

Physical NN Compact Model for Flexible Modeling

Jen-Hao Chen
BSIM Group, UC Berkeley

Outline

- Introduction
- Physical Neural Network Model
- Results
- Conclusion



Introduction

- Neural network (NN) model is a promising alternative for transistor compact modeling.
- Pure data-driven NN models face challenges in terms of fitting flexibility and variability.
- A physical and accurate gate current model is lacking in many NN models.
- Training NN models across a wide range of transistor geometries has been rarely discussed in the previous literature.
- A physical neural network (PNN) compact model is proposed to overcome these challenges.



Physical NN Model of I_D

- By drawing an analogy to the drain current equation, we can derive a physical pre-processing equation for the drain current data.

$$I_D = W\mu \frac{dV}{dy} Q \approx W\mu \frac{V_{DS}}{L} Q_0 e^{\frac{q\phi}{kT}} \longrightarrow I_D = W\mu_r \frac{V_{DS}}{1 + \alpha L} e^{y_1}$$

- y_1 is a trainable output of the NN, and α is a pre-training parameter used to capture the non-ideal trend with respect to L .

Input:

$(V_{GS} - V_T)/mT$

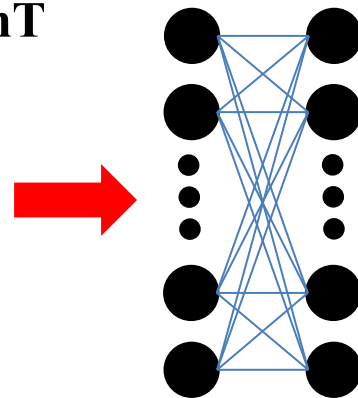
V_{DS}

L

W

EOT

T



I_D network

Output: y_1



$(V_{GS} - V_T)/mT$ is used to consider threshold voltage and subthreshold swing.

Physical NN Model of I_G

- A simple I_G model is developed to ensure zero current at zero bias.

$$I_G = WT^\beta (V_{GS} - \gamma V_{DS}) e^{y_2}$$

- y_2 is a trainable output of the NN. β and γ are pre-training parameters.

Input:

V_{GS}

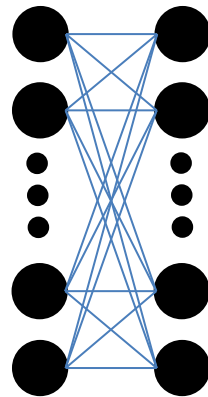
V_{DS}

L

W

EOT

T



I_G network

Output: y_2

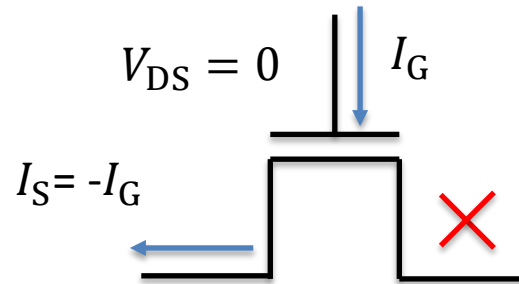


Gate Current Partition (I)

- The gate current should be equally partitioned at $V_{DS} = 0V$. ($I_{GD} = I_{GS}$)
- However, in the previously released Verilog-A code,

$$I(d, s) < + I_D \quad \text{where} \quad I_D = W\mu_r \frac{V_{DS}}{1 + \alpha L} e^{y_1} = 0$$

$$I(g, s) < + I_G \quad \text{where} \quad I_G = WT^\beta (V_{GS} - \gamma V_{DS}) e^{y_2}$$



- The gate current directed to drain side at $V_{DS} = 0V$ is forced to be 0 because V_{DS} appears in the numerator.



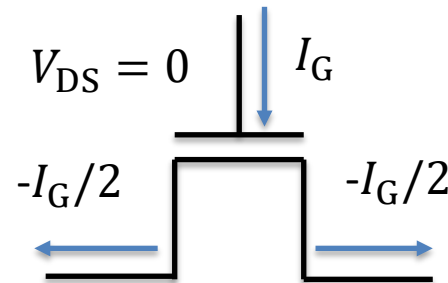
Gate Current Partition (II)

- A smoothing factor is introduced to guarantee equal I_G partition at $V_{DS} = 0$ V.

$$I(d, s) < + I_D$$

$$I(g, d) < + I_G \frac{\exp(-KV_{DS}^2)}{2}$$

$$I(g, s) < + I_G \left(1 - \frac{\exp(-KV_{DS}^2)}{2}\right)$$



- K is a large value such that the smoothing factor approaches to 0 when V_{DS} is not 0V.



Physical NN Model of QV

- The Q preprocessing equations are separated into intrinsic and parasitic components.

$$Q_G = WLy_{3G} + WC_p(2V_{GS} - V_{DS})$$

$$Q_D = -WLy_{3D} - WC_p(V_{GS} - V_{DS})$$

$$Q_S = -Q_G - Q_D$$

Input:

V_{GS}

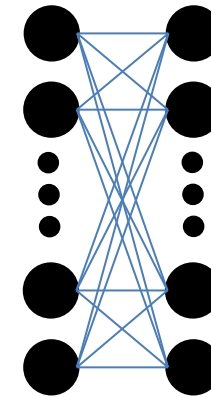
V_{DS}

L

W

EOT

T



Q network



Output:

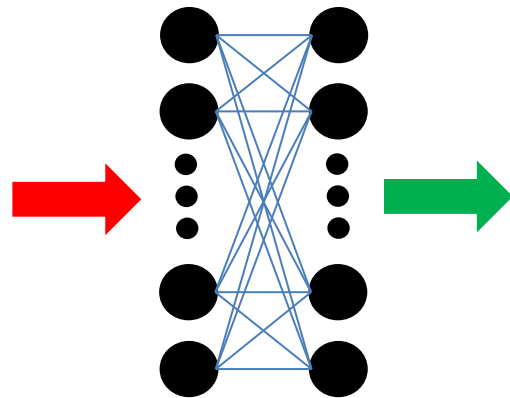
y_{3G}, y_{3D}

- $y_{3G(D)}$ is a trainable output representing the intrinsic charges.
- C_p is the parasitic capacitance parameter that can be used for variability modeling.

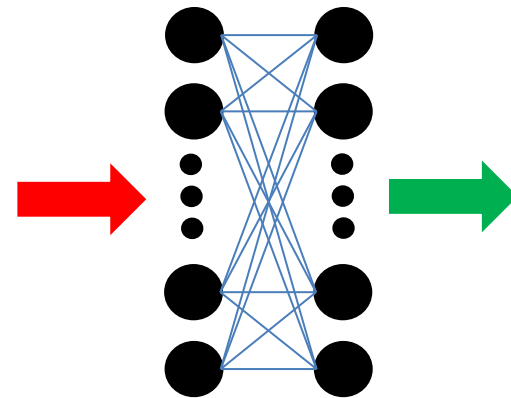


Large Scale Training

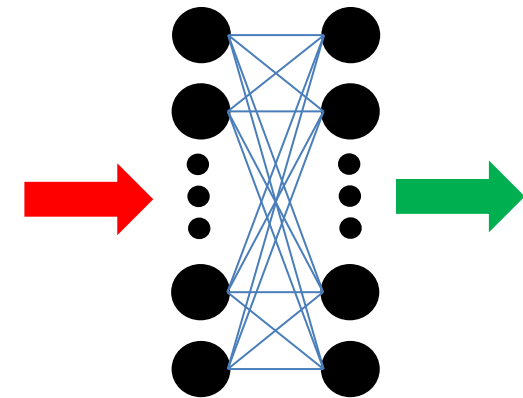
- The accuracy of NN models trained over a large geometry range is usually limited.
- We developed a PNN model covering a wide range of transistor geometry ($L = 12 \sim 302 \text{ nm}$, $W = 10 \sim 50 \text{ nm}$) by using multiple bin networks.



Short L network
(12 ~ 27.5nm)



Medium L network
(27.5 ~ 125nm)



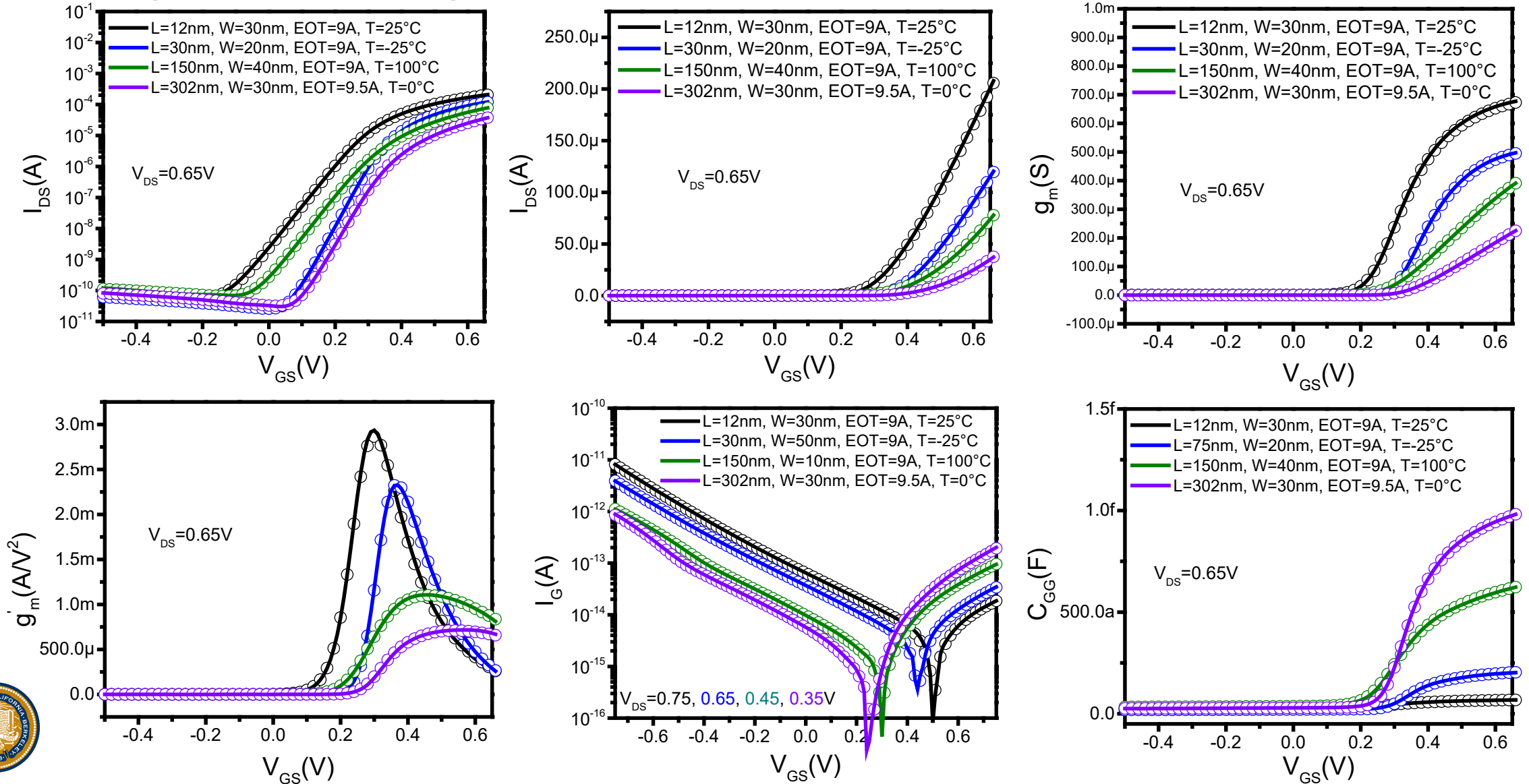
Long L network
(125 ~ 302nm)

- The training dataset is BSIM-CMG data calibrated to IRDS 1.5nm GAAFET. [1]



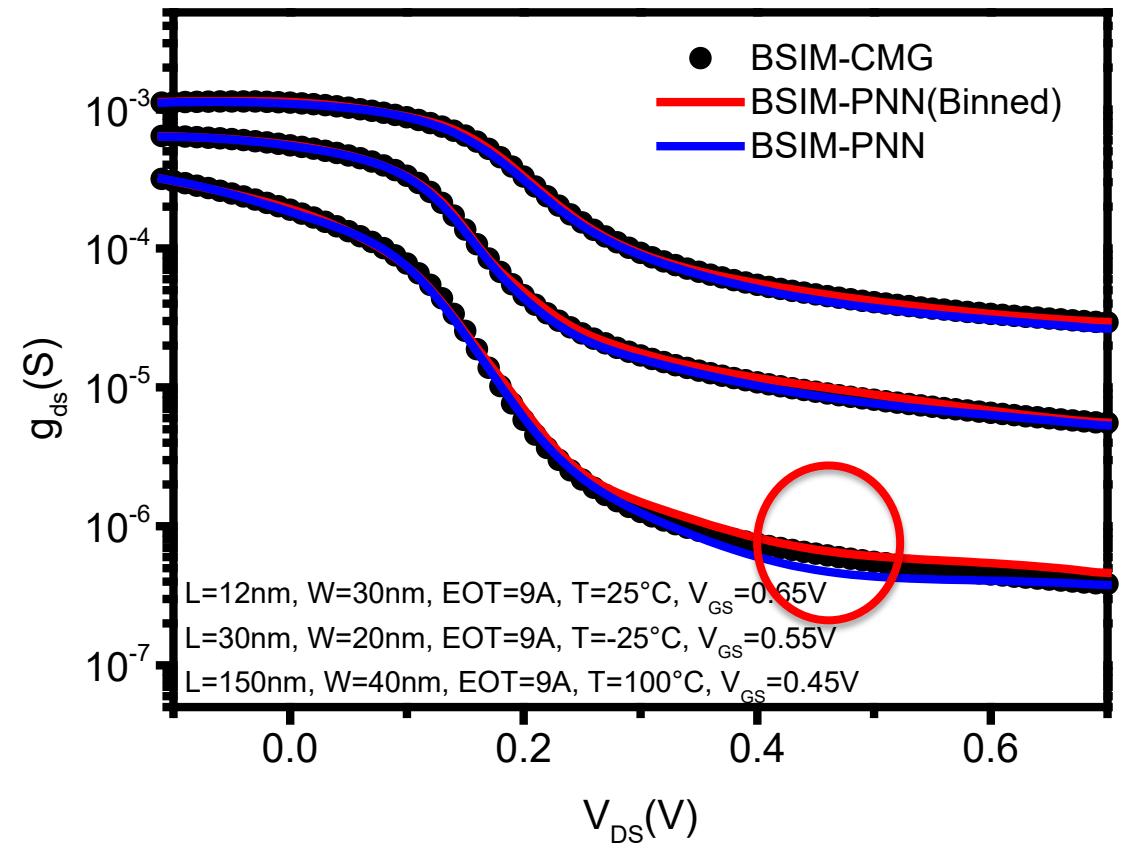
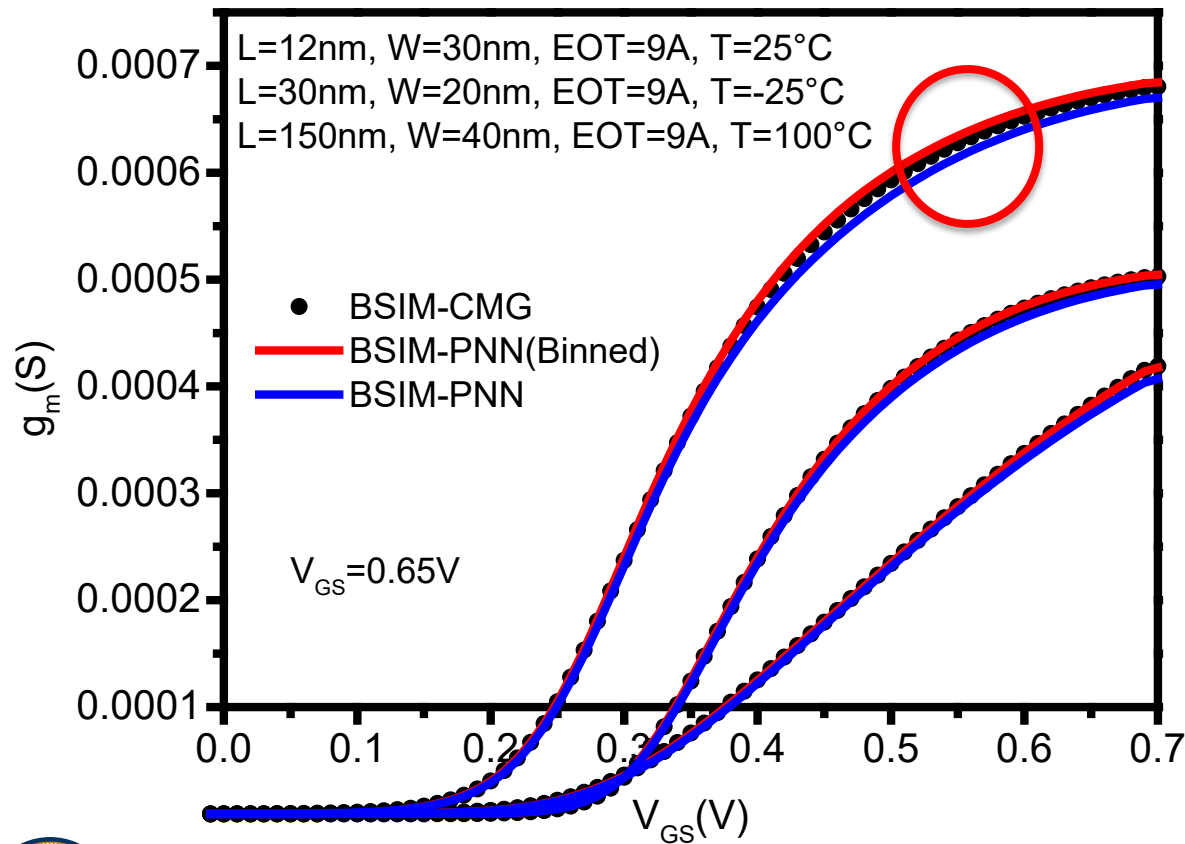
Model Fitting Result (I)

- The PNN model with multiple bins can achieve high accuracy across a wide range of transistor geometries.



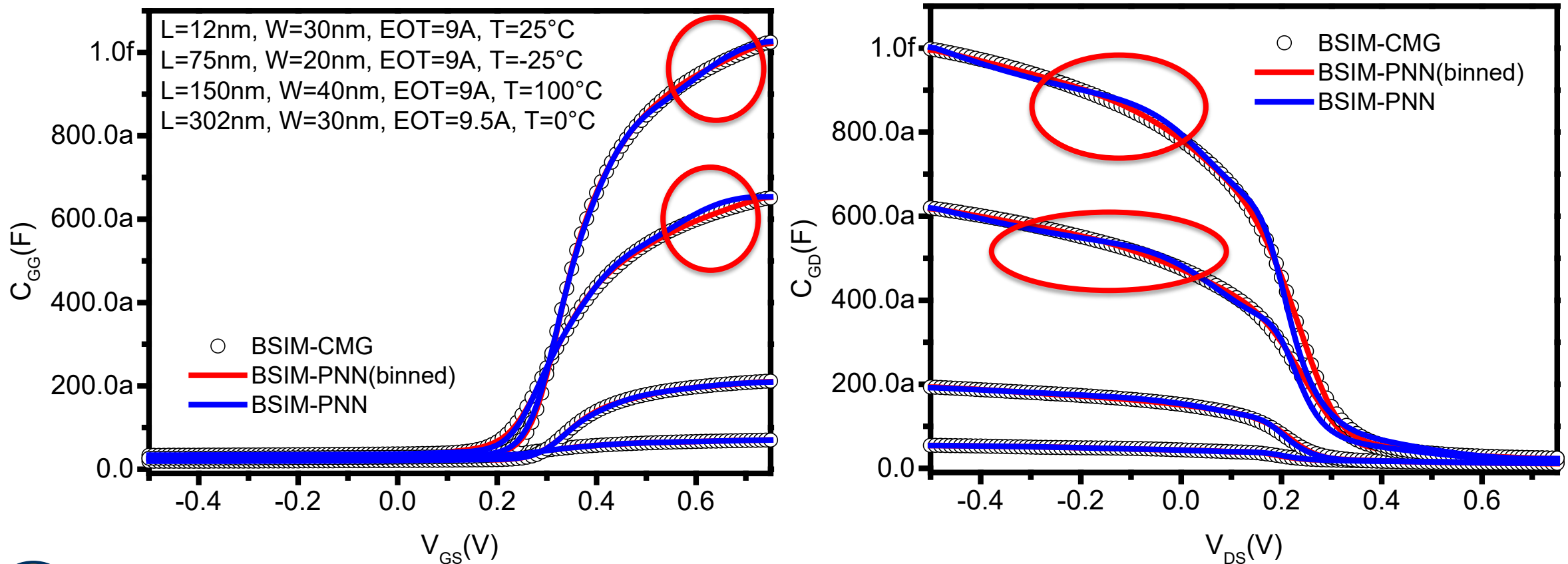
Model Fitting Result (II)

- Compared to a single bin, using multiple bins can achieve higher accuracy in g_m , g_{ds} and CV characteristics with the same network size.



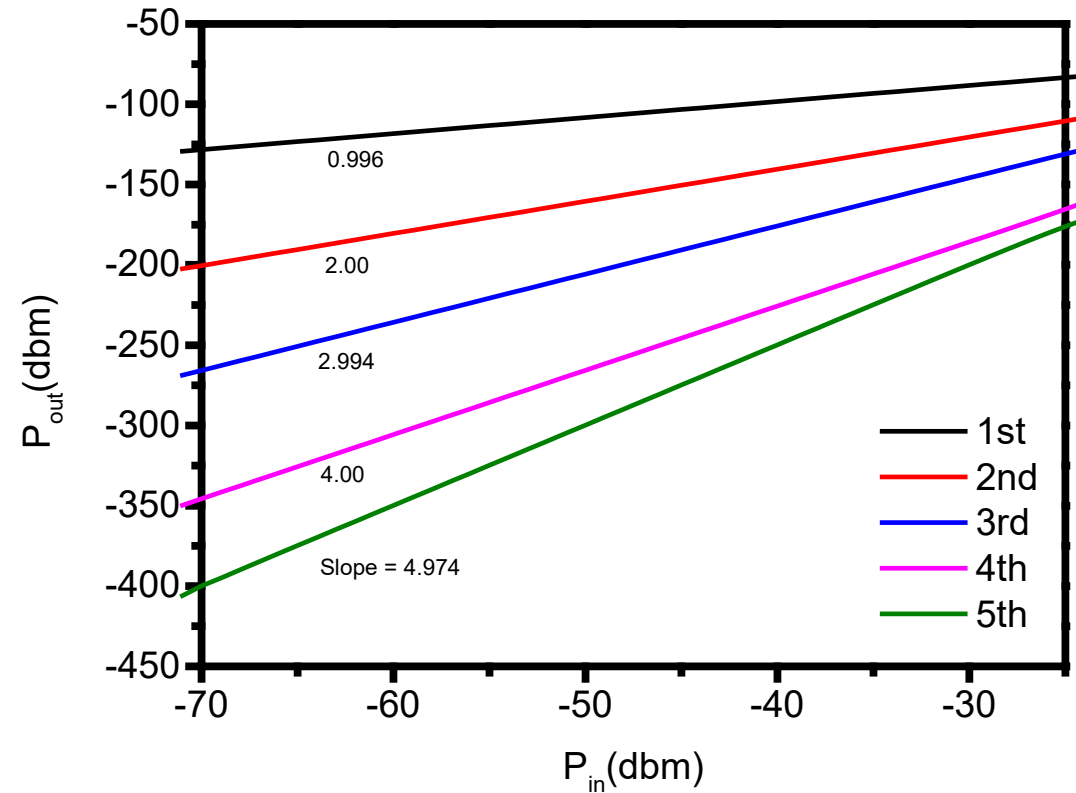
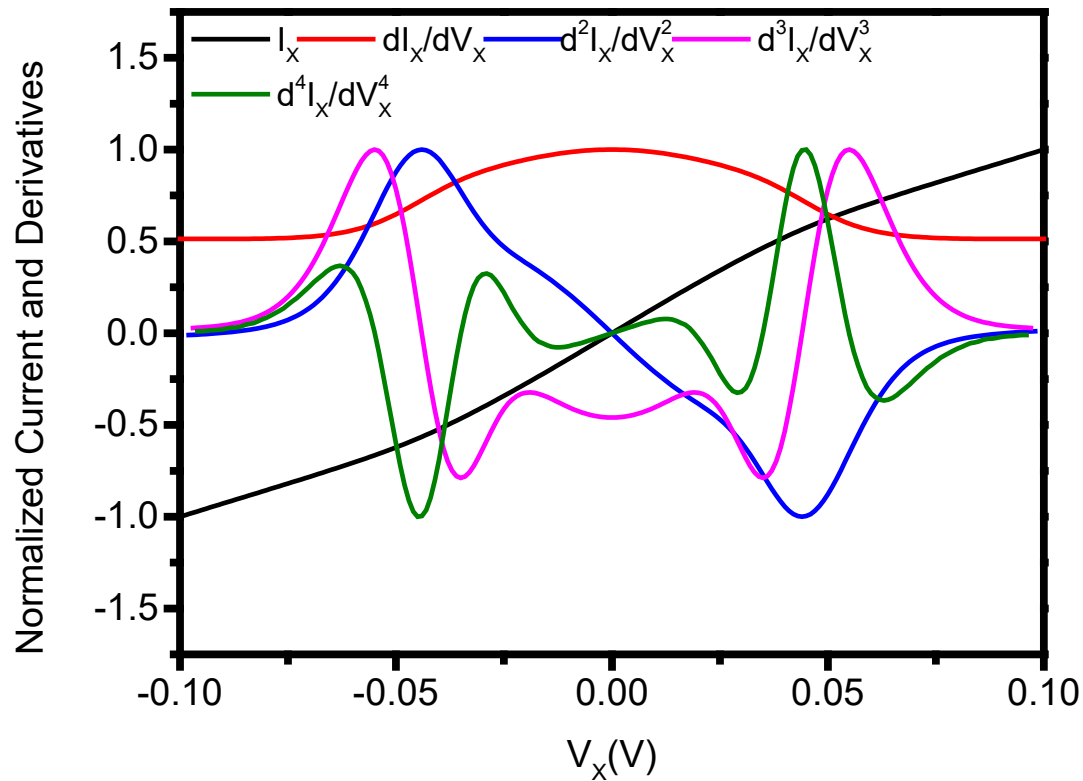
Model Fitting Result (III)

- Compared to a single bin, using multiple bins can achieve higher accuracy in g_m , g_{ds} and CV characteristics with the same network size.



Benchmark Tests

- Both Gummel symmetry test and Harmonic balance test are performed on this PNN model.



Post-Training Processing

- The PNN model includes post-training parameters for more fitting flexibility and variability modeling.
- We use the following factor as a drain-current multiplier to account for variations in key physical parameters, such as mobility, saturation velocity and channel length modulation.

$$M = \frac{\mu_{r0} \sqrt{1 + DSAT \cdot \phi_{DS} \cdot (1 + DLAMBDA \cdot V_{DS})}}{1 + DUA \cdot E_{\perp} + \frac{DUD}{1 + qig}}$$

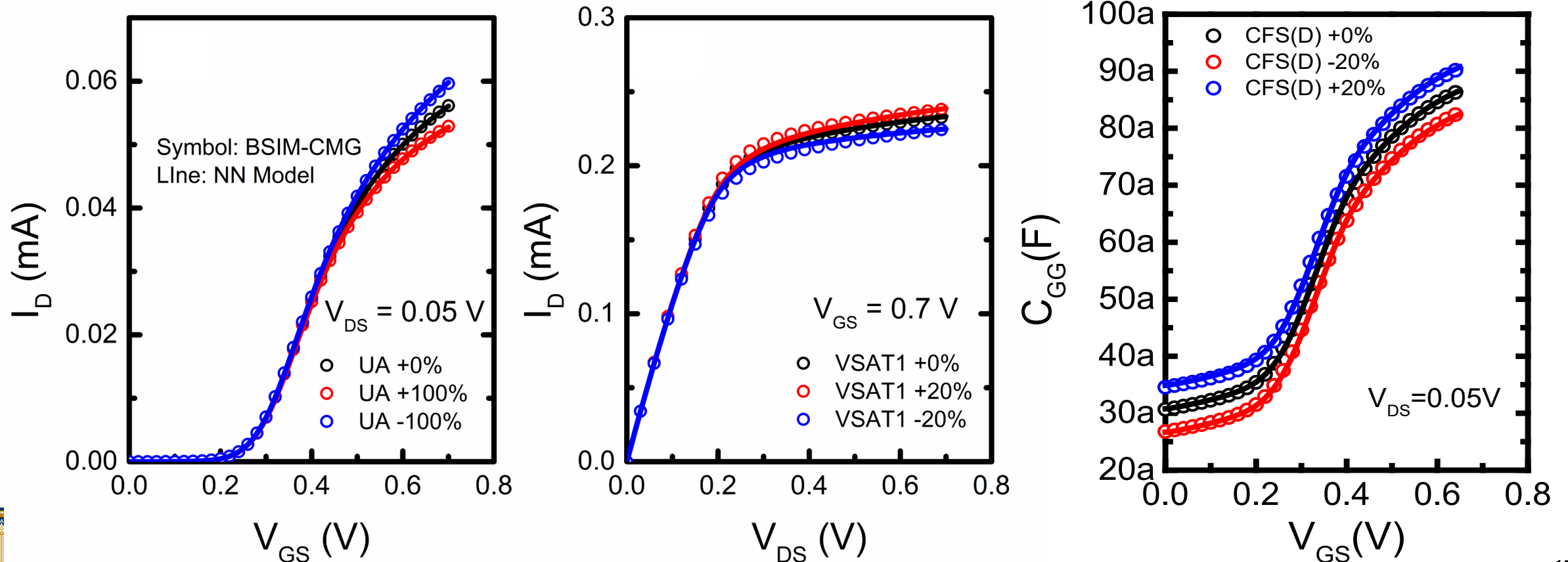
- A V_{TH} shift is added to the V_{GS} input to account for flexibility in DIBL effect.

$$\Delta V_{TH} = -DETA \cdot V_{DS}$$



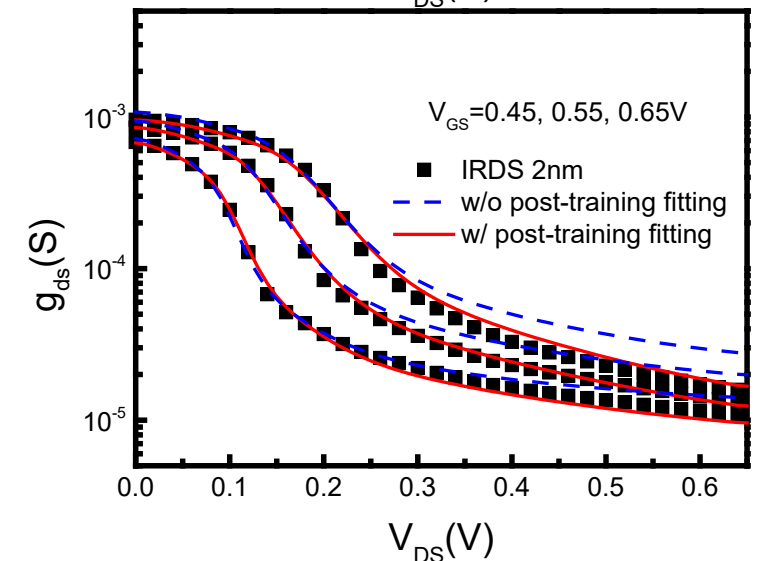
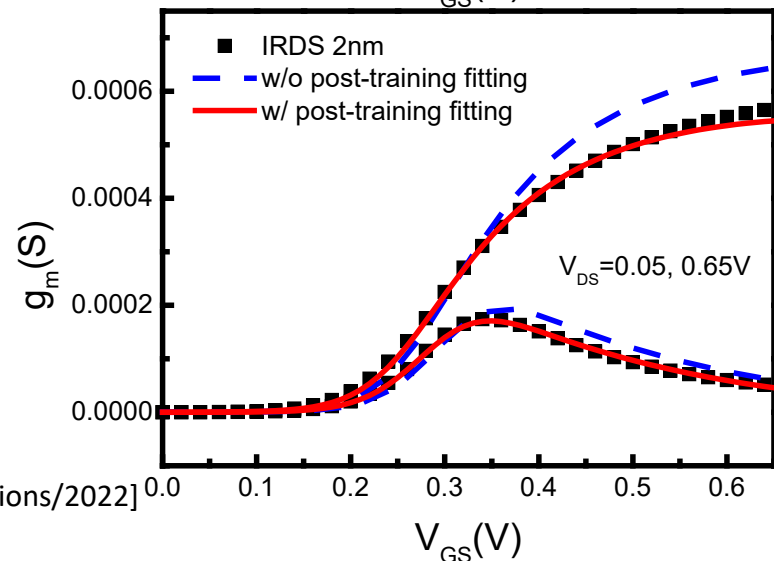
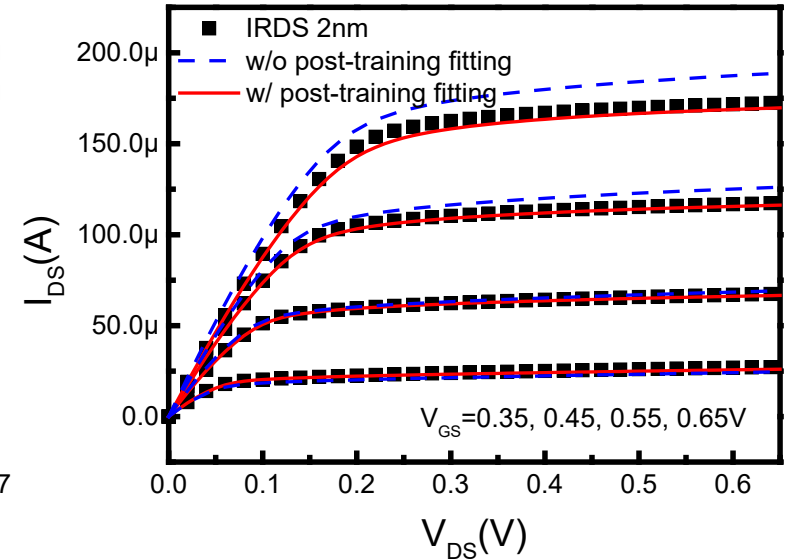
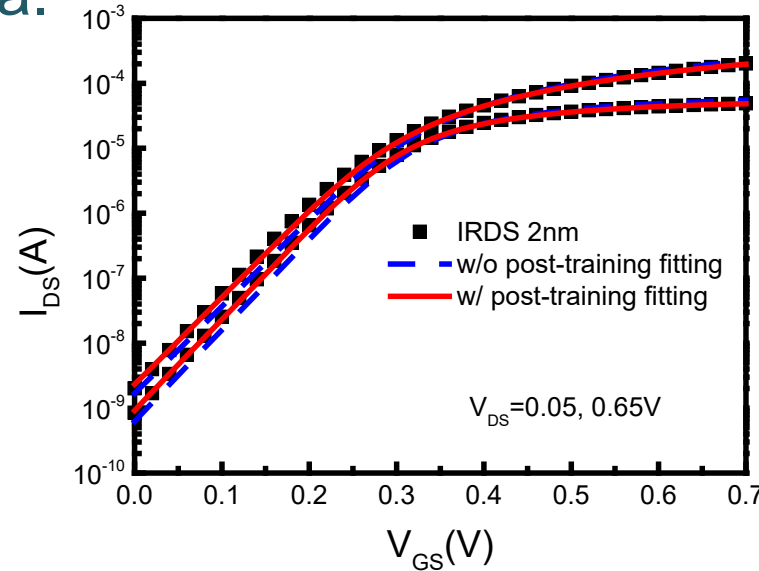
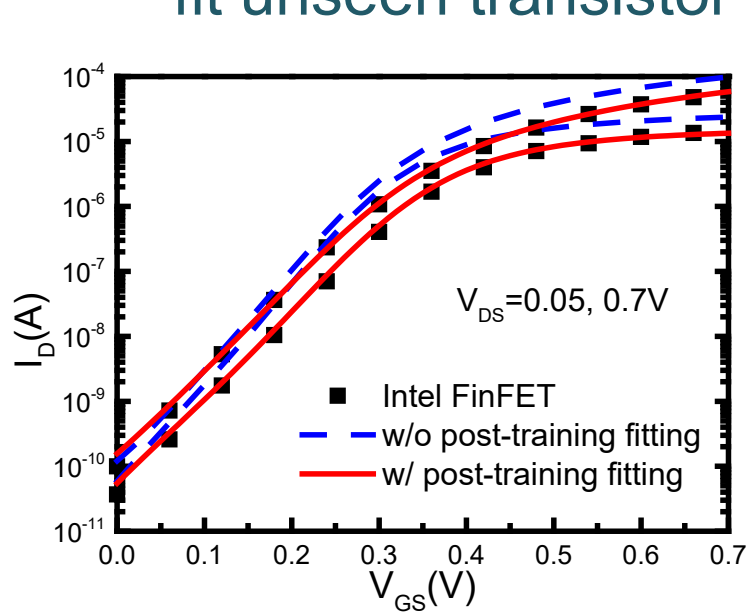
Variability Test

- A variability test on DUA, DSAT and Cp in the PNN is conducted.
- The result generated by the PNN can match BSIM-CMG variability.



Fitting Flexibility

- With parameters m , V_T , $DETA$, μ_{r0} , $DSAT$, DUA and $DLAMBDA$, the PNN can fit unseen transistor data.



The PNN model can fit transistor data across different geometries, including FinFET and GAAFET.

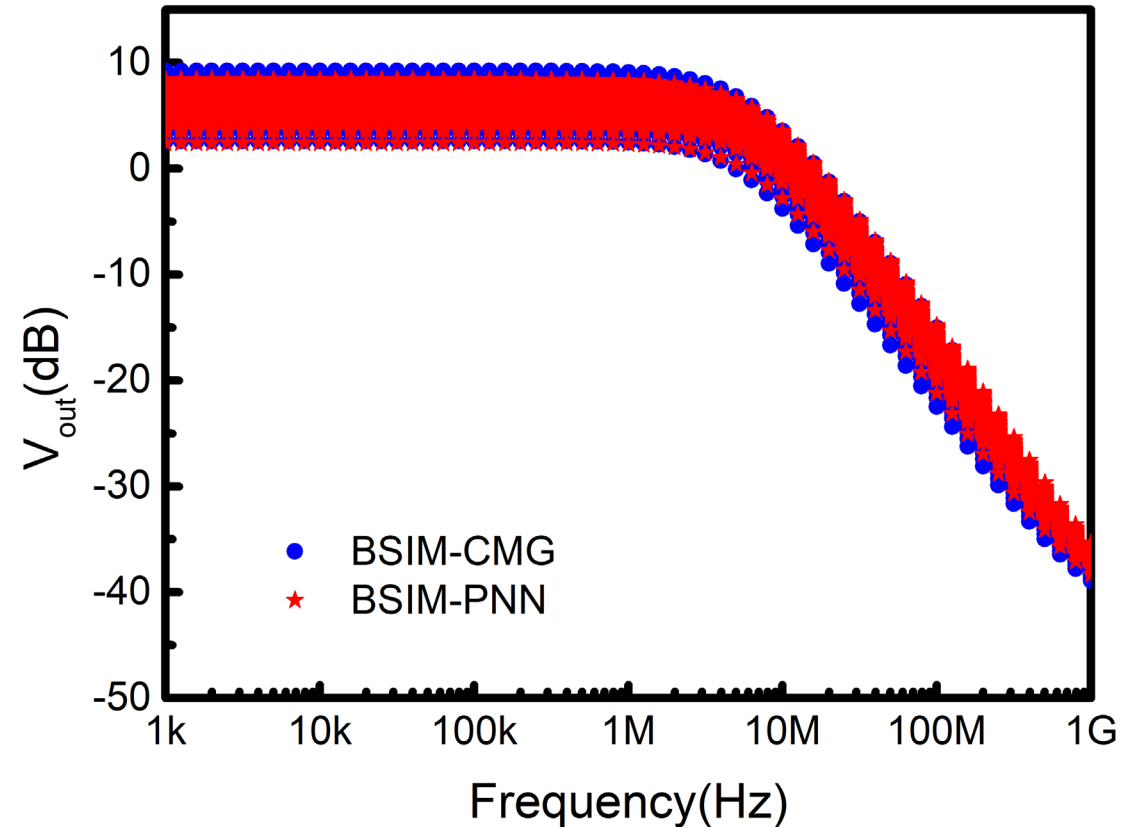
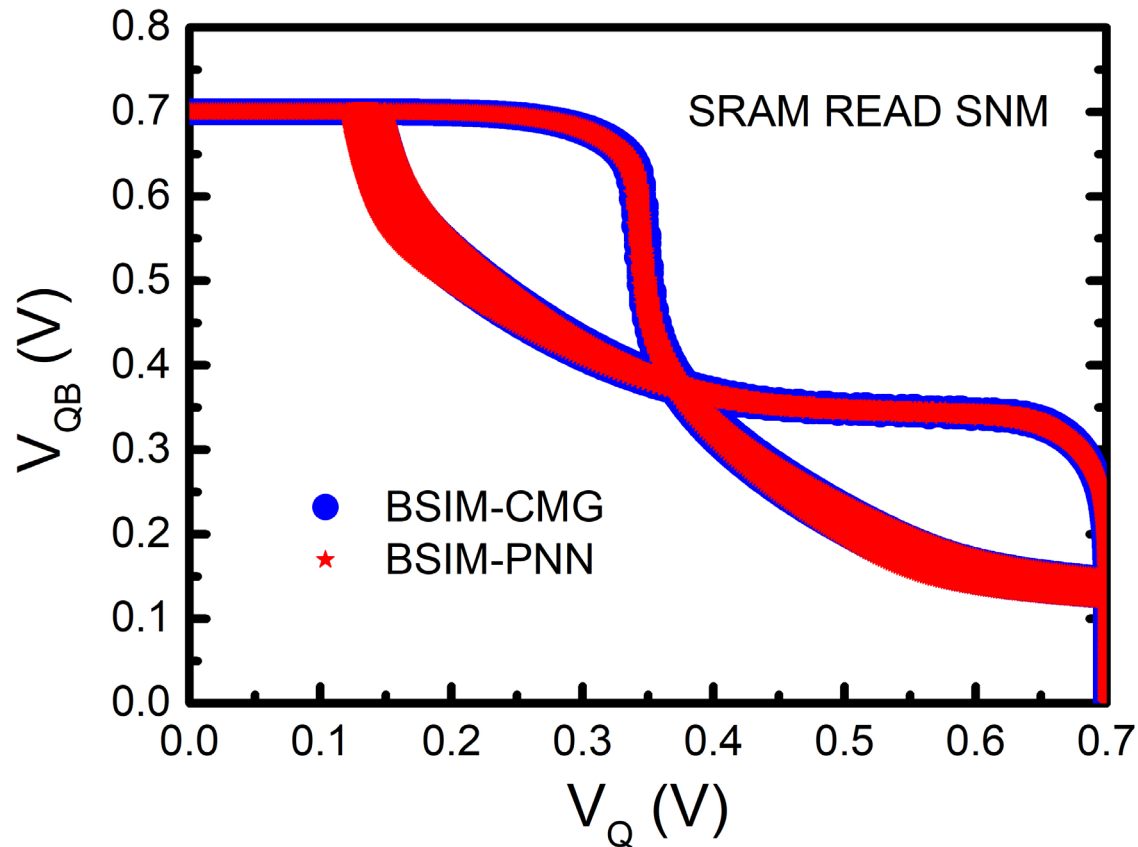


[1] IRDS 2022 Edition, 2022. [<https://irds.ieee.org/editions/2022>]

[2] C. Auth et al. 2017 IEEE IEDM, p. 29.1.1-4.

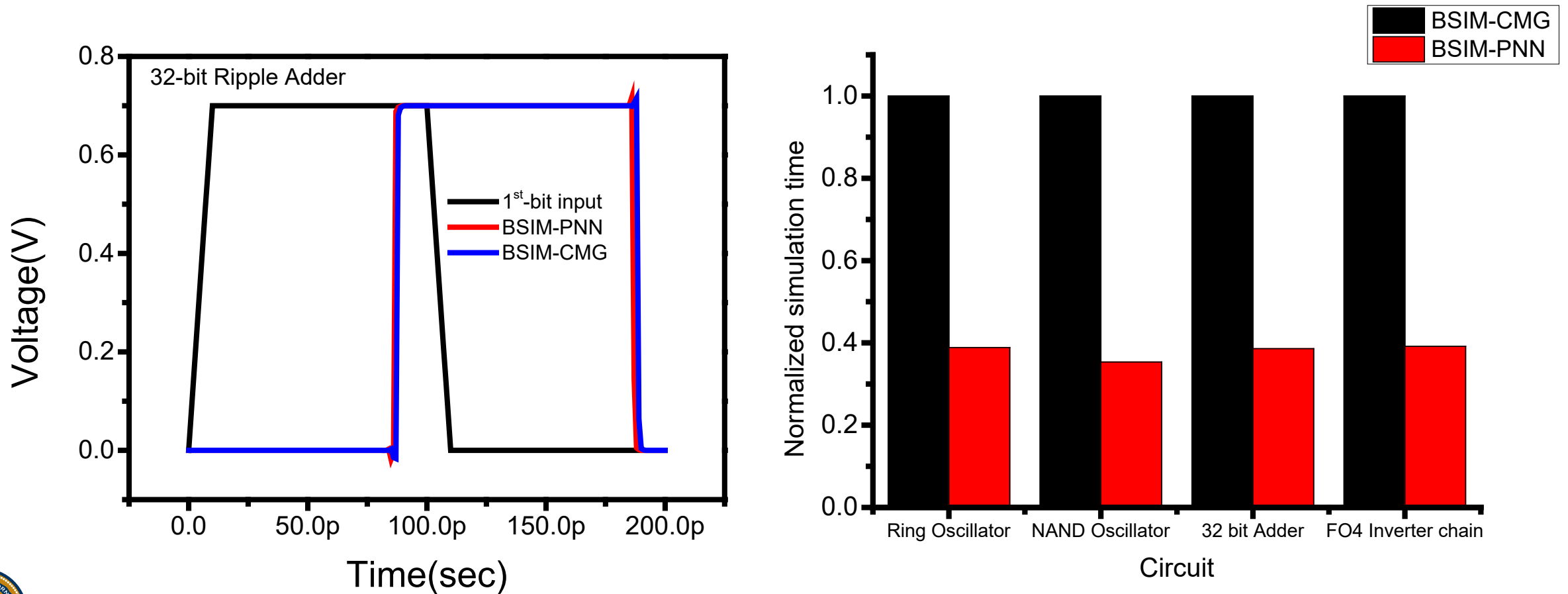
Circuit Simulation

- Monte Carlo simulations of SRAM and OTA are simulated by varying U_0 , U_A , U_D , V_{SAT1} in BSIM-CMG and the corresponding parameters in PNN.



Simulation Speed

- Some other circuits are also simulated. The simulation time is reduced by ~60% compared to BSIM-CMG with the PNN model in this work.



Conclusion

- A physical neural network (PNN) model incorporating simple device physics has been developed.
- The post-training parameters is introduced to enhance fitting flexibility and enable variability modeling.
- A more physical gate current model is proposed for zero bias condition and gate current partition.
- A binned PNN model is proposed for modeling transistor across a wide range of geometry, achieving improved accuracy.



The background features a series of white lines connecting several small white dots, creating a network-like or geometric pattern. The dots are positioned at various points, forming a series of interconnected triangles and polygons. The overall effect is a clean, modern, and technical aesthetic.

Thank You