{DS}$ values of 0.1, 0.5, and 1.1 V. These sweeps generate the $I_D - V_{DS}$ and $C_{GG} - V_{DS}$ curves, as well as the $I_D - V_{GS}$ and $C_{GG} - V_{GS}$ curves, respectively. The target data consist of all data points not included in these input sweeps. Specifically, the decoder is tasked with predicting the entire range of $V_{DS}$ (from 0 to 1.1 V) and $V_{GS}$ (from −0.3 to 1.1 V) at an interval of 0.05 V. These data points cover the full operational space of the devices, extending beyond the fixed values used in the input data sweeps.

In our experiments, we trained our model for a total of 5,000 epochs using a batch size of 64. The learning rate was initially set to $10^{-4}$ and decayed by a factor of 0.8 every 200 epochs. We utilized the Adam optimizer to minimize the total loss, which comprises a combination of Mean Squared Error (MSE) loss and L1 regularization, with a coefficient of $10^{-5}$. The encoder was configured with a Multihead Attention (MHA) mechanism, including 4 attention heads and 3 encoder layers, all utilizing the ELU activation function. The decoder was constructed with 6 layers, each possessing 256 hidden nodes, and incorporated skip connections to enhance performance.

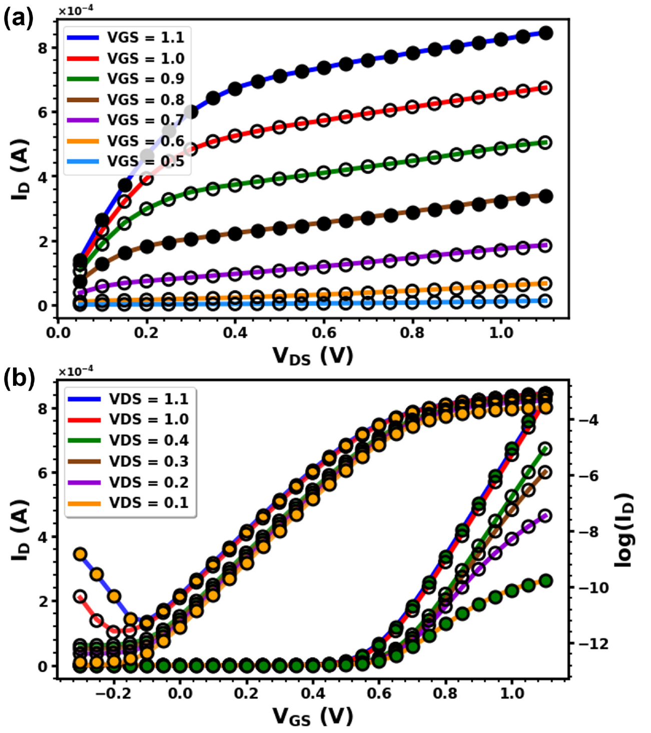

FIGURE 2. Performance on I-V Curve. Solid circles denote input data points; open circles represent predictions on test data, highlighting the model's generalization capabilities.

Figure 2 exhibits the performance of our encoder-decoder architecture, demonstrating its capability to accurately capture the output characteristics shown in (a), and the transfer characteristics depicted in (b). In this figure, solid circles represent the provided input data points, while open circles indicate unseen test data. The model demonstrated high precision on both the provided input and unseen test data, emphasizing its robustness and validity.

## IV. LATENT DIMENSIONS ANALYSIS

This section explores the latent space of each trained I-V and C-V model, examining the role of each dimension through interpolation and heatmap analysis. For interpolation, an instance of device input is encoded into a latent vector, with each dimension interpolated while holding others constant, to understand the function of each dimension. Heatmap analysis computes the gradient of each latent dimension with respect to the input data samples, highlighting input points that significantly influence the decisions of the ANNs. The gradients are normalized and displayed in a color spectrum ranging from gray to red, where a red marker indicates a high gradient value for the target latent dimension, signifying a strong impact on the model's output. Furthermore, by commenting on the explanations of the interpolated curves and potential device physics reasons, we aim to underscore that these interpolated curves are not merely visually similar, but they also represent physically plausible device characteristics.

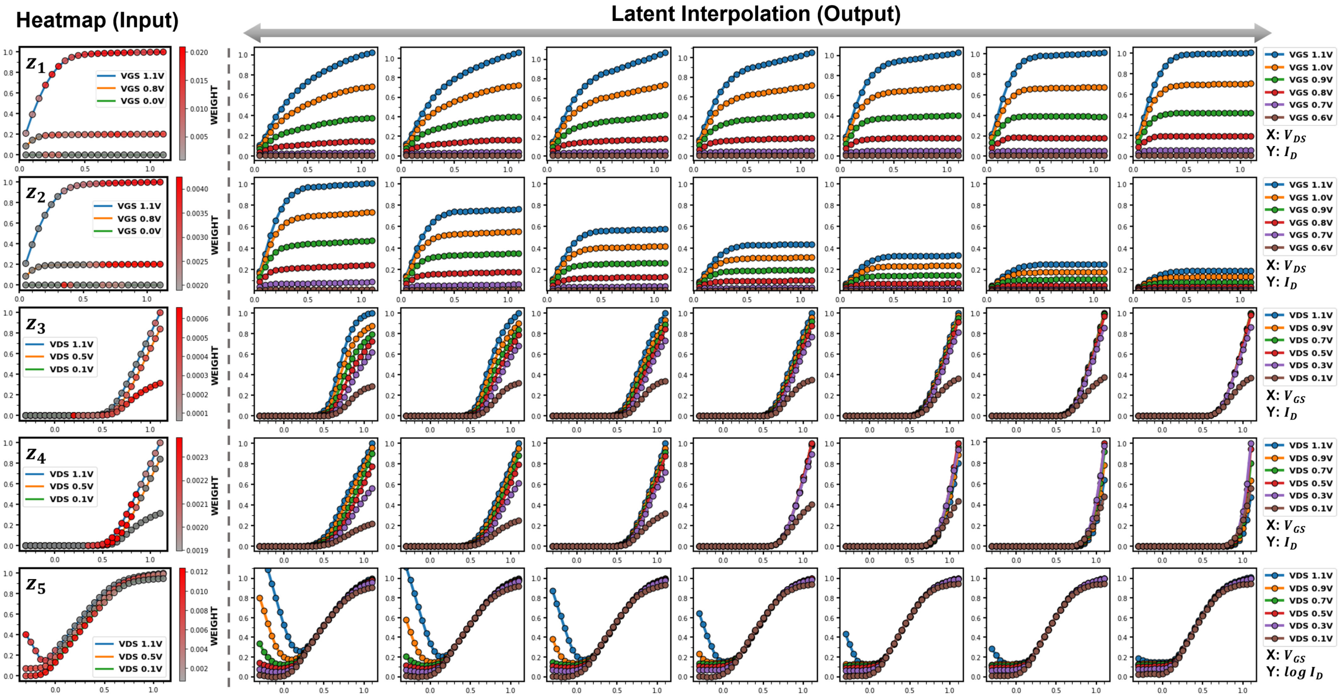

**FIGURE 3.** **Latent Interpolation of I-V Characteristics: Each row demonstrates variations in I-V curves influenced by a distinct latent dimension ($z_1$ to $z_5$). The disentangled representation emphasizes the unique, interpretable role of each latent variable.**

### A. I-V MODEL INTERPRETATION

As depicted in Figure 3, we examine the influence of five specific latent dimensions, $z_1$ to $z_5$, on the model's output. It is crucial to note that the numbering of these dimensions is for illustration purposes and does not represent a fixed order, as their ordering is determined automatically during training. To focus on shape variations and facilitate clearer interpretation, we normalize $I_D$ and $\log I_D$ in the plots by their respective maximums, except for $z_2$, which is normalized using the device with the largest $I_{D_{max}}$.

Upon examining $z_1$, it is evident that in the saturation region, the output conductance, $g_{ds}$, tends to flatten as $z_1$ increases. The I-V model, trained with diverse device sizes and temperatures, needs to capture the unique behaviors associated with Channel Length Modulation (CLM) and thermal effects, as both significantly influence $g_{ds}$ in the saturation region. To effectively differentiate between devices and make accurate predictions in unseen regions of devices, the I-V model has encoded information related to $g_{ds}$ in the saturation region within the $z_1$ latent dimension. The heatmap shows that regions with high $V_{GS}$ and $V_{DS}$ values are highlighted in red, indicating that the encoder is effectively capturing information related to $g_{ds}$ from these input saturation regions.

In the plots for $z_2$, a consistent decrease in the $I_D$ scale is observed as $z_2$ ascends, while the curve shapes remain unchanged. Given the direct relationship between the $I_D$ scale and gate width, $z_2$ likely corresponds to this device parameter. The heatmap supports this by showing the encoder's emphasis on regions of maximal $I_D$ across varying $V_{GS}$ values to determine scales for unseen $I_D - V_{DS}$ sweeps.

Moving to $z_3$, we notice two distinct behaviors with changing values. First, with an increase in $z_3$, $I_D$ shifts from saturation to a linear rise with $V_{GS}$. This shift can be attributed to factors like velocity saturation or early channel pinch-off. Second, a noticeable spread of curves with $V_{DS}$ variations appears evident at lower $z_3$ values, while curves tend to converge at elevated $z_3$ values. Phenomena such as CLM and DIBL can explain these variances. Given that these effects occur after the device is turned on, the model predominantly extracts data from high $V_{GS}$ input regions.

Observations related to $z_4$ indicate that the threshold voltage ($V_{TH}$) rises as $z_4$ increases. This rise may result from the channel lengthening, which mitigates short-channel effects and alters $V_{th}$. The heatmap highlights the model's focus on inputs where $I_D$ shows a rapid change. Lastly, for $z_5$, the data indicates a reduction in Gate Induced Drain Leakage (GIDL) intensity as $z_5$ grows. This effect becomes more evident with larger $V_{DS}$ values, a consequence of the strengthened electric field near the drain-to-channel junction. It's also worth noting that shorter channel lengths can exacerbate GIDL. The model pays more attention to input data where the device is essentially off, i.e., $V_{GS}$ values close to zero or negative.

### B. C-V MODEL INTERPRETATION

Referring to Figure 4, we investigate the effect of specific latent dimensions, ranging from $z_1$ to $z_3$, on the predictions of the model. Following an approach analogous to the I-V

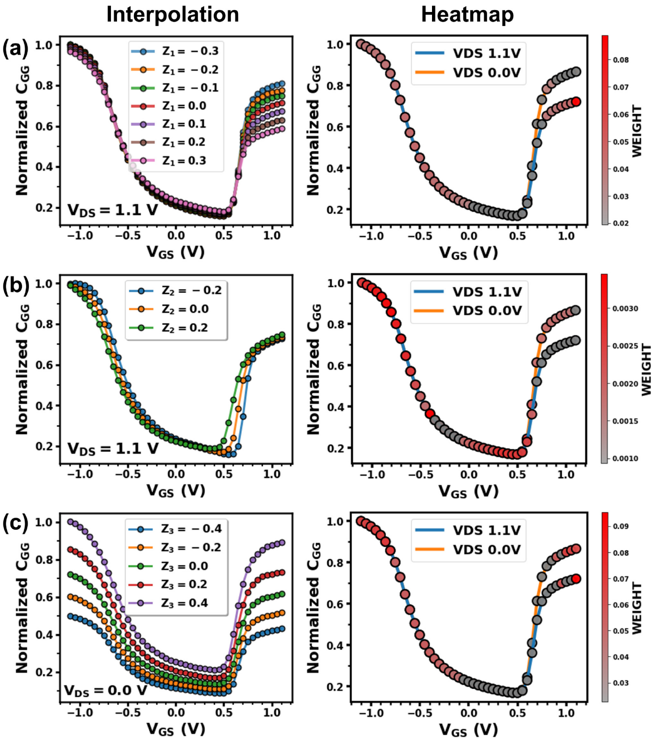

**FIGURE 4.** C-V Characteristics in Latent Space. Interpretations of each latent dimension: (a) $C_{GG}$ modulation in the saturation region, (b) Shifts in $V_{TH}$ and $V_{FB}$, (c) Scaling of $C_{GG}$.

analysis, $C_{GG}$ values in the plots are normalized based on their individual max values. However, for $z_3$, normalization is carried out using the device exhibiting the highest $C_{GG_{max}}$.

For $z_1$, as its value increases, there's a clear reduction in the capacitance values in the inversion, especially within the saturation region. This trend is influenced by various device characteristics. The oxide thickness, doping concentration, and channel length play a pivotal role in shaping the capacitance values in the inversion. The heatmap apparently emphasizes the saturation regions, with a marked focus at $V_{DS} = 1.1$ V. Moreover, the observed low $C_{GG}$ in saturation at high $V_{DS}$ can be explained by the device's transition into the saturation mode, leading to channel pinch-off and consequently, a reduced effective channel area contributing to $C_{GG}$.

As $z_2$ increases, we notice a concurrent reduction in both $V_{TH}$ and $V_{FB}$. This correlated behavior suggests that there are shared underlying factors influencing these changes, such as a change in the work function of the gate material, charges in the gate oxide, or variations in the substrate doping. All of these can affect both $V_{TH}$ and $V_{FB}$ at the same time. The model appears to focus more on the vicinity of $V_{FB}$ and $V_{TH}$, highlighting their importance in the latent dimension $z_2$.

In the case of $z_3$, increasing values bring about a clear rise in the magnitude of $C_{GG}$, while the overall shape, including features like $V_{TH}$ and $V_{FB}$, remains consistent across different $z_3$ values. This highlights the capability of our method in capturing a disentangled representation, allowing for straightforward interpretation of device characteristics curves. The encoder mainly attends to the regions with high capacitance, both in inversion and accumulation, which are crucial in determining the scales of unseen $C_{GG} - V_{GS}$ curves. Furthermore, since the magnitude of $C_{GG}$ is closely associated with the gate area, represented by $W \times L$, $z_3$ could potentially correspond to this *gate area* parameter, providing further insights into its influence on device characteristics.

In this section, we have examined the latent spaces of the I-V and C-V models, pinpointing how each dimension reflects specific attributes of semiconductor devices. Utilizing L1 regularization has resulted in a sparse yet disentangled representation, with each latent dimension reflecting a unique characteristic. The analysis through interpolation shows that the model generates curves that are both realistic and informative, aligning with known device behaviors. Heatmap analysis further reveals the selective focus of our neural networks, identifying the input data elements that are most influential in shaping the output. This investigation underscores our ability to elucidate the internal mechanisms of neural compact models, offering clear and interpretable insights that can enhance understanding in the field of device physics.

## V. DISCUSSION

We conducted additional experiments to explore the relationship between latent dimensions and the actual electrical parameters of semiconductor devices. We selected fifteen random devices from a 32nm technology node and encoded their input data into a latent space. Specifically, we focused on interpolating the $z_4$ latent dimension from $-0.3$ to $+0.3$, while keeping the other dimensions constant, effectively creating a series of virtual devices. We then measured the threshold voltages, $V_{TH_{LIN}}$ in the linear region with $V_{DS}$ at 0.1 V, and $V_{TH_{SAT}}$ in the saturation region with $V_{DS}$ at 1.1 V, for these virtual devices.

In Figure 5, each color represents a latent vector corresponding to a different actual device, ensuring that identical colors across both $V_{TH_{LIN}}$ and $V_{TH_{SAT}}$ indicate the same device. The data shows that both $V_{TH_{LIN}}$ and $V_{TH_{SAT}}$ exhibit a near-linear relationship with the $z_4$ latent dimension, suggesting a strong correlation. This observation suggests that $z_4$ plays a pivotal role in modulating the threshold voltage, a parameter of critical importance in defining device performance. Notably, for lower values of $z_4$, $V_{TH_{SAT}}$ is observed to be lower than $V_{TH_{LIN}}$ and this trend reverses as $z_4$ increases. This phenomenon can be attributed to DIBL, which becomes more pronounced in shorter-channel devices, indicating that with an increase in $z_4$, the virtual devices transition from short to long-channel characteristics.

## VI. CONCLUSION

In this study, we introduce a framework for NCMs that emphasizes interpretable and reliable simulations. Our approach involves regularizing latent spaces to achieve

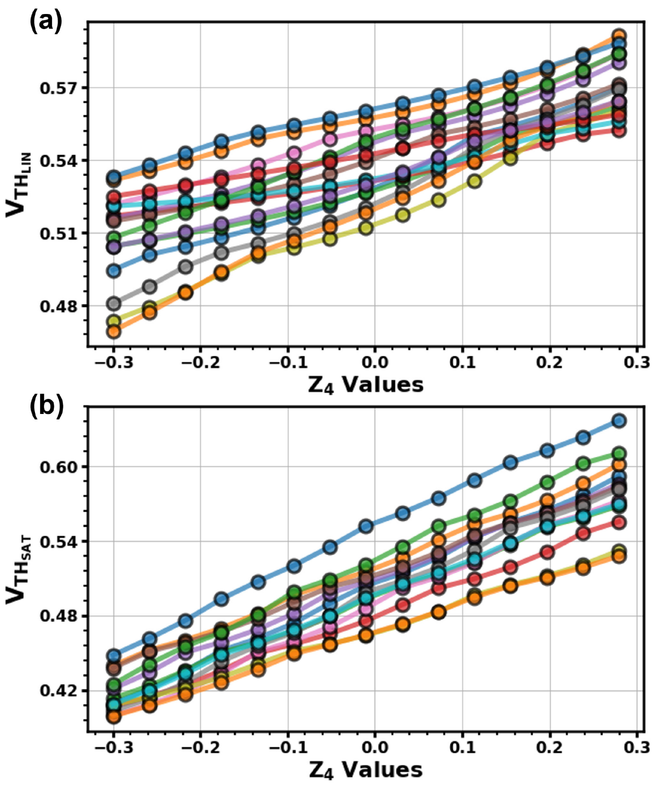

**FIGURE 5.** **Correlation between $z_4$ and Threshold Voltages: A linear trend in threshold voltages with $z_4$ confirms disentangled and interpretable representation.**

disentangled representations, which ensures a transparent understanding of model outcomes. By interpolating latent dimensions, we clarify their impact on the device's I-V and C-V characteristics, while heatmaps provide visual and intuitive insights into the NCMs' internal operations. Our work integrates explainable AI into device modeling, promoting the development of transparent and efficient NCMs that are trusted and widely adopted in Design-Technology Co-Optimization (DTCO), thereby advancing semiconductor device research.

## REFERENCES

[1] M.-Y. Kao, F. Chavez, S. Khandelwal, and C. Hu, "Deep learning-based BSIM-CMG parameter extraction for 10-nm FinFET," *IEEE Trans. Electron Devices*, vol. 69, no. 8, pp. 4765–4768, Aug. 2022.

[2] J. P. Duarte et al., "BSIM-CMG: Standard FinFET compact model for circuit design," in *Proc. 41st Eur. Solid-State Circuits Conf. (ESSCIRC)*, 2015, pp. 196–201.

[3] G. Gildenblat et al., "PSP: An advanced surface-potential-based MOSFET model for circuit simulation," *IEEE Trans. Electron Devices*, vol. 53, no. 9, pp. 1979–1993, Sep. 2006.

[4] A. Dasgupta et al., "BSIM compact model of quantum confinement in advanced nanosheet FETs," *IEEE Trans. Electron Devices*, vol. 67, no. 2, pp. 730–737, Feb. 2020.

[5] M. Li, O. İrsoy, C. Cardie, and H. G. Xing, "Physics-inspired neural networks for efficient device compact modeling," *IEEE J. Explor. Solid-State Computat. Devices Circuits*, vol. 2, no. 1, pp. 44–49, Dec. 2016.

[6] L. Zhang and M. Chan, "Artificial neural network design for compact modeling of generic transistors," *J. Comput. Electron.*, vol. 16, pp. 825–832, Sep. 2017.

[7] 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. 2020.

[8] J. Wang, Y.-H. Kim, J. Ryu, C. Jeong, W. Choi, and D. Kim, "Artificial neural network-based compact modeling methodology for advanced transistors," *IEEE Trans. Electron Devices*, vol. 68, no. 3, pp. 1318–1325, Mar. 2021.

[9] K. Sheelvardhan, S. Guglani, M. Ehteshamuddin, S. Roy, and A. Dasgupta, "Machine learning augmented compact modeling for simultaneous improvement in computational speed and accuracy," *IEEE Trans. Electron Devices*, vol. 71, no. 1, pp. 239–245, Jan. 2024.

[10] C. Park, P. Vincent, S. Chong, J. Park, Y. S. Cha, and H. Cho, "Hierarchical mixture-of-experts approach for neural compact modeling of MOSFETs," *Solid-State Electron.*, vol. 199, Jan. 2023, Art. no. 108500.

[11] Y. S. Cha et al., "A novel methodology for neural compact modeling based on knowledge transfer," *Solid-State Electron.*, vol. 198, Dec. 2022, Art. no. 108450.

[12] H. Shao, L. Wang, R. Chen, H. Li, and Y. Liu, "Safety-enhanced autonomous driving using interpretable sensor fusion transformer," in *Proc. Conf. Robot Learn.*, 2023, pp. 726–737.

[13] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "MDNet: A semantically and visually interpretable medical image diagnosis network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 3549–355.

[14] Y.-S. Yang, Y. Li, and S. R. Kola, "A physical-based artificial neural networks compact modeling framework for emerging FETs," *IEEE Trans. Electron Devices*, vol. 71, no. 1, pp. 223–230, Jan. 2024.

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*,, 2017, pp. 618–626.

[16] S. Yao et al., "ReAct: Synergizing reasoning and acting in language models," in *Proc. ICLR*, 2022 pp. 1–33.

[17] Y. Kim, S. Myung, J. Ryu, C. Jeong, and D. S. Kim, "Physics-augmented neural compact model for emerging device technologies," in *Proc. Int. Conf. Simul. Semicond. Process. Devices (SISPAD)*, 2020, pp. 257–260.

[18] C.-T. Tung, M.-Y. Kao, and C. Hu, "Neural network-based and modeling with high accuracy and potential model speed," *IEEE Trans. Electron Devices*, vol. 69, no. 11, pp. 6476–6479, Nov. 2022.

[19] Q. Yao et al., "A novel prediction technology of output characteristics for IGBT based on compact model and artificial neural networks," *IEEE Trans. Electron Devices*, vol. 70, no. 9, pp. 4885–4891, Sep. 2023.

[20] M.-Y. Kao, H. Kam, and C. Hu, "Deep-learning-assisted physics-driven MOSFET current-voltage modeling," *IEEE Electron Device Lett.*, vol. 43, no. 6, pp. 974–977, Jun. 2022.

[21] *International Technology Roadmap for Semiconductors (ITRS)*, Semicond. Ind. Assoc., Washington, DC, USA, 2007.