npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

# Factorial design analytics on effects of material parameter uncertainties in multiphysics modeling of additive manufacturing

Amanda Giam[1,2], Fan Chen[2], Jiaxiang Cai[3]✉ and Wentao Yan[1,2]✉

A bottleneck in Laser Powder Bed Fusion (L-PBF) metal additive manufacturing (AM) is the quality inconsistency of its products. To address this issue without costly experimentation, computational multi-physics modeling has been used, but the effectiveness is limited by parameter uncertainties and their interactions. We propose a full factorial design and variable selection approach for the analytics of main and interaction effects arising from material parameter uncertainties in multi-physics models. Data is collected from high-fidelity thermal-fluid simulations based on a 2-level full factorial design for 5 selected material parameters. Crucial physical phenomena of the L-PBF process are analyzed to extract physics-based domain knowledge, which are used to establish a validation checkpoint for our study. Initial data visualization with half-normal probability plots, interaction plots and standard deviation plots, is used to assess if the checkpoint is being met. We then apply the combination of best subset selection and the LASSO method on multiple linear regression models for comprehensive variable selection. Analytics yield statistically and phyiscally validated findings with practical implications, emphasizing the importance of parameter interactions under uncertainty, and their relation to the underlying physics of L-PBF.

*npj Computational Materials* (2023)9:51 ; https://doi.org/10.1038/s41524-023-01004-9

## INTRODUCTION

Laser Powder Bed Fusion (L-PBF) is a commonly used metal additive manufacturing (AM) process that is capable of manufacturing products with complex geometries[1]. A major barrier that hinders a wide application of L-PBF in industry is the quality inconsistency of its products. In practice, it is difficult to measure some parameters to high precision due to constraints such as powder oxidation and temperature fluctuation. This introduces substantial uncertainties to input parameters such as the absorbed laser power and surface tension temperature sensitivity. Consequently, these input uncertainties in reality cause variations in the quality of the L-PBF products. These variations may be amplified by interaction effects arising from these uncertainties. To alleviate this issue, the AM community resort to multi-physics modeling, where input parameters can be precisely set, and costly experimentation can be circumvented. Multi-physics modeling refers to the application of high-fidelity mathematical models, numerical tools, and software technologies that closely approximates the actual L-PBF process by incorporating simultaneous physical phenomena of the process, e.g., heat transfer, fluid flow, powder melting and solidification[2]. Nevertheless, input parameter uncertainties cannot be eliminated in multi-physics models due to the lack of knowledge on the exact values of the parameters[3,4]. As a result, the uncertainties from the inputs and their interactions, propagate to essential model outputs such as the melt pool dimensions. The melt pool dimensions are key performance indices (KPIs) of the L-PBF process because the melt pool influence microstructure[5], thus affecting the structural integrity and quality of the final product[6]. This drives the need for model-based uncertainty quantification (UQ), which is the process of investigating the effects of uncertainty sources on the output quantities of interests (QoIs) in computational models[7].

UQ is an interdisciplinary field that involves both physical and statistical aspects. The physical aspect of UQ entails multi-physics modeling of the L-PBF process and/or actual L-PBF experimentation to collect data for subsequent analysis. The statistical aspect of UQ encompasses the application of statistical techniques before or after data collection, such as the design of experiments (DOE), sensitivity analyses and/or surrogate modeling, which are often used to mitigate or bypass heavy computational cost. The UQ studies for AM in literature mostly obtain data from simulation models such as continuum-based thermal models using the Finite Element Method (FEM) or semi-analytical thermal-conduction models based on the homogeneous continuum assumption, which are less accurate in the physical aspect as compared to computational fluid dynamic (CFD) models resolving the thermal fluid-flow behaviors of individual powder particles[8]. The limited accuracy of the low-fidelity models hinders the effectiveness of the previous UQ studies. For example, Moges et al.[3] has employed a fractional factorial DOE to analyze the main and interaction effects of input parameters in semi-analytical and finite element models. By performing a normal probability plot of the data from the simulation models, the absorbed laser power and thermal conductivity were identified as the most significant input parameters. The major advantage of his study is the reasonable computational cost. However, the semi-analytical and FEM models are not the most accurate, and the use of a high-fidelity model, i.e., the thermal fluid-flow model, will be better instead. Tapia et al.[9] has studied the influence of laser parameters on melt pool characteristics by applying the polynomial chaos expansion (PCE) framework on data from two simulation models—where the first

[1]Integrative Sciences and Engineering Programme, NUS Graduate School, National University of Singapore, Singapore 119077, Singapore. [2]Department of Mechanical Engineering, National University of Singapore, Singapore 117575, Singapore. [3]Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore 117576, Singapore. ✉email: jiaxiang@u.nus.edu; mpeyanw@nus.edu.sg

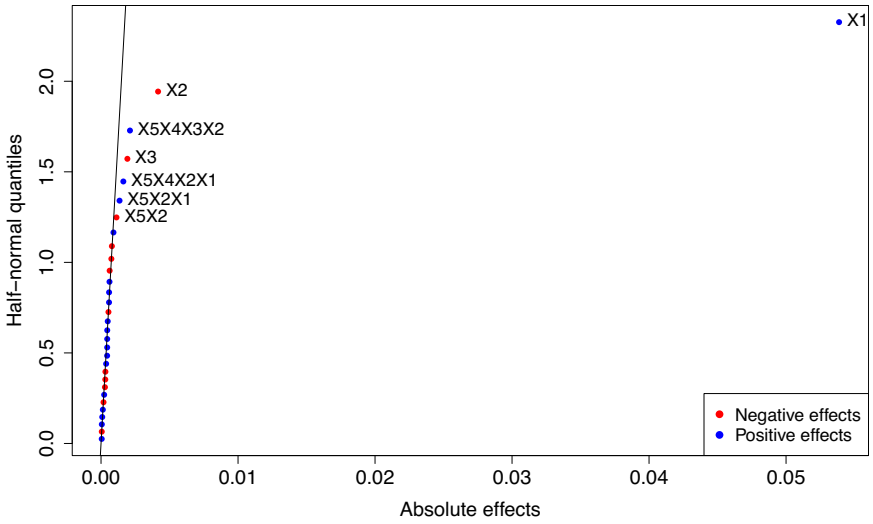


**Fig. 1 Half-normal plot of all effects, where input factors ($P_A$, $\lambda$, $\mu$, $\gamma$, $-d\gamma/dT$), are represented by ($X_1$, $X_2$, $X_3$, $X_4$, $X_5$), respectively.** The X-axis represents the absolute value of the effects, and the Y-axis shows the corresponding probability of observing such an effect. Only the outliers (i.e. significant effects) are labeled on the plot.

model is a reduced order thermal model (Eagar-Tsai model), and the second model is a finite element thermal model. Although the PCE framework is a decent tool for UQ, these simulation models are less accurate than high-fidelity thermal fluid-flow models. Wang et al.[10] has utilized the Gaussian Process (GP) surrogate model to perform a global sensitivity analysis on parameters affecting microstructure. Despite the GP model being a robust surrogate for UQ, the simulation data used to train the surrogate model came from a finite-element based thermal model, which is nevertheless not as accurate as thermal fluid-flow models.

Most of the previous studies perform UQ based on lower-fidelity simulation models[3,9–11]—which although incur lower computational cost, do not accurately capture the complexities of the physical L-PBF process. Such a limitation can make it difficult to apply the UQ results in an industrial setting. Additionally, there is a lack of studies on interaction effects, which are suspected to be significant[3]. Hence there is a pressing demand for a UQ study anchored with a high-fidelity multi-physics model, which can provide practical insights at a reasonable computational cost. To bridge this gap, we propose the use of a computationally efficient factorial design, and a comprehensive variable selection approach, to analyze the effects arising from input parameter uncertainties and their interactions in a high-fidelity multi-physics model, i.e., the thermal-fluid model. Through the use of analytics coupled with the strength of high-fidelity multi-physics modeling, we aim to provide practical insights for the AM community. The choice of the thermal fluid model achieves sufficient accuracy for the physical aspect of UQ. In addition, our methodology also accounts for the statistical aspect of UQ through the application of DOE, surrogate modeling, sensitivity analysis and uncertainty analysis. Moreover, the statistical results of this work is carefully evaluated with physics-based domain knowledge to demonstrate result consistency and attain statistical-physical validation. These jointly validated results then provide practical guidance to the simulation and experimental groups directly. Overall, the well-established techniques employed in the study are straightforward for the different communities in UQ such as simulation groups, industrial practitioners, and data analysts. As such, the ease of result interpretation and facilitation of common understanding across the communities is made possible through the use of these techniques. The largest benefit of the factorial design and analysis is its capability to yield consistent, practical insights with low computational cost and complexity.

This paper aims to obtain practical insights by using the proposed factorial design along with variable selection and model analytics, to characterize the uncertainties due to five input material parameters (or factors) in the thermal fluid model, namely the laser power absorption ($P_A$), thermal conductivity ($\lambda$), viscosity ($\mu$), surface tension coefficient ($\gamma$), surface tension temperature sensitivity ($-d\gamma/dT$), quantifying their respective influences on the selected output variable—the melt pool depth ($Y$). Justification on the selection of these five material input parameters is provided in Section "Methods" . The remainder of the paper is organized as follows. Section "Results and discussion" reports the key findings of this paper from both statistical and physical perspectives with practical follow-up directions for simulation groups and industrial practitioners in metal AM. Section "Methods" presents the comprehensive methodology used in this paper including the design of experiments, thermal-fluid simulations, data visualization, variable selection and statistical-physical validation.

## RESULTS AND DISCUSSION

### Data visualization

The half-normal plot for all the effects of the $2^5$ factorial is displayed in Fig. 1, where the five input factors: ($P_A$, $\lambda$, $\mu$, $\gamma$, $-d\gamma/dT$), are denoted by ($X_1$, $X_2$, $X_3$, $X_4$, $X_5$), respectively for ease of representation. From Fig. 1, it can be seen that the main effect of $P_A$ is an obvious outlier, implying that it is highly suspected to be a significant factor. Such an observation agrees well with literature—it is universally agreed in the AM community that the absorbed laser power is a major factor with large influence on the melt pool geometry. The second most important factor that demonstrates considerable deviation from the fitted line is the main effect of $\lambda$. In addition, the main effect of $\mu$, 2-factor interaction effect of $\lambda * -d\gamma/dT$, as well as the higher-order interactions ($P_A * \lambda * -d\gamma/dT$, $P_A * \lambda * \gamma * -d\gamma/dT$, $\lambda * \mu * \gamma * -d\gamma/dT$) also exhibit deviations from the line. As these observed deviations are smaller, a considerable amount of subjectivity is involved in the assessment of their significance due to their proximity to the fitted line. Therefore, further analysis with a quantitative basis, i.e., hypothesis tests, will be conducted in Section "Variable selection and model analytics" to validate our prior conclusions. Since the half-normal plot has identified potentially significant interactions such as $\lambda * -d\gamma/dT$, we analyze all possible 2-factor interactions in detail using interaction plots.

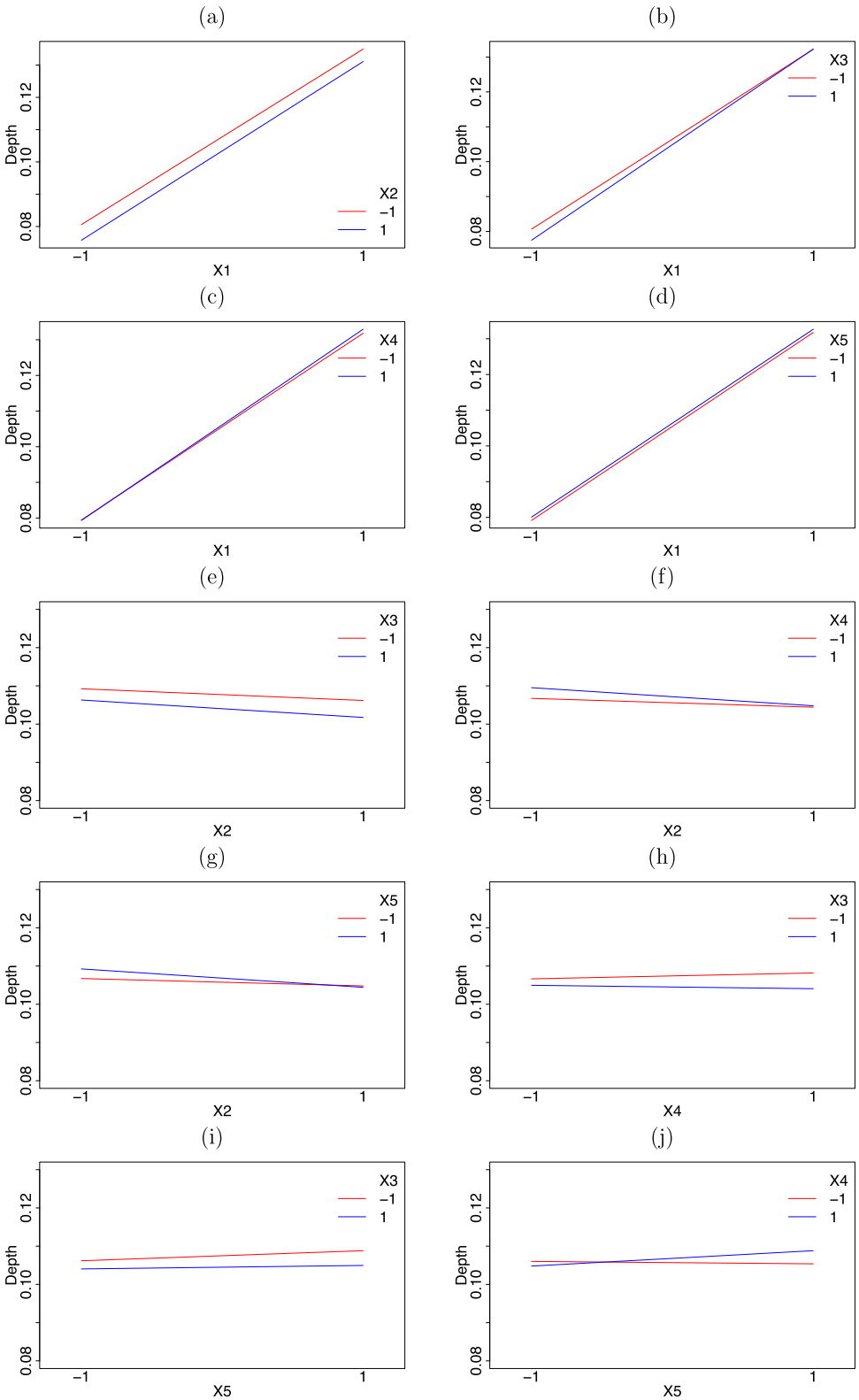


**Fig. 2 Interaction plots with the melt pool depth on the Y-axis and an input factor on the X-axis. a** Interaction effect of $P_A * \lambda$. **b** Interaction effect of $P_A * \mu$. **c** Interaction effect of $P_A * \gamma$. **d** Interaction effect of $P_A * -d\gamma/dT$. **e** Interaction effect of $\lambda * \mu$. **f** Interaction effect of $\lambda * \gamma$. **g** Interaction effect of $\lambda * -d\gamma/dT$. **h** Interaction effect of $\mu * \gamma$. **i** Interaction effect of $\mu * -d\gamma/dT$. **j** Interaction effect of $\gamma * -d\gamma/dT$.

The interaction plots for all $\binom{5}{2}$ 2-factor interactions are shown in Fig. 2, where the five input factors: ($P_A$, $\lambda$, $\mu$, $\gamma$, $-d\gamma/dT$), are denoted by ($X_1$, $X_2$, $X_3$, $X_4$, $X_5$), respectively for ease of representation. We first observe that the magnitude of interactions involved with $P_A$ are much larger than that of other factors, which is expected since the dominating influence of $P_A$ has already been established. The $P_A * \lambda$ interaction is not significant, as seen from the two parallel lines in Fig. 2a. Hence

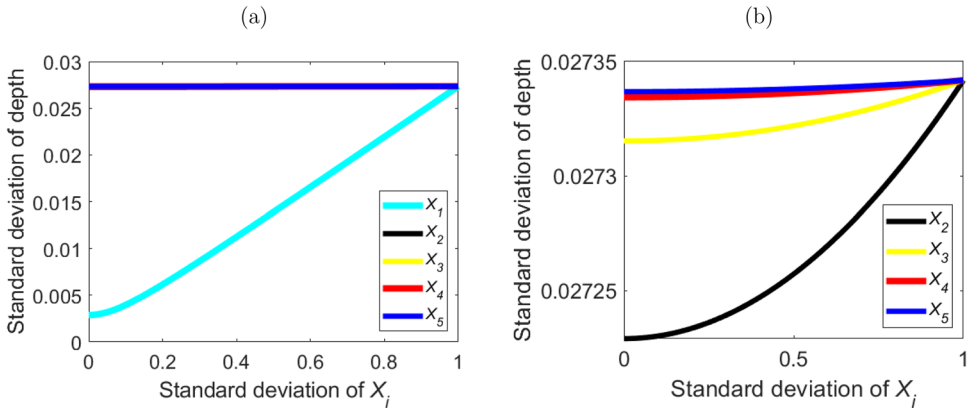


**Fig. 3 Standard deviation plots with melt pool depth deviation $\sigma_Y$ on the Y-axis, and standard deviation of main effect $X_i$ on the X-axis.** **a** All factor deviations, ($\sigma_{P_A}$, $\sigma_\lambda$, $\sigma_\mu$, $\sigma_\gamma$, $\sigma_{-\mathrm{d}\gamma/\mathrm{d}T}$). **b** Factor deviations, ($\sigma_\lambda$, $\sigma_\mu$, $\sigma_\gamma$, $\sigma_{-\mathrm{d}\gamma/\mathrm{d}T}$).

**Table 1.** Regression output for five MLR models formed by manual selection of variables.

| | Terms | Significant Factors at $\alpha = 0.1$ | Adjusted $R^2$ | Residual normality |
|---|---|---|---|---|
| $Y_{X_{i1}X_{i2}X_{i_3}X_{i_4}X_{i_5}}$, full model | 32 | NA | NA | NA |
| $Y_{X_{i1}X_{i2}X_{i_3}X_{i_4}}$, 4-factor interactions model | 31 | $P_A$, $\lambda$, $\mu$, $P_A * \mu$, $\lambda * -\mathrm{d}\gamma/\mathrm{d}T$, $P_A * \lambda * -\mathrm{d}\gamma/\mathrm{d}T$, $\lambda * \mu * \gamma * -\mathrm{d}\gamma/\mathrm{d}T$, $P_A * \lambda * \gamma * -\mathrm{d}\gamma/\mathrm{d}T$ | 0.99982 | Severe Violation |
| $Y_{X_{i1}X_{i2}X_{i_3}}$, 3-factor interactions model | 26 | $P_A$, $\lambda$ | 0.98642 | Some Violation |
| $Y_{X_{i1}X_{i2}}$, 2-factor interactions model | 16 | $P_A$, $\lambda$, $\mu$ | 0.99251 | Mild violation |
| $Y_{X_{i1}}$, main effects model | 6 | $P_A$, $\lambda$, $\mu$ | 0.993785 | Little violation |

the amount of absorbed laser power does not interact with the thermal conductivity of IN625, and these two factors may be calibrated independent of each other in experiments or simulations. On the other hand, the interactions of $P_A * \mu$, $\lambda * \gamma$ and $\lambda * -\mathrm{d}\gamma/\mathrm{d}T$ $\lambda * -\mathrm{d}\gamma/\mathrm{d}T$are likely significant due to the considerable degree of non-parallelism for the lines observed in Fig. 2b, f, and g, respectively. This implies that the absorbed laser power interacts with the viscosity of the IN625 material. The analysis also reveals an interaction between the thermal conductivity of IN625 and other material properties of IN625, such as the surface tension coefficient and temperature sensitivity of the surface tension. Therefore, these interactions should be taken into account in the calibration of L-PBF simulations and experiments. Further elaboration and follow-up guidance on the significant interactions can be found in Section "Practical interpretation with joint statistical-physical validation". As for the rest of the interactions, their significance are rather inconclusive, as there is subjectivity in determining the extent of non-parallelism of the lines. Due to the subjectivity, the significance of all interactions will be validated with a quantitative basis through the regression analysis in Section "Variable selection and model analytics". We next study the relationship between the input uncertainties of the main effects ($P_A$, $\lambda$, $\mu$, $\gamma$, $-\mathrm{d}\gamma/\mathrm{d}T$) and the output uncertainty of the melt pool depth.

Standard deviation plots based on the uncertainty function in Eq. (7), have been constructed in Fig. 3 to study the overall influence of each input parameter's standard deviation onto the output standard deviation. In these plots, the Y-axis represents the standard deviation of the response melt pool depth, while the X-axis represents the standard deviation of a coded input factor. A plot of the output depth's standard deviation, $\sigma_Y$ against the input standard deviations of the five factors ($\sigma_{P_A}$, $\sigma_\lambda$, $\sigma_\mu$, $\sigma_\gamma$, $\sigma_{-\mathrm{d}\gamma/\mathrm{d}T}$) is shown in Fig. 3a. In addition, another plot excluding $P_A$ is illustrated in Fig. 3b, where we consider $\sigma_Y$ against the four input standard deviations ($\sigma_\lambda$, $\sigma_\mu$, $\sigma_\gamma$, $\sigma_{-\mathrm{d}\gamma/\mathrm{d}T}$).

It is observed from Fig. 3a that the standard deviation of factor $P_A$ propagates the largest uncertainty to the output uncertainty, dominating the uncertainties propagated by the rest of the variables. A change in the input standard deviation of $P_A$ causes the largest change (approximately 0.02) in output standard deviation of the depth. This result aligns well with the prior results of the half-normal plot. It is intuitive that the most influential factor ($P_A$) will naturally contribute the largest uncertainty to the response melt pool depth. The propagation of uncertainty from the other four input variables ($\lambda, \mu, \gamma, -\mathrm{d}\gamma/\mathrm{d}T$) shown in Fig. 3b, is approximately in the order of magnitude of $10^{-5}$. In the absence of $\sigma_{P_A}$, the uncertainty propagated to the output depth from the four input variables ($\lambda, \mu, \gamma, -\mathrm{d}\gamma/\mathrm{d}T$) in descending order is: $\sigma_\lambda > \sigma_\mu > \sigma_{-\mathrm{d}\gamma/\mathrm{d}T} > \sigma_\gamma$.

Overall, data visualization through the half-normal plot, interaction plots and standard deviation plots validates our small-sample based analysis, as it correctly identifies $P_A$ as the most significant factor with a dominating influence on the response melt pool depth, which is consistent with existing literature.

## Variable selection and model analytics

Table 1 summarises the key output of the five MLR models formed via the systematic manual selection of variables, based on the recommended analysis procedure for a full factorial design[12]. The respective $p$ values and regression coefficients of the five models are provided in Supplementary Table 1. The full model, $Y_{X_{i1}X_{i2}X_{i_3}X_{i_4}X_{i_5}}$ has no meaningful results due to the lack of replicates for the deterministic simulation, causing zero degrees of freedom for the standard error (SE) of the coefficient estimates.

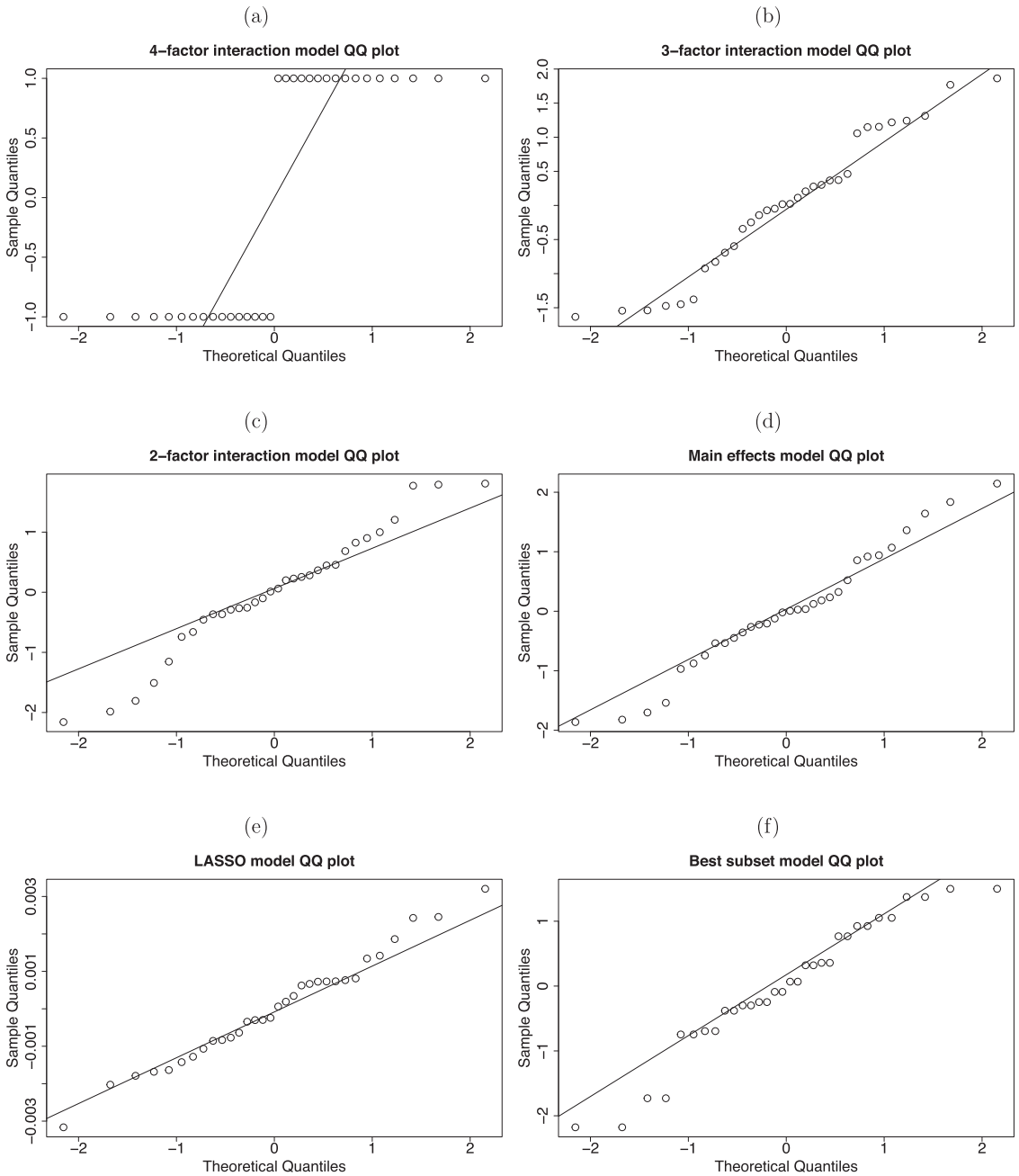


**Fig. 4 Quantile-quantile (QQ) plots for six MLR models. a** QQ plot of 4-factor interaction model, $Y_{X_{i1}X_{i2}X_{i_3}X_{i_4}}$. **b** QQ plot of 3-factor interaction model, $Y_{X_{i1}X_{i2}X_{i_3}}$. **c** QQ plot of 2-factor interaction model, $Y_{X_{i1}X_{i2}}$. **d** QQ plot of main effects model, $Y_{X_{i1}}$. **e** QQ plot of LASSO model, $Y_{X_L}$. **f** QQ plot of best subset model, $Y_{X_k}$.

Thus we analyze the reduced models instead. It is observed that the results on variable significance are not consistent across the reduced models in Table 1. From the QQ plots of the models provided in Fig. 4, it is found that most models have some degree of violation of the residual normality assumption, with the main effects model having the least. All four reduced models have decent adjusted $R^2$ values, with the 4-factor interaction model, $Y_{X_{i1}X_{i2}X_{i_3}X_{i_4}}$, attaining the highest value. However, $Y_{X_{i1}X_{i2}X_{i_3}X_{i_4}}$ is not an appropriate model choice due to severe violation of residual normality as seen in Fig. 4a. Among the four reduced models, the main effects model, $Y_{X_{i1}}$, appears to be the most appropriate model choice as it strikes the best balance between residual normality and adjusted $R^2$. Nevertheless, $Y_{X_{i1}}$ is an oversimplified model that cannot provide insight on interactions. Hence forming an optimal model for result interpretation via the manual selection of variables is challenging. This motivates automated variable selection such as best subset selection, to perform an exhaustive search for a model containing the optimal number of variables ($k$) and the best combination of variables.

The optimal $k$ is determined by criterion such as Mallow's $C_p$ and the adjusted $R^2$ as shown in Fig. 5. It is observed that the residual sum of squares (RSS) converges to a minimum value when $k$ is approximately equal to 23. The value of $C_p$ converges to a minimum when $k = 25$. For the adjusted $R^2$, the maximum value is attained at approximately $k = 27$. In order to strike a balance between goodness of fit and not overfitting the model, the optimal number of variables is determined as $k = 25$. The best

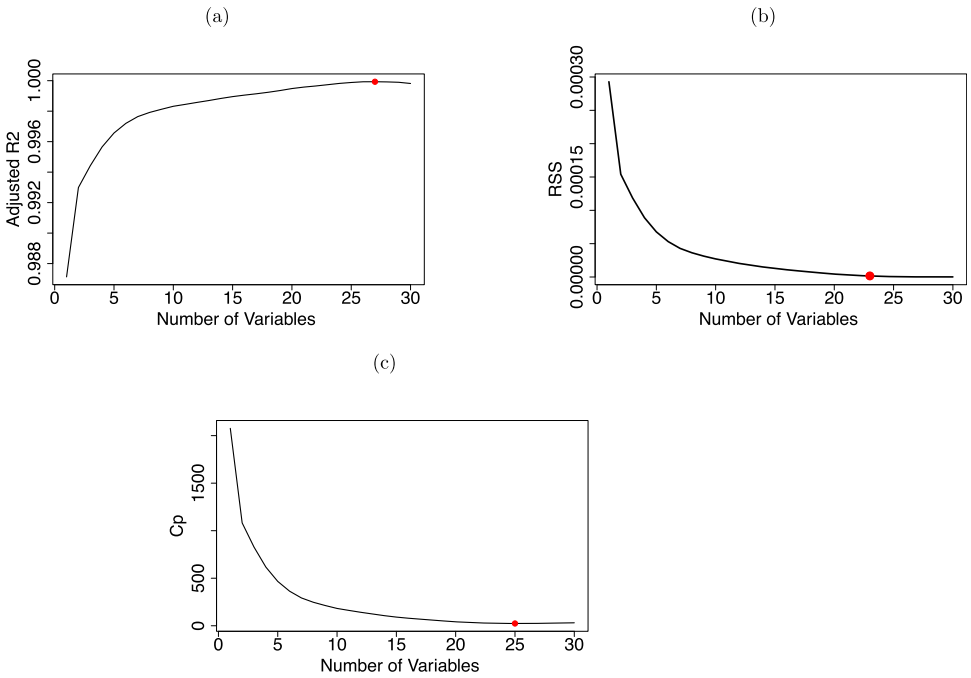


**Fig. 5 Criterion to determine optimal number of variables *k* for the best model. a** Adjusted $R^2$ values for different values of *k*. **b** Residual sum of squares (RSS) values for different values of *k*. **c** $C_p$ values for different values of *k*.

combination of 25 variables is also selected from the best subset selection algorithm as displayed in Table 2.

In the best subset model $Y_{X_k}$, all terms are statistically significant at $\alpha = 0.1$, and the model has the highest adjusted $R^2$ value of 0.9999. The QQ plot in Fig. 4f reflects that most of the model's residuals are reasonably scattered around the best fit line, apart from slight deviations at the tail ends. Thus, the best subset model has the best performance metrics compared to the other models formed via manual selection of variables. However, it is challenging to interpret some of its results regarding the significance of higher-order interactions, i.e., $P_A * \lambda * -\mathrm{d}\gamma/\mathrm{dT}$, $\lambda * \mu * \gamma * -\mathrm{d}\gamma/\mathrm{dT}$, $P_A * \lambda * \gamma * -\mathrm{d}\gamma/\mathrm{dT}$ as it is rare for higher-order interactions involving 3 factors and above to be significant according to literature[12]. As there is currently no evidence from physical domain knowledge to support the presence of higher order interactions involving more than 3 factors, we focus on interpreting the main effects and 2-factor interactions instead. The parameter ranking of the statistically significant main effects and 2-factor interactions is provided in Table. 3. It can be seen that the top three factors that influence the melt pool depth are the main effects of: $P_A$, $\lambda$ and $\mu$. This result is in agreement with that of the main effects model and half-normal plots. It is observed that some 2-factor interactions such as $\lambda * -\mathrm{d}\gamma/\mathrm{dT}$, $P_A * \mu$, $\lambda * \mu$, $\lambda * \gamma$, $P_A * -\mathrm{d}\gamma/\mathrm{dT}$, $P_A * \gamma$ might be more significant than the main effects of $\gamma$ and $-\mathrm{d}\gamma/\mathrm{d}T$. As for the significant higher-order interactions, there are a couple of possible reasons that could explain the result. A possibility is that standardization of the input variables affected the relative magnitude of the regression coefficients of the interactions with respect to the main effects. Hence the interpretation of standardized regression weights should always be conducted with caution, using domain knowledge as a reference to evaluate the statistical results[13,14]. Another reason for the significant higher-order interactions could be overfitting, where the model fits to the noise in the training data rather than the underlying pattern. Due to the considerably large number of model parameters relative to the sample size, $Y_{X_k}$ is prone to overfitting. To check this, we have used leave-one-out cross-validation (LOOCV) to compute both the train and test MSE of the best subset model. The resulting ratio of the test MSE and train MSE for the best subset model is then benchmarked it against that of the main effects model, $Y_{X_{i1}}$, to assess if overfitting occurs. From Table 4, it can be seen that both the train and test MSE values of $Y_{X_k}$ are small with a magnitude of $1.68*10^{-8}$ and $4.77*10^{-7}$ respectively. The train and test MSE values of $Y_{X_k}$ are also smaller than that of $Y_{X_{i1}}$, which could imply that $Y_{X_k}$ has better predictive power. However, when we examine the ratio of the test and train MSE, we observe that the best subset model's test MSE is around 28 times larger than that of its train MSE. This is a sign of overfitting since the performance on the training set significantly outperforms that of the test set, meaning the best subset model may not generalize well to unseen data. As a result, model $Y_{X_k}$ exhibits high variance and low bias, which undermines the interpretability of the model. Therefore, we will implement regularized regression to address overfitting to achieve a better balance between bias and variance.

To achieve repeatability and stability of our results, we use LASSO regression with bootstrapping to estimate the regularized regression coefficients and corresponding confidence intervals (CIs). The results are provided in Table 5. The optimal value of the regularization parameter, $\lambda_{reg}$, for the LASSO regression model is determined to be 0.0003398815 through LOOCV. The adjusted $R^2$ of the model is 0.9972896, indicating a good fit. Residual normality of the model is also satisfactory, as shown in Fig. 4e.

We first apply the LASSO method to the full model, and observe that the variable selection results are consistent with those of the best subset model, particularly for the main effects and 2-factor interactions. All variables identified as significant by the LASSO model $Y_{X_L}$ have also been identified as significant by the best subset model $Y_{X_k}$. The parameter ranking from the LASSO model as reported in Table 6 is nearly identical to that of the best subset model, with the exception of $\gamma$, $-\mathrm{d}\gamma/\mathrm{d}T$, $P_A * \gamma$, $P_A * -\mathrm{d}\gamma/\mathrm{dT}$. Although these four terms are not identified as significant in the LASSO model, we expect some tradeoff between bias and variance to occur from the regularization process, which could explain why these variables were not selected. The consistent parameter ranking adds validity to the previous results. It is observed that the application of LASSO to the full model provides confirmation of the results from the best subset model. The result consistency for both models increases confidence in the selected

**Table 2.** Best subset model regression coefficients and $p$ values of selected variables using a significance level of 0.1.

| Terms | Include? | Regression coefficients |
|---|---|---|
| (Intercept) | TRUE | 1.05839E−01*** |
| $P_A$ | TRUE | 2.69350E−02*** |
| $\lambda$ | TRUE | −2.08196E−03*** |
| $\mu$ | TRUE | −9.61593E−04*** |
| $\gamma$ | TRUE | 1.14047E−04* |
| $-d\gamma/dT$ | TRUE | 2.26852E−04*** |
| $P_A * \lambda$ | FALSE | 0 |
| $P_A * \mu$ | TRUE | 4.53708E−04*** |
| $P_A * \gamma$ | TRUE | 2.44496E−04*** |
| $P_A * -d\gamma/dT$ | TRUE | 3.08903E−04*** |
| $\lambda * \mu$ | TRUE | −3.96380E−04*** |
| $\lambda * \gamma$ | TRUE | −3.77203E−04*** |
| $\lambda * -d\gamma/dT$ | TRUE | −5.63809E−04*** |
| $\mu * \gamma$ | FALSE | 0 |
| $\mu * -d\gamma/dT$ | FALSE | 0 |
| $\gamma * -d\gamma/dT$ | FALSE | 0 |
| $P_A * \lambda * \mu$ | TRUE | 1.86421E−04** |
| $P_A * \lambda * \gamma$ | FALSE | 0 |
| $P_A * \lambda * -d\gamma/dT$ | TRUE | 6.75741E−04*** |
| $P_A * \mu * \gamma$ | TRUE | −1.55266E−04** |
| $P_A * \mu * -d\gamma/dT$ | TRUE | 2.25628E−04*** |
| $P_A * \gamma * -d\gamma/dT$ | TRUE | 2.24396E−04*** |
| $\lambda * \mu * \gamma$ | TRUE | 2.93934E−04 *** |
| $\lambda * \mu - d\gamma/dT$ | TRUE | 2.92017E−04*** |
| $\lambda * \gamma * -d\gamma/dT$ | TRUE | −1.42671E−04** |
| $\mu * \gamma * -d\gamma/dT$ | TRUE | −3.14491E−04*** |
| $P_A * \lambda * \mu * \gamma$ | TRUE | 2.22234E−04*** |
| $P_A * \lambda * \mu * -d\gamma/dT$ | TRUE | −2.71283E−04*** |
| $P_A * \lambda * \gamma * -d\gamma/dT$ | TRUE | 8.13880E−04*** |
| $P_A * \mu * \gamma * -d\gamma/dT$ | TRUE | −1.59792E−04** |
| $\lambda * \mu * \gamma * -d\gamma/dT$, | TRUE | 1.05624E−03*** |
| Adj. $R^2$ | 0.99988206 | |
| ***$p < 0.01$; **$p < 0.05$; *$p < 0.1$. | Standard error = 0.00005287 | for all coefficient estimates |

**Table 3.** Parameter ranking of main effects and 2-factor interactions in the best subset model.

| Ranking | Parameter |
|---|---|
| 1 | $P_A$ |
| 2 | $\lambda$ |
| 3 | $\mu$ |
| 4 | $\lambda * -d\gamma/dT$, |
| 5 | $P_A * \mu$ |
| 6 | $\lambda * \mu$ |
| 7 | $\lambda * \gamma$ |
| 8 | $P_A * -d\gamma/dT$ |
| 9 | $P_A * \gamma$ |
| 10 | $-d\gamma/dT$ |
| 11 | $\gamma$ |

variables, identifying a stable and interpretable set of variables that are relevant to the response.

Another interesting finding is that applying LASSO to the best subset model yields the same regression coefficients as those obtained from applying LASSO on the full model as presented in Table 5. The only discernible difference is in the width of the confidence intervals of the coefficient estimates, as shown in Fig. 6, but the variation is minimal and almost negligible. This finding suggests that the LASSO method effectively selects the most important variables regardless of which model it is applied on, making it a valuable tool not only for variable selection, but also for model refinement. The agreement between the coefficients obtained by applying LASSO on the full model and LASSO on the best subset model provides additional evidence of the robustness of the selected variables and the stability of the model, which can be seen as an added value of using LASSO and best subset selection in combination.

To further assess the performance of the LASSO model, we also use LOOCV to compare the ratio of test mean squared error (MSE) to train MSE for the LASSO model. The LASSO model has generally low values for both train and test MSE, with a test-to-train MSE ratio of approximately 2.4. This ratio of test-to-train MSE is comparable to that of the main effects model, and notably superior to that of the best subset model. Thus, the LASSO model mitigates overfitting, providing better balance between bias and variance. The result is a more parsimonious model with improved interpretability and enhanced generalizability to unseen data.

Overall, the combination of LASSO regression and best subset selection proves to be an effective tool for a comprehensive variable selection for a small sample size. Given the consistent results between the LASSO model $Y_{X_L}$ and the best subset model $Y_{X_k}$, as well as the improved balance between bias and variance, reduced complexity, and improved interpretability, the LASSO model is the most suitable for result interpretation. Based on the information provided by the parameter ranking in Table 6, (e.g., $\lambda * -d\gamma/dT$, $\lambda * \mu$, $P_A * \mu$ being more significant than main effects of $\gamma$ and $-d\gamma/dT$), valuable insights can be obtained. These insights have practical implications for our understanding of the physical phenomena in the L-PBF process. In the following section, we will discuss how we can use these results to bridge the missing links in our knowledge of these phenomena and provide directions for future research.

### Practical interpretation with joint statistical-physical validation

The significance and uncertainties of the input factors as well as their interactions are evaluated from both physical and statistical perspectives. The physical perspective corresponds to the inferences drawn from domain knowledge in Section "Inferences from physics-based domain knowledge". On the other hand, the statistical perspective comprises of the results from the data visualization and variable selection in Sections "Data visualization to Variable selection and model analytics". The joint statistical-physical evaluation is presented in Table 7. In general, there is result consistency between factors identified to be significant from both physical and statistical perspectives. Upon studying the overlapping significant variables from both the physical and statistical perspectives, we can use our statistical findings to contribute insights to the existing physical domain knowledge. Before delving into the insights, we first establish the credibility of our work by demonstrating the achievement of statistical-physical validation. The initial data visualization through statistical plots correctly identifies $P_A$ as a dominant factor contributing to the largest uncertainty, which aligns with well-established physical conclusions in the AM community. Hence this cross-domain validation serves as a substantial source of credibility for our findings, enabling us to pass the statistical-physical validation checkpoint despite the constraint of a small sample size, and

**Table 4.** Cross-validation error for the models.

| | Train MSE | Test MSE | Ratio of test MSE:train MSE |
|---|---|---|---|
| Main effects model | 3.82934428E−06 | 5.80066352E−06 | 1.51479290 |
| Best subset model | 1.67695916E−08 | 4.77001718E−07 | 28.4444444 |
| LASSO model | 1.99120001E−06 | 4.77208228E−06 | 2.39658611 |

**Table 5.** LASSO method applied to the full model, $Y_{X_{i1}X_{i2}X_{i3}X_{i4}X_{i5}}$, for variable selection[a].

| | Variable | Regression coefficient | Confidence interval of estimate |
|---|---|---|---|
| 1 | (Intercept) | 1.05840E−01 | |
| 2 | $P_A$ | 2.65950E−02 | [2.6481E−02, 2.7403E−02] |
| 3 | $\lambda$ | −1.74210E−03 | [−2.5763E−03, −1.6261E−03] |
| 4 | $\mu$ | −6.21710E−04 | [−1.2434E−03, −5.1081E−04] |
| 5 | $\gamma$ | 0 | [−1.1174E−04, 1.5449E−04] |
| 6 | $-\mathrm{d}\gamma/\mathrm{d}T$ | 0 | [−1.3260E−04, 1.7679E−04] |
| 7 | $P_A$: $\lambda$ | 0 | [−1.9592E−04, 1.6435E−04] |
| 8 | $P_A$: $\mu$ | 1.13830E−04 | [−4.3454E−05, 3.1418E−04] |
| 9 | $P_A$: $\gamma$ | 0 | [−1.3641E−04, 1.6228E−04] |
| 10 | $P_A$: $-\mathrm{d}\gamma/\mathrm{d}T$ | 0 | [−1.5428E−04, 1.0928E−04] |
| 11 | $\lambda$: $\mu$ | −5.64980E−05 | [−1.8704E−04, 5.5667E−05] |
| 12 | $\lambda$: $\gamma$ | −3.73220E−05 | [−1.6863E−04, 9.2135E−05] |
| 13 | $\lambda$: $-\mathrm{d}\gamma/\mathrm{d}T$ | −2.23930E−04 | [−4.4786E−04, −9.0059E−05] |
| 14 | $\mu$: $\gamma$ | 0 | [−2.1030E−04, 1.4835E−04] |
| 15 | $\mu$: $-\mathrm{d}\gamma/\mathrm{d}T$ | 0 | [−1.8330E−04, 9.5482E−05] |
| 16 | $\gamma$: $-\mathrm{d}\gamma/\mathrm{d}T$ | 0 | [−1.5408E−04, 1.7311E−04] |
| 17 | $P_A$: $\lambda$: $\mu$ | 0 | [−9.0821E−05, 1.2955E−04] |
| 18 | $P_A$: $\lambda$: $\gamma$ | 0 | [−1.7375E−04, 1.5250E−04] |
| 19 | $P_A$: $\lambda$: $-\mathrm{d}\gamma/\mathrm{d}T$ | 3.35860E−04 | [2.1058E−04, 6.7172E−04] |
| 20 | $P_A$: $\mu$: $\gamma$ | 0 | [−1.1278E−04, 1.5428E−04] |
| 21 | $P_A$: $\mu$: $-\mathrm{d}\gamma/\mathrm{d}T$ | 0 | [−1.6792E−04, 1.3970E−04] |
| 22 | $P_A$: $\gamma$: $-\mathrm{d}\gamma/\mathrm{d}T$ | 0 | [−1.4926E−04, 1.1707E−04] |
| 23 | $\lambda$: $\mu$: $\gamma$ | 0 | [−1.1622E−04, 1.5245E−04] |
| 24 | $\lambda$: $\mu$: $-\mathrm{d}\gamma/\mathrm{d}T$ | 0 | [−1.8964E−04, 1.6005E−04] |
| 25 | $\lambda$: $\gamma$: $-\mathrm{d}\gamma/\mathrm{d}T$ | 0 | [−1.6836E−04, 1.3746E−04] |
| 26 | $\mu$: $\gamma$: $-\mathrm{d}\gamma/\mathrm{d}T$ | 0 | [−1.6794E−04, 1.3532E−04] |
| 27 | $P_A$: $\lambda$: $\mu$: $\gamma$ | 0 | [−1.2429E−04, 1.1270E−04] |
| 28 | $P_A$: $\lambda$: $\mu$: $-\mathrm{d}\gamma/\mathrm{d}T$ | 0 | [−1.0911E−04, 1.6172E−04] |
| 29 | $P_A$: $\lambda$: $\gamma$: $-\mathrm{d}\gamma/\mathrm{d}T$ | 4.74000E−04 | [3.2954E−04, 9.4800E−04] |
| 30 | $P_A$: $\mu$: $\gamma$: $-\mathrm{d}\gamma/\mathrm{d}T$ | 0 | [−1.3084E−04, 1.1765E−04] |
| 31 | $\lambda$: $\mu$: $\gamma$: $-\mathrm{d}\gamma/\mathrm{d}T$ | 7.16360E−04 | [5.6841E−04, 1.4327E−03] |
| 32 | $P_A$: $\lambda$: $\mu$: $\gamma$: $-\mathrm{d}\gamma/\mathrm{d}T$ | 0 | [−1.4869E−04, 1.4495E−04] |

[a]A bootstrap size of 500 is used to estimate the regularized regression coefficients with the corresponding 90% confidence intervals for the bootstrapped estimates.

proceed on to offer further practical insights. Subsequently, the robust results of variable selection obtained from the combination of best subset selection and LASSO regression, implies that this comprehensive approach is a powerful tool for variable selection in small sample sizes, which can successfully identify a stable and interpretable set of variables. Hence this approach can be considered as a viable solution for other high-fidelity multi-physics models that face the constraint of high computational cost and small sample size, such as in phase-field models of microstructural evolutions or residual stress models. In addition, it is noteworthy that the parameter ranking of the optimal LASSO model reveals interactions effects such as $\lambda * -\mathrm{d}\gamma/\mathrm{dT}$, $P_A * \mu$, $\lambda * \mu$, $\lambda * \gamma$, $P_A * -\mathrm{d}\gamma/\mathrm{dT}$, $P_A * \gamma$ to be more significant than main effects of $\gamma$ and $-\mathrm{d}\gamma/\mathrm{d}T$. This finding highlights the importance of incorporating these interaction effects in sensitivity analysis, rather than solely focusing on the main effects of $\gamma$ and $-\mathrm{d}\gamma/\mathrm{d}T$, thus the AM community should account for these interactions in future design of experiments. Specifically for physical experiments, it is recommended to use experimental designs which can support further investigation of interaction effects involving the thermal conductivity with other material properties of IN625, as well as the interactions of laser power

absorption with viscosity and surface tension. In simulations, more research should be invested on the physics driving the interactions between:

- laser power absorption and viscosity
- laser power absorption and surface tension related parameters
- thermal conductivity and viscosity
- thermal conductivity and surface tension related parameters

Next, the interaction effects of $\lambda * \mu$ and $\lambda * -\mathrm{d}\gamma/\mathrm{dT}$ are validated to be significant by the prandtl and marangoni numbers, respectively. We further outline the association of these effects with the *Pr* and *Ma* numbers as follows.

- The variability in *Pr* can be inferred as a joint effect—which involves the 2 main effects of $\lambda, \mu$, and the $\lambda * \mu$ interaction, because these three terms have been found to be significant from the statistical perspective.
- The statistical significance and ranking of importance of $\lambda * -\mathrm{d}\gamma/\mathrm{dT}$ helps us to identify it as the most prominent interaction out of the $\binom{3}{2}$ possible interaction terms that could contribute to the variability in *Ma*.

**Table 6.** Parameter ranking of main effects and 2-factor interactions of the LASSO model.

| Ranking | Parameter |
|---|---|
| 1 | $P_A$ |
| 2 | $\lambda$ |
| 3 | $\mu$ |
| 4 | $\lambda * -\mathrm{d}\gamma/\mathrm{dT}$ |
| 5 | $P_A{}^*\mu$ |
| 6 | $\lambda * \mu$ |
| 7 | $\lambda * \gamma$ |

Given that the interactions between $\lambda * \mu$ and $\lambda * -\mathrm{d}\gamma/\mathrm{dT}$ may be the key contributors to the variability in *Pr* and *Ma*, the AM community should consider channeling resources for further investigation of these interactions. For instance, instead of varying *Pr* and *Ma* in simulations, it could be more informative to vary $\lambda * \mu$ and $\lambda * -\mathrm{d}\gamma/\mathrm{dT}$ instead. Furthermore, the evident interactions of the thermal conductivity $\lambda$ with other material properties imply that it is important to take note of potential enhancement or counteracting effects of different factor combinations for the four factors: $\lambda, \mu, \gamma, -\mathrm{d}\gamma/\mathrm{d}T$. We should calibrate one factor's level while considering another factor's level instead of calibrating them independently. Another important interaction that requires more attention from the AM community is $P_A * \mu$. Further investigations should be conducted on the $P_A * \mu$ interaction as it may indicate the presence of a previously unidentified physical phenomenon in AM, or a possible relation to an existing physical phenomenon that has yet to be fully understood. Since the significance of the $P_A * \mu$ interaction falls between that of $\lambda * \mu$ and $\lambda * -\mathrm{d}\gamma/\mathrm{dT}$, which are related to the two significant physical effects of *Pr* and *Ma*, respectively, it is likely that the potential physical phenomenon associated with the $P_A * \mu$ interaction may also be a key player in the field.

Some general discussion points for the five main effects ($P_A$, $\lambda$, $\mu$, $\gamma$ and $-\mathrm{d}\gamma/\mathrm{d}T$) are provided as follows. Firstly, the AM community should pay careful attention to the laser power absorption, by

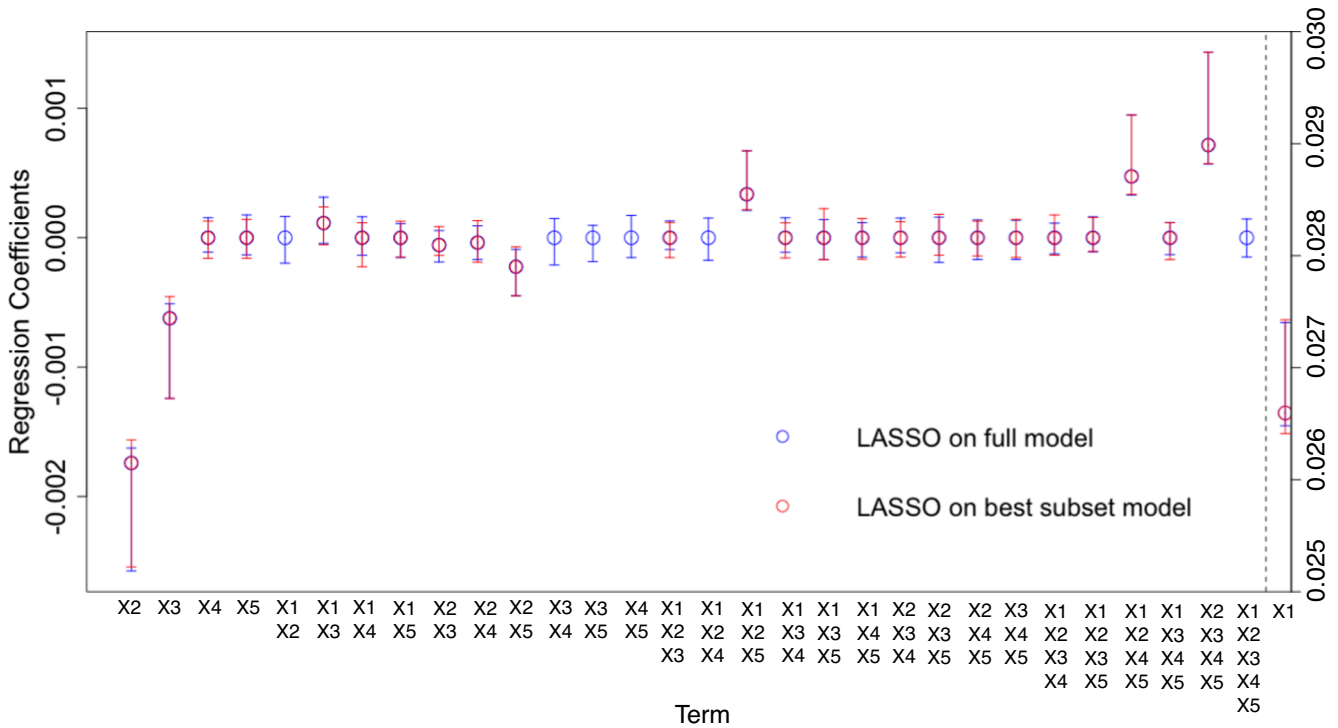


**Fig. 6 90% confidence intervals of the regularized regression coefficients from applying LASSO to the full model (marked in blue) and the best subset model (marked in red) respectively.** The X-axis contains the labels for the model terms, where $P_A$, $\lambda$, $\mu$, $\gamma$, $-\mathrm{d}\gamma/\mathrm{d}T$, are denoted by $X_1$, $X_2$, $X_3$, $X_4$, $X_5$ respectively. The plot illustrates the regression coefficients on a dual Y-axis scale, with the Y-axis on the right displaying the range of values for the regression coefficient of $X_1$, and the Y-axis on the left representing the range of values for the regression coefficients of all other variables. The regression coefficient for $X_1$ exhibits a significantly different magnitude from other variables. A dotted line demarcates the boundary between the two Y-axes, with values on the left corresponding to the left Y-axis, and values on the right corresponding to the right Y-axis.

**Table 7.** Significant factors identified from the physical and statistical perspectives.

| Perspective | Physical | Statistical |
|---|---|---|
| Results | Physical phenomena such as the marangoni effect, effective energy input, and heat transport mechanisms significantly influence melt pool geometry. It is universally agreed in the AM community that the laser power absorption exhibits a dominating influence on the melt pool dimensions. The crucial physical parameters based on the aforementioned domain knowledge are summarized as follows:<br>• $Pr = \frac{c_p \mu}{\lambda}$<br>• $Ma = -\frac{d\gamma}{dT}\frac{w\Delta T}{\mu a}$<br>• $Pe = \frac{UL}{\alpha}$<br>• $Re = \frac{Pe}{Pr}$<br>• $\eta$<br>• $\mu$<br>• $-d\gamma/dT$ | • Data visualization via Figs. 1, 2, 3 consistently identify $P_A$ as a dominating factor with the largest uncertainty, and suspect the significance of some 2-factor interactions.<br>• Both variable selection methods (best subset selection and LASSO regression) identify $P_A$, $\lambda$, $\mu$, $\lambda * -d\gamma/dT$, $P_A * \mu$, $\lambda * \mu$, $P_A * \mu * -d\gamma/Td$, $P_A * \lambda * \gamma * -d\gamma/dT$, $\lambda * \mu * \gamma * -d\gamma/dT$, as statistically significant variables.<br>• The LASSO model achieves the best bias-variance tradeoff. The corresponding parameter ranking of significant main effects and 2-factor interactions from the optimal LASSO model in descending order is as follows: $P_A$, $\lambda$, $\mu$, $\lambda * -d\gamma/dT$, $P_A * \mu$, $\lambda * \mu$, $\lambda * \gamma$, $P_A * -d\gamma/dT$, $P_A * \gamma$. |
| Interpretation | • Factors $P_A, \lambda, \mu, -d\gamma/dT$ are significant<br>• Suspected significant interactions:<br>$\lambda * \mu$, $\lambda * -d/dT$, $\mu * d\gamma/dT$, $\lambda * \mu * -d\gamma/dT$ | • Factors $P_A, \lambda, \mu$ are significant.<br>• Confirmed significant interactions:<br>$\lambda * \mu$, $\lambda * \gamma$, $\lambda * -d\gamma/dT$, $P_A * \mu$ |

accurately deriving the absorbed laser power from fundamental physics in simulations, such as implementing a ray-tracing model to achieve physically-informed absorptivity, and/or calibrating against experiments, to accurately predict the melt pool dimensions. Investing immense efforts into accurately measuring the absorbed laser power for physical experiments is also crucial, by carefully controlling its potential variations caused by surface roughness, powder oxidation, and temperature fluctuation since these may result in significant interactions. Secondly, the thermal conductivity $\lambda$ and viscosity $\mu$, which also play significant roles, should be carefully controlled when determining processing parameter windows for materials. For instance, the processing parameter windows of some materials with higher thermal conductivity, i.e., copper, are very different from those of commonly used materials with lower thermal conductivity, i.e., stainless steel. Hence when exploring the processing window for any material, it is advisable to look up a similar material with known thermal conductivity and viscosity values, as reference for the calibration of $\lambda$ and $\mu$ to avoid trial-and-error variations. Finally, as the surface tension coefficient ($\gamma$) and its temperature sensitivity ($-d\gamma/dT$) are involved in substantial interactions with the thermal conductivity, this implies that material compositions or impurities that alter the temperature sensitivity of surface tension is worthy of attention during experiments as it may affect the interactions. For example, the surface tension temperature sensitivity of Invar36 alloy is susceptible to oxygen content, which is affected by powder type, i.e., oxidation effects in reused powder. Thus reused powder may cause variations in $-d\gamma/dT$, possibly leading to different interaction effects, and this could lead to very different molten pool flow behaviors as observed in X-ray imaging[15]. In contrast, for another material such as S17-4 PH stainless steel powder, which has a surface tension temperature sensitivity that is not susceptible to oxygen content, the mechanical properties of the L-PBF specimens made from fresh state powder do not exhibit obvious changes from those made from powder that has been recycled multiple times[16]. Therefore, it is important to take into account the potential variations and interaction effects for the surface tension temperature sensitivity of different materials during experiments. The aforementioned conclusions are valid for the provided ranges of energy density and material parameters in Table. 8, which correspond to the conduction mode heating. However, they may not always apply in vastly different ranges corresponding to different modes of melting, such as the keyhole mode.

In summary, we use a comprehensive data analytics approach on a full factorial design to work within the constraint of our small dataset from a high-fidelity multi-physics model. The results are consistent for the main and interaction effects of ($P_A, \lambda, \mu, \gamma, -d\gamma/dT$) from both statistical and physical perspectives, despite the limited sample size. The domain knowledge validation, coupled with the strength of the high-fidelity simulations, yields insightful results at a reasonable computational cost with low complexity. The conclusions are summarised as follows:

- The combination of best subset selection and LASSO regression is a comprehensive variable selection approach that may be effective on small sample sizes with many variables, and can potentially be applied to other high-fidelity multi-physics models, such as phase-field models of microstructural evolutions or residual stress models.
- The hybrid variable selection approach consistently identifies a stable and interpretable set of variables relevant to the response, including main effects and 2-factor interactions such as $P_A$, $\lambda$, $\mu$, $\lambda * -d\gamma/dT$, $P_A * \mu$, $\lambda * \mu$, $\lambda * \gamma$ in descending order of significance.
- The parameter ranking suggests that interactions such as $\lambda * -d\gamma/dT$, $P_A * \mu$, $\lambda * \mu$, $\lambda * \gamma$ might be more significant than main effects of $\gamma$ and $-d\gamma/dT$. Hence the AM community should shift their focus of sensitivity analysis to incorporate these interactions instead and account for them in future DOE.
- Further investigation on the $P_A * \mu$ interaction is necessary, as the significant interaction could be related to an existing physical phenomenon in AM.

The comprehensive variable selection and joint statistical-physical interpretation gives practical guidance to both the simulation community and industrial practitioners in AM on resource allocation, understanding underlying physics, future design of experiments, and potential application in other fields. These insights have the potential to improve quality consistency of L-PBF products with careful control of the significant variables and their interactions. Future work should consider designs that provide more detailed analysis of interactions.

## METHODS

### Overview

Consider a high-fidelity multi-physics model such as the thermal-fluid model in Fig. 7, which has the five material input parameters: the absorbed laser power ($P_A$), thermal conductivity ($\lambda$), viscosity ($\mu$), surface tension coefficient ($\gamma$), surface tension temperature sensitivity ($-d\gamma/dT$), and the melt pool depth, $Y$, as the response variable of interest. The input uncertainties ($\Delta P_A$, $\Delta\mu$...etc.)

**Table 8.** Full factorial design of experiments for input factors ($P_A$, $\lambda$, $\mu$, $\gamma$, $-$ dγ/dT) with the corresponding output melt pool depth values.

| Factor | $P_A$ | $\lambda$ | $\mu$ | $\gamma$ | -dγ/dT | $D$ |
|---|---|---|---|---|---|---|
| Name | Absorbed Laser Power | Thermal Conductivity | Viscosity | Surface Tension Coefficient | Surface Tension Temperature Sensitivity | Melt Pool Depth |
| Lower Level | 58.5 | 18.24 | 0.004 | 1.504 | 0.00008 | N.A. |
| Higher Level | 97.5 | 27.36 | 0.006 | 2.256 | 0.00012 | N.A. |
| Nominal Value | 78 | 22.8 | 0.005 | 1.88 | 0.0001 | 0.1064 |
| Unit | W | $W\ m^{-1}\ K^{-1}$ | $kg(m{\cdot}s)^{-1}$ | $kgs^{-2}$ | $kgs^{-2}K^{-1}$ | mm |
| Relative Error | 25% | 20% | 20% | 20% | 20% | N.A. |
| Case 1 | − | − | − | − | − | 0.08162 |
| Case 2 | − | − | − | − | + | 0.08162 |
| Case 3 | − | − | − | + | − | 0.07827 |
| Case 4 | − | − | − | + | + | 0.08588 |
| Case 5 | − | − | + | − | − | 0.07917 |
| Case 6 | − | − | + | − | + | 0.08077 |
| Case 7 | − | − | + | + | − | 0.08037 |
| Case 8 | − | − | + | + | + | 0.08042 |
| Case 9 | − | + | − | − | − | 0.07912 |
| Case 10 | − | + | − | − | + | 0.079696 |
| Case 11 | − | + | − | + | − | 0.081869 |
| Case 12 | − | + | − | + | + | 0.07448 |
| Case 13 | − | + | + | − | − | 0.07576 |
| Case 14 | − | + | + | − | + | 0.07452 |
| Case 15 | − | + | + | + | − | 0.07571 |
| Case 16 | − | + | + | + | + | 0.07319 |
| Case 17 | + | − | − | − | − | 0.1346 |
| Case 18 | + | − | − | − | + | 0.1326 |
| Case 19 | + | − | − | + | − | 0.1353 |
| Case 20 | + | − | − | + | + | 0.138 |
| Case 21 | + | − | + | − | − | 0.13184 |
| Case 22 | + | − | + | − | + | 0.13722 |
| Case 23 | + | − | + | + | − | 0.13587 |
| Case 24 | + | − | + | + | + | 0.13318 |
| Case 25 | + | + | − | − | − | 0.13167 |
| Case 26 | + | + | − | − | + | 0.13184 |
| Case 27 | + | + | − | + | − | 0.13049 |
| Case 28 | + | + | − | + | + | 0.13175 |
| Case 29 | + | + | + | − | − | 0.13041 |
| Case 30 | + | + | + | − | + | 0.12915 |
| Case 31 | + | + | + | + | − | 0.12772 |
| Case 32 | + | + | + | + | + | 0.13273 |

A scan speed of $v = 0.6\ ms^{-1}$ is used for all cases.

propagate directly to the depth $Y$, and may also result in interaction effects such as $P_A * \mu$ and $\lambda * \mu$ that further contribute to output uncertainty of the depth ($\Delta Y$). We aim to understand the effects of these input uncertainties, but a constraint is the high computational cost of high-fidelity simulations—which take 2 to 3 days on average per simulation. Therefore, a full factorial design and variable selection approach is proposed for analytics on the effects of these input uncertainties at a reasonable computational cost.

The thermal-fluid model is selected as our multi-physics model due to its ability to capture major physical phenomena of the L-PBF process[8], hence offering better accuracy than analytical models and heat transfer models which use the finite element method. Material parameters are specifically chosen in this study since most studies in the literature focus on the effects of process parameters, such as laser power, scan speed, and beam radius[9,17,18]. Though reported to be highly sensitive[19], there are limited studies on material parameters such as surface tension, viscosity, thermal conductivity and the absorbed laser power—which is determined by the energy absorptivity of the material while the input laser power is kept constant. For instance, it has been found that the energy absorptivity variation is related to the likelihood of keyhole pore formation, which drastically changes the melt pool geometry[8,20]. The thermal conductivity, viscosity, and surface tension parameters play a critical role in thermal-fluid simulation results as they are related to the flow properties and thermal properties of the material, which in turn control the hydrodynamics and transport phenomena of the melt pool[19].

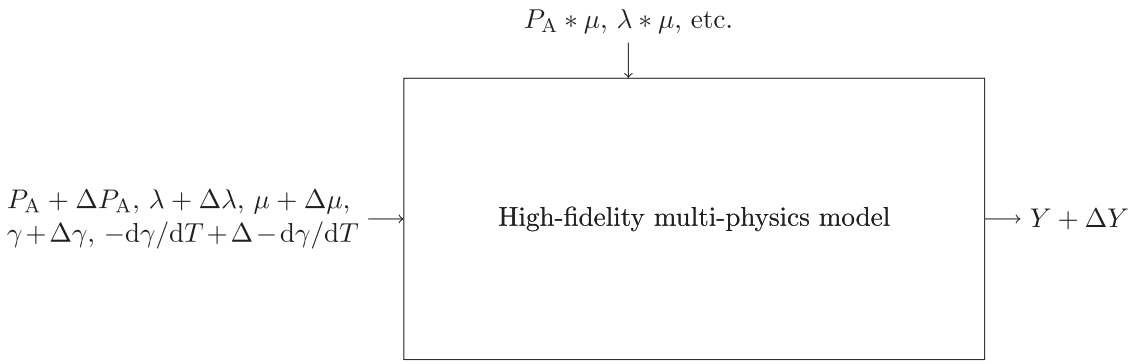


**Fig. 7** A black box representation of the propagation of uncertainty Δ, from the input parameters ($P_A, \lambda, \mu, \gamma, -d\gamma/dT$) to the output ($Y$) in a high-fidelity multi-physics model.

Hence this study focuses on the effects of highly sensitive material parameters on the output QoI, which is selected as the melt pool depth, $Y$. Of the three output melt pool dimensions (length, width, depth), the depth is most crucial since it is related to lack of fusion defects, which affect the mechanical properties of L-PBF products[20]. The nickel-based alloy, Inconel 625 (IN625), is selected as the model material, since it is popular in many AM applications due to its high strength and good fabricability[1].

The methodology used in this work is detailed as follows. A full factorial design of experiments is conducted for the five material parameters ($P_A, \lambda, \mu, \gamma, -d\gamma/dT$), where the upper and lower factor levels of the design are selected to represent their respective input uncertainties. Data of melt pool dimensions is generated using high-fidelity thermal-fluid simulations. A thorough review of the physical phenomena in the L-PBF process is conducted to serve as domain knowledge for verifying the statistical results obtained in this study. Given the limited size of our data set, we define a validation checkpoint based on domain knowledge to ensure the credibility of our subsequent statistical analysis. The selected statistical-physical validation criterion is $P_A$ having a dominating influence on the response depth, since it is a well established fact in the AM community. Initial data visualization using half-normal plots, interaction plots, and standard deviation plots are being conducted to confirm that our prior results meet the validation checkpoint, and that our data set is suitable for further analysis. Subsequently, multiple linear regression analysis is employed to further investigate the findings from the initial visualization, through hypothesis testing of seven different models. A combination of variable selection techniques such as best subset selection and the least absolute shrinkage and selection operator (LASSO) regression is used to identify the significant variables and provide a parameter ranking. Measures such as adjusted $R^2$, Mallow's $C_p$, the ratio of test mean squared error to train mean squared error, and residual normality are used to assess model performance. Finally, the overall statistical results of the study are jointly evaluated with physical domain knowledge to provide statistical-physical validation for our critical findings. These findings are then used to draw practical insights for simulation and experimental communities in AM.

## Design of experiments for simulations

A full factorial DOE is constructed for the five factors: $P_A$, $\lambda$, $\mu$, $\gamma$, and $-d\gamma/dT$ due to its ability to study not only the main effect of a single factor, but also interaction effects between any two factors on the output QoI[21,22]. A major advantage of the full factorial design is its ability to comprehensively examine all possible combinations of input factors[23]. This allows us to study important factor interactions, which are suspected to have substantial influence on melt pool geometry[3,24]. Here we consider the 2-level design, and the simulations are conducted by taking all possible combinations of each factor's high level (+) or low level (−)[25]. The high and low levels of the factors: $P_A, \mu, \lambda, \gamma, -d\gamma/dT$, are taken as 25%, 20%, 20%, 20%, 20% above and below the nominal values of the factors respectively. These relative error percentages of the factors represent their respective input uncertainties in multi-physics modeling, which arise due to the lack of knowledge on their exact values. Since it is not possible to explicitly determine these exact values, the nominal values for the factors, along with their variations (or uncertainties), are chosen based on prior research and domain knowledge[3,24]. For instance, the commonly used nominal value for energy absorptivity under the laser power of 195W and scan speed of 0.6 $ms^{-1}$, is 0.4[3]. Hence the absorbed laser power $P_A$—which involves a multiplication of laser power and energy absorptivity, has a selected nominal value of 78W. In addition, the relative error of $P_A$ has a selected value of 25% since previous studies have estimated the uncertainty of the absorption coefficient to be larger than that of other input factors, with a variation of at least 25%[3]. The total number of simulations to be run is $2^5$, and they are performed using the thermal fluid-flow model discussed in Section "Thermal fluid-flow model". The constructed DOE with complete design information, along with the simulated depth values, are given in Table 8.

## Thermal fluid-flow model

The thermal-fluid simulation is utilized to build the dataset for the subsequent data analysis[26–29]. Based on the assumption of the incompressible laminar flow, the governing equations of mass continuity with the incompressiblity condition, momentum conservation and energy conservation are given as

$$\begin{cases} \nabla \cdot (\mathbf{v}) & = 0, \\ \frac{\partial}{\partial t}(\rho \mathbf{v}) + \nabla \cdot (\rho \mathbf{v} \otimes \mathbf{v}) & = \nabla \cdot (\mu \nabla \mathbf{v}) - \nabla p + \rho \mathbf{g} + \mathbf{f_B}, \\ \frac{\partial}{\partial t}(\rho h) + \nabla \cdot (\rho \mathbf{v} h) & = q + \nabla \cdot (\lambda \nabla T), \end{cases} \quad (1)$$

which are related to the five selected material input parameters in this study ($P_A, \lambda, \mu, \gamma, -d\gamma/dT$)[26]. The absorbed laser power, $P_A$, is contained in term $q$ of the energy conservation equation. The thermal conductivity, $\lambda$, is incorporated in the energy conservation equation. Viscosity, $\mu$, is incorporated in the momentum conservation equation. The boundary conditions for the momentum conservation equation incorporate surface tension, recoil pressure and Marangoni forces[20], which account for the surface tension coefficient, $\gamma$, and surface tension temperature sensitivity, $-d\gamma/dT$. The thermal boundary conditions incorporate the surface radiation and energy loss by evaporation. In addition, $\mathbf{v}$ represents the velocity vector, $p$ represents the pressure, and $T$ represents the temperature. In the energy conservation equation,

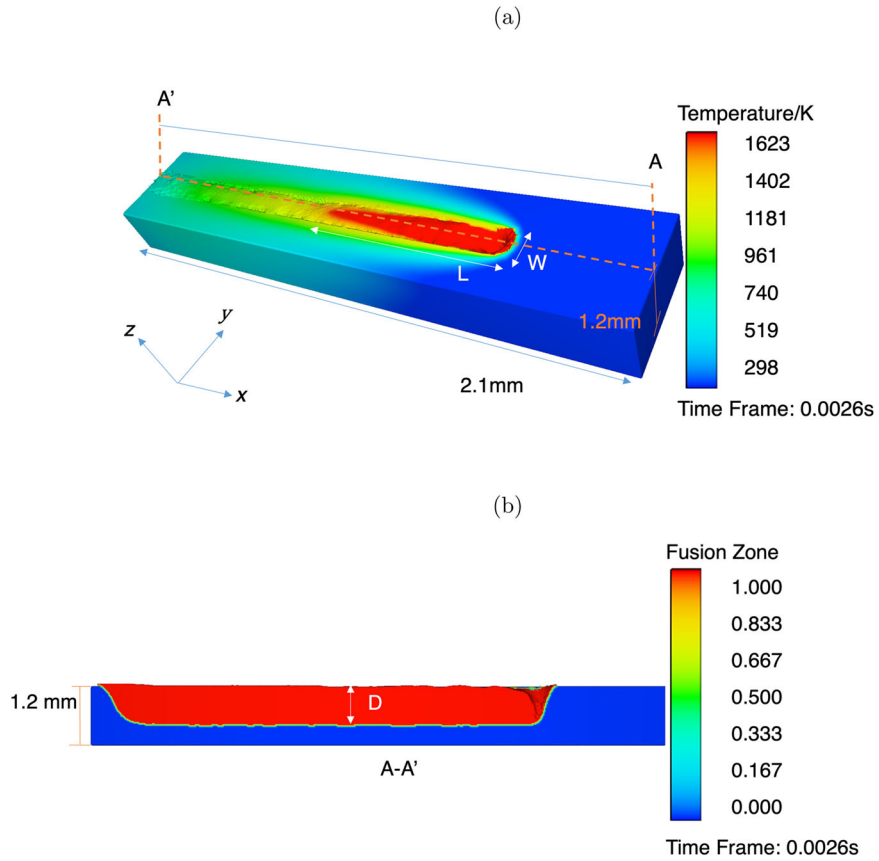


**Fig. 8 Typical simulation results with corresponding colour scales. a** Isometric view with temperature scale, where *L* and *W* denote the melt pool length and width respectively. **b** Vertical cross sectional view of melt pool depth with fusion scale---where 1 represents the region that has ever been melted, while 0 represents the region that has never been melted.

$h$ is the specific enthalpy denoted by $h = c_pT + (1 - f_s)L$, where $c_p$, $f_s$ and $L$ represent the specific heat, solid fraction and latent heat of melting, respectively. The momentum equation incorporates gravity (**g**) and buoyancy ($\mathbf{f_B}$, Boussinesq approximation).

The free surface at each time increment is tracked and reconstructed by the Volume of Fluid (VOF) method[30], given as

$$\frac{\partial F}{\partial t} + \nabla \cdot (F\mathbf{v}) = 0, \tag{2}$$

where $F$ denotes the fluid fraction. The model is able to provide output melt pool dimensions from simulations performed at different input factor settings. In these simulations, the powder layer is not incorporated to minimize melt pool fluctuation caused by the randomly-packed powder layer. This enables better focus on the output variation caused by the material input parameters, and more details about the model can be referred to[26,28].

The full set of simulation results is provided in Table 8. Two variables, namely the temperature and the fusion zone, are used to measure melt pool dimensions, as shown in Fig. 8. The temperature plot is used to measure melt pool length. The fusion zone represents the region that has ever been melted, i.e., temperatures exceeding the melting temperature of IN625 (1623K), and is used to measure melt pool width and depth. It displays the entire melted region, which includes both molten and solidified states of the material along the scan track.

## Inferences from physics-based domain knowledge

In this section, crucial physics that occur in the L-PBF process based on the knowledge of field experts, e.g., simulation teams, industrial practitioners, are discussed and analyzed. Some physical phenomena found to significantly influence melt pool geometry are the marangoni effect, effective energy input, and heat transport mechanisms, i.e., via conduction, convection, diffusion[31].

The Marangoni Effect refers to the phenomenon of mass transfer along an interface between two fluids driven by a surface tension gradient[32,33]. It can be quantified with the Marangoni Number, *Ma*, which compares the rate of transport of fluid due to Marangoni flows, with the rate of transport of diffusion[31]. It contains 3 input factors of interest in this study, namely $-\mathrm{d}\gamma/\mathrm{d}T$, $\mu$, $\lambda$, and is defined by: $Ma = -\frac{\mathrm{d}\gamma}{\mathrm{d}T}\frac{w\Delta T}{\mu\alpha}$, where $\alpha$ is the thermal diffusivity of the alloy, given by $\alpha = \frac{\lambda}{\rho c_p}$. The constants $c_p$ and $\rho$ stand for the specific heat and density respectively, while $w$ is the characteristic length of the melt pool, which is taken as melt pool width. The difference between the maximum temperature inside the pool and the solidus temperature of an alloy is denoted by $\Delta T$.

Dimensionless numbers related to the different types of heat transfer mechanisms are the Prandtl Number (*Pr*), the Peclet Number (*Pe*) and the Reynolds Number (*Re*)[19,31,34]. The Prandtl Number, *Pr*, is a fluid property, which reflects the ratio of kinematic viscosity and heat diffusivity. It contains 2 input factors of interest in this study, namely $\mu$ and $\lambda$, and is defined as: $Pr = \frac{c_p\mu}{\lambda}$. It provides a gauge of the relative effects of momentum diffusivity and thermal diffusivity[34]. The Peclet Number, *Pe*, signifies the ratio of the convection rate associated with the scanning speed and the rate of conduction[31]. It is related to input factor $\lambda$ of this study, and is defined as: $Pe = \frac{UL}{\alpha}$, where $U$ is the characteristic velocity, $\alpha$ is the thermal diffusivity of the alloy, and $L$ is the characteristic length—which is taken as the melt pool length[31]. The dimensionless parameters *Pr* and *Pe* are related by the Reynolds Number, which is defined as: $Re = Pe/Pr$. It is not an independent parameter since it is a ratio of the Peclet Number and Prandtl Number. Holding the Reynolds number constant is equivalent to holding laser diameter and scanning velocity constant[31]. This

**Table 9.** Inferences based on domain knowledge.

| Physical Parameters | Observation from domain knowledge for: $P_A, \lambda, \mu, \gamma, -d\gamma/dT$ | Physics-based inferences for $P_A, \lambda, \mu, \gamma, -d\gamma/dT$ |
|---|---|---|
| Prandtl Number, $Pr = \frac{c_p \mu}{\lambda}$ | • High *Pr* values result in wider and shallower melt pools, while low *Pr* values cause narrow and deep melt pools[35,36].<br>• Materials with high *Pr* values result in melt pools with strong Marangoni convection[35]. | Parameter *Pr* has a significant effect on melt pool dimensions—where this effect is attributed to one or more of the following main effect(s) and/or interaction(s) of the input factors:<br>• Main effects of: $\lambda, \mu$<br>• The $\binom{2}{2}$ possible interaction terms of $\lambda$ and $\mu$.<br>Inference: Main effects of $\lambda$, $\mu$ and/or interaction effect $\lambda * \mu$ is significant |
| Marangoni Number, $Ma = -\frac{d\gamma}{dT}\frac{w\Delta T}{\mu a}$ | • Marangoni flow in the melt-pool could significantly influence the melt-pool shape in a nonlinear manner. A higher *Ma* number results in a wider and shallower melt pool[35].<br>• A high value of *Ma* implies larger liquid metal velocity and more efficient convective heat transfer. This causes a larger melt pool with higher aspect ratio and length[31].<br>• Large values of *Ma* indicate a strong driving force for convection at the melt pool surface, causing rapid heat transfer in the radial direction by convective heat transport and consequently, a higher value of *Pe*[36]. | Parameter *Ma* has a significant effect on melt pool dimensions—where this effect is attributed to one or more of the following main effect(s) and/or interaction(s) of the input factors:<br>• Main effects of: $\lambda, \mu, -d\gamma/dT$<br>• The $\binom{3}{2}$ possible 2-factor interactions of $\lambda, \mu, -d\gamma/dT$<br>• The $\binom{3}{3}$ 3-factor interaction of $\lambda, \mu, -d\gamma/dT$<br>Inference: Factors $\lambda, \mu, -d\gamma/dT$ and/or interaction effects $\lambda * \mu, \mu * -d\gamma/dT, \lambda * d\gamma/dT, \lambda * \mu * -d\lambda/dT$ is significant |
| Peclet Number, $Pe = \frac{UL}{a}$ | • A low value of *Pe* causes a narrow and deep melt pool[36].<br>• Higher values of *Pe* imply smaller conduction-induced melt zone in the absence of thermocapillary motion[34].<br>• A high value of *Pe* results in increased aspect ratio due to more dominant Marangoni convection[34]. | Parameter *Pe* has a significant effect on melt pool dimensions—where this effect is attributed to:<br>• Main effect of: $\lambda$<br>Inference: Factor $\lambda$ is significant. |
| Absorption coefficient, $\eta$ | $P_A$ has significant effect on melt pool depth[20] | Parameter $\eta$ has a significant effect on melt pool dimensions—where this effect is attributed to:<br>• Main effect of: $P_A$<br>Inference: Factor $P_A$ is significant |
| Surface Tension Temperature Sensitivity, $-d\gamma/dT$ | A higher value for $-d\gamma/dT$ result in convective heat transfer and nearly round melt pool shape[19]. | The main effect of factor $-d\gamma/dT$ has a significant effect on melt pool dimensions.<br>Inference: Factor $-d\gamma/dT$ is significant. |
| Viscosity, $\mu$ | • A higher value of $\mu$ increases melt pool length[19]. | The main effect of factor $\mu$ has a significant effect on melt pool dimensions.<br>Inference: Factor $\mu$ is significant. |

provides a convenient basis of comparison for different parameter effects—with constant Re, material effects are associated with *Pr*, and heat input effects are attributed to $-d\gamma/dT$[31,34]. Another crucial parameter related to the effective laser energy input is the laser absorption coefficient of the material ($\eta$). It represents the percentage of laser power that is actually absorbed by the material for a specific experimental set up[4].

The observations reported by the simulation groups as well as industrial practitioners provide useful inferences regarding the main effects as well as potential interaction effects of input factors ($P_A, \lambda, \mu, \gamma, -d\gamma/dT$) onto the response melt pool dimensions. Table 9 summarizes the observations made by the domain experts, along with the corresponding inferences for the input factors $P_A, \lambda, \mu, \gamma, -d\gamma/dT$[19,19,31,34–37]. The inferences are drawn based on the assumption that all other variables, apart from the input factors of our study ($P_A, \lambda, \mu, \gamma, -d\gamma/dT$), are kept as constants in the physical parameters. Since $P_A, \lambda, \mu, \gamma, -d\gamma/dT$ contribute to the main source of variability in these physical parameters, their effects on the output can be associated with the main effects of $P_A, \lambda, \mu, \gamma, -d\gamma/dT$ and/or their interactions. For instance, it has been reported in literature that *Pr* has a substantial effect on the melt pool aspect ratio[35,36]. Such an observation could be caused by the main effect of $\lambda$, $\mu$ and/or the combined effect of both factors. Hence, a possible inference for *Pr* is that the factors $\lambda$, $\mu$ and/or interaction $\lambda * \mu$ is significant. The same reasoning is applied for the rest of the physical parameters to yield the respective inferences as shown in Table. 9.

The research conducted thus far will serve as domain knowledge that can be used to complement the interpretation of the statistical results subsequently, allowing us to jointly evaluate our results from a statistical-physical perspective.

## Half-normal probability plots

Data visualization plays a crucial role in analytics as it serves as a common ground for understanding the data, and serves as a quick method to assess if the defined validation checkpoint is being met. Additionally, it allows for the detection of any unusual trends in the data. To achieve this, we will utilize half-normal plots, interaction plots and standard deviation plots for our data visualization process. If the validation checkpoint is met through the prior results of the data visualization, we will proceed with further analysis using multiple linear regression.

In a general $2^k$ factorial, if there are no replicates, Montgomery[12] recommended the use of a normal probability plot for analysis. The normal probability plot (NPP) works on the assumption that changes in input factor levels have no effect on the response, and that the variation in the response variable occurs by chance, i.e., random fluctuation of the response variable about a mean. This implies that all 32 effects—which are the main effects and interaction effects of the five input factors, are initially assumed to have roughly normal distributions centred at zero, and should form a straight line when plotted as points on a normal probability scale[14]. Hence points (effects) that fit reasonably well on the straight line agree with this assumption, and are concluded as not significant. However, effects that deviate from the line are not easily explained as chance occurrences, and are suspected to be significant. According to the aforementioned working principle, the following steps are used to produce the NPP and perform prior analysis of all the factor effects.

(1) Calculation of the 32 effects—The effects are calculated and sorted in standard order using Yates' Algorithm[14,38,39].
(2) The ordered effect values then undergo a rankit approximation—which estimates the expected values of the effects'

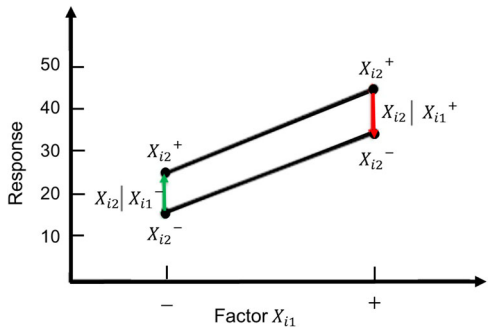


**Fig. 9 Interaction plot of an input factor $X_{i_1}$ on the X-axis and the response variable on the *Y*-axis.** Another input factor $X_{i_2}$ is varied simultaneously to produce two lines corresponding to the low and high levels of $X_{i_2}$ respectively.

order statistics from the standard normal distribution, to yield their respective cumulative probabilities and corresponding *z*-test statistics according to

$$\begin{aligned} z_i &= \Phi^{-1}\left(\tfrac{i-a}{n+1-2a}\right), \\ &\text{for } i = 1, 2, \ldots, n, \\ a &= 3/8 \text{ if } n \le 10 \text{ and } \quad 0.5 \text{ for } n{>}10, \end{aligned} \tag{3}$$

where $\Phi^{-1}$ represents the standard normal quantile function, $i$ represents the rank order of each effect, $n$ represents the total number of effects. Since $n = 32$, the corresponding value of $a$ is 0.5[40,41]

(3) The NPP plot is generated with the corresponding *z*-statistic for each effect on the vertical axis, and the absolute value of the effect on the horizontal axis.
(4) A best fit straight line is drawn and outliers are identified as significant effects.

The half-normal plot shares the same working principles as the NPP, except that it considers only the magnitude of these effects.

The benefit of using the NPP or half-normal plot is that it allows for a quick and convenient method of analyzing all the factor effects, since outliers can be identified from visual inspection. However, a drawback is the lack of a clear-cut measure for significance, requiring a large dose of subjective judgement. Therefore, the half-normal plots will be complemented with statistical hypothesis testing using regression models, and also jointly evaluated with physical domain knowledge found previously in Section "Inferences from physics-based domain knowledge" to validate the results of the visual analysis.

### Interaction plots

In an interaction plot, the response variable of interest *Y*, i.e., melt pool depth, is plotted on the vertical axis. An input factor $X_{i_1}$ is plotted on the horizontal axis with the domain spanning from its low to high level. Another input factor $X_{i_2}$ that has suspected interaction with $X_{i_1}$ is also varied simultaneously from its low to high level. This yields two separate lines characterizing the effect of $X_{i_1}$ on the response, corresponding to the low and high levels of $X_{i_2}$ respectively as shown in Fig. 9. Simple visual inspection of the two lines allow us to quickly study how interactions affect the relationship between the factors and the response. If the two lines are parallel, this indicates that there is no interaction between factors $X_{i_1}$ and $X_{i_2}$. This is because the effect of $X_{i_2}$ on the response variable when $X_{i_1}$ is at its low level, $X_{i_2}|X_{i_1}{}^-$, is the same as that when $X_{i_1}$ is at its high level, $X_{i_2}|X_{i_1}{}^+$. Hence this shows that the interaction effect, $X_{i_1} * X_{i_2} = (X_{i_2}|X_{i_1}{}^+ - X_{i_2}|X_{i_1}{}^-)/2 = 0$, and the factor level of $X_{i_2}$ does not affect the effect of $X_{i_1}$ on the response. On the other hand, non-parallel lines indicate the presence of an interaction effect between $X_{i_1}$ and $X_{i_2}$.

### Standard deviation plots

The main effects model, $Y_{X_{i1}}$, is a parsimonious model that can be used to obtain quick results on uncertainty propagation via the one-factor-at-a-time (OFAT) method.

$$Y_{X_{i1}} = \beta_0 + \beta_1 P_A + \beta_2 \lambda + \beta_3 \mu + \beta_4 \gamma + \beta_5(-\mathrm{d}\gamma/\mathrm{d}T) + \epsilon_1. \tag{4}$$

The model coefficients in $\boldsymbol{\beta}$ are assumed to be fixed[42], and act as weights attached to the individual $X_i$'s which are normally distributed:

$$X_i \sim N(0, \sigma_{X_i}) \quad i = 1, 2, \ldots n. \tag{5}$$

Hence, the response *Y*—melt pool depth, will also be normally distributed with a mean and variance of

$$\begin{aligned} \overline{Y} &= \textstyle\sum_{i=1}^{n} \beta_i \overline{x}_i, \\ \sigma_Y &= \sqrt{\textstyle\sum_{i=1}^{n} \beta i^2 \sigma_{X_i}^2}. \end{aligned} \tag{6}$$

In addition, $\boldsymbol{\beta}$ can be defined by the partial derivative of the response *Y* versus the individual variable $X_i$, $\frac{\partial Y}{\partial X_i}$[42]. By applying (6) to the input factors in our study ($P_A, \lambda, \mu, \gamma, -\mathrm{d}\gamma/\mathrm{d}T$), we have

$$\sigma_Y = \sqrt{\left(\frac{\partial y}{\partial P_A}\right)^2 \left(\sigma_{P_A}^2\right) + \left(\frac{\partial y}{\partial \lambda}\right)^2 \left(\sigma_\lambda^2\right) + \left(\frac{\partial y}{\partial \mu}\right)^2 \left(\sigma_\mu^2\right) + \left(\frac{\partial y}{\partial \gamma}\right)^2 \left(\sigma_\gamma^2\right) + \left(\frac{\partial y}{\partial(-\mathrm{d}\gamma/\mathrm{d}T)}\right)^2 \left(\sigma_{-\mathrm{d}\gamma/\mathrm{d}T}^2\right)}. \tag{7}$$

Equation (7) serves as a function to approximate uncertainty propagation, on which we apply an OFAT approach to investigate how each factor's input uncertainty—represented by the factor's standard deviation, affects the output uncertainty, i.e., standard deviation of melt pool depth $\sigma_Y$. This approach involves varying each input standard deviation $\sigma_{X_i}$ in steps of 0.01 to analyze the influence of small perturbations in input standard deviation onto the output standard deviation. Plots of input-output standard deviation are then made to assess how the input uncertainties propagate to the output uncertainty. The graphical results are displayed in Section "Data visualization".

### Multiple linear regression

A multiple linear regression (MLR) model is an empirical model that relates the chosen response variable of interest, melt pool depth *Y*, to the $p$ predictors stored in vector $X$, where each predictor can be a main or interaction effect[12]. The MLR model takes on the general form of

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{8}$$

where

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \text{ and } \quad \boldsymbol{\epsilon} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}.$$

In (8), the total number of observations from the thermal-fluid simulations is denoted by $n$, which is equal to 32. An actual observation obtained from the thermal-fluid simulation is denoted by lowercase $y_i$ for $i = 1,..., n$, and $\epsilon$ is the error term. The error term is represented by the vector of residuals $e_i$ for $i = 1, \ldots, n$, where $e_i = y_i - \hat{y}_i$, which is the difference between each observation from the thermal-fluid simulation, $y_i$, and the corresponding fitted value from the regression, $\hat{y}_i$. The parameters $\beta_j$, $j = 0, 1, \ldots, p$ are the predictor coefficients, which are calculated using the least squares estimator, $\hat{\beta}$. The vector of $\hat{\beta}$ minimizes the sum of square of errors, $\sum_{i=1}^{n} e_i^2$ and is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{9}$$

Let $X_i$ denote a main effect, where $i = 1,...5$, and all main effects in this study are coded inputs that have been scaled according to

$$\frac{X_i - (X_{i_{\text{low}}} + X_{i_{\text{high}}})/2}{(X_{i_{\text{high}}} - X_{i_{\text{low}}})/2}, \tag{10}$$

to be between the range of [−1, 1], with a mean of zero and standard deviation of one[12].

In this study, we consider seven different MLR models—the full model, the 4-factor interactions model, the 3-factor interactions model, the 2-factor interactions model, the main effects model, the best subset model and the LASSO model. Each model contains $p$ predictors, where $p = 31$, 30, 25, 15, 5, 25, 10 for the seven models respectively, and $m$ denotes the total number of main effects in the model.

Let $Y_{X_{i1}}$ represent the main effects only model given by

$$Y_{X_{i1}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon_1, \tag{11}$$

where this model comprises the intercept term $\beta_0$, and the 5 main effects of input factors: $P_A, \lambda, \mu, \gamma, -\mathrm{d}\gamma/\mathrm{d}T$ (denoted by $X_1, X_2, X_3, X_4, X_5$ respectively).

Let $Y_{X_{i1}X_{i2}}$ represent the 2-factor interaction model given by

$$Y_{X_{i1}X_{i2}} = \beta_0 + \sum_{i1=1}^{5} \beta_{i1} X_{i1} \quad + \sum_{\substack{(i_1, i_2) \in [m]^2 \\ i_1 \neq i_2}}^{10} \beta_{i_1,i_2} X_{i_1} X_{i_2} + \epsilon_2, \tag{12}$$

where this model comprises the intercept term, the 5 main effects and all $\binom{5}{2}$ possible 2-factor interactions, $X_{i1} * X_{i2}$, (e.g. $P_A * \lambda$).

Let $Y_{X_{i1}X_{i2}X_{i_3}}$ represent the 3-factor interaction model given by

$$Y_{X_{i_1}X_{i_2}X_{i_3}} = \beta_0 + \sum_{i1=1}^{5} \beta_{i1} X_{i1} \quad + \sum_{\substack{(i_1, i_2) \in [m]^2 \\ i_1 \neq i_2}}^{10} \beta_{i_1,i_2} X_{i_1} X_{i_2} + \sum_{\substack{(i_1, i_2, i_3) \in [m]^3 \\ i_1 \neq i_2 \neq i_3}}^{10} \beta_{i_1,i_2,i_3} X_{i_1} X_{i_2} X_{i_3} + \epsilon_3, \tag{13}$$

where this model comprises the intercept term, the 5 main effects, all $\binom{5}{2}$ 2-factor interactions, and all $\binom{5}{3}$ possible 3-factor interactions, $X_{i1} * X_{i2} * X_{i3}$, (e.g. $P_A * \lambda * \mu$).

Let $Y_{X_{i1}X_{i2}X_{i_3}X_{i_4}}$ represent the 4-factor interaction model given by

$$Y_{X_{i_1}X_{i_2}X_{i_3}X_{i_4}} = \beta_0 + \sum_{i1=1}^{5} \beta_{i1} X_{i1} + \sum_{\substack{(i_1, i_2) \in [m]^2 \\ i_1 \neq i_2}}^{10} \beta_{i_1,i_2} X_{i_1} X_{i_2} + \sum_{\substack{(i_1, i_2, i_3) \in [m]^3 \\ i_1 \neq i_2 \neq i_3}}^{10} \beta_{i_1,i_2,i_3} X_{i_1} X_{i_2} X_{i_3} + \sum_{\substack{(i_1, i_2, i_3, i_4) \in [m]^4 \\ i_1 \neq i_2 \neq i_3 \neq i_4}}^{5} \beta_{i_1,i_2,i_3,i_4} X_{i_1} X_{i_2} X_{i_3} X_{i_4} + \epsilon_4, \tag{14}$$

where this model comprises the intercept term, the 5 main effects, all $\binom{5}{2}$ 2-factor interactions, all $\binom{5}{3}$ 3-factor interactions, and all $\binom{5}{4}$ possible 4-factor interactions, $X_{i1} * X_{i2} * X_{i3} * X_{i4}$, (e.g. $P_A * \lambda * \mu * \gamma$).

Let $Y_{X_{i1}X_{i2}X_{i3}X_{i4}X_{i_5}}$ represent the full model given by

$$Y_{X_{i_1}X_{i_2}X_{i_3}X_{i_4}X_{i_5}} = \beta_0 + \sum_{i1=1}^{5} \beta_{i1} X_{i1} + \sum_{\substack{(i_1, i_2) \in [m]^2 \\ i_1 \neq i_2}}^{10} \beta_{i_1,i_2} X_{i_1} X_{i_2} + \sum_{\substack{(i_1, i_2, i_3) \in [m]^3 \\ i_1 \neq i_2 \neq i_3}}^{10} \beta_{i_1,i_2,i_3} X_{i_1} X_{i_2} X_{i_3} + \sum_{\substack{(i_1, i_2, i_3, i_4) \in [m]^4 \\ i_1 \neq i_2 \neq i_3 \neq i_4}}^{5} \beta_{i_1,i_2,i_3,i_4} X_{i_1} X_{i_2} X_{i_3} X_{i_4} + \sum_{\substack{(i_1, i_2, i_3, i_4, i_5) \in [m]^5 \\ i_1 \neq i_2 \neq i_3 \neq i_4 \neq i_5}}^{1} \beta_{i_1,i_2,i_3,i_4,i_5} X_{i_1} X_{i_2} X_{i_3} X_{i_4} X_{i_5} + \epsilon_5, \tag{15}$$

where this model comprises the intercept term, the 5 main effects, all $\binom{5}{2}$ 2-factor interactions, all $\binom{5}{3}$ 3-factor interactions, all $\binom{5}{4}$ 4-factor interactions and the 5-factor interaction, $X_{i1} * X_{i2} * X_{i3} * X_{i4} * X_{i5}$ (i.e., $P_A * \lambda * \mu * \gamma * -\mathrm{d}\gamma/\mathrm{d}T$).

### Best subset selection

Since these 5 MLR models have been formed via manual selection of variables, it is possible that all five candidate models do not contain the optimal number of variables and the best combination for them. Hence, the best subset selection algorithm is used to search through all possible combinations of variables, choosing the best model with the optimal number and combination of variables. The best subset model, denoted by $Y_{X_k}$, has been formed via the best subset selection algorithm as follows[43].

**Algorithm 1**. Best Subset Selection Procedure

1. Let $Y_0$ denote the null model, which contains no predictors.
2. **for** $j = 1, 2, \ldots p$ : **do**

(a) Fit all $\binom{p}{j}$ models that contain exactly $j$ predictors.

(b) Pick the best among these $\binom{p}{j}$ models, and term it as $Y_j$.

(Here best is defined as having the smallest RSS, or equivalently largest $R^2$.)

**end for**

3. Select a single best model from among $Y_0, \ldots, Y_p$ using performance metrics such as $C_p$ or adjusted $R^2$, and term this model as $Y_{X_k}$.

In Algorithm 1, a separate least squares regression is fitted for each possible combination of $p$ predictors, producing $2^p$ models in total. In our study, the maximum value for $p = 30$ due to the lack of degrees of freedom for fitting the full model, and this will be further discussed in Section "Variable selection and model analytics".

Step 2 identifies the best model for each subset size ($j$) based on the smallest residual sum of squares (RSS), hence reducing the number of models for consideration to $p + 1$. The RSS is defined as

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (e_i)^2. \end{aligned} \tag{16}$$

Among these $p + 1$ models, a single best model is selected using Mallow's $C_p$ and adjusted $R^2$, which serve as performance metrics for assessing goodness of model fit. The selected model with the optimal metrics is then termed as the best subset model $Y_{X_k}$. The Mallow's $C_p$ is defined as

$$C_p = \frac{1}{n}\left(\text{RSS} + 2j\hat{\sigma}^2\right). \tag{17}$$

The adjusted $R^2$ is defined as

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n - j - 1)}{\text{TSS}/(n - 1)}. \tag{18}$$

A low $C_p$ and high adjusted $R^2$ indicates a good model fit.

### LASSO regression

The LASSO (Least Absolute Shrinkage and Selection Operator) is a regularized regression method that shrinks the coefficients of less important variables towards zero, and can effectively perform variable selection as well. Given a set of predictors $X$ and a response variable $Y$, the LASSO coefficients, $\hat{\beta}_{\lambda_{\text{reg}}}$ are obtained by minimizing the following optimization problem shown in Eq. (19):

$$\hat{\beta}_{\lambda_{\text{reg}}} = \arg\min_{\beta} \frac{1}{2n} |\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}|_2^2 + \lambda_{\text{reg}} |\boldsymbol{\beta}|_1 \tag{19}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients, $n$ is the number of observations, $\lambda_{\text{reg}}$ is a tuning parameter that controls the strength of regularization[43]. The L1 norm, represented by $|\boldsymbol{\beta}|_1$, performs the regularization of the coefficients. The L2 norm, denoted by

$\frac{1}{2n}|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}|_2^2$, corresponds to the residual sum of squares (RSS) of the model. The value of $\lambda_{reg}$ controls the strength of regularization, and determines the number of variables included in the final model. A smaller value of $\lambda_{reg}$ will result in a model with more variables, while a larger value of $\lambda_{reg}$ will result in a model with fewer variables. The choice of $\lambda_{reg}$ can be obtained using cross-validation techniques such as leave-one-out cross-validation. The LASSO regression is chosen over other regularization methods such as the ridge regression due to it's variable selection capability, which can complement the best subset selection method to provide a comprehensive variable selection with robust results. Both techniques can be used to identify a smaller set of variables that are less prone to overfitting, which may then be compared to check if the same variables are selected by both methods. This can help to increase the confidence in the selected variables, yielding a stable and interpretable set of predictor variables that are relevant to the response. By removing variables, the LASSO method can also perform model refinement, resulting in a more parsimonious and intepretable model. We denote the LASSO-regularized regression model as $Y_{X_L}$.

### Cross-validation

Cross-validation (CV) is a technique used to evaluate the performance of a model by dividing the data into a training set and a test set. The model is trained on the training set and its performance is evaluated on the test set. A popular cross-validation method is leave-one-out cross-validation (LOOCV), which is particularly well-suited for the full factorial design used in this study, as it does not require the typical train-test split of the data. Given our small sample size, the use of LOOCV allows us to utilize all available data to fit our model. Hence we employ the LOOCV technique to evaluate the performance of our models. It is also used to determine the optimal value of $\lambda_{reg}$ for the LASSO regression. In LOOCV, the data is split into a train and test set by leaving out one observation from the dataset as the test set, and using the remaining observations as the training set. This process is repeated for each observation in the dataset, resulting in $n$ test sets and $n$ corresponding train sets, where $n$ is the number of observations in the dataset.

In cross-validation, the Mean Squared Error (MSE) is a commonly used performance metric for the test set. The MSE measures the average of the squared differences between the predicted values from the model fitted using the train set, and the true values from the test set, as shown in Eq. (20):

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{20}$$

where $y_i$ is the true value, $\hat{y}_i$ is the predicted value, and $n$ is the number of observations. The MSE for the test set (test MSE) provides an estimate of the model's performance on unseen data, while the MSE for the training set (train MSE) assesses how well the model is fitting the training data. The ratio of the test MSE to the train MSE is used to evaluate model performance and assess overfitting. A model that is overfitting will have a higher ratio of test MSE to train MSE, indicating that the model performs well on the training set, but has poor performance on the test set. Conversely, a model that is not overfitting will have a lower ratio of test MSE to train MSE, which suggests that the model performs well on both the training and test sets.

### Parameter ranking

In the coded variable analysis, the magnitudes of the model coefficients in $\boldsymbol{\beta}$, are directly comparable since they are dimensionless. This allows us to determine the relative sizes of factor effects. These standardized model or regression coefficients measure the effect of changing each design factor over a one-unit interval, and are equivalent to the partial derivatives of the model response with respect to each input factor[12]. The raw values of regression coefficients are seldom used as they incorporate the original units of design factors, which makes the results difficult to interpret[12]. According to Saltelli et al.[42], $\boldsymbol{\beta}$ can be a robust and reliable measure of sensitivity. Therefore, the respective regression coefficients in $\boldsymbol{\beta}$ is used to rank all the parameters in terms of their relative importance on the response variable. In addition, the corresponding $p$-values of the model coefficients can indicate statistical significance of the terms.

## DATA AVAILABILITY

All data generated or analyzed during this study are included in this published article.

## CODE AVAILABILITY

The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

Received: 6 October 2022; Accepted: 16 March 2023;
Published online: 03 April 2023

## REFERENCES

1. Davis, J. R. *Asm specialty handbook: nickel, cobalt, and their alloys.* (ASM International, Member/Customer Service Center, Materials Park, OH 44073-0002, USA, 2000. 442 (2000).
2. Cross, M. et al. Computational modelling of multi-physics and multi-scale processes in parallel. *Int. J. Comput. Methods Eng. Sci. Mech.* **8**, 63–74 (2007).
3. Moges, T. et al. Quantifying uncertainty in laser powder bed fusion additive manufacturing models and simulations. In *Solid Freeform Fabrication Symposium An Additive Manufacturing Conference* (2018).
4. Ning, J., Sievers, D. E., Garmestani, H. & Liang, S. Y. Analytical modeling of in-process temperature in powder bed additive manufacturing considering laser power absorption, latent heat, scanning strategy, and powder packing. *Materials* **12**, 808 (2019).
5. Yang, M., Wang, L. & Yan, W. Phase-field modeling of grain evolutions in additive manufacturing from nucleation, growth, to coarsening. *npj Comput. Mater.* **7**, 1–12 (2021).
6. Tang, H., Huang, H., Liu, C., Liu, Z. & Yan, W. Multi-scale modelling of structure-property relationship in additively manufactured metallic materials. *Int J. Mech. Sci.* **194**, 106185 (2021).
7. Hu, Z. & Mahadevan, S. Uncertainty quantification and management in additive manufacturing: current status, needs, and opportunities. *Int. J. Adv. Manuf. Technol.* **93**, 2855–2874 (2017).
8. Yan, W. et al. Data-driven characterization of thermal models for powder-bed-fusion additive manufacturing. *Addit. Manuf.* **36**, 101503 (2020).
9. Tapia, G. et al. Uncertainty propagation analysis of computational models in laser powder bed fusion additive manufacturing using polynomial chaos expansions. *J. Manuf. Sci. Eng.* **140**, 121006 (2018).
10. Wang, Z. et al. Uncertainty quantification in metallic additive manufacturing through physics-informed data-driven modeling. *JOM* **71**, 2625–2634 (2019).
11. McMillan, M., Leary, M. & Brandt, M. Computationally efficient finite difference method for metal additive manufacturing: A reduced-order dfam tool applied to slm. *Mater. Design* **132**, 226–243 (2017).
12. Montgomery, D. C. *Design and analysis of experiments* (John wiley & sons, 2017).
13. Aiken, L. S., West, S. G. & Reno, R. R. *Multiple regression: Testing and interpreting interactions* (sage, 1991).
14. Box, G. E., Hunter, W. H., Hunter, S. et al. *Statistics for experimenters*, vol. 664 (John Wiley and sons New York, 1978).
15. Leung, C. L. A. et al. The effect of powder oxidation on defect formation in laser additive manufacturing. *Acta Materialia* **166**, 294–305 (2019).
16. Jacob, G. et al. *Effects of powder recycling on stainless steel powder and built material properties in metal powder bed fusion processes* (US Department of Commerce, National Institute of Standards and Technology, 2017).
17. Criales, L. E. et al. Laser powder bed fusion of nickel alloy 625: experimental investigations of effects of process parameters on melt pool size and shape with spatter analysis. *Int. J. Mach. Tools Manuf.* **121**, 22–36 (2017).

18. Kempen, K. et al. Process optimization and microstructural analysis for selective laser melting of alsi10mg. *Solid Freeform Fabrication Symposium* **22**, 484–495 (2011).
19. Shrestha, S., Cheng, B. & Chou, K. An investigation into melt pool effective thermal conductivity for thermal modeling of powder-bed electron beam additive manufacturing. In *Proceedings of the 27th Annual International Solid Freeform Fabrication Symposium*, 207–218 (2016).
20. Wang, L., Zhang, Y., Chia, H. Y. & Yan, W. Mechanism of keyhole pore formation in metal additive manufacturing. *npj Comput. Mater.* **8**, 1–11 (2022).
21. George, E. et al. *Statistics for experimenters: design, innovation, and discovery* (Wiley New York, 2005).
22. Collins, L. M., Dziak, J. J., Kugler, K. C. & Trail, J. B. Factorial experiments: efficient tools for evaluation of intervention components. *Am. J. Prev. Med.* **47**, 498–504 (2014).
23. Kamath, C. Data mining and statistical inference in selective laser melting. *Int. J. Adv. Manuf. Technol.* **86**, 1659–1677 (2016).
24. Ma, L. et al. Using design of experiments in finite element modeling to identify critical variables for laser powder bed fusion. In *International solid freeform fabrication symposium*, 219–228 (Laboratory for Freeform Fabrication and the University of Texas Austin, TX, USA, 2015).
25. Oehlert, G. Comparing models: The analysis of variance. *A First Course in Design and Analysis of Experiments. WH Freeman and Co., New York, NY* 44–52 (2000).
26. Yan, W. et al. Multi-physics modeling of single/multiple-track defect mechanisms in electron beam selective melting. *Acta Materialia* **134**, 324–333 (2017).
27. Yan, W. et al. Data-driven multi-scale multi-physics models to derive process–structure–property relationships for additive manufacturing. *Comput. Mech.* **61**, 521–541 (2018).
28. Yan, W. et al. Meso-scale modeling of multiple-layer fabrication process in selective electron beam melting: inter-layer/track voids formation. *Mater. Design* **141**, 210–219 (2018).
29. Hojjatzadeh, S. M. H. et al. Pore elimination mechanisms during 3d printing of metals. *Nat. Commun.* **10**, 1–8 (2019).
30. Hirt, C. W. & Nichols, B. D. Volume of fluid (vof) method for the dynamics of free boundaries. *J. Comput. Phys.* **39**, 201–225 (1981).
31. Mukherjee, T., Manvatkar, V., De, A. & DebRoy, T. Dimensionless numbers in additive manufacturing. *J. Appl. Phys.* **121**, 064904 (2017).
32. Getling, A. V. *Rayleigh-B nard Convection: Structures and Dynamics*, vol. 11 (World Scientific, 1998).
33. Cai, Y. & Zhang Newby, B.-m. Marangoni flow-induced self-assembly of hexagonal and stripelike nanoparticle patterns. *J. Am. Chem. Soc.* **130**, 6076–6077 (2008).
34. Chan, C., Mazumder, J. & Chen, M. A two-dimensional transient model for convection in laser melted pool. *Metall. Trans. A* **15**, 2175–2184 (1984).
35. Fotovvati, B., Wayne, S. F., Lewis, G. & Asadi, E. A review on melt-pool characteristics in laser welding of metals. *Adv. Mater. Sci. Eng.* **2018**, 1–18 (2018).
36. Robert, A. & Debroy, T. Geometry of laser spot welds from dimensionless numbers. *Metall. Mater. Trans. B* **32**, 941–947 (2001).
37. Van Elsen, M., Al-Bender, F. & Kruth, J.-P. Application of dimensional analysis to selective laser melting. *Rapid Prototyp. J.* **14**, 15–22 (2008).
38. Yates, F. *The design and analysis of factorial experiments* (Imperial Bureau of Soil Science Harpenden, UK, 1978).
39. Drum, M. Yates's algorithm. *Encyclopedia of Biostatistics* **8**, 6195–6196 (2005).
40. Chambers, J. M. *Graphical methods for data analysis* (CRC Press, 2018).
41. Gygi, C. & Williams, B. *Six sigma for dummies* (John Wiley & Sons, 2012).
42. Saltelli, A. et al. *Global sensitivity analysis: the primer* (John Wiley & Sons, 2008).
43. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An introduction to statistical learning*, vol. 112 (Springer, 2013).

## ACKNOWLEDGEMENTS

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (MOE-T2EP50121-0017). We would like to thank Assoc. Prof. Zhisheng Ye for his valuable advice and words of wisdom. Insightful discussions with Dr. Padmeya Indurkar and Prof. Goh Thong Ngee are also sincerely acknowledged.

## AUTHOR CONTRIBUTIONS

A.G. collected the simulation data under the guidance of W.Y., ideated the data analytics approaches, executed the coding and result interpretation, drafted and revised the manuscript. F.C. contributed to the high-fidelity thermal fluid-flow simulations used in the paper. J.C. checked the data analytics approaches and results, provided guidance and revised the manuscript. W.Y. conceived the project, executed the thermal-fluid flow simulations, provided guidance, revised the manuscript, and led the study.

## COMPETING INTERESTS

The authors declare no competing interests.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41524-023-01004-9.

**Correspondence** and requests for materials should be addressed to Jiaxiang Cai or Wentao Yan.

**Reprints and permission information** is available at http://www.nature.com/reprints

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