

RESEARCH ARTICLE

Multi-source data-driven model to predict the deposition rate of CVD-SiC coatings with improved accuracy and robustness

Ruizhao Liu¹ | Chaokun Song² | Zhenya Lu¹ | Kang Guan¹  | Xudong Wang² |
Pei Cao² | Xiaohui Yang²  | Longteng Bai²

¹School of Materials Science and Engineering, South China University of Technology, Guangzhou, China

²Xi'an Aerospace Propulsion Institute, Xi'an, China

Correspondence

Zhenya Lu and Kang Guan, School of Materials Science and Engineering, South China University of Technology, Guangzhou, 510640, China.

Email: zhylu@scut.edu.cn and mshguan@scut.edu.cn; kanguan123@gmail.com

Funding information

Major Basic Research Projects on Equipment, Grant/Award Number: 909010106-301; National Natural Science Foundation of China; 52472096; Guangdong Basic and Applied Basic Research Foundation, Grant/Award Numbers: 2023A1515012156, 2024A1515011656; Funding of Xi'an Aerospace Propulsion Institute, Grant/Award Number: 040020006120250054

Abstract

Chemical vapor deposition (CVD) is widely used for the preparation of silicon carbide (SiC) coatings. The growth of controllable SiC coatings at desired efficiencies is of critical importance in industrial production. In this study, we developed a multi-source data-driven model to predict the deposition rate of CVD-SiC coatings. To address the challenge of missing values in the training data, we introduced a simulation-based filling (SBF) method. The correlation between parameters and the deposition rate was analyzed using machine learning (ML) techniques. Our results demonstrate that leveraging a multi-source dataset reduces the prediction error from 12% to 3% and improves the generalizability beyond boundary conditions. Comprehensive Sobol sensitivity analysis quantitatively confirmed temperature as the dominant parameter, accounting for 63% of output variance directly and 78% when including interactions. The incorporation of multi-source dataset improves the accuracy and robustness of the ML model, which also is beneficial in optimization efficiency. The framework presented here highlights the potential of data-driven approaches for guiding the CVD-SiC process, offering a novel perspective for future research.

KEYWORDS

chemical vapor deposition (CVD), data driven, deposition rate, machine learning (ML), silicon carbide (SiC)

1 | INTRODUCTION

Silicon carbide (SiC), a third-generation semiconductor material, exhibits considerable potential use in power electronic devices and the aerospace industry due to its exceptional thermal conductivity, electron mobility, breakdown voltage, and chemical stability.^{1–5} The capability of SiC to maintain performance under extreme conditions, such as high temperatures and corrosive environments, significantly enhances the reliability and efficiency of

components in demanding applications.^{6–9} To fully leverage these remarkable properties, it is crucial to produce high-quality SiC coatings and films.

Chemical vapor deposition (CVD) has become the preferred method for preparing SiC due to its ability to provide uniformity, high purity, and precise stoichiometry control.^{10,11} Zhao et al.¹² prepared a controllable SiC coating on carbon fiber (CF) through low-pressure chemical vapor deposition (LPCVD), which significantly reinforced hydroxyapatite (HA) composites

(SiC-CF/HA). When 0.5 wt% of SiC-CF was added, the bending strength increased to 28.4 MPa, nearly three times that of pure HA sintered under identical conditions. Wen et al.¹³ showed that CVD-SiC coating played a crucial role in enhancing the flexural strength of PIP-C/SiC composites. In the absence of coating, PIP-C/SiC composites lost 29.2% of their mass, retaining only 6.74% of the original flexural strength. In contrast, the coated composites retained 10.2% of their mass and 23.4% of their flexural strength. Zhan et al.¹⁴ conducted mechanical tests on continuous silicon carbide fiber-reinforced silicon carbide ceramic (SiC_f/SiC) composite tubes, with and without external CVD-SiC coatings, using a C-shaped annular specimen geometry. They concluded that CVD-SiC coatings and the interface between the coatings and the SiC_f/SiC composites effectively protect the composites, improving annular tensile strength and maintaining the integrity of the wound tube. Baux et al.¹⁵ reported that pure and dense CVD-SiC coatings substantially improve the mechanical strength, oxidation resistance, and thermal diffusivity of the material applied in solar receivers. However, the CVD process often involves long processing times and high production costs, creating significant challenges in process optimization and reproducibility.^{16,17} These limitations stem from the process's inherent complexity, where multiple variables—such as temperature, pressure, and gas flow rates—interact in intricate ways to influence the final product's characteristics.^{18,19} Given these challenges, there is a pressing need for more efficient and reliable methods to optimize the SiC deposition process.

Multi-physics field simulation promotes enhanced CVD process efficiency and product quality. Bijjargi et al.²⁰ presented a detailed 3D model consisting of the compressive transport models and the reaction mechanism to investigate the SiC film growth over a fuel combustion nozzle in the commercial hot-wall CVD reactor. Ogawa et al.²¹ reported a computational fluid dynamics (CFD) simulation of CVD-SiC in the CH₃SiCl₃/H₂ system, and the MTS decomposition model was coupled with the SiC growth model. Their coupled model successfully reproduced multiple datasets through validation studies. Cao et al.²² optimized the CVD process by CFD simulation. The effects of graphite plate placement, diameter, deposition temperature, reaction gas flow, and graphite plate rotational speed on the deposition rate, deposition uniformity, surface morphology, and cross-section morphology of the SiC coating were thoroughly studied. Zou et al.²³ investigated an industrial-scale multi-substrate CVD reactor for producing coated glass products using a porous media-based modeling approach. The simulation results showed that the gas reactant mixing in the reactor could be improved by implementing a top-injection scheme,

introducing natural convection, and redistributing the substrate surfaces. However, multi-physics field simulations are computationally expensive and often rely on expert knowledge. Therefore, developing a simple yet effective model for process optimization is crucial.

Machine learning (ML) excels at uncovering complex patterns and relationships within large datasets, and the trained model can be employed for the parameter optimization, making it especially suitable for the analysis and optimization of processes like CVD.^{24,25} Kimaev et al.²⁶ used trained artificial neural networks (ANNs) to model the roughness and growth rate of thin films produced by CVD, achieving orders-of-magnitude computational savings. Rajak et al.²⁷ applied offline reinforcement learning (RL) to predict optimal synthesis schedules for monolayer CVD-MoS₂. The RL agent, trained on 10,000 computational synthesis simulations, learned the threshold temperatures and chemical potentials for the onset of chemical reactions after and predicted previously unknown synthesis schedules that successfully produced well-sulfidized crystalline, phase-pure MoS₂. Xu et al.²⁸ implemented supervised ML for the CVD synthesis of high-quality quasi-1D few-layered WTe₂ nanoribbons, demonstrating the effectiveness of ML in guiding the synthesis of 1D nanostructures. Tang et al.²⁹ developed a classification ML model for chemical-vapor-grown MoS₂ that optimized synthesis conditions to achieve a higher success rate. A major challenge in applying ML to CVD processes is the scarcity of comprehensive training data, which can hinder the development of accurate and reliable predictive models. Costine et al.³⁰ manually constructed a database from literature data on MoS₂ thin films grown via CVD. ML techniques were then applied to extract underlying trends from the data, facilitating the formation of MoS₂ monolayers. To overcome the challenge of limited data, a multi-source approach to data collection was adopted.

The theoretical foundation for applying ML to materials processing prediction tasks relies on the principle that material behavior follows physical laws that create recognizable patterns in experimental data. While traditional physics-based modeling requires explicit formulation of these laws, data-driven approaches can implicitly extract these patterns without requiring complete mechanistic understanding. The efficacy of such approaches depends critically on data representativeness, feature selection validity, and model architectural suitability. In CVD processes specifically, the complex interplay between transport phenomena, reaction kinetics, and thermodynamics creates multidimensional parameter spaces that are particularly amenable to ML exploration, provided sufficient data quality and domain-knowledge-guided feature engineering are implemented.

This work utilizes a multi-source, data-driven ML approach to explore the complex relationship between the process parameters and the deposition rate of CVD-SiC. The training data were sourced from a combination of literature data and numerical simulation outcomes based on a limited set of experimental data. To address the missing values of literature data, we introduced a simulation-based filling (SBF) method. After feature engineering, model training, and hyperparameter optimization, a ML model capable of target prediction was developed. Feature importance was extracted from the trained model, which can be further applied for mechanistic analysis and process optimization of CVD-SiC. The multi-source data-driven ML model combines the accuracy of numerical simulation with the efficiency and robustness of ML, providing new insights into the exploration of CVD-SiC.

2 | DATASET

The process parameters can be broadly classified into precursor, reactor, substrate, and growth parameters (Figure 1A). These parameters collectively describe the entire CVD process and can serve as inputs for an ML model. As an efficient data analysis method, ML plays an increasingly important role in contemporary research. The accuracy of a ML model depends on the quantity and quality of the training data. However, the prohibitive costs and extended time scale of CVD-SiC experiments have resulted in a lack of sufficient training data, posing a significant obstacle to the development of ML-based data analysis model.

To address the aforementioned challenges, a multi-source dataset was constructed (Figure 1B). A total of 3345 data points were extracted from open-source literature.^{31–68} Given the inherent variability in research content and experimental equipment, the presence of missing features is inevitable. Incomplete features can degrade model performance, making this another issue that needs to be addressed. To tackle this issue, we proposed a simulation-based filling (SBF) method for missing value imputation. The implementation steps are as follows:

1. conducting typical CVD-SiC experiments to obtain experimental data;
2. modelling of CVD-SiC reactor and optimization of model parameters based on experimental data;
3. parametric scanning of the model to obtain simulation data with complete features;
4. missing value filling based on simulation data.

2.1 | Experiment

Honeycomb graphite blocks with a diameter of 20 mm and a thickness of 10 mm were employed as the deposition substrate. The blocks were polished, ultrasonically cleaned, and dried before use. MTS was selected as the source gas, with H₂ as the carrier gas, and Ar as the diluent. MTS was then heated and vaporized into the reaction chamber with H₂. The reaction chamber was cylindrical in shape, with dimensions of $\Phi_{in}42 \times \Phi_{out}46 \times 60$ mm. The deposition temperature varied from 1573 to 1673 K, pressure from 6 to 24 kPa, and α from 5 to 20. The reactor equipment diagram of CVD is shown in Figure 2.

2.2 | Multi-physics simulation

The epitaxial growth of SiC is a complex process that exhibits a dependence on various factors, including the temperature, flow, and chemical reaction. Consequently, a multi-physics simulation was conducted, employing a thermal-fluid-chemical model and the following equations: continuity, momentum, energy, and species transport.

The continuity equation and momentum equation are as follows:

$$\nabla \cdot (\rho \mathbf{u}) = 0 \quad (1)$$

$$\begin{aligned} \rho (\mathbf{u} \cdot \nabla) \mathbf{u} \\ = -\nabla p + \nabla \cdot \left[\mu \left(\nabla \mathbf{u} + (\nabla \mathbf{u})^T \right) - \frac{2}{3} \mu (\nabla \cdot \mathbf{u}) \mathbf{I} \right] + \mathbf{F} \end{aligned} \quad (2)$$

where ρ is the mass density, $\mathbf{u}$ is the gas mixture velocity, p is the total fluid pressure, μ is the dynamic viscosity, $\mathbf{I}$ is the unit tensor, and $\mathbf{F}$ represents the gravity.

The energy equation is expressed as follows:

$$\rho C_p \mathbf{u} \cdot \nabla T + \nabla \cdot \mathbf{q} = Q \quad (3)$$

$$\mathbf{q} = (-k \nabla T) \quad (4)$$

where C_p is the fluid heat capacity, T is the deposition temperature of the fluid, $\mathbf{q}$ is the Fourier heat conduction, k is the thermal conductivity, and Q is the internal heat source term, including reaction heat, radiant heat, and other user-defined terms.

Species transport equation can be expressed as follows:

$$\nabla \cdot \mathbf{j}_i + (\mathbf{u} \cdot \nabla) c_i = R_i \quad (5)$$

$$\mathbf{j}_i = -D_i \nabla c_i \quad (6)$$

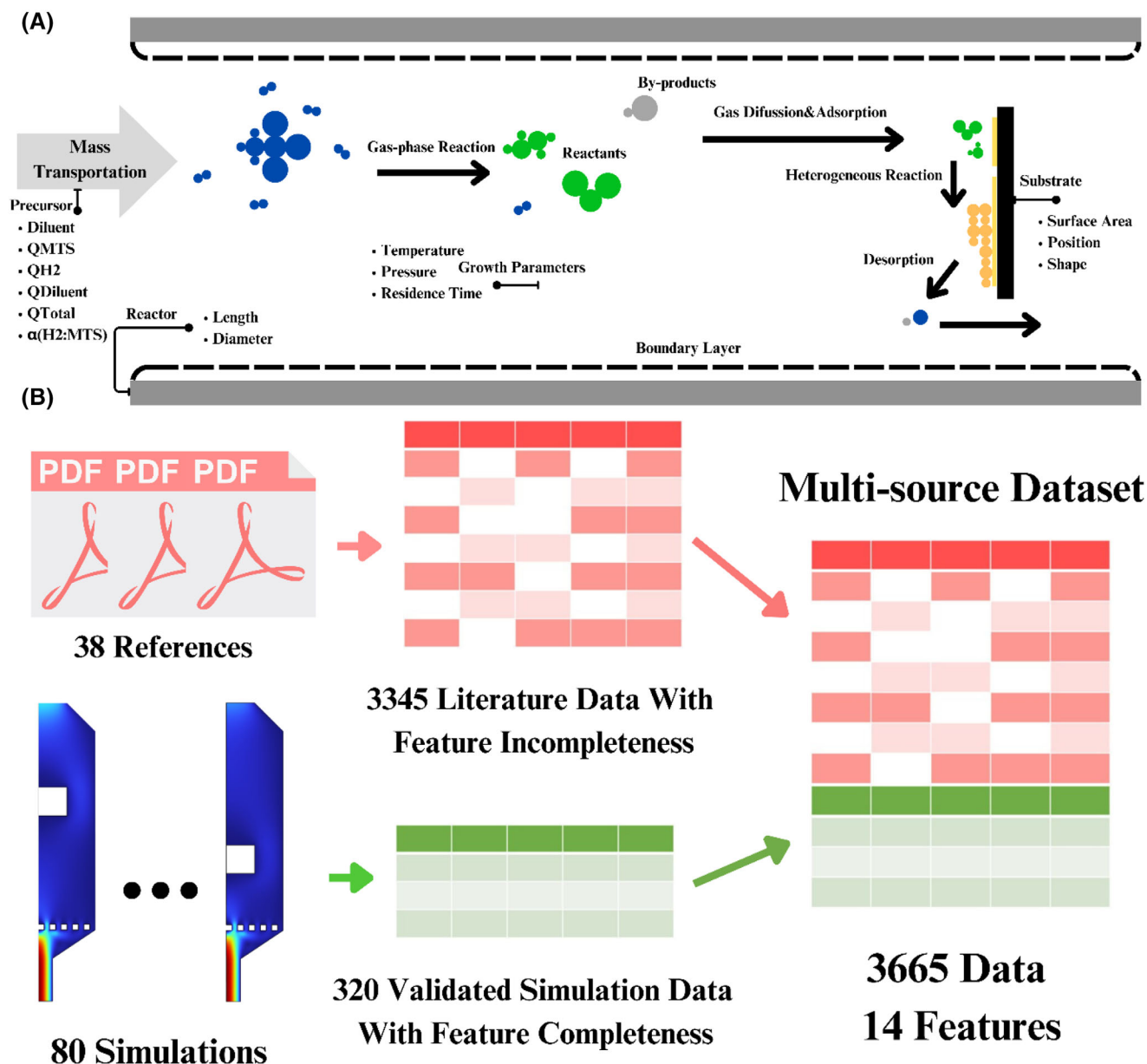


FIGURE 1 (A) Schematic of the chemical vapor deposition (CVD) process; (B) establishment of multi-source dataset.

where D_i , ∇c_i , and R_i represent the diffusion coefficient, molar concentration, and reaction rate of gas component i , respectively.

The density of the mixed gas is expressed by the ideal gas equation:

$$\rho = \frac{PM}{RT} \quad (7)$$

$$M = \sum_{i=1}^N x_i M_i \quad (8)$$

where M is the molar mass of the mixed gas.

At the substrate–gas interface, we implemented comprehensive boundary conditions to accurately capture the physical phenomena. For momentum transport, a no-slip wall condition ($u = 0$, $v = 0$) was imposed. For species transport, we applied a flux boundary condition coupling surface reactions with bulk diffusion: $-D_i \nabla c_i = R_i$, where R_i represents the temperature-dependent surface reaction rate. To ensure accurate resolution of the steep thermal and concentration gradients near the substrate surface, we employed boundary layer mesh refinement validated through grid independence studies.

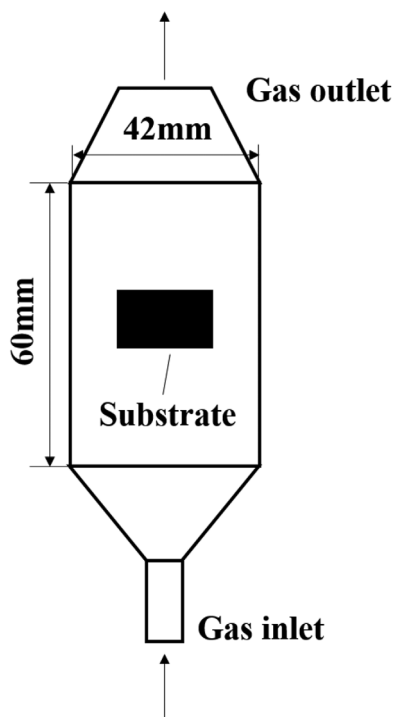


FIGURE 2 Chemical vapor deposition (CVD) reactor equipment diagram.

2.3 | Imputation

A variety of imputation methods were employed to address missing values, including simple imputation, multiple imputation by chained equations (MICE),^{69–71} linear regression imputation,⁷² K-Nearest Neighbor (KNN) regression imputation,⁷³ and random forest regression imputation.⁷⁴

In the case of simple imputation, missing values are filled using measures of central tendency, such as the mean, median, or mode. MICE is a Bayesian-based method that employs a series of iterative predictive models to impute missing values. In this approach, each iteration uses the remaining variables in the dataset to estimate the variable with missing values, continuing until convergence. Linear regression imputation addresses missing values by creating a linear combination of features, aiming to minimize the error between imputed and actual values through the identification of an optimal line or plane/hyperplane. KNN regression imputation calculates the distances between the sample in the descriptor hyper-space and the training data, determining the target value as the average of the *k*-nearest neighbors. Random forest regression imputation constructs multiple decision trees, which are then aggregated through weighted averaging to determine the final target value. It should be noted that regression-based imputation methods are only effective when a complete feature dataset is available.

The statistical validity of each imputation method was carefully evaluated. Among iterative methods, MICE was implemented with 300 iterations to ensure convergence, given the moderate dataset size and our feature engineering approach. The error propagation between imputation methods and subsequent modeling was systematically analyzed through a comparative evaluation matrix of all 63 possible combinations (7 imputation methods $\times$ 9 training models), with performance assessed via fivefold cross-validation to ensure robust evaluation.

3 | MACHINE LEARNING

In addition to the previously mentioned regression methods—linear regression,⁷⁵ KNN regression, and random forest regression, several other commonly used regression algorithms were employed in this study.^{76,77} One such method is support vector regression (SVR),⁷⁸ which constructs a model of minimal complexity by introducing the ϵ -insensitive loss function. This approach ensures that the most prediction errors fall within the ϵ -range, focusing only on errors that exceed this threshold. Ensemble learning methods such as AdaBoost and gradient boosting decision trees (GBDTs) are widely used to enhance the accuracy of regression tasks.⁷⁹ AdaBoost works by adjusting the weights of samples, enabling subsequent models to concentrate more on difficult-to-predict samples, ultimately combining multiple weak learners into a strong learner. In contrast, GBDT employs a gradient boosting strategy, where decision tree incrementally fits the residuals of the previous models, and the results of all trees are weighted collectively to yield the final prediction. Both methods aim to generate more powerful models by augmenting weak learners to improve regression accuracy. Neural networks (NNs),⁷⁰ with their multi-layer architecture comprising input, hidden, and output layers, automatically learn complex nonlinear relationships between input features and output values. By using forward propagation to calculate predictions and evaluating errors with a loss function, NNs adjust weights through backpropagation and gradient descent. This iterative optimization process minimizes errors, facilitating the efficient prediction of continuous target values.

For handling potential collinearity between input parameters, particularly between gas flow rates, we implemented method-specific approaches. Linear models utilized regularization techniques (Ridge and Lasso) to mitigate multicollinearity effects. NN models incorporated Dropout layers to prevent overfitting to potentially collinear features. Tree-based models naturally mitigate collinearity concerns through their splitting mechanisms that implicitly perform feature selection.

TABLE 1 The models employed for training and their hyper-parameters.

Model	Hyper-parameters
Ridge	"alpha": (1e-2, 1e3)
Lasso	"alpha": (1e-2, 1e2)
RandomForestRegressor	"n_estimators": (10, 200), "max_features": ["sqrt", "log2", None], "max_depth": (3, 20), "min_samples_split": (2, 11), "min_samples_leaf": (1, 5)
SVR	"C": (1e-1, 1e2), "kernel": ["rbf", "linear"], "gamma": (1e-4, 1e0)
KNeighborsRegressor	"n_neighbors": (1, 20), "weights": ["uniform", "distance"], "p": [1, 2] # Manhattan or Euclidean distance
AdaBoostRegressor	"n_estimators": (50, 300), "learning_rate": (1e-3, 1), "loss": ["linear", "square", "exponential"]
GradientBoostingRegressor	"n_estimators": (50, 300), "learning_rate": (1e-3, 1), "max_depth": (3, 15), "subsample": (0.5, 1), "min_samples_split": (2, 21), "min_samples_leaf": (1, 11)
PyTorchRegressor	"hidden_dim": (16, 256), "hidden_layers": (1, 5), "activation": ["relu", "tanh"], "optimizer": ["adam", "rmsprop"], "learning_rate": (1e-4, 1e-2), "batch_size": (16, 256), "epochs": (2000, 4000)

The selection of hyperparameters, including maximum depth, learning rate, and network architecture, directly affects model performance. Hyperparameter optimization helps identify the optimal configuration that maximizes the model's learning potential. A systematic search and adjustment of the hyperparameters can effectively mitigate overfitting and underfitting, thereby enhancing the model's generalization capacity and improving its predictive accuracy. In this study, hyperparameter optimization was conducted via random search, which involves constructing models based on randomly sampled combinations of hyperparameters across the hyperparameter space. The models employed for training and their hyper-parameters are shown in Table 1.

The dataset is randomly shuffled before splitting to ensure a more uniform distribution of data points. This

step helps mitigate the impact of any inherent biases or non-uniform patterns in the dataset. After shuffling, the dataset is split into training, validation, and test sets using a random partitioning method. To further enhance the robustness of the model evaluation, cross-validation is employed. In cross-validation, the dataset is divided into multiple folds (e.g., five-fold), and the model is trained and validated iteratively on different subsets of the data. This approach provides a more reliable estimate of the model's performance by reducing the influence of any specific data split and ensuring that the evaluation is representative of the entire dataset. By combining random shuffling, random partitioning, and cross-validation, the process significantly reduces the risk of non-uniform data distribution affecting the model's performance and generalizability.

A total of 50 training iterations were conducted, with the results evaluated using five-fold cross-validation to assess the model performance. Mean absolute percentage error (MAPE) and root mean square error (MSE) were employed to assess the closeness of prediction values and measured values.

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^N \frac{|E_i - P_i|}{E_i} \quad (9)$$

$$\text{RMSE} = \sqrt{\frac{100}{n} \sum_{i=1}^N (E_i - P_i)^2} \quad (10)$$

where E_i is the measured experimental value, P_i indicates the ML model predicted result, and n refers to the total number of sample points.

Parameter analysis was conducted using the Mantel test and SHapley Additive exPlanations (SHAP). The Mantel test is a statistical method used to determine if there is a non-random association between two sets of variables.^{80–83}

This test involves the following steps:

1. construct distance matrices for each of the two sets of variables;
2. compute the correlation between these distance matrices using Pearson's correlation coefficient or Spearman's rank correlation coefficient;
3. randomly shuffle the row and column labels of one distance matrix and recalculate the correlation with the other matrix. Repeat this process multiple times to generate a distribution of correlation coefficients under random conditions;
4. compare the original correlation coefficient (from step 2) with the distribution of random correlation coefficients. If the original coefficient is significantly higher than most of the random coefficients

(Mantel's $p < 0.05$), it indicates a significant association between the two distance matrices.

SHAP is a game theory-based method used to explain the contributions of individual features to the predictions of an ML model.^{84–87} The basic idea of SHAP is to use Shapley values to quantify the impact of each feature on the output of the model. A positive SHAP value indicates that the feature has a positive influence on the prediction, while a negative SHAP value indicates a negative influence. The SHAP value Φ_i of a feature x_i can be calculated using the following formula:

$$\Phi_i = \sum_{S \subseteq N[i]} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (11)$$

where N is the set of all features and $f(S)$ is the predicted value of the model on a subset S .

Inverse parameter optimization (IPO) in this study focuses on identifying optimal input parameters to maximize the predicted output of a pre-trained ML model, reversing the conventional predictive modeling workflow. Unlike forward prediction models, IPO searches the input parameter space for combinations that optimize deposition rate. The selected model, chosen through R^2 -based evaluation, ensures robust predictive performance. The algorithm incorporates dynamic iteration scaling that increases search intensity when performance improves by over 70% and stagnation detection, with termination criteria set at either reaching a maximum iteration or 10 consecutive iterations without significant improvement. A random search strategy is employed over gradient-based methods due to its ability to handle mixed continuous and discrete parameters, avoid local maxima in non-convex optimization through stochastic sampling, maintain compatibility with black-box models, and eliminate gradient computation challenges, ensuring stability and flexibility across diverse model architectures.

4 | RESULTS AND DISCUSSION

The distribution of target data (deposition rate in log scale) for the original dataset is shown in Figure 3. The histogram demonstrates a skewed distribution with the mean (24.19) greater than the median (6.25), confirming the right-skewed nature of the data. The standard deviations and extreme values indicate considerable variability, suggesting a more dispersed distribution. The Kernel Density Estimation (KDE) curve and the cumulative distribution function both indicate that the target data are more densely populated in the low value region (to the left), indicating that the majority of data points are clustered at smaller

deposition rates. However, a few higher outliers raise the mean values.

The training results obtained using different regression models after simple imputation (mean, mode, and median) are presented in Figure 4. The model evaluation metrics are the R^2 values after fivefold cross-validation. As shown in the figure, the training results can be grouped into three categories: the first category is random forest, KNN, GBDT and NN, with R^2 values around 0.8; the second category includes SVR and AdaBoost, with R^2 values below 0.6; and the last category consists of linear regression, ridge regression, Lasso regression, with R^2 values below 0.4. The simple imputation strategy has a relatively minor impact on the ultimate performance of the model. The model that demonstrated superior performance was the NN model with mode imputation, which exhibited an R^2 value of 0.783.

The deposition rate was predicted using the NN with mode imputation, and subsequently fitted to the experimental values. The resulting fit is presented in Figure 5, which shows an R^2 value of 0.7986 and MAPE of 12%. The relatively high error highlights the limitations of simple imputation. While simple imputation is computationally straightforward, it often overlooks the true distributional characteristics, which can reduce variability and, in turn, degrade model performance. Furthermore, simple imputation is insufficient for capturing the correlations between variables, particularly in multivariate analyses. To optimize the imputation process, we proposed SBF (see Section 2.3 for details).

Figure 6 shows the concentration distribution of the key reactants under typical experimental conditions. Following the transfer of the input gas through the distribution plate, a uniform concentration distribution is achieved. The MTS is then conveyed along the flow direction by the combined effects of fluid convection and material diffusion, with continuous decomposition and generation of intermediate substances. The gas-phase and surface reaction model employed for the reactor model is presented in Table 2.

The selection of this simplified four-reaction mechanism was based on comprehensive analysis of dominant pathways in CVD-SiC formation. While MTS decomposition generates numerous intermediate products, including C_2H_4 and $SiCl_2$ as key species efficiently captures the primary reaction kinetics while maintaining computational tractability. The pre-exponential factors and activation energies were initially estimated from literature data and subsequently optimized through iterative validation against experimental deposition rates until the overall error was reduced below 15%. This simplified approach prioritizes computational efficiency while preserving predictive accuracy for the deposition rate. However, it is

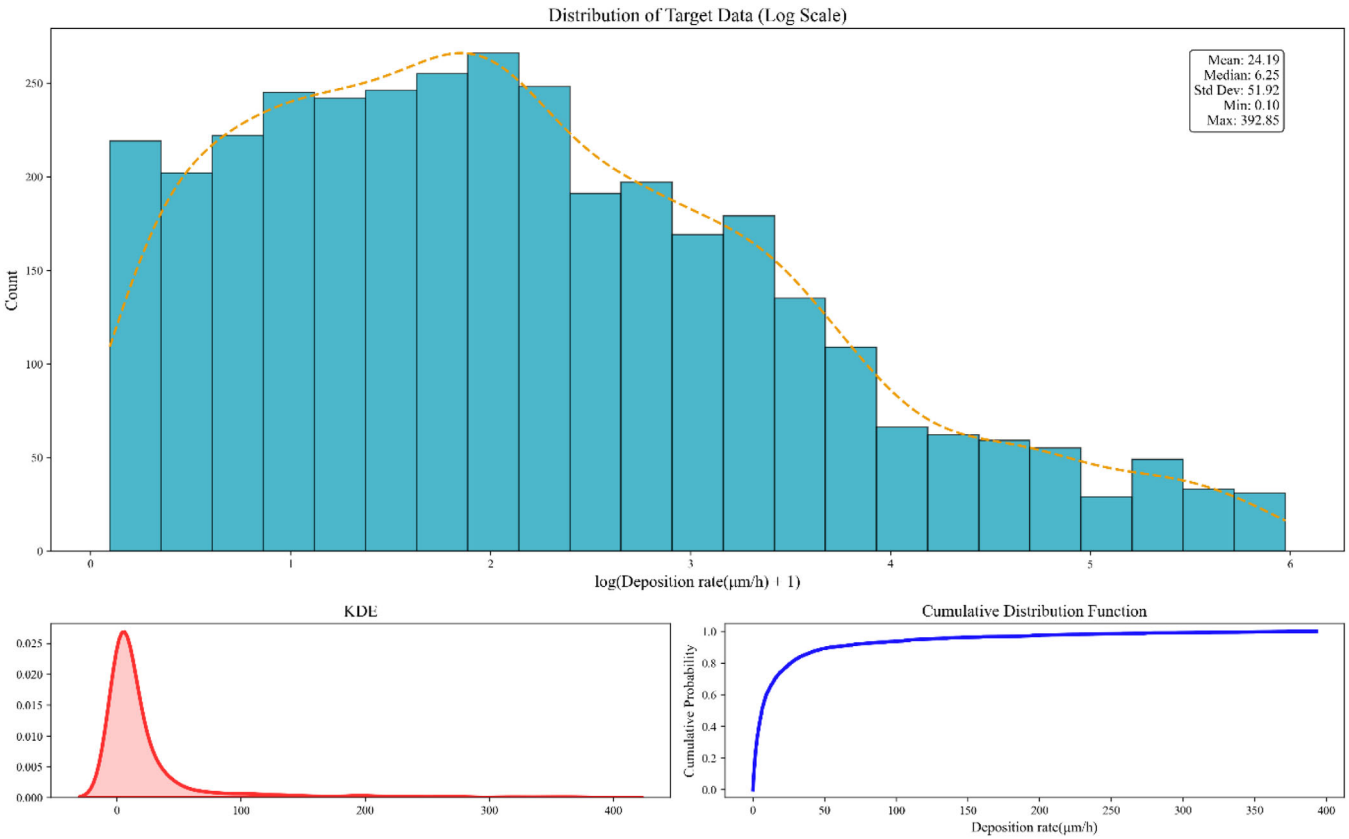


FIGURE 3 Distribution of target data.

TABLE 2 Gas phase and surface reaction.

No.	Reaction	A (s^{-1} or $m^3/(mol \cdot s)$)	n	E ($kJ \cdot mol^{-1}$)
1	$CH_3SiCl_3 \Rightarrow CH_3 + SiCl_3$	$7.63E14$	0	$2.9E5$
2	$2SiCl_3 + H_2 \Rightarrow 2SiCl_2 + 2HCl$	$3E9$	0	$2E5$
3	$2CH_3 \Rightarrow C_2H_4 + H_2$	$8E11$	0	$8.5E4$
4	$C_2H_4 + 2SiCl_2 \Rightarrow 2SiC(ads) + 4HCl$	$1.53E5$	0	$3.5E5$

important to acknowledge the limitations of this simplification, which may include reduced ability to predict detailed microstructural evolution, incomplete representation of minor reaction pathways that could become significant under extreme processing conditions, and potential inaccuracies in modeling transient phenomena. Nevertheless, for the target application of predicting deposition rates under typical processing conditions, these limitations have minimal impact on the model's overall predictive capability.

The intermediate gas concentration in Figure 6 exhibits greater variability of the substance in the region of the substrate. One reason for this is that the presence of the substrate hinders the transport of gases, leading to lower concentrations of substances on surfaces distant from the inlet. The other factor is the surface reaction. As illus-

trated in Figure 1A, intermediate gases such as C_2H_4 and $SiCl_2$ undergo a series of processes on the surface of the substrate, including adsorption, migration, reaction, and desorption. The control mechanisms of these processes vary under different conditions, resulting in the depletion of intermediate gases and a significant concentration gradient.

The deposition rate simulation data obtained from the optimized multi-physics field simulation model exhibited a strong correlation with experimental observations under typical boundary conditions, with an R^2 value of 0.9735 and MAPE of 4% (Figure 7). The simulation was then refined in the data-sparse region to generate additional data with complete features. The following experimental parameters were simulated: temperature in 1173–1573 K, pressure in 6–12 kPa, and α in 6–15. A total of 320 data with

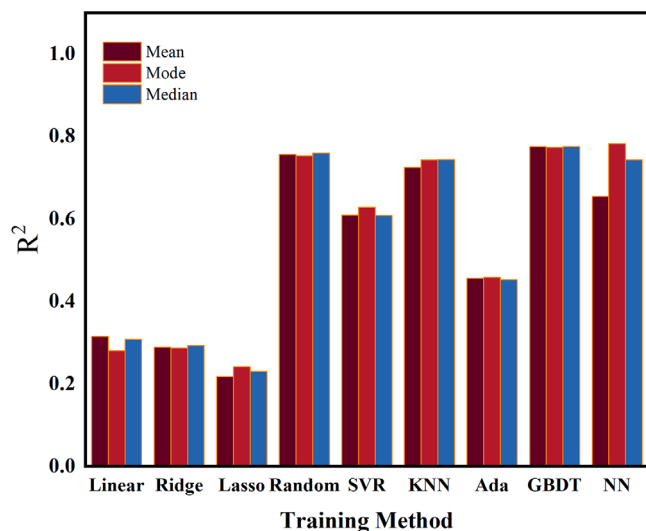


FIGURE 4 R^2 of training model with simple imputation.

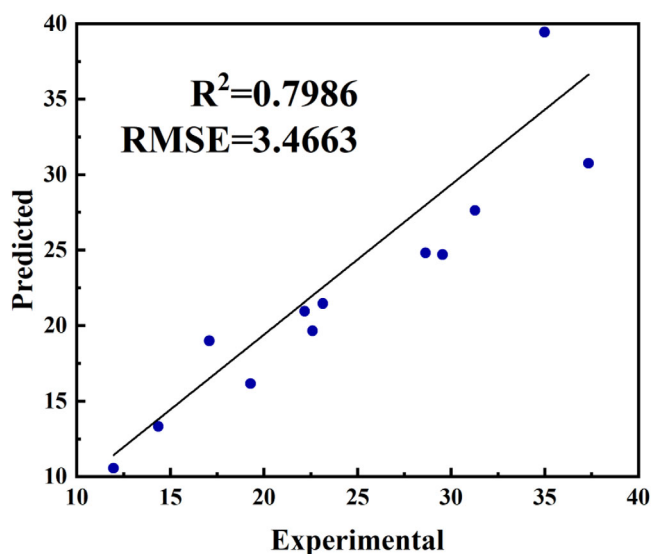


FIGURE 5 Goodness of fit between experimental value and predicted values by neural network (NN) with mode imputation.

feature completeness were generated. Missing values were imputed using a variety of regression strategies based on the complete feature set. The performance of the training model is presented in Figure 8.

The fit of each training model to the dataset was enhanced following model-based imputation, as depicted in Figure 8. The best model was the GBDT model with random forest imputation, which achieved an R^2 value of 0.9853 and a reduced MAPE of 3%. Its fit to the experimental data is shown in Figure 9, with an improved R^2 value of 0.9899.

The training dataset is composed of simulation data and literature data. The boundary conditions of the simulation data ($C_{\text{simulation}}$) are subordinate to those of the litera-

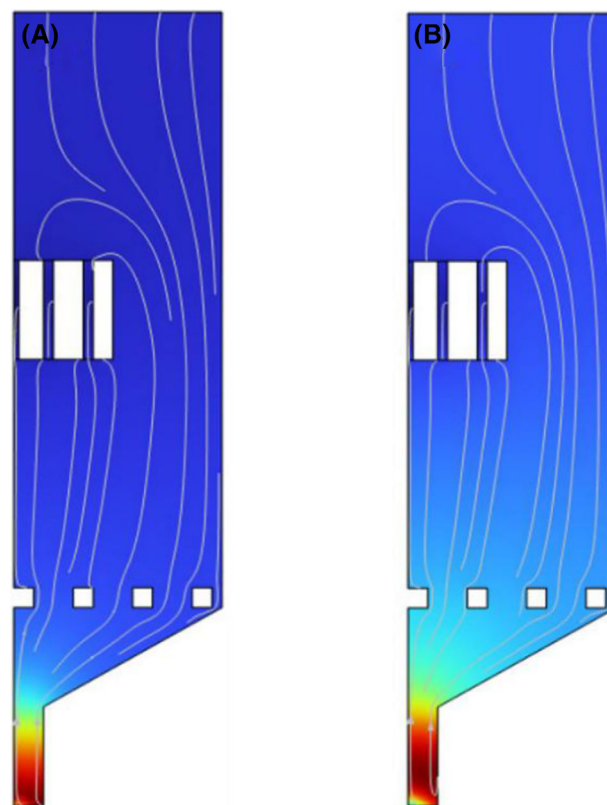


FIGURE 6 Concentration distribution in reactor model of (A) C_2H_4 ; (B) $SiCl_2$ at $T = 1373$ K, $P = 12$ kPa, $\alpha = 10$, and $Q_{MTS} = 100$ sccm.

ture data ($C_{\text{literature}}$), where $C_{\text{simulation}}$ belongs to $C_{\text{literature}}$ ($C_{\text{simulation}} \subseteq C_{\text{literature}}$). To validate the robustness of the predictive model, we compared the performance of the model beyond the boundary conditions of the multi-physics simulation (complement of $C_{\text{simulation}}$ in $C_{\text{literature}}$). Figure 10 presents a comparison of the fits with and without the introduction of simulation data beyond the boundary conditions of the multi-physics simulation. As illustrated, the R^2 beyond the boundary conditions also exhibits an improvement to 0.9746 after the incorporation of the simulation data, which suggests an enhancement in model robustness.

To rigorously evaluate model robustness, we conducted a comprehensive sensitivity analysis using Sobol indices to quantify parameter influence on deposition rate predictions. Temperature emerged as the dominant parameter ($S_1 = 0.63$, $ST = 0.78$), directly accounting for 63% of output variance and explaining 78% when including interactions. Secondary parameters like residence time ($S_1 = 0.07$, $ST = 0.10$) and Q_{mts} ($S_1 = 0.06$, $ST = 0.16$) showed amplified influence through interactions, while parameters including pressure, substrate position, Q_{total} , and reactor length collectively contributed less than 10% to output uncertainty. This analysis confirms that our model

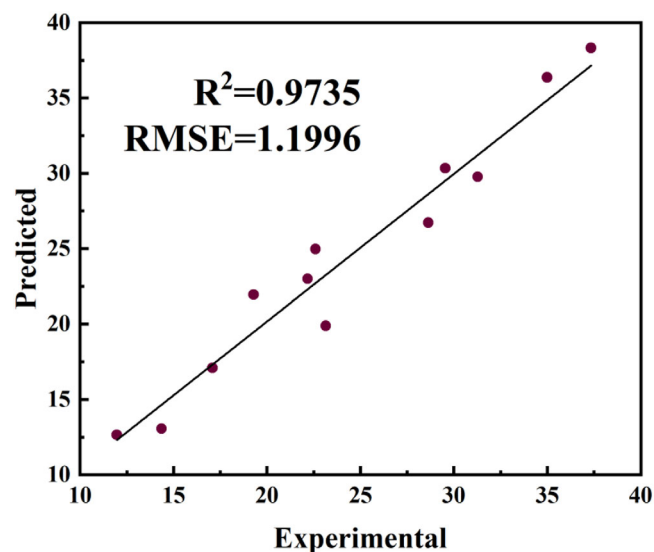


FIGURE 7 Goodness of fit between experimental and predicted values by reactor model.

correctly captures the physical dominance of temperature in the CVD-SiC process.

To further validate prediction reliability, we statistically analyzed model residuals. The systematic error (bias) was 0.0342, indicating a negligible tendency to slightly overestimate deposition rates. The random error was 0.7974, demonstrating low variability in predictions. Normality of residuals was confirmed through multiple statistical tests (Shapiro–Wilk, Kolmogorov–Smirnov, and Anderson–Darling), with all p -values exceeding the 0.05 significance threshold.

Model-based imputation offers significant advantages over simple imputation methods, particularly in four key areas: accuracy, generalization ability, robustness, and interpretability. First, regression models can capture the intricate relationships between data points by fitting mathematical equations, allowing for a more accurate representation of the non-linear characteristics of the data. Second, these models exhibit improved generalization capabilities, maintaining high prediction accuracy when applied to new, unseen data. Furthermore, regression models can effectively reduce the impact of noise and outliers through regularization techniques, resulting in greater robustness. Finally, model-based imputation enhances interpretability, as the underlying relationships between features are more transparent compared to simple imputation methods. These advantages arise from the regression model's capacity to adaptively modify its parameters to accommodate the complexities of the data. This flexibility enables a more in-depth and thorough data analysis through mathematical modeling. It is evident that model-based imputation is a more appropriate and effective approach in scenarios that necessitate high accuracy, robust gener-

alization capabilities, and a comprehensive understanding of the underlying mechanisms of the data.

Figure 11 presents the Mantel test of CVD-SiC process parameters. The heat map visually illustrates the correlation between distinct features, with lines indicating the correlation between individual parameters and the deposition rate. As shown in the heat map, there is a strong correlation between QH2 (sccm), Qdiluent (sccm) and Qtotal (sccm). In the CVD process, the gas flow rate is typically optimized to achieve specific deposition conditions, stabilizing the ratio of the component gases and thereby creating a high degree of correlation. The p -value (Mantel's p) is to determine whether the r -value correlation coefficient is statistically significant. A p -value less than 0.05 indicates a strong correlation between the two variables, making the r -value significant. Parameters that are statistically significant with respect to the deposition rate are denoted by orange lines, with the line thickness representing the strength of the correlation. This finding was further confirmed by the subsequent feature analysis.

To elucidate the influence of each parameter on the deposition rate in the CVD process, a comprehensive investigation was conducted to examine the distinct contribution of each variable and its intricate impact on the model. Figures 12 and 13 illustrate the feature importance and SHAP plots of CVD-SiC, respectively. Both plots indicate that temperature is the primary influencing factor. In the SHAP plot, the temperature parameter shows a transition from a positive to a negative influence as the temperatures increases. This illustrates the non-linear relationship between temperature and deposition rate, highlighting the importance of temperature fluctuation in the CVD-SiC process. At lower temperatures, a moderate increase facilitates the nucleation and growth of SiC crystals while reducing impurities and defects, thereby benefiting the deposition. Temperature is thus considered a beneficial factor in this phase. Elevated temperatures often lead to an increase in by-products, impurities, and defects, which hinder nucleation and reduce the coating quality. This shift from a positive to a negative effect underscores the importance of precise temperature control when optimizing reaction parameters, as extremes in temperature can have detrimental effects on coating deposition.

The residence time describes the time that the reactive gas flows over the surface of the substrate and is calculated as follows:

$$\tau = \frac{V}{Q} \cdot \frac{T_0}{T} \cdot \frac{P}{P_0} \quad (12)$$

where V is the effective volume ($V = \pi r^2 \cdot L$) and L is the substrate position; Q is the reaction gas flow;

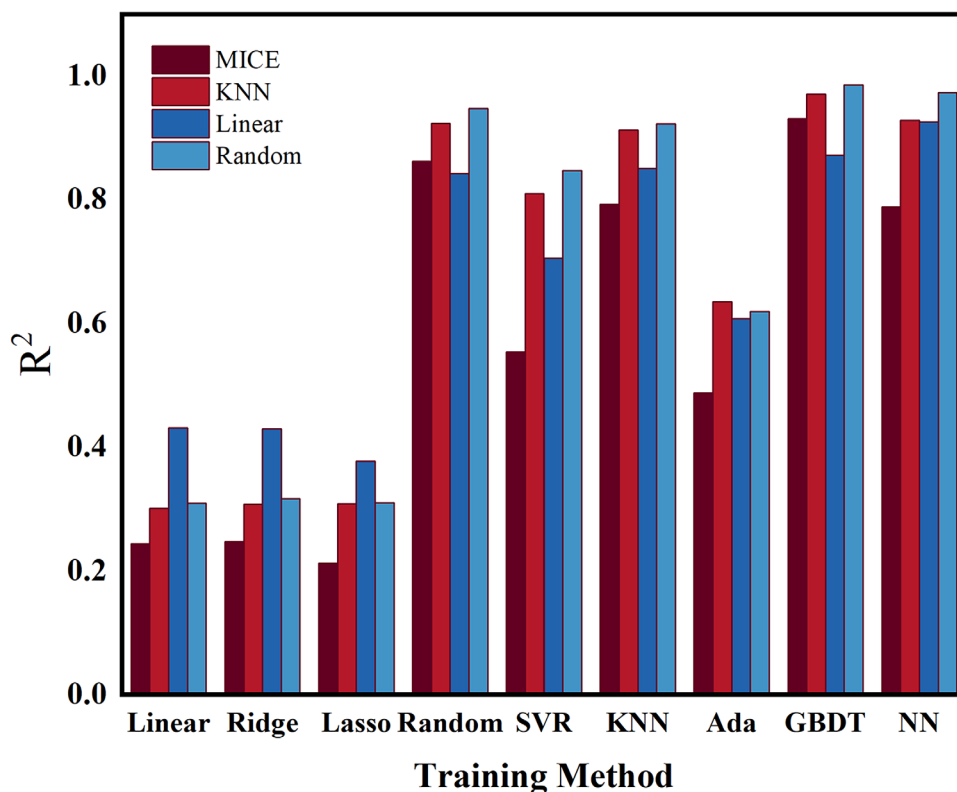


FIGURE 8 R^2 of training model with model-based imputation.

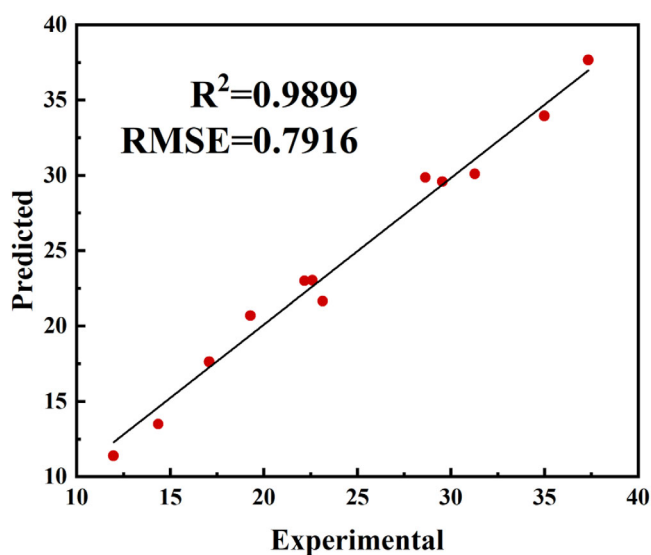


FIGURE 9 Goodness of fit between experimental values and predicted values by gradient boosting decision tree (GBDT) with random forest imputation.

T , P are the reaction temperature and total pressure, respectively; $T_0 = 298$ K; and $P_0 = 101$ kPa. The residence time is dependent on a number of factors, including the substrate position, the rate of gas flow, temperature, and pressure. Variations in residence time result in different extents of chemical reactions, leading to a different com-

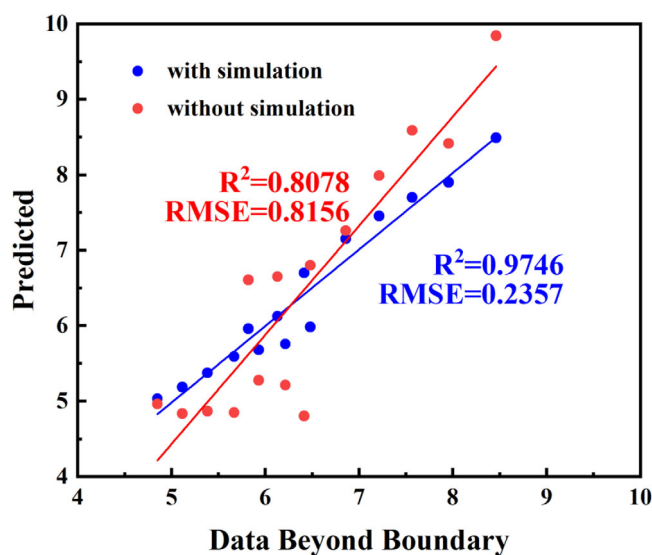


FIGURE 10 Goodness of fit between data and predicted values beyond the boundary conditions of the multi-physics simulation.

position and concentration of species in the gas phase. Consequently, the effective reactive species and their concentrations in the formed material are also altered, which ultimately impacts the properties of the prepared material. In light of the above, residence time can also be defined as “effective reaction time”.

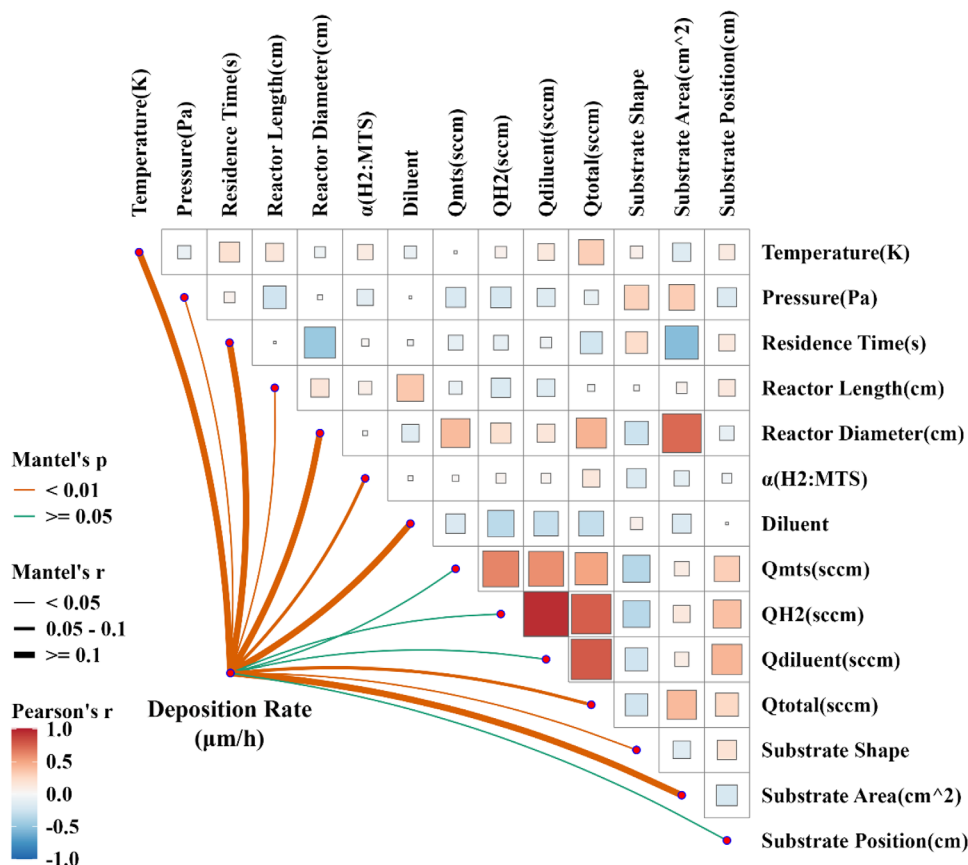


FIGURE 11 Mantel test of chemical vapor deposition-silicon carbide (CVD-SiC) process parameters.

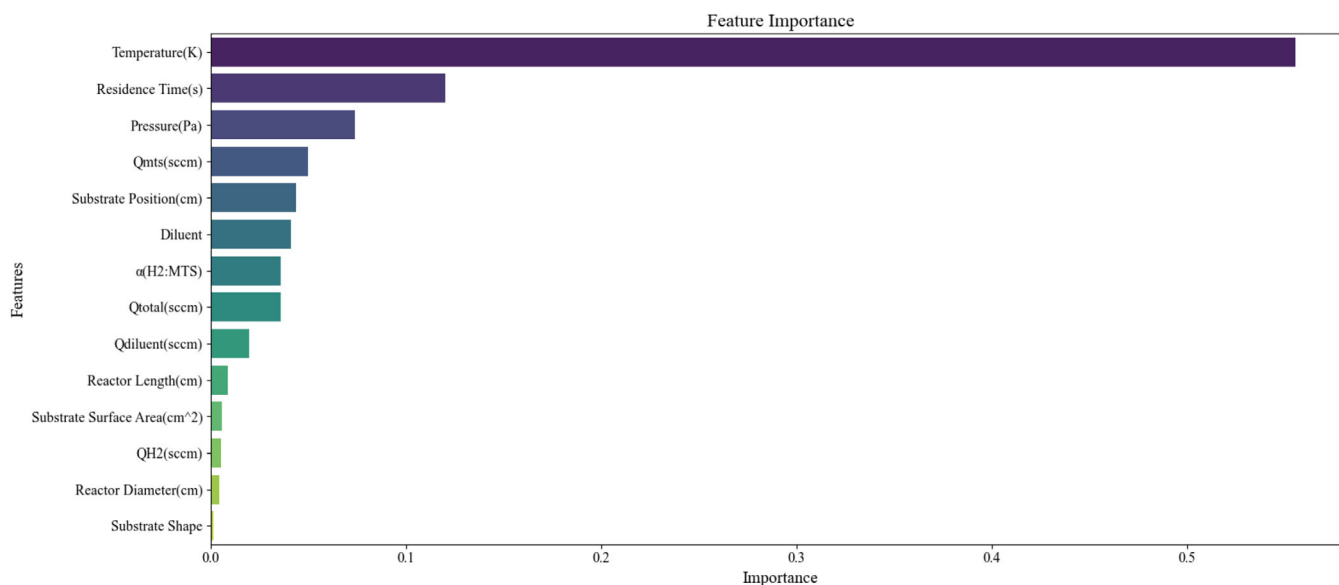


FIGURE 12 Feature importance of chemical vapor deposition-silicon carbide (CVD-SiC) process parameters.

In multiphysics simulations, a laminar flow assumption was applied to the fluid. The data were fitted using the binning method.⁸⁸ The code implements an adaptive polynomial regression framework with three core

components: (1) An adaptive binning preprocessor that dynamically segments data into spatial bins (default = 50 bins) to compute mean values, effectively filtering high-frequency noise while preserving underlying trends; (2)

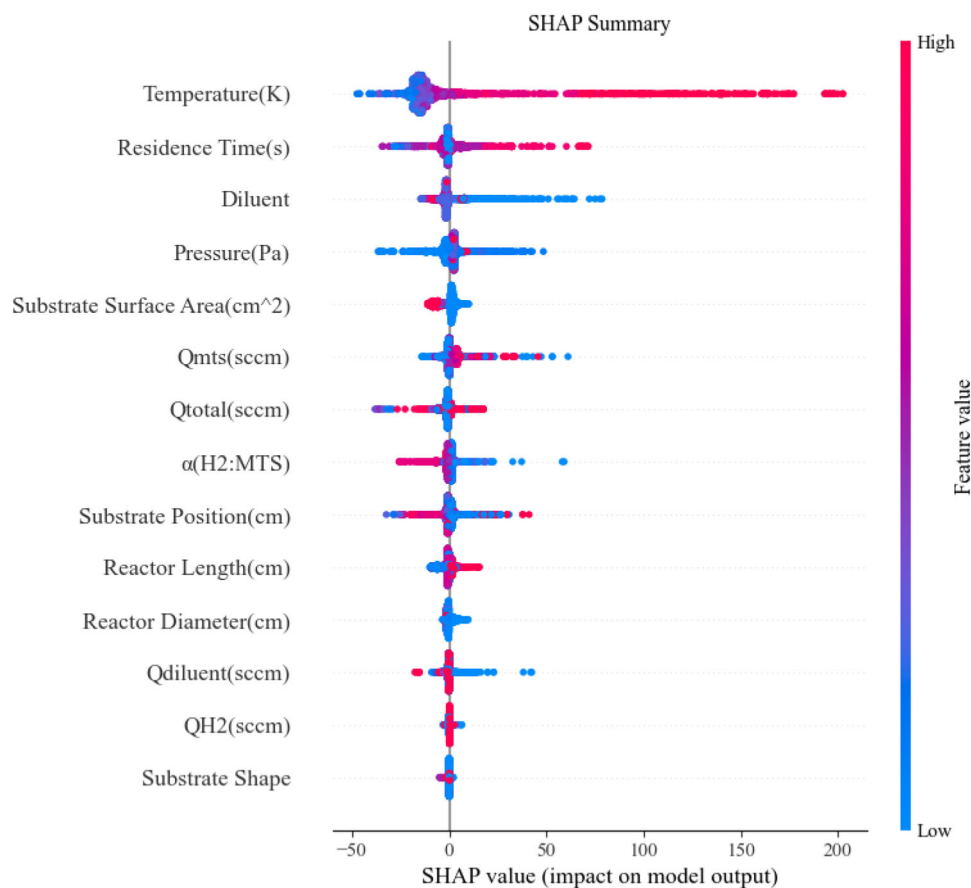


FIGURE 13 SHapley Additive exPlanations (SHAP) of chemical vapor deposition-silicon carbide (CVD-SiC) process parameters.

a polynomial feature generator that systematically constructs polynomial terms up to specified degrees (tested up to 30th order), coupled with rigorous data quality checks including length validation and NAN detection; (3) a model evaluator that quantifies goodness-of-fit through R^2 metrics and generates comparative visualizations of polynomial order effects. The systematic testing across degrees 1–30 reveals optimal model complexity through R^2 maximization, with results automatically exported to structured Excel reports. This implementation particularly addresses overfitting through spatial regularization via binning rather than conventional mathematical regularization methods. The variation of R^2 with the maximum degree of the polynomial is shown below.

The results demonstrate a nonlinear relationship between polynomial degree (max degree) and model performance (R^2). Initially, R^2 increases with higher degree, peaking at degree 12 (Figure 14), indicating improved fit with added complexity. Beyond this point, R^2 declines gradually (e.g., 0.686 at degree 12 $\rightarrow$ 0.232 at degree 30), suggesting overfitting where the model captures noise rather than the underlying trend. Minor fluctuations (e.g., temporary R^2 increases at degrees 16 and 20) may reflect localized data patterns but do not alter the overall

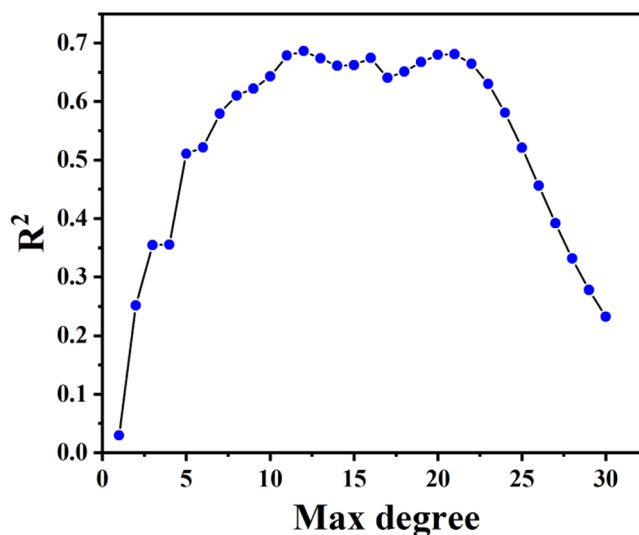


FIGURE 14 Variation of R^2 with the maximum degree of the polynomial.

downward trajectory. Optimal performance balances complexity and generalization, with degree 12 achieving the highest R^2 . Further increasing complexity harms predictive power, emphasizing the need for regularization

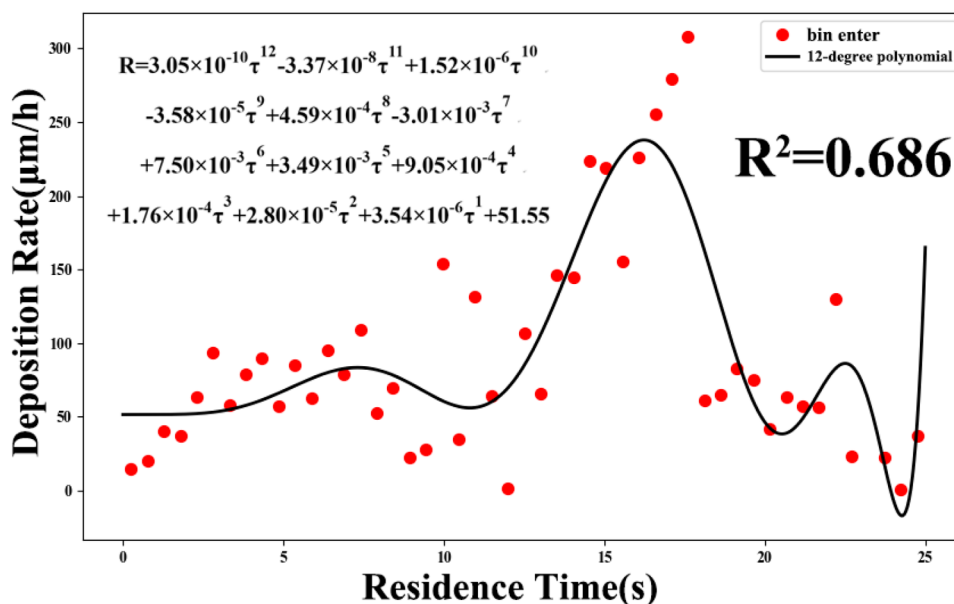


FIGURE 15 The fitting of residence time and deposition rate with 12-degree polynomial.

or early stopping in high-degree polynomial models. The coefficients of the polynomial ($R = \sum_{i=1}^{12} a_i \tau^i$) with a max degree of 12 are listed in Figure 15, where R denotes deposition rate ($\mu\text{m/h}$) and τ denotes residence time (s).

An increase in pressure within the reaction chamber reduces the average distance between reactant molecules, leading to a higher frequency of collisions. This typically enhances the reaction rate. However, as pressure increases, there is also a rise in the accumulation of reaction by-products. If the deposition process is controlled by a surface chemical reaction, the Langmuir isothermal adsorption principle ($\theta_i = K_i \cdot \theta \cdot P_i$) suggests that an increase in total pressure will elevate the concentration of reactants, the amount of reactive material adsorbed on the substrate surface, and the deposition rate. However, once the active adsorption sites on the substrate surface become saturated, the concentration of reactants no longer affects the deposition rate, which then remains constant.

Moreover, the positions of the substrate within the heating zone can result in the formation of concentration gradients and temperature gradients. A larger surface area may lead to concentration gradients on the substrate, especially when the reactant supply is limited, resulting in non-uniform film thickness. A smaller reactor radius promotes a more uniform flow, temperature, and concentration field. In contrast, reactor length and substrate shape have a lesser impact on the deposition rate.

The model's performance was further evaluated across different substrate geometries and positions to assess its generalizability. For geometries, prediction errors were as follows: circular (2.43%), rectangular (2.91%), square

(2.86%), and nozzle (3.36%), showing slightly better performance for simpler geometries. Across substrate positions, errors were consistently low: 2 cm (2.68%), 3 cm (3.03%), 4 cm (2.92%), and 5 cm (2.19%). The model performed exceptionally well in the kinetic-to-mass-transport transition region (1473–1573K), achieving an R^2 of 0.9910 and MAPE of 2.86%. This demonstrates the model's robustness across varying process conditions, though further optimization could improve predictions for complex geometries like nozzles.

5 | CONCLUSIONS

In this study, the potential of artificial intelligence in the context of the CVD-SiC process was explored using ML. A multi-source dataset was compiled by combining literature data, numerical simulation outcomes, which were fitted with experimental results. Using this dataset, a data-driven ML model was developed to predict the deposition rate of CVD-SiC coatings. The incorporation of simulation data enhances the precision and robustness of ML models. The SBF method reduced the prediction error from 12% to 3%. Global sensitivity analysis revealed temperature as the dominant parameter, accounting for 63% of output variance directly and 78% when including interactions, providing quantitative validation of experimental observations in CVD-SiC processing.

Beyond deposition rate prediction, the established computational framework provides a foundation for future work on predicting and optimizing additional coating

properties. While the current implementation focuses on deposition rate, the methodology can be extended to incorporate coating uniformity, microstructure, and mechanical properties as multi-objective optimization targets. This multi-source data-driven model offers forward guidance for experiments, helping reduce exploration costs and enhance efficiency.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude for the financial support provided by the Major Basic Research Projects on Equipment (Grant No. 909010106-301), the National Natural Science Foundation of China (Grant No. 52472096), and the Guangdong Basic and Applied Basic Research Foundation (Grants No. 2023A1515012156 and 2024A1515011656). We would also like to thank the Xi'an Aerospace Propulsion Institute (Contract No. 040020006120250054) for their generous support throughout this research. Their funding has been instrumental in the successful completion of this work.

ORCID

Kang Guan  <https://orcid.org/0000-0001-5619-6290>

Xiaohui Yang  <https://orcid.org/0000-0001-8805-8351>

REFERENCES

- Liu G, Tuttle BR, Dhar S.. Silicon carbide: a unique platform for metal-oxide-semiconductor physics. *Appl Phys Rev*. 2015;2:021307.
- Lee H, Smet V, Tummala R. A review of SiC power module packaging technologies: challenges, advances, and emerging issues. *IEEE J Emerg Sel Top Power Electron*. 2020;8:239–55.
- She X, Huang AQ, Lucia O, Ozpineci B. Review of Silicon carbide power devices and their applications. *IEEE Trans Indust Electron*. 2017;64:8193–205.
- He R, Zhou N, Zhang K, Zhang X, Zhang L, Wang W, et al. Progress and challenges towards additive manufacturing of SiC ceramic. *J Adv Ceram*. 2021;10:637–74.
- Terrani KA. Accident tolerant fuel cladding development: promise, status, and challenges. *J Nucl Mater*. 2018;501:13–30.
- Whitaker B, Barkley A, Cole Z, Passmore B, Martin D, McNutt TR, et al. High-efficiency, isolated on-board vehicle battery charger utilizing silicon carbide power devices. *IEEE Trans Power Electron*. 2014;29:2606–17.
- Kimoto T. Material science and device physics in SiC technology for high-voltage power devices. *Jpn J Appl Phys*. 2015;54:040103.
- Roccaforte F, Fiorenza P, Greco G, Lo Nigro R, Giannazzo F, Iucolano F, et al. Emerging trends in wide band gap semiconductors (SiC and GaN) technology for power devices. *Microelectron Eng*. 2018;187:66–77.
- Castelletto S, Johnson BC, Ivády V, Stavrias N, Umeda T, Gali A, et al. A silicon carbide room-temperature single-photon source. *Nat Mater*. 2014;13:151–56.
- Choy KL. Chemical vapour deposition of coatings. *Progr Mater Sci*. 2003;48:57–170.
- Zhang J, Wang F, Shenoy VB, Tang M, Lou J. Towards controlled synthesis of 2D crystals by chemical vapor deposition (CVD). *Mater Today*. 2020;40:132–39.
- Zhao XN, Wang XD, Xin H, Zhang L, Yang JJ, Jiang GZ. Controllable preparation of SiC coating protecting carbon fiber from oxidation damage during sintering process and SiC coated carbon fiber reinforced hydroxyapatite composites. *Appl Surf Sci*. 2018;450:265–73.
- Wen J, Zhou S, Yi L, Sun BZ, Wang Y, Li GD, et al. Oxidation behavior and high-temperature flexural property of CVD-SiC-coated PIP-C/SiC composites. *Ceram Int*. 2018;44:16583–88.
- Zhan CT, Liu YQ, Chen JP, Yang RK, Liu Y, Zhang ZX, et al. Hoop tensile properties of SiC fiber-reinforced SiC matrix composite tubes with/without CVD-SiC coating. *J Eur Ceram Soc*. 2024;44:116707.
- Baux A, Goillot A, Jacques S, Heisel C, Rochais D, Charpentier L, et al. Synthesis and properties of macroporous SiC ceramics synthesized by 3D printing and chemical vapor infiltration/deposition. *J Eur Ceram Soc*. 2020;40:2834–54.
- Abid N, Khan AM, Shujait S, Chaudhary K, Ikram M, Imran M, et al. Synthesis of nanomaterials using various top-down and bottom-up approaches, influencing factors, advantages, and disadvantages: a review. *Adv Colloid Interface Sci*. 2022;300:102597.
- Cai X, Luo Y, Liu B, Cheng HM. Preparation of 2D material dispersions and their applications. *Chem Soc Rev*. 2018;47:6224–66.
- Kukushkin SA, Osipov AV. Theory and practice of SiC growth on Si and its applications to wide-gap semiconductor films. *J Phys D: Appl Phys*. 2014;47:313001.
- Fraga M, Pessoa R. Progresses in synthesis and application of SiC films: from CVD to ALD and from MEMS to NEMS. *Micromachines (Basel)*. 2020;11:799.
- Bijjargi YS, Shinde VM, Mudgal A, Kumar H, Prasad NE. CFD modeling and optimal design of SiC deposition on the fuel combustion nozzle in a commercial CVD reactor. *Ceram Int*. 2022;48:11043–55.
- Ogawa T, Fukumoto K, Machida H, Norinaga K. CFD simulation of CVD reactors in the CH₃SiCl₃(MTS)/H₂ system using a two-step MTS decomposition and one-step SiC growth models. *Heliyon*. 2023;9:e15061.
- Cao K, Li H, Xia S, Liu H, Yang T, Zheng Y, et al. A batch preparation of large-size graphite plate/SiC coating by CVD:CFD simulation and experimental. *Ceram Int*. 2024;50:16798–812.
- Zou S, Xiao J, Wu V, Chen XD. Analyzing industrial CVD reactors using a porous media approach. *Chem Eng J*. 2021;415:129038.
- Morgan D, Jacobs R. Opportunities and challenges for machine learning in materials science. *Annu Rev Mater Res*. 2020;50:71–103.
- Moosavi SM, Jablonka KM, Smit B. The role of machine learning in the understanding and design of materials. *J Am Chem Soc*. 2020;142:20273–87.
- Kimaev G, Ricardez-Sandoval LA. Artificial neural network discrimination for parameter estimation and optimal product design of thin films manufactured by chemical vapor deposition. *J Phys Chem C*. 2020;124:18615–27.
- Rajak P, Krishnamoorthy A, Mishra A, Kalia R, Nakano A, Vashishta P. Autonomous reinforcement learning agent for chemical vapor deposition synthesis of quantum materials. *npj Comput Mater*. 2021;7:108.

28. Xu M, Tang B, Lu Y, Zhu C, Lu Q, Zhu C, et al. Machine learning driven synthesis of few-layered WTe₂ with geometrical control. *J Am Chem Soc*. 2021;143:18103–13.
29. Tang ZR, Zhao SB, Li J, Zuo YH, Tian J, Tang HY, et al. Optimizing the chemical vapor deposition process of 4H-SiC epitaxial layer growth with machine-learning-assisted multiphysics simulations. *Case Stud Thermal Eng*. 2024;59:104507.
30. Costine A, Delsa P, Li T, Reinke P, Balachandran PV. Data-driven assessment of chemical vapor deposition grown MoS₂ monolayer thin films. *J Appl Phys*. 2020;128:235303.
31. Streitwieser DA. Kinetic Investigation of the Chemical Vapor Infiltration and Reaction (CVI-R) Process for the Production of SiC and TiC Biomimetic Ceramics from Paper Preforms. Germany: Friedrich-Alexander-Universitaet Erlangen-Nuernberg 2004.
32. Lu C, Cheng L, Zhao C, Zhang L, Xu Y. Kinetics of chemical vapor deposition of SiC from methyltrichlorosilane and hydrogen. *Appl Surf Sci*. 2009;255:7495–99.
33. Josiek A, Langlais F. Kinetics of CVD of stoichiometric and Si-excess SiC in the system MTS/H₂ at medium decomposition of MTS. *Chem Vapor Deposition*. 1996;2:141–46.
34. Josiek A, Langlais F, Bourrat X. A study of the transition between growth of stoichiometric and silicon-excess silicon carbide by CVD in the system MTS/H₂. *Chem Vapor Deposition*. 1996;12:5.
35. Papasouliotis GD, Sotirchos SV. Hydrogen chloride effects on the CVD of silicon carbide from methyltrichlorosilane. *Chem Vapor Deposition*. 1998;4:235–46.
36. Fukushima Y, Sato N, Funato Y, Sugiura H, Hotozuka K, Momose T, et al. Multi-scale analysis and elementary reaction simulation of SiC-CVD using CH₃SiCl₃/H₂. *ECS J Solid State Sci Technol*. 2013;2:P492–97.
37. Funato Y, Sato N, Fukushima Y, Sugiura H, Momose T, Shimogaki Y. Fundamental evaluation of gas-phase elementary reaction models for silicon carbide chemical vapor deposition. *ECS J Solid State Sci Technol*. 2017;6:P399–P404.
38. Kostjulin IM, Sotirchos SV. Codeposition of SiC and C from mixtures of methyltrichlorosilane and ethylene in hydrogen. *Ind Eng Chem Res*. 2001;40:11.
39. Venugopalan R, Prakash J, Nuwad J, Pillai CGS, Tyagi AK, Dakshinamoorthy S. Morphological study of SiC coating developed on 2D carbon composites using MTS precursor in a hot-wall vertical reactor. *Int J Mater Res*. 2012;103:1251–56.
40. Li B, Huang H, He T, Mao B, Wang X. Effects of process parameters on deposition rate of SiC nanowires by chemical vapor deposition. *J Chem Eng Japan*. 2020;53:273–79.
41. Loumagne F, Langlais F, Naslain R. Experimental kinetic study of the chemical vapour deposition of SiC-based ceramics from CH₃SiCl₃/H₂ gas precursor. *J Cryst Growth*. 1995;155:7.
42. Sone H, Kaneko T, Miyakawa N. In situ measurements and growth kinetics of silicon carbide chemical vapor deposition from methyltrichlorosilane. *J Cryst Growth*. 2000;219:245–52.
43. Josiek A, Langlais F. Residence-time dependent kinetics of CVD growth of SiC in the MTS/H₂ system. *J Cryst Growth*. 1996;160:253–60.
44. Kaneko T, Okuno T. Growth kinetics of silicon carbide CVD. *J Cryst Growth*. 1988;91:599–604.
45. Loumagne F, Langlais F, Naslain R. Reactional mechanisms of the chemical vapour deposition of SiC-based ceramics from CH₃SiCl₃/H₂ gas precursor. *J Cryst Growth*. 1995;155:205–13.
46. Matsunami H, Nishino S, Odaka M, Tanaka T. Epitaxial growth of α -SiC layers by chemical vapor deposition technique. *J Cryst Growth*. 1975;31:72–75.
47. Mollick PK, Venugopalan R, Srivastava D. CFD coupled kinetic modeling and simulation of hot wall vertical tubular reactor for deposition of SiC crystal from MTS. *J Cryst Growth*. 2017;475:97–109.
48. Chiu CC, Desu SB, Tsai CY. Low pressure chemical vapor deposition (LPCVD) of β -SiC on Si(100) using MTS in a hot wall reactor. *J Mater Res*. 2011;8:2617–26.
49. Chiu CC, Desu SB, Tsai CY. Low pressure chemical vapor deposition (LPCVD) of β -SiC on Si(100) using MTS in a hot wall reactor. *J Mater Res*. 1993;8:2617–26.
50. Lee Y-J. The effect of diluent gases on the growth behavior of CVD SiC films with temperature. *J Mater Sci*. 2000;35:4519–26.
51. Choi BJ, Kim DR. Growth of silicon carbide by chemical vapour deposition. *J Mater Sci Lett*. 1991;10:860–62.
52. Su PY, Hyon KM, Young LJ. The effect of substrate materials on the microstructure and phases of silicon carbide prepared by chemical vapour deposition. *J Mater Sci Lett*. 1989;8:321–25.
53. Mousavipour SH, Saheb V, Ramezani S. Kinetics and mechanism of pyrolysis of methyltrichlorosilane. *J Phys Chem A*. 2004;108:1946–52.
54. Besmann TM, Sheldon BW, III TSM, Kaster MD. Depletion effects of silicon carbide deposition from methyltrichlorosilane. *J Am Ceram Soc*. 1992;75:2899–903.
55. G.T.D, Deposition and microstructure of vapor-deposited silicon carbide. *J Am Ceram Soc*. 1967;51:4.
56. Papasouliotis GD, Sotirchos SV. Gravimetric investigation of the deposition of SiC films through decomposition of methyltrichlorosilane. *J Electrochem Soc*. 1995;142:3834–44.
57. Seo J-W, Choi K. Application of CFD simulation to silicon carbide deposition for nozzles with funnel. *J Korean Ceram Soc*. 2020;58:184–91.
58. Langlais F, Prebende C, Tarride B, Naslain R. On the kinetics of the CVD of Si from SiH₂Cl₂/H₂ and SiC from CH₃SiCl₃/H₂ in a vertical tubular hot-wall reactor. *Le Journal de Physique Colloques*. 1989;50:C5-93–C5-103.
59. Langlais F, Loumagne F, Lespiaux D, Schamm S, Naslain R. Kinetic processes in the CVD of SiC from CH₃SiCl₃-H₂ in a vertical hot-wall reactor. *Le Journal de Physique IV*. 1995;05:C5-105–C5-112.
60. Parretta A, Giunta G, Cappelli E, Adoncccchi V, Vittori V. Influence of substrate and process parameters on the properties of CVD-SiC. *Mater Res Soc*. 1990;168:227–32.
61. Delhaes C, Neuschütz D. Rate of SiC deposition from methyltrichlorosilane and influence of HCL addition. *Mater Res Soc*. 1995;363:63–68.
62. Minford E, Stevens RE, Magnotta VL, Dyer PN, Watkins TR, Green DJ. Effects of deposition conditions on the microstructure and properties of CVD SiC. *Mater Res Soc*. 1989;168:6.
63. Shishkin Y, Myers-Ward RL, Sadow SE, Galyukov A, Vorob'ev AN, Brovin D, et al. Analysis of SiC CVD growth in a horizontal hot-wall reactor by experiment and 3D modelling. *Mater Sci Forum*. 2007;556–557:61–64.

64. Radmilovic V, Gao D, Maboudian R, Dahmen U. Growth mechanism of (111)-textured 3C-SiC films on Si(100) substrates. *Microsc Microanal*. 2003;9:664–65.
65. Lee YJ, Choi DJ, Kim SS, Lee HL, Kim HD. Comparison of diluent gas effect on the growth behavior of horizontal CVD SiC with analytical and experimental data. *Surf Coat Technol*. 2004;177–178:415–19.
66. Long Y, Javed A, Shapiro I, Chen Z-k, Xiong X, Xiao P. The effect of substrate position on the microstructure and mechanical properties of SiC coatings on carbon/carbon composites. *Surf Coat Technol*. 2011;206:568–74.
67. Dong-Joo K, Doo-Jin C, Young-Wook K. Effect of reactant depletion on the microstructure and preferred orientation of polycrystalline SiC films by chemical vapor deposition. *Thin Solid Films*. 1995;266:6.
68. Chin JM, Gantzel P, Hudson R. The structure of chemical vapor deposited silicon carbide. *Thin Solid Films*. 1977;40:57–72.
69. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *Am J Epidemiol*. 2014;179:764–74.
70. Sun Y, Li J, Xu Y, Zhang T, Wang X. Deep learning versus conventional methods for missing data imputation: a review and comparative study. *Expert Syst Appl*. 2023;227:120201.
71. Mera-Gaona M, Neumann U, Vargas-Canas R, Lopez DM. Evaluating the impact of multivariate imputation by MICE in feature selection. *Plos One*. 2021;16.
72. Rajakarunakaran SA, Lourdu AR, Muthusamy S, Panchal H, Alrubaie AJ, Jaber MM, et al. Prediction of strength and analysis in self-compacting concrete using machine learning based regression techniques. *Adv Eng Software*. 2022;173:103267.
73. Kejiang Z, Achari G, Hua L, Zargar A, Sadiq R. Machine learning approaches to predict coagulant dosage in water treatment plants. *Int J Syst Assurance Eng Manage*. 2013;4:205–14.
74. Ao Y, Li H, Zhu L, Ali S, Yang Z. The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. *J Petrol Sci Eng*. 2019;174:776–89.
75. Wu Y. Linear regression in machine learning. *Proc SPIE*. 2022;12163:121634T.
76. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. *Knowl Inf Syst*. 2008;14:1–37.
77. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–30.
78. Bobbili R, Madhu V. A machine learning model for failure of perforated plates under impact. *Mech Based Des Struct Mach*. 2022;50:2582–90.
79. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86:2278–324.
80. Vera-Badillo FE, Templeton AJ, de Gouveia P, Diaz-Padilla I, Bedard PL, Al-Mubarak M, et al. Androgen receptor expression and outcomes in early breast cancer: A systematic review and meta-analysis. *J Natl Cancer Instit*. 2014;106:djt319.
81. Meirmans PG. genodive version 3.0: easy-to-use software for the analysis of genetic data of diploids and polyploids. *Mol Ecol Resour*. 2020;20:1126–31.
82. Zheng Q, Hu Y, Zhang S, Noll L, Boeckle T, Dietrich M, et al. Soil multifunctionality is affected by the soil environment and by microbial Choo community composition and diversity. *Soil Biol Biochem*. 2019;136:107521.
83. Li C, Gan Y, Zhang C, He H, Fang J, Wang L, et al. “Microplastic communities” in different environments: differences, links, and role of diversity index in source analysis. *Water Res*. 2021;188:116574.
84. Ekanayake IU, Meddage DPP, Rathnayake U. A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Stud Constr Mater*. 2022;16:e01059.
85. Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Comput Methods Programs Biomed*. 2022;214:106584.
86. Li Z. Extracting spatial effects from machine learning model using local interpretation method: an example of SHAP and XGBoost. *Comput Environ Urban Syst*. 2022;96:101845.
87. Rodriguez-Perez R, Bajorath J. Interpretation of machine learning models using Shapley values: application to compound potency and multi-target activity predictions. *J Comput-Aided Mol Des*. 2020;34:1013–26.
88. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. New York: Springer, 2009.

How to cite this article: Liu R, Song C, Lu Z, Guan K, Wang X, Cao P, et al. Multi-source data-driven model to predict the deposition rate of CVD-SiC coatings with improved accuracy and robustness. *J Am Ceram Soc*. 2025;108:e20616. <https://doi.org/10.1111/jace.20616>