

Physics-Informed Neural Networks for Stable Drift-Diffusion Simulation and Inverse Parameter Extraction in 2D Semiconductor Devices

Satofumi Souma

Department of Electrical and Electronic Engineering
Kobe University Kobe 657-8501, Japan
Email: ssouma@harbor.kobe-u.ac.jp

Akiko Ueda

Semiconductor Frontier Research Center,
National Institute of Advanced Industrial Science
and Technology (AIST), Tsukuba, 305-8569, Japan
Email: ueda-akiko@aist.go.jp

Abstract—While Physics-Informed Neural Networks (PINNs) offer a promising paradigm for solving partial differential equations directly, their application to semiconductor drift-diffusion (DD) models is fundamentally hindered by severe gradient pathologies arising from exponential carrier statistics. To overcome these difficulties, we propose a highly robust PINN framework tailored for ultra-thin 2D semiconductors. By introducing a dual-network architecture that synergistically co-predicts quasi-Fermi potentials and current densities, alongside the native embedding of analytical 2D carrier statistics, we effectively circumvent the high-order derivative instabilities inherent in standard PINNs. Using a representative 2D MoS₂ p-n junction, we demonstrate that our framework reliably predicts forward I-V characteristics without external training data and successfully extracts the microscopic carrier lifetime directly from macroscopic electrical measurements.

Index Terms—Physics-Informed Neural Networks (PINNs), semiconductor device modeling, drift-diffusion model, inverse problems, parameter extraction

I. INTRODUCTION

As electronic devices aggressively scale down, efficient design via modeling and simulation becomes critical [1]. Recently, “Digital Twins” have gained attention, demanding fast surrogate models to replace computationally expensive solvers [2]. While conventional data-driven neural networks (NNs) successfully accelerate quantum transport calculations [3], [4], their reliance on massive high-fidelity training data remains a bottleneck. Physics-Informed Neural Networks (PINNs), which embed physical laws directly into the learning process, offer a promising alternative [5], [6].

The drift-diffusion (DD) framework is the standard for carrier transport modeling, even for emerging ultra-thin two-dimensional (2D) materials like transition metal dichalcogenides (TMDCs) [7], [8]. Although initial PINN-based DD models have been reported [10], [11], applying standard PINNs to DD models remains fundamentally challenging. Specifically, the exponential dependence of carrier densities on potential causes severe gradient pathologies, making charge conservation computationally unstable under forward biases where recombination is active.

To address this, we present a PINN framework tailored for 2D semiconductors that stabilizes the DD equations by em-

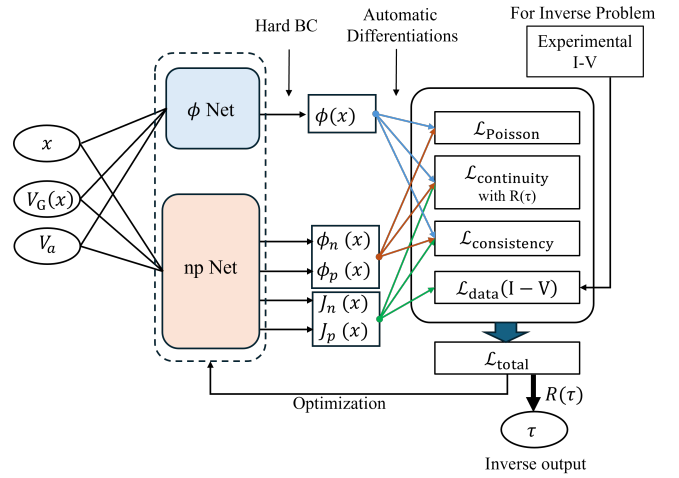


Fig. 1. Schematic illustration of the proposed dual-NN PINN framework. The architecture employs the ϕ Net for the electrostatic potential and the np Net for the synergistic co-prediction of quasi-Fermi levels and current densities, effectively circumventing high-order derivative computations. Hard boundary conditions (Hard BC) are strictly enforced via the network ansatz, embedding the exact Dirichlet boundary conditions dictated by the applied contact biases. The continuity loss incorporates the Shockley-Read-Hall recombination rate $R(\tau)$ to evaluate charge conservation. The total loss ($\mathcal{L}_{\text{total}}$) minimizes physical PDE residuals alongside macroscopic data loss ($\mathcal{L}_{\text{data}}$), enabling simultaneous forward simulation and the inverse extraction of physical parameters such as the carrier lifetime τ .

bedding physical consistencies via automatic differentiation. This approach enables robust forward simulation and effective inverse parameter extraction.

II. PHYSICAL MODEL AND NETWORK ARCHITECTURE

A. Mathematical Model and Scaling

To model the carrier transport in ultra-thin two-dimensional (2D) semiconductor devices (e.g., transition metal dichalcogenides (TMDCs)), we implement a modified 1D drift-diffusion framework. By integrating the vertical electrostatic fields across the channel thickness t_{ch} , we obtain an effective

2D Poisson equation, which is solved simultaneously with the current continuity equations:

$$\varepsilon_0 \varepsilon_r \frac{d^2 \phi}{dx^2} = -q(p - n) + \frac{C_{\text{ox}}}{t_{\text{ch}}} (\phi - V_g(x)), \quad (1)$$

$$\frac{dJ_n}{dx} = qR(n, p), \quad \frac{dJ_p}{dx} = -qR(n, p), \quad (2)$$

where ϕ is the electrostatic potential, V_g is the gate voltage, C_{ox} is the oxide capacitance, and $R(n, p)$ is the Shockley-Read-Hall (SRH) recombination rate.

To ensure numerical stability during optimization, we scale all variables to a dimensionless domain using the thermal voltage $V_t = k_B T / q$, device length L_{device} , and reference density n_{scale} (e.g., $x' = x / L_{\text{device}}$, $\phi' = \phi / V_t$). The dimensionless governing equations become:

$$\frac{1}{\kappa^2} \frac{d^2 \phi'}{dx'^2} + (p' - n' + \alpha(V'_g(x') - \phi')) = 0, \quad (3)$$

$$\frac{dJ'_n}{dx'} = R', \quad \frac{dJ'_p}{dx'} = -R', \quad (4)$$

where the control parameters are $\kappa^2 = q n_{\text{scale}} L_{\text{device}}^2 / (\varepsilon_0 \varepsilon_r V_t)$ and $\alpha = C_{\text{ox}} V_t / (q t_{\text{ch}} n_{\text{scale}})$. Furthermore, the dimensionless SRH rate is evaluated as $R' = (n' p' - n_i'^2) / [\tau'(n' + p' + 2n_i')]$, utilizing the scaled carrier lifetime τ' and intrinsic density n_i' .

B. Analytical 2D Statistics and Mathematical Equivalence

Because the density of states in a 2D electron gas is energy-independent above the band edge, the 0-th order Fermi-Dirac integral can be evaluated analytically without numerical approximations. In the scaled domain, the exact electron density is expressed as $n' = N'_c \ln[1 + \exp(\phi' - \phi'_n)]$, where N'_c is the effective 2D density of states and ϕ'_n is the quasi-Fermi potential.

Crucially, this mathematical structure naturally corresponds to the `Softplus` activation function, $\text{softplus}(z) = \ln(1 + \exp(z))$, widely implemented in machine learning frameworks. We directly embed this analytical formulation into the network ansatz as:

$$n' = N'_c \cdot \text{softplus}(\phi' - \phi'_n), \quad p' = N'_c \cdot \text{softplus}(\phi'_p - \phi'). \quad (5)$$

By implementing this analytical alignment, we effectively circumvent the classical Boltzmann exponential approximation. This eliminates the numerical overflow pathologies in heavily accumulated regimes and enables robust automatic differentiation. Consequently, this embedding stabilizes the Poisson solver, allowing our PINN model to stably predict continuous finite-bias I-V characteristics.

C. Dual-Network Architecture and Consistency Formulations

Evaluating the drift-diffusion currents ($J'_n = -n' d\phi'_n / dx'$) directly within the continuity equation (4) requires second-order automatic differentiation ($\frac{d^2 \phi'_n}{dx'^2}$) operating on highly nonlinear exponential fields. In standard PINNs, this inevitably

causes optimization stall or gradient explosion under forward bias. To overcome this, we propose a dual-network co-prediction framework (Fig. 1) comprising a ϕ -Net and an np -Net.

To strictly satisfy the contact boundary conditions, we embed a generalized ansatz into both networks. The ϕ -Net outputs a correction term $\mathcal{N}_\phi$, while the np -Net outputs a four-dimensional vector $\mathcal{N}_{np} = [\phi_n, \phi_p, J_n, J_p]^T$. The physical potentials are then constructed as:

$$\phi'(x') = \phi'_{\text{eq}}(0)(1 - x') + (\phi'_{\text{eq}}(1) + V'_a)x' + x'(1 - x')\mathcal{N}_\phi, \quad (6)$$

$$\phi'_{n,p}(x') = V'_a x' + x'(1 - x')\mathcal{N}_{np}^{(1,2)}, \quad (7)$$

where ϕ'_{eq} denotes the equilibrium flat-band potentials, and V'_a is the scaled applied bias. Crucially, the current densities are treated as direct decoupled outputs ($J'_{n,p} = \mathcal{N}_{np}^{(3,4)}$). This effectively reduces the required order of automatic differentiation for the continuity equations from second-order to first-order ($\frac{dJ'_n}{dx'}$).

To restore the mathematical link between these decoupled variables, we minimize a consistency loss ($\mathcal{L}_{\text{cons}}$) incorporating a localized damping factor ($\delta = 10^{-2}$) to prevent singularity in the depletion regions:

$$\mathcal{L}_{\text{cons}} = \frac{1}{N} \sum_{i=1}^N \sum_{\nu \in \{n,p\}} \mathcal{H} \left(\frac{J'_\nu - (-\nu' \frac{d\phi'_\nu}{dx'})}{\nu' + \delta} \right), \quad (8)$$

where $\mathcal{H}(\cdot)$ denotes the Huber loss function, and ν represents the carrier type.

Finally, the complete PDE physics is enforced via the composite loss $\mathcal{L}_{\text{total}} = \lambda_{\text{PDE}} \mathcal{L}_{\text{Poisson}} + \lambda_{\text{cons}} \mathcal{L}_{\text{cons}} + \lambda_{\text{cont}} \mathcal{L}_{\text{cont}}$, which is optimized via a hybrid Adam/L-BFGS loop. Here, $\mathcal{L}_{\text{cont}}$ incorporates both (4) and the total current conservation ($\frac{d}{dx'}(J'_n + J'_p) = 0$). In addition, by treating the SRH lifetime τ as a trainable parameter and adding a macroscopic data loss ($\|J_{\text{total}} - J_{\text{target}}\|^2$) to $\mathcal{L}_{\text{total}}$, the optimizer autonomously extracts the true τ directly from the I-V curves.

III. SIMULATION RESULTS AND DISCUSSION

To evaluate the predictive capability of our proposed framework, we investigate the spatial profiles of a representative MoS₂ p-n junction. For the physical implementation, we configure the device with a channel length of $L = 1 \mu\text{m}$, an effective thickness of $t_{\text{ch}} = 10 \text{ nm}$, and typical room-temperature carrier mobilities ($\mu_n = 200$ and $\mu_p = 50 \text{ cm}^2/\text{Vs}$).

Figure 2 illustrates the spatial distributions of the physical variables under various applied biases, ranging from thermal equilibrium ($V_a = 0.0 \text{ V}$) to high-injection forward bias conditions (up to $V_a = 0.3 \text{ V}$). Across all bias regimes, the PINN predictions for the electrostatic potential, quasi-Fermi levels, and local carrier densities exhibit a robust agreement with the reference solutions obtained via the conventional boundary value problem (BVP) solver. Notably, the left panels demonstrate that the quasi-Fermi level splitting accurately matches the applied boundary bias, smoothly transitioning across the depletion region. Furthermore, the center panels of

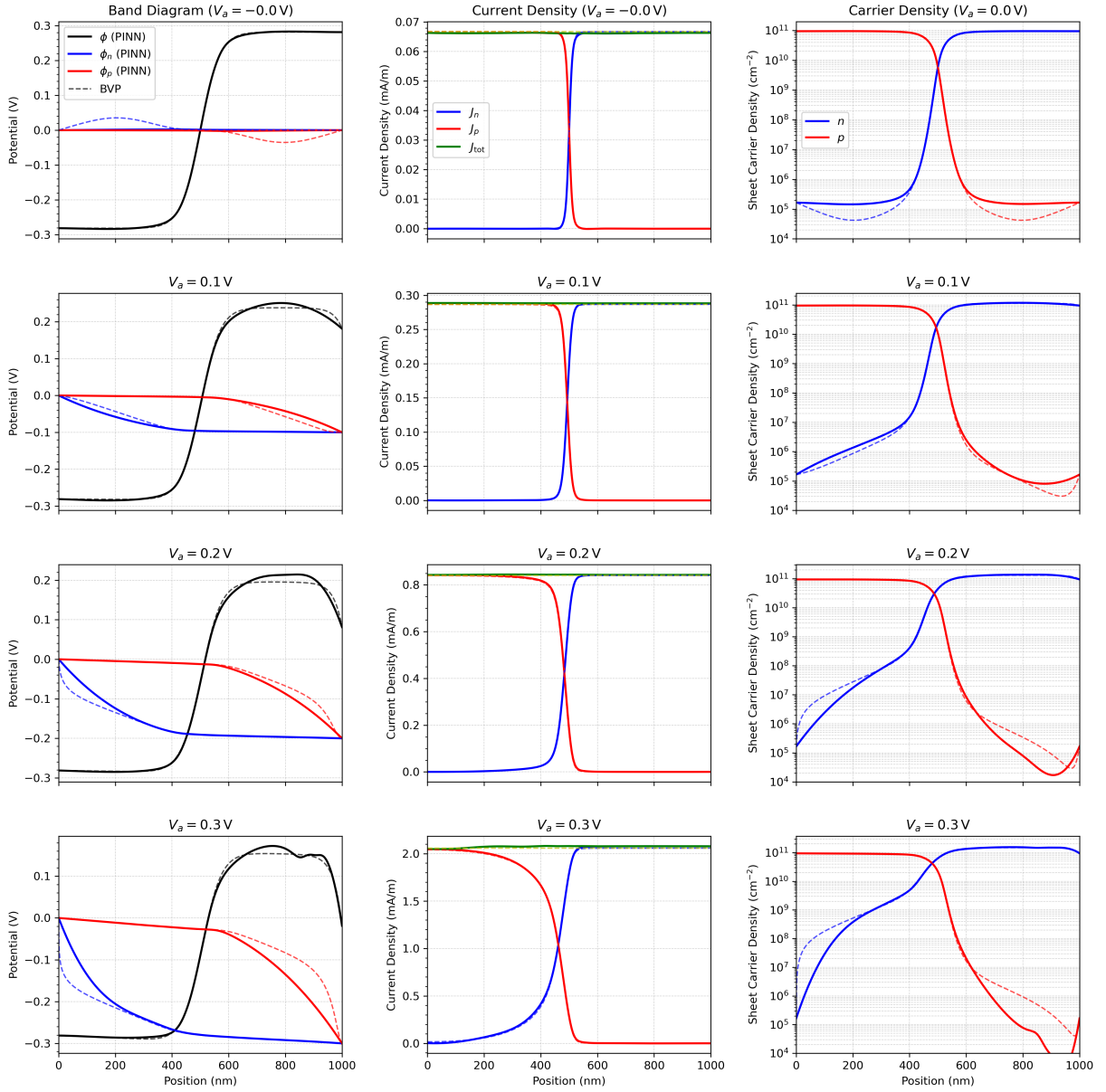


Fig. 2. Spatial profiles of the PINN solution for a MoS₂ p-n junction under various forward bias conditions (from $V_a = 0.0$ V to 0.3 V). (Left) Band diagrams showing the electrostatic potential ϕ and quasi-Fermi potentials ϕ_n , ϕ_p . (Center) Electron and hole current density profiles J_n , J_p , and the strictly conserved total current J_{tot} . (Right) Carrier densities n and p on a logarithmic scale. Dashed lines indicate reference solutions obtained from a conventional boundary value problem (BVP) solver.

Fig. 2 confirm that the total current density ($J_{\text{tot}} = J_n + J_p$) is strictly conserved across the entire channel, even as the injection level increases by orders of magnitude at $V_a = 0.3$ V. This underscores that the dual-network architecture effectively captures the highly nonlinear transport physics without suffering from the numerical divergence typically observed in standard PINNs.

Furthermore, we evaluate the macroscopic electrical properties and the inverse parameter extraction capability of the model. Figure 3 demonstrates the forward I–V characteristics simulated across various carrier lifetimes, where the PINN

reliably reproduces the exponential current behavior inherent in the DD equations. For the inverse problem, we implement the macroscopic I–V data (from $V_a = 0.0$ V to 0.3 V) as a training target to dynamically extract the microscopic SRH lifetime τ . As illustrated in Fig. 1, these training data are evaluated through the data loss $\mathcal{L}_{\text{data}}$, which links the macroscopic measurements to the microscopic physics and enables the determination of the SRH lifetime parameter τ . As depicted in the convergence history (Fig. 4) and the resulting error distribution (Fig. 5), the estimated τ robustly converges to the true value within a 10% error margin, regardless of

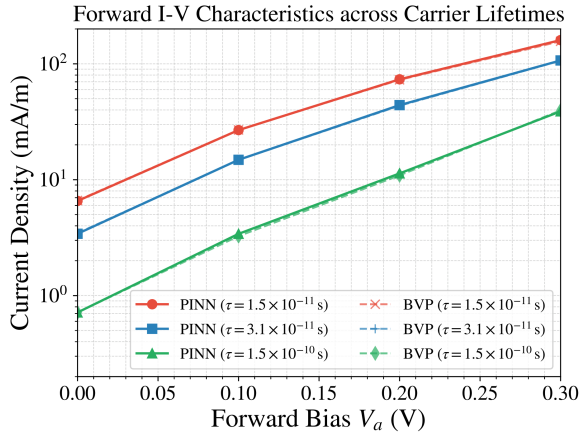


Fig. 3. Forward macroscopic I-V characteristics simulated across various carrier lifetimes, comparing the PINN predictions with the BVP reference solutions.

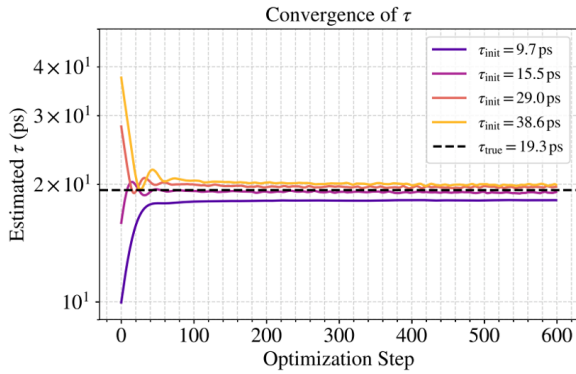


Fig. 4. Convergence history of the estimated carrier lifetime τ for different initial values τ_{init} .

the initial guess τ_{init} . This result confirms that our framework successfully extracts the internal physical parameters directly from the macroscopic measurements, providing a stable and highly reliable tool for 2D device characterization.

IV. CONCLUSION

By strategically decoupling the electrostatic potential and transport variables and embedding the analytical 2D carrier statistics, our framework provides a stable, fully differentiable solver for 2D semiconductor device modeling. This synergistic co-prediction strategy effectively circumvents the high-order derivative instabilities that plague standard PINNs under forward bias conditions. Consequently, it enables not only robust forward I-V simulations but also high-fidelity inverse parameter extractions, successfully estimating the microscopic SRH lifetime directly from macroscopic electrical measurements.

Although the initial offline training incurs a computational overhead, the subsequent online inference evaluates the electrical responses virtually instantaneously, bypassing the iterative matrix inversions of conventional BVP solvers. Future work will focus on extending this framework to transient simulations

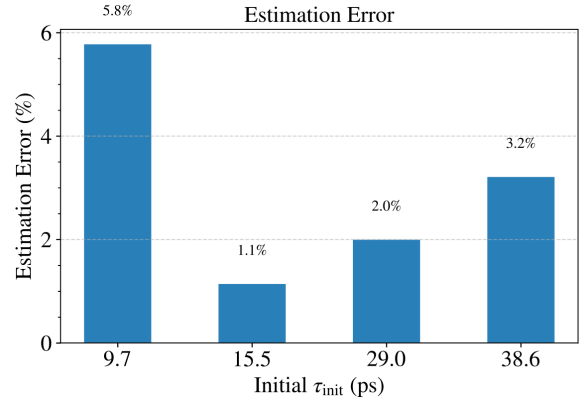


Fig. 5. Estimation error of τ as a function of τ_{init} , with the 10% error threshold indicated by the dashed red line.

and integrating non-isothermal models to capture self-heating effects. Furthermore, deploying this high-speed differentiable solver onto edge computing platforms, such as FPGAs, holds great potential for realizing real-time feedback control in semiconductor manufacturing digital twins.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI Grant Number JP22K04244 and the 2025 Research Grant from the Hagiwara Foundation.

REFERENCES

- [1] International Roadmap for Devices and Systems (2019) <https://irds.ieee.org/>.
- [2] P. Raut, D. K. Panda, and A. K. Goyal, "A Comprehensive Review on Next-Generation Modeling and Optimization for Semiconductor Devices," *IEEE Access*, vol. 13, pp. 123724–123742, 2025.
- [3] S. Souma and M. Ogawa, "Acceleration of nonequilibrium Green's function simulation for nanoscale FETs by applying convolutional neural network model," *IEICE Electron. Express*, vol. 17, no. 4, pp. 1–6, 2020.
- [4] S. Souma, "Neural network model for implementation of electron-phonon scattering in nanoscale device simulations based on NEGF method," in *2021 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, 2021, doi: 10.1109/SISPAD54002.2021.9592557.
- [5] 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," *J. Comput. Phys.*, vol. 378, pp. 686–707, 2019.
- [6] K. M. Mamun, N. Pala, and M. S. A. Shawkata, "Comprehensive Review of Machine Learning Approaches for Semiconductor Device Modeling and Simulation," *IEEE Access*, vol. 13, pp. 162969–162991, 2025.
- [7] A. Ueda, H. Imamura, and H. Fukidome, "High-frequency characteristics of ultra-short gate MoS₂ transistors," *Appl. Phys. Express*, vol. 18, no. 3, 034005, 2025.
- [8] A. Nipane, J. T. Teherani, and A. Ueda, "Demystifying the role of channel region in two-dimensional transistors," *Appl. Phys. Express*, vol. 14, no. 4, 044003, 2021.
- [9] R. Li, E. Lee, and T. Luo, "Physics-Informed Deep Learning for Solving Coupled Electron and Phonon Boltzmann Transport Equations," *Phys. Rev. Applied*, vol. 19, 064049, 2023.
- [10] B. Kim and M. Shin, "A Novel Neural-Network Device Modeling Based on Physics-Informed Machine Learning," *IEEE Trans. Electron Devices*, vol. 70, no. 11, pp. 6021–6025, 2023.
- [11] R. Riganti, M. G. C. Alasio, E. Bellotti, and L. Dal Negro, "DDNet: A Unified Physics-Informed Deep Learning Framework for Semiconductor Device Modeling," *arXiv preprint arXiv:2509.08073*, 2025.