Systems & Control Transactions

Original Research Article
Peer Reviewed Conference Proceeding
ESCAPE 36 - European Symposium on Computer Aided Process Engineering
Sheffield, UK. 21-24 June 2026

PSE PRESS

Enhancing plasma etching efficiency via physics-based modeling and machine learning

Eneri Boniakou[a], Yao Xue[b], Tzannis Vasileiadis[c], Sotiris Mouchtouris[ad], Katerina Oikonomou[e], Chloi Zormpa[d], Antonios Armaou[cf*], Vassilios Constantoudis[d], Evangelos Gogolides[d], George Kokkoris[a]

[a] National Technical University of Athens, Department of Chemical Engineering, Athens 15780, Greece
[b] The Pennsylvania State University, Department of Mechanical Engineering, University Park, PA 16802, USA
[c] University of Patras, Department of Chemical Engineering, Rio, 26504, Greece
[d] National Centre for Scientific Research Demokritos, Institute of Nanoscience and Nanotechnology, Athens 15341, Greece
[e] National Centre for Scientific Research Demokritos, Institute of Informatics and Telecommunications, Athens 15341, Greece
[f] The Pennsylvania State University, Department of Chemical Engineering, University Park, PA 16802, USA
* Corresponding Author: armaou@psu.edu; armaou@upatras.gr.

ABSTRACT

Modern semiconductor manufacturing requires extreme precision as yield margins narrow in the "More-than-Moore" era. While physics-based models (PBMs) provide high-fidelity insights into plasma etching, their computational intensity—often requiring hours per simulation—renders them impractical for direct iterative optimization. This work demonstrates a hybrid framework that utilizes data-driven surrogate models to enable rapid, cost-effective process optimization. A 2D axisymmetric fluid model of an inductively coupled O_2 plasma (ICP) reactor was developed to generate a training dataset for two neural architectures: a Multi-Layer Perceptron (MLP) and a Kolmogorov-Arnold Network (KAN). These surrogates predict radial etching rates across a wide operating window of power, pressure, gas flow, and bias voltage. By replacing the expensive PBM with these high-speed surrogates, derivative-free optimization algorithms (Nelder-Mead and Powell) successfully identified a profit-maximizing operating point (2000 W, 10 mTorr) orders of magnitude faster than direct physical simulation. The results confirm that surrogate-based optimization effectively captures dominant physical trends, such as ion-flux limited regimes, while providing a "Confidence Gap" through model disagreement to flag epistemic uncertainty. This methodology offers a scalable blueprint for reducing the computational burden of process design, transitioning from expensive trial-and-error to efficient, physics-validated autonomous discovery.

Keywords: Modelling and Simulations, Machine Learning, Optimization, Industry 4.0, Plasma process

INTRODUCTION

The semiconductor industry is currently navigating a “More-than-Moore” era. Until recently, critical dimensions kept shrinking until they reached a scale below the 5 nm node, where the economic and technical constraints on device fabrication have intensified. Yield margins are becoming increasingly narrow, requiring strict process control and optimization. Thus, the ability to precisely model and optimize fabrication steps is no longer merely advantageous but essential for maintaining economic viability in high-volume manufacturing [1].

Low pressure plasma, an electrically neutral ionized gas, is formed in a low pressure environment (typically 1-100 mTorr) by applying voltage between two electrodes or through coil excitation. Characterized by complex physical and chemical interactions, these discharges enable unique surface modifications unmatched by other commercial methods [2]. Among the various configurations, inductively coupled plasmas (ICPs) are of particular importance in semiconductor manufacturing [3].

Plasma etching is an essential dry fabrication process in modern integrated circuit production. It uses reactive species and energetic ions to selectively remove material from surfaces, enabling highly anisotropic and controllable pattern transfer from lithographic masks [2]. In an ICP reactor, an oscillating magnetic field is generated by a coil which in turn induces a strong electric field

capable of sustaining high plasma densities. Furthermore, wafer biasing controls the energy of ions reaching the wafer, enabling high etching rates and selectivity [4].

Numerical simulations of these reactors have become an important tool [5], which enables an improved understanding of the process and allows replacing in-situ measurements which are expensive or even impossible [6]. However, accurate predictions of spatially varying etching rates are challenging due to the complex interplay of the governing physics. This challenge is compounded by the significantly nonlinear nature of the problem, which elevates the complexity of optimization process requiring multiple iterations of the computationally expensive physics-based model (PBM).

To address these challenges, the industry is increasingly looking toward "virtual metrology" and digital twin technologies [1]. Currently, two distinct modeling paradigms exist. PBMs offer high fidelity and interpretability but are computationally expensive, often requiring hours to simulate a single operating point. Data-driven Machine Learning (ML) model approaches offer rapid inference speeds suitable for real-time control but typically suffer from poor generalizability outside their training data, a lack of physical interpretability and reliability which is directly correlated to the accuracy and the volume of the training data.

PHYSICS BASED MODEL

Fluid model plasma simulations have been proved to give valid approaches of plasma processes with typical ICP chambers of pressures in the order of $10\ mTorr$ [6].

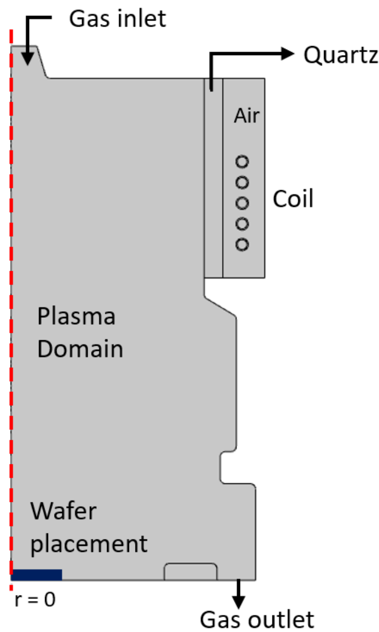

Figure 1. Computational domain of the *COBRA* reactor showing its major parts.

In this paper, a fluid model is developed, for the simulation of the plasma gas phase at low pressure conditions, focusing on etching PMMA blanket samples by O_2 gas within Oxfords Instruments' *PlasmaPro 100 Cobra* ICP plasma reactor. In this system, a 5-turn coil, placed behind a dielectric window, driven by a 13.56 MHz RF generator, couples energy electromagnetically into the plasma.

In the fluid model, species densities, electron density, and electron energy evolve in response to key operating conditions: coil power, pressure, gas feed rate, and wafer bias voltage. Given the reactor's axial symmetry, the problem is modeled in 2D axisymmetric coordinates. Figure 1 illustrates the computational domain representing the reactor geometry.

Briefly, for the electromagnetic field, Maxwell's equation set is solved. Ampere's law [eqs. (1)-(2)] is used to calculate the magnetic vector potential in the plasma region, in the dielectric window, in the coil domain and in the air contained in the domain covering the antenna

$$\nabla \times \mathbf{B} = \mu_0 \mu_r \mathbf{J_{tot}} \quad (1)$$

$$\mathbf{B} = \nabla \times \mathbf{A} \quad (2)$$

where $\mathbf{A}$ is the magnetic vector potential, μ_0 and μ_r are the permeability of free space and the domain material (quartz or air) respectively, and $\boldsymbol{B}$ is the magnetic flux density. $\mathbf{J_{tot}}$ is the vector of the total current density in each domain. For the electric potential, Poisson's equation [eq. (3)] is utilized

$$-\nabla^2 V = \frac{\rho}{\epsilon_0 \epsilon_r} \quad (3)$$

where ϵ_0 and ϵ_r is the permittivity of free space and material respectively, and ρ is the charge density in the plasma domain.

The electrical conductivity of each domain is a key model parameter. Known values are used for air and quartz, while copper's conductivity is assigned to the coil. For the plasma domain, the conductivity is updated dynamically at each time step according to eq. (4)

$$\sigma_{plasma} = \frac{N_e q^2}{m_e(\upsilon + j\omega)} \quad (4)$$

where N_e refers to the electron density, q is its charge, j is the unit imaginary number and ω stands for the angular frequency of the antenna. The total collision frequency υ is computed by summing all electron impact reaction rates and dividing by the electron density.

Regarding the electron impact reactions, their rate constants are calculated by integrating the product of the electron energy distribution function (EEDF) and the reaction-specific cross-section over the electron energy spectrum as given by eq. (5). For the O_2 plasma, the EEDF is approximated as Maxwellian, consistent with ref.

[7], thus

$$k_{e,i} = \int_0^\infty \sqrt{\frac{2q}{m_e}} \epsilon \sigma_i f(\epsilon) d\epsilon \quad (5)$$

where ϵ is the electron energy, σ_i is the i-th electron impact reaction cross section and f is the EEDF.

The mass balances for heavy species (neutral or ionic) are also solved in the model as shown, for a species k, in eq. (6)

$$\frac{\partial(\rho w_k)}{\partial t} = R_k - \nabla \cdot \mathbf{\Gamma}_k - \nabla \cdot (\rho \boldsymbol{u} w_k) \quad (6)$$

where ρ, w_k, R_k, and $\mathbf{\Gamma_k}$ are the gas mixture density, the mass fraction of species k, the production/consumption rate due to reactions in the gas phase, and the flux of k respectively. $\boldsymbol{u}$ is the fluid velocity. The flux is given from the drift-diffusion approximation [5] and the velocity is computed from the continuity and momentum conservation equations. The gas temperature is calculated by solving the gas energy balance, which accounts for heating contributions from electron–neutral collisions, thermochemical reactions, Joule heating, and the Frank–Condon mechanism, based on ref. [8]. The PBM is implemented and numerically solved using COMSOL Multiphysics.

The set of reactions employed for the O_2 plasma is based on ref. [7]. A comprehensive description of the underlying reaction mechanisms is omitted for brevity. Briefly, it includes elastic collisions of O_2 and O with electrons, electron impact ionization, O_2 dissociation, and excitation reactions. At the walls, ions are neutralized and excited states return to their corresponding ground state. Electrically, the inner walls are set to ground potential, while the inlet and outlet boundaries maintain zero net current. A no-slip condition is applied to all wall surfaces, and the inlet and outlet are modeled with a fully developed velocity profile.

At the wafer level, the PMMA etching surface model follows the framework of ref. [9]. The dominant mechanism is ion-enhanced etching, driven by the interaction of adsorbed neutral oxygen with ions. Under typical plasma etching conditions, ion enhanced etching dominates over physical sputtering and pure chemical etching. In this mechanism, the oxidation of carbon in the PMMA polymer acts as the rate-determining step, while the desorption of volatile reaction products is rapid. Consequently, the overall etching rate depends on the fluxes of both ions and neutrals, as well as the ion energy.

SURROGATE MODELING

Multi-Layer Perceptron

The trained Multi-Layer Perceptron (MLP) has as input the 4 operating conditions and predicts 10 etching rates across the wafer radius.

Table 1: Process operation parameters' range

Parameter	Lower Bound	Higher Bound
ICP Power (W)	500	2000
Pressure (mTorr)	10	100
Feed (sccm)	40	150
Voltage Bias (V)	0	700

The PBM was run iteratively 2000 times under varying operating conditions to generate an ensemble of data for model training. The simulation inputs were sampled from the operation windows presented in Table 1. The solution of the PBM for a single set of operating conditions requires ~1h (Threadripper 3960X, 264GB RAM). Following each simulation, 10 radial etching rate values along the wafer are exported. The dataset was partitioned into training (60%), validation (20%), and testing (20%) sets. To rigorously assess model generalization, the test set deliberately occupied a distinct subspace comprising edge cases of the parameter space. This ensured the model was evaluated on challenging, unseen conditions rather than interpolated data points.

To ensure stable and efficient network training, all input and output variables were min-max normalized to a $[0, 1]$ range [10]. This scaling prevents features with wider numerical ranges (e.g., etching rate) from dominating the learning process and helps the optimizer converge faster by mitigating gradient instability. The implemented model was a fully connected MLP with a rather simple architecture of an input layer with 4 neurons, two hidden layers with 20 neurons each, and an output layer with 10 neurons. The hidden layers utilized the Rectified Linear Unit (ReLU) activation function to introduce non-linearity while maintaining stable gradient flow during backpropagation. Experimental fine-tuning indicated that ReLU hidden layers with an Exponential Linear Unit (ELU) [11] activated output layer, had the best performance with a smooth, continuous regression at the output.

Weighted mean squared error (MSE) was implemented with L1 regularization which is ideal for regression tasks as a loss function [12]. The L1 penalty term, controlled by the hyperparameter λ, was applied to network weights to improve generalization. During training λ was set to 10^{-3}. The function is formalized as

$$L = \frac{1}{N} \sum_{m=1}^{N} \sum_{j=1}^{10} \frac{1}{10} \left(y_{m,j} - \hat{y}_{m,j}\right)^2 + \lambda |W_k| \quad (7)$$

where 10 is the set of output etching rate points, N the batch size, W_k refers to the model weights and $y_{m,j}, \hat{y}_{m,j}$ are the true and predicted values for the sample m and radial position j.

Training was performed using the Adam optimizer [13], with an initial learning rate of $3 \cdot 10^{-3}$ and a weight decay of $2 \cdot 10^{-5}$. A scheduler was configured to monitor the validation loss and reduce the learning rate by a factor of 10 if no improvement was observed over 5 epochs.

To prevent overfitting, early stopping was employed based on validation loss, to terminate the training if no improvement was observed, reverting the model to the state with the best validation performance. The batch size of training was set to 32. The initial learning rate, the weight decay and the batch size were optimized via grid search based on minimizing the loss. Figure 2 shows the MLP model accuracy with high R^2 (0.97). Points cluster tightly along the ideal line, confirming the model captures process physics well. Minor divergence at data edges is expected due to sparse training in those regions and greater uncertainty in the PBM itself. The training and inference times are quite low; 10 min and < 10^{-3} s respectively (Threadripper 3960X, 264GB RAM).

Table 2: Comparison of KAN and MLP architectures

Aspect	MLP	KAN
Activation Location	On nodes	On edges
Activation Type	Fixed (ReLU, etc.)	Learnable (B-splines)
Nonlinearity source	Stacked layers	Flexible edge functions
Smoothness	Not guaranteed	Inherently smooth
Parameter efficiency	Lower	Higher
Interpretability	Limited	Improved

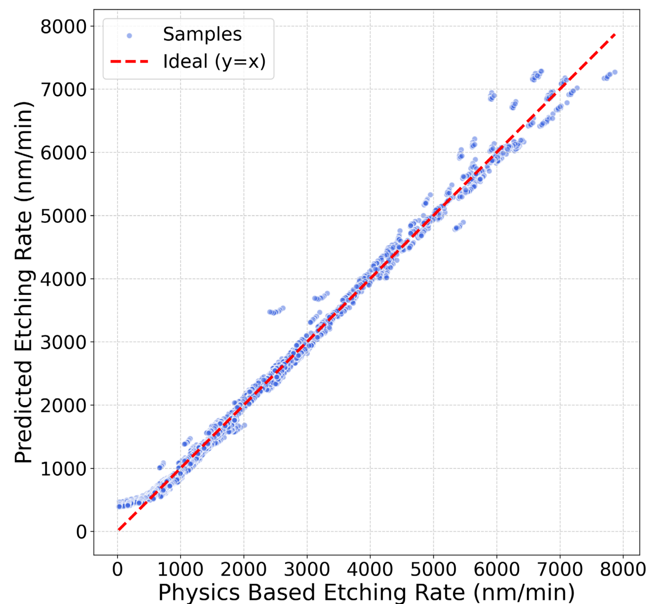

Figure 2. Scatter plot of the MLP model (4 inputs, 10 outputs). Model scores: $R^2 = 0.97$ and a MSE of 0.002.

Kolmogorov-Arnold Network

We also employed Kolmogorov-Arnold Networks (KANs) [14] as our base architecture for etch rate prediction. This model's training time was 5 minutes on a GeForce Titan RTX, requiring about 1GB of GRAM.

Unlike traditional MLPs that apply fixed activation functions (e.g., ReLU, sigmoid) on nodes, KANs place learnable activation functions on edges. Each edge activation is parameterized by B-spline basis functions

$$\phi(x) = \sum_{j=1}^{G+k} c_j B_j^k(x) \quad (8)$$

where B_j^k are B-spline basis functions of order k defined on a grid with G intervals, and c_j are learnable coefficients. This allows KANs to adaptively learn activation shapes suited to specific input-output relationships, rather than relying on predefined fixed nonlinearities.

Advantages of B-Spline Parameterization

The B-spline parameterization offers several advantages for scientific modeling tasks, such as smoothness, local support and parameter efficiency. B-splines are inherently smooth functions, making them well-suited for approximating continuous physical quantities where abrupt discontinuities are not expected. Also, each basis function has compact support, enabling efficient local adaptation to input variations without causing global interference in other regions of the input space. Lastly, KANs can achieve comparable or superior accuracy with fewer parameters than MLPs, as the flexible learnable activation functions can capture complex nonlinear relationships more efficiently than stacking many layers with fixed activations.

Plasma-Encoded Formulation

To enable continuous spatial prediction, we adopt a position-encoded formulation

$$\hat{r}(p) = KAN([x, p]) \quad (9)$$

where the radial position $p \in [-1, 1]$ is concatenated with process parameters as network input, $x \in R^4$ denotes the same operating parameters as previously, and $\hat{r}(p)$ is the predicted etch rate at normalized radial position p.

This pointwise formulation transforms the problem from predicting a fixed-length output vector to learning a continuous function over spatial coordinates. Rather than training a model to output $[r_1, r_2, \ldots, r_{10}]$ simultaneously, we train it to predict the etching rate at any given position, allowing:

- **Arbitrary resolution:** Prediction can be made at any spatial location, not just the measurement points.
- **Implicit spatial coherence:** The continuous function representation naturally enforces smooth transitions between nearby positions.
- **Better generalization:** The network learns the underlying continuous mapping rather than memorizing discrete point correspondences.

The KAN's smooth B-spline activations further

promote smooth spatial variations in the output, which is consistent with physical expectations for plasma etching processes where etching rates vary continuously across the wafer surface.

OPTIMIZATION

Objective Function Formulation

The optimization problem is formulated to maximize the net economic value generated by the etching process per unit time (€/s). The objective function, $J(x, y)$, accounts for both the value of the processed wafer and the operational expenditures. It is defined as

$$J(x, y) = P_{Efficiency} - P_{Cost} \quad (10)$$

$$x = Power, Pressure, Feed, VBias \quad (11)$$

$$y = Er_{radial} \quad (12)$$

where the decision vector x includes the controllable operation parameters (Source Power, Pressure, Feed Flow, Bias Voltage) and y represents the predicted radial etching rate profile derived from the surrogate models.

The efficiency term $(P_{Efficiency})$ maps the physical output to monetary gain. It is defined as the product of the mean etching rate (Er_{mean}) normalised to the wafer area, the effective process yield (Y), and the economic value per unit of etched thickness (G_h):

$$P_{Efficiency} = Y \cdot Er_{mean} \cdot G_h \quad (13)$$

In this context, yield represents the fraction of wafer surface area that falls within strict uniformity tolerance bounds; while high etching rates increase throughput, the function degrades rapidly if the radial profile deviates from the wafer center maximum, ensuring speed is not prioritized at the expense of critical dimension uniformity.

The cost term (P_{Cost}) consists of the power, material and environmental harm costs. The power cost (C_{Power}) includes the RF power (source and bias) and facility baseloads. We observe that the total power cost variance is minimal across the parameter space, as the fixed power consumption of auxiliary systems (e.g., pumps, cleanroom infrastructure) outweighs the variable RF load.

$$P_{Cost} = C_{Power} + C_{Mat} + C_{Env} \quad (14)$$

The material cost (C_{Mat}) accounts for the consumption of process gases. While the optimization strictly controls the flow of the reactive precursor (Feed), the total material cost is largely driven by the steady-state consumption of inert carrier gases required for the reactor's operation. Similar to the power term, the relative variance of material costs remains low across the parameter bounds, despite constituting a significant fraction of the absolute operational expenditure. The environmental term (C_{Env}) calculates costs associated with the Global Warming Potential (GWP) of the effluent. In the studied O_2 plasma, GWP costs are negligible, though the framework is designed to scale for more complex chemistries where environmental abatement is a dominant cost driver.

The objective function can be readily augmented to include upstream lithography parameters (e.g., mask selectivity interactions) or auxiliary economic terms such as equipment maintenance cycles, thereby enabling a holistic optimization of the broader patterning module without altering the fundamental control logic.

Optimization Strategy

The optimization of the reactor setpoints is implemented within Python as a derivative-free optimization (DFO) problem. Given the non-convex nature of the neural network response surfaces and the discontinuities introduced by the yield penalty function (Y), gradient-based methods are prone to instability or premature convergence to local optima. Consequently, we validate our findings by executing the optimization task using two independent direct-search algorithms: Nelder-Mead (NM) and Powell's Conjugate Direction (PCD) methods [15, 16].

Rather than relying on a single solver, conducting separate, independent optimization searches with these distinct methods serves as a critical verification. The simplex-based approach of NM and the directional line-search approach of PCD operate via different mechanisms; their convergence to a common basin of attraction, despite their differing search trajectories, increases our confidence in the optimality of the identified solution and independence from algorithmic artifacts.

The optimization is constrained within a hyper-rectangle defined by the physical limits of the training data, presented in Table 1. To facilitate unbiased exploration, both solvers are initialized using a "Warm Start" strategy originating from the centroid of the hyperparameter space. This neutral initialization point prevents the optimizer from being prejudiced toward specific high-power or high-pressure regimes, allowing the objective function gradient to naturally drive the trajectory toward the most profitable operating window.

Optimization Results

The optimization results for both surrogate architectures are summarized in Tables 3 and 4. For each respective surrogate model, both the NM and PCD algorithms converged to the same operating point. Combined with the multistart method, this confirms that the identified solutions represent the true global maximums of the respective neural network response surfaces, independent of the optimization method employed.

Table 3: MLP Optimization Results

Optimal Conditions	$\boldsymbol{J}$ [=] €/$\boldsymbol{s}$	**10.55**

Power (W)	2000	$P_{Efficiency}$	13.64
Pressure (mTorr)	10	C_{Power}	$3.15\ e^{-4}$
Feed (sccm)	40	C_{Mat}	3.09
Voltage Bias (V)	700	C_{Env}	$2.18\ e^{-8}$

Table 4: KAN Optimization Results

Optimal Conditions		$\boldsymbol{J}\ [=]\ \text{€}/\boldsymbol{s}$	**10.47**
Power (W)	2000	$P_{Efficiency}$	13.57
Pressure (mTorr)	10	C_{Power}	$3.14\ e^{-4}$
Feed (sccm)	105	C_{Mat}	3.09
Voltage Bias (V)	686	C_{Env}	$5.72\ e^{-8}$

Figure 3 illustrates the optimization convergence, highlighting the prohibitive cost of direct optimization. Since a PBM prediction takes about an hour, the numerous iterations required would be computationally impractical to run. In contrast, the surrogate models allowed the entire process to complete on a mid-tier CPU in approximately 3-4 minutes per method employed.

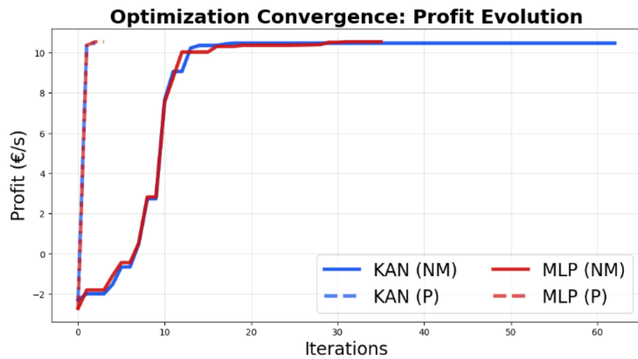

Figure 3. Optimization Convergence of each method

To validate the accuracy of these solutions, the optimal etching rate profiles predicted by both surrogates are compared against the PBM ground truth in Figure 4. The PBM reference profile was generated using the optimal conditions identified by the MLP surrogate. As illustrated, both the MLP and KAN surrogates yield radial profiles that closely align with the physical baseline, despite the divergence in their secondary Feed Flow parameters.

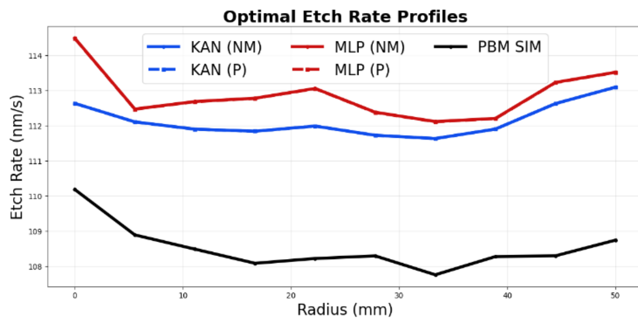

Figure 4. Etching rate profiles for both models and optimization methods, compared to PBM profile.

Sensitivity Analysis and Exploration

While the optimization algorithms converged effectively, the divergence in the final setpoints necessitates a closer examination of the surrogate response surfaces. To understand the reasons behind the different optimal conditions, local sensitivity sweeps are performed around the high-power/low-pressure $(2000\ W, 10\ mTorr)$ area where both models agreed.

As established in the optimization results, both surrogates identified Power and Pressure as the primary control levers. Sensitivity sweeps confirm that both the MLP and KAN correctly approximate the PBM gradients in these dimensions, identifying the upper bound of power and the lower bound of pressure as the trajectory for maximizing etching rate. This alignment confirms that both architectures successfully captured the global trend of the dominant variables.

A minor discrepancy was observed in the Bias Voltage, where the MLP model converged to $700\ V$ while the KAN settled at $686\ V$. Despite this offset, the behavior of both models remains largely consistent, reaffirming the parameter's critical influence on the objective function. To verify this, a sensitivity sweep was conducted by varying Voltage Bias across its design limits while holding the rest of the parameters at their optimal values, shown in Fig. 5. This isolation reveals the distinct response surfaces of each surrogate: while both models correctly and similarly identify the parameter importance, the KAN response exhibits a stagnation effect near the upper bound, whereas the MLP maintains a steeper gradient.

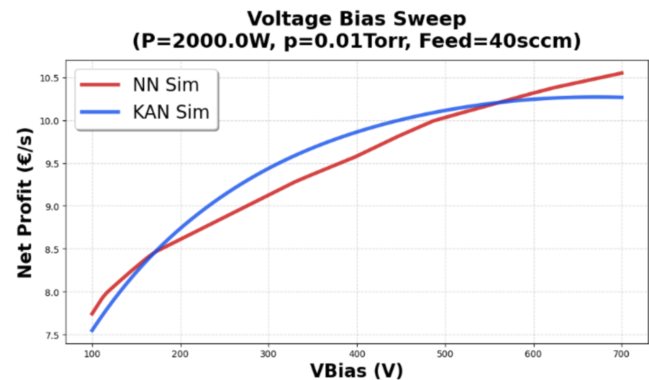

Figure 5. VBias response profiles: MLP(NN) and KAN

The most significant divergence occurred for the Feed parameter $(40\ sccm\ vs\ 105\ sccm)$. To analyze this, the Feed was swept across the design space while holding other inputs at their optimal values, shown in Figure 6; it reveals a distinct conflict in functional approximation. The Ground Truth for the surrogates, i.e., PBM, behaves near-linearly, showing a slight preference for lower flow rates. It also showcases the impact of the slight profile

difference between the PBM and the surrogates, as shown in Figure 4, on the objective function. The MLP surrogate is closer to mimicking the PBM linear behavior. It correctly identifies the negative slope and directs the optimizer toward the lower bound, effectively matching the true physical optimal point. The KAN surrogate introduces an artificial curvature. While its predictions are closer to the PBM at the boundaries, it significantly deviates in the center of the domain, creating a parabolic shape that drives the optimizer towards the middle.

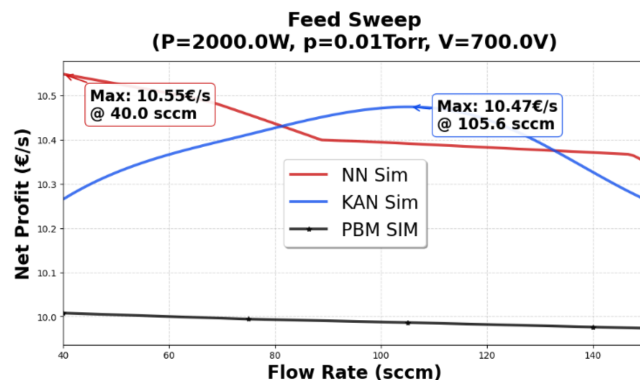

Figure 6. Feed Flow response profiles: MLP(NN) and KAN surrogates vs. PBM.

DISCUSSION

Data Sparsity

The divergence in the Feed Flow optimization highlights a fundamental difference in how the two architectures handle interpolation in sparse data regions. Despite both models being trained on the exact same dataset, the KAN architecture likely attempted to fit a higher-order non-linearity where none existed, whereas the MLP maintained a simpler linear approximation that proved more robust in this specific scenario.

This behavior could be attributed to the "Edge Effect." The optimal operating point $(2000\ W, 10\ mTorr)$ lies at the corner of the training hypercube. Neural networks are notoriously prone to high variance at the boundaries of their training domain. The KAN architecture, which utilizes learnable spline functions, is designed to be highly flexible in capturing local non-linearities. However, the results indicate that in this specific boundary region, this flexibility may have resulted in the model capturing stochastic noise rather than the true physical trend. On the other hand, the MLP appears to have defaulted to a smoother, more global approximation. In this instance, the simpler structure of the MLP likely acted as a constraint, preventing the model from inferring a local optimum where the physical landscape is effectively flat.

Physics Validation

From a plasma physics perspective, the PBM's insensitivity to Feed (and the MLP's resulting preference for the lower bound) is physically consistent with an ion-driven etching regime. At low pressures, the mean free path of species increases, and the etching mechanism becomes dominated by directional ion bombardment (controlled by Power and Bias) rather than chemical etching by neutral radicals (controlled by Feed Flow).

Consequently, variations in the gas flow have a negligible impact on the etching rate compared to the ion density. The PBM correctly reflects this "flat" landscape. The fact that the optimizer identified the correct high-power/low-pressure window proves that the framework successfully isolated the primary physics drivers, even if the secondary parameters had a diminished effect.

The Confidence Gap

The disagreement between the models in the Feed value raises a critical question regarding reliability. If a single model had been used, the optimizer would have returned a result with high apparent confidence, while it could still be misleading. The existence of this "Confidence Gap" between the two models is important. It suggests that in industrial applications, model disagreement should be treated as a proxy for epistemic uncertainty. This underscores the necessity for future frameworks to integrate Uncertainty Quantification. By flagging regions where the MLP and KAN diverge, the system could automatically trigger "active learning, " requesting targeted PBM simulations to resolve the ambiguity before finalizing the optimization, while also improving the models.

CONCLUSIONS

This work presents a framework for the optimization of plasma etching processes, aimed at bridging the fidelity of Physics-Based Models (PBMs) with the computational efficiency of data-driven surrogates. The core contribution of this study extends beyond the specific etching recipe identified; it demonstrates that integrating diverse neural architectures into the optimization loop provides a necessary proxy for prediction reliability, a critical step toward fully autonomous process discovery.

The application of this framework successfully identified a high-value operating regime characterized by high power $(2000\ W)$ and low pressure $(10\ mTorr)$. This result is physically consistent with an ion driven etching mechanism, where maximizing ion energy and density dominates value creation. Furthermore, the use of two distinct surrogate architectures (MLP and KAN) confirmed the robustness of these setpoints; both models agreed on the primary drivers, while disagreeing regarding the rest.

Beyond the specific framework, this study validates the feasibility and efficiency of surrogate-based optimization for complex industrial processes. We

demonstrated that by training neural networks on a targeted dataset of PBM simulations, it is possible to achieve global optimality orders of magnitude faster than direct physics simulation. This workflow offers a practical blueprint for reducing the computational burden of process design, allowing engineers to explore vast parameter spaces without the prohibitive time costs associated with traditional physics solvers or experimental trial-and-error.

Future work will transition from static modeling to a hybrid active learning loop. The PBM could be integrated directly into the optimization process, utilizing prediction uncertainty as a trigger. Instead of relying solely on pre-trained surrogates, the system will dynamically query the PBM whenever the certainty of a prediction drops. This creates a self-correcting workflow that autonomously improves model accuracy in high-value regions, ensuring that the final operating point is validated by physics while keeping computational costs minimal.

ACKNOWLEDGEMENTS

Support of the project "Enhancing plasma etching efficiency, repeatability, and environmental footprint via AI-based modeling and optimization (plasmaAI)" of the program "AI-Aware Pathways to Sustainable Semiconductor Process and Manufacturing Technologies", Intel Corporation & Merck KGaA, Darmstadt is gratefully acknowledged.

AUTHOR IDENTIFIERS

Author ORCIDs:
Boniakou E.: 0009-0002-7080-2837
Xue Y.: 0009-0001-1200-6157
Vasileiadis T.: 0009-0003-7122-5995
Mouchtouris S.: 0000-0001-7580-4554
Oikonomou K.: 0000-0001-7537-7310
Zormpa C.: 0009-0005-6797-9844
Armaou A.: 0000-0002-8592-7934
Kokkoris G.: 0000-0003-4507-7311
Gogolides E.: 0000-0002-1870-5629
Constantoudis V.: 0000-0003-3164-977X

REFERENCES

1. Oehrlein GS, Brandstadter SM, Bruce RL, Chang JP, DeMott JC, Donnelly VM, Dussart R, Fischer A, Gottscho RA, Hamaguchi S, Honda M, Hori M, Ishikawa K, Jaloviar SG, Kanarik KJ, Karahashi K, Ko A, Kothari H, Kuboi N, Kushner MJ, Lill T, Luan P, Mesbah A, Miller E, Nath S, Ohya Y, Omura M, Park C, Poulose J, Rauf S, Sekine M, Smith TG, Stafford N, Standaert T, Ventzek PLG. Future of plasma etching for microelectronics: challenges and opportunities. Journal of Vacuum Science & Technology B 42: (2024). https://doi.org/10.1116/6.0003579
2. Lieberman MA, Lichtenberg AJ. Principles of plasma discharges and materials processing. Wiley (2005). https://doi.org/10.1002/0471724254
3. Graves DB, Labelle CB, Kushner MJ, Aydil ES, Donnelly VM, Chang JP, Mayer P, Overzet L, Shannon S, Rauf S, Ruzic DN. Science challenges and research opportunities for plasma applications in microelectronics. Journal of Vacuum Science & Technology B 42: (2024). https://doi.org/10.1116/6.0003531
4. Tachi S, Tsujimoto K, Arai S, Kure T. Low-temperature dry etching. Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films 9:796-803 (1991). https://doi.org/10.1116/1.577364
5. Mouchtouris S, Kokkoris G. A hybrid model for low pressure inductively coupled plasmas combining a fluid model for electrons with a plasma-potential-dependent energy distribution and a fluid-monte carlo model for ions. Plasma Sources Sci. Technol. 25:025007 (2016). https://doi.org/10.1088/0963-0252/25/2/025007
6. Brezmes AO, Breitkopf C. Fast and reliable simulations of argon inductively coupled plasma using COMSOL. Vacuum 116:65-72 (2015). https://doi.org/10.1016/j.vacuum.2015.03.002
7. Gudmundsson JT, Thorsteinsson EG. Oxygen discharges diluted with argon: dissociation processes. Plasma Sources Sci Technol 16:399-412 (2007) https://doi.org/10.1088/0963-0252/16/2/022
8. Kiehlbauch MW, Graves DB. Inductively coupled plasmas in oxygen: modeling and experiment. Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films 21:660-670 (2003). https://doi.org/10.1116/1.1564024
9. Memos G, Lidorikis E, Gogolides E, Kokkoris G. A hybrid modeling framework for the investigation of surface roughening of polymers during oxygen plasma etching. J. Phys. D: Appl. Phys. 54:175205 (2021). https://doi.org/10.1088/1361-6463/abdb0b
10. Al Shalabi L, Shaaban Z. Normalization as a preprocessing engine for data mining and the approach of preference matrix. 2006 International Conference on Dependability of Computer Systems :207-214 (2006). https://doi.org/10.1109/depcos-relcomex.2006.38
11. Clevert DA, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (ELUs). arXiv preprint arXiv:1511.07289 (2015)
12. Wu S, Li G, Deng L, Liu L, Wu D, Xie Y, Shi L. $L1$ -

norm batch normalization for efficient training of deep neural networks. IEEE Trans. Neural Netw. Learning Syst. 30:2043-2051 (2019). https://doi.org/10.1109/tnnls.2018.2876179

13. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Liu Z, Wang Y, Vaidya S, Ruehle F, Halverson J, Soljačić M, Hou TY, Tegmark M. KAN: Kolmogorov-Arnold networks. arXiv preprint arXiv:2404.19756 (2024) http://arxiv.org/abs/2404.19756
15. Powell MJD. An efficient method for finding the minimum of a function of several variables without calculating derivatives. The Computer Journal 7:155-162 (1964). https://doi.org/10.1093/comjnl/7.2.155
16. Nelder JA, Mead R. A simplex method for function minimization. The Computer Journal 7:308-313 (1965). https://doi.org/10.1093/comjnl/7.4.308

© 2026 by the authors. Licensed to PSEcommunity.org and PSE Press. This is an open access article under the creative commons CC-BY-SA licensing terms. Credit must be given to creator and adaptations must be shared under the same terms. See https://creativecommons.org/licenses/by-sa/4.0/