A Physics-Informed Cascade Surrogate for Likelihood-Aligned Rare-Event Monte Carlo of SPAD Avalanche Breakdown

Mathieu Vandwalle
X-FAB Global Services GmbH, 99097 Erfurt, Germany
mathieu.vandwalle@xfab.com
ORCID: 0000-0003-2925-1850

***Abstract*—Monte Carlo carrier-transport simulations used to estimate the spatial avalanche breakdown probability in single-photon avalanche diodes (SPADs) are essential for accurate photon detection probability (PDP) modeling, yet remain computationally prohibitive due to stochastic transport and rare impact-ionization events. A physics-informed cascade neural surrogate is introduced that emulates the Monte Carlo avalanche simulation workflow by decomposing it into three physically ordered stages: spatial electric-field reconstruction, excess bias inference, and breakdown-probability prediction. The final stage is trained directly on Monte Carlo injection counts using a binomial negative log-likelihood, yielding calibrated spatial probability maps rather than mean-based regression outputs. Across variations in process, geometry, temperature, and applied bias, converged Monte Carlo breakdown-probability maps are accurately reproduced while runtime is reduced by up to 10^6-fold. More broadly, the cascade and likelihood-aligned training provide a reusable template for accelerating rare-event Monte Carlo in semiconductor transport and reliability, including high-voltage semiconductor device avalanche breakdown and radiation-induced degradation.**

***Index Terms*—Single-Photon Avalanche Diodes (SPADs), avalanche breakdown probability, rare-event Monte Carlo, physics-informed neural networks, probabilistic neural surrogates.**

I. Introduction

Silicon Single-Photon Avalanche Diodes (SPADs) are key enabling photodetectors for low-light sensing applications, including time-of-flight imaging, fluorescence lifetime microscopy, and quantum communication [1]–[3]. These devices are operated above their breakdown voltage (V_{BD}), where high electric fields accelerate carriers to energies sufficient for impact ionization, allowing the absorption of a single photon to trigger a self-sustaining avalanche multiplication process. The probability that a photo-generated carrier initiates an avalanche, referred to as the avalanche breakdown probability (p_t), directly determines the photon detection probability.

Accurate modeling of avalanche breakdown probability relies on Monte Carlo carrier-transport simulations that explicitly capture stochastic drift–diffusion dynamics, impact ionization, and recombination processes [4], [5]. While physically accurate, these simulations are intrinsically rare-event: statistically converged estimates of p_t require large numbers of carrier injection trials across the device grid, leading to prohibitive computational cost, especially when spatially resolved probability maps are required across wide ranges of bias, temperature, geometry, and process conditions. Here, rare-event refers to low-probability avalanche triggering, where only a few breakdowns are observed out of the N_{MC} carrier-injection trials at a given spatial location and operating point. This limitation significantly restricts design-space exploration and optimization in practical SPAD development. As illustrated in **Fig. 1**, increasing Monte Carlo simulation time gradually reveals additional avalanche-triggering regions, but at the expense of rapidly increasing wall-clock runtime.

This challenge is not unique to SPADs. Computationally intensive breakdown and degradation phenomena also arise in high-voltage power semiconductor devices [6] and in radiation-exposed semiconductor technologies [7], [8], where the cost of physically accurate simulation similarly limits predictive modeling and reliability analysis.

Recent work has explored machine-learning acceleration of charge-transport and Monte Carlo models, including kinetic Monte Carlo in disordered materials and neural solvers for Boltzmann transport [9]–[12]. However, many surrogates treat Monte Carlo outputs as point targets and ignore finite-sample uncertainty, which becomes critical in rare-event regimes where calibration matters [13].

In SPAD modeling, prior machine-learning work has focused on PDP prediction or inverse design, where avalanche behavior is captured indirectly through electrostatics or device-level surrogates [14], [15]. Direct prediction of spatially resolved breakdown probability from Monte Carlo injection counts across operating conditions has not been explicitly addressed in prior work.

In this work, a physics-informed cascade neural surrogate is proposed as a likelihood-aligned approximation of rare-event Monte Carlo avalanche simulations in SPADs. Here, likelihood-aligned denotes that the training objective matches the finite-sample binomial statistics of the Monte Carlo injection process, rather than treating breakdown probabilities as deterministic point targets. The surrogate mirrors the physical simulation pipeline by decomposing the problem into three sequential stages: spatial electric-field reconstruction, excess bias inference, and probabilistic breakdown prediction. By exposing physically meaningful intermediate quantities and training the final stage using a binomial negative log-likelihood

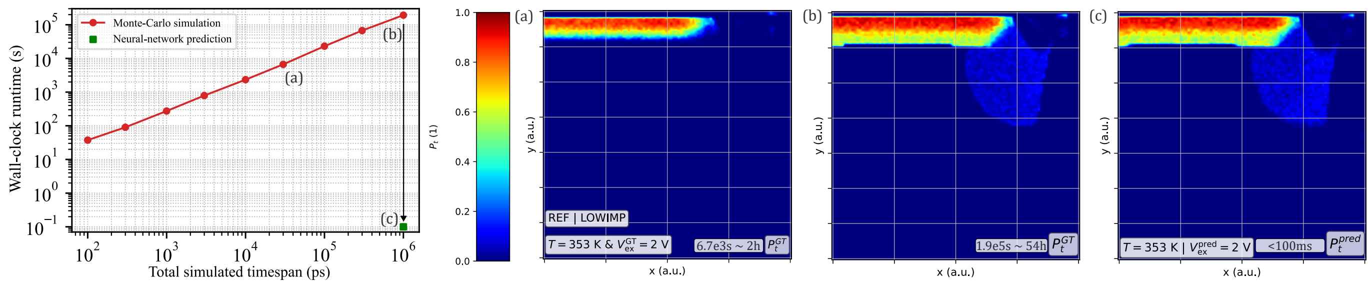

Fig. 1: Rare-event convergence of spatial breakdown probability $p_t(x, y)$. **Left:** runtime versus simulated timespan for Monte Carlo (circle) and surrogate inference (square). **Right:** p_t maps for increasing Monte Carlo budgets (a–b) and the Stage 3 surrogate prediction (c), showing closely matched distributions at orders-of-magnitude lower computational cost.

on Monte Carlo breakdown counts, physical causality and injection statistics are preserved. Calibration is assessed using spatially averaged $\langle p_t \rangle$ trends versus excess bias voltage (V_{ex}) and temperature (T) across different process conditions and device areas, where agreement with Monte Carlo is observed within expected statistical uncertainty.

The main contributions of this work are as follows:

(i) A physics-ordered causal factorization (electric field $\rightarrow$ excess bias $\rightarrow$ p_t) enabling stage-wise verification against TCAD and Monte Carlo references.

(ii) Likelihood-aligned learning of spatial breakdown probability from finite Monte Carlo injection counts via a binomial negative log-likelihood, improving calibration in the rare-event regime.

(iii) Fully self-consistent inference in which breakdown probability is predicted exclusively from surrogate-generated intermediate quantities, without TCAD or Monte Carlo inputs at test time.

(iv) Orders-of-magnitude acceleration (up to 10^6-fold) while preserving converged spatial breakdown probability maps and global $\langle p_t \rangle$ trends.

II. Background

A. Physics of avalanche breakdown in SPADs

The simulation workflow uses Sentaurus TCAD [16], [17] to extract dopant maps $N_{A,D}$ and electric-field maps E, then runs Python Monte Carlo carrier transport following [4], [5]. In this framework, carrier transport is modeled as a stochastic drift–diffusion process using an Euler–Maruyama Monte Carlo scheme [18]. Carrier trajectories are therefore represented as a drift–diffusion random walk:

$$X_{n+1} = X_n + v_d(X_n)\, dt + \sigma(X_n)\, W_n, \tag{1}$$

$$\sigma(X_n) = \sqrt{2D(X_n)\, dt}, \tag{2}$$

$$v_d(X_n) = \pm\mu_{n,p}(T, |E|)\; E(X_n), \tag{3}$$

$$D_{n,p}(X_n) = \mu_{n,p}(T, |E|)\; \frac{k_B T}{q}. \tag{4}$$

Here, X_n denotes the carrier position at iteration n, and dt is the simulation time step, with k_B the Boltzmann constant and q the elementary charge. The drift velocity v_d and diffusion coefficient $D_{n,p}$, defined through the mobility $\mu_{n,p}$, follow the Arora low-field mobility model [19] combined with the Canali high-field saturation model [20] for electrons (n) and holes (p). At each iteration, the random variables W_n are drawn from a normal distribution $\mathcal{N}(0, \sqrt{dt})$ and model Brownian motion associated with random scattering events.

Once carrier transport is modeled, the likelihood of impact ionization must be evaluated to determine whether a self-sustaining avalanche can be triggered. The cumulative impact-ionization probability along a carrier trajectory evolves as

$$p_{\text{ion}}^{n+1} = p_{\text{ion}}^{n} + |v_d(X_n)|\; dt\; \alpha_{n,p}(X_n), \tag{5}$$

where $\alpha_{n,p}$ denotes the impact-ionization coefficients for n and p. These coefficients follow the Chynoweth law [21], which models impact ionization by expressing the field- and temperature-dependent number of electron–hole pairs generated per unit distance traveled, yielding for electrons and holes the following expressions:

$$\alpha_{n,p}(|E|, T) = \gamma(T)\, a_{n,p}\, \exp\left(-\frac{\gamma(T)\, b_{n,p}}{|E|}\right), \tag{6}$$

$$\gamma(T) = \frac{\tanh\left(\frac{\hbar\omega_{\text{op}}}{2k_B T}\right)}{\tanh\left(\frac{\hbar\omega_{\text{op}}}{2k_B \cdot 300}\right)}, \tag{7}$$

The parameters $a_{n,p}$ and $b_{n,p}$ are material-dependent ionization coefficients [22], while $\gamma(T)$ accounts for phonon-assisted scattering via the optical phonon energy $\hbar\omega_{\text{op}}$, where $\hbar$ is the reduced Planck constant.

Impact ionization occurs when carriers gain sufficient kinetic energy from the electric field to generate electron–hole pairs. Increasing lattice temperature enhances carrier–phonon scattering, shortening the mean free path and reducing the probability of reaching the ionization threshold. Carrier multiplication is therefore further constrained by cumulative drift–diffusion energy gating [23], with ionization permitted only when both electron and hole energies exceed their respective thresholds [24].

In addition, carrier recombination is governed by the minority carrier lifetime [25], [26], which decreases with increasing doping concentration due to a higher density of recombination centers and with increasing temperature due to thermally activated recombination mechanisms. The temperature- and doping-dependent lifetime is modeled as

$$\tau_n = \tau_0 \left(\frac{T}{300\,\text{K}}\right)^{-\frac{3}{2}} \Big/ \left[1 + \frac{N_A + N_D}{N_0}\right], \tag{8}$$

Table I: Shared simulation and design conditions.

Factor	Values (shared)
Process	{LOWIMP, EPI1, EPI2}
Area	{REF/3, REF/2, REF/1.5, REF, REF×2}
Temperature T (K)	{253, 273, 300, 313, 333, 353}
Excess bias voltage V_{ex} (V)	{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5}

The parameters τ_0 and N_0 are fitted to experimental data [25], [26]. The cumulative recombination probability along a trajectory is then updated as

$$p_{\mathrm{rec}}^{n+1} = p_{\mathrm{rec}}^{n} + \frac{dt}{\tau_n}, \tag{9}$$

$$p_{\mathrm{rec}}^{n+1} = 1 - \exp\left(-p_{\mathrm{rec}}^{n+1}\right), \tag{10}$$

which models recombination as a stochastic Poisson process with exponential decay competing with impact ionization.

At each iteration step, two independent random thresholds are drawn for ionization and recombination, ensuring statistically independent events and preserving the Markov (memoryless) property of the process. When these random thresholds are crossed, impact ionization (electron–hole pair generation) or carrier recombination occurs. For each spatial mesh point, the Monte Carlo simulation is repeated over N_{MC} independent carrier injection trials. Self-sustaining carrier avalanche is declared when the number of active carriers exceeds a predefined threshold, and the local breakdown probability $p_t(x,y)$ is estimated as the fraction of trials that lead to breakdown. This discrete-count supervision directly motivates the likelihood-aligned formulation adopted in Stage 3 of the proposed surrogate. With the physical background and statistical estimation procedure established, the following section details the experimental setup used in this study.

III. Experimental Setup

A. TCAD and Monte-Carlo Dataset

The dataset is generated from Sentaurus TCAD simulations providing dopant distributions and electric-field maps, combined with Monte Carlo avalanche breakdown simulations. All samples share the process and operating factors summarized in **Table I**, and the dataset spans their Cartesian product, yielding approximately 10^3 spatial device maps corresponding to distinct geometry, process and operating conditions.

For each configuration, TCAD simulations provide the applied voltage V_{applied} and the corresponding breakdown voltage V_{BD}, enabling direct determination of the excess bias $V_{\mathrm{ex}} = V_{\mathrm{applied}} - V_{\mathrm{BD}}$. The simulated breakdown voltage is consistent with experimentally measured values reported in prior studies [5].

Monte Carlo simulations perform N_{MC} independent carrier injection trials per spatial location, yielding a breakdown count $k(x,y) \in \{0,\ldots,N_{\mathrm{MC}}\}$ and an empirical avalanche breakdown probability $p_t(x,y) = k(x,y)/N_{\mathrm{MC}}$. The avalanche breakdown probability has also been indirectly validated via photon detection probability (PDP) measurements reported in prior studies, since PDP is given by the product of the quantum efficiency (QE) and the avalanche breakdown probability [5]. Moreover, the temperature and bias dependence of PDP is primarily governed by the breakdown probability, motivating its explicit modeling in this study.

All quantities are stored as paired tensors (X,Y) with fixed channel ordering. The input tensor $X \in \mathbb{R}^{H\times W\times 5}$ contains spatially resolved, process-dependent device descriptors and operating-condition parameters, with channel ordering $(x,\ y,\ N_{A,D},\ T,\ V_{\mathrm{applied}})$, corresponding to spatial coordinates, doping concentration, temperature, and applied voltage. The target tensor $Y \in \mathbb{R}^{H\times W\times 6}$ contains spatial coordinates, electric-field components, excess bias, and breakdown probability with channel ordering $(x,\ y,\ E_x,\ E_y,\ V_{\mathrm{ex}},\ p_{\mathrm{t}})$. The breakdown voltage V_{BD} is stored separately for temperature-dependent calibration but is not predicted, since it corresponds to $V_{\mathrm{ex}} = 0$.

B. Train/Development/Test Splits

To ensure reproducible evaluation, the dataset is partitioned at the sample level into 80% training, 10% development, and 10% test sets using a fixed random seed, which is also used for network initialization and optimization.

The training set is used to fit all surrogate models, the development set is used for model selection and evaluation of downstream stages, and the test set is reserved exclusively for final performance assessment. For Stage 1 only, early stopping is performed using a small validation subset drawn from the training data, while the development set remains fully held out. All neural networks are implemented and trained in Python using TensorFlow [27].

C. Monte Carlo Protocol and Probability Resolution

Monte Carlo breakdown probabilities are estimated from a finite number of independent carrier injection trials per pixel. In this work, $N_{\mathrm{MC}} = 100$ trials are used, consistent with the variance–cost trade-off reported in [4]. This finite-sample binomial setting directly motivates the use of a binomial negative log-likelihood loss for training the Stage 3 surrogate. With this probabilistic formulation established, the following section introduces the proposed cascade surrogate and its three learning stages.

IV. Proposed Method

A. Cascade Neural Surrogate Overview

The cascade structure is chosen for three reasons: (i) it preserves the physical ordering of causally related quantities (electric field → excess bias → avalanche probability), (ii) it removes a direct operating-condition→ p_t pathway, improving robustness near breakdown, and (iii) it enables stage-wise validation through physically interpretable intermediate quantities that can be directly compared with TCAD- and Monte-Carlo-derived references. As illustrated in **Fig. 2**, this structure yields a three-stage surrogate that substantially reduces the computational burden of Monte Carlo simulation while preserving full spatial resolution.

Specifically, **Stage 1** predicts the spatial electric-field components (E_x, E_y) from device geometry, doping, and operating conditions. **Stage 2** infers the excess bias V_{ex} from the

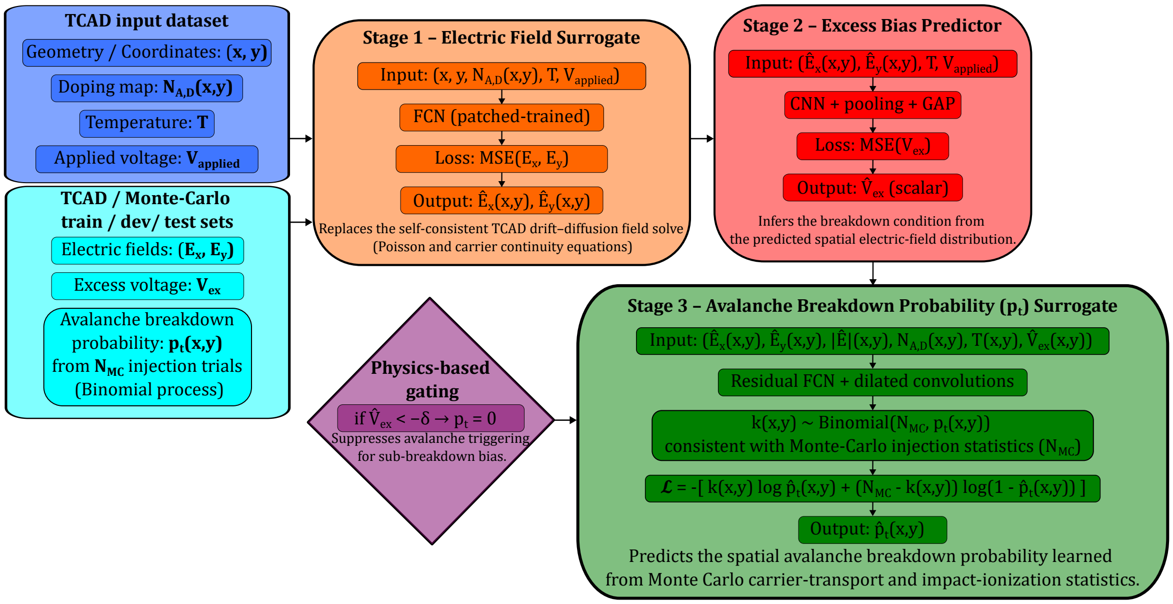

Fig. 2: Cascade neural surrogate for SPAD breakdown modeling. **Stage 1** predicts (E_x, E_y), **Stage 2** infers V_{ex}, and **Stage 3** predicts p_t using binomial supervision from Monte Carlo injections. The staged structure enforces physical causality and improves robustness near breakdown.

predicted electric-field distribution, providing a scalar descriptor of the operating point across sub- and super-breakdown regimes. **Stage 3** predicts the spatial avalanche breakdown probability $p_t(x, y)$ using the quantities inferred by the preceding stages and is trained to remain consistent with Monte Carlo injection statistics. At inference time, all stages operate exclusively on predicted quantities, ensuring self-consistent evaluation of breakdown probability.

B. Stage 1: Electric-Field Surrogate

The first stage of the cascade surrogate approximates the TCAD electrostatic solution by predicting the spatial electric-field components (E_x, E_y). This stage provides the electric-field distribution required for carrier transport and avalanche modeling in subsequent stages.

a) Architecture and training: A fully convolutional network (FCN) is used to model the electric field as a spatially resolved map. The network takes as input the spatial coordinates (x, y), the doping map $N_{A,D}(x, y)$, the temperature T, and the applied voltage V_{applied}, and outputs the electric-field components $(\hat{E}_x, \hat{E}_y)$ at each spatial location. The architecture is trained using spatial patches, allowing large device domains to be handled efficiently while preserving local spatial correlations. The electric-field surrogate is trained using TCAD-generated electric-field maps as ground truth. A mean-squared-error (MSE) loss is applied to the two-channel output, jointly penalizing errors in both x- and y-components of the electric field. Training, development, and test sets are constructed from disjoint TCAD simulations spanning process variations, geometrical scalings, temperatures, and applied voltages. During training, only TCAD quantities are used as targets, and no information from later cascade stages is introduced.

b) Results: **Fig. 3** reports representative electric-field predictions obtained with the Stage 1 surrogate. From left to right, the panels show the TCAD ground-truth (GT) spatial maps of the electric-field components (E_x, E_y), the corresponding surrogate predictions, and the associated prediction error, for a representative operating condition from the test set. The surrogate accurately reproduces both the magnitude and spatial distribution of the electric field, including the high-field region critical for impact ionization and the lower-field tail governing carrier transport. The error maps indicate that discrepancies remain localized and are small compared to the absolute field values.

Fig. 4 further evaluates the surrogate by comparing one-dimensional electric-field profiles extracted along the dashed line in **Fig. 3** across multiple temperatures at fixed applied bias and across multiple applied biases at fixed temperature. Predicted profiles closely follow the TCAD reference for both E_x and E_y across all temperatures and applied biases, preserving both the position and amplitude of the peak and tail electric-field regions. These results show that the Stage 1 surrogate provides an accurate approximation of the TCAD electrostatic field solution and can be used as input for subsequent stages of the cascade. Having established accurate field reconstruction, the next stage maps the predicted field distribution to a scalar excess bias descriptor used to condition avalanche triggering.

C. Stage 2: Excess Bias Prediction

Stage 2 aggregates the predicted electric-field distribution into a scalar excess bias voltage V_{ex}, providing a compact operating-point descriptor used to condition avalanche breakdown in the subsequent stage.

a) Architecture and training: The excess bias predictor takes as input the predicted electric-field components $(\hat{E}_x, \hat{E}_y)$

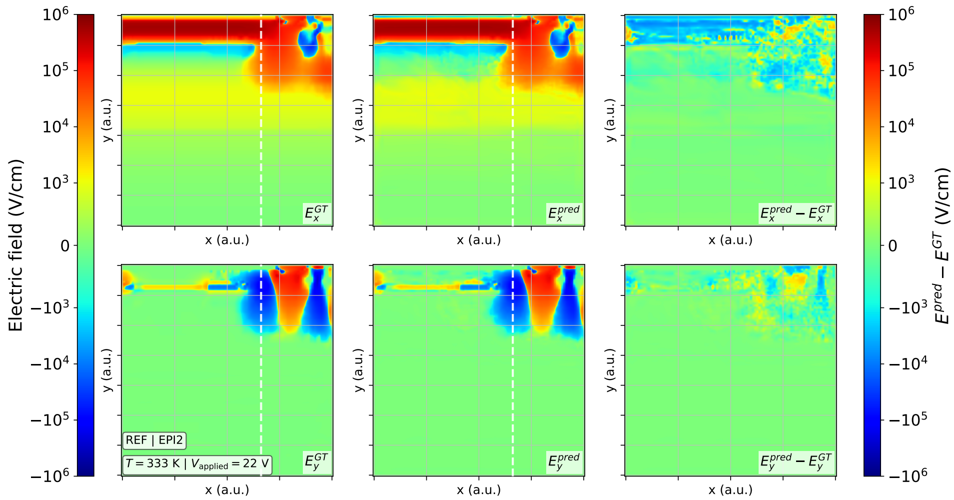

Fig. 3: **Stage 1** electric-field prediction example showing TCAD ground truth (GT), surrogate prediction (pred), and residuals for E_x (top) and E_y (bottom) at a representative operating condition. **Stage 1** accurately reproduces the high-field region governing impact ionization.

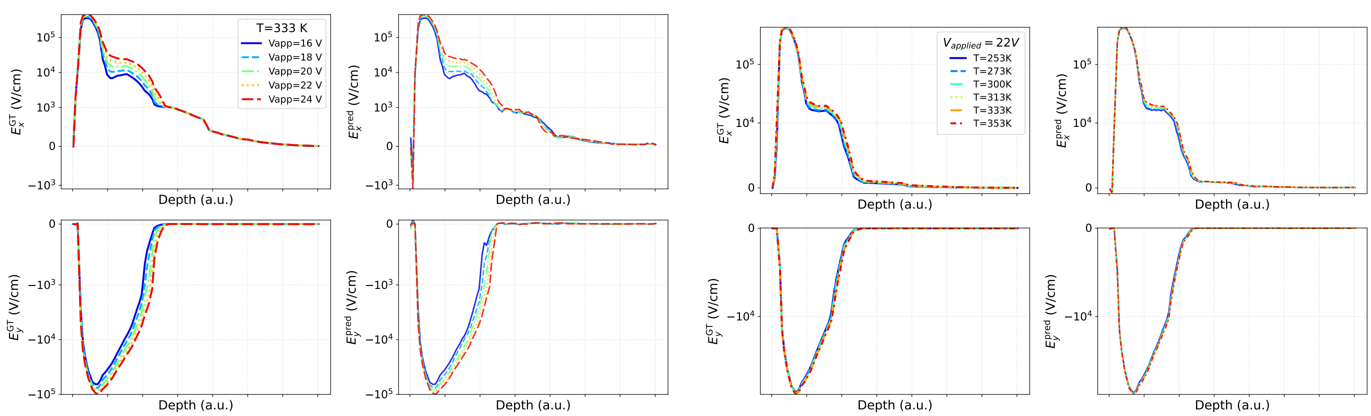

Fig. 4: **Stage 1** electric-field depth profiles along the indicated cutline in **Fig. 3**. **Left:** varying V_{applied} at fixed T; **Right:** varying T at fixed V_{applied}, comparing TCAD and predictions. **Stage 1** preserves the depth evolution of the electric field, including both peak location and magnitude as well as the transport-dominant field tail, across operating conditions.

from Stage 1 together with temperature and applied voltage. A convolutional neural network (CNN) with max-pooling layers, a final global average pooling (GAP) layer, and fully connected output layers is used to aggregate the spatial electric-field information into a scalar representation. The model is trained using a mean-squared-error loss between the predicted excess bias and the TCAD reference value, ensuring consistency with the corresponding electrostatic field solution. Stage 2 is formulated using Stage 1-predicted fields, preserving cascade consistency at inference time.

b) Results: **Fig. 5** reports parity plots comparing predicted and TCAD excess bias voltages over the test set. Prediction accuracy is quantified using the mean absolute error (MAE), root mean squared error (RMSE), and mean signed error (bias). Across the full range of operating conditions, spanning excess biases from -5 V to $+5$ V, the predictor accurately recovers the excess bias, achieving a MAE of 0.03 V, an RMSE of 0.04 V, and negligible bias. The left panel shows predicted versus ground-truth excess bias V_{ex}, while the right panel compares the corresponding breakdown voltage inferred from the surrogate with the TCAD reference, with samples tightly clustered around the identity line across applied voltages and temperatures.

Fig. 6 further evaluates the Stage 2 surrogate by reporting the breakdown voltage as a function of temperature for multiple device configurations. Across all area and process variants, the surrogate correctly captures both the temperature dependence of V_{BD} and the relative ordering between devices. This demonstrates that the excess bias predictor generalizes across structural and thermal variations while preserving physically meaningful trends. Overall, these results confirm that Stage 2 provides a robust and physically consistent mapping between spatial electric-field distributions and scalar operating conditions, enabling reliable conditioning of the avalanche breakdown probability in the subsequent stage.

c) Ablation study: An ablation study is conducted to assess the impact of the electric-field magnitude $|E|$ on excess bias prediction. As illustrated in **Table II**, removing $|E|$ from

Table II: **Stage 2** ablation study for $\hat{V}_{\mathrm{ex}}$ prediction.

Variant	$\mathrm{MAE}_{\mathrm{dev}}$	$\mathrm{RMSE}_{\mathrm{dev}}$	$\mathrm{MAE}_{\mathrm{test}}$	$\mathrm{RMSE}_{\mathrm{test}}$
4ch_noEmag	**0.0330**	**0.0404**	**0.0367**	**0.0428**
5ch_withEmag	0.0355	0.0440	0.0381	0.0495

the input leads to consistently lower MAE and RMSE on both development and test sets compared to the 5-channel variant. This indicates that the excess bias is sufficiently encoded by the electric-field components (E_x, E_y) and operating conditions, and that explicitly including $|E|$ introduces mild redundancy without improving accuracy. The reduced 4-channel configuration is therefore retained for the final Stage 2 model. The predicted excess bias $\hat{V}_{\mathrm{ex}}$ and electric-field distribution then serve as physically meaningful inputs for the breakdown-probability surrogate in Stage 3.

D. Stage 3: Breakdown Probability Surrogate

The third stage of the cascade surrogate predicts the spatial avalanche breakdown probability $p_t(x, y)$ from the physical quantities predicted by the previous stages. This stage replaces the explicit Monte Carlo injection process by a learned probabilistic surrogate that remains consistent with carrier-transport and impact-ionization statistics.

a) Architecture and training: The third stage implements a likelihood-aligned surrogate for rare-event Monte Carlo avalanche simulations by predicting the spatial breakdown probability $p_t(x, y)$ from the predicted electric-field components $(\hat{E}_x, \hat{E}_y)$, the resulting electric-field magnitude $|\hat{E}|$, the doping distribution, the temperature, and the predicted excess bias $\hat{V}_{\mathrm{ex}}$. This ensures that Stage 3 operates exclusively on surrogate-predicted quantities, enabling fully self-consistent cascade inference. The inclusion of $|\hat{E}|$ reflects the dominant dependence of impact-ionization rates on the electric-field magnitude, as described by (6). A residual fully convolutional network (FCN) with dilated convolutions is used to capture the non-local spatial correlations governing avalanche triggering. The model is trained using a binomial negative log-likelihood (BinomNLL) loss, denoted by $\mathcal{L}$ and defined in **Fig. 2**, consistent with Monte Carlo injection statistics, where each spatial location corresponds to N_{MC} independent Bernoulli trials, with observed breakdown counts $k(x, y) \in \{0, \ldots, N_{\mathrm{MC}}\}$ modeled as a binomial random variable. This formulation preserves the discrete nature and variance of the Monte Carlo process, yielding physically interpretable probabilities and improved calibration compared to mean-based regression losses. At inference time, a physics-based gating rule sets $\hat{p}_t(x, y) = 0$ when the predicted excess bias falls below breakdown ($\hat{V}_{\mathrm{ex}} < 0$), preventing non-physical avalanche triggering in sub-breakdown conditions.

b) Results: **Fig. 7** compares ground-truth and predicted spatial breakdown probability maps, together with the corresponding prediction error. The surrogate accurately reproduces both the magnitude and spatial localization of the breakdown probability within the high-field region, including its depth profile and lateral periphery. Residual errors remain small and spatially uncorrelated, indicating that the model does not introduce spurious avalanche-triggering sites.

Fig. 8 reports one-dimensional breakdown probability profiles extracted along the dashed line shown in **Fig. 7**, for multiple temperatures at fixed excess bias and for multiple excess bias values at fixed temperature. Predicted profiles closely follow the Monte Carlo reference across operating conditions, preserving both the peak probability and the spatial extent of the multiplication region. This agreement demonstrates that the surrogate captures the combined influence of excess bias and temperature on avalanche triggering.

Fig. 9 further evaluates the surrogate by comparing the spatially averaged breakdown probability, obtained by averaging p_t over the simulated device region, as a function of excess bias and temperature. The predicted trends match the Monte Carlo reference across all tested conditions, with deviations remaining within statistical uncertainty. These results confirm that the Stage 3 surrogate preserves both local and global breakdown statistics. Overall, the third stage provides a statistically consistent and physically constrained approximation of Monte Carlo avalanche simulations, completing the cascade surrogate and enabling efficient, self-consistent prediction of

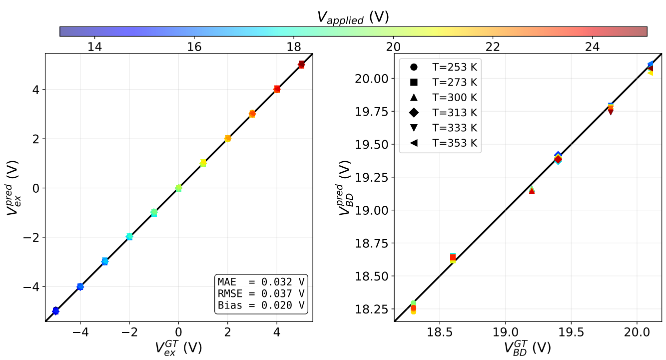

Fig. 5: **Stage 2** parity for excess and breakdown voltages, with applied voltage encoded by color and temperature by marker shape. **Left:** predicted versus ground-truth V_{ex}; **Right:** predicted versus ground-truth V_{BD}. **Stage 2** accurately recovers the device operating point.

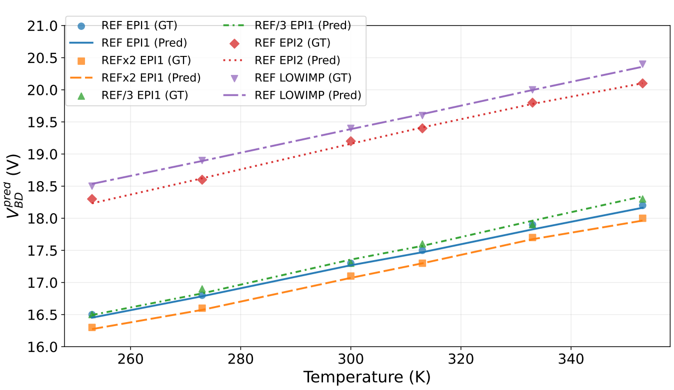

Fig. 6: **Stage 2** breakdown voltage versus temperature for selected devices. Markers show TCAD values and lines show predictions. **Stage 2** captures temperature trends and device-to-device ordering.

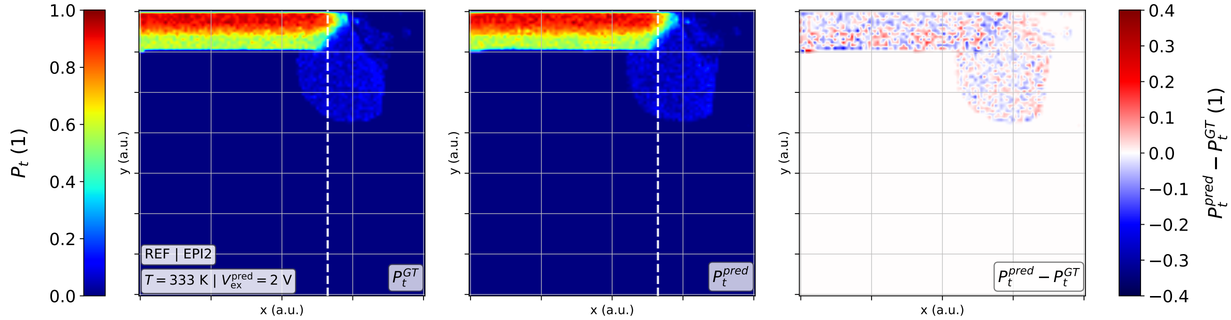

Fig. 7: **Stage 3** spatial breakdown probability maps. Shown are Monte Carlo ground truth p_t^{GT}, prediction p_t^{pred}, and residual at a representative condition. **Stage 3** reproduces both magnitude and localization of avalanche triggering.

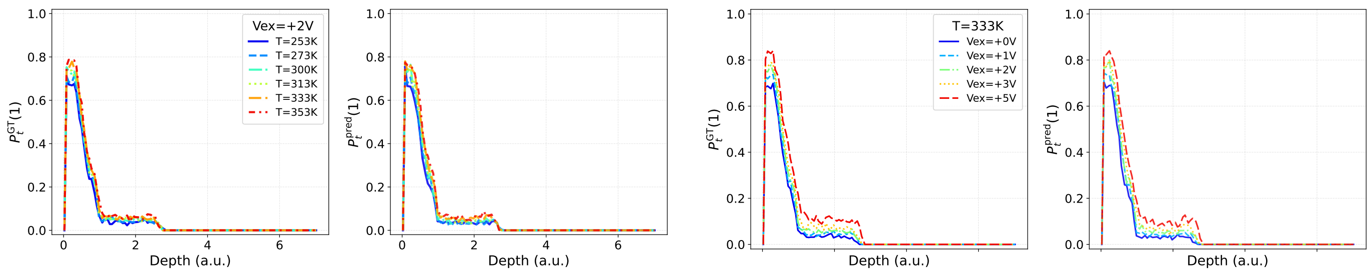

Fig. 8: **Stage 3** breakdown-probability depth profiles along the indicated cutline in **Fig. 7**. **Left:** varying T at fixed V_{ex}; **Right:** varying V_{ex} at fixed T. **Stage 3** accurately reproduces the depth profile of the breakdown probability.

breakdown probability across operating conditions.

c) Ablation study: **Table III** reports a comprehensive ablation study of the Stage 3 breakdown-probability surrogate, assessing the impact of excess bias handling, input conditioning, and loss formulation. All models are trained on the same dataset and use the predicted excess bias from Stage 2, enabling a controlled comparison across configurations. For the binomial negative log-likelihood loss, the baseline model achieves the lowest development and test errors. In contrast, removing information through input ablations or training constraints leads to systematic performance degradation. In particular, explicitly snapping the predicted excess bias V_{ex} to zero under near-breakdown conditions degrades both MAE and loss ($\mathcal{L}$), indicating that hard thresholding suppresses informative near-threshold variations and introduces a non-physical discontinuity. Similarly, restricting training to $V_{\mathrm{ex}} \geq 0$ yields the worst performance, highlighting the importance of exposing the model to sub-breakdown operating points for robustness and calibration near avalanche onset.

The mean-based (Huber) regression loss is included as a complementary baseline to assess whether the observed trends depend on the probabilistic formulation. Similar trends are observed under the Huber loss, with consistently higher errors than the binomial formulation. In particular, the performance gap between the baseline and the V_{ex}-snapped variant persists across both losses, confirming that smooth excess bias conditioning is beneficial irrespective of the optimization objective. Across all ablations, the binomial loss achieves lower test error and better generalization, reflecting its ability to more faithfully model the underlying carrier-injection statistics.

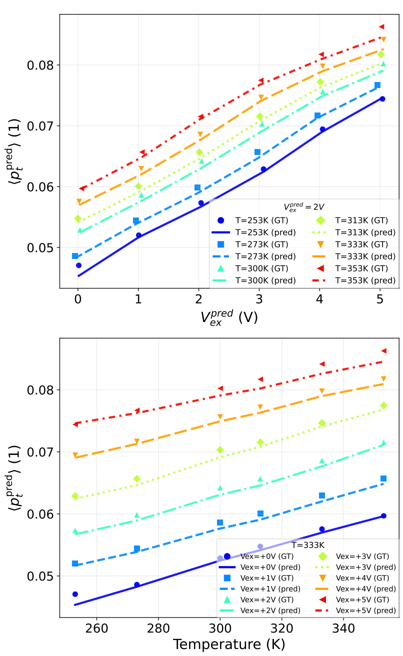

Fig. 9: **Stage 3** calibration using spatial averages $\langle p_t^{pred} \rangle$. **Top:** versus predicted V_{ex}. **Bottom:** versus temperature. **Stage 3** preserves global breakdown trends within Monte Carlo uncertainty.

V. Conclusions

This work presents a physics-informed cascade neural surrogate for efficient prediction of avalanche breakdown probability in SPADs. By decomposing the Monte Carlo simulation pipeline into sequential learning stages, the surrogate preserves

Table III: **Stage 3** ablation study for $\hat{p}_t(x, y)$ prediction.

Loss	Variant	$\mathrm{MAE}_{\mathrm{dev}}$	$\mathcal{L}_{\mathrm{dev}}$	$\mathrm{MAE}_{\mathrm{test}}$	$\mathcal{L}_{\mathrm{test}}$
BinomNLL	**Baseline**	**0.00296**	**4.024**	**0.00308**	**4.089**
	No V_{applied}	0.00298	4.028	0.00313	4.092
	No $\lvert E \rvert$	0.00303	4.029	0.00316	4.093
	V_{ex} snap	0.00357	4.082	0.00392	4.164
	$V_{\mathrm{ex}} \geq 0$ train	0.00394	4.098	0.00429	4.177
Mean (Huber)	Baseline	0.00298	4.058	0.00309	4.109
	No V_{applied}	0.00299	4.036	0.00318	4.107
	V_{ex} snap	0.00323	4.108	0.00352	4.186
	No $\lvert E \rvert$	0.00337	4.124	0.00375	4.206
	$V_{\mathrm{ex}} \geq 0$ train	0.00351	4.148	0.00399	4.232

the physical coupling between electric field, excess bias voltage, and avalanche triggering while reducing computational cost by up to 10^6-fold. The first two stages reconstruct the electrostatic field and operating conditions from device geometry and bias, replacing the TCAD field solve and breakdown extraction. The third stage predicts spatial breakdown probability maps using a binomial negative log-likelihood loss that explicitly models Monte Carlo injection statistics, yielding statistically consistent and well-calibrated predictions.

Across variations in process, geometry, temperature, and bias, the cascade surrogate closely matches converged Monte Carlo results at a fraction of the runtime, enabling rapid design-space exploration and breakdown-aware optimization in practical SPAD development. Overall, this work demonstrates that physically structured neural surrogates can faithfully approximate stochastic, rare-event device simulations and can be transposed to a wide class of semiconductor transport problems. Future work will focus on learning automated diagnostics of electrostatic barriers and carrier-trapping regions directly from electric-field vector structure (e.g., non-uniformity and directional consistency), and on jointly optimizing these diagnostics with V_{BD} and $\langle p_t \rangle$ for scalable, breakdown-aware device optimization.

Acknowledgements

The author thanks the X-FAB Opto & Sensors team across Erfurt, Germany, and Kuching, Malaysia, for their support in SPAD process development.

References

[1] F. Piron, B. Morrison, M. Yuce, and J.-M. Redouté, "A review of single-photon avalanche diode time-of-flight imaging sensor arrays," *IEEE Sensors Journal*, 2021.

[2] C. Bruschini, H. Homulle, S. Burri, and E. Charbon, "Single-photon avalanche diode imagers in biophotonics," *Light: Science & Applications*, vol. 8, p. 87, 2019.

[3] F. Ceccarelli, G. Acconcia, A. Gulinatti, M. Ghioni, I. Rech, and R. Osellame, "Recent advances and future perspectives of single-photon avalanche diodes for quantum photonics applications," *Advanced Quantum Technologies*, 2020.

[4] M. Sicre, "Study of the noise aging mechanisms in single-photon avalanche photodiode for time-of-flight imaging," Ph.D. dissertation, INSA Lyon, 2023. [Online]. Available: https://theses.hal.science/tel-04524484

[5] M. Vandwalle, A. Greve, D. Gabler, M. Skadell, H. Schmidt, A. Zimmer, V. A. J. Kumar, and O. Y. Ling, "Drift-diffusion monte carlo simulation for predictive photon detection probability in fsi spads across uv–vis–ir," in *2025 IEEE International Electron Devices Meeting (IEDM)*, 2025, pp. 1–4.

[6] H. Zhou, X. Wang, S. Wang, C. Liu, D. Chen, J. Li, L. Ma, and G. Zhang, "Deep learning method for breakdown voltage and forward i–v characteristic prediction of silicon carbide schottky barrier diodes," *Micromachines*, vol. 16, no. 5, p. 583, 2025.

[7] Y. Zhang, X. Li, S. Wang, and D. Chen, "Machine learning-based modeling of total ionizing dose effects in mos devices," *IEEE Transactions on Nuclear Science*, vol. 69, no. 8, pp. 1895–1902, 2022.

[8] C. Liu, X. Wang, and G. Zhang, "Neural network surrogate modeling for radiation effects in semiconductor devices," *Microelectronics Reliability*, vol. 148, p. 115018, 2023.

[9] O. Çaylak, A. Yaman, and B. Baumeier, "Evolutionary approach to constructing a deep feedforward neural network for prediction of electronic coupling elements in molecular materials," *Journal of Chemical Theory and Computation*, vol. 15, no. 3, pp. 1777–1784, 2019.

[10] O. T. Unke and M. Meuwly, "Machine learning for charge transport in molecular materials," *Advanced Theory and Simulations*, vol. 1, no. 4, p. 1800136, 2018.

[11] Q. Lin, C. Zhang, X. Meng, and Z. Guo, "Monte carlo physics-informed neural networks for multiscale heat conduction via phonon boltzmann transport equation," *arXiv preprint arXiv:2408.10965*, 2024. [Online]. Available: https://arxiv.org/abs/2408.10965

[12] Z. Li, Y. Chen, and Z. Hu, "Differentiable solvers for the boltzmann transport equation," *npj Computational Materials*, vol. 11, p. 42, 2025.

[13] O. F. Erdem, D. P. Broughton, J. Svoboda, C. Huang, and M. I. Radaideh, "On the impact of monte carlo statistical uncertainty on surrogate-based design optimization," *arXiv preprint arXiv:2506.00018*, 2025. [Online]. Available: https://arxiv.org/abs/2506.00018

[14] X. Qian, W. Jiang, and M. J. Deen, "Enhanced photon detection probability model for single-photon avalanche diodes in tcad with machine learning," in *2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2022.

[15] B. Zhang, Z. Li, Z. Wang, Y. Yu, H. H. Tan, C. Jagadish, D. Dong, and L. Fu, "Physics-aware inverse design for nanowire single-photon avalanche detectors via deep learning," *arXiv*, 2025. [Online]. Available: https://arxiv.org/abs/2502.18857

[16] Synopsys, Inc., *Sentaurus Process User Guide*, Synopsys, Inc., 2024.

[17] ——, *Sentaurus Device User Guide*, Synopsys, Inc., 2024.

[18] G. A. Pavliotis, *Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations*. Springer, 2014.

[19] N. Arora, J. Hauser, and D. Roulston, "Electron and hole mobilities in silicon as a function of concentration and temperature," *IEEE Transactions on Electron Devices*, vol. 29, no. 2, pp. 292–295, 1982.

[20] C. Canali, G. Majni, R. Minder, and G. Ottaviani, "Electron and hole drift velocity measurements in silicon and their empirical relation to electric field and temperature," *IEEE Transactions on Electron Devices*, vol. 22, no. 11, pp. 1045–1047, 1975.

[21] A. G. Chynoweth, "Ionization rates for electrons and holes in silicon," *Physical Review*, vol. 109, no. 5, pp. 1537–1540, 1958.

[22] R. Van Overstraeten and H. De Man, "Measurement of the ionization rates in diffused silicon p–n junctions," *Solid-State Electronics*, vol. 13, no. 5, pp. 583–608, 1970.

[23] S. Selberherr, *Analysis and Simulation of Semiconductor Devices*. Springer-Verlag, 1984.

[24] C. H. Tan, J. C. Clark, J. P. R. David, G. J. Rees, S. A. Plimmer, R. C. Tozer, D. C. Herbert, D. J. Robbins, W. Y. Leong, and J. Newey, "Avalanche noise measurement in thin si p^+–i–n^+ diodes," *Applied Physics Letters*, vol. 76, no. 26, pp. 3926–3928, 2000.

[25] M. Tyagi and R. Van Overstraeten, "Minority carrier transport and impact ionization in silicon," *Solid-State Electronics*, vol. 26, no. 6, pp. 577–587, 1983.

[26] J. A. del Alamo, S. E. Swirhun, and R. M. Swanson, "Measurement of velocity–field characteristics of silicon," *Solid-State Electronics*, vol. 30, no. 11, pp. 1127–1136, 1987.

[27] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: https://www.tensorflow.org/