Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Neural SPICE Surrogate Modeling and Sensitivity Analysis for Semiconductor MOSFETs: A Deep Learning Framework for Accelerated VLSI Simulation

Vrishab Sooram
Indian Institute of Technology Roorkee
Mayank Goswami
mayankfph@iitr.ac.in

Indian Institute of Technology Roorkee https://orcid.org/0000-0001-8829-9980

Method Article

Keywords: MOSFET compact modeling, neural surrogate modeling, SPICE simulation, sensitivity analysis, deep learning, VLSI CAD

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11728512/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Neural SPICE Surrogate Modeling and Sensitivity Analysis for Semiconductor MOSFETs: A Deep Learning Framework for Accelerated VLSI Simulation

Vrishab Sooram[1] and Mayank Goswami [1, 2, #]
[1] Department of Physics, IIT Roorkee
[2]*Mehta Family School of Data Science and Artificial Intelligence*, IIT Roorkee,
Roorkee, Uttarakhand-247667, India.
[#]mayank.goswami@ph.iitr.ac.in

Abstract—Modern Very Large Scale Integration (VLSI) electronic design automation (EDA) workflows rely heavily on circuit simulation engines (SPICE) to resolve large systems of non-linear differential-algebraic equations. As semiconductor dimensions scale, computing drain current characteristics iteratively via traditional numerical techniques introduces severe computational latency, particularly across extensive Monte Carlo process-variation sweeps. In this work, we propose a deep learning surrogate modeling framework to approximate device-level non-linear transport physics. The surrogate establishes a direct forward mapping F[Vgs, Vds, Vth, μ0, λ]--> Id across all operating regimes (cutoff, linear, and saturation) utilizing a high-capacity Multi-Layer Perceptron (MLP) architecture equipped with continuous Sigmoid Linear Unit (SiLU) activations. Trained on synthetically generated physical parameter profiles, the model demonstrates rapid convergence, driving means squared loss from an initial 0.122 down to 0.0078 with a coefficient of determination R^2 > 0.98 on unobserved test partitions. Furthermore, a normalized sensitivity analysis identifies gate bias Vgs and threshold voltage Vth as primary driving variables, fully consistent with metal-oxide-semiconductor field-effect transistor (MOSFET) electrostatics. The proposed surrogate achieves evaluation latencies under 2 milliseconds, providing a scalable path toward accelerated analog, digital, and post-silicon statistical simulations.

Keywords—MOSFET compact modeling, neural surrogate modeling, SPICE simulation, sensitivity analysis, deep learning, VLSI CAD.

I. Introduction

The aggressive downscaling of silicon complementary metal-oxide-semiconductor (CMOS) technology has enabled high functional integration densities and processing bandwidths in contemporary digital and mixed-signal integrated circuits (ICs) [1], [5]. However, entering nanoscale technology nodes has magnified the impact of intrinsic process variations, line-edge roughness, random dopant fluctuations, and short-channel effects [5], [6]. In modern Electronic Design Automation (EDA) flows, circuit verification, yield estimation, and statistical corner exploration necessitate millions of SPICE (Simulation Program with Integrated Circuit Emphasis) evaluations [4].

Standard SPICE engines compute branch currents and nodal voltages by resolving systems of coupled non-linear equations using numerical algorithms such as the Newton-Raphson iteration [2], [4]. Within these loops, the underlying device compact models (e.g., BSIM family) are repeatedly evaluated to update the circuit Jacobian matrix [4], [6]. When repeated over vast Monte Carlo sweeps or large-scale multi-transistor matrices, this iterative evaluation forms an acute computational bottleneck [4].

To alleviate this latency, compact modeling research has explored mathematical approximations, such as lookup tables (LUTs) and multi-dimensional polynomial surface fittings [2]. Nevertheless, LUT approaches face high memory footprints and poor multi-dimensional scalability, while high-order polynomials are susceptible to Runge's phenomenon and numerical divergence outside narrow operational envelopes. Conversely, solving the inverse problem—extracting constitutive semiconductor device parameters directly from measured current-voltage (I-V) characteristics—remains ill-posed due to non-uniqueness and parameter entanglement [3].

To address these challenges, this study presents a robust, forward-mapping deep learning surrogate framework. By projecting a combined operational and physical parameter vector—comprising gate-to-source bias (Vgs), drain-to-source bias (Vds), threshold voltage (Vth), low-field mobility (μ0), and channel-length modulation parameter (λ)—directly to the output drain current (Id), the forward neural operator bypasses iterative convergence checks while preserving physical fidelity [2], [3]. Using an optimized Multi-Layer Perceptron (MLP) with Sigmoid Linear Unit (SiLU) activation functions, this framework offers continuous, differentiable current profiles, enabling rapid sensitivity evaluations and real-time execution in accelerated analog and digital circuit simulations.

II. Mathematical Background & Physics Formulations

A. Governing MOSFET Analytical Formulations

The analytical transport physics governing a long-channel n-channel MOSFET can be formulated using the standard Shichman-Hodges formulation with continuous transitions [5]. The device behavior is partitioned across three core operating regimes dictated by the terminal biases:

1. Subthreshold / Cutoff Regime (Vgs <Є$Є$ Vth):
   In the deep subthreshold regime, the inversion layer charge density is negligible compared to the depletion charge. The drift current is negligible, leaving diffusion as the primary transport mechanism:

Id ≈ 0

2. Linear / Triode Regime (Vgs > Vth and Vds < Vgs - Vth):
   When the gate bias forms a continuous inversion channel between the source and the drain, drift current dominates under a longitudinal electric field:

Id = μ0 Cox (W/L) left[(Vgs - Vth)Vds - (1/2)Vds^2] (1 + λ Vds)

where Cox = frac{∈ox}{tox} represents the gate dielectric capacitance per unit area, W is the channel width, L is the physical channel length, μ0 denotes the low-field surface electron mobility, and λ represents the channel-length modulation factor accounting for pinch-off point movement [5].

3. Saturation Regime ($Vgs > Vth$ and Vds > Vgs - Vth): At the pinch-off threshold V_{ds,sat} = Vgs - Vth, the inversion layer charge at the drain edge approaches zero. The current saturates, governed by:

Id = (1/2) μ0 Cox (W/L) (Vgs - Vth)^2 (1 + λ Vds)

B. Process Transconductance and Formal Forward Operator

Defining the intrinsic process transconductance parameter as k' = μ0 Cox and the structural device aspect ratio as beta = k' (W/L), the system defines a mapping operator {F}: R^5 to R. The forward parameter state is represented as:

f{x} = [Vgs, Vds, Vth, μ0, λ]^T in Omega subset R^5

The non-linear forward mapping executed by the underlying semiconductor physics maps $f(x)$ to the scalar drain current:

$$Id = F(f(x)) = F(Vgs, Vds, Vth, \mu0, \lambda)$$

The surrogate modeling objective is to synthesize a parameterized deep neural operator $\hat{F}(f(x); \Theta)$ that approximates $F(f(x))$ over the entire compact operational domain Ω, such that:

$$\|F(f(x)) - \hat{F}(f(x); \Theta)\|2 \leq \varepsilon$$

for an arbitrarily small tolerance $\varepsilon > 0$ [1], [4].

III. Methodology & Neural Network Architecture

A. Synthetic Dataset Generation and Normalization

To train the neural surrogate across a physically realistic operational space, input vectors were sampled uniformly across defined bounds typical of sub-micron CMOS devices:

- Gate-to-Source Voltage (Vgs): 0.0 V ≤ Vgs ≤ 1.8 V
- Drain-to-Source Voltage (Vds): 0.0 V ≤ Vds ≤ 1.8 V
- Threshold Voltage (Vth): 0.3 V ≤ Vth ≤ 0.6 V
- Low-Field Electron Mobility (μ0): 0.03 m²/V·s ≤ μ0 ≤ 0.06 m²/V·s
- Channel Length Modulation (λ): 0.01 V⁻¹ ≤ λ ≤ 0.10 V⁻¹
- Geometric Channel Aspect Ratio: (W/L) = 2.0 with effective tox = 3.5 nm (Cox ≈ 9.86 × 10⁻³ F/m²).

A comprehensive dataset of N = 25,000 points was generated and partitioned using a 70% training (Ntrain = 17,500), 15% validation (Nval = 3,750), and 15% testing (Ntest = 3,750) split. Input features were standardized using z-score normalization:

$$\tilde{f}(x)i = (f(x)i - \mu x) / \sigma x$$

where μx and σx are the vector-wise empirical mean and standard deviation of the training split, preventing gradient vanishing or exploding during backpropagation [3].

B. Network Topology and Optimization

The parameterized neural operator $\hat{F}(f(x); \Theta)$ is constructed as a deep multi-layer feedforward network with a 5-128-256-128-1 topology:

$h(1) = \sigma(W(1)\, \tilde{f}(x) + b(1))$ $h(2) = \sigma(W(2)\, h(1) + b(2))$ $h(3) = \sigma(W(3)\, h(2) + b(3))$ $\hat{Id} = W(4)\, h(3) + b(4)$

where W(l) and b(l) represent the layer weights and biases comprising parameter set Θ. The activation function $\sigma(\cdot)$ is the Sigmoid Linear Unit (SiLU), defined as:

$SiLU(z) = z \cdot sigmoid(z) = z / (1 + e{-}z)$

The SiLU activation was selected over standard Rectified Linear Units (ReLU) because it is infinitely differentiable ($C\infty$-smooth), non-monotonic, and self-gated. These properties avoid discontinuous second derivatives, ensuring realistic device transconductance ($gm = \partial Id/\partial Vgs$) and output conductance ($gds = \partial Id/\partial Vds$) profiles critical for analog circuit analysis [1], [4].

The network was implemented in PyTorch and optimized using the Adam algorithm with an initial learning rate $\eta = 10^{-3}$, weight decay parameter $\gamma = 10^{-5}$, mini-batch size $B = 64$, and a mean squared error (MSE) objective function:

$L(\Theta) = (1/B)\, \Sigma i{=}1B\, (Id,i - \hat{Id},i(\tilde{f}(x)i; \Theta))^2$

Training was conducted over 250 epochs with a cosine annealing learning rate scheduler and an early stopping criterion bounded by validation loss stagnation over 20 consecutive epochs.

IV. Parameter Sensitivity Analysis & Method Justification

A. Normalized Sensitivity Metric

To quantify how perturbations in operational and physical parameters affect the output drain current, we compute a normalized dimensionless sensitivity index Sx:

$$Sx = \lim \Delta x \to 0\, ((\Delta Id / Id) / (\Delta x / x)) = (\partial Id / \partial x) \cdot (x / Id)$$

A centered finite-difference scheme with a ±10% parameter perturbation was applied around a nominal operational bias point ($f(x)0 = [Vgs = 1.2\ V, Vds = 1.2\ V, Vth = 0.4\ V, \mu 0 = 0.045\ m^2/V{\cdot}s, \lambda = 0.05\ V^{-1}]T$):

$$Sx \approx (Id(f(x)0 + \Delta x\, ek) - Id(f(x)0 - \Delta x\, ek)) / (2\, \Delta x) \cdot (x0,k / Id(f(x)0))$$

B. Sensitivity Rankings and Physical Alignment

The resulting normalized sensitivity metrics reflect the underlying semiconductor physics:

1. Gate-to-Source Overdrive ($\{S\}\{V\{gs\}\} \approx +2.14$): In the saturation regime, the current exhibits a quadratic dependence on gate overdrive (Vgs - Vth)^2, making Vgs the most sensitive driver of Id.
2. Threshold Voltage ($S\{V\{th\}\} \approx -1.08$): Variations in threshold voltage directly shift the gate overdrive, yielding an inverse, high-magnitude sensitivity profile.
3. Low-Field Mobility ($\{S\}_\{\mu 0\} \approx +1.00$): The drain current scales linearly with carrier drift mobility across linear and saturation regions, yielding a direct unit sensitivity index.

4. Drain-to-Source Bias ({S}{V{ds}} ≈ +0.06): Because the nominal point is in saturation, the direct channel-bias dependency is weak and mediated primarily through channel-length modulation.
5. Channel-Length Modulation ({S}_λ ≈ +0.05): Acts as a weak perturbation on saturation output resistance (r_o), producing a modest first-order sensitivity profile.

C. Comparison with Conventional Techniques

The neural surrogate method balances continuous physical differentiability with computational speed, avoiding common trade-offs found in traditional approaches:

Evaluation Metric / Feature	**SPICE (Iterative Newton-Raphson)**	**High-Order Polynomial Fitting**	**D W**
Computational Complexity	High: O(Niter · M³) matrix inversions	Low: {O}(K) polynomial tensor ops	Lo Gl
Convergence Guarantees	Subject to non-convergence failures	Deterministic; unstable at boundaries	D
High-Dimensional Scaling	Rigid; computationally expensive	Prone to Runge's phenomenon (D > 4)	Sc ba
Inference Latency	15.0 - 120.0 text{ ms} per point	0.05 - 0.20 text{ ms} per point	0.
Physical Continuity (C^∞)	Piecewise continuous across regimes	Continuous, but oscillates off-grid	In
Process Variation sweeps	Computationally prohibitive for corners	Poor extrapolation outside fit points	Ro

V. Experimental Results & Discussion

A. Training Trajectory and Convergence

The neural network converged steadily across training epochs. Starting from an initial mean squared error of Ltrain = 0.122 at epoch 1, the loss dropped below 1.5×10^{-2} by epoch 45. Cosine learning rate scheduling enabled stable refinement, reaching a final test MSE loss of:

$$L_{test} = 0.0078$$

No evidence of pathological overfitting was observed, with validation error tracking test loss within a 3.2% margin across all epochs.

B. Parity Plot and Goodness-of-Fit Analysis

Model predictions on the independent test split (Ntest = 3,750) were validated against ground-truth analytical calculations across all three operating regimes. The coefficient of determination was computed as:

$$R^2 = 1 - \left(\Sigma_{i=1}^{N_{test}} (I_{d,i} - \hat{I}_{d,i})^2 \right) / \left(\Sigma_{i=1}^{N_{test}} (I_{d,i} - \bar{I}_d)^2 \right) = 0.9842$$

The test dataset aligns closely along the ideal parity line ($y = x$) across four orders of magnitude (10^{-7} A to 1.2×10^{-2} A). The mean absolute percentage error (MAPE) across the active inversion region remained below 1.45%.

C. Computational Inference Latency

Inference latency benchmarks were executed on an Intel Core i7 workstation alongside an NVIDIA RTX 4070 GPU host. Batch-mode neural forward passes processed 10^4 device operational evaluations in 12.4 ms, corresponding to an average single-device inference time of $\tau inf \approx 1.24$ μs. Single-instance sequential CPU evaluations took 1.82 ms, representing a substantial speedup over iterative SPICE solver convergence times (15–120 ms depending on matrix conditioning and dynamic timestep selection) [2], [4].

VI. Coursework & Physics Domain Relevance

The principles and techniques developed in this work interface directly with key areas of academic physics and electrical engineering curricula:

- Solid-State Semiconductor Physics: Reinforces quantum and classical carrier transport principles, including electron drift-diffusion, surface mobility degradation under vertical electric fields, transverse field inversion, and band-bending phenomena [5].
- Analog Integrated Circuit Design: Directly affects small-signal transconductance calculations (gm), output dynamic resistance (ro = 1/gds), and intrinsic device gain (Av = −gm ro), which govern operational transconductance amplifier (OTA) design [5], [6].
- Digital CMOS VLSI Design: Informs fundamental switching dynamics, dynamic power dissipation (Pdyn = α CL Vdd^2 f), and propagation delay metrics characterized by the Alpha-Power Law through saturation current modeling [4].
- Computational Physics & Scientific Machine Learning: Provides practical exposure to high-dimensional optimization, non-linear regression, loss surface topology, and surrogate modeling for partial differential and non-linear algebraic systems [1], [3].

VII. Conclusion & Limitations

This paper presented a deep learning surrogate modeling framework designed to accelerate semiconductor device evaluations in circuit simulations. By mapping the operational and physical state [Vgs, Vds, Vth, μ0, λ]^T directly to drain current Id using a SiLU-activated MLP architecture, the surrogate achieved high fidelity (R^2 > 0.98, MSE convergence to 0.0078) and sub-2 millisecond inference speeds without non-linear solver convergence iterations. A normalized sensitivity analysis confirmed that gate overdrive variables dominate device current, in agreement with standard semiconductor physics.

Limitations & Future Scope

While the current framework demonstrates the efficacy of the neural surrogate approach, several areas remain for future extension:

1. Advanced Industrial Compact Models: The present formulation is trained on analytical Shichman-Hodges formulations. Incorporating comprehensive industrial BSIM-4 or BSIM-CMG model suites will be essential to capture high-order short-channel effects, such as drain-induced barrier lowering (DIBL), velocity saturation, and gate leakage currents [5], [6].

2. Multi-Gate Architectures: Extending the architecture to multi-gate non-planar topologies (e.g., FinFETs, Nanosheet GAAFETs) will require accounting for 3D electrostatic volume inversion and quantum confinement effects [5], [6].
3. Extreme Operational Environs: Incorporating temperature dependence (T) as an explicit input variable will support cryogenic CMOS simulations for quantum computing readouts and space-grade electronics.
4. Physics-Informed Neural Networks (PINNs): Integrating semiconductor drift-diffusion partial differential equations directly into the neural loss function can enforce thermodynamic consistency and conservation laws across extrapolation regimes [1], [3].

VIII. References

- [1] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, Feb. 2019, doi: 10.1016/j.jcp.2018.10.045.
- [2] W. Huang and Y. Wang, "Machine learning for electronic design automation: A survey of surrogate modeling in circuit simulation," *ACM Transactions on Design Automation of Electronic Systems*, vol. 28, no. 4, pp. 1–32, May 2023, doi: 10.1145/3580512.
- [3] Z. Chen, Y. Liu, and K. Saurabh, "AutoPINN: Automated physics-informed neural networks for parameter extraction and device modeling of advanced semiconductor transistors," *Micromachines*, vol. 14, no. 6, p. 1120, May 2023, doi: 10.3390/mi14061120.
- [4] L. Zhang, X. Li, and C. Cheng, "Artificial neural network based compact modeling methodology for advanced nanoscale MOSFETs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 10, pp. 3412–3425, Oct. 2022, doi: 10.1109/TCAD.2021.3134211.
- [5] S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 2006.
- [6] Y. S. Chauhan, D. D. Lu, V. Sriramkumar, S. Khandelwal, C. C. Hu, C. Xu, and J. P. Duarte, *FinFET Modeling for IC Simulation and Design: Using the BSIM-CMG Standard*, 1st ed. London, UK: Academic Press, 2015.