

A Deep Learning Framework for Fast, Unified Process and Device TCAD Simulation

Konstantina Kovani^{*†}, Greta Chiaravalli^{*†}, Francesco Di Stefano^{*†}, Michal Dyrek^{*}, Semin Kwak^{*}, Vamsi Spandan Arza^{*}, Ryota Hachiya[†], Yonggwi Cho[†], Hideki Minari[†], Jun Komachi[†], Shinya Yamakawa[†], Lorenzo Servadei^{*}

^{*}AI for Chip Design Team, Sony AI, Sony Europe, Zurich, Switzerland

[†]Sony Semiconductor Solutions Corp., Kanagawa, Japan

Emails: konstantina.kovani@sony.com, greta.chiaravalli@sony.com, francesco.distefano@sony.com, michal.dyrek@sony.com, semin.kwak@sony.com, vamsispandan.arza@sony.com, ryota.hachiya@sony.com, yonggwi.cho@sony.com, hideki.minari@sony.com, jun.komachi@sony.com, shinya.yamakawa@sony.com, lorenzo.servadei@sony.com

[†]Equal Contributors

Abstract—We introduce a unified multi-stage graph-based surrogate modeling framework for semiconductor Technology Computer-Aided Design (TCAD) that spans the full electrical design flow—from process simulation (implantation and diffusion) through device simulation and the extraction of electrical characteristics at the current density stage. The proposed architecture operates directly on unstructured meshes and predicts spatially resolved, node-level physical quantities with high fidelity. By learning the complete field solution rather than a small set of scalar metrics, the proposed architecture can be retrained across device variants and operating conditions, enabling designers to compute arbitrary metrics a posteriori from the predicted nodal fields.

Index Terms—TCAD, deep learning, device simulation, process simulation, surrogate model

I. INTRODUCTION

Technology Computer-Aided Design (TCAD) is a cornerstone of modern semiconductor development as it allows engineers to simulate and optimize fabrication processes and device performance before costly manufacturing. By combining process simulations (e.g., implantation, etching, diffusion) with device simulations (e.g., measuring electrical characteristics), TCAD enables the analysis of a device’s electrical behavior after fabrication steps, minimizing expensive trial-and-error and improving reliability. However, conventional TCAD workflows are computationally intensive, with simulation runtimes ranging from minutes to hours per configuration, posing significant bottlenecks in iterative design processes and large-scale design-space explorations.

To mitigate these computational challenges, neural surrogates have been increasingly investigated, particularly for device simulation tasks. Vanilla models based on Multi-Layer Perceptrons (MLPs) [1]–[3] and Convolutional Neural Networks (CNNs) [4], [5] have been applied to approximate device physics with reduced runtime complexity. However, CNN-based pipelines typically rely on mesh-to-grid conversions, which are computationally expensive and less flexible for the irregular geometries common in semiconductor devices. Graph-based surrogates, including Graph Neural Networks (GNNs) [6] and latest attention-enhanced variants [7], offer a more adaptable alternative by learning directly

Process Simulation – 625 simulations			
Symbol	Description	Range	Unit
AsDose	Arsenic Dose	$1 \cdot 10^{14} - 1 \cdot 10^{15}$	cm^{-2}
BDose	Boron Dose	$9 \cdot 10^{12} - 2 \cdot 10^{13}$	cm^{-2}
AsEnergy	Arsenic Energy	1–3	keV
BEnergy	Boron Energy	10–30	keV
Device Simulation – 1296 simulations			
Symbol	Description	Range	Unit
V_D	Drain Voltage	0.0–1.0	V
V_G	Gate Voltage	0.0–1.0	V

TABLE I: Input parameter ranges for process and device simulations.

on unstructured meshes, capturing local and global relationships without requiring interpolation or transformation. These models have demonstrated significant advances in predictive accuracy and robustness, especially on geometrically complex domains. Despite these advancements, current research has primarily focused on isolated tasks, leaving the end-to-end modeling of the full process–device prediction chain largely unaddressed.

In this work, we propose a unified graph-based TCAD surrogate that performs multi-stage prediction across the entire TCAD pipeline, from process conditions to device solutions and derived electrical characteristics. Unlike traditional surrogates that regress a fixed set of task-specific figures of merit, our model predicts high-fidelity 3D nodal fields directly on unstructured meshes. This enables the flexible computation of arbitrary performance metrics post hoc, without requiring additional ad-hoc models, while supporting seamless transfer across a broad range of device architectures and fabrication conditions. We further demonstrate that the proposed multi-stage surrogate integrates process and device stages into a unified framework, ensuring consistent field-level predictions while controlling error propagation. Benchmarking against state-of-the-art graph baselines highlights our model’s strong field-level accuracy and its robust performance in both staged and unified inference modes.

II. METHODS

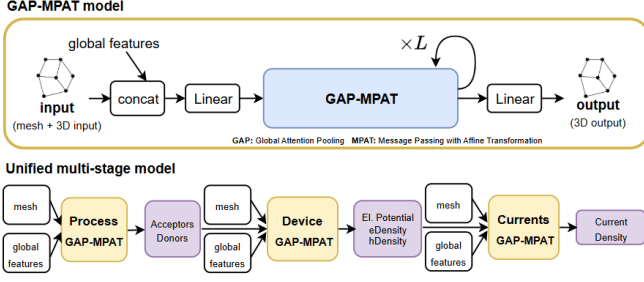


Fig. 1: Top: GAP-MPAT backbone schematic. Bottom: Process-to-device inference pipeline showing chained predictions across all three stages.

A. Multi-Stage Model Formulation

We formulate a unified surrogate model as a sequence of three physically motivated stages—process, device, and current density, each corresponding to a distinct step in the TCAD simulation flow. In unified mode, these stages are executed sequentially so that upstream predictions condition downstream inference, enabling end-to-end process-to-performance estimation.

Process stage models ion implantation and diffusion: given device geometry and process descriptors, it predicts spatial dopant concentration fields (nodal donor and acceptor concentrations). Binary masks identifying the source/drain contacts and substrate region are included as spatial context inputs. To preserve modularity and enable independent training for each stage, we further decompose the full device simulation into a device stage and a current density stage. The device stage takes as input the geometry, dopant fields (predicted or reference), bias conditions—drain voltage V_D and gate voltage V_G —and binary region masks, and produces electrostatic potential, electron density, and hole density. The current density stage then receives geometry, electrostatic potential, carrier densities, binary region masks, and the quasi-Fermi potentials reconstructed from carrier density and electrostatic potential through the Maxwell-Boltzmann relation [9], and predicts the current density vector. Fig. 4 summarizes this chained formulation and its stage outputs: the process block reports dopant fields (acceptor and donor concentrations), the device block reports electrostatic potential and carrier densities, and the current density block reports the vector current field from which direction and magnitude can be analyzed.

B. Model Architecture

The proposed surrogate model, Global Attention Pooling - Message Passing with Affine Transformation (GAP-MPAT), leverages a local-to-global ad-hoc framework schematized in Fig. 1. Node-level features are concatenated with global process descriptors and projected into a latent space by a linear input layer. A stack of GAP-MPAT blocks then refines

Process Surrogate Model						
Method	Acceptor Conc. [cm ⁻³]		Donor Conc. [cm ⁻³]			
	NMRE	logMAE	NMRE		logMAE	
GAP-MPAT	0.88	0.04	0.89		0.01	
Device Surrogate Model						
Method	El. Potential [V]		eDensity [cm ⁻³]		hDensity [cm ⁻³]	
	NMRE	logMAE	NMRE	logMAE	NMRE	logMAE
GAP-MPAT	0.30	0.02	0.79	0.08	0.04	0.71
multi-stage	0.38	0.03	1.04	0.09	0.04	1.08
Current Density Surrogate Model						
Method	J _x [A/cm ²]		J _y [A/cm ²]		J _z [A/cm ²]	
	NMRE	logMAE	NMRE	logMAE	NMRE	logMAE
GAP-MPAT	2.10	2.51	1.77	1.72	2.17	3.27
multi-stage	2.21	2.61	2.25	1.73	2.41	3.17

TABLE II: Error metrics for process (top), device (middle), and current density (bottom) surrogates for GAP-MPAT model with 64 hidden dimensions and 12 layers. Normalized Mean Relative Error is defined as $\text{NMRE}[\%] = \frac{100}{N} \sum_{i=1}^N \frac{|y_i - \hat{y}_i|}{\max_j(|y_j|)}$. Logarithmic Mean Absolute Error is defined as $\text{logMAE} = \frac{1}{N} \sum_{i=1}^N |\log(y_i + \epsilon) - \log(\hat{y}_i + \epsilon)|$.

the latent representation: MPAT performs one-hop degree-normalized message passing to capture local mesh interactions, together with a parallel affine self-update for preserving node-specific information. After each message-passing step, a global attention pooling layer aggregates sample-level mean and max statistics and injects this global context back to node embeddings through learned attention weights. Repeating this block over L -layers yields enriched node representations, which are finally mapped by a linear decoder to the target 3D nodal fields. The same backbone is used across process, device, and current density stages, enabling a unified process-to-performance surrogate without mesh-to-grid conversion.

III. EXPERIMENTS & RESULTS

We evaluate GAP-MPAT on a TCAD simulation dataset generated using Sentaurus sprocess and sdevice for a 45nm CMOS technology node, covering the joint process-device parameter space summarized in Table I. The process split comprises 625 simulations spanning Arsenic and Boron implant doses and energies; the device split comprises 1296 simulations spanning drain voltage V_D and gate voltage V_G , each in the range $[0, 1]V$. These parameters serve as global input features to the different model stages. For each task, 90% of samples are used for training and 10% for evaluation.

A. Single-Stage Model Performance

GAP-MPAT is first evaluated in a single-stage configuration across all three stages of the TCAD flow. As shown in Table II, normalized mean relative error (NMRE) stays below 1% for both donor and acceptor concentrations in the process stage, and remains at or below 1% for electrostatic potential and carrier densities in the device stage. In the current density stage, NMRE rises to at most 3%, reflecting the greater

complexity of predicting a vectorial quantity whose values span more than 20 orders of magnitude. Alongside NMRE, we report logMAE—the mean absolute error computed on log-scaled fields—to capture accuracy across the wide dynamic range of physical quantities.

GAP-MPAT is further compared against state-of-the-art baselines—Graph Convolutional Networks (GCN) [6], Graph Attention Network-v2 (GATv2) [7], and Multi-Hop Affine Transformation (MHAT) [8]—under matched model capacities, where d denotes the hidden dimension and L the number of message-passing layers (configurations: $64d/12L$ and $32d/12L$). Evaluations focus on the device stage, with insights extended to the process and current density stages. As shown in Table III, GAP-MPAT achieves consistently low NMRE and logMAE for electrostatic potential, electron density, and hole density across both model sizes, outperforming all baselines in most comparisons with the strongest gains on electrostatic potential and electron density, and remaining comparable in the remaining cases. Its stable accuracy across quantities with different physical scales confirms its suitability for coupled semiconductor-field reconstruction on non-uniform meshes.

B. Unified Multi-Stage Model Performance

For the unified multi-stage evaluation, the process, device, and current density modules are composed into a chained inference pipeline, and Table II quantifies error propagation across stages. The key result is that switching from single-stage to unified inference introduces only modest and controlled degradation—demonstrating that the framework successfully avoids the error amplification commonly seen in cascaded models. In the device stage, NMRE increases by at most 0.25 percentage points, and hole density is virtually unaffected. In the current density stage, per-component NMRE rises by less than 0.5 percentage points, remaining well below 3% throughout. These small absolute shifts are physically meaningful: they confirm that the predicted fields preserve the correct spatial structure—concentration gradients, potential distributions, and current flow patterns—required for accurate downstream analysis. Fig. 4 further supports these findings through spatial inspection across all stages. Predicted donor and acceptor fields preserve the main concentration gradients relative to TCAD ground truth; electrostatic potential and carrier-density patterns remain consistent in both high- and low-gradient regions; and the vectorial current output is reproduced with consistent flow orientation and magnitude distribution. As shown in Figs. 2 and 3 (right), larger residuals are mainly localized near geometric corners and small-volume regions, while most of the domain preserves close agreement with ground truth—explaining why global error metrics remain stable under chained inference.

C. Derived Metrics Estimation

Beyond field-level accuracy, we evaluate physically derived quantities computed from the predicted nodal fields against TCAD ground truth. Channel depth is extracted as the position at the channel midpoint where the electron density equals

		Device Surrogate Model					
		El. Pot. [V]		eDen. [cm^{-3}]		hDen. [cm^{-3}]	
Config	Method	NMRE	logMAE	NMRE	logMAE	NMRE	logMAE
64d/12L	GAP-MPAT	0.30	0.02	0.79	0.08	0.04	0.71
	GCN [6]	0.80	0.05	1.34	0.18	0.06	0.94
	GATv2 [7]	0.56	0.04	1.11	0.13	0.04	0.79
	MHAT [8]	0.68	0.05	1.19	0.15	0.06	0.95
32d/12L	GAP-MPAT	0.33	0.02	0.77	0.08	0.03	0.86
	GCN [6]	1.13	0.09	2.00	0.26	0.10	0.98
	GATv2 [7]	0.82	0.05	1.66	0.20	0.06	0.85
	MHAT [8]	1.06	0.08	1.87	0.22	0.10	1.13

TABLE III: Error comparison of GAP-MPAT, Graph Convolutional Networks (GCN) [6], Graph Attention Network-v2 (GATv2) [7], and Multi-Hop Affine Transformation (MHAT) [8] for two model sizes ($64d/12L$ and $32d/12L$).

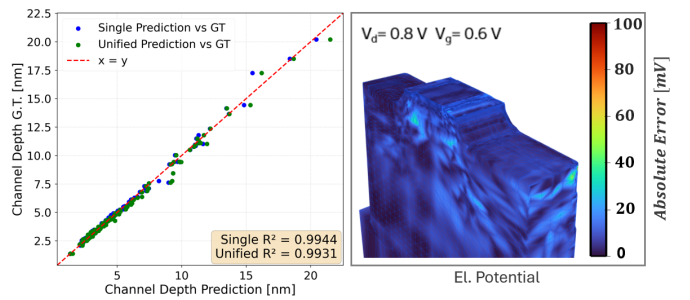


Fig. 2: Left: Comparison of single-stage and unified predictions against ground-truth channel depth across the test set. Right: 3D visualization of electrostatic potential Absolute Error [mV] at high-bias condition.

the acceptor concentration; samples without a valid crossing (i.e., no formed channel) are excluded. The parity plot in Fig. 2 (left) shows near-ideal agreement for both single-stage and unified inference, indicating that the surrogate accurately preserves physically meaningful spatial transitions. We further evaluate drain current by integrating the predicted current density over the drain contact cross-section. As shown in Fig. 3 (left), both inference settings closely follow the ideal

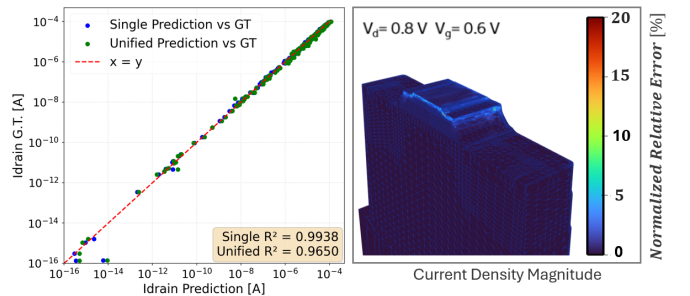


Fig. 3: Left: Comparison of single-stage and unified predictions against ground-truth drain current across the test set. Right: 3D normalized relative error of current density magnitude at a representative bias condition.

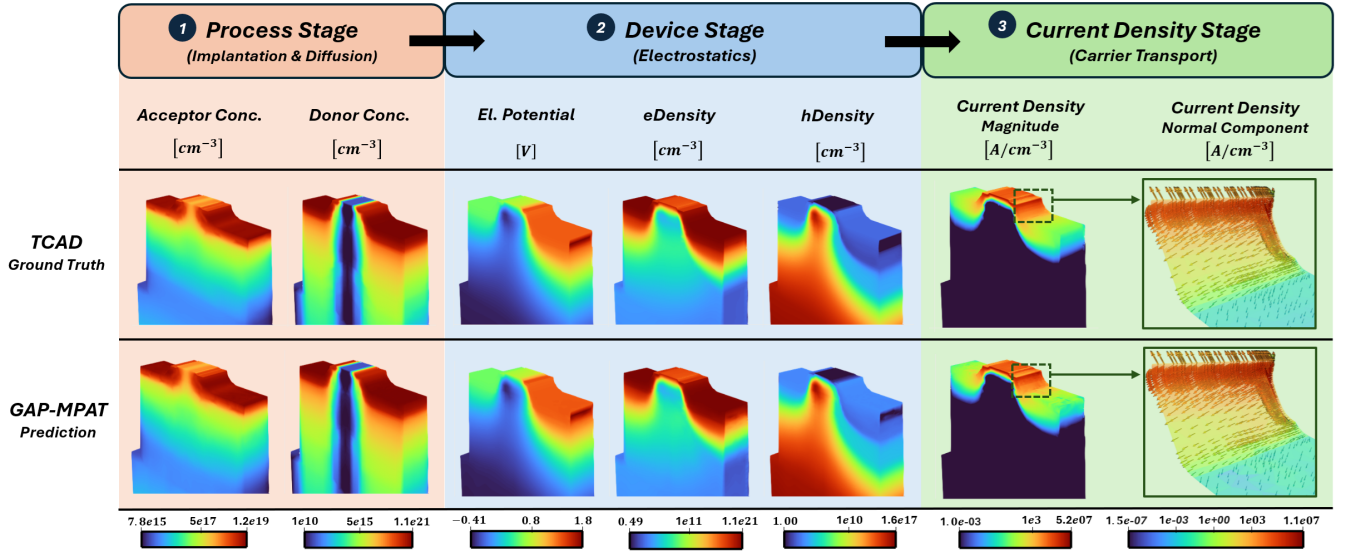


Fig. 4: End-to-end comparison of the unified multi-stage surrogate across process, device, and current density stages. Columns report stage outputs, while rows compare TCAD ground truth (top) and unified GAP-MPAT predictions (bottom).

$y = x$ trend across bias conditions, with deviations primarily in the low-current regime. Overall, both channel depth and drain current exhibit excellent agreement with TCAD results, achieving consistently high R^2 values above 0.96. While the unified setting shows a modest reduction in accuracy for current prediction, likely due to the greater sensitivity of current-derived quantities to localized field errors under chained inference, the strong parity trends confirm that physically meaningful electrical metrics can be reliably recovered from the predicted fields.

IV. CONCLUSION

We present GAP-MPAT, a unified graph-based neural surrogate for end-to-end TCAD process-to-device simulation. The model effectively controls error propagation across its three-stage pipeline (dopant reconstruction, device-field prediction, and current density estimation), ensuring physically consistent nodal fields throughout. This consistency prevents the common performance cliff seen in cascaded models and enables accurate numerical recomputation of derived metrics, eliminating the need for ad-hoc surrogates in workflows with diverse figures of merit. With an inference time of ~ 40 ms per sample, GAP-MPAT delivers orders-of-magnitude speedup over traditional TCAD solvers (minutes to hours per simulation). By directly operating on unstructured meshes, it accelerates TCAD workflows without requiring mesh-to-grid conversions, making it practical for irregular and complex device geometries. These features enable rapid design-space exploration, process/electrical co-optimization, and broader applicability in semiconductor design. Moreover, GAP-MPAT's local-to-global architecture supports the proposed framework in modeling diverse spatial scales and physical quantities in TCAD workflows, from dopant concentrations to current densities.

Its ability to handle this diversity while maintaining high accuracy highlights its utility for advancing TCAD pipelines. Overall, GAP-MPAT sets a new benchmark for combining speed, accuracy, and scalability in predictive modeling, paving the way for future innovations in semiconductor design.

REFERENCES

- [1] H. Dhillon, K. Mehta, M. Xiao, B. Wang, Y. Zhang, and H. Y. Wong, "TCAD-augmented machine learning with and without domain expertise," *IEEE Transactions on Electron Devices*, vol. 68, no. 11, pp. 5498–5503, Nov. 2021, doi: 10.1109/TED.2021.3073378.
- [2] K. Mehta and H. Wong, "Prediction of FinFET current-voltage and capacitance-voltage curves using machine learning with autoencoder," *IEEE Electron Device Letters*, vol. 42, no. 2, pp. 136–139, Feb. 2021, doi: 10.1109/LED.2020.3045064.
- [3] S. C. Han and S. M. Hong, "Deep neural network for generation of the initial electrostatic potential profile," in *Proc. Int. Conf. Simul. Semicond. Processes Devices (SISPAD)*, Sep. 2019, pp. 1–4, doi: 10.1109/SISPAD.2019.8870521.
- [4] S. Han, J. Choi, and S. Hong, "Acceleration of three-dimensional device simulation with the 3D convolutional neural network," in *Proc. Int. Conf. Simul. Semicond. Processes Devices (SISPAD)*, Sep. 2021, pp. 52–55, doi: 10.1109/SISPAD54002.2021.9592540.
- [5] W. J. Lee, W. T. Hsieh, B. H. Fang, K. H. Kao, and N.-Y. Chen, "Device simulations with a U-Net model predicting physical quantities in two-dimensional landscapes," *Scientific Reports*, vol. 13, no. 1, p. 731, 2023, doi: 10.1038/s41598-023-27599-z.
- [6] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, Toulon, France, Apr. 24–26, pp. 1–14, 2017. Available: arXiv:1609.02907. doi: 10.48550/arXiv.1609.02907.
- [7] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022, doi: 10.48550/arXiv.2105.14491.
- [8] W. Jang, S. Myung, J. M. Choe, Y.-G. Kim, and D. S. Kim, "TCAD Device Simulation With Graph Neural Network," *IEEE Electron Device Letters*, vol. 44, no. 8, pp. 1368–1371, Aug. 2023, doi: 10.1109/LED.2023.3290930.
- [9] R. S. Muller, T. I. Kamins, and M. Chan, *Device Electronics for Integrated Circuits*, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 2003.