Journal of Intelligent Manufacturing
https://doi.org/10.1007/s10845-023-02082-8

Machine learning enabled optimization of showerhead design for semiconductor deposition process

Zeqing Jin[1] · Dahyun Daniel Lim[1] · Xueying Zhao[2] · Meenakshi Mamunuru[2] · Sassan Roham[2] · Grace X. Gu[1]

Received: 28 September 2022 / Accepted: 18 January 2023

Abstract
In semiconductor fabrication, the deposition process generates layers of materials to realize insulating and conducting functionality. The uniformity of the deposited thin film layers' thickness is crucial to create high-performance semiconductor devices. Tuning fabrication process parameters (e.g., for evenly distributed gas flow on the semiconductor wafer) is one of the dominant factors that affect film uniformity, as evidenced by both experimental and numerical studies. Conventional trial and error methods employed to change and test a range of fabrication conditions are time-consuming, and few studies have explored the effect of changing the geometry of hardware components, such as the showerhead. Here, we present a design optimization of the showerhead for flow uniformity based on numerical simulation data using machine learning surrogate models. Accurate machine learning models and optimization algorithms are developed and implemented to achieve 10% more flow uniformity compared to a benchmark traditional showerhead design. Moreover, the developed Bayesian optimization method saves 10-fold computational cost in reaching the optimal showerhead designs compared to conventional approaches. This machine learning enabled optimization platform shows promising results which could be implemented for other optimization problems in various manufacturing systems such as semiconductor fabrication and additive manufacturing.

Keywords Semiconductor deposition process · Showerhead design · Machine learning · Surrogate model · Bayesian optimization

Introduction

Semiconductor manufacturing is an indispensable part of contemporary technology to create essential devices and components closely related to people's daily life and the operation of modern society. Semiconductor fabrication utilizes cutting-edge techniques from multiple fields including thermal, control, chemistry, and material sciences. In semiconductor fabrication, the deposition process is a crucial step. A typical deposition setup includes a reaction chamber, a substrate holder (pedestal), and a showerhead. During operation, the showerhead distributes reactant gas from a set of holes above the substrate, which is placed on the pedestal and film grows on the substrate under the reaction (Fig. 1(a)) (Ding et al., 2019; Li et al., 2018). Film thickness uniformity is a critical factor in evaluating the quality of film growth as a non-uniform film may cause non-functioning die and hence reduce fabrication yields. From the perspective of experimental hardware design, critical components such as the showerhead faceplate, stem, and porous baffle have a large impact on the final performance of deposition. Specifically, the film thickness is largely dominated by the profile of gas flow in the chamber, which is related to many design parameters such as the diameter of the stem, the position of the porous baffle, as well as hole size and distribution on the showerhead faceplate. Among these components, the showerhead is believed to be one of the most dominant and fundamental parts that affect flow uniformity (Liao et al., 2018; Xia et al., 2014). Current showerhead designs with symmetrical hole patterns are likely to experience undesired stagnated flows along the lines of symmetry. On the other hand, proposed non-symmetrical

✉ Grace X. Gu
ggu@berkeley.edu

[1] Department of Mechanical Engineering, University of California, Berkeley, CA 94720-1740, USA

[2] Lam Research Corporation, 4650 Cushing Pkwy, Fremont, CA 94538, USA

Published online: 15 February 2023

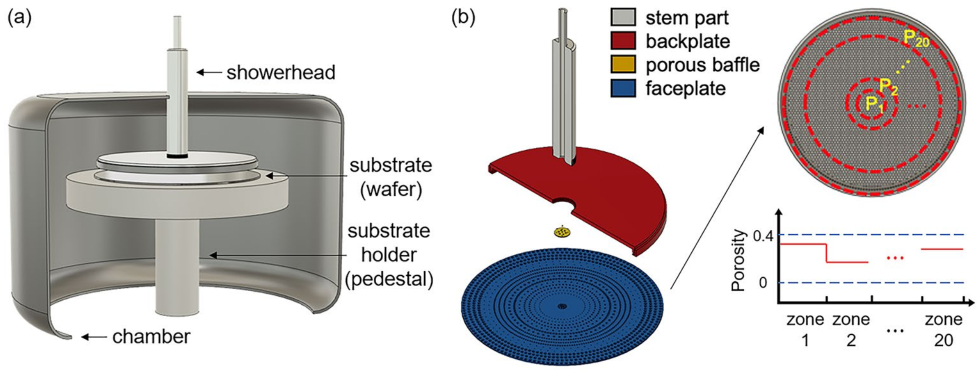

Fig. 1 (a) CAD representation of the deposition system. (b) Parameterization of the showerhead design

designs using spiral shapes lack practical demonstrations (Selep et al., 2019). Therefore, studies have been actively conducted to achieve better fabrication performance (e.g. film uniformity) by optimizing the manufacturing system (e.g. hardware design, process parameters). For example, researchers have applied full factorial and Taguchi orthogonal arrays methods to characterize and improve the low pressure chemical vapor deposition (LPCVD) process (DePinto & Wilson, 1990). A variety of factors including gas ratio, total gas volume, tube conditioning, and wafer direction were studied to obtain optimal film thickness uniformity and particle generation. With the aid of advanced computer-aided engineering (CAE) tools, researchers utilized computational fluid dynamics (CFD) simulations to improve the evenness of thin films by optimizing the rotation speed of substrates for metal-organic chemical vapor deposition (Li et al., 2018). Design of experiment (DOE) and a response surface model (RSM) are used to obtain the relationship between inputs (rotation speeds) and outputs (deposition rates). Although previous studies have actively adopted physical experiments, numerical simulations, and optimization algorithms to improve the performance of fabrication, challenges persist in the inaccuracy and inefficiency of exploring optimal designs, especially when considering the optimization of geometries of fabrication parts. Therefore, efficient optimization of the manufacturing system at a more fundamental level, identifying and compartmentalizing key geometries and parts, is vital to achieving better fabrication performance.

State-of-the-art artificial intelligence (AI) algorithms show great potential in utilizing large amounts of data generated during the manufacturing process to efficiently find solutions for optimization problems. Advances in AI techniques, specifically machine learning (ML) algorithms, have been providing remarkable achievements and success in various fields including autonomous vehicles, materials design, additive manufacturing, and biological applications (Chen & Gu, 2021; Janai et al., 2020; Jin et al., 2020, 2021; Jumper et al., 2021; Lee et al., 2022; Silver et al., 2016; Yu et al., 2022). An ML-powered approach can explore the underlying relationships between the input information and output objectives as learning is done on the given data. Moreover, ML algorithms can make accurate predictions on the objective when given unseen inputs based on the obtained analytical relationship. Hence, ML methods have been used to understand the complicated relationship between fabrication performance and system inputs such as process parameters. For instance, neural networks (NNs) were implemented to understand the relationship between process parameters (e.g. the flow rate of infill gases, the substrate temperature, and the pressure) and final product quality including film thickness and refractive index of silicon dioxide films for plasma-enhanced chemical vapor deposition (PECVD) (Chen et al., 2007). The performance of the NN method reached satisfying performance with a mean absolute deviation of 10.503 Å for thickness prediction and an error rate of 2.67% for the refractive index value. Recently, researchers have utilized recurrent neural networks (RNNs) to predict deposition behavior with less than 5% deviation from the CFD simulation results for plasma-enhanced atomic layer deposition (PEALD). Accurate predictions from the ML algorithms only take a few seconds compared to a day of computing the results using a multiscale CFD model. Although the current methods have achieved great progress in terms of predicting desired fabrication performance and accelerating the forward physical modeling process, challenges still exist in the following aspects. First, a gap in the literature exists regarding optimized hardware designs for the deposition process to improve fabrication performance. Although tuning process parameters can reach a local optimal performance, it is believed that smart hardware design is another major way to achieve better fabrication quality. Second, the choice of input design variables requires prior experience with physical experiments, which is usually nontrivial. Moreover, the importance of different design variables with respect to desired outputs is not clearly understood. Lastly, the training of ML algorithms takes a large number of data points, which requires heavy computational resources upstream to generate training data from numerical simulations.

In this paper, we aim to address the three challenges mentioned above through a case study of showerhead design optimization using efficient machine learning methods based on CFD simulations. The objective of the optimization is to improve the gas flow uniformity of the deposition process, where the input design variables are showerhead hole pattern and size, as well as the fillet on the stem part of the showerhead. Machine learning regression models are implemented to obtain the relationship between flow uniformity and design variables and sensitivity analysis is conducted to quantify the importance of the inputs. In addition, optimal showerhead designs are proposed through multiple optimization algorithms including genetic algorithm (GA), stochastic gradient descent (SGD), and Bayesian optimization (BO). The best design reaches 10% more flow uniformity compared to the baseline showerhead design at 10% of the computational cost of conventional optimization algorithms when applying BO. Our work achieves the following contributions: (1) Development of an ML-powered framework capable of optimizing hardware designs for semiconductor fabrication based on CFD simulation; (2) Implementation of an efficient optimization algorithm that can interact with numerical simulation and ML models during optimization iteration; (3) Automation of generating multiple optimal designs with high efficiency showing superior performance compared to baseline prototype. It is believed that such a framework demonstrated in our case study can adapt to other optimization problems in all kinds of manufacturing fields. The paper is organized as follows. 'Materials and methods' section discusses detailed information and methods regarding the implemented CAE process, ML models, optimization algorithms, and prototype fabrication. 'Results and discussion' section shows the ML performance and optimization results of the developed framework as well as comparisons between different optimization algorithms. 'Conclusions and perspectives' section summarizes the work, raises current challenges, and proposes future plans.

Materials and methods

This section deals with details for establishing the CAD and CFD models, training machine learning surrogate models, implementing optimization algorithms, as well as fabricating showerhead prototypes.

Parameterized CAD model and efficient CFD simulation In order to realize the design parameter study in CFD simulation, a CAD model for the deposition process is built based on a showerhead patent design developed by Lam Research (Chandrasekharan et al., 2019) and parameterized with 30 design variables including dimensions related to the stem part, back plate, porous baffle, and faceplate as shown in Fig. 1(b). Most of the design variables deal with the design of faceplate including the hole pattern and inlet stem fillet, which are considered two major factors affecting the flow distribution. Specifically, the faceplate is divided into 20 evenly spaced concentric circles, from zone 1 to zone 20. For each zone, a porosity value will be determined and treated as a design variable (Fig. 1(b)). The porosity of each zone is defined as the total area of holes over the area of the zone and the average porosity of the patent design is about 0.2. Here, the upper and lower bounds of porosity are set to 0.4 and 0 to have sufficient exploration design space. In terms of the inlet stem fillet design, two positions are considered: corner of inlet flow encountering porous baffle and faceplate. Each design has a design space of fillet radius from 0 to 3 mm.

CFD simulation is conducted through Ansys Fluent software. For the CFD simulation setup, ideal Argon gas with 3 L/min flow is used as inlet and zero-gauge pressure is set as outlet conditions. The operating pressure is set to 7.5 Torr (1000 Pa). The temperature boundary condition for the pedestal and reactor wall are 400 °C and 75 °C, respectively. Furthermore, the faceplate is viewed as a porous media with porosity values determined by the design. To ensure the flow is passing vertically across the faceplate same as the physical condition, viscous resistance value in the horizontal direction is set 1000 times larger than the value in the vertical direction. In terms of the CFD simulation outcome, the plane of interest is set 1.5 mm above the pedestal with the same circular shape as the pedestal (radius = 200 mm). On the plane of interest, 50 spokes with equal included angles are sampled and 50 sampling points (equally separated) are taken on each spoke. The non-uniformity of the flow is indicated by the standard deviation (σ) of velocity profile through these 2500 sampling points. To align with industrial conventions, three-sigma (3σ) value is used, which is a statistical calculation showing the data within three standard deviations from the mean. Ranges of 3σ velocity magnitude and axial velocity value are calculated as outputs, which are treated as evaluation metrics indicating the non-uniformity of the flow. The 3σ values calculated based on the patent design are regarded as baseline results.

Surrogate models Surrogate models are analytical models to approximate the outcome of complex physical models and simulations such as FEA and CFD. The main purpose of implementing surrogate models is to accelerate the generation or prediction of the target objectives with high accuracy, which provides greater efficiency for optimization tasks using surrogate models. Taking the CFD simulation in this paper as an example, it takes around 10 min to run the numerical simulation on a desktop with Intel i5 CPU. While a typical prediction given by the surrogate model only costs 1 millisecond, which is at least 5 orders of magnitude faster. This fast prediction feature is one reason numerous surrogate models have been developed in several fields for a variety of applications. For example, deep neural networks (DNNs) are trained to predict the deformation of a soft pneumatic joint by given pressure, force, and torque inputs with high accuracy and efficiency (Zhang et al., 2022). In this study, regression models including linear, polynomial, and Gaussian process (GP) regression are applied considering the number of input design variables and complexity of the surrogate modeling (Galton, 1886; Quinonero-Candela & Rasmussen, 2005; Stigler, 1974). The linear regression model fits a linear relationship between the scalar output and multiple input variables. Weight coefficients are assigned and multiplied to each input variable and the estimation is the sum of all products with an intercept coefficient added at the end to offset the results. The weight coefficients and intercept coefficients are derived until the difference between estimation and true value is minimized. Polynomial regression is an elevated regression model with a non-linear relationship between the input variables, where the fitted relationship is correlated to a polynomial order of input variables such as the product of multiple input variables or the power of a single input variable. The GP regression constructs the relationship through a joint probability distribution over the input variables. The prediction of GP interpolates the observations through a pre-defined kernel, which controls the shape of the Gaussian function at specific points based on the similarity between actual true values and predictions. The GP prediction is also probabilistic, which provides empirical confidence intervals associated with the predicted values. This feature enables further refitting and exploration in the region of interest. For ML implementation, the Python package scikit-learn is used to fit the regression models mentioned above (Pedregosa et al., 2011). The collected CFD data points are randomly split into two datasets with an 8:2 ratio corresponding to the training to testing dataset ratio. The ML models are fitted on the training dataset and evaluated using the testing dataset.

Optimization algorithms With the obtained surrogate models, the performance of the output metrics can be improved by analyzing the analytical models. Common optimization algorithms can be classified into two categories based on objective functions. For differentiable objective functions, where the derivative can be calculated for any given point in the input space (e.g. DNNs, regression models), optimization algorithms such as gradient descent (GD), Newton's method, and line search can be applied. For non-differentiable objective functions, where the derivative cannot be easily calculated, population algorithms such as genetic algorithms (GA) provide a feasible choice. In addition to the conventional optimization process being fully iterated based on the surrogate model, elevated optimization methods taking advantage of CFD simulation results during the optimization iteration are also developed. Adopting this method, the interaction between the CFD simulation and optimization algorithms is fulfilled during the iterative exploration process. Specifically, a small portion of data is needed to build a GP regression model, which approximates the analytical relationships using Gaussian distribution. By defining a proper evaluation (acquisition) function to trade off exploitation and exploration, a subsequent design will be proposed and feedbacked into the CFD simulation to get ground truth performance. The corresponding result will be concatenated into the initial batch of data points and a new GP model will be fitted to find the next design. This iterative on-the-fly method is called Bayesian optimization (BO), which has achieved successful results in various optimization applications (Gongora et al., 2020; Snoek et al., 2012).

In this paper, random search, GA, stochastic gradient descent (SGD), and BO methods are applied to find the optimal faceplate designs that maximize flow uniformity. Here, random search is regarded as a baseline model and the collected dataset is used to find the best result for the random search algorithm. For GA and SGD methods, the trained ML surrogate model will be utilized. The flow of GA works as follows: (1) Randomly generate S candidates and evaluate through the ML model; (2) Rank the candidates based on performance, keep the top P candidates serving as parents, and discard the remaining $S - P$ candidates; (3) Generate P children based on a linear combination of P parents; (4) Randomly generate $S - P - P$ offspring in the new generation to keep the total generation size constant; (5) Iterate the previous 4 steps until convergence. The flow of SGD works as follows: (1) Initialize an initial design point to start; (2) Calculate the gradient with respect to each input variable; (3) Update the inputs toward the negative direction of the gradient; (4) Randomly generate another initial design point and iterate the previous 3 steps until convergence. The advantage of the conventional optimization methods including GA and SGD is that they have a faster convergence rate compared to random search due to their exploration core based on either evolutionary search or gradient evaluation. Additionally, the BO approach has more advantages when the final objectives are expensive to evaluate and there is no access to the derivatives (e.g., CFD physics simulations) as the algorithm explores the design field based on the probability of interest location.

Showerhead prototype and fabrication The porosity of the faceplate is modulated through the faceplate hole patterns by manipulating the diameter and number of holes in individual zones. Holes are positioned at the middle line of each zone, owing to the concentric circular pattern (around a center point). Hole diameters are sized between 1 and 4 mm depending on the porosity level. The showerhead prototype is built to show the potential to fabricate samples with various porosity distributions for conducting experiments. The prototype is fabricated mainly using 3D-printing and laser cutting. The faceplate is built using a laser cutter (ULTRA R9000) with casted acrylic; stem, porous baffle, and back plates are fabricated with polylactic acid (PLA) material using fused filament fabrication (FFF) 3D-printer (Ultimaker 3).

Results and discussion

Machine learning model implementation and sensitivity analysis With the parameterized CAD model and CFD simulation, both inputs and outputs of the ML model are defined. The inputs have 22 variables, which are comprised of 20 porosity values corresponding to each zone on the faceplate plus 2 radius values for the fillet design. The outputs are the 3σ values calculated based on the velocity profile at the plane of interest. Figure 2(a) shows the velocity magnitude at the plane of interest for the baseline faceplate design. For the data collection, a batch of designs (1000 data points, Fig. 2(b)) are generated randomly and fed into Ansys Fluent software to obtain the flow uniformity performance automatically. The 3σ value of the axial velocity is treated as ground truth and used to fit a regression model based on design input variables. Here, the axial velocity is determined as objective due to its dominant effect on the flow uniformity normal to the plane of interest compared to using velocity magnitude which includes a radial component. Both linear and polynomial regression models are implemented and reach R^2 values over 99%, which indicates a decent fit of the relationship. The regression model reaches its best performance (R^2 = 99.8%) when a second-order polynomial regression model is used (Fig. 2(c)). Here, all the predicted and ground truth values are normalized by the baseline value (1.0 indicates the baseline result). A smaller number indicates more flow uniformity (less velocity deviation).

Sensitivity analysis is performed to better understand the machine learning model, especially the importance of different design variables. Here, each variable's gradient (weight coefficient) is obtained from the linear regression model and shown in Fig. 2(d). A high absolute value of the weight coefficient indicates a higher influence on the objective output. From the curve, we can see the porosity value of the inner zones and fillet 2 have larger coefficients meaning more dominance with respect to the uniformity of the flow. Intuitively, these design variables are all major components in direct contact with the incoming flow and are believed to be the most critical in terms of design space. It is worth noting that sensitivity analysis also helps determine the number of zones on the faceplate design. Initially, 10 zones (10 porosity values) are considered to fully describe the design space. As sensitivity analysis concludes the importance of inner zones, a finer division (20 zones) is applied. For the inner zones, from zone 1 to zone 10, different porosity values are assigned to each zone; for the outer zones, from zone 11 to zone 20, identical porosity values are assigned to pairs of adjacent zones (e.g., zones 11 and 12 share the same porosity value, zones 13 and 14 share another porosity value). This division maintains the same spacing of concentric circles and the assignment of porosity values sufficiently controls the design space as well.

Conventional optimization approach and optimized faceplate designs With the trained ML surrogate models, the analytical relationship between input design variables against output objective is obtained. Hence, optimal designs can be generated by evaluating their performance based on the surrogate model by applying state-of-the-art optimization algorithms. Here, two optimization methods GA and SGD are utilized to find the optimal faceplate design. To simplify the optimization process, fillet design is not included in this subsection (fillet radii are zero). Results with fillet design will be demonstrated with an elevated optimization method in the next subsection. Figure 3(a) shows the optimization process of GA using and as defined in the material and methods section. We can see a stepped downward curve for the performance profile indicating more uniformity (better results) reached by new generations. Figure 3(b) shows the performance profile for an iteration of SGD method achieving the optimal result. As the inputs (porosity values) keep decreasing during the gradient descent updates, a stop limit of 0.01 is set to prevent the porosity value from becoming too small or negative. The iteration stops when four porosity values reach the stop limit considering the feasibility of fabricating the designed prototype. The vertical axis of the optimization performance curve is the normalized non-uniformity value (3σ value) by dividing the baseline result. Both optimization methods find their top design reaching 5.7% more uniformity in their best optimization iteration. Taking the best design obtained from GA as an example, the porosity distribution with zone information is shown in Fig. 3(c) and its schematic design is displayed using 1600

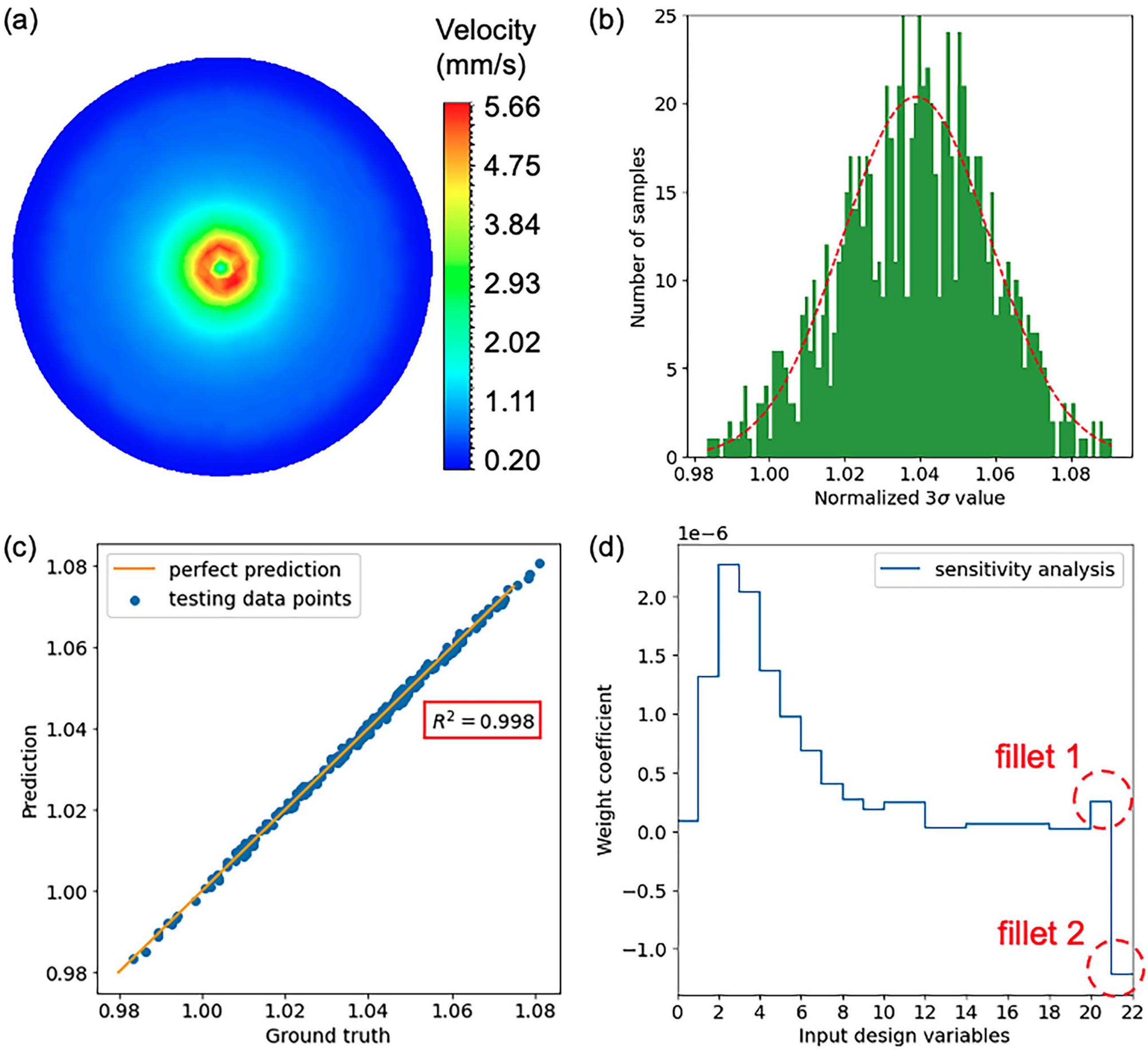

Fig. 2 (a) CFD results showing velocity profile on the plane of interest. (b) The distribution of collected data points with horizontal axis showing normalized 3σ value against the baseline. The red dashed curve shows a fitted normal distribution of the samples. (c) ML model predictions compared to ground truth of testing data points showing high accuracy. (d) Sensitivity analysis of the input design variables including 20 porosity values and 2 fillet radius designs

holes with diameters ranging from 1 to 4 mm shown in Fig. 4. The physical prototype is fabricated using laser cutting and 3D-printing method mentioned in the material and methods section and the size of the assembled prototype is 240 mm · 240 mm · 350 mm, which is a 73% scaled version of the original model due to the build size constraints of the 3D-printer and laser cutter (Fig. 4). Additionally, superior designs and inferior designs are presented in Fig. 5(a) for both optimization methods. The first row shows three inferior designs obtained by GA algorithm. We can see dramatic variations in the porosity values across the zones, which is not ideal when aiming to achieve a uniform flow distribution. The second and third rows show three superior designs for GA and SGD methods. They all share a similar porosity tendency at the inner zones as porous baffle is in the center and directly interacts with the inlet flow. For inner zones, a large porosity value is first assigned to zone 1 and has a gradually increased tendency in the direction away from the origin. This indicates that such porosity pattern is beneficial to generate a more uniform flow distribution. Besides showing the results of faceplate designs, the flow profiles of different designs are also compared at the plane of interest.

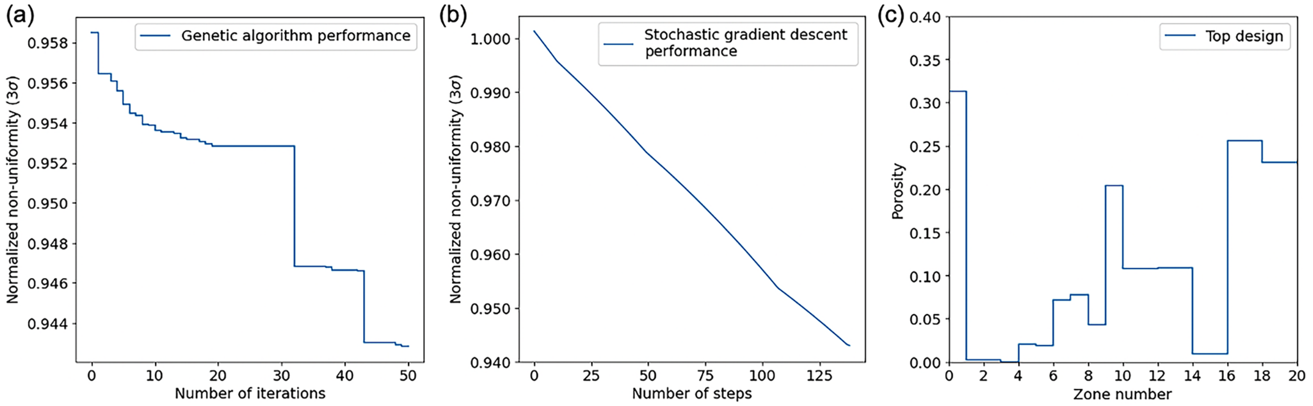

Fig. 3 (a, b) Optimization performance of GA and SGD method. (c) The top design obtained from GA method showing the porosity profile with respect to zone number

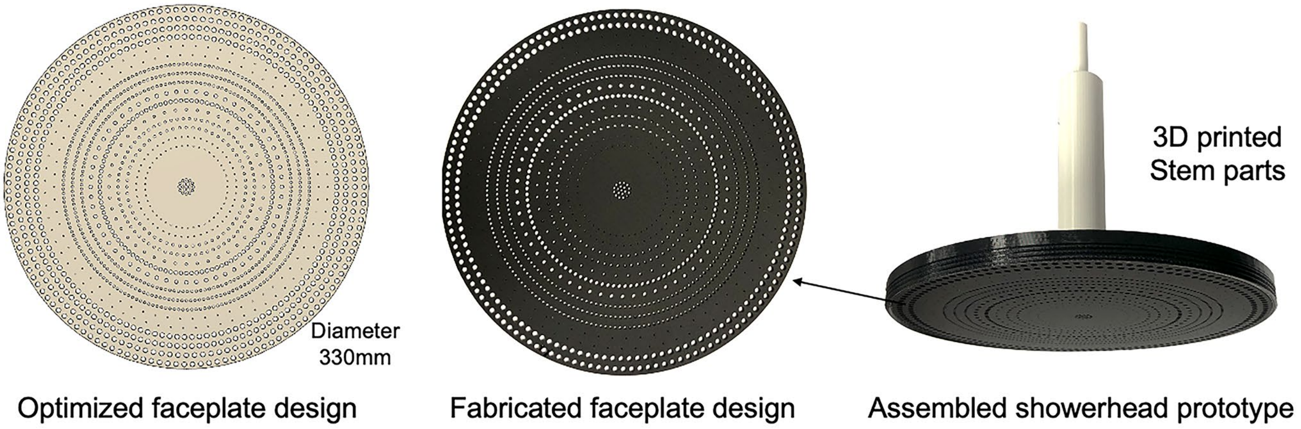
Diameter 330mm
Optimized faceplate design
Fabricated faceplate design
3D printed Stem parts
Assembled showerhead prototype

Fig. 4 Fabricated faceplate prototype based on proposed top faceplate design from genetic algorithm

In Fig. 5(b), axial velocity distribution along the radial distance is presented for the baseline design, an inferior design, and a superior design from GA. The green curve from the optimized faceplate design shows the best performance for flow uniformity. The velocity at a radial distance close to the origin is effectively lowered with the optimized faceplate design, which also leads to more flow uniformity at the full domain.

Bayesian optimization (BO): an elevated approach of higher efficiency at a lower computational cost By using the conventional optimization methods introduced in the previous section, a series of optimal designs are obtained showing remarkable improvement in flow uniformity. However, before the optimization step, a large training dataset is required (1000 data points) to fit an accurate ML surrogate model, which is time-consuming.

During the optimization iteration, the proposed design is generated solely from the ML surrogate model. The lack of communication between the CFD simulation and ML model during the optimization process can potentially cause propagating errors in the performance of the proposed designs. Hence, to accelerate and improve the optimization process, a BO based approach is applied which saves a large amount of data preparation cost and computational time (Gongora et al., 2020; Snoek et al., 2012). Here, a small portion of data (50 data points in our case) is initially used to build a GP model. A radial basis function (RBF) kernel is used due to its Gaussian form and wide applicability. The probability of improvement (PI) function is applied as the acquisition function to select the next exploration point, which has the highest probability of superior performance. When the subsequent design is proposed, it is fed back into the CFD simulation to get the ground truth performance. The corresponding result is concatenated into the initial batch of data points and a new GP model is fitted to propose the next design. Figure 6(a) shows the BO learning performance for 50 iterations and reaches 5.5% more uniformity when no fillet design is involved. When fillet designs are taken into consideration, based on the sensitivity analysis shown in Fig. 2(d), fillet 1 has a positive correlation with non-uniformity, while fillet 2 has a negative correlation

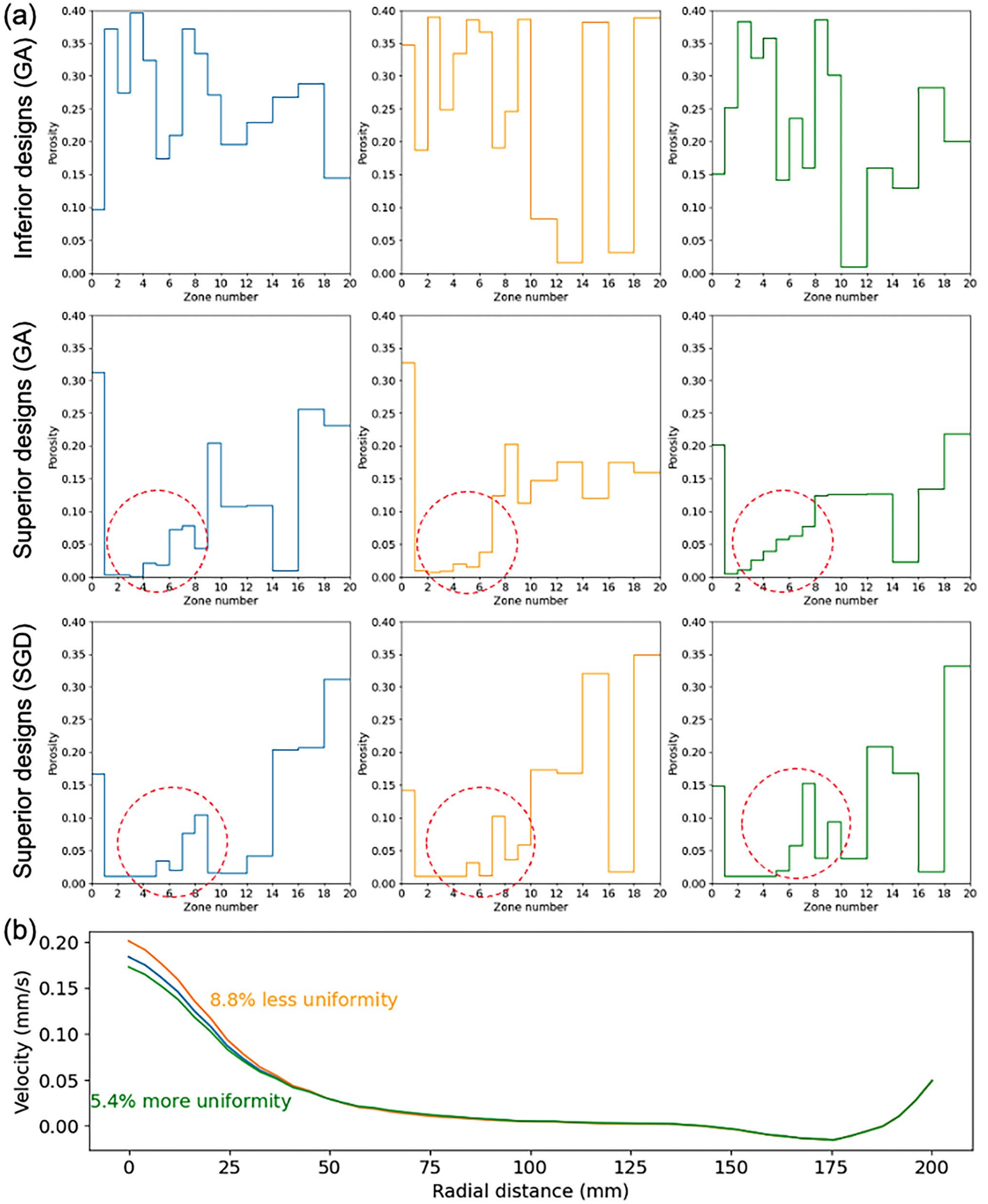

Fig. 5 (a) Comparison of superior and inferior designs generated from GA and SGD algorithm. (b) Axial velocity profiles along the radial distance are presented for the baseline design (blue curve), an inferior design (yellow curve), and a superior design (green curve)

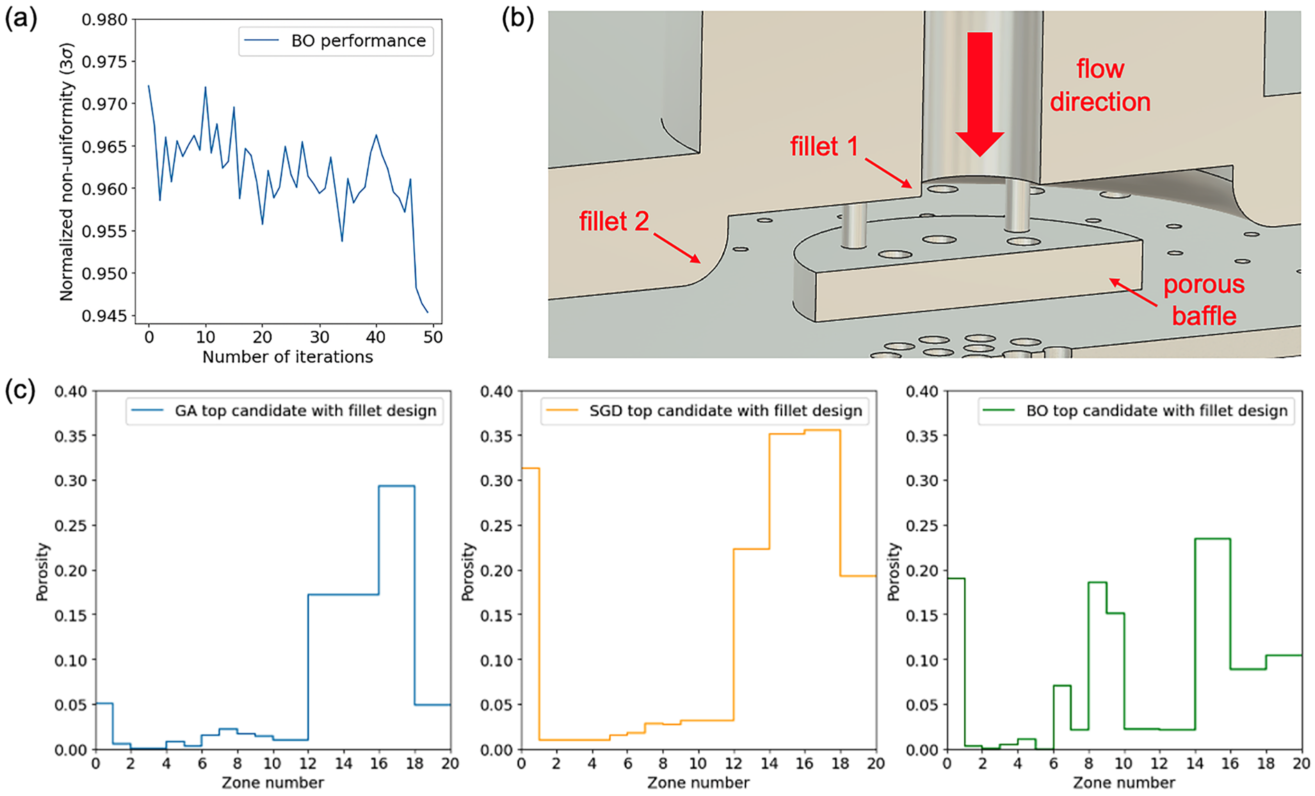

Fig. 6 (a) Optimization performance of BO method. (b) CAD model showing the best fillet designs with fillet 1 = 0 mm and fillet 2 = 3 mm. (c) Top candidate designs for three optimization methods (GA, SGD, and BO) with fillet designs

Table 1 Summary of the top designs from four optimization approaches with and without fillet designs

Algorithms	Improvement of top design with faceplate		Improvement of top design with faceplate + fillet	
	ML model prediction	CFD simulation	ML model prediction	CFD simulation
Random search	/	+ 1.7%	/	+ 3.6%
Genetic algorithm	+ 5.7%	+ 5.3%	+ 10.0%	+ 9.9%
Stochastic gradient descent	+ 5.7%	+ 5.4%	+ 9.7%	+ 9.6%
Bayesian optimization	/	+ 5.5%	/	+ 9.6%

with non-uniformity. Hence, the value of fillet 1 needs to be minimized (set to 0 mm) and the value of fillet 2 should be maximized (set to 3 mm considering the design space limitation) (Fig. 6(b)). We conclude that adopting such an optimized pair of fillet designs is beneficial for flow uniformity. With the determined fillet designs creating the optimal flow uniformity, a top candidate design with 9.6% more uniformity is obtained through the BO method. The porosity profile of the best candidate from all three algorithms (GA, SGD, and BO) are presented in Fig. 6(c) with their detailed flow uniformity performance shown in Table 1. The added fillet design leads to another boost in generating a more uniform flow for all the optimization algorithms and reaching an optimal of 10% increased uniformity using GA method. Although BO does not show the best performance in the fillet design case, it still reaches competitive results considering that it only uses a tenth of the data points compared to the other algorithms. It is believed that with more learning data and optimization iterations, BO method can achieve the same or even better performance compared to 10% more flow uniformity. By comparing the porosity profile obtained with fillet design and non-fillet design, the porosity values of the inner zones tend to be zero (a solid region). The results match the sensitivity analysis as the inner zones have a positive correlation with non-uniformity, which needs to be minimized to increase uniformity. The fillet design, too, largely contributes to the spreading of inlet flow to the remote end of the showerhead, which weakens the effect of porous baffle and the inner zones of the faceplate below it.

Conclusions and perspectives

In summary, a hardware optimization framework is established through a case study of showerhead design for the semiconductor deposition process. ML surrogate models are trained based on data collected from an efficient CFD simulation model. Coefficient of determination score R^2 =99.8% is reached for the best ML regression model indicating a near-perfect fit of the relationship between design variables and flow uniformity. Sensitivity analysis is also conducted showing greater importance of decreasing the porosity values in the inner zone and fillet designs. Conventional optimization methods such as random search, GA, and SGD approaches are conducted, and multiple optimal showerhead designs are proposed. The top candidate design with refined fillet radii reaches 10% more uniformity compared to the baseline showerhead prototype. Additionally, Bayesian optimization is further implemented showing a competitive performance (9.6% more uniformity) with only one-tenth of the data points required against conventional methods. If the time of optimization iteration is considered negligible compared to the data collection process, the BO approach accelerates the exploration of optimal designs 10 times faster than the conventional optimization algorithms. Lastly, the proposed top candidate design is fabricated using laser cutting and 3D-printing for demonstration of manufacturing feasibility and potential use in future experimental validation. An experimental setup based on the deposition process will be built for validating the performance of showerhead designs in our future work. A qualitative comparison of the flow uniformity performance between proposed optimal designs and baseline prototypes can be studied by visualizing and measuring the flow velocity using anemometers. Moreover, the established framework is envisioned to be augmented to encompass multiple objective functions (e.g., flow and thermal uniformity) and more design variables (e.g., hardware design parameters and process parameters) based on problem settings. Finally, prior knowledge obtained from CFD simulations can be used as guidance to design better acquisition functions used to propose new designs during the BO process. A similar approach has been validated in optimizing the toughness of 3D-printed structures using prior simulation results obtained from finite element analysis (FEA) (Gongora et al., 2021). A faster convergence rate and better performance are achieved using the FEA-informed BO method. This method can be transferred to our framework by involving discrepancy of simulation ground truth and GP surrogate model prediction during the calculation of acquisition function to explore a better subsequent design. It is believed that the developed framework can be expanded and adapted to a platform with multiple objectives and advanced algorithms to find solutions for different kinds of optimization problems in manufacturing systems such as semiconductor fabrication and additive manufacturing.

Acknowledgements This work used the Extreme Science and Engineering Discovery Environment (XSEDE) Bridges system, which is supported by National Science Foundation (Fund number ACI-1548562). The authors acknowledge support from the National Science Foundation (Fund Number: DMREF-2119276) and Lam Research Unlock Ideas Program.

References

Chandrasekharan, R., Sangplung, S., Swaminathan, S., Pasquale, F., Kang, H., Lavoie, A., Augustyniak, E., Sakiyama, Y., Baldasseroni, C., & Varadarajan, S. (2019). *Low volume showerhead with faceplate holes for improved flow uniformity*.

Chen, W. C., Lee, A. H. I., Deng, W. J., & Liu, K. Y. (2007). 2007/05/01/). The implementation of neural network for semiconductor PECVD process. *Expert Systems with Applications*, *32*(4), 1148–1153. https://doi.org/10.1016/j.eswa.2006.02.013.

Chen, C. T., & Gu, G. X. (2021). Learning hidden elasticity with deep neural networks. *Proceedings of the National Academy of Sciences, 118*(31).

DePinto, G., & Wilson, J. (1990). Optimization of LPCVD silicon nitride deposition process by use of designed experiments. IEEE/SEMI Conference on Advanced Semiconductor Manufacturing Workshop

Ding, Y., Zhang, Y., Ren, Y. M., Orkoulas, G., & Christofides, P. D. (2019). 2019/11/01/). Machine learning-based modeling and operation for ALD of SiO2 thin-films using data from a multiscale CFD simulation. *Chemical Engineering Research and Design*, *151*, 131–145. https://doi.org/10.1016/j.cherd.2019.09.005.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263.

Gongora, A. E., Xu, B., Perry, W., Okoye, C., Riley, P., Reyes, K. G., Morgan, E. F., & Brown, K. A. (2020). A bayesian experimental autonomous researcher for mechanical design. *Science advances*, *6*(15), eaaz1708.

Gongora, A. E., Snapp, K. L., Whiting, E., Riley, P., Reyes, K. G., Morgan, E. F., & Brown, K. A. (2021). Using simulation to accelerate autonomous experimentation: a case study using mechanics. *Iscience*, *24*(4), 102262.

Janai, J., Güney, F., Behl, A., & Geiger, A. (2020). Computer vision for autonomous vehicles: problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, *12*(1–3), 1–308.

Jin, Z., Zhang, Z., Demir, K., & Gu, G. X. (2020). Machine learning for advanced additive manufacturing. *Matter*, *3*(5), 1541–1556.

Jin, Z., Zhang, Z., Ott, J., & Gu, G. X. (2021). Precise localization and semantic segmentation detection of printing conditions in fused filament fabrication technologies using machine learning. *Additive Manufacturing*, *37*, 101696.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., & Potapenko, A. (2021). Highly accurate protein structure prediction with AlphaFold. *nature*, *596*(7873), 583–589.

Lee, S., Zhang, Z., & Gu, G. X. (2022). Generative machine learning algorithm for lattice structures with superior mechanical properties. *Materials Horizons*, *9*(3), 952–960.

Li, J., Fei, Z., Xu, Y., Wang, J., Fan, B., Ma, X., & Wang, G. (2018). Study on the optimization of the deposition rate of planetary GaN-MOCVD films based on CFD simulation and the corresponding surface model. *Royal Society Open Science*, *5*(2), 171757. https://doi.org/10.1098/rsos.171757.

Liao, C. C., Hsiau, S. S., & Chuang, T. C. (2018). 2018/01/01). Modeling and designing a new gas injection diffusion system for metalorganic chemical vapor deposition. *Heat and Mass Transfer*, *54*(1), 115–123. https://doi.org/10.1007/s00231-017-2110-8.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Quinonero-Candela, J., & Rasmussen, C. E. (2005). A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, *6*, 1939–1959.

Selep, M. J., Borth, A. J., Wiltse, J. M., Slevin, D. M., & Madsen, E. (2019). *Chemical vapor deposition shower head for uniform gas distribution*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., & Lanctot, M. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, *529*(7587), 484–489.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms.Advances in neural information processing systems, *25*.

Stigler, S. M. (1974). Gergonne's 1815 paper on the design and analysis of polynomial regression experiments. *Historia Mathematica*, *1*(4), 431–439.

Xia, H., Xiang, D., & Mou, P. (2014). Simulation-Based optimization of a Vector Showerhead System for the control of Flow Field Profile in a Vertical Reactor Chamber. *Advances in Mechanical Engineering*, *6*, 525102. https://doi.org/10.1155/2014/525102.

Yu, C. H., Wu, C. Y., & Buehler, M. J. (2022). Deep learning based design of porous graphene for enhanced mechanical resilience. *Computational Materials Science*, *206*, 111270.

Zhang, Z., Jin, Z., & Gu, G. X. (2022). Efficient pneumatic actuation modeling using hybrid physics-based and data-driven framework. *Cell Reports Physical Science*, *3*(4), 100842.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.