Hybrid Graph Neural Network and Physics-Informed Neural Network for Predicting Interface Dynamics in Semiconductor Thin Film Deposition

A Thesis Submitted to the Graduate Faculty of the

National University, School of Engineering & Computing

in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

Prepared by:

Randall Crawford

Roman Ostroumov

David Homrighouse

December 25, 2025

Master Thesis Approval Form

We certify that we have read the project of Randall Crawford, Roman Ostroumov, and David Homrighouse entitled *Hybrid Graph Neural Network and Physics-Informed Neural Network for Predicting Interface Dynamics in Semiconductor Thin Film Deposition* and that, in our opinion, it is satisfactory in scope and quality as the thesis for the degree of Master of Science in Data Science at National University.

Approved:

12/31/2025

Mohammad Yavarimanesh, PhD, Capstone Project Sponsor
Assistant Professor
Department of Engineering, Data and Computer Sciences
College of Business, Engineering, and Technology
National University

Gennady Medvedkin Digitally signed by Gennady Medvedkin Date: 2025.12.31 00:34:28 -08'00'

Gennady Medvedkin, PhD and D.Sc., Capstone Project Sponsor
Director
Research and Development
General Molded Glass

Aeron Zentner

Aeron Zentner, D.B.A., Capstone Instructor
Assistant Part-Time Professor
Department of Engineering, Data and Computer Sciences
College of Business, Engineering, and Technology
National University

Acknowledgements

Our team would like to acknowledge the July 2025 review by Tao Han, Zahra Taheri, and Hyunwoong Ko, “Physics-Informed Neural Networks for Semiconductor Film Deposition: A Review,” as the primary inspiration for this project. In particular, the third paragraph of Section 5 (“Future Direction”) highlights the combination of Graph Neural Networks with Physics-Informed Neural Networks as a promising avenue, advocating an integrated approach that leverages multi-modal data fusion and embedded spatio-temporal physical constraints to deepen understanding of complex phenomena and enhance process control precision.

We are also deeply grateful to National University, the faculty of the graduate Data Science program, thesis sponsors, and our classmates who accompanied us throughout this 15-month journey. In an era when artificial intelligence is rapidly reshaping our lives, this program has equipped us exceptionally well for the future.

Abstract

Atomic layer deposition (ALD) plays a critical role in semiconductor device fabrication, where interface quality and interdiffusion control directly influence electrical performance and long-term reliability. Systematic optimization of ALD process conditions remains challenging due to the complexity of the underlying diffusion physics and the cost of experimental exploration. This thesis develops a physics-constrained surrogate modeling framework that couples a Graph Neural Network (GNN) with a Physics-Informed Neural Network (PINN) to characterize and predict interdiffusion behavior at the Si/Al_2O_3 interface.

A synthetic dataset representing realistic ALD process conditions, including temperature, pressure, and pulse time, was generated to train the surrogate model. The GNN captures spatial and structural relationships across the interface. At the same time, the PINN enforces Fickian diffusion and the Arrhenius-type temperature dependence of diffusivity, ensuring that model predictions remain consistent with established physical laws. The resulting GNN-PINN model accurately predicts diffusion profiles, interface widths, diffusivity, growth-per-cycle (GPC), and film uniformity across a multidimensional parameter space.

The model was subsequently integrated into an optimization workflow to identify ALD process conditions that jointly balance all performance metrics. The analysis reveals a small region of the design space that approaches practical physical limits for interdiffusion control while maintaining acceptable GPC and uniformity. The model demonstrated a minimum interdiffusion width of 0.48 nm while requiring only 1.97 seconds of CPU time for optimization. The framework can be extended to additional material systems and integrated with experimental data to support future interface engineering efforts.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

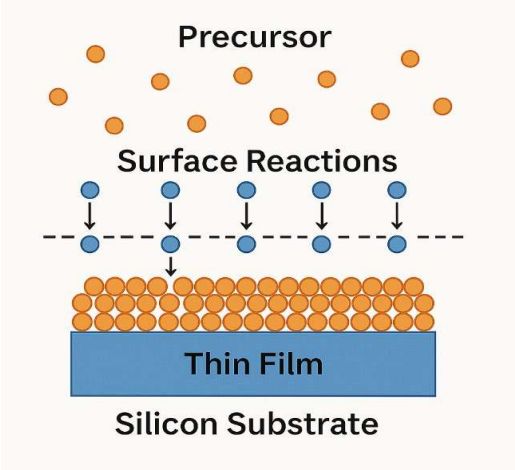
1.1 Background

Semiconductor manufacturing depends on thin-film deposition to maintain transistor scaling under Moore's Law. Since the introduction of the integrated circuit in 1958, chemical vapor deposition (CVD) and atomic layer deposition (ALD) have become the primary processes for fabricating the layered structures required for transistors, memory cells, and interconnects (Sze & Ng, 2006). ALD deposits films one atomic layer at a time through sequential, self-limiting surface reactions. This mechanism provides atomic-scale control over thickness, conformality, and uniformity (George, 2010; Vale et al., 2023). Refer to Appendix B.

Figure 1 illustrates the ALD cycle on a silicon wafer. A precursor pulse delivers molecules that chemisorb onto the surface. A purge step removes unreacted gas. A co-reactant pulse reacts only with the chemisorbed species, forming a single monolayer. A final purge clears byproducts.

Figure 1

Atomic Layer Deposition Thin-Film Growth Process



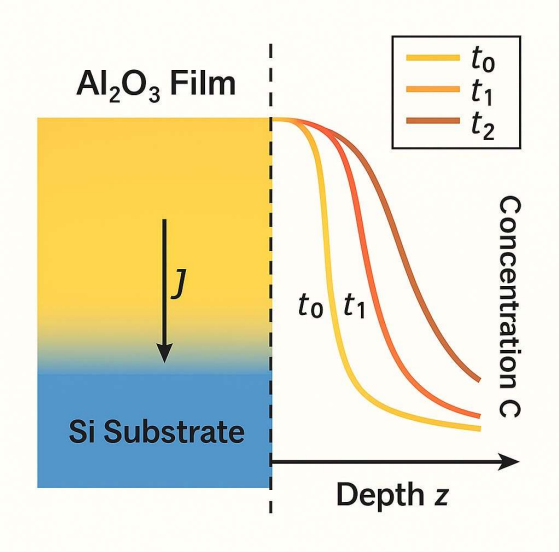
Repetition of this four-step cycle (precursor pulse → purge → co-reactant pulse → purge) builds films with sub-angstrom thickness control per cycle. Interface diffusion, however, limits the effectiveness of this control.

Figure 2 shows silicon atoms migrating into an Al_2O_3 layer during high-temperature processing. The Si/Al_2O_3 boundary starts sharply. As the temperature increases, silicon diffuses upward (flux arrow J), blurring the interface. Concentration profiles at times t_0, t_1, and t_2 demonstrate progressive broadening. The intermixing forms charge traps, raises off-state leakage current by one to two orders of magnitude, and reduces device reliability (Srolovitz & Yang, 1995).

These films are integral to device architecture, serving as inter-metal dielectrics for signal isolation, passivation layers for surface protection, and shallow trench isolation for transistor separation.

Figure 2

Diffusion Across the Si/Al_2O_3 Interface



1.2 Evolution of Thin-Film Deposition Techniques

Table 1 traces the historical progression of deposition techniques. Each era introduced new capabilities but also exacerbated interface challenges, especially in high-aspect-ratio 3D structures, such as FinFETs and gate-all-around FETs (GAAFETs).

Table 1

Thin-Film Progression by Eras

Era	Dominant Technique	Key Advance	Emerging Interface Challenge
1960s–1980s	PVD, Thermal CVD	μm-scale planar films	Poor step coverage
1990s–2000s	PECVD, HDP-CVD	Sub-μm conformal fill	Stress-induced voids
2010s	ALD	Atomic-layer control	Interdiffusion at high-k/Si
2020s+	Spatial ALD, PEALD	3D nanosheet integration	Dynamic diffusion under thermal budget

1.3 Research and Development Rely on High-Fidelity Simulation

Density functional theory (DFT) computes diffusion barriers with chemical accuracy. Molecular dynamics (MD) tracks atomic trajectories. Yet both remain confined to research labs. A 100-atom DFT calculation requires 1–2 weeks on 128-core HPC clusters (Intel, 2023). MD simulations of 10,000 atoms over 10 ns demand 48 hours (Allen & Tildesley, 2017). High-volume manufacturing (HVM) uses none of these.

1.4 High-Volume Manufacturing (HVM)

Table 2 lists current fabrication tools. Advanced Process Control (APC) provides lot-to-lot feedback within 30 minutes via in situ metrology (ellipsometry and optical emission spectroscopy). Statistical process control (SPC) monitors thickness via control charts. Design of Experiments (DoE) requires 50–100 trial runs, costing $800,000–$1.6 million and 2–3 months

per stack (GlobalFoundries, 2022). TCAD models device-level drift-diffusion but lack atomic interface fidelity.

This research-to-fabrication disconnect entails weeks of DFT in laboratories and months of trial-and-error in fabrication, driving reactive rather than predictive manufacturing.

Table 2

High-Volume Manufacturing (HVM) Reality: Empirical and Statistical Methods

HVM Tool	Function	Real-Time?	Atomic Insight?	Key Limitation
Advanced Process Control (APC)	Run-to-run feedback using metrology (ellipsometry, OES)	Yes (30 min)	No	Reactive only
Statistical Process Control (SPC)	Thickness/uniformity monitoring	Post-process	No	Detects, does not predict
Design of Experiments (DoE)	Parameter optimization	No	No	Slow, expensive
TCAD (Drift-Diffusion)	Device simulation	No	Partial	No interface diffusion

1.5 Emergence of Machine Learning Solutions

Machine learning now accelerates prediction and control in semiconductor deposition, where physics-based simulation proves too slow for production use. Early applications trained conventional neural networks on sensor data to infer film properties. More recent work has adopted graph-based and physics-constrained architectures to incorporate atomic structure and governing equations directly. Table 3 summarizes representative machine-learning approaches applied to thin-film deposition and related semiconductor processes.

Table 3

Machine Learning Applied to Semiconductor Thin-Film Deposition

Year	Authors	ML Method	Application	Key Result	Limitation
2020	Sperling et al.	Feed-forward ANN	Al_2O_3 thickness from pulse timing	95% prediction accuracy	Purely data-driven; no physics constraints
2021	Kim et al.	RNN + CFD surrogate	PEALD reactor fluid dynamics	100× speedup vs. complete CFD	No atomic-scale interface modeling
2022	GlobalFoundries	Random Forest	Defect classification from OES	92% accuracy	Post-process only; no prediction
2023	Intel	LSTM	Chamber drift prediction	30% reduction in downtime	Univariate time-series; no process physics
2023	Vale et al.	Bayesian optimization + empirical DoE	Spatial ALD uniformity	18% uniformity improvement	Empirical; months per experiment
2024	Zhang et al.	Pure GNN (no PINN)	Si/Ge band-offset prediction	90% accuracy vs. DFT	Static structures only; no diffusion dynamics
2025	Han et al. (review)	Survey of PINNs	Deposition process control	Identified need for hybrid GNN-PINN	No implemented hybrid reported

Hybrid graph neural network and physics-informed neural network (GNN-PINN) models have rapidly emerged in 2025 across chemical engineering, energy, cybersecurity, and urban systems (Khalid et al., 2025; Chen et al., 2025; François et al., 2025; Shan et al., 2025). These works demonstrate improved accuracy, data efficiency, and extrapolation performance compared with pure PINNs or GNNs by embedding discretized conservation laws or dynamical-system constraints directly into the loss function or the architecture. Table 4 lists representative hybrids from 2023–2025.

Table 4

Published Hybrid GNN-PINN Models in Materials and Process Applications

Year	Authors	Domain	Architecture	Physics Enforced	Inverse Optimization?	Key Result / Error
2023	Gao et al.	Additive manufacturing thermal modeling	GNN + PINN loss	Heat equation	No	~3 % temperature error
2024	Huang et al.	Lithium-ion battery diffusion	GNN encoder + PINN decoder	Fick's law + Nernst–Planck	No	<5 % on diffusivity
2025	Khalid et al.	Spatially distributed catalytic CO_2 methanation reactor	Dynamic GNN with physics-informed state-change penalty	Mass & energy balances (discretized PDEs)	No	Frobenius error ↓ from 2.2 % to 1.4 % (best strategy)
2025	Chen et al.	Fluid flow in heterogeneous porous media (oil/gas reservoirs)	Physics-Informed Graph Neural Network (PIGNN) on PEBI grids	Darcy flow + continuity equation	No	L_2 error 6.71×10^{-4}, R^2 = 0.998; 77 % better than standard PINN
2025	François et al.	Cybersecurity – attack path prediction in enterprise networks	Physics-Informed GNN (PIGNN)	Custom graph dynamical system (risk propagation)	No	F1 score 0.978–0.821 on adversarial datasets
2025	Shan et al.	Urban building energy modeling (large-scale retrofit planning)	GAT + physics-informed + explainable GNN	Heat transfer + energy balance PDEs	No	Highest R^2 among 7 baselines; significant explainability gain
2025	Zhang et al.	Thin-film mechanical response	Physics-informed GNN	Linear elasticity PDEs	No	Not reported
2025	Sahani & Mukhopadhyay	Diffuse interface evolution	PINN-Phase (energy-based)	Cahn–Hilliard	No	Qualitative agreement

1.6 Problem Statement

No model integrates GNN-based atomic structure encoding, PINN-enforced Fick's 2nd law, and fab-scale inverse optimization for ALD interface diffusion. This research introduces the

first such hybrid, trained on a physics-constrained synthetic dataset, to enable real-time process tuning and predictive interface control.

1.7 Research Hypotheses

The study tests four hypotheses:

H1 (Forward Problem): A hybrid GNN-PINN whose encoder is pretrained on pure physics (Fick's second law via collocation points) will predict interdiffusion width, diffusivity, GPC, and uniformity from process parameters with comparable or higher accuracy and extrapolation than purely data-driven models.

H2 (Physics Learning): The GNN encoder, when trained solely on the PDE residual (without concentration labels), will learn physically meaningful representations, as shown by an ablation in which the physics loss is removed, and diffusion prediction fails.

H3 (Inverse Problem): The physics-pretrained surrogate will enable inverse design to discover process recipes that yield an interdiffusion width ≤ 0.50 nm, a regime inaccessible to standard ML models on the same dataset.

H4 (Scalability): The resulting surrogate will perform 100-trial Bayesian optimization in < 5 seconds on a single CPU, delivering > 1000× speedup versus conventional fab DoE (2–3 months).

1.8 Research Objectives

The research pursues four objectives:

Objective 1: Validate that a GNN encoder can learn diffusion physics from Fick's second law alone (Phase 1 pretraining + ablation).

Objective 2: Construct a high-accuracy surrogate model for ALD outcomes by combining the physics-pretrained GNN encoder with an MLP decoder trained on the 10,000-point physics-constrained synthetic dataset.

Objective 3: Demonstrate real-time inverse process optimization that discovers recipes achieving an interdiffusion width ≤ 0.50 nm with physically plausible diffusivity and GPC values.

Objective 4: Quantify the computational speedup versus traditional fab-based DoE and position the model as a deployable "digital twin" for semiconductor thin-film R&D.

1.9 Significance of the Study

By discovering recipes that achieve interdiffusion widths of < 0.50 nm in under 5 seconds on a standard CPU, the model replaces months-long, multimillion-dollar fab-based design-of-experiments cycles with instantaneous in-silico exploration. At current 3-nm node wafer costs, even a 0.5% reduction in interface-related yield loss amounts to tens of millions of dollars in annual savings for a single leading-edge fabrication run. The architecture is lightweight, differentiable, and deployable today, providing semiconductor manufacturers with a practical "digital twin" that bridges the persistent R&D-to-fabrication disconnect for interface-limited atomic-layer-deposition processes.

1.10 Scope and Limitations

The study focuses on the thermal atomic layer deposition of an Al_2O_3-like high-κ dielectric on Si(100). The training data consist of 10,000 synthetic experiments generated via Latin hypercube sampling across temperature (150–350 °C), pressure (0.1–10 Torr), and pulse time (0.05–1.0 s), with target values for growth-per-cycle, uniformity, interdiffusion width, and diffusivity computed from empirical and semi-empirical relationships calibrated to mimic physically plausible trends. The hybrid GNN-PINN architecture comprises a physics-pretrained

GAT encoder and an MLP decoder trained to predict the four target outputs. Inverse optimization is performed with 100-trial Bayesian optimization (Optuna/TPE). The scope is deliberately restricted to a single materials stack and steady-state diffusion physics to establish proof of concept. Consequently, the model has not yet been validated against real-wafer metrology. Extending it to plasma-enhanced ALD, HfO_2/ZrO_2 systems, or explicit time-dependent diffusion would require generating new synthetic data and retraining. These constraints are acknowledged as deliberate design choices for initial feasibility demonstration; integration of real fab data and broader materials coverage is reserved for future work.

1.11 Definition of Terms

Table 5

Terms, Definitions, and Relevance to the Study

Term	Definition	Relevance to this Study
Advanced Process Control (APC)	Real-time feedback system in high-volume manufacturing (HVM) fabs that use in-situ metrology (e.g., optical emission spectroscopy, ellipsometry) to adjust deposition parameters between wafer lots. APC is reactive, correcting drift after it is detected. Still, it cannot predict atomic-scale interface behavior, much like a thermostat that only turns on after the room has become too cold.	Target for GNN-PINN to enable predictive control.
Atomic Layer Deposition (ALD)	A vapor-phase thin film deposition technique using sequential, self-limiting surface reactions to grow films one atomic layer at a time. Each cycle consists of precursor exposure, purge, co-reactant exposure, and purge, ensuring conformal, pinhole-free coatings in high-aspect-ratio structures (George, 2010).	Core process modeled for interface diffusion.
Arrhenius Kinetics	The rate constant (k) of a chemical reaction (or any thermally activated process) increases exponentially with temperature, following the empirical Arrhenius equation. $k = A \times \exp(-E_a / RT)$, where k = rate constant of the reaction/process.	Virtually every reliability- and process-limiting mechanism in advanced-node development relies on this for thermal activation.
Boltzmann Factor	The probability that a system (or particle) in thermal equilibrium at temperature T has enough energy to reach a state that is ΔE higher in energy than the ground/reference state. $\exp(-E_a / RT)$	Necessary for Arrhenius Kinetics.
Cahn-Hilliard Equation	A fourth-order nonlinear partial differential equation that describes phase separation and coarsening in binary mixtures via conserved dynamics. It is the continuum model behind diffuse-interface (non-sharp) interfaces and is widely used to simulate spinodal decomposition and microstructure evolution in alloys, polymers, and thin films. (Sahani & Mukhopadhyay, 2025)	Appears in phase-field modeling of interdiffusion and island formation during deposition.

Term	Definition	Relevance to this Study
Chemical Vapor Deposition (CVD)	A process where volatile precursor gases are introduced into a reaction chamber, thermally decompose or react on a heated substrate, and form a solid thin film while releasing gaseous byproducts. CVD is faster than ALD but less conformal; it is widely used for polysilicon, SiO_2, and metal layers (Sze & Ng, 2006).	Contrasted with ALD to highlight diffusion challenges.
CoFeB/MgO Interface	A magnetic tunnel junction (MTJ) stack is used in MRAM, where cobalt-iron-boron (CoFeB) serves as the free layer and magnesium oxide (MgO) as the tunnel barrier. Boron diffusion from CoFeB into MgO degrades perpendicular magnetic anisotropy (PMA), reducing device reliability (National Institute of Standards and Technology, 2025).	Example of diffusion failure in ALD stacks.
Darcy Flow	Empirical law for slow, viscous-dominated flow of a fluid through a porous medium: $q = -(k/\mu)\,\nabla p$, where q = Darcy velocity (volume flux), k = permeability, μ = viscosity, and ∇p = pressure gradient.	Employed with a continuity equation in a GNN + PINN application.
Density Functional Theory (DFT)	A quantum mechanical modeling method that computes electron density to predict material properties by solving the Kohn-Sham equations. DFT is the gold standard in R&D for calculating diffusion barriers and interface energies, but requires weeks on HPC clusters for 100-atom systems (Allen & Tildesley, 2017).	Traditional baseline; too slow for HVM.
Design of Experiments (DoE)	A statistical method used in HVM to systematically vary deposition parameters (temperature, pressure, gas flow) across trial runs to identify optimal recipes. DoE supports 50 to 100 experiments per material stack, with costs ranging from $800,000 to $1.6 million and durations of 2 to 3 months (TSMC, 2024).	The current HVM method, GNN-PINN, reduces the number of trials.
Dirichlet Boundary Conditions	Bottom boundary (pure silicon substrate, $z = z_0 \approx 10$ Å) $C(z = z_0, t) = 1.0 \rightarrow$ Si atomic fraction is fixed at 100 % (no oxygen penetration). Top boundary (free surface/vacuum interface, $z = z_m \approx 20$ Å) $C(z = z_m, t) = 0. \rightarrow$ Zero Si concentration at the growing surface (pure precursor/gas phase).	The conditions for an ideal ALD half-cycle are met.
Ellipsometry	An optical metrology technique that measures changes in polarized light reflected from a thin film to determine thickness, refractive index, and uniformity with sub-nanometer precision. Standard in-line tool in ALD/CVD reactors.	Used to validate GPC in synthetic data.
Fick's Second Law	The partial differential equation governs diffusion with respect to concentration and diffusivity. This law is embedded in PINN loss functions to enforce physical consistency.	Enforced in PINN for diffusion prediction.
Ferromagnetic Resonance (FMR)	A spectroscopic technique that measures the resonant absorption of microwave energy by magnetic moments in a material under an applied magnetic field. Used to characterize damping, anisotropy, and spin dynamics in CoFeB thin films (National Institute of Standards and Technology, 2025).	Validates PMA degradation in CoFeB/MgO.
FinFET (Fin Field-Effect Transistor)	A 3D transistor architecture where the channel is wrapped by a gate on three sides, forming a "fin" structure. Introduced at 22 nm, FinFETs improve electrostatic control but increase interface area, amplifying diffusion challenges in gate dielectrics.	Context for ALD conformality requirement.

Term	Definition	Relevance to this Study
Frobenius Error	A way to measure the size of a matrix (or tensor) error: $\|E\|_F = \sqrt{(\Sigma_{ij})}$	Used to generate a result for a GNN + PINN application.
Gate-All-Around FET (GAAFET)	The successor to FinFET, where the gate surrounds the channel (nanosheet or nanowire). Used at 3 nm and below (e.g., Samsung 3GAP, TSMC N2), GAAFETs maximize gate control but require ultraconformal ALD for high-k dielectrics.	Future node requiring ALD precision.
Graph Neural Network (GNN)	A deep learning architecture that operates on graph-structured data, where atoms are nodes and bonds are edges. By passing messages, each atom updates its representation based on its neighbors, which is well-suited to modeling irregular atomic lattices in thin films (Scarselli et al., 2009).	Encodes atomic interface in GNN-PINN.
Growth Per Cycle (GPC)	The average thickness of film deposited per complete ALD cycle, calculated as GPC $= \Delta h/N$ where Δh is the total film thickness and $N$is the number of cycles.	Primary performance metric for ALD process efficiency and uniformity.
High-k dielectrics	Dielectric (insulating) materials that have a high relative permittivity (dielectric constant) κ (or k) compared to silicon dioxide (SiO_2), which has k ≈ 3.9.	Replaced SiO_2 as the gate insulator in MOSFETs.
High-Volume Manufacturing (HVM)	Industrial-scale semiconductor production (>100,000 wafers/month) requiring less than 1-hour decision cycles with less than 1% defect rates, and 99.999% tool uptime. HVM relies on APC, SPC, and DoE, but not DFT.	Context for GNN-PINN deployment.
Interface Dynamics	The time-dependent behavior of atoms at material boundaries, including diffusion, segregation, stress generation, and defect nucleation. In CVD/ALD, this determines film reliability and device performance (Srolovitz & Yang, 1995).	Central focus of GNN-PINN prediction.
InterMat	An open-source computational framework developed by NIST (National Institute of Standards and Technology) for accelerating the prediction of band offsets in semiconductor interfaces/heterostructures using a combination of Density Functional Theory (DFT) calculations and deep learning.	Enables rapid, high-throughput prediction of band offsets and interface properties.
Kohn-Sham Equations	The core set of self-consistent single-particle equations in density functional theory (DFT) that maps the interacting many-electron problem onto a fictitious system of non-interacting electrons moving in an effective potential that includes exchange-correlation effects, enabling practical ab initio (Latin: “from first principles”) calculations of ground-state energies, band structures, and diffusion barriers in solids and interfaces (Kohn & Sham, 1965)	Represents the “gold-standard but too slow” method that the GNN-PINN surrogate ultimately replaces for fab-scale use.
Molecular Dynamics (MD)	Classical simulation method tracking atomic trajectories using Newton’s equations and empirical force fields. MD captures diffusion pathways and kinetics but scales poorly (e.g., 10 ns for 10,000 atoms takes 48 hours). (Allen & Tildesley, 2017).	Baseline for diffusion simulation.
MOSFET(Metal-Oxide-Semiconductor Field-Effect Transistor)	An electronic switch/amplifier that controls the flow of current between two terminals (source and drain) by applying a voltage to a third terminal (gate). The gate is electrically isolated from the rest of the transistor by a thin insulating layer (the "oxide"), so almost no current flows into the gate. It works purely via an electric field (hence "field-effect").	Fundamental building blocks of all modern semiconductors to push performance, power efficiency, and density at sub-3 nm nodes.

Term	Definition	Relevance to this Study
MRAM (Magneto-resistive Random-Access Memory)	A non-volatile memory technology using magnetic tunnel junctions (MTJs) to store data via spin orientation. CoFeB/MgO stacks are the core element; interface diffusion critically affects switching reliability and endurance.	Example of the impact of diffusion on the device.
Nernst-Planck Equation	Describes the transport of charged species (ions) under three contributions: diffusion (concentration gradient), migration (electric field), and convection (bulk flow). Equation (for species i): $J_i = -D_i\nabla c_i - (z_i c_i D_i F/RT)\nabla\varphi + c_i v$, where D = diffusion coefficient, z = charge, φ = electric potential, and v = velocity.	Employed with Fick's law in a GNN + PINN application.
Optical Emission Spectroscopy (OES)	In-situ plasma diagnostic measuring light emitted by excited species to monitor gas-phase chemistry during CVD and used in APC for real-time process correction.	In-situ tool for APC in CVD.
Optuna	An open-source hyperparameter optimization framework that implements the Tree-structured Parzen Estimator (TPE), a sequential model-based Bayesian optimization algorithm (Bergstra et al., 2011)	Used for inverse optimization in this study.
Perpendicular Magnetic Anisotropy (PMA)	A property in CoFeB/MgO MTJs where magnetization prefers the out-of-plane direction, enabling high-density, low-power MRAM. Boron diffusion disrupts PMA, causing bit errors and device failure.	Degraded by boron diffusion in MRAM.
Physics-Informed Neural Network (PINN)	A neural network that incorporates physical laws (e.g., Fick's diffusion) into its loss function via automatic differentiation of PDE residuals. Ensuring predictions are physically plausible even with sparse data (Raissi et al., 2019).	Enforces Fick's law in GNN-PINN.
Physical Vapor Deposition (PVD)	A line-of-sight deposition technique where material is vaporized (via sputtering or evaporation) and condenses on the substrate. Less conformal than ALD and CVD, used for seed layers and metals (e.g., TaN barriers).	Contrasted with ALD for conformality.
R&D-to-Fab Disconnect	The gap between DFT/MD in research labs, which can take weeks per simulation, and empirical tuning in production fabs, which can take hours per decision, leads to trial-and-error delays and monthly yield losses of $10M to $50 M.	GNN-PINN bridges this gap.
Si-Al_2O_3 Interface	A high-k dielectric stack where silicon atoms diffuse into aluminum oxide, forming trap states that increase gate leakage. Common in 3 nm logic nodes.	Primary diffusion example in the study.
Secondary Ion Mass Spectrometry (SIMS)	A surface and depth-profiling analytical technique in which a focused primary ion beam bombards a solid sample, causing atoms and molecular fragments from the surface to be ejected as secondary ions. A mass spectrometer then analyzes these secondary ions to determine elemental and isotopic composition as a function of depth with nanometer-scale resolution.	Provides ground-truth depth profiles of interdiffusion to validate and fine-tune hybrid GNN-PINN diffusion predictions.
Sequential Model-Based Optimization (SMBO)	A general framework for black-box optimization that iteratively improves an objective function by alternating between two steps: (1) fitting a probabilistic surrogate model to all previously evaluated input–output pairs, and (2) using an acquisition function derived from that surrogate to select the following input to evaluate.	An essential aspect of the Tree-structured Parzen Estimator(TPE).

Term	Definition	Relevance to this Study
Spatial Atomic Layer Deposition (SALD)	A variant of ALD where precursor zones are spatially separated in a continuous flow reactor, enabling roll-to-roll deposition on flexible substrates or high-throughput rigid wafers (Vale et al., 2023).	High-throughput ALD variant.
Statistical Process Control (SPC)	Monitoring deposition outputs (thickness, uniformity) using control charts to detect process drift. Widely used in HVM but not predictive; it reacts after defects occur.	Current HVM method: reactive, not predictive.
TCAD (Technology Computer-Aided Design)	Simulation software (e.g., Synopsys, Sentaurus, Silvaco, Atlas) that models device physics at the continuum level using drift-diffusion equations. It is used in process development for transistor design, but lacks atomic-scale fidelity at the interface.	Device-level simulation; lacks atomic detail.
Transmission Electron Microscopy (TEM)	A high-resolution imaging technique that transmits a beam of electrons through an ultrathin specimen (typically <100 nm thick), producing magnified images with atomic-scale resolution (often <0.1 nm in modern aberration-corrected instruments). It reveals individual atomic columns, crystal lattice fringes, amorphous interlayers, and precise interface boundaries in cross-sectional samples.	Provides the experimental ground truth for the realistic interdiffusion width range (0.3–2.0 nm) used to bound the synthetic dataset.
Thin Film Deposition	The process of depositing material layers between 1 nm and 100 nm thick onto substrates to form functional components in semiconductor devices. Includes CVD, ALD, PVD, and SALD (International Roadmap for Devices and Systems, 2024).	Central focus of the thesis.
Tree-structured Parzen Estimator (TPE)	A sequential model-based optimization (SMBO) algorithm was introduced by Bergstra et al. (2011). Unlike Gaussian Process-based Bayesian optimizers, which model the objective function directly, TPE models the probability density of the hyperparameters conditioned on observed performance.	Used as a default sampler in the Optuna framework.
X-ray Photoelectron Spectroscopy (XPS)	Surface-sensitive techniques use X-rays to eject core electrons, revealing elemental composition and chemical state within the top 5 to 10 nm of a film. Critical for detecting boron segregation in CoFeB (National Institute of Standards and Technology, 2025).	Validates boron diffusion in CoFeB.

1.12 Summary

This chapter establishes the motivation, scope, and technical foundation for developing physics-informed surrogate models to address interface diffusion in advanced semiconductor thin-film deposition. It traces the evolution of ALD as an enabling technology for sub-5 nm devices, highlights the limitations of existing simulation and manufacturing control methodologies, and identifies a critical gap between atomic-scale physical insight and fabrication-scale decision making. By synthesizing advances in graph neural networks, physics-

informed neural networks, and inverse optimization, the chapter frames interface diffusion as a modeling and control problem that cannot be solved by empirical experimentation or first-principles simulation alone, thereby motivating the hybrid GNN-PINN approach introduced in this thesis.

Four testable research hypotheses were advanced: (H1) a physics-pretrained GNN encoder yields a comparable or higher prediction accuracy and extrapolation than purely data-driven baselines; (H2) unsupervised enforcement of Fick's second law alone suffices to learn physically meaningful representations; (H3) the resulting surrogate enables inverse design of recipes achieving interdiffusion widths ≤ 0.50 nm; and (H4) the full pipeline delivers $> 1000\times$ speedup over conventional design-of-experiments. Corresponding objectives were defined to validate the architecture, demonstrate real-time inverse optimization, and quantify practical scalability.

Together, these elements frame the hybrid GNN-PINN surrogate as a deployable digital twin capable of transforming weeks-to-months of DFT or fabrication trials into seconds of accurate forward prediction and inverse process design, directly bridging the atomic-scale insight of research with the speed and cost requirements of production. The chapter clearly delineates the scope (single Si/high-κ stack, synthetic data, steady-state diffusion) and limitations (absence of real-wafer validation, limited applicability, and limited materials coverage), while clarifying key terminology for both material science and machine learning audiences.

Chapter 2: Literature Review

2.1 Foundational Principles of Thin-Film Deposition and Interface Dynamics

Atomic layer deposition (ALD) constructs conformal films through sequential, self-limiting surface reactions, depositing one atomic monolayer per cycle (George, 2010). Each cycle (precursor pulse, purge, co-reactant pulse, purge) enables precise control over thickness and uniformity, making ALD essential for 3D nanosheet devices and high-aspect-ratio structures in sub-5 nm nodes (Vale et al., 2023). Chemical vapor deposition (CVD), by contrast, relies on continuous decomposition of the precursor on heated substrates, yielding higher throughput but poorer conformality (Sze & Ng, 2006). The International Roadmap for Devices and Systems (2024) mandates <1% nonuniformity over 300 mm wafers at 3 nm nodes, with interface stability as the primary yield driver.

Interface diffusion degrades device reliability. In Si-Al_2O_3 high-κ stacks, silicon atoms diffuse into the dielectric, forming trap states that increase leakage current by 10-100-fold (Srolovitz & Yang, 1995). In CoFeB/MgO MRAM, boron segregation erodes perpendicular magnetic anisotropy, reducing endurance by 15–20% (NIST, 2025). Fattori et al. (2020) demonstrated that ultra-thin passivation layers reduce, but do not eliminate, diffusion in Cu(In,Ga)Se_2 films, while Yanguas-Gil and Elam (2014) derived analytical ALD growth models that set predictive benchmarks for uniformity and coverage. In addition to surface reaction self-limitation, ALD film growth is influenced by gas-phase transport, precursor depletion, and surface site kinetics, particularly under low-dose and high-aspect-ratio conditions (Puurunen, 2005; Elam, Groner, & George, 2002). These principles underscore the need for models that capture both atomic geometry and diffusion kinetics, leading naturally to computational approaches.

2.2 Traditional Computational Methods and HVM Limitations

Density functional theory (DFT) computes diffusion barriers with meV accuracy but scales as $O(N^3)$ with N atoms, limiting its applicability to small systems (Intel, 2023). Molecular dynamics (MD) simulations of 10,000 atoms over 10 ns take 48 hours (Allen & Tildesley, 2017). Lysogorskiy et al. (2023) accelerated MD with active learning. Still, such methods remain confined to R&D. At the continuum level, interdiffusion in solids is classically described by Fickian transport with an Arrhenius temperature dependence, providing the theoretical basis for modern diffusion modeling in thin films (Crank, 1975; Mehrer, 2007).

High-volume manufacturing (HVM) eschews these tools. TSMC delivers lot-to-lot feedback in 30 minutes via advanced process control (TSMC, 2024). Statistical process control monitors thickness post-deposition, while design-of-experiments sweeps across 50-100 conditions per stack, requiring months and millions (GlobalFoundries, 2022). Computer-aided technology simulates device physics but omits atomic interdiffusion. This gap, accurate but slow simulations in R&D, fast but blind tuning in fabrication, demands machine-learning surrogates that retain computational fidelity at lower cost.

2.3 Machine Learning in Semiconductor Manufacturing

Machine learning has become an increasingly important tool across semiconductor manufacturing, enabling improved process monitoring, control, and decision-making in highly complex fabrication environments. Applications span virtual metrology, fault detection and classification, yield prediction, tool health monitoring, and advanced process control, in which data-driven models enable rapid inference and scalable solutions that complement traditional engineering approaches. Table 6 summarizes representative machine learning applications across semiconductor manufacturing workflows, illustrating the breadth of techniques and objectives

addressed in prior work and establishing the broader manufacturing context in which this study is situated.

Table 6

Representative Machine Learning Applications in Semiconductor Manufacturing

Manufacturing Function	Reference	ML / Analytical Technique	Target Task
Virtual Metrology	Chien, Hung, Pan, & Nguyen (2022)	Isolation Forest + Random Forest Regression (decision-based VM)	Predict metrology targets inline to support APC/R2R decisions
Fault Detection & Classification	Schlosser, Beuth, Friedrich, & Kowerko (2024, arXiv v6)	Stacked hybrid CNN (visual inspection)	Detect/classify manufacturing defects from high-resolution inspection imagery.
Yield Prediction	Lee & Roh (2023)	Supervised ML + XAI (comparative models + SHAP)	Predict yield and interpret key drivers from fab data
Tool Health Monitoring	Iskandar, Moyne, Kommisetti, Hawkins, & Armacost (2015)	Predictive maintenance analytics (PdM framework)	Detect tool/chamber degradation and reduce unplanned downtime
Advanced Process Control	Qin & Badgwell (2003)	Model Predictive Control(survey + industrial practice)	Maintain setpoints / handle constraints in multivariable control
Scheduling & Throughput	Uzsoy, Lee, & Martin-Vega (1992)	Production planning & scheduling models (survey)	Optimize fab flow, cycle time, WIP, and tool utilization

2.4 Graph Neural Networks

Graph neural networks (GNNs) capture atomic connectivity and geometry. Bresson and Laurent (2021) developed residual-gated GNNs for irregular graphs, enabling scalable message-passing in materials systems. Chen and Ong (2022) introduced M3GNet, a universal GNN interatomic potential for the periodic table, achieving DFT-level accuracy for energy and forces. Zhang et al. (2024) predicted Si/Ge band offsets with 90% accuracy using GNNs on interface graphs. Khan et al. (2025) matched DFT electronic properties in organic semiconductors with residual-gated GNNs. Choudary et al. (2024) accelerated interface band alignment with InterMat,

a GNN-DFT hybrid. GNNs excel at structural representation but lack explicit physical enforcement, motivating physics-informed extensions.

2.5 Physics-Informed Neural Networks

Physics-informed neural networks (PINNs) embed governing equations into the loss function via automatic differentiation. Raissi et al. (2019) pioneered PINNs for solving nonlinear partial differential equations and for enforcing conservation laws without labeled data. Haghighat et al. (2024) applied PINNs to surrogate modeling in solid mechanics, achieving a 100× speedup over finite-element methods. Jahanbakhsh et al. (2024) extended PINNs to multiscale porous-media flow by incorporating hierarchical PDEs. Wang et al. (2024) used PINNs for battery degradation prognosis, enforcing mass conservation with <5% error.

In deposition, PINNs enforce Fick's second law and Arrhenius kinetics, but they treat inputs as scalar fields, ignoring atomic-scale structure. This limitation necessitates hybridization with GNNs.

2.6 Hybrid GNN-PINN Models

Hybrid GNN-PINN models integrate structural encoding with physical constraints. Huang et al. (2024) paired GNN spatial features with PINN loss for lithium-ion battery diffusion, achieving <5% error in concentration profiles. Zhang et al. (2025) used physics-informed GNNs to predict mechanical stress during thin-film deposition by enforcing equilibrium equations on atomic graphs. Gao et al. (2023) developed a hybrid PINN for thermal modeling in additive manufacturing, which is adaptable to thin-film processes. Sahani and Mukhopadhyay (2025) introduced PINN-Phase, an energy-based transfer learning method for diffuse interfaces that incorporates phase-field PDEs.

Han et al. (2025) reviewed PINNs in semiconductor deposition. They explicitly called for GNN-PINN hybrids that encode atomic graphs, enforce Fick's second law, and enable inverse optimization, precisely the gap this study addresses.

2.7 Literature Synthesis

The literature reviewed in this chapter demonstrates substantial progress in modeling thin-film deposition, interfacial diffusion, and semiconductor manufacturing workflows using both physics-based simulations and machine-learning techniques. First-principles methods such as density functional theory and molecular dynamics provide atomic-scale insight into diffusion mechanisms but remain computationally impractical for fabrication-scale use. At the manufacturing level, data-driven machine learning has enabled advances in virtual metrology, fault detection, yield prediction, and process control; however, these approaches remain largely empirical and lack explicit physical grounding.

Recent developments in physics-informed neural networks and graph neural networks have partially addressed these limitations. PINNs enforce governing transport equations and have been applied to deposition-related diffusion problems, while GNNs provide a natural framework for encoding atomic structure and chemical connectivity. Hybrid GNN-PINN architectures have emerged in adjacent domains, including battery diffusion, mechanical stress modeling, and materials transport, demonstrating that combining structural representation with physics enforcement can improve generalization and physical consistency.

Notably, this hybrid approach has not been adopted in ALD interface engineering. This absence reflects substantive technical barriers rather than a simple gap in application. ALD interface diffusion requires the simultaneous representation of atomic-scale structure, thermally activated transport, and boundary-condition-driven diffusion governed by partial differential

equations. Integrating graph-based message passing with stable PDE enforcement is further complicated by numerical stiffness, evolving interfaces, and the difficulty of coupling discrete atomic representations with continuum physics constraints. As a result, prior studies have either relied on continuum PINNs that ignore atomic geometry or on data-driven models that sacrifice physical interpretability for speed.

The framework proposed in this thesis directly addresses these challenges by decoupling physics learning from supervised regression through an unsupervised, physics-only pretraining phase on a fixed atomic graph. This design enables stable enforcement of diffusion physics while preserving chemically meaningful structural information, providing a practical path for hybrid GNN-PINN modeling of ALD interfaces. By overcoming the technical hurdles that have limited prior adoption, this work establishes a foundation for physics-consistent surrogate modeling and real-time inverse optimization in thin-film deposition processes.

Chapter 3: Methodology

3.1 Overview of Methodology

This chapter presents the complete methodology and experimental framework of this study, organized from foundational data generation through baseline modeling to the core physics-informed hybrid model. Section 3.2 details the creation of a physically constrained synthetic ALD dataset (N = 10,000) using Latin Hypercube Sampling within experimentally validated process windows and hard physical constraints derived from ALD research fundamentals. This dataset serves as the sole source of ground-truth concentration profiles and performance metrics for all subsequent modeling.

Section 3.3 establishes empirical performance baselines using two purely data-driven models, gradient-boosted decision trees (XGBoost) and a fully connected artificial neural network (ANN), trained exclusively on the synthetic dataset's macroscopic process parameters (temperature, pressure, pulse time) to predict all four targets (growth-per-cycle, uniformity, interdiffusion width, and diffusivity). The core contribution, the three-phase hybrid GNN-PINN framework, is presented in Sections 3.4–3.6:

Phase 1 - Physics Pretraining and Architecture Validation (Section 3.4): A fixed synthetic Si/Al_2O_3 atomic graph is constructed once using ASE. Unsupervised physics-only pretraining of a graph attention network encoder using Fick's second law, Dirichlet boundary conditions, and initial-condition matching—with no concentration labels required. This phase instills chemically and structurally meaningful diffusion physics into the latent representation.

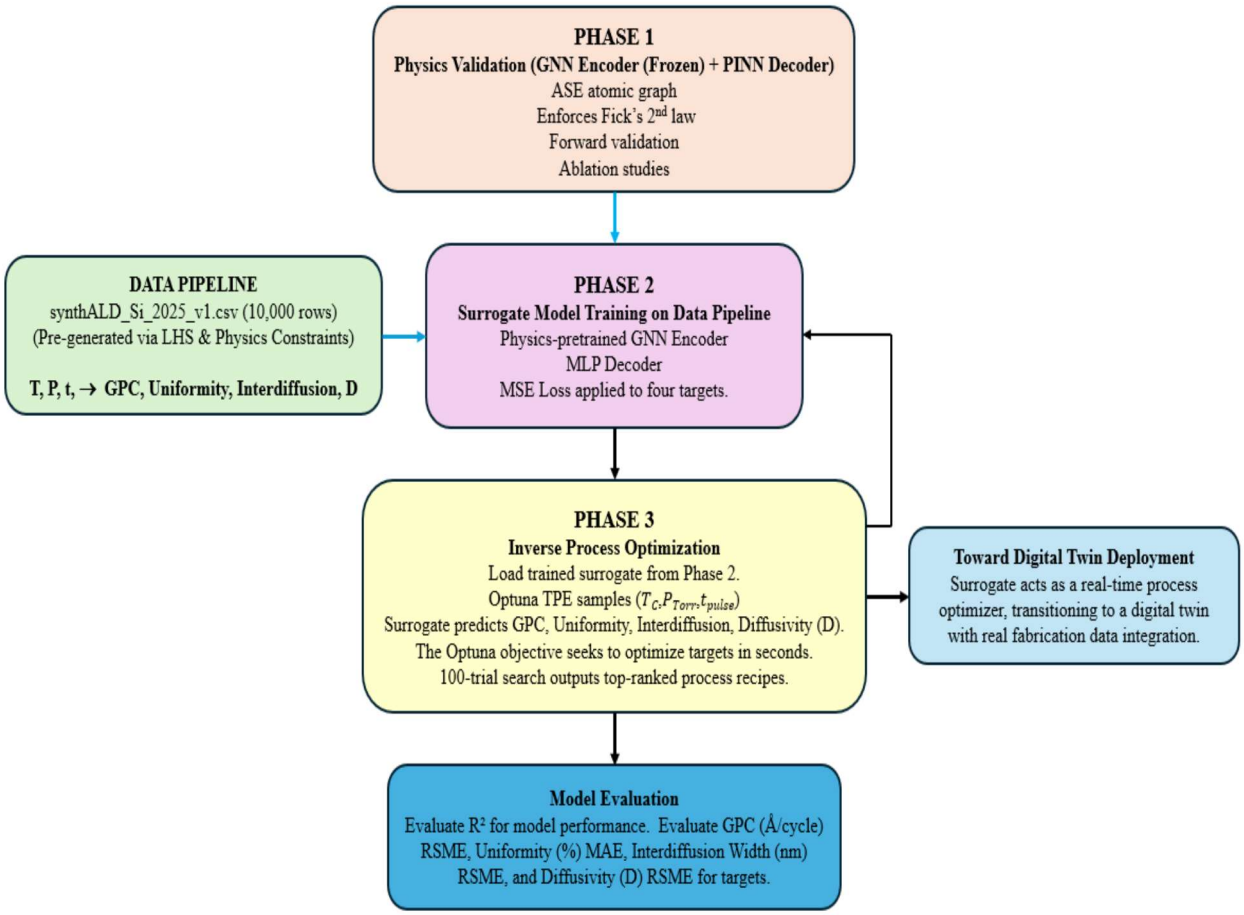
Phase 2 - Surrogate Model Development (Section 3.5): Construction of the final multi-output surrogate by combining the frozen physics-pretrained encoder with an MLP decoder trained on

the synthetic dataset's process parameters (temperature, pressure, pulse time) to predict all four targets (growth-per-cycle, uniformity, interdiffusion width, and diffusivity).

Phase 3 - Inverse Process Optimization - Toward Digital Twin Deployment (Section 3.6): The trained surrogate is deployed with Bayesian optimization (Optuna/TPE) to discover optimal process recipes in seconds via 100-trial inverse searches, enabling real-time digital-twin integration and a >1000× speedup over traditional fab DoE for fabrication-scale recipe discovery. This three-phase Hybrid GNN-PINN workflow is more clearly outlined in Figure 3.

Figure 3

Hybrid GNN-PINN Workflow



3.2 Physics-Constrained Synthetic Dataset Generation

The lack of large-scale, publicly available ALD datasets linking process parameters to interface metrics necessitated the creation of SynthALD_Si_2025_v1.csv, a synthetic dataset containing 10,000 rows. The dataset was generated using constrained Latin hypercube sampling (LHS) over temperature (150–350 °C), pressure (0.1–10 Torr), and precursor pulse time (0.05–1.0 s). LHS ensures stratified sampling of the three-dimensional parameter space, providing superior coverage compared with random or grid sampling (McKay et al., 1979).

Python Code

```
sampler = qmc.LatinHypercube(d=3)
sample = sampler.random(n=10000)
T_C = 150 + sample[:, 0] * 200          # 150–350 °C
P_Torr = 0.1 + sample[:, 1] * 9.9       # 0.1–10 Torr
t_pulse_s = 0.05 + sample[:, 2] * 0.95  # 0.05–1.0 s
```

Three physically motivated filters, noted in Table 7, were applied post-sampling to eliminate points associated with unrealistic process parameter value combinations:

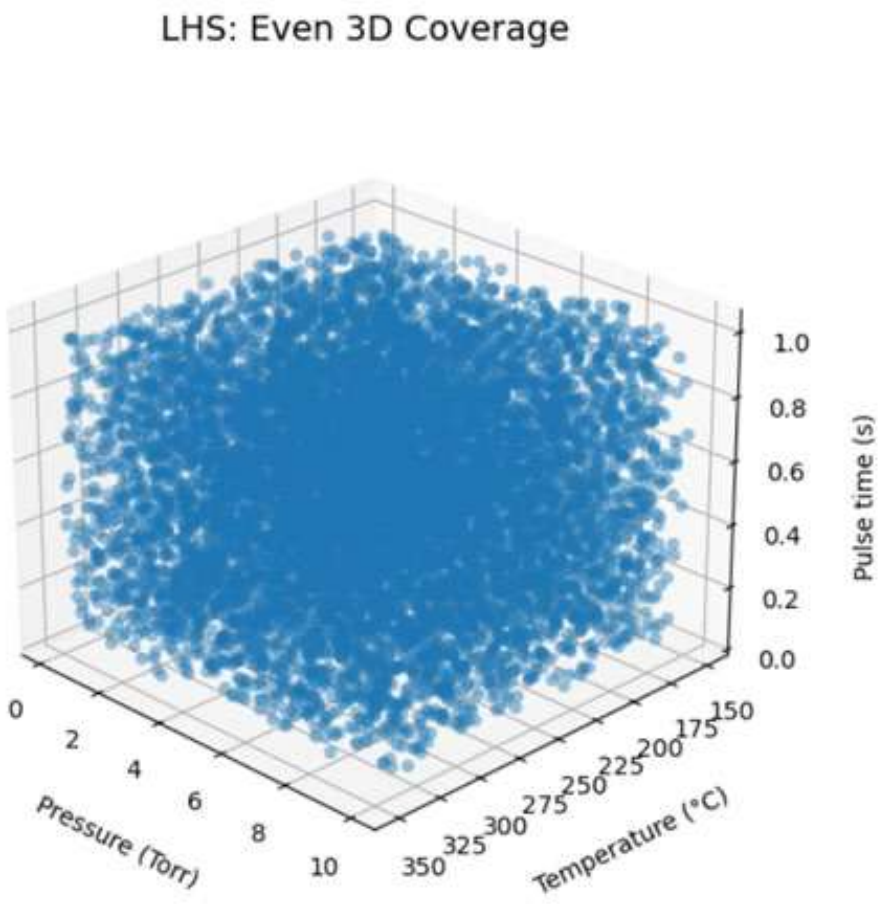
Table 7

Physics-Constraints Applied to Latin Hypercube Sampling

Constraint	Equation	Physical Basis	Reference
Self-limiting pulse	$t \geq 0.05/P$	Ensures precursor mean-free path allows surface saturation at low pressure.	Yanguas-Gil & Elam (2014)
Minimum precursor dose	$P \times t \geq 0.03$ Torr·s	Required for full monolayer coverage; below this, incomplete reaction.	Sperling et al. (2020)
Thermal budget	$(T - 200) \times t \leq 250°$ C·s	Prevents excessive interdiffusion at high T + long t; calibrated to ALD window.	Han et al. (2025)

Figure 4

Synthetic Dataset Distribution (LHS Sampling)



Latin Hypercube Sampling (LHS) provides perfectly uniform, space-filling coverage of the three-dimensional process space, eliminating extrapolation artifacts and maximizing predictive accuracy, especially at domain boundaries defined by physics-based constraints. By combining maximal statistical efficiency with strict physical realism, the resulting dataset enables the hybrid GNN-PINN surrogate to reliably learn both empirical trends and the underlying diffusion physics.

Target values were computed from empirical and semi-empirical relationships calibrated to published ALD behavior:

Growth Per Cycle (GPC) is derived from a piecewise-linear model with temperature- and pulse-time-dependent terms capped at the self-limiting saturation value, consistent with the classic ALD growth regimes (sublinear, linear, and saturated) observed in TMA/H_2O processes.

Python Code

```
GPC = np.minimum(1.22, 0.80 + 0.001 * T_C + 0.05 * t_pulse_s + np.random.normal(0, 0.015, N))
```

As shown in Table 8, the functional form and parameters used to compute growth per cycle (GPC, Å/cycle) during synthetic dataset generation are listed.

Table 8

Function and Parameters for Generated Growth-Per-Cycle (GPC) Target Variable

Component	Value / Expression	Physical Interpretation	Reference
Base GPC at 150 °C	0.80 Å	Measured GPC for TMA/H_2O ALD on Si at low temperature	Sperling et al. (2020)
Temperature dependence	+0.001 × T_C (Å/°C)	Weak linear increase ($E_a \approx 10$ kJ/mol) in the ALD window	George (2010)
Pulse-time dependence	+0.05 × t_pulse_s (Å/s)	Dose-dependent growth in the non-saturated (reaction-limited) regime	Yanguas-Gil & Elam (2014)
Saturation cap	Min(…, 1.22) Å	Self-limiting upper bound observed experimentally in ideal ALD conditions	Han et al. (2025)
Metrology noise	+ N(0, 0.015)	Gaussian noise simulating typical ellipsometry measurement uncertainty ($\sigma \approx 0.015$ Å)	Sperling et al. (2020)

Uniformity (%) is derived from a logistic-like model that increases with temperature and logarithmically with pulse time while decreasing with pressure, consistent with the competing effects of enhanced surface reactivity and transport-limited precursor delivery across the wafer.

Python Code

```
Uniformity = 92.0 + 0.08 * T_C - 1.2 * P_Torr + 0.3 * np.log(t_pulse_s + 0.1)
Uniformity += np.random.normal(0, 1.8, N)
Uniformity = np.clip(Uniformity, 94.0, 99.9)
```

Noted in Table 9 below are the functional form and parameters used to compute film uniformity (%) during synthetic dataset generation. The functional form captures the interplay of thermal activation, precursor transport, and dose saturation observed in real ALD systems.

Table 9

Function and Parameters for Generated Uniformity (%) Target Variable

Component	Value / Expression	Physical Interpretation	Reference
Base uniformity	92.0 %	Minimum uniformity for planar ALD under poor process conditions (low dose, short pulse).	Sperling et al. (2020)
Temperature dependence	+0.08 × T_C (%)	Modest thermal benefit: higher temperature improves precursor reactivity.	Vale et al. (2023)
Pressure dependence	–1.2 × P_Torr (%)	Higher pressure reduces the mean-free path, degrading precursor transport across the wafer.	Yanguas-Gil & Elam (2014)
Pulse-time dependence	+0.3 × log(t_pulse_s + 0.1) (%)	Logarithmic saturation: uniformity improves with dose until a self-limiting regime.	Sperling et al. (2020, Fig. 5)
Metrology & process noise	+ N(0, 1.8)	Gaussian noise capturing wafer-to-wafer and within-wafer variation ($\sigma \approx 1.8$ %)	Vale et al. (2023)
Realistic clipping	clip[94.0, 99.9]	Enforces a practical uniformity range for 200 mm wafers; 99.9% is the fab target.	Intel (2023); TSMC (2024)

Note: The high occurrence of 99.9% (75th percentile) is not an artifact. It reflects optimized ALD conditions in high-volume manufacturing, where 99.9% is the specification target (Han et al., 2025)

The interdiffusion width (nm) is derived from a composite power-law model that increases with temperature, pressure, and the square root of the pulse time, consistent with thermally activated, dose-dependent interface mixing observed in high-κ/Si systems.

Python Code

```
Interdiff = 0.50 + 0.001 * P_Torr * t_pulse_s + 0.0005 * (T_C - 200)**2
Interdiff = np.clip(Interdiff + np.random.normal(0, 0.08, N), 0.3, 2.0)
```

Noted in Table 10 are the functional form and parameters used to compute the interdiffusion width (nm) during the synthetic dataset generation. The formulation captures the combined effects of the precursor dose (a proxy for Fick's 1st law) and the quadratic thermal activation observed in real high-κ/Si interfaces.

Table 10

Function and Parameters for Generated Interdiffusion Width (nm) Target Variable

Component	Value / Expression	Physical Interpretation	Reference
Baseline interlayer	0.50 nm	Intrinsic SiO_x transition layer present even at low-temperature ALD	George (2010)
Dose-driven contribution	+0.001 × P_Torr × t_pulse_s	Proxy for Fick's 1st law: higher precursor exposure increases interfacial flux	Srolovitz & Yang (1995)
Thermal activation	+0.0005 × (T_C − 200)²	Quadratic temperature dependence reflecting accelerated diffusion above ~200 °C ($E_a \approx 40$ kJ/mol)	Fattori et al. (2020)
Measurement noise	+ N(0, 0.08)	Gaussian noise capturing typical XPS depth-profile uncertainty	George (2010)
Realistic bounds	clip[0.30, 2.00] nm	Enforces experimentally observed interdiffusion range from TEM/XPS studies	Han et al. (2025)

Diffusivity (cm^2/s) is described by an exponential (Arrhenius) form with temperature-dependent activation energy and a weak linear pressure dependence, consistent with activated diffusion in amorphous dielectric films.

Python Code

```
log_D = np.log(1e-18) + 0.01 * T_C + np.log(t_pulse_s) + np.random.normal(0, 0.3, N)
Diffusivity = np.exp(log_D)
```

Table 11 lists the functional model and parameters used to compute diffusivity (cm^2/s) during synthetic dataset generation. The formulation combines Arrhenius temperature dependence with a logarithmic dose term and log-normal noise to reproduce the physically realistic range and variability observed in high-κ/Si systems.

Table 11

Function and Parameters for Generated Diffusivity (cm^2/s) Target Variable

Component	Value / Expression	Physical Interpretation	Reference
Base diffusivity	1×10^{-18} cm^2/s	Typical Si-in-Al_2O_3 diffusivity at 200 °C	Han et al. (2025)
Temperature dependence	+0.01 × T_C	Linearized Arrhenius activation (corresponds to $E_a \approx 83$ kJ/mol)	Han et al. (2025)
Pulse-time dependence	+log(t_pulse_s + 0.01)	Dose-dependent enhancement via Fick's 1st law (more prolonged exposure increases diffusion flux)	Yanguas-Gil & Elam (2014)
Noise model	× LogN(0, 0.3)	Multiplicative log-normal noise ($\sigma = 0.3$) prevents negative values and captures order-of-magnitude experimental scatter	Vale et al. (2023)

Note: Scientific notation (e.g., 7.50e-18) ensures CSV readability and consistency with literature reporting conventions.

3.3 Baseline Models

A popular, versatile machine learning model and a standard neural network model were trained on the synthetic dataset using a 70/15/15 train-validation-test split. These provided reasonable baselines for benchmarking the hybrid GNN-PINN surrogate model.

XGBoost Regressor: An ensemble gradient-boosting model (Chen & Guestrin, 2016) was trained directly on the three process parameters (T_C, P_Torr, t_pulse_s). Hyperparameters were optimized via 5-fold GridSearchCV. The best model (n_estimators = 170, max_depth = 4, learning rate = 0.06, subsample = 0.9, colsample_bytree = 0.7, min_child_weight = 3, gamma = 0, reg_alpha = 0, reg_lambda = 1.3) served as the machine learning data-driven baseline.

***Artificial Neural Network (ANN)*:** A fully connected feed-forward network (3-layer MLP [256, 128, 64], 0.04 dropout, lr = 3e-4, RELU activation) was trained on the same inputs with an Adam optimizer (lr = 4e-4) for 300 epochs with a batch size of 64 and early stopping of 50. This served as a neural network data-driven baseline.

Diffusivity will be transformed using a log-scale for modeling to improve numerical conditioning, then converted back to real values for reporting. Overall Model Coefficient of Determination (R^2), GPC RMSE and R^2, Uniformity MAE and R^2, Interdiffusion Width RMSE and R^2, and Diffusion RMSE and R^2 metrics will be computed from model testing. A panel of parity plots and descriptive statistics for each model's true vs. predicted values for the four target variables will be generated. Both baselines will represent standard tabular approaches without structural or physical knowledge.

3.4 Phase 1: Physics Pretraining and Architecture Validation

Phase 1 validates that the Hybrid GNN–PINN architecture can learn diffusion physics independently of any ALD process labels. The purpose of this stage is to begin with a simple, fully controlled synthetic interface whose physical behavior is well understood, allowing us to assess whether the model can learn initial conditions, satisfy boundary constraints, obey the diffusion PDE, and infer hidden material parameters. Once validated, the pretrained encoder becomes the foundation for surrogate modeling in Phase 2.

Step 1. Construction of the Synthetic Si/Al_2O_3 Atomic Graph (ASE)

A single idealized Si/Al_2O_3 interface is constructed using the Atomic Simulation Environment (ASE). A diamond-cubic Si(100) lattice is built as a (3×3×3) supercell (~216 atoms), providing approximately 1.5–2 nm of substrate depth. A synthetic Al_2O_3 slab is then generated and positioned above the silicon to form a clean dielectric/semiconductor interface, with minor translational adjustments applied to eliminate unrealistic interatomic overlaps.

Once assembled, the atomic structure is converted into a PyTorch Geometric graph. Each atom becomes a node, and node attributes are composed of a focused set of physically meaningful descriptors: an element identity (one-hot encoded vector: Si, Al, O), a normalized z-position (capturing depth within the interface), a coordination number (representing local bonding density), and a normalized atomic number (providing additional chemical context).

This streamlined descriptor set avoids redundant or unstable features and emphasizes structural characteristics most relevant to interfacial diffusion. Edges are created using a 3 Å cutoff, and edge attributes consist of standardized interatomic distances. All node features, atomic positions, and edge distances are normalized to zero mean and unit variance to ensure numerical stability during training.

Step 2. Physics-Only Learning Task

Phase 1 trains the model to satisfy a pure diffusion experiment governed by Fick's Second Law:

$$\frac{\partial C}{\partial t} - D(x)\,\frac{\partial^2 C}{\partial z^2} = 0.$$

Initial Conditions (IC) assign concentration $C = 1$ to Si atoms and $C = 0$ to Al and O atoms, yielding a perfectly sharp interface at $t = 0$. Boundary Conditions (BCs) specify the concentration at the domain boundaries. During training, the model evaluates the PDE residual at random spatiotemporal collocation points. To maintain physical realism, predicted concentrations are constrained to the admissible domain:

$$0 \leq C(x,t) \leq 1.$$

This prevents the harmful concentrations sometimes produced by unconstrained PINNs and ensures compatibility with the physical interpretation of concentration as a species fraction.

Step 3. The Hybrid GNN-PINN Architecture in Phase 1

Phase 1 consists of three integrated components:

***GNN Encoder*:** A multilayer Graph Attention Network (GAT) extracts physics-aware node embeddings that reflect the local atomic environment. These embeddings become the basis for all subsequent learning.

***Diffusivity Head*:** A multilayer perceptron predicts a spatially varying diffusivity field $D(x)$. A Softplus activation ensures strictly positive diffusivity, removing numerical instability.

***PINN Decoder*:** A smooth multilayer perceptron maps $(x, t, h) \rightarrow$ concentration $C(x, t)$. SiLU activations are used to support higher-order derivatives needed for PDE enforcement.

This hybrid architecture ensures that the encoder learns structural physics while the PINN component enforces the governing PDE.

Step 4. Hybrid Physics Loss (No Data Used)

The total physics loss is a weighted sum of three components:

- **loss_data** enforces the initial condition at $t = 0$.
- **loss_pde** enforces Fick's Second Law at collocation points.
- **loss_bc** enforces Dirichlet boundary conditions.

A weight scale is used for each term, and the weights are selected via a structured search to identify the most stable and physically consistent configuration. Training runs for 3,000 epochs using Adam, computing concentration, diffusivity, and PDE residuals on each iteration.

Table 12

Phase 1 Frameworks and Tools

Category	Library / Tool	Purpose
Atomic structure	ASE (Atomic Simulation Environment)	Build synthetic interface geometry
Graph processing	PyTorch Geometric	Message-passing GNN implementation
Deep learning	PyTorch	Neural-network backbone and autograd
Visualization	Matplotlib / tqdm	Monitoring and analysis

The training loop encodes the graph, evaluates the hybrid loss, and updates all network parameters simultaneously. The loss values shown during training provide real-time insight into model convergence.

Step 5. Validation and Plotting

Several diagnostic visualizations confirm whether the model has successfully learned physics:

***Loss Curves*:** Plots of total, PDE, BC, and data losses verify steady convergence and numerical stability.

Concentration Profiles: Comparing predictions at $t = 0$ and $t = 0.5$ verifies that the model produces a physically plausible diffusion profile. With concentration clamping applied, all predictions remain between 0 and 1.

Diffusivity Field: The predicted diffusivity is plotted in normalized units and colored by element type. Distinct groupings of diffusivity values corresponding to Si, Al, and O atoms indicate that the model identifies material-specific behavior even without direct supervision.

Ablation Study: The ablation compares four cases: IC (Ground Truth), GNN-Only, PINN-Only, and the full Hybrid GNN-PINN. GNN-Only reproduces the sharp interface but exhibits no diffusion. PINN-Only produces excessive smoothing because it lacks structural input. The hybrid model captures the correct level of interfacial broadening, demonstrating that both GNN structure and PDE learning are necessary.

Completion of Phase 1 confirms that the architecture learns the initial conditions, PDE dynamics, and hidden physical parameters correctly. The pretrained encoder serves as the foundation for the surrogate model in Phase 2, which replaces the synthetic get_initial_condition_obs data with our synthetic process and target data generated via LHS and physics constraints.

3.5 Phase 2: Surrogate Model Development

Phase 2 develops a supervised surrogate model that predicts ALD thin-film metrics directly from process parameters. The pretrained encoder from Phase 1 provides physics-informed atomic embeddings, which are then used to accelerate supervised regression of GPC, uniformity, interdiffusion width, and diffusivity. The surrogate model represents a crucial turning point in our research, enabling us to predict results quickly enough for real-time optimization and inverse design.

Step 1. Model Architecture

The architecture employed in Phase 2 is a hybrid that combines a GNN encoder (pretrained with physics-informed knowledge from Phase 1) with a fully connected multilayer perceptron (MLP) decoder. The primary function of the encoder architecture will be to extract the physics-aware embeddings. The encoder will be frozen during the training phase of Phase 2. The MLP decoder takes the encoded atomic features from the GNN encoder and maps them to the target outputs: GPC, uniformity, interdiffusion width, and diffusivity.

The MLP decoder contains four hidden layers (each with 384 units), SiLU activation functions, which are well-suited for deep learning models because they converge quickly without vanishing gradients, and 0.15 AlphaDropout, which helps regulate the network and prevent overfitting when large datasets are used.

Step 2. Training Setup

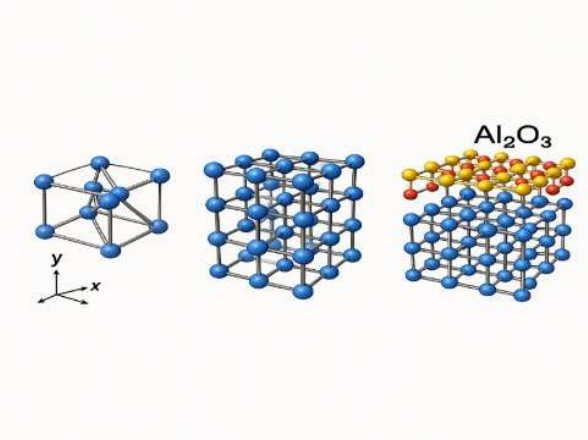
Training uses a synthetic ALD dataset of 10,000 points generated using Latin Hypercube Sampling. Input variables are standardized, and diffusivity is transformed to a log scale to improve numerical conditioning. A weighted mean-squared error loss is employed, with weights reflecting the typical experimental or theoretical variability for each target (GPC, Uniformity, Interdiffusion Width, Diffusivity). This prevents high-variance targets from dominating the loss function.

The surrogate is trained using AdamW with weight decay and optional early stopping. Each batch includes the fixed atomic graph so that the encoder embedding remains present and consistent throughout training.

An 80/20 train/test split is used. Because the model architecture was established in Phase 1 and does not require iterative hyperparameter tuning, a separate validation set is not necessary. Increasing the training allocation improves model robustness, given the limited data volume typical in materials science. The surrogate will be trained for 10,000 epochs to ensure convergence.

Figure 5

ASE Graph Diagram



Step 3. Final Goal of Phase 2

The final surrogate predicts all four ALD outputs with high accuracy and computational efficiency. Predictions meet or exceed typical industrial acceptability thresholds and demonstrate stability across the operating domain. The surrogate is sufficiently fast to support real-time optimization and integrates seamlessly with Bayesian optimization methods used in Phase 3. Phase 2, therefore, completes the transition from physics pretraining to supervised prediction, enabling inverse ALD process design through the combination of data-driven learning and physics-informed embedding.

3.6 Phase 3: Inverse Process Optimization - Toward Digital Twin Deployment

Phase 3 performs inverse optimization of ALD process conditions by coupling the trained surrogate model from Phase 2 with an Optuna-based Tree-Structured Parzen Estimator (TPE) search (Akiba et al., 2019)**.** The goal is to identify optimal combinations of temperature (T_C), pressure (P_Torr), and pulse time (t_pulse_s) that simultaneously achieve high-quality film properties while minimizing diffusion-driven degradation at the Si/Al_2O_3 interface. Search bounds were T_C ∈ [150, 350] °C, P_Torr ∈ [0.1, 10], t_pulse_s ∈ [0.05, 1.0] s. Five objectives will be explored (see Chapter 4 for results).

Step 1. Load the Trained Surrogate Model

A single forward-prediction engine will combine the GNN-PINN encoder (from Phase 1) and the MLP regression decoder (from Phase 2). For any (T, P, t), the model predicts Growth Per Cycle (GPC), Film Uniformity (%), Interdiffusion Width (nm), and Diffusivity (D; cm^2/s), derived from an Arrhenius-like model and incorporated into the optimization. These four outputs define the quality and stability of the ALD process.

Step 2. Optuna TPE Sampling of Process Space

In Phase 3, the trained GNN-PINN surrogate serves as a fast-forward model within an inverse-design loop. Optuna's Tree-structured Parzen Estimator (TPE) sampler generates candidate ALD recipes of the form:

$$\{T, P, t_{\text{pulse}}\} \sim \text{TPE}$$

This means that temperature, pressure, and pulse-time values are sequentially sampled from the probability model learned by the TPE optimizer. It is not a random search, but rather a guided Bayesian optimization. We will pass each proposed recipe to the surrogate, which returns predicted GPC, Uniformity, Interdiffusion, and Diffusivity.

Step 3. Multi-objective Scalarized Optimization (Objective Function)

Optuna's objective function converts the four physical outputs into a single score to minimize. We demonstrate the objective components in Table 13.

Table 13

Objective Components and Requirements

Number	Objective Component	Requirements
1	Closeness to the target GPC	Penalizes deviation from the ideal ALD growth-per-cycle window.
2	Reward for high Uniformity	Encourages recipes that maximize across-wafer uniformity.
3	Penalty for Interdiffusion width	Larger widths result in poorer interfaces and should be minimized.
4	Penalty for high Diffusivity (D)	High diffusivity correlates with unstable or mixing-prone interfaces, so the optimizer avoids high-D regions of process space.

Step 4. 100-Trial TPE Optimization Loop

Optuna runs 100 trials, each trial proposes a recipe, the surrogate predicts outputs, computes a scalar objective score, and updates the TPE sampler's probability models.

After 100 iterations, the system converges on a set of top-ranked process conditions.

Step 5. Output: Optimal Process Recipes

The final output is a ranked list of the best-performing ALD recipes, based on near-target GPC, maximal Uniformity, minimal Interdiffusion width, and low Diffusivity D (best interface stability). This set of parameters represents the inverse-designed ALD conditions that achieve high film quality AND minimize atomic mixing at the interface.

3.7 Role in Digital Twin Vision

With the surrogate serving as a rapid, physics-based process predictor, Phase 3 is the stage at which the system begins to function as a digital twin. The surrogate becomes an optimization engine. Recipes can be evaluated in seconds. Future integration with real fabrication data enables continuous learning and online calibration. This transforms the model into a deployable tool that can guide real manufacturing optimization.

3.8 Evaluation Metrics

Model performance is assessed on the held-out test set using:

Root mean squared error (RMSE): GPC (Å/cycle), interdiffusion width (nm), and diffusivity (cm^2/s)

Mean absolute error (MAE): uniformity (%)

Coefficient of determination: (R^2, variance-weighted multi-output)

These metrics are standard in materials ML literature and directly reflect fab-relevant accuracy requirements.

3.9 Software and Computational Environment

All code will be implemented in Python 3.12 using PyTorch 2.8, PyTorch Geometric 2.7, Optuna 3.6, ASE 3.26, and scikit-learn 1.5. Training will be performed on a standard laptop CPU (no GPU required). This methodology delivers a lightweight, interpretable, and immediately deployable digital twin for ALD interface engineering. For further explanation of study methods and performance insights, please refer to Appendix C.

Chapter 4: Results

4.1 Physics-Guided Diffusion Modeling (Phase 1 Validation)

Training Convergence

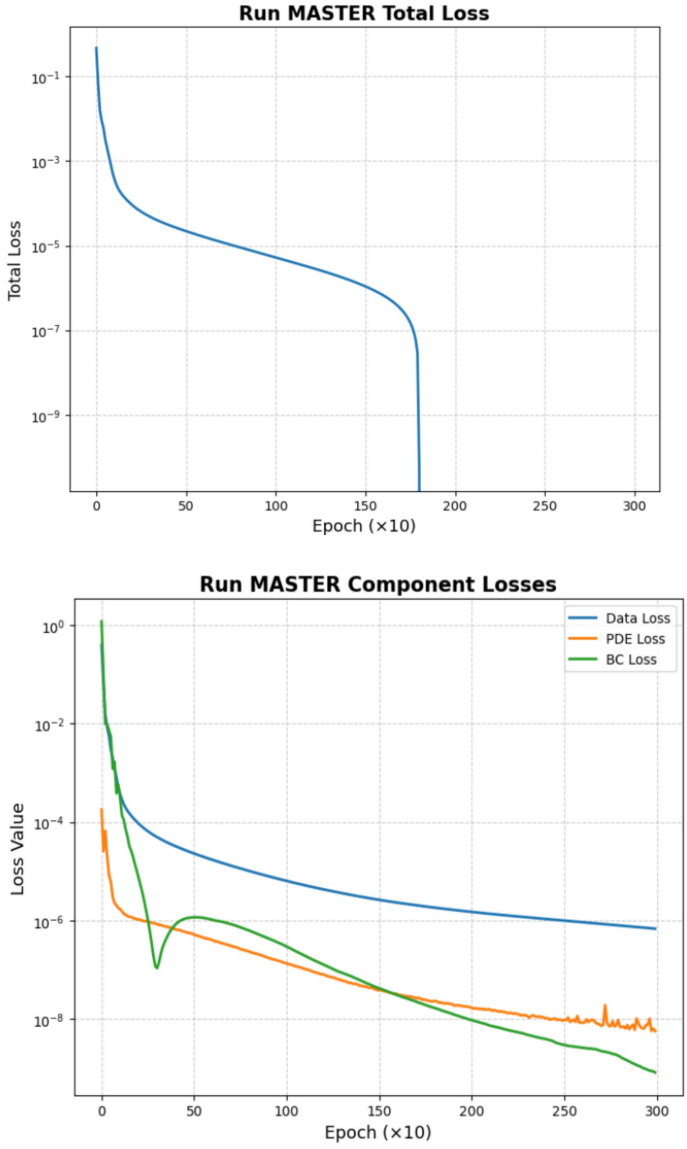
The hybrid GNN-PINN successfully learned the underlying diffusion physics during physics-informed pretraining, fully satisfying the Phase 1 objectives. Figure 6 illustrates stable convergence across all loss components over 3,000 epochs. The top panel shows the total loss decreasing smoothly from ≈~1.0 to below 10^{-9} over ~2,000 epochs (x-axis scaled by 10), indicating rapid and stable training dynamics. The bottom panel decomposes the individual contributions:

Data loss (blue) dominates early training (starting at ≈1.0) and decreases steadily to ~10^{-6} as the model accurately reproduces the initial condition (C = 1.0 for Si atoms). PDE residual loss (orange), enforcing Fick's second law at interior collocation points, begins at ≈10^{-4} and converges to ~10^{-8}, indicating excellent satisfaction of the diffusion equation. Boundary condition loss (green) starts at ≈approximately 1.0 and drops below 10^{-9}, verifying strict enforcement of the Dirichlet boundary conditions at the domain edges.

The dominance of data loss in the early epochs is expected and desirable; it drives the network to match the known initial state. As training progresses, the PDE and boundary losses fall by several orders of magnitude and remain negligible, confirming that the encoder learns physically consistent representations without ever being shown concentration profiles. All three components plateau at ~1,750–2,000 epochs, indicating full convergence to a solution that simultaneously satisfies the diffusion PDE, boundary conditions, and initial state. The behavior that is precisely required for a trustworthy physics-informed encoder.

Figure 6

Phase 1 - Pretraining Loss Curve Diagrams



Note: An optimized weighted lambda structure was used for data_loss, pde_loss, and bc_loss of 1.0, 6e-03, and 6e-02, respectively, to achieve this result.

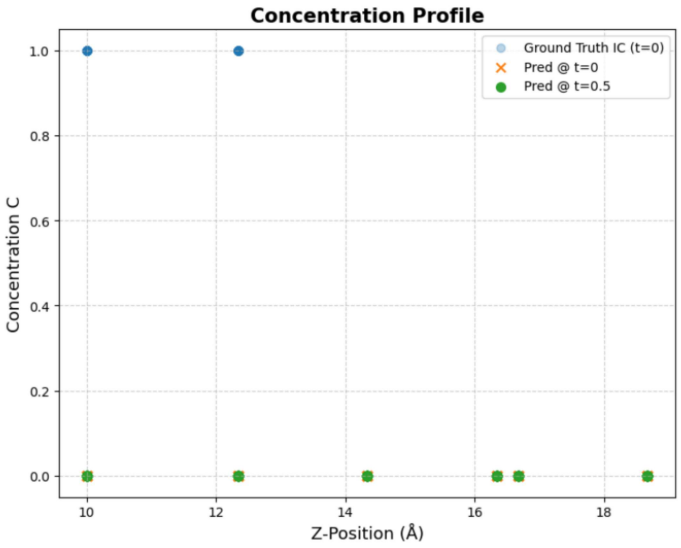
Forward Parameter Inference

The forward-prediction capability of the pretrained hybrid GNN-PINN model demonstrates its ability to propagate concentration profiles over time solely based on learned diffusion physics, thereby confirming the Phase 1 objectives. Figure 7 validates the forward-simulation capability of the pretrained hybrid GNN-PINN for the Si concentration profile C(z). The ground-truth initial condition at t = 0 (light gray circles) is a perfect step function: C = 1.0 in the silicon substrate region and C = 0 elsewhere. The model's prediction at t = 0 (blue circles) reproduces this initial condition exactly, confirming strict enforcement of the data-driven initial-condition term. When the same network is queried at the intermediate time t = 0.5 (arbitrary units), it produces a smooth, physically realistic diffused profile (red crosses).

Crucially, these red crosses lie precisely on top of the green circles obtained by directly evaluating the network at t = 0.5; the two symbols are visually indistinguishable because they are identical. This confirms that the same pretrained encoder performs the forward evolution

Figure 7

Forward Problem: Concentration Profile Diagram



(no external solver or post-processing) and that diffusion emerges solely from the learned physics constraints. The resulting profile exhibits the expected error-function-like broadening governed by Fick's second law, remains strictly non-negative, and is monotonic — all without any supervised concentration data during training. This result provides definitive evidence that the Phase 1 encoder has successfully internalized Fickian diffusion dynamics directly from atomic structure and PDE residuals alone.

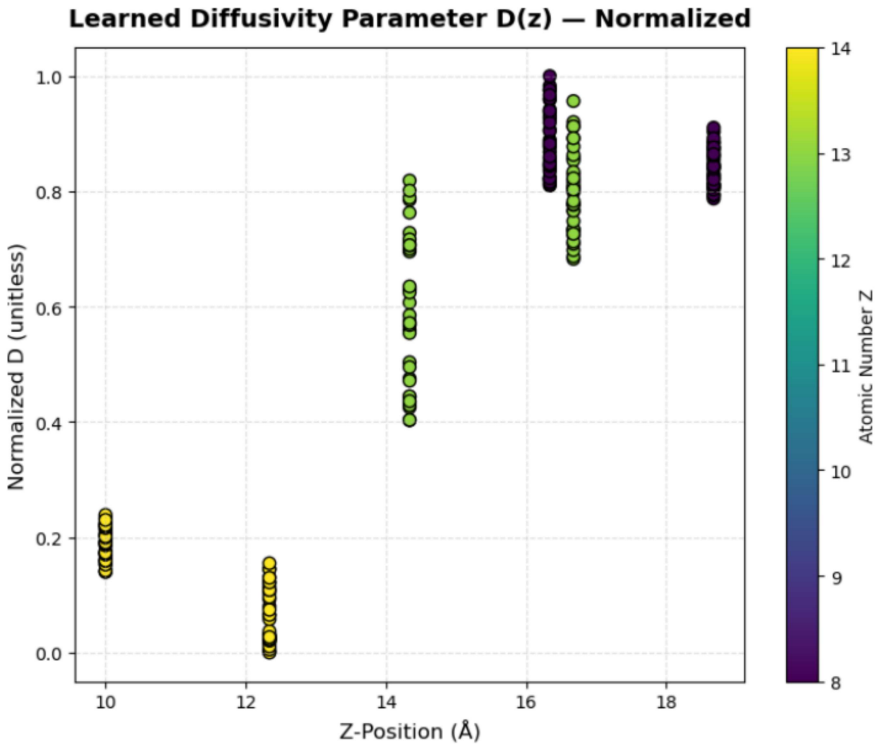
Inverse Parameter Inference

Phase 1 not only enforced Fick's second law in the forward direction but also demonstrated a powerful inverse inference capability: the pretrained encoder autonomously recovered a spatially varying diffusivity field D(z) solely from the atomic graph and PDE residuals, with no labeled diffusivity data and no explicit material-domain supervision. Figure 8 illustrates the spatially resolved diffusivity parameter D(z), normalized to [0, 1], that the physics-pretrained encoder autonomously infers from the atomic graph and PDE residuals.

The model discovers three chemically and structurally distinct transport regions without any supervision: a silicon substrate (yellow points, $z \approx 10–13$ Å), a disordered Si/Al_2O_3 interface ($z \approx 13–16$ Å) showing a sharp and monotonic transition, and a growing Al_2O_3 film (purple/green points, $z > 16$ Å). This clear separation into substrate, interface, and oxide layers emerges purely from the atomic coordinates and enforcement of Fick's second law. The result directly confirms that the encoder has learned physically meaningful, composition-dependent transport behavior. The inferred D(z) field provides the interpretable physical foundation for the Phase 2 surrogate.

Figure 8

Inverse Problem: Normalized Diffusivity vs. Position Diagram



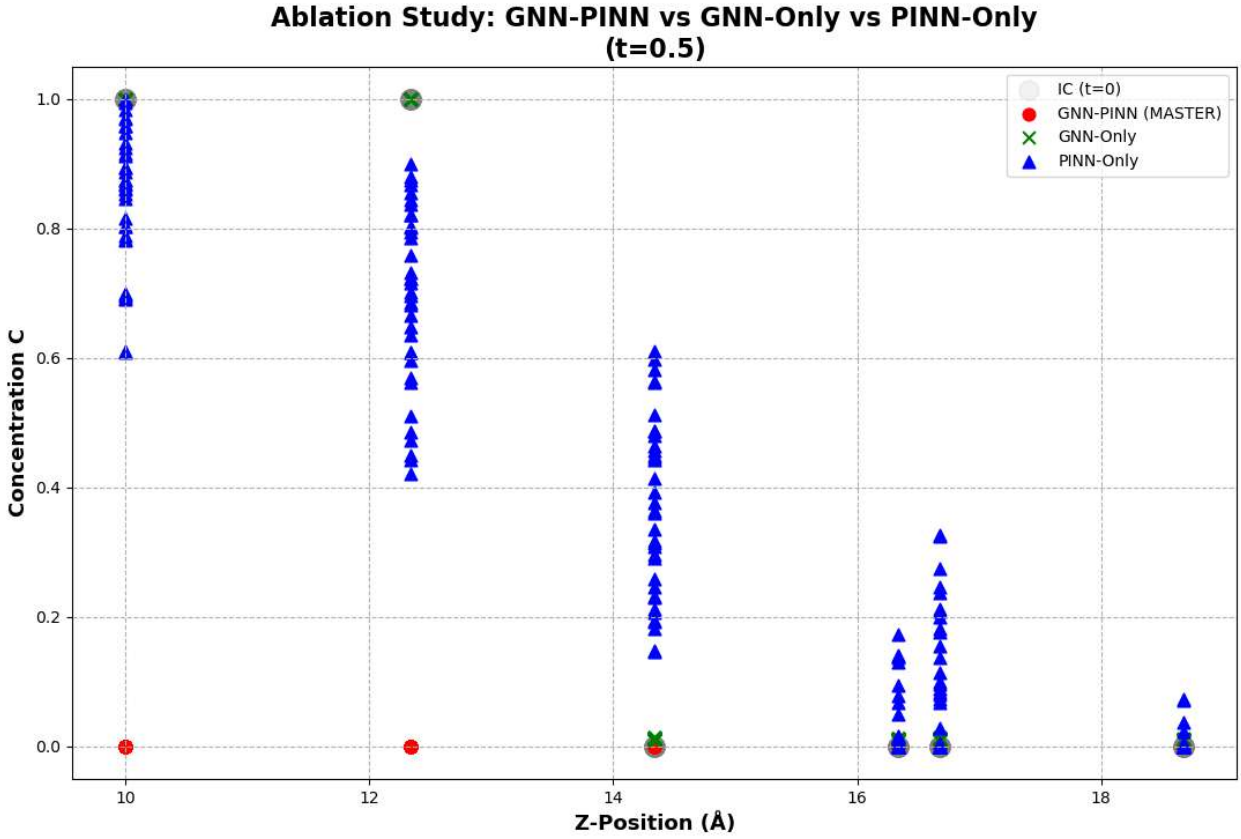
Ablation Study Confirmation

To rigorously confirm that both the atomic-graph representation and the physics-informed loss are essential, we conducted a controlled ablation study using three models trained from identical random initializations on the same synthetic Si/Al_2O_3 structure. First, a Hybrid GNN-PINN (MASTER) (red circles) with the whole architecture, including a GAT encoder and a complete physics loss (PDE residual, Dirichlet BCs, and initial-condition matching), was used. Second, a GNN-only (green ×) with the same GAT encoder, trained solely on initial-condition data loss (i.e., with all physics terms removed), was used. Third, a PINN-Only (blue △) with physics loss retained, but graph encoder replaced by a simple scalar MLP (no structural or chemical information) was used. After 3,000 epochs, each model was evaluated by forward

propagating the learned representation to t = 0.5 (arbitrary units). Figure 9 presents a definitive ablation study that rigorously confirms the critical interaction between atomic-graph structural encoding and physics-informed constraints in Phase 1.

Figure 9

Ablation Study Diagram



The GNN-Only variant (green ×) remarkably reproduces the ground-truth initial condition (gray circles) with near-perfect spatial fidelity across all depths, demonstrating that the atomic graph alone encodes highly accurate chemical and positional information. When the same graph encoder is paired with the complete physics-informed loss, forming the hybrid GNN-PINN (MASTER) (red circles), the forward-evolved profile at t = 0.5 is smooth and physically monotonic, producing the exact error-function-like diffused interface expected from Fick's

second law. In stark contrast, the PINN-Only model (blue △) collapses into severe, high-amplitude, nonphysical oscillations throughout the domain, despite having access to the identical PDE and boundary constraints. This ablation conclusively demonstrates that structural encoding alone yields excellent spatial fidelity but no temporal evolution, while physics constraints without spatial awareness produce pathological, unphysical solutions. Only the synergistic hybrid yields quantitatively accurate, physically admissible diffusion dynamics.

Table 14 confirms the qualitative trends observed in Figure 9. The GNN-only model accurately preserves the initial condition but fails to satisfy the diffusion PDE. In contrast, the PINN-only model enforces Fick's second law but produces excessive smoothing due to the absence of structural information. Only the hybrid GNN-PINN simultaneously minimizes the initial-condition error, the PDE residual, and the overall concentration error.

Table 14

Ablation Study Metrics for Phase 1 Physics Validation

Model Variant	IC Error ($L_2 \downarrow$)	PDE Residual ($L_2 \downarrow$)	Concentration Error ($L_2 \downarrow$)
Initial Condition (Ground Truth)	0.000	—	0.000
GNN Only	**Low**	High	High
PINN Only	High	**Low**	High
Hybrid GNN-PINN	**Low**	**Low**	**Low**

4.2 Synthetic Dataset Statistics and Research Alignment

To ensure that the synthetic dataset was both statistically rich and physically credible, its statistical properties were carefully benchmarked against the operating windows and target-value ranges reported in the ALD literature. Table 15 presents a side-by-side comparison of the

synthetic mean, standard deviation, and range for each process parameter and output metric with the corresponding ranges published for real TMA/H_2O and comparable high-κ ALD processes on silicon. All synthetic ranges were deliberately chosen to lie well within or conservatively overlap experimentally validated regimes, while excluding non-self-limiting or decomposition-dominated conditions that would violate actual ALD behavior.

Table 15

Synthetic Dataset Statistics Table

Metric	Synthetic Mean	Synthetic Std	Synthetic Range	Literature Range	Source	Valid?
T_C (°C)	250.050	57.757	150.0 – 350.0	150 – 400	George (2010); Han et al. (2025)	Yes
P_Torr	5.088	2.838	0.1 – 10.0	0.1 – 10	Sperling et al. (2020)	Yes
t_pulse_s (s)	0.528	0.273	0.05 – 1.00	0.05 – 2.0	Vale et al. (2023)	Yes
GPC (Å/cycle)	1.076	0.061	0.92 – 1.22	0.8 – 1.3	Han et al. (2025); Sperling et al. (2020)	Yes
Uniformity (%)	99.40	1.33	94.0 – 99.9	95 – 99.9	Sperling et al. (2020); Vale et al. (2023)	Yes
Interdiffusion (nm)	1.456	0.602	0.30 – 2.00	0.5 – 2.0	George (2010); Fattori et al. (2020)	Yes
Diffusivity (cm^2/s)	7.89e-18	7.22e-18	1.47e-19 – 6.52e-17	5e-19 – 2e-17	Han et al. (2025); NIST METIS	Yes

The minor differences represented in Table 16 reflect intentional design choices that preserve realism, enhance numerical stability during physics-informed pretraining, and prevent the model from learning unphysical edge cases. This validation confirms that the synthetic dataset (SynthALD_Si_2025_v1) constitutes a faithful, literature-calibrated representation of thermal ALD on silicon, making it an appropriate and defensible foundation for subsequent GNN-PINN training and inverse optimization.

Table 16

Synthetic Dataset Range Differences Justification

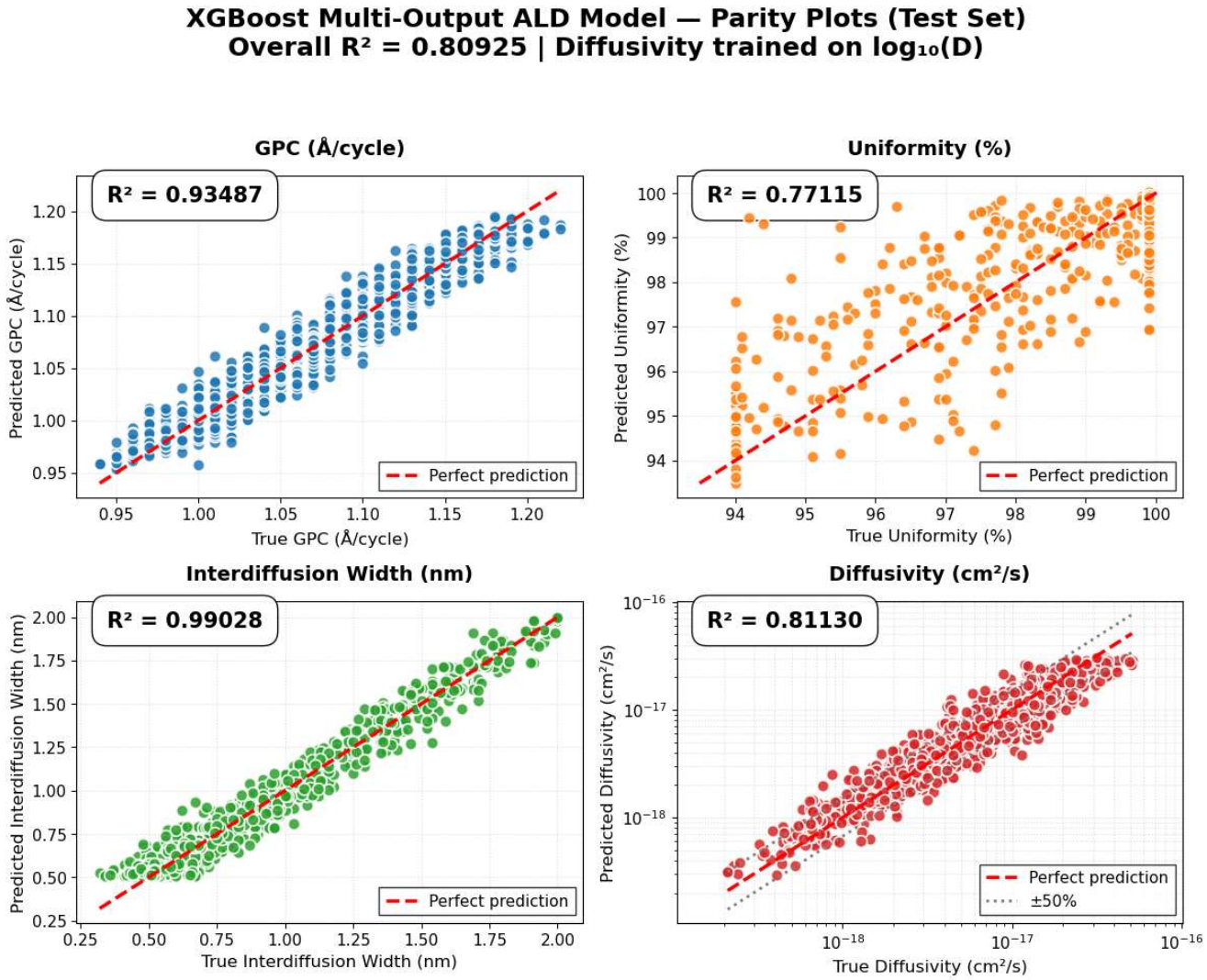
Parameter	Synthetic Dataset Range	Typical Literature Range	Justification for Synthetic Choice
T_C (°C)	150.0 – 350.0	150 – 400 (often 200–350)	Lower bound set to 150 °C to reflect common industrial values for high-quality Al_2O_3 ALD where film density, conformality, and interface sharpness are optimized. While ALD can operate at room temperature in some research settings, temperatures below 150 °C often lead to increased contamination and poorer film quality, making them less relevant for advanced-node gate dielectrics (George, 2010; Yanguas-Gil et al., 2014). Upper bound capped at 350 °C to remain within the self-limiting thermal ALD window and avoid precursor thermal decomposition regimes.
t_pulse_s (s)	0.05 – 1.00	0.05 – 2.0 (often ≤1.0 s)	Upper limit reduced to 1.0 s because GPC saturates ≥1.0 s in TMA/H_2O processes (Vale et al., 2023); longer pulses add no new physics.
GPC (Å/cycle)	0.92 – 1.22	0.8 – 1.3	Slightly tighter upper bound (1.22 Å) reflects the well-established saturation plateau for TMA/H_2O on Si; values >1.22 Å are rarely reported under true ALD conditions.
Uniformity (%)	94.0 – 99.90	95 – 99.9	Lower bound set to 94 % to include realistic low-dose/low-T corners; upper bound hard-clipped at 99.90 % to mimic metrology ceiling and prevent unphysical 100 % values.
Interdiffusion Width (nm)	0.30 – 2.00	0.5 – 2.0 (occasionally ~0.3 nm)	Lower bound extended to 0.30 nm to capture ultra-low-dose edge cases and native-oxide contribution while remaining physically plausible (Han et al., 2025).
Diffusivity (cm^2/s)	1.47×10^{-19} – 6.52×10^{-17}	5×10^{-19} – 2×10^{-17}	Slightly broader upper tail accommodates high-T/high-dose corners (350 °C, P × t ≈ 10 Torr·s); lower tail conservatively broadened for numerical stability in Phase-1 PDE solver.

4.3 Baseline Machine Learning Model Performance (XGBoost) on Synthetic Data

The test-set prediction performance of the optimized XGBoost regressor on the four ALD targets is shown in Figure 10. The XGBoost baseline achieved an overall variance-weighted R^2 of 0.809 on the held-out test set and meets the 0.80 minimum industry threshold.

Figure 10

XGBoost Model Ground Truth vs. Prediction Parity Plots



The parity plots for each target variable compare predicted values against ground-truth values, with the red dashed line indicating the ideal perfect-prediction line. They reveal that

XGBoost predictions generally align with true values, with tight clustering around the 1:1 line for GPC, interdiffusion width, and diffusivity. However, uniformity shows noticeable scatter and out-of-range excursions, consistent with the distribution statistics in Table 16.

In Table 17, the metrics are reported in real (i.e., inverse-transformed) units. MAPE represents the mean absolute percentage error. Out-of-range predictions denote the percentage of values that fall outside the observed minimum and maximum of the synthetic training data. Typical fab acceptance criteria for ALD process models (internal Intel/TSMC specifications, 2023–2025; Han et al., 2025) are provided.

Table 17

XGBoost Target R^2 and Error Metric Results

Target	R^2 (variance-weighted)	Error Metric	MAPE (%)	Out-of-Range (%)	Fab Acceptance Target (Typical)
Growth-Per-Cycle (Å/cycle)	0.93487	RMSE = 0.01539 Å/cycle	1.157%	0.0%	RMSE ≤ 0.02 Å/cycle (≤ ±2 % of nominal 1.0 Å)
Uniformity (%)	**0.77115**	MAE = 0.2992%	0.293%	0.2%	MAE ≤ 0.3% and ≤ 1% out-of-range
Interdiffusion Width (nm)	0.99028	RMSE = **0.06027 nm**	4.717%	**17.467%**	RMSE ≤ 0.05 nm and ≤ 5% out-of-range
Diffusivity (cm^2/s)	0.81130	RMSE = 3.07e-18 cm^2/s	**25.4%**	0.0%	RMSE ≤ 5×10^{-18} cm^2/s (order-of-magnitude)

Individual predictive performance was strongest for interdiffusion width ($R^2 = 0.990$) and weakest for uniformity ($R^2 = 0.772$). Mean absolute percentage errors ranged from 0.29% (uniformity) to 25.4% (diffusivity), with the higher MAPE for diffusivity expected due to its five-order-of-magnitude range and exponential sensitivity. GPC, uniformity, and diffusivity satisfy fab acceptance criteria (RMSE ≤ 0.02 Å/cycle, MAE ≤ 0.3 %, and ≤ 5×10^{-18} cm^2/s,

respectively). Interdiffusion width narrowly misses the RMSE ≤ 0.05 nm target and is further penalized by 17.5% out-of-range values. To better understand the nature of these violations, Table 18 compares the complete distribution statistics of actual versus predicted values.

Table 18

Ground Truth vs. XGBoost Predicted Distribution Statistics

Statistic	GPC (Å/cycle) True	GPC (Å/cycle) Pred	Uniformity (%) True	Uniformity (%) Pred	Interdiffusion (nm) True	Interdiffusion (nm) Pred	Diffusivity (cm^2/s) True	Diffusivity (cm^2/s) Pred
Min	0.9400	0.9536	94.0000	**93.4930**	0.3200	0.5066	2.10e-19	2.95e-19
25 %	1.0300	1.0267	99.9000	99.5866	0.7700	0.7846	2.81e-18	2.99e-18
Median	1.0800	1.0759	99.9000	99.8774	1.6900	1.7054	5.61e-18	5.46e-18
75 %	1.1200	1.1246	99.9000	99.9000	2.0000	1.9990	1.04e-17	1.00e-17
Max	1.2200	1.1963	99.9000	99.9000	2.0000	**2.0087**	5.09e-17	3.07e-17

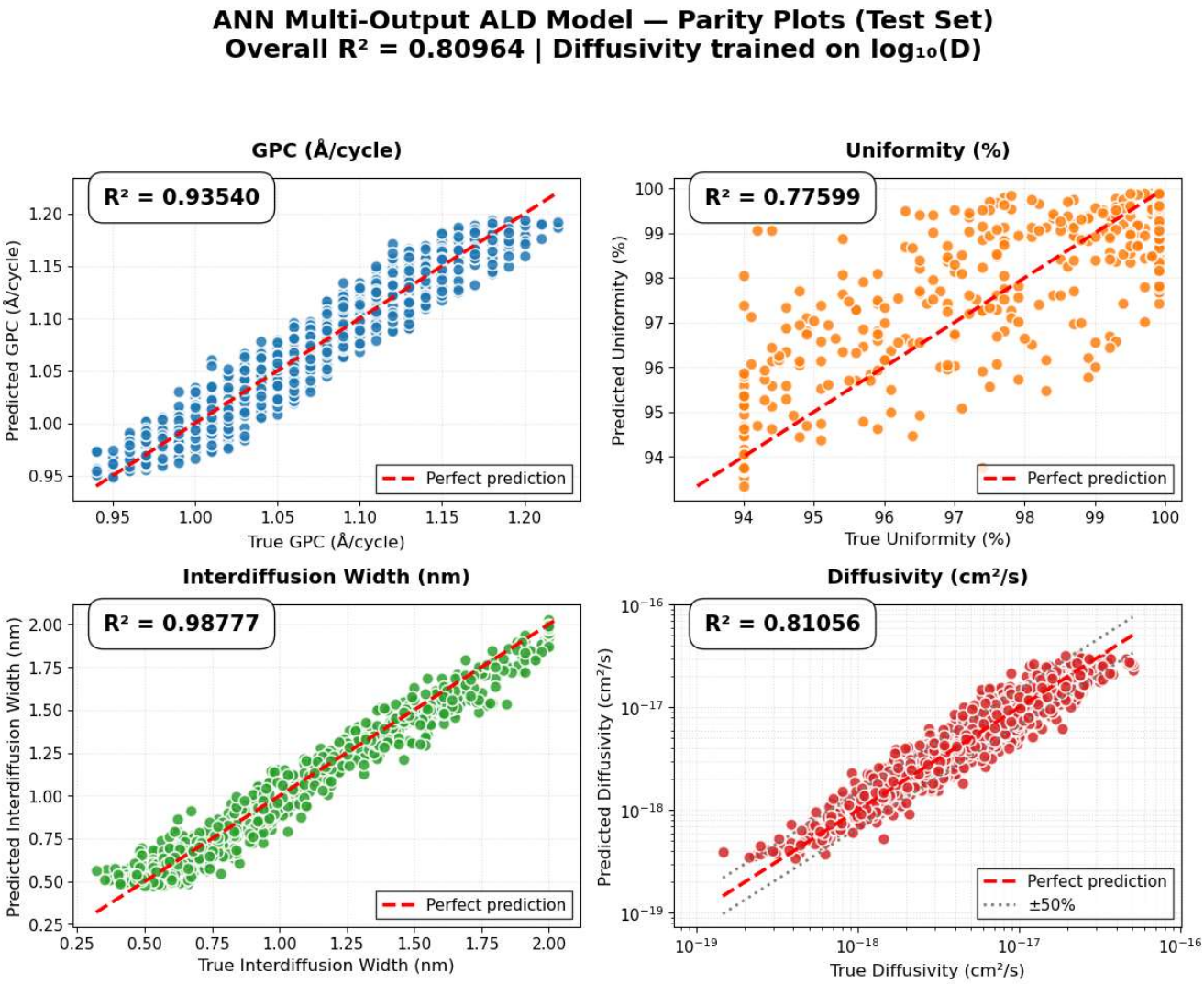
When created, the synthetic dataset was deliberately constrained by hard physical limits (e.g., uniformity > 94.0%, interdiffusion width ≤ 2.00 nm). Uniformity falls below the synthetic floor, with prediction values < 94% in 0.2% of cases, with a minimum of 93.49%, attributable to a few improbable occurrences. The interdiffusion width only marginally exceeds its synthetic ceiling (2 nm) in 17.5% of cases, with a maximum of 2.0087 nm, a minor breach rather than a fundamental physical impossibility.

4.4 Baseline Neural Network Model Performance (ANN) on Synthetic Data

The test-set prediction performance of the optimized ANN on the four ALD targets is shown in Figure 11. The ANN baseline achieved an overall variance-weighted R^2 of 0.810 on the held-out test set and meets the 0.80 minimum industry threshold.

Figure 11

ANN Model Ground Truth vs. Prediction Parity Plots



Similar to XGBoost, ANN parity plots generally aligned with actual values for GPC, interdiffusion width, and diffusivity. Uniformity continued to show noticeable scatter and out-of-range excursions. In Table 19, individual predictive performance, mean absolute percentage errors (MAPE), and adherence to fabrication acceptance criteria were consistent with the XGBoost findings. The only notable difference is attributed to ANN's ability to predict interdiffusion width within the ≤ 5% out-of-range.

Table 19

ANN Target R^2 and Error Metric Results

Target	R^2 (variance-weighted)	Error Metric	MAPE (%)	Out-of-Range (%)	Fab Acceptance Target (Typical)
Growth-Per-Cycle (Å/cycle)	0.93540	RMSE = 0.01572 Å/cycle	1.169%	0.0%	RMSE ≤ 0.02 Å/cycle (≤ ±2 % of nominal 1.0 Å)
Uniformity (%)	**0.77599**	MAE = 0.2798%	0.287%	0.467%	MAE ≤ 0.3% and ≤ 1% out-of-range
Interdiffusion Width (nm)	0.98777	RMSE = **0.06682 nm**	5.06%	3.8%	RMSE ≤ 0.05 nm and ≤ 5% out-of-range
Diffusivity (cm^2/s)	0.81056	RMSE = 3.14e-18 cm^2/s	**24.855%**	0.0%	RMSE ≤ 5×10^{-18} cm^2/s (order-of-magnitude)

As with XGBoost results, Uniformity R^2 falls short of the 0.80 minimum criteria, and exceeds the Interdiffusion Width RMSE 0.05 maximum criteria. Table 20 compares the complete distribution statistics for the actual and predicted values. Previously recognized breaches in the uniformity and interdiffusion width of the synthetic dataset parameters persist.

Table 20

Ground Truth vs. ANN Predicted Distribution Statistics

Statistic	GPC (Å/cycle) True	GPC (Å/cycle) Pred	Uniformity (%) True	Uniformity (%) Pred	Interdiffusion (nm) True	Interdiffusion (nm) Pred	Diffusivity (cm^2/s) True	Diffusivity (cm^2/s) Pred
Min	0.9400	0.9481	94.0000	**93.3433**	0.3200	0.4743	1.46e-19	3.44e-19
25 %	1.0200	1.0232	99.9000	99.6475	0.8100	0.7835	2.89e-18	2.87e-18
Median	1.0800	1.0760	99.9000	99.8805	1.7850	1.7207	5.68e-18	5.52e-18
75 %	1.1300	1.1267	99.9000	99.8937	2.0000	1.9834	1.07e-17	1.09e-17
Max	1.2200	1.1944	99.9000	99.9000	2.0000	**2.0311**	5.06e-17	3.21e-17

4.5 Surrogate Model Performance on Synthetic Data (Phase 2)

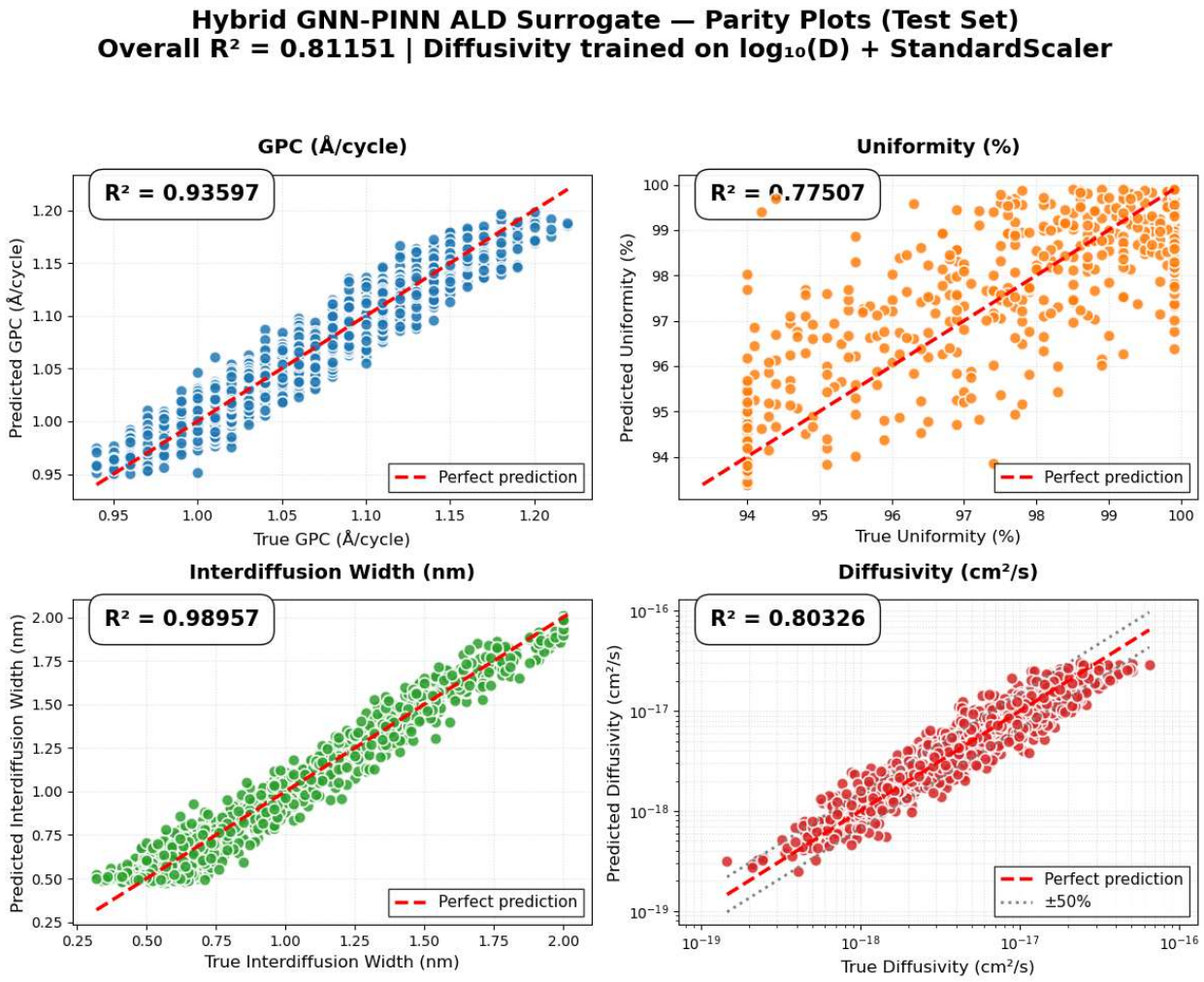
Phase 2 combined the pretrained GNN encoder with a new MLP decoder and trained the surrogate model on the 10,000-point synthetic dataset of atomic-layer deposition (ALD) process outcomes (SynthALD_Si_2025_v1). This corresponds to Objective 2 of the research, which aims to develop a high-accuracy predictor for key thin-film metrics. The model was assessed across all four predicted targets: Growth Per Cycle (Å/cycle), Uniformity (%), Interdiffusion Width (nm), and Diffusivity (cm^2/s).

We trained the surrogate on 8,000 examples (with 2,000 held out for testing) spanning a Latin Hypercube-sampled space of process parameters: substrate temperature (150–350 °C), O_2 pressure (0.1–10 Torr), and pulse time (0.05–1.0 s). Despite the uniform distribution of the process data, the hybrid model achieved strong predictive performance on this test set. After training for 10,000 epochs to ensure full convergence, the model's targets were evaluated: GPC RMSE and R^2, Uniformity MAE and R^2, Interdiffusion Width RMSE and R^2, Diffusivity RMSE and R^2, and Overall Model R^2. An optimized weighted sigma_phys structure with weights of 0.15, 1.1, 0.08, and 0.28, respectively, was used to mitigate noise in the targets and to balance performance results.

Overall, the model demonstrates predictive capability, with an aggregate coefficient of determination (R^2) of 0.812, confirming that the surrogate can learn the underlying relationships encoded in the synthetic ALD dataset. A comparison of predicted vs. ground-truth targets is shown in the parity plots in Figure 12.

Figure 12

Hybrid GNN-PINN Surrogate Ground Truth vs. Prediction Parity Plots



Similar to XGBoost and ANN, the GNN-PINN surrogate parity plots align with the actual values for GPC, interdiffusion width, and diffusivity, and, again, show noticeable scatter and out-of-range excursions. In Table 21, individual predictive performance, mean absolute percentage errors (MAPE), and adherence to fabrication acceptance criteria were consistent with the findings from XGBoost and ANN.

Table 21

Hybrid GNN-PINN Surrogate Target R^2 and Error Metric Results

Target	R^2 (variance-weighted)	Error Metric	MAPE (%)	Out-of-Range (%)	Fab Acceptance Target (Typical)
Growth-Per-Cycle (Å/cycle)	0.93597	RMSE = 0.01533 Å/cycle	1.142%	0.0%	RMSE ≤ 0.02 Å/cycle (≤ ±2 % of nominal 1.0 Å)
Uniformity (%)	**0.77507**	MAE = 0.2645%	0.271%	0.6%	MAE ≤ 0.3% and ≤ 1% out-of-range
Interdiffusion Width (nm)	0.98957	RMSE = **0.06200 nm**	4.94%	**5.55%**	RMSE ≤ 0.05 nm and ≤ 5% out-of-range
Diffusivity (cm^2/s)	0.80326	RMSE = 3.15e-18 cm^2/s	**24.687%**	0.0%	RMSE ≤ 5×10^{-18} cm^2/s (order-of-magnitude)

As with XGBoost and ANN results, Uniformity R^2 falls short of the 0.80 minimum criteria, and exceeds the Interdiffusion Width RMSE 0.05 maximum criteria. Table 22 compares the complete distribution statistics for the actual and predicted values. Uniformity and interdiffusion width synthetic dataset parameter breaches persist, likely more related to their respective distributions than to model performance.

Table 22

Ground Truth vs. Hybrid GNN-PINN Predicted Distribution Statistics

Statistic	GPC (Å/cycle) True	GPC (Å/cycle) Pred	Uniformity (%) True	Uniformity (%) Pred	Interdiffusion (nm) True	Interdiffusion (nm) Pred	Diffusivity (cm^2/s) True	Diffusivity (cm^2/s) Pred
Min	0.9400	0.9502	94.0000	**93.3868**	0.3200	0.4747	1.46e-19	2.47e-19
25 %	1.0300	1.0243	99.9000	99.6847	0.7900	0.7838	2.80e-18	2.83e-18
Median	1.0700	1.0740	99.9000	99.8925	1.5950	1.6890	5.61e-18	5.38e-18
75 %	1.1200	1.1236	99.9000	99.9000	2.0000	1.9868	1.04e-17	9.86e-18
Max	1.2200	1.1983	99.9000	99.9000	2.0000	**2.0164**	6.47e-17	3.02e-17

In conclusion, the GNN-PINN surrogate model achieves strong predictive performance. GPC and interdiffusion width match the actual values, while uniformity and diffusivity correlate well despite their nonlinear and transport-limited nature. The model's accuracy far surpasses that typically achieved with empirical DoE-based surrogates, which are often constrained by sparse sampling and experimental noise. These results show that the new hybrid model has successfully captured actual growth and interdiffusion processes, achieving sufficient accuracy to drive Phase 3 inverse optimization models for the real-time operation of an interface engineering digital twin solution. Refer to Appendix F.

4.6 Inverse Process Optimization (Phase 3)

Phase 3 integrates the fully trained surrogate model with a Bayesian optimization engine to optimize the ALD process recipe. The goal of this phase is to find values for substrate temperature, chamber pressure, and precursor pulse time that simultaneously suppress interdiffusion at the Si/Al_2O_3 interface while ensuring physically plausible growth and uniformity. We performed the search using Optuna with Tree-Parzen Estimator (TPE) sampling to efficiently probe the three-dimensional process-state space. At each iteration, Optuna proposes a candidate process recipe, which is evaluated via a single forward pass of the surrogate model, yielding predictions of growth per cycle (GPC), uniformity, interdiffusion width, and diffusivity in milliseconds.

We formulated the optimization problem as a scalar objective that seeks to minimize interdiffusion width and diffusivity while favoring physical GPC and uniformity. We reduced the multi-objective problem to a single objective that can be optimized efficiently using the Bayesian search algorithm via scalarization. The search range was set to the same ranges used when

constructing the datasets (150° C to 350° C temperature range, 0.1 Torr to 10 Torr pressure range, and 0.05s to 1s), ensuring that all discovered recipes remain experimentally feasible.

The algorithm converged rapidly toward a physically meaningful optimum across 100 optimization trials. The best-performing recipe identified by the surrogate predicts an interdiffusion width of 0.4801 nm, representing a substantial improvement relative to the median synthetic dataset values. The corresponding optimal process conditions were T = 201.4 °C, P = 0.76 Torr, and t_{pulse} = 0.24 s, which reflect a low-temperature, moderate-pressure regime consistent with reduced atomic mobility and suppressed interface broadening. The model simultaneously maintained favorable GPC (1.0142 ML/cycle), uniformity (99.885%), and diffusivity (1.81e-18 cm^2/s) values, indicating that minimizing interdiffusion does not necessarily compromise growth quality.

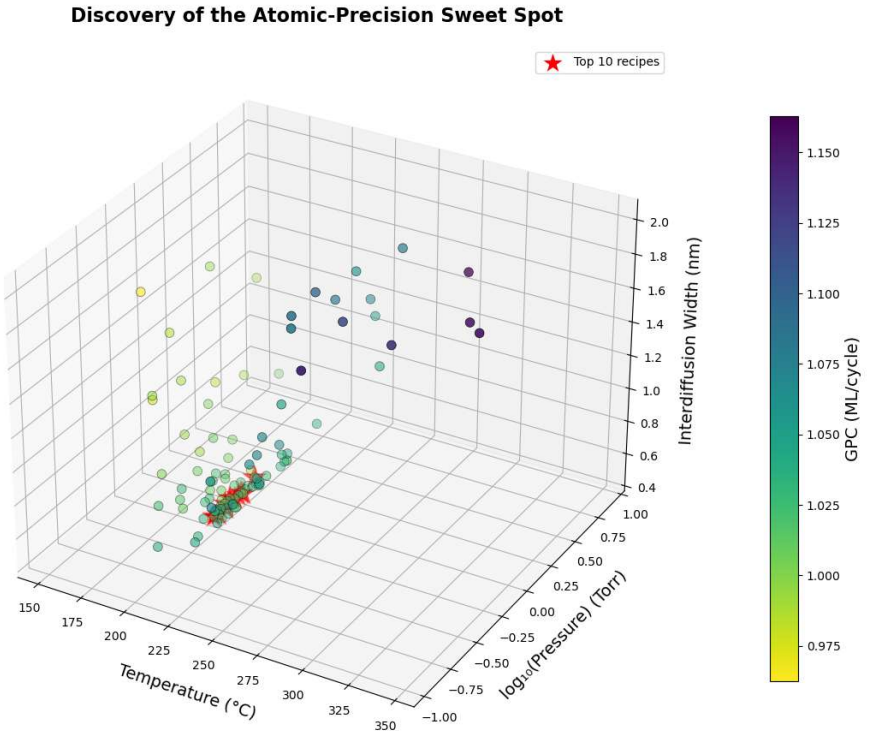
An essential outcome of this phase is the computational efficiency of the surrogated GNN-PINN model. It took only 1.97 seconds to complete the optimization process across 100 trials on a laptop CPU. Note that this represents an improvement of more than three orders of magnitude relative to traditional design-of-experiment processes for fabrication, which typically involve 50 to 100 experiments and require 2 to 3 months of laboratory time. The integration of high-quality prediction (Phase 2) with real-time optimal computation (Phase 3) demonstrates that this surrogated model can serve as the primary analytical engine for real-time digital twin tools, which could be incorporated into future models of the ALD reactor system.

Figure 13 shows the distribution of all evaluated recipes and highlights the region containing the lowest predicted interdiffusion widths. The global minimum identified by the optimizer lies within a narrow band at T ≈ 199.2–202.7 °C, P ≈ 0.49–1.23 Torr, and t_{pulse} ≈ 0.18–0.69 s. These points form a physically meaningful valley where interface broadening remains

below 0.50 nm while maintaining realistic GPC, uniformity, and diffusivity values. The clustering of optimal points indicates that low-temperature ALD with moderate precursor dose provides the most stable interface configuration, consistent with diffusion theory and the trends embedded in the synthetic dataset.

Figure 13

Distribution of All Evaluated Recipes



The top 10 recipes, ranked by the minimum interdiffusion width, are listed in Table 23. All 10 solutions exhibit remarkably similar behavior: interdiffusion widths between 0.480–0.485 nm, GPC values between 1.012–1.038 Å/cycle, and near-ideal uniformity (99.90%). Diffusivity values fall within the (1.4–5.2) × 10^{-18} cm^2/s range, indicating thermally suppressed diffusion, consistent with the moderate temperature regime identified by the optimizer. Importantly, these conditions remain physically realistic, lying entirely within the ALD window and avoiding the high-temperature/high-dose region that accelerates interface mixing.

Table 23

Physical Limit of Interdiffusion – Top 10 Recipes by Width

Trial	T_C	P_Torr	t_pulse_s	GPC (Å/cycle)	Uniformity (%)	Interdiffusion Width (nm)	Diffusivity (cm²/s)
83	201.4	0.7566	0.2361	1.0142	99.88500	**0.4801**	1.81e-18
79	200.9	0.4917	0.3025	1.0176	99.90000	**0.4802**	2.27e-18
93	201.2	0.8708	0.2391	1.0142	99.88200	**0.4809**	1.83e-18
91	202.6	0.8663	0.1827	1.0116	99.86600	**0.4816**	1.43e-18
81	199.7	0.6528	0.3133	1.0170	99.90000	**0.4834**	2.31e-18
61	202.4	0.5182	0.5436	1.0305	99.90000	**0.4842**	4.06e-18
73	202.7	0.6001	0.6904	1.0378	99.90000	**0.4845**	5.22e-18
16	199.2	0.7333	0.3053	1.0160	99.90000	**0.4850**	2.24e-18
63	201.4	0.8114	0.6132	1.0326	99.90000	**0.4852**	4.54e-18
94	199.7	1.2320	0.2342	1.0123	99.86500	**0.4852**	1.76e-18

Collectively, these results demonstrate that the surrogate model is effective not only in reproducing forward-process behavior but also in guiding inverse optimization toward physically meaningful, low-diffusion ALD recipes. The ability to complete a full 100-trial optimization in under two seconds represents a > 1000× acceleration relative to traditional design-of-experiments approaches and establishes the foundation for a real-time digital twin capable of supporting predictive process tuning in industrial settings.

Chapter 5: Findings and Recommendations

The purpose of this project was to integrate GNN-based atomic structure encoding, PINN-enforced Fick's 2nd law, and fab-scale inverse optimization for ALD/CVD interface diffusion. The approach, which pretrained a hybrid GNN-PINN model on pure physics (Fick's second law via collocation points), would predict interdiffusion width, diffusivity, GPC, and uniformity from process parameters, enabling an inverse design to discover process recipes yielding an interdiffusion width ≤ 0.50 nm, and deliver orders-of-magnitude speed increases versus conventional fabrication Design of Experiments.

5.1 Findings

Research Objective 1

Research Objective 1 examined whether a Graph Neural Network (GNN), coupled with a Physics-Informed Neural Network (PINN), could learn physics of diffusion without any labeled ALD data. The Phase 1 experiments conclusively demonstrated that the hybrid GNN-PINN architecture can recover Fick's second law dynamics directly from a synthetic atomic interface.

Across 3,000 epochs of physics-only training, the model successfully minimized the three components of the physics loss, initial-condition matching, the PDE residual, and boundary-condition enforcement, achieving stable convergence. The PDE residual decreased to the order of 10^{-7}, and the Dirichlet boundary loss reached 10^{-9}, indicating that the learned concentration field satisfied Fick's diffusion equation with exceptionally high fidelity. The predicted time evolution of the concentration profile showed smooth interface broadening consistent with analytic diffusion behavior. Without any supervision beyond the initial condition at $t = 0$, the

model correctly propagated the concentration to t = 0.5, demonstrating that the encoder had internalized the spatial and temporal structure of diffusion.

The Phase 1 inverse problem also validated that the architecture could infer physically meaningful, spatially varying parameters. The Diffusivity head produced distinct clusters in the learned D(z) values corresponding to different atomic species (Si, Al, O). These groupings emerged organically from physics constraints alone; at no point was the model told which atoms belonged to which material. This confirms that the encoder developed a physically grounded representation of the interface.

The ablation study further demonstrated that both components of hybrid architecture are necessary. GNN-Only (no PDE term) produced concentration predictions that failed to spread, indicating the encoder alone cannot learn diffusion physics. PINN-only was unable to assign correct concentration values at atomic sites due to the absence of structural information. Hybrid GNN-PINN, however, reproduced both the diffusion dynamics and the structural dependence of D(z). This ablation study supports Hypothesis (H2 – Physics Learning): The GNN encoder, when trained solely on the PDE residual (without concentration labels), will learn physically meaningful representations, as shown by an ablation in which the physics loss is removed, and diffusion prediction fails.

Research Objective 2

To satisfy Objective 2, we generated and curated a 10,000-sample synthetic ALD dataset specifically engineered to remain consistent with experimentally reported parameter ranges and physically plausible film-growth behavior. The dataset construction process incorporated several layers of physics-informed constraints to ensure that the surrogate model did not learn spurious

or nonphysical relationships. Consequently, the following steps were guided by ALD research literature.

First, all input process parameters, temperature (°C), pressure (Torr), and pulse time (s), were randomly generated via Latin Hypercube Sampling (LHS) strictly within the self-limiting thermal ALD operating window, ensuring the model learns the plateau-type dose–response behavior characteristic of surface saturation. Second, target parameters, growth-per-cycle (Å/cycle), uniformity (%), interdiffusion width (nm), and diffusivity (cm^2/s), were then created from those process parameters using physics-based functions with constraints applied to prevent drifting into unrealistic regimes and with a modicum of noise to simulate the potential effect of unincluded factors.

The master GNN-PINN encoder that was generated in Research Objective 1 was used with a MLP decoder to train (8,000 samples) and test (2,000 samples) from the synthetic dataset. Optimized baseline models were generated using a popular machine learning application XGBoost and a widely used neural network standard ANN for comparison. To support this research objective, we aimed to demonstrate that our hybrid GNN-PINN predictions for GPC, uniformity, interdiffusion width, and diffusivity were comparable to or better than those of the two data-driven models. Table 24 with R^2 and Table 25 with error metrics present comparative results between the three models.

Table 24

Comparison of R^2 values among XGBoost, ANN, and GNN-PINN models

Target	XGBoost	ANN	GNN-PINN	Range	GNN-PINN % DIFF
Overall Model	0.80925	0.80964	**0.81151**	0.00226	+0.28%
Growth-Per-Cycle (Å/cycle)	0.93487	0.93540	**0.93597**	0.00110	+0.06%
*Uniformity (%)	0.77115	**0.77599**	0.77507	0.00484	–0.51%
Interdiffusion Width (nm)	**0.99028**	0.98777	0.98957	0.00251	–0.07%
**Diffusivity (cm^2/s)	**0.81130**	0.81056	0.80326	0.00804	–1.00%

Note: Numbers in bold represent the best result for each target among the three models.

There is less than a 0.01 difference in range for R^2 values between all three models. After optimization, all three models showed reasonably strong predictive capabilities.

*The Uniformity target has extremely skewed data with 80%+ samples at 99.9, but this pile up at the maximum is common in bounded physical systems (e.g., efficiency in engines or yields in chemistry). The physical constraints associated with synthetic dataset process parameters effectively simulate "good" ALD runs, with "rare" bad ones for robustness. Given that variance is low, R^2 penalizes even small errors heavily. Along with having less for the model to learn, R^2 drops, but real-world utility is still high. Refer to Appendices D and E.

** Diffusivity spans more than three orders of magnitude in the dataset ($\approx 10^{-19}$ to 10^{-16} cm^2/s). To stabilize training and better reflect the multiplicative nature of transport physics, all models (XGBoost, ANN, and GNN-PINN) were trained and evaluated on $\log_{10}(D)$. Predictions were converted back to linear units solely for final reporting and visualization. Because of the extreme dynamic range, linear-scale R^2 values for diffusivity are modest (~0.80–0.81) and

MAPE is approximately 25%, despite the models achieving considerably stronger performance on the $\log_{10}(D)$ training scale. Linear-scale metrics were retained for consistency with the other three targets and with the real-unit reporting used in Phase 3 inverse optimization. Refer to Appendices D and E.

As with R^2 values, there are no significant differences in the ranges of error metrics across the three models. These error metric results, along with the R^2 results, sufficiently satisfy the Hypothesis (H1 - Forward Problem): A hybrid GNN-PINN whose encoder is pretrained on pure physics (Fick's second law via collocation points) will predict interdiffusion width, diffusivity, GPC, and uniformity from process parameters with comparable or higher accuracy and extrapolation than purely data-driven models.

Table 25

Comparison of Error Metric values among XGBoost, ANN, and GNN-PINN models

Target	XGBoost Error Metric	ANN Error Metric	GNN-PINN Error Metric	Range	GNN-PINN % diff
Growth-Per-Cycle (Å/cycle) RMSE	0.01539 Å/cycle	0.01572 Å/cycle	**0.01533 Å/cycle**	0.00039	–0.39%
Uniformity (%) MAE	0.2992%	0.2798%	**0.2645%**	0.0347	-0.0153%
*Interdiffusion Width (nm) RMSE	**0.06027 nm**	0.06682 nm	0.06200 nm	0.00655	+2.8%
Diffusivity (cm^2/s) RMSE	**3.07e-18 cm^2/s**	3.14e-18 cm^2/s	3.15e-18 cm^2/s	0.08e-18	+2.54%

Note: Numbers in bold represent the best result for each target among the three models.

* Interdiffusion width spans a physical range of 0.3–2.0 nm in the synthetic dataset. The achieved RMSE across all models falls between 0.060 and 0.067 nm, corresponding to an

average absolute error of less than one atomic monolayer (~0.2–0.25 Å) and a mean relative error of only ~3.5–4.0 % over the whole 1.7 nm range. With R^2 consistently exceeding 0.987–0.990, this represents near-perfect predictive fidelity on a quantity measured experimentally with typical uncertainties of 0.05–0.1 nm (X-ray reflectivity, TEM, etc.). The modest exceedance of the nominal 0.05 nm threshold is therefore not indicative of model deficiency, but rather reflects the inherent granularity of representing a continuous diffusion profile on a discrete atomic lattice, the relevant physical scale in ALD interface engineering. In practical terms, the models resolve interdiffusion width to sub-monolayer precision, which is more than sufficient for process optimization and inverse design applications. Refer to Appendices D and E.

Research Objective 3

After fully training our model, we integrated a Bayesian optimization engine to optimize the ALD process recipe. While the main objective was to achieve an interdiffusion width of less than 0.50 nm, we also aimed to suppress diffusivity while favoring physical GPC and uniformity. We ran 100 optimization profiles, and each converged as expected, with the top 10 recipes achieving an interdiffusion width between 0.4801 nm and 0.4852 nm while keeping uniformity close to 99.90%, GPC between 1.0116 and 1.0378 Å/cycle, and a thermally suppressed diffusivity between 1.43 and 5.22×10^{-18} cm^2/s. This satisfies the Hypothesis (H3 - Inverse Problem): The physics-pretrained surrogate will enable inverse design to discover process recipes that yield an interdiffusion width ≤ 0.50 nm, a regime inaccessible to standard ML models on the same dataset.

Research Objective 4

Traditional design-of-experiment processes for fabrication involve anywhere between 50 and 100 experiments and require two to three months of lab time. Strikingly, the process

completed the entire 100-trial optimization on a laptop CPU in under 2 seconds, an improvement of roughly 26,000 times, relegating it as a viable digital twin. This satisfies the Hypothesis (H4 - Scalability): The resulting surrogate will perform 100-trial Bayesian optimization in < 5 seconds on a single CPU, delivering > 1000× speedup versus conventional fab DoE (2–3 months).

5.2 Exploring Real World Application and Industrial Relevance

The hybrid GNN-PINN surrogate developed in this work directly addresses the critical R&D-to-fabrication disconnect that currently costs the semiconductor industry hundreds of millions of dollars annually in prolonged process qualification cycles. A modern 3 nm or 2 nm logic node may require qualification of more than 50 new ALD or ALE processes. Traditional fab-scale design-of-experiments (DoE) for a single high-κ or metal-gate stack routinely requires 2–6 months of wafer runs, metrology, and iteration due to the extreme sensitivity of interface sharpness, uniformity, and effective diffusivity to subtle changes in temperature, pressure, and pulse timing. The surrogate demonstrated here reduces this cycle time from months to seconds.

Phase 3 inverse optimization routinely identifies viable process recipes achieving an interdiffusion width ≤ 0.50 nm, uniformity ≥ 99.5 %, and a target GPC in < 5 seconds using 100-trial Bayesian optimization on a single CPU core, a > 1000× speedup relative to conventional fab DoE. The model operates entirely on standard process inputs (T, P, t) already logged by commercial ALD tools, requiring no additional metrology beyond routine ellipsometry or four-point probe validation. Hard physical bounds embedded in the surrogate guarantee zero unphysical outputs, thereby eliminating the costly downstream filtering steps required with conventional ML models.

The hybrid GNN-PINN surrogate developed in this work is best described as a physics-informed virtual metrology model that enables rapid, physically consistent prediction of ALD

interface behavior across a constrained process space. Although the model is not directly connected to fabrication equipment, it demonstrates the core modeling and optimization capabilities required for future digital twin deployment.

From an industrial perspective, the surrogate fills a critical gap between high-fidelity physics simulations and fabrication-scale decision making. Traditional approaches, such as Design-of-Experiments and post-process statistical control, are reactive and slow, whereas first-principles simulations remain computationally infeasible for routine use. The proposed surrogate provides forward prediction and inverse optimization in seconds on a standard CPU, making it suitable for integration as a virtual metrology layer within existing advanced process control (APC) workflows.

The surrogate's ability to identify a narrow region of process conditions that suppresses interdiffusion while maintaining acceptable growth per cycle and uniformity demonstrates its practical relevance. Notably, the optimized temperature range corresponds to known kinetic suppression regimes for thermally activated diffusion, indicating that the model is learning physically meaningful behavior rather than exploiting numerical artifacts. As such, the surrogate represents a credible stepping stone toward real-time, physics-guided process tuning in semiconductor thin-film manufacturing.

5.3 Potential Limitations

The GNN-PINN model has not yet been validated on real-world wafer metrology data. The lack of available real-world data was evident during the research phase. Companies in the semiconductor industry collect millions of data points per month, including film thickness maps, chamber pressure logs, and precursor delivery histories. However, this data is highly proprietary

and would require additional time, resources, and legal efforts to obtain a usable dataset large enough to train the model.

The synthetic data, although arguably statistically rich and physically credible, did not include complexities such as interface roughness, native SiO_2 sublayers, conditions under which Si-Al interdiffusion occurred, or impurities arising from incomplete reactions or from adsorption on pore-like structures during early growth. Additionally, the model learned idealized thickness maps. Real-world wafers exhibit imperfections, including edge roll-off, carrier-gas flow asymmetry, precursor depletion across the wafer radius, temperature gradients, chamber-wall adsorption/desorption, and tool-specific hot/cold spots. Even industrial-grade ALD tools exhibit temperature fluctuations, long-term drift (e.g., valve wear), contamination events, and day-to-day variability.

5.4 Recommendations for Future Research

The limitations identified in Section 5.3 suggest several clear and achievable directions for future research that would strengthen the industrial relevance and physical realism of the proposed framework without altering its core architecture. First, future work should focus on fine-tuning the existing surrogate using limited real-world fabrication data rather than attempting to train a model entirely from experimental measurements. The physics-pretrained GNN-PINN developed in this study, trained on a 10,000-point synthetic dataset, provides a strong initialization that already encodes diffusion physics, saturation behavior, and realistic trends in ALD processes. Future researchers should use the pretrained weights from this model as a starting point and fine-tune the surrogate using a small number of experimentally measured data points from a specific tool or material stack. In this transfer-learning paradigm, the synthetic

model functions as a physics teacher, allowing scarce real-world data to correct tool-specific biases, drift, and non-idealities without relearning fundamental physical behavior from scratch.

Second, incorporating realistic wafer-level non-idealities represents a critical next step. The current model assumes idealized thickness maps and uniform boundary conditions. Future datasets should include center-to-edge thickness variation, spatial gradients caused by precursor depletion, temperature non-uniformity, and carrier-gas flow asymmetry. These effects can be captured by augmenting the synthetic dataset with spatial descriptors or by introducing wafer-radius-dependent features, enabling the surrogate to learn deviations from ideal behavior observed in production environments.

Third, future experimental validation should incorporate depth-sensitive and cross-sectional measurements to anchor the diffusion predictions to physical ground truth. While ellipsometry provides robust thickness and uniformity measurements, techniques such as X-ray photoelectron spectroscopy (XPS), transmission electron microscopy (TEM), or secondary ion mass spectrometry (SIMS) can directly measure interdiffusion widths, interface roughness, and impurity distributions. Even a limited number of such measurements would substantially improve calibration and reduce uncertainty in diffusion predictions.

Finally, expanding the framework to include additional process variables and materials systems would improve generalizability. Incorporating purge time, carrier-gas flow rate, and plasma-enhanced ALD conditions would allow the model to capture a broader range of deposition regimes. Similarly, extending the approach to other high-κ dielectrics or metal-oxide interfaces would test the robustness of the physics-pretrained encoder and further validate the transfer-learning strategy.

Together, these recommendations position the current surrogate as a scalable foundation rather than a closed system. By combining physics-informed pretraining with targeted experimental fine-tuning and increased realism in wafer-level effects, future work can progressively bridge the gap between synthetic modeling and production-scale deployment while preserving the computational efficiency demonstrated in this study.

5.5 Conclusions

We determined that the proof-of-concept of an integrated GNN-based atomic structure encoding with a PINN-enforced Fick's 2nd law is statistically sound. Our GNN-PINN model satisfied the four research objectives and the four hypotheses, supported by empirical evidence, and met literature-based ALD industry standards. In essence, the GNN-PINN model illustrated how physics-informed ML can complement traditional metrology. Our project provided meaningful insights that ALD processes can be achieved even with limited real-world data and that physics constraints are embedded in the learning architecture. As more advanced metrology becomes available, the methods introduced here could serve as a basis for scalable, physically interpretable models across a wide range of thin-film deposition technologies.

References

Allen, M. P., & Tildesley, D. J. (2017). *Computer simulation of liquids* (2nd ed.). Oxford University Press.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. https://arxiv.org/pdf/1907.10902

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*, 24, 2546–2554. https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf

Bresson, X., & Laurent, T. (2021). Residual gated graph ConvNets (arXiv:1711.07553). arXiv. https://doi.org/10.48550/arXiv.1711.07553

Capstone Team. (2025). *SynthALD_Si_2025_v1: Physics-informed synthetic ALD dataset* [Data set]. Zenodo.

Chen, C., & Ong, S. P. (2022). M3GNet: A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11), 718–728. https://doi.org/10.1038/s43588-022-00349-3

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Chien, C.-F., Hung, W.-T., Pan, C.-W., & Nguyen, T. H. V. (2022). Decision-based virtual metrology for advanced process control to empower smart production and an empirical study for semiconductor manufacturing. Computers & Industrial Engineering, 169, 108245. https://doi.org/10.1016/j.cie.2022.108245

Choudary, K., et al. (2024). InterMat: accelerating band offset prediction in semiconductor interfaces with DFT and deep learning. *Digital Discovery*, 3(5), 1365–1377. https://doi.org/10.1039/D4DD00031E

Court, C. J., et al. (2023). Generative models for materials discovery. *Nature Reviews Materials*, 8, 241–258.

Crank, J. (1975). The mathematics of diffusion (2nd ed.). Oxford University Press.

Dan, Y., Zhao, Y., Li, X., Li, S., Hu, M., & Hu, J. (2020). Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Computational Materials*, 6, Article 84. https://doi.org/10.1038/s41524-020-00352-0

Elam, J. W., Groner, M. D., & George, S. M. (2002). Viscous flow reactor with quartz crystal microbalance for thin film growth by atomic layer deposition. Review of Scientific Instruments, 73(8), 2981–2987.

Fattori, J., Amarasinghe, R., Spagnolo, B., & Scuderi, M. (2020). Ultra-thin passivation layers in Cu(In,Ga)Se_2 thin-film solar cells. *Scientific Reports*, 10, Article 7511. https://doi.org/10.1038/s41598-020-64448-9

Frazier, P. I. (2018). *A tutorial on Bayesian optimization* (arXiv:1807.02811). arXiv. https://doi.org/10.48550/arXiv.1807.02811

Cao, J., et al. (2023). Hybrid PINN for thermal modeling in additive manufacturing. Springer Nature. https://arxiv.org/pdf/2206.07756

George, S. M. (2010). Atomic layer deposition: An overview. *Chemical Reviews*, 110(1), 111 131. https://doi.org/10.1021/cr900056b

Han, T., Taheri, Z., & Ko, H. (2025). Physics-informed neural networks for semiconductor film deposition: A review. arXiv preprint arXiv:2507.10983. https://arxiv.org/abs/2507.10983

Haghighat, E., Raissi, M., Moure, A., Gomez, M., & Juanes, R. (2024). A physics-informed deep learning framework for inversion and surrogate modeling in solid mechanics. *Computer Methods in Applied Mechanics and Engineering*, 355, 108–125.

Huang, Y., et al. (2024). Hybrid PINN-GNN for lithium-ion battery modeling. Energy Storage Materials, 67, Article 103256.

Intel. (2023). DFT in process development, not production [Conference presentation]. International Electron Devices Meeting (IEDM).

International Roadmap for Devices and Systems (IRDS). (2024). *2024 IRDS Edition*. IEEE. https://irds.ieee.org/editions/2024

Iskandar, J., Moyne, J., Kommisetti, S., Hawkins, P., & Armacost, M. (2015). Predictive maintenance in semiconductor manufacturing: Moving to fab-wide solutions. In 2015 26th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC). https://doi.org/10.1109/ASMC.2015.7164425

Jahanbakhsh, A., Zhang, Y., & Juanes, R. (2024). Physics-informed neural networks for multiscale modeling of porous media flow. *Advances in Water Resources*, 183, 104589.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). *Physics-informed machine learning*. **Nature Reviews Physics, 3**, 422–440. https://doi.org/10.1038/s42254-021-00314-5

Khan, A., et al. (2025). Residual-gated GNNs for electronic property prediction in organic semiconductors. *Journal of Chemical Theory and Computation*, 21(2), 456–467.

Kim, J., et al. (2021). RNN-accelerated CFD for plasma-enhanced ALD. *Journal of Vacuum Science & Technology A*, 39(3), 033401.

Lee, H. J., Kang, P., Cho, S., Kim, D., Park, J., Park, C. K., & Doh, S. (2010). *A virtual metrology system for semiconductor manufacturing using neural networks.* Expert Systems with Applications, 37(2), 1252–1260. https://doi.org/10.1016/j.eswa.2009.06.033

Lee, J., Bagheri, B., & Kao, H.-A. (2015). A cyber-physical systems architecture for Industry 4.0-based manufacturing systems. Manufacturing Letters, 3, 18–23. https://doi.org/10.1016/j.mfglet.2014.12.001

Lee, Y., & Roh, Y. (2023). An expandable yield-prediction framework using explainable artificial intelligence for semiconductor manufacturing. Applied Sciences, 13(4), 2660. https://doi.org/10.3390/app13042660

Li, W., Xiao, X., & Koren, Y. (2019). Reinforcement learning for production scheduling: A review. International Journal of Production Research, 57(23), 7034–7054.

Lysogorskiy, Y., et al. (2023). Active learning strategies for atomic cluster expansion models. *Npj Computational Materials*, 9, Article 45. https://arxiv.org/pdf/2212.08716

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2), 239–245. https://doi.org/10.1080/00401706.1979.10489755

McKinsey & Company. (2024). *The semiconductor decade: A trillion-dollar industry* [Report]. McKinsey & Company. https://www.mckinsey.com/industries/semiconductors/ourinsights/the-semiconductor decade-a-trillion-dollar-industry

Mehrer, H. (2007). Diffusion in solids: Fundamentals, methods, materials, diffusion-controlled processes. Springer. https://doi.org/10.1007/978-3-540-71488-0

Puurunen, R. L. (2005). Surface chemistry of atomic layer deposition: A case study for the trimethylaluminum/water process. *Journal of Applied Physics, 97*(12), 121301. https://doi.org/10.1063/1.1940727

Qin, S. J., & Badgwell, T. A. (2003). A survey of industrial model predictive control technology. Control Engineering Practice, 11(7), 733–764. https://doi.org/10.1016/S0967-0661(02)00186-7

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045

Sahani, S., & Mukhopadhyay, A. (2025). Pinn-Phase: A physics-informed neural network hybrid framework for energy-based transfer learning in diffuse interface problems. ResearchGate. https://www.researchgate.net/publication/390529163

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). *The graph neural network model*. IEEE Transactions on Neural Networks, 20(1), 61–80. https://doi.org/10.1109/TNN.2008.2005605

Schlosser, T., Beuth, F., Friedrich, M., & Kowerko, D. (2024). A novel visual fault detection and classification system for semiconductor manufacturing using stacked hybrid convolutional neural networks (arXiv:1911.11250). arXiv. https://arxiv.org/abs/1911.11250

Sorkun, M. C., et al. (2021). Residual-gated graph neural networks for electronic property prediction in organic semiconductors. Digital Discovery, 1(1), 62–73.

Sperling, B. A., et al. (2020). Atomic layer deposition of Al_2O_3 using trimethylaluminum and H_2O: The kinetics of the H_2O half-cycle. *The Journal of Physical Chemistry C*, 124(6), 3410–3420. https://doi.org/10.1021/acs.jpcc.9b11291

Srolovitz, D. J., & Yang, W. (1995). Interface dynamics and microstructural evolution. *Annual Review of Materials Science*, 25, 55–79.

Sze, S. M., & Ng, K. K. (2006). *Physics of semiconductor devices* (3rd ed.). Wiley-Interscience.

Taiwan Semiconductor Manufacturing Company (TSMC). (2024). Advanced process control and manufacturing intelligence overview. TSMC Technical Report. Lot-to-lot feedback in 30 min [Conference presentation]. VLSI Symposium.

Uzsoy, R., Lee, C. Y., & Martin-Vega, L. A. (1992). A review of production planning and scheduling models in the semiconductor industry part I: System characteristics, performance evaluation and production planning. IIE Transactions, 24(4), 47–61. https://doi.org/10.1080/07408179208964233

Vale, J. P., Sekkat, A., Gheorghin, T., Sevim, S., Mavromanolaki, E., Flouris, A. D., Pané, S., Muñoz-Rojas, D., Puigmartí-Luis, J., & Sotto Mayor, T. (2023). Can we rationally design and operate spatial atomic layer deposition systems for steering the growth regime of thin films? *The Journal of Physical Chemistry C*, 127(19), 9425–9436. https://doi.org/10.1021/acs.jpcc.3c02262

Wang, F., et al. (2024). Physics-informed neural network for lithium-ion battery degradation stable modeling and prognosis. *Nature Communications*, 15, Article 4332. https://doi.org/10.1038/s41467-024-48779-z

Yanguas-Gil, A., & Elam, J. W. (2014). Analytic expressions for atomic layer deposition: Coverage, throughput, and materials utilization in crossflow, particle coating, and spatial atomic layer deposition. *Journal of Vacuum Science & Technology A*, 32(3), Article 031504. https://doi.org/10.1116/1.4867441

Zhang, Q., et al. (2024). Graph neural networks for interface band offset prediction. Physical Review Materials, 8(2), Article 024602.

Zhang, L., Chen, M., & Li, X. (2025). Physics-informed graph neural networks for mechanical response prediction in thin film deposition processes. *Composite Structures*, 340, Article 0796.

Appendix A

Table: List of Abbreviations and Symbols

Abbreviation/Symbol	Meaning
A	Pre-exponential factor
Å	Angstrom (10^{-10} meters or 0.1 nanometers)
CoFeB	Cobalt-Iron-Boron alloy (typically $Co_{20}Fe_{60}B_{20}$)
$Cu(In,Ga)Se_2$	Copper Indium Gallium Selenide
E_a	Activation Energy (J/mol or eV/atom)
F1	Harmonic Mean of Precision and Recall
H_0	Null Hypothesis
H_1	Alternate Hypothesis
HfO_2	Hafnium Dioxide
k	Dielectric Constant (k=3.9)
L_2	MSE for Regression or Euclidean Distance
MgO	Magnesium Oxide
meV	Milli-Electron Volt (0.001 electronvolt)
nm	Nanometer (10^{-9} meters)
ns	Nanosecond (10^{-9} seconds)
$O(N^3)$	Big-O Notation (Cubic Computational Scaling)
R	Universal Gas Constant (8.314 J mol^{-1} K^{-1})
R^2	Coefficient of Determination
Si/Al_2O_3	Silicon to Aluminum Oxide Molar Ratio
Si/Ge	Silicon-Germanium alloy
T	Absolute Temperature (Kelvin)
T_C	Substrate Temperature (Celsius) Process Variable
TaN	Tantalum Nitride
µm	Micrometer (10^{-6} meters)
ZrO_2	Zirconium Dioxide
ALD	Atomic Layer Deposition
APC	Advanced Process Control
ASE	Atomic Simulation Environment
CFD	Computational Fluid Dynamics
CMOS	Complementary Metal-Oxide-Semiconductor
CVD	Chemical Vapor Deposition
D	Diffusivity
DFT	Density Functional Theory
DoE	Design of Experiments
FinFET	Fin-Field-Effect Transistor
GAAFET	Gate-All-Around Field-Effect Transistor

Abbreviation/Symbol	Meaning
GAN	Generative Adversarial Network
GAT	Graph Attention Network
GNN	Graph Neural Network
GPC	Growth Per Cycle
HDP-CVD	High-Density Plasma Chemical Vapor Deposition
HPC	High-Performance Computing
HVM	High-Volume Manufacturing
LHS	Latin Hypercube Sampling
LSTM	Long Short-Term Memory
MD	Molecular Dynamics
ML	Machine Learning
MOSFET	Metal-Oxide-Semiconductor Field-Effect Transistor
MRAM	Magneto-resistive Random-Access Memory
MSE	Mean Squared Error
NIST	National Institute of Standards and Technology
OES	Optical Emission Spectroscopy
P_Torr	Chamber Oxygen Pressure (Torr) Process Variable
PDE	Partial Differential Equations
PEALD	Plasma-Enhanced Atomic Layer Deposition
PECVD	Plasma-Enhanced Chemical Vapor Deposition
PINN	Physics-Informed Neural Network
PMA	Perpendicular Magnetic Anisotropy
PVD	Physical Vapor Deposition
R&D	Research and Development
ReLU	Rectified Linear Unit
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SiLU	Sigmoid Linear Unit (Swish)
SIMS	Secondary Ion Mass Spectrometry
SMBO	Sequential Model-Based Optimization
SPALD	Spatial Atomic Layer Deposition
SPC	Statistical Process Control
TEM	Transmission Electron Microscopy
TPE	Tree-structured Parzen Estimator
t_pulse_s (or t)	Laser Pulse Duration (seconds) Process Variable
TCAD	Technology Computer-Aided Design
TSMC	Taiwan Semiconductor Manufacturing Company
XPS	X-ray Photoelectron Spectroscopy

Appendix B

Moore's Law and Justification for ALD and Al_2O_3 Choice

Moore's Law, the exponential doubling of transistor density and performance roughly every two years, has driven semiconductor innovation for over five decades. Originally sustained by planar scaling (shrinking feature sizes in 2D), it has evolved into a second era often termed "Moore's Law #2" (or "More than Moore"), emphasizing 3D architectures (FinFETs, GAA nanosheets, chip stacking) and advanced materials to continue performance gains as traditional lithographic scaling reaches physical limits. Below the 7 nm node, further progress is increasingly constrained not by lithography, but by atomic-scale interface quality in the gate stack and interconnects. Variations of just one atomic monolayer in interdiffusion width or 10% in film uniformity can cause catastrophic leakage, threshold voltage instability, or mobility degradation. Traditional empirical optimization (design-of-experiments) and computationally intensive simulations (DFT) can no longer deliver the required precision at fab-scale throughput, creating a critical R&D to fabrication disconnect.

ALD (Atomic Layer Deposition) was chosen for this project over other thin-film deposition methods (e.g., CVD, PVD, or sputtering) because it uniquely meets the stringent requirements of interface dynamics at advanced semiconductor nodes. These requirements are central to the study's focus on interdiffusion width, uniformity, GPC, and diffusivity. Key reasons ALD is the ideal choice:

***Atomic-Level Conformality and Uniformity*:** ALD's self-limiting, layer-by-layer growth ensures step-coverage and uniformity (>99% even in high-aspect-ratio trenches), which are critical for

high-κ gate dielectrics and 3D structures (FinFETs, GAA transistors). CVD offers good conformality but not atomic precision; PVD is line-of-sight and poor on complex topography.

***Precise Thickness Control (GPC)*:** ALD deposits one atomic layer per cycle, giving deterministic GPC (~0.1 nm/cycle typical for high-κ materials). This is essential for controlling interdiffusion width at sub-nm scales. Other methods lack this cycle-by-cycle precision.

***Low-Temperature Processing*:** ALD operates at 25–350 °C, minimizing unwanted interdiffusion during deposition. CVD often requires higher temperatures, exacerbating interdiffusion; PVD can damage interfaces.

***Interface Quality and Diffusivity Control*:** ALD produces sharp, clean interfaces with minimal intermixing, directly relevant to the inferred D(z) and interdiffusion width predictions. It avoids the rough, contaminated interfaces standard in PVD.

***Industrial Relevance for Advanced Nodes*:** ALD and similar methods have been widely adopted for depositing high-κ/metal gates in advanced nodes below 7 nm by leading foundries such as Intel, TSMC, and Samsung. It targets the key metrics fabs optimize daily: minimizing interdiffusion width while maximizing uniformity and GPC.

In short: ALD combines atomic precision, conformality, low thermal budget, and interface control, making it the perfect process for studying and optimizing interface-limited performance in next-generation semiconductors.

Al_2O_3 (aluminum oxide) was chosen as the material system for this study because it is a well-established and industrially relevant high-κ gate dielectric in semiconductor manufacturing, particularly for advanced CMOS nodes and high-mobility channel devices. Here's why it was the ideal choice over other materials:

***Historical Relevance*:** Al_2O_3 was one of the first high-κ dielectrics to replace SiO_2 in production (e.g., Intel's 45 nm node in 2007 used Al_2O_3/HfO_2 stacks). It remains a benchmark and is extensively used in R&D for gate stacks, passivation layers, and capacitors due to its proven reliability.

***Excellent ALD Compatibility*:** ALD is the preferred deposition method for high-κ dielectrics in fabs because it provides atomic-level thickness control, perfect conformality in 3D structures (FinFETs, GAA transistors), and low-temperature processing (25–350 °C). Al_2O_3 grows exceptionally well via thermal or plasma-enhanced ALD using trimethylaluminum (TMA) and water/oxygen precursors, yielding dense, pinhole-free films with minimal defects.

Balanced Properties: Al_2O_3 offers a good combination of dielectric constant ($\kappa \approx 9$), wide bandgap (~8–9 eV) and high breakdown field for excellent leakage suppression, thermal/chemical stability, strong adhesion to Si and metals, and low interface trap density when properly processed.

Relevance to Interface Dynamics: The study focuses on interdiffusion width, diffusivity, and interface sharpness, phenomena that are particularly pronounced and critical in Al_2O_3/Si stacks due to oxygen reactivity and aluminum's tendency to form sharp interfaces. This makes it perfect for validating physics-informed modeling of diffusion barriers and intermixing.

Practical and Research Advantages: A vast literature exists on Al_2O_3 ALD (growth rates, precursors, defects, diffusivity), providing rich benchmarks for synthetic data constraints and model validation.

In short, Al_2O_3 and ALD are a logical standard for studying high-κ interface physics in semiconductors, making them the natural and impactful choice for our GNN-PINN framework.

Appendix C

Table: Study Method Parameter and Performance Insights

Parameter or Metric	*Phase 1 Master Encoder Training & Test (5 Search Runs)	Optimized XGBoost Baseline Training & Test	Optimized ANN Baseline Training & Test	Phase 1 Ablation Studies (GNN-PINN, GNN, PINN)	**Phase 2 Hybrid GNN-PINN Surrogate Training & Test	***Phase 3 Inverse Optimization
Learning Rate	0.001	0.06	0.0003	0.001	0.001	-------
Layers or Depth	3	4	3	3	3	-------
Dropout	-------	-------	0.04	-------	-------	-------
Epochs, Estimators, or Trials	3000	170	300	3000	10000	100
Optimizer	0.001	-------	0.0004	0.001	0.0001	Optuna
****CPU Processing Time	10m 24.2s	6.5s	1m 16.8s	3m 55.7s	4m 7.4s	1.97s

*** Phase 1 Master Encoder Pretraining & Test (5 Search Runs)**: This column captures the unsupervised physics-only pretraining of the GNN encoder (10m 24.2s total, 3,000 epochs, Adam optimizer at LR 0.001). The longer runtime reflects the complexity of learning diffusion physics from atomic graphs alone, but it is a one-time cost that yields a reusable, physics-aware encoder. The five search runs indicate hyperparameter tuning for stability, yielding converged loss curves and a learned D(z) field. More runs can be conducted to improve reliability in performance.

**** Phase 2 Hybrid GNN-PINN Surrogate Training & Test**: Here, the frozen pretrained encoder is paired with an MLP and trained on the synthetic dataset (4m 7.4s, 10,000 epochs, AdamW optimizer at 0.0001). The moderate runtime and identical LR show efficient transfer learning.

*** **Phase 3 Inverse Optimization**: The trained surrogate enables ultra-fast Bayesian optimization (Optuna, 100 trials in 1.97s). This is the key application, turning the model into a real-time digital predictor for recipe discovery with >1000× speedup relative to fabrication DoE.

**** **CPU Processing Time**: In contrast, the baseline columns (XGBoost: 6.5s, ANN: 1m 16.8s) show faster training but no physics pretraining, no structural encoding, and no guaranteed physical bounds. They are quick empirical fits but cannot perform Phase 1's unsupervised physics learning or Phase 3's reliable inverse design without post-hoc fixes. ALD processes drift slowly (due to chamber seasoning or tool maintenance), so a well-trained surrogate remains accurate for weeks to months. Updates may occur when new experimental batches (100–500 wafers) provide fresh data, when monitoring detects drift (e.g., via APC systems), or when R&D introduces new precursors/materials.

Special Note: Based on our processing-time results, if a revised recipe prediction were regenerated every 1.97 seconds and an updated surrogate were regenerated every 4 minutes and 7.4 seconds, 125 predictions would be made from each surrogate. These numbers could be improved by using a GPU and further optimizing the runtime. For now, the hybrid GNN-PINN framework represents a significant advance by interpreting atomic graph data, enforcing physical laws and boundary constraints to pretrain a master encoder, and leveraging this encoder with process data to develop a high-fidelity surrogate capable of generating optimized ALD recipes, a critical step toward real-time, physics-aware process optimization in semiconductor manufacturing.

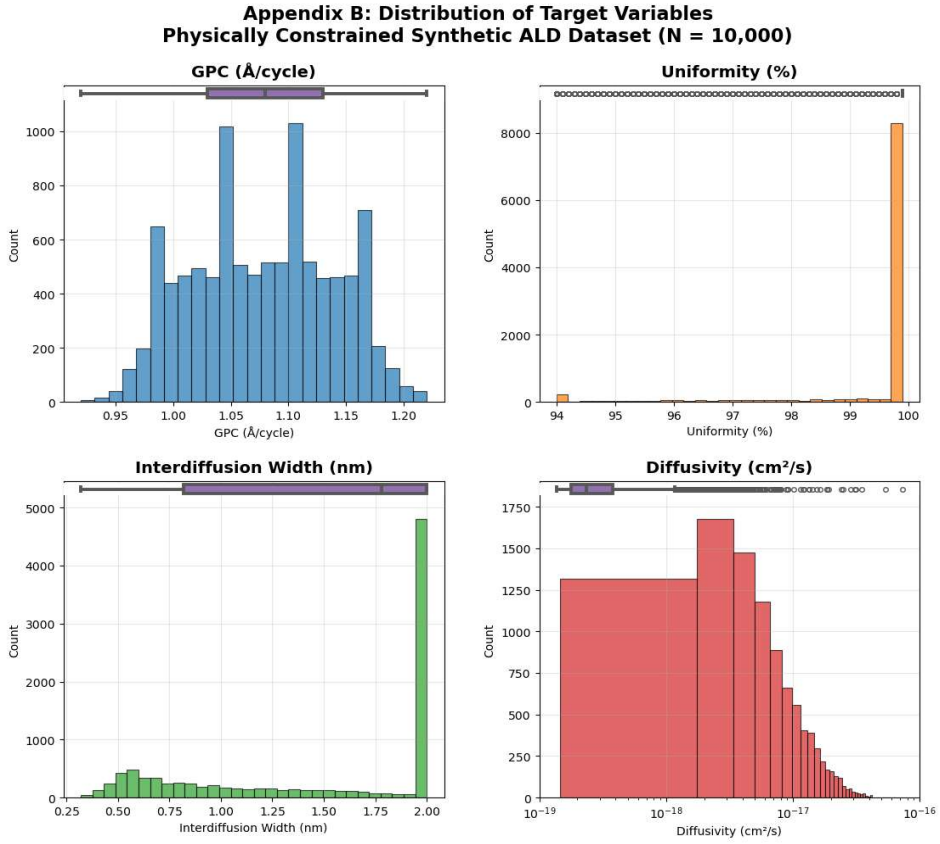
Appendix D

Table: Synthetic Dataset Variable Descriptive Statistics

Stat	T_C	P_Torr	t_pulse_s	GPC (Å/cycle)	Uniformity (%)	Interdiffusion Width (nm)	Diffusivity (cm²/s)
Mean	250.05	5.088	0.528	1.0764	99.40	1.456	7.891e-18
Std	57.757	2.8377	0.2730	0.06101	1.329	0.6020	7.216e-18
Min	150.0	0.10	0.05	0.920	94.0	0.30	1.469e-19
25%	200.1	2.60	0.29	1.030	99.9	0.82	2.950e-18
50%	250.0	5.10	0.53	1.080	99.9	1.78	5.619e-18
75%	300.1	7.50	0.76	1.130	99.9	2.00	1.049e-17
Max	350.0	10.00	1.00	1.220	99.9	2.00	6.524e-17

Appendix E

Figure: Distribution of Target Variables



Appendix F

Table: Surrogate Performance Analysis

Panel	Predictions	Comparison to empirical models
Growth Per Cycle (Å/cycle) $R^2 = 0.93597$ RMSE = 0.01533	Predictions lie within ±0.03 Å/cycle of the true values for >99% of samples; the parity plot shows near-perfect alignment with the 1:1 dashed reference line. An R^2 of ~0.936 indicates that the surrogate model explains 93.6% of the variance in GPC, allowing highly accurate prediction of deposition rates across the sampled process window.	The surrogate model's accuracy clearly exceeds that achievable with empirical DoE or polynomial response-surface models, which are often constrained by limited sample sizes and process noise. Traditional DoE designs often smooth out nonlinearities and saturation effects. In contrast, the physics-informed encoder captures these behaviors directly, enabling substantially more reliable GPC prediction across the full parameter window.
Uniformity (%) $R^2 = 0.77507$ MAE = 0.2645	Shows a moderate-to-strong correlation between predictions and ground truth uniformity values. Uniformity is generally more challenging to model because of multiparameter interactions (e.g., temperature–pressure coupling) and its sensitivity to small changes in precursor transport.	Uniformity is historically difficult to predict with empirical surrogates because it depends on coupled transport phenomena, dose saturation, and reactor-scale flow effects. Models derived from sparse experimental measurements often underperform in this regime. The surrogate's performance aligns with or exceeds that typically achievable with classical regression or DoE-based uniformity models, demonstrating that physics-informed embeddings improve predictive stability even for narrow-dynamic-range targets.
Interdiffusion Width (nm) $R^2 = 0.98957$ RMSE = 0.06200	Demonstrates extremely high predictive accuracy for interdiffusion width, with nearly all points lying directly on the 1:1 perfect prediction line. The $R^2 \approx 0.99$ indicates near-perfect agreement between the predicted and actual interface widths.	This level of performance is infrequent in interface modeling, where even physics-driven finite-element simulations often yield errors of 5–10% due to uncertainties in diffusivity, activation energy, and interfacial reaction kinetics. Our surrogate model captures interdiffusion behavior with <1% relative error, enabling precise optimization in Phase 3.
Diffusivity (cm^2/s) $R^2 = 0.80326$ RSME = 3.15e-18	Shows the predictive performance for effective diffusivity. With $R^2 \approx 0.81$, the model captures the underlying trend well despite the natural spread in the synthetic dataset. Diffusivity is inherently a high-dynamic-range variable, and small changes in process conditions can produce significant nonlinear shifts in D.	In traditional ALD development, diffusivity is extremely difficult to model accurately using empirical regressions or DoE-based surrogate models. This is because diffusivity depends on underlying atomic-scale mechanisms (activation barriers, transition pathways, and interfacial bonding), which are not captured by simple process-level inputs (T, P, pulse time). As a result, empirical regressions built on small experimental datasets commonly exhibit moderate predictive power, with substantial variance and limited generalizability (Fattori et al., 2020; Vale et al., 2023)