Machine Learning-augmented High-efficient TCAD on Accurate Characteristics of Gate-all-around Transistors with Quantum Effect

Haoqing Xu[1,2], Guohui Zhan[1,2], Shixin Li[1,2], Jiahao Wu[2,3], Kun Luo[1], Yu Liu[4,*], Chao He[5], and Zhenhua Wu[1,2,*]
[1]Institute of Microelectronics, Chinese Academy of Sciences, 100029 Beijing, China;
[2]University of Chinese Academy of Sciences, 100049 Beijing, China;
[3]Institute of Computing Technology, Chinese Academy of Sciences, 100190 Beijing, China;
[4]Inspur Group Co., Ltd, 250000, Jinan, China;
[5]HiSilicon Technology Co., Ltd, 201206 Shanghai, China;
*email: wuzhenhua@ime.ac.com, liuyubj@inspur.com;

***Abstract*—A machine learning (ML)-augmented TCAD framework is proposed to build an adaptive density gradient (DG) model for the ultra-scaled gate-all-around (GAA) devices. First, to capture the impact of quantum confinement on Silicon nanowire and nanosheet GAA FETs, the multi-subband k·p model is calibrated with the first-principles calculations. In parallel, a fully-connected multi-layer neural network, i.e., Multilayer Perceptron (MLP), is trained to learn the empirical quantum correction potential parameters from drift-diffusion-equation (DD) based Technology Computer-Aided Design (TCAD). Then the MLP is incorporated into the ML-augmented TCAD framework to obtain an adaptive DG model regarding to the prepared k·p results. This ML-augmented adaptive DG model extend the scope of the applications of the DD based TCAD, approaching the level of the subband Schrodinger Poisson solver.**

***Index Terms*—Gate-all-around Transistor, Machine Learning, Quantum Effect**

I. Introduction

The Silicon nanosheet GAA transistor is a promising candidate for the sub-3-nm technology node due to its good performance in gate control. However, the ultra-scaled nanosheet GAA transistor exhibits severe quantum effects, which challenge the accuracy of empirical DD TCAD. The DG model, a quantum correction model, has been introduced to provide reasonable electron statistics for moderate-size FinFET or GAAFET channels. However, as the channel size decreases, calibration becomes more tedious, and artificial overfitting may occur. The current quantum correction models do not update empirical parameters timely, which can result in reduced quantum effects [1], [2].

To address this issue, a ML-augmented TCAD framework has been proposed to build an adaptive density gradient (DG) model. The framework is geometry-aware and stress-aware, and the k·p numerical simulation is calibrated from the first-principles calculation. Furthermore, the proposed framework significantly reduces computational cost and uncovers intrinsic knowledge of the reference k·p results.

This work was supported in part by the MOST under Grant 2021YFA1200502 and the NSFC under Grant 12174423.

II. Experiments and Discussion

The proposed ML-augmented TCAD framework is shown in Fig. 1. The workflow is composed of three phases: Hamiltonian Phase, Train Phase and Inference Phase. A multi-layer perceptron (MLP) is employed to learn the quantum correction potential related parameters from the carrier density profile in the cross-sections of GAA devices with various shapes and/or different stress. Once the training finished, the NN could infer a set of proper parameters for DG model from a carrier density profile calculated using more accurate theory like k·p. With the updated DG models parameters, new carrier density profiles can be generated by TCAD and data augmentation in standard self-supervised learning can be realized. By minimalizing the loss function between the original carrier density profile from k·p theory and the TCAD iteratively, an adaptive DG model is obtained for each bias, geometry and stress, et.al.

A. Hamiltonian Phase

The Hamiltonian is one of the most important physical quantities in nanostructures. For a material like Silicon, first-principles calculations are performed to obtain accurate band dispersion. Then an appropriate band model, i.e., the DKK model of the valence band, is chosen to obtain the k·p Hamiltonian with tunned parameters [3]. The effect of strain, electric, and magnetic fields also be included in the model. The effect of strain, electric, and magnetic fields can also be included in the model. Then, the k·p Hamiltonian is discretized along some surface or direction. Finally, we obtain the pseudo-tight binding Hamiltonian adaptive to the same mesh used in TCAD simulations. The two cross-sections with different transport direction along $\langle 110 \rangle$ and $\langle 100 \rangle$ are shown in Fig. 2(a) and (b). Also, we calculate the band structure along$\langle 110 \rangle$ and $\langle 100 \rangle$ directions of different sizes of Si nanowires, and present the band comparison of the k·p method and DFT results. Then, the k·p Hamiltonian is validated. The results in Fig. 2 indicate that the geometry dependence of the quantum effect is not neglected for ultra-scaled devices.

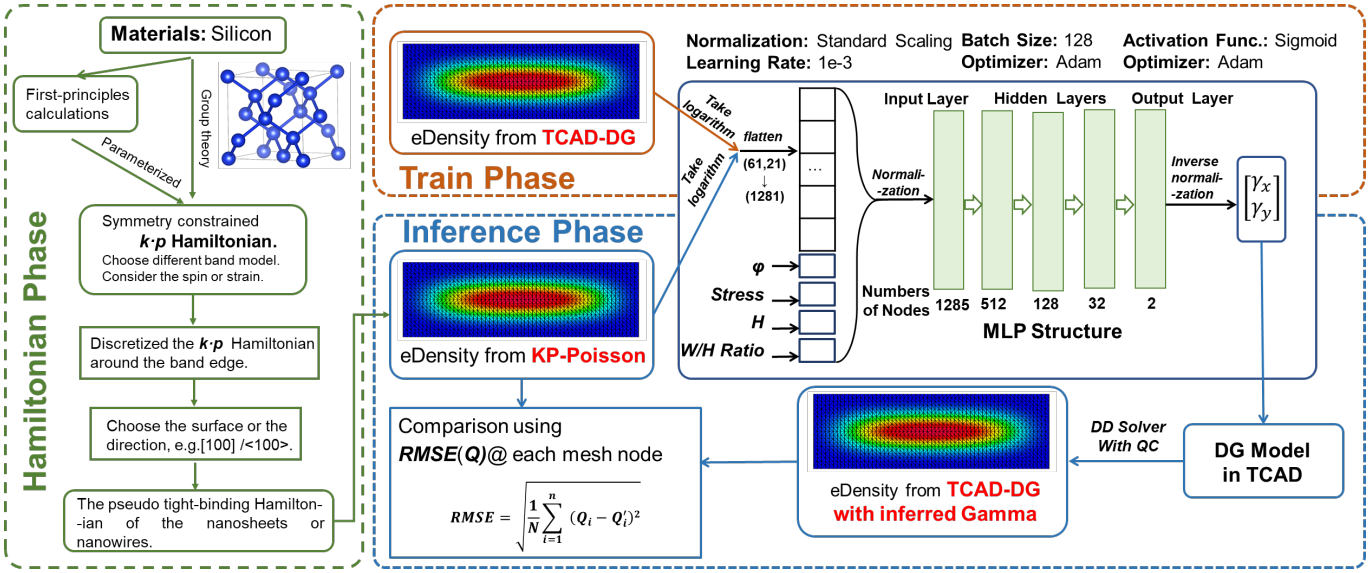


Fig. 1. Schematic of overall ML-augmented TCAD framework. The workflow is composed of three phases: Hamiltonian Phase, Train Phase and Inference Phase.

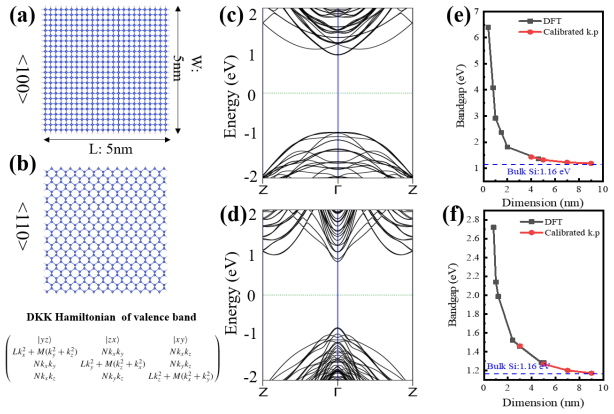


Fig. 2. The cross-section of Si nanowires along (a) ⟨110⟩ and (b) ⟨100⟩ directions. (c)-(d) denote the nanowires band structures of 2 nm diameter. (e)-(f) The bandgap as a function of dimension, black and red denote the DFT and TCAD results respectively. The HSE06 level is used in these DFT calculations.

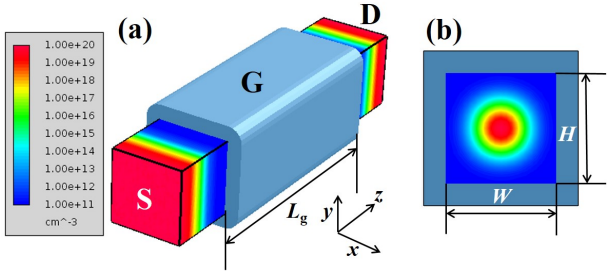


Fig. 3. (a) The proposed gate-all-around device in this work and (b) the cross-section of the proposed device simulated by the Poisson-k.p solver. Undoped channel is set and the metal gate Fermi energy is set to align with the middle line of the gap of the Silicon.

B. Data Generation

Nanowire n-type FETs with L_g=15 nm and oxide thickness T_{ox} of 1 nm are built using the GTS framework (Fig. 3(a)), the cross-section of which is also generated for device simulations (see Fig. 3(b)). The nanowire height H, width ratio W/H, channel stress, and bias condition are split for device simulations. The nanowire height H is split from 3 nm to 9 nm with a 2 nm increment for each step. The width-height ratio varies from 1.0 to 3.0 for device simulations. The channel stress is applied along the transport direction (z-axis) and ranges from 0 to 0.8 GPa. The gate voltage is swept from 0 to 0.8 V with a 0.2V increment for each step.

The DG model with anisotropy empirical parameters γ is employed for quantum correction potential λ, which is added to the DD model,

$$J_n = q\cdot\mu_n\cdot n\cdot(grad(\frac{\epsilon_C}{q}-\psi-\lambda_n)+\frac{k_BT}{q}\cdot\frac{N_{C,0}}{n}\cdot grad(\frac{n}{N_{C,0}})) \quad (1)$$

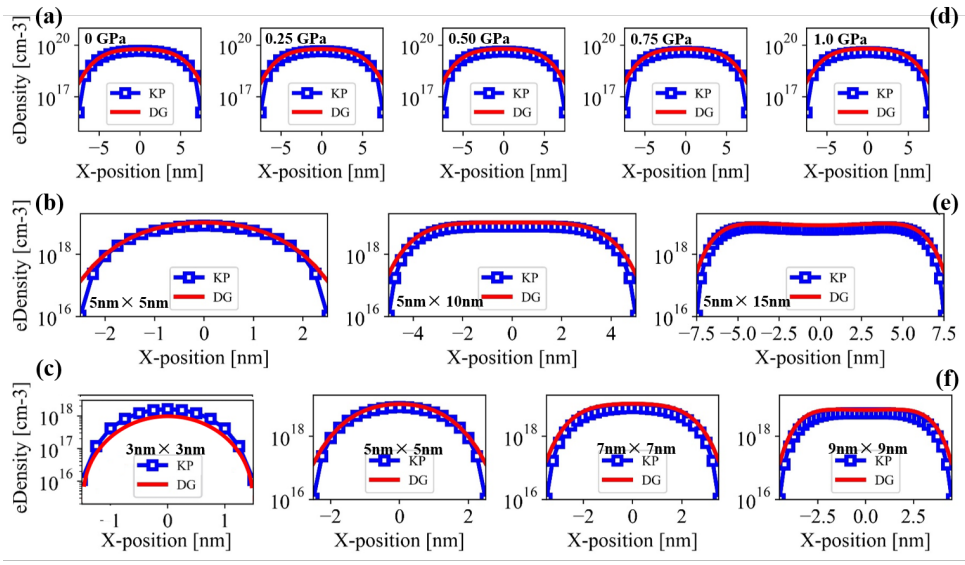
The simplified first order approximation of the quantum potential derived from Wigner's equation is employed as

$$\lambda_n = \frac{\hbar^2}{12\cdot\lambda_n\cdot m_0}\cdot divgrad\frac{\psi+\gamma_n-\epsilon_C/q}{k_BT} \quad (2)$$

The two component parameters of γ, that is, γ_x and γ_y, are split from 0.1 to 0.5 with 0.1 increment each step for dataset generation.

C. Neural Network Training and Test

As the dataset is generated from TCAD simulation using drift-diffusion equation with DG model, the electron density profile on device cross-section is extracted with corresponding parameters including surface potential ϕ, stress, nanosheet height H, and width-height ratio W/H. ϕ is extracted from the



	Stress [GPa]	RMSE
5nm×5nm Vg=0.8V	0.00	4.2×10^{18}
	0.25	5.2×10^{18}
	0.50	6.1×10^{18}
	0.75	7.0×10^{18}
	1.00	7.8×10^{18}

	W/H	RMSE
H=5nm Vg=0.8V Stress=0.6GPa	1.0	4.4×10^{17}
	2.0	6.4×10^{17}
	3.0	7.3×10^{17}

	H [nm]	RMSE
W/H = 1 Vg=0.8V Stress=0.6GPa	3.0	3.2×10^{17}
	5.0	4.4×10^{17}
	7.0	6.8×10^{17}
	9.0	7.6×10^{17}

Fig. 4. The 1-D cut of electron density profile of devices along X-direction at the center position of H under (a) 0 to 0.8 GPa tensile stress, (b) 1.0 to 3.0 width-height ratio, and (c) 3.0 to 9.0 nanowire height. The corresponding RMSE of charge profiles between k.p solutions and DG with inferred gamma solutions at each mesh node is shown in (d)(e)(f).

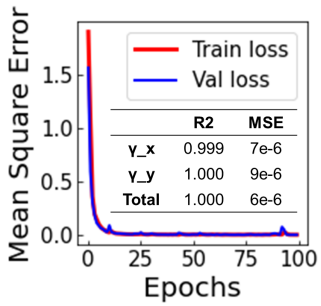


Fig. 5. MSE on the train set and validation set in the MLP training and inserted table shows the performance of the MLP on the test set.

potential at the position 0.1 nm under the GAA channel/oxide interface, which is related to V_g, work function W_f, and channel doping concentration N_{ch}.

The data pre-processing is required for better training performance, composed of two steps. First, the electron density profile is taken logarithm. Second, the profile matrix is flattened into a 1-dimensional vector. Together with four parameters ϕ, stress, H, and W/H, the normalization is performed with a standard scaler, where centering and scaling take place independently on each feature. After data preprocessing, the dataset is split into 60%, 20%, and 20% for the train set, validation set, and test set respectively [4], [5]. The MLP structure is shown in Fig. 1, which is composed of one input layer with 1285 nodes, one output layer with 2 nodes, and three hidden layers with 512, 128, and 32 nodes respectively. The MLP is implemented in Python with the Pytorch framework and trained using the back-propagation algorithm. The mean square error (MSE) is used as the loss function in the training and optimized by the Adam optimizer with a learning rate of 1.0×10-3. The activation function is sigmoid and the batch size is 128. After 100 epochs, the MSE on the train set converges and the performance of the MLP is evaluated on the test set as shown in the inserted table in Fig. 5. The coefficient of determination R^2 reaches 1.0. Total MSE reaches 6×10^{-6}, which indicates the MLP is capable to predict accurate quantum effect-related empirical parameters γ_x and γ_y.

D. Inference Phase

The trained ANN is now used as a DG-parameter solver as shown in Fig. 1. A cross-section of the proposed device is solved using a self-consistent loop with the Poisson equation coupled with the k·p method, from which the electron density profile could be obtained as input to the trained ANN. In this way, the corresponding DG parameters γ_x and γ_y are inferred by the ANN. The electron density profile calculated by the DG model with inferred γ_x and γ_y see (red lines in Fig. 4(a)(b)(c)) are compared with that calculated by the k·p theory (see blue dots in Fig. 4(a)(b)(c)) in terms of 1-D cut distribution along the x-direction, which indicates a good prediction is achieved for DG parameters. The performance of the proposed framework is verified by the following two parts:

1) Stress-aware: Firstly, the proposed framework is tested under different channel stress. The 5nm×5nm cross-section of GAA devices are used. The γ inference process is similar to the process mentioned above. The stress condition range from 0 to 1.0 GPa and the bias condition is 0.8V. The accurate 1D-cut profiles (blue dots in Fig. 4(a)) and predicted profile (red lines in Fig. 7(a)) match well, and the RMSE of charge at each mesh node is relatively small around 10^{18} (see Fig. 4(d)).

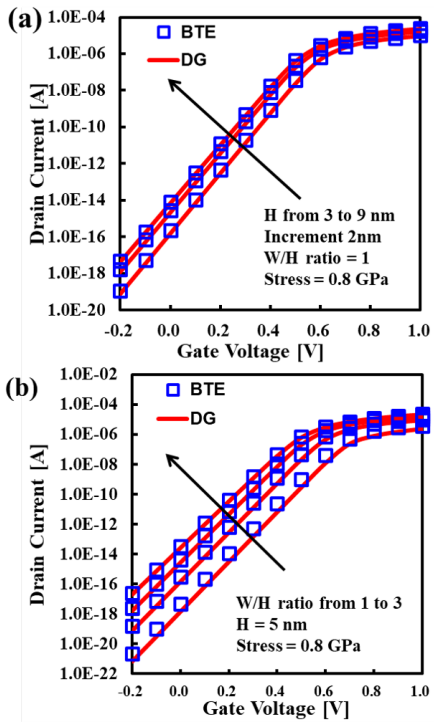


Fig. 6. BTE simulation results and the proposed DG results of GAA devices with various channel geometry (a) Nanowire diameter rises from 3 to 9 nm and (b) W/H ratio increases from 1 to 3.

2) Geometry-aware: The geometry dependence of the framework is also checked using the GAA devices with various width-height ratios and nanowire diameters. As we could see from Fig. 4(b), the electron density profile of three GAA devices of H=5 nm and W=5, 10, and 15 nm are compared under the tensile stress of 0.8 GPa and the gate voltage of 0.6V. A good prediction of electron concentration is performed see Fig.4(b) and (e). Besides, the nanowire devices with diameters of 3, 5, 7, and 9 nm are simulated and analyzed using the same approach as shown in Fig. 4(c)(f).

These results show that the proposed neural network exhibits stress-aware and geometry-aware accuracy for quantum confinement effect prediction. Our ML-augmented framework works well for quantum correction by finding adaptive parameters of the DG model, which is demonstrated on both wide cross-section devices and ultra-scaled cross-section (the smallest cross-section is only 3nm×3nm).

E. Discussion

To check our framework in 3-D TCAD simulation, an L_g=15 nm nanowire n-type FET is employed for comparison of Multi-Subband Boltzman Equation (MSBTE) simulation and the proposed ML-augmented DD simulation with inferred adaptive DG model. The I-V curve is shown in Fig. 6(a) and (b), where nanowire diameter varies in (a) and the W/H varies in (b). The good match between I-V from the two solvers indicates that our proposed ML-augmented TCAD framework could do accurate characteristics of quantum effect in the state-of-art GAA devices with much less computational cost than MSBTE method.

III. Conclusion

A ML-augmented TCAD simulation framework, that can effectively extend the scope of application of DD based TCAD for ultra-scaled GAA transistors, is proposed. It is able to capture quantum confinement accurately by constructing an adaptive DG model, which is demonstrated to be both stress-aware and geometry-aware. The I-V characteristics match well with the accurate but more time-consuming MSBTE method.

References

[1] Neophytou, Neophytos, et al. "Bandstructure effects in silicon nanowire electron transport." IEEE Transactions on Electron Devices 55.6 (2008): 1286-1297.

[2] Dasgupta, Avirup, et al. "Compact modeling of cross-sectional scaling in gate-all-around FETs: 3-D to 1-D transition." IEEE Transactions on Electron Devices 65.3 (2018): 1094-1100.

[3] Dresselhaus, Gene. "Spin-orbit coupling effects in zinc blende structures." Physical Review 100.2 (1955): 580.

[4] Han, Seung-Cheol, Jonghyun Choi, and Sung-Min Hong. "Acceleration of semiconductor device simulation with approximate solutions predicted by trained neural networks." IEEE Transactions on Electron Devices 68.11 (2021): 5483-5489.

[5] Myung, Sanghoon, et al. "Restructuring TCAD System: Teaching Traditional TCAD New Tricks." 2021 IEEE International Electron Devices Meeting (IEDM). IEEE, 2021.