UNIVERSITY OF DEBRECEN
FACULTY OF ENGINEERING
DEPARTMENT OF
MECHANICAL ENGINEERING

MACHINE LEARNING-BASED INVESTIGATION OF THE COMPRESSIVE BEHAVIOR OF TI6AL4V LATTICE STRUCTURES

THESIS

Name: Maliha Binte Hasan
BSC „Mechanical Engineering" student
Supervisor(s): Dr. Mankovits Tamás
Head of department:Dr. Mankovits Tamás

Automotive Production Process Control Specialization

Debrecen
2026

Table of Contents

Table of notations

y_i	Actual Value
$\hat{y}_i$	Predicted Value
n	Total Number of Prediction
p	Number of Predictors (independent variables)
$F\ [N]$	Force
X	Original Data Value
X'	Normalised Value
μ	Mean of the Feature
σ	Standard Deviation

1 Introduction

In the recent world of biomedical implants, the use of Titanium alloys to make lattice structures and use them as an implant for bones has become widespread. These implants are required more because of fractures, osteoporosis and other degenerative bone diseases [1]. Another reason is that in modern healthcare bone implant operations are becoming more common [1]. It is because of the good biocompatibility of the Ti6Al4V[1]. This means Ti6Al4V can integrate with the surrounding bone tissue which improves implant stability [2]. It is possible for the alloy to integrate easily because of its high resistance to corrosion in biological environments and high resistance to fatigue under cyclic loadings [2]. These implants are additively manufactured with the titanium alloy[2]. Highly complex geometries can be 3D modelled through Additive manufacturing. This enables efficient production of lattice structures of which the porosity is controlled accurately [10]. Based on each patients' anatomical requirements, customised implants with customised porosity can be produced [10]. This is immensely helpful as porous structures improve the bone in growth through the implants [2], [14]. As the implants replace the bones in the body, it is very crucial for them to have the same mechanical properties as the human bone[3]. If there is any kind of mechanical mismatch between the implant and the actual bone, then the load on the implant and the bone will differ [9]. This leads to localised stresses [9]. Matching the stiffness of implant and bone lengthens the implant's life [9]. For this reason, the accurate optimisation, and prediction of the stiffness of Ti6Al4V lattice structures are very important in the biomedical field[3].

Finite Element Analysis helps in simulating mechanical behaviour under stress and thus predicting the effective young's modulus of a certain lattice structure based on lattice variables like strut thickness and strut length[3], [4] . This means FEA allows us to evaluate the mechanical properties of lattice structures before manufacturing them [3]. This numerical simulation subsequently reduces the cost of manufacturing and experimentations as the structural and mechanical response can analysed from given loading conditions [3]. And thus, deformations and stress concentration zones can also be identified [4]. The issue is each simulation is computationally extensive and costly[5]. This is due to the rise of computational with increasing geometric combinations [5]. Moreover, to study large designs for simulations it is required to carry out repeated simulations. That means; to generate more data for the investigation of which variables are more suited as implants, we must compromise on data efficiency and computational cost[5]. Expansion of the implant lattice parameters becomes difficult through only using FEA [5]. That is why to reduce simulations carried out while maintaining prediction accuracy, efficient sampling is needed [36].

Machine learning models in this regard can open a new door for us as it can predict from a given set of data[6]. Machine learning models creates relationships between given input and outputs [6]. Because of this, Machine learning models such as regression models can be used to predict the mechanical behaviour of implants like effective young's modulus[6]. Machine learning has gained popularity in material sciences as quick prediction of the outcomes

improves design efficiency [7]. Machine learning enables swifter comparison of geometric designs [7]. But the problem is, these machine learning models also has many lacking, for example too much data can diverge them from the accurate prediction ability[7]. To explain it more, large input and output parameters mean large dataset to feed to the machine learning models. Even though theoretically increasing dataset should mean more precise prediction of the machine learning models, redundant sample may originate prediction noise [7]. That means performance of the models becomes dependant on the type and size of parameters [36]. Efficient data selection becomes important for accurate machine learning model prediction [36].

That is why this study will first collect data from Finite Element Analysis and then use them to find out exactly how much data is needed for each model to be able to predict the Young's modulus from the strut inputs and investigate how valid these results are. Total of four regression models are compared in different aspects to compare their prediction capability and data efficiency. The four regression models to be investigated are Linear Regression, Polynomial Regression, Gaussian Process Regression, and Support Vector Regression models. The comparison will be done by their performances which is calculated using statistical metrics such as RMSE and R^2 . To ensure the reliability and consistency of the performance of the models, Cross-validation method is applied across different training subsets. After that the study will analyse the locations of the learning points, which have the best training ability, selected by models. The pattern of the selected parameters from the regression models will tell us if there is any certain geometric combination of strut thickness and strut length that provide greater training capability for the models. This will help us to reduce the number of required simulations which in return reduce the computational cost all while maintaining prediction accuracy. By this, the study aims to solve the problem of large data collection. The methodology proposed will contribute toward efficient data-driven Ti6Al4V lattice structure optimisation for the biomedical implants.

2 Literature Review

2.1 Background: Medical Mismatch between bone and implants

One of the major issues with orthopaedic implant design is when the design of bone structures must compromise between adequate porosity and mechanical stability to provide support for bone tissue regeneration without inducing stress shielding [8]. Also, due to stress shielding effect, periprosthetic osteolysis occurs, which means improper stress distribution resulting in bone density losses in underloaded zones and bone overproduction in stress localized zones[9]. The root cause of this is the difference in the elastic modulus between the implant and the surrounding bone tissue, which causes the implant to loosen and bone to be reabsorbed into body[2], [10].

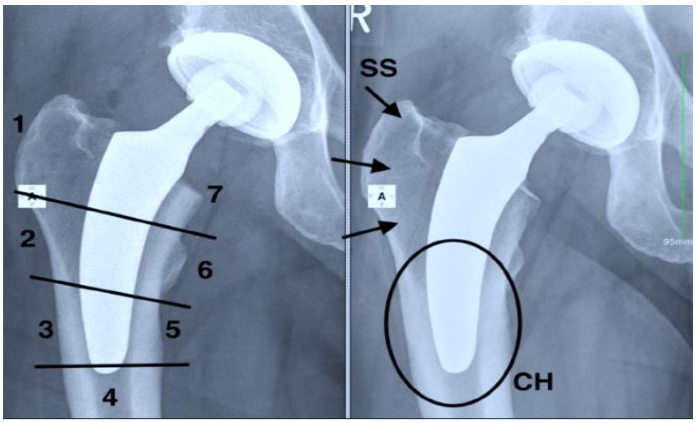

Figure 1 Stress Shielding indicated by SS[41].

2.1.1 Importance of TiAl4V implants

Cui et al. (2024) demonstrated that titanium based alloy was superior choice to use it for load bearing purposes due to its high corrosion resistance, biocompatibility and the absence of limitations like toxicity, high elastic modulus and rapid degradation, which are prominent in other commonly used metal alloys like stainless steel, magnesium-based alloys, cobalt-base alloys[2]. Additionally, TiAl4V out of the titanium-based alloys have been shown to be best choice for implant as their mechanical properties closely resemble of as of human bones[1].

2.1.2 Why lattice structures are used

Lattice structures have high strength to weight ratios, versatile biomechanical properties, and osteointegration and internal bone growth ability, which increases their secondary stability[11]. These structures are customizable, resistant to impact and maintains stiffness even with filler material[12]. Moreover, properties like stiffness, yield strength, energy absorption are often highly sensitive to the design variables of cell type, size and porosity. Functionally graded and TPMS derived lattice structures offer maximum strength with minimum weight property across biomedical, aerospace and automotive application[12], [13].

2.1.3 Importance of accurate mechanical behaviour prediction

The mechanical behaviour of orthopaedic implants can be tailored to an individual patient need by manipulation of the porosity and strut geometry of additively manufactured lattice structures, enabling tuning of the effective elastic modulus and reducing the risk of stress shielding[11]. Given the critical function of implants as functional body parts, it is essential that their mechanical and biomechanical behaviour with the surrounding tissues is effectively predicted to ensure safety and performance[3]. In addition, the mechanical behaviour of lattice structures produced by selective laser melting (SLM) is significantly affected by variations in the structure, and therefore, a robust modelling approach such as finite element analysis (FEA) needs to be used to include manufacturing variability[11].

2.1.4 Key properties of Ti6Al4V

As mentioned before, Ti6Al4V (ti64) titanium alloys are the most preferred biomaterial for bone implants as they have great biocompatibility, strength corrosion resistance and fatigue behaviour[2], [14]. Taking advantage of the properties of lattice structures, parameters such as unit cell geometry, pore dimensions and structural topology of ti6al4v lattice implants can be changed and designed to achieve enhanced osseointegration[15].

2.1.5 Influence of geometrical parameters of lattice structures on mechanical behaviour

Mechanical behavior of Ti6Al4V lattice structures is greatly affected by geometry dependent parameters like strut thickness, strut length and unit cell size. The effect on behavior include the changes in properties like yield strength, elastic modulus and surface area to volume ratio; Bittredge et al(2022) concluded in a compression test that an optimal structure with 1 mm strut diameter and 5 mm length showed 200 MPa yield strength and 5 GPa elastic modulus, which can be a potential application for shoulder implants[10]. The mechanical properties like strength or stiffness of the lattice structures depend on the relative density and this density is determined by the size, thickness and how the strut rods are connected in unit cells.

2.1.6 Additive manufacturing methods for ti6al4v

Additive manufacturing (AM) methods such as Selective Laser Melting (SLM) and Electron Beam Melting (EBM) have significantly transformed the production of Ti6Al4V biomedical implants to enable the development of complex, patient-specific lattice structures with enhanced mechanical compatibility and osseointegration[2]. From these, SLM has been extensively applied to the fabrication of porous scaffolds and cellular structures, as demonstrated in various studies on the compressive behaviour and structural precision of Ti6Al4V lattices fabricated by this process[1][16].

2.1.7 Importance of FEA in lattice structure

Finite element analysis has been largely used to carry out investigation on the mechanical performance of Ti6AL4V lattice structures, especially to find out the compressive strength, effective young's modulus and stress distribution when the lattice structure is subjected to loading[3], [13]. FEA simulation tests have shown good association with the values obtained

from laboratory compression tests which establishes the application of numerical methods for predicting the mechanical behavior in biomedical implant design[3].

2.1.8 Boundary conditions, meshing strategy and model validation

A commonly used meshing type is tetrahedral meshing type as it provides greater control over the mesh by locally controlling the strut diameter, strut intersection rounding cell size and thus the density of the lattice and guarantees a suitable mesh throughout the geometry which is crucial for optimizing mechanical properties[4], [17]. Also the patch independent method inside the tetrahedral meshing is used for its efficiency among faceted bodies and accuracy in analyzing the stress and deformation of complex lattice structures[4], [17]. Together with tetrahedral meshing and patch independent algorithm ensures the accuracy and reliability of the FEA results[17]. The boundary conditions are applied in FEA simulations to mimic the physical compression experiments[4]. Model verification is done by comparing the results obtained through simulation and the experiment data and less than 8% error reporting to be a well calibrated model[18]. Stiffness matrix simulation and physical testing can further confirm the validity of the FEA prediction[18].

2.1.9 Limitations of FEA

Despite FEA contributing significantly to testing of mechanical behavior assessment, limitation in modeling level and boundary condition still prevails which hinders the accuracy of the FEA results as it is directly dependent of how well these aspects imitate real world scenario[5], [18]. Another limitation of finite element simulation is the computation cost when dealing with a large set of datasheets[5]. Last but not the least, though finer meshes improve the resolution of the results, they also become computationally extensive[18]. These issues motivate the adoption of machine learning approaches[5].

2.2 Machine learning and regression techniques in Material Engineering

2.2.1 Benefits that machine learning brings

Despite FEA contributing significantly to testing of mechanical behavior assessment, limitation in modeling level and boundary condition still prevails which hinders the accuracy of the FEA results as it is directly dependent on how well these aspects imitate real world scene. Overcoming the limitations that the traditional FEA method brings, machine learning has proven to be the rising interest in the field of materials science and computer science due to its ability to predict the mechanical performance of materials, such as yield strength, young's modulus and compressive strength as well as to predict material designs such as lattice constants of lattice strictures[5], [19]. Integrating machine learning and information of materials derived from FEA to analyze large data set simplifies the workflow, optimizes the processes and saves time and effort for accurate data prediction or identification, which in turns alleviates the computational costs[6], [20], [21]. These properties make ML capable of handling high design efficiencies required in biomedical and engineering fields[6].

2.2.2 Types of Regression based models for property prediction

Machine learning methods like Support Vector Regression (SVR), Artificial Neural Networks (ANN), Gaussian Process Regression (GPR), Decision Trees (DT), polynomial regression(PR)and Linear Regression (LR) have been applied for the prediction of mechanical properties and designing with acceptable accuracy[22], [23]. ANN use neurones, interconnected processing elements, to carry out complex tasks and are highly effective in preserving the complex, non linear lattice geometry[6], [22], [24]. SVR, A variant of support vector machine, uses kernels to solve nonlinear problems and is specifically adapted to solve regression problems by finding a hyperplane which minimizes the distance between hyperplane and training data[22]. GPR uses Bayes' rule and provided training data to update the probabilities of each function which represents model and is a non parametric regression technique[22]. Linear Tree uses linear regression, whereas polynomial regression incorporates polynomial terms to solve non linear patterns and when linear method is insufficient[24].

2.2.3 Model Validation and Verification methods

To evaluate the accuracy and validity of the machine learning and regression algorithms, the performance of these models against new data is evaluated using various types of metrics after training them with a known dataset[3], [25]. Some of the repeatedly used metrics are:

- $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
- $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- Coefficient of Determination, $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$
- Adjusted $R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$

[19], [22], [25], [26], [27]

Also, K-fold cross validation like tenfold cross validation, fivefold cross validation and leave one out cross validation (LOOCV) are applied on the machine learning models to establish that the models are generalized, avoid overfitting thus becoming more robust models [22], [28], [29], [30].

2.2.4 Comparison of different regression models from literature

Multiple studies have shown that SVR, GPR and ANN had higher accuracy levels when predicting young's modulus compared to other linear models. For example, Liu et al(2024) used SVR to estimate statistic modulus for sandstone and the validation results were R^2 = 0.98 and RMSE= o.11 GPa.[23] In another study, GPR achieved R=0.89 for predicting young's modulus whereas the linear regression model achieved r=0.78.[31] ANN presented accuracy of 99.97% for estimating the equivalent elastic modulus of multicoated lattices using 4 input parameters[24], [26]. On the other hand DTR and MLR show less flexibility for modeling strong non nonlinear relationship as opposed to SVR, GPR and ANN who outperformed with their robustness and accuracy when dealing with scarce dataset, complex nonlinear data [19], [23], [24], [26], [32].

However, several studies have reported that the accuracy of models heavily depends on the dataset size and sampling strategy of data [7], [23], [25], [33], which urges us to look into data efficiency and learning point optimization, final and most important part of this literature review.

2.2.5 Limitations of large or unoptimized datasets

Even though, theoretically it is expected that large datasets can improve the accuracy of the machine learning regression models, large or unoptimized datasets often bring limitations in computational efficiency and predictive accuracy real life[7], [34], [35]. such as:

- Computational Cost and inefficiency: To collect each new data for large dataset, each simulation in FEA consists of meshing, applying boundary condition, and solving multiple load steps, which requires high level computational effort and thus, large dataset results in being both time and resource consuming[36], [37]. Furthermore, in the study of predicting mechanical properties of lattice structures, these limitation is even more prominent as each geometry variant needs different model preparation and meshing[38].
- Curse of dimensionality: A phenomenon which arises when there are many inputs in regression models and leads to sparsity due to feature expansion that reduces generalization capability in regression models[7], [35]. Stability of SVR and GPR models gets negatively affected the most as they heavily depend on the data density in feature space for prediction process[7], [34].
- Overfitting and instability: Too much untreated data can clutter the feature space, and the models end up using unnecessary data, making themselves unable to generalize to new data[7], [23], [36], [38].
- Kernel sensitivity: when dataset is unoptimized and large, kernel tuning parameter becomes more complex and sensitive to local data variations, which increases training time for kernel-based models like SVR and GPR[35], [39]. Even ANN faces the same consequence, for example higher computational time and resources used, when the dataset is large and unoptimized[34].

2.2.6 Approaches to minimize training data requirements

Firstly, systematic sampling strategy ensures that the entire input space, which is the combination of strut length and thickness in case of lattice structures, is properly represented and gives better performance in prediction accuracy for ensemble-based models with limited data available compared to simple random sampling[34], [35], [37], [40]. Tao et al(2024) depicts how sampling strategies reduce redundancy and increase representativeness of training points[34].

Another effective way is sensitive based feature selection method, which analyses the effect of each input on the output and ranks the features accordingly and is described to have improved the feature reliability, showing how reducing feature dimensionality can lower the required data size[29]. thus, lowering the ‘curse of dimensionality’.

Last but not the least, Active learning, based on model uncertainty, predicted error and exploitation or exploration trade off, identifies the most informative data points for model training. Liu et al (2024) demonstrated that given a proper sampling and representation strategy, active learning needs only as much as 5% of all available simulations for structure

property regression tasks[36]. Since each new data sample is used for maximum information gain in active learning, active learning allows the machine learning models to perform well even with the small data size[35].

2.2.7 Research gap and Motivation for present study

While methods like active learning and feature selection help reduce the required training points[34], [35], [36], very few studies investigated on the exact number of learning points truly necessary for the accurate prediction of Youngs modulus in Ti6al4v lattice structures. In addition, most of the existing research studied general data efficiency or model accuracy without investigating the optimal spatial distribution of learning points withing design space. In the limitation section it was mentioned that poorly distributed data samples can reduce the model generalization and accuracy[36], [38].

Finally, even though regression models like SVR, GPR, ANN have shown to perform well even in small sample [7], [35], [38], there is not enough evidence on which model's performance is better and more robust against varying dataset sizes and distributions for lattice's material properties. These gaps highlight the need for investigation into the topics:

- Optimum number of learning points needs to accurately predict the young's modulus of lattice structures.
- Ideal location and distribution of these learning points in the input space (strut thickness vs strut length).
- Comparison of the different regression models' performance under data-efficient condition.

3 Methodology

Goal: Evaluate the impact on the effective young's modulus of the titanium lattice structures from the strut length and thickness and find the optimal number of learning points as well as their locations needed to train the machine learning based predictive model.

3.1 Data Generation Via FEA

3.1.1 Geometry Creation

Tool: Ansys Workbench 2025 R2

In the static structural section of the work bench, 1st the material type was selected as the **Ti-6Al-4V** alloy. The titanium alloy was assumed to be linear and isotropic with young modulus of 106247MPa and Poisson's ratio of 0.34. Then 20*20*20 3d infill lattice structures with varying strut lengths (0.7mm - 0.5mm) and strut thicknesses (0.5 mm-0.3 mm) were modeled in ANSYS Space claim. Total 25 lattice structures were modeled and investigated.

3.1.2 Boundary Condition

The next step was comprised of assigning loading and constraints in Ansys mechanical for the structural analysis of the lattice structure. On one plane compressive loading of 2300N along the y axis was applied and a fixed support was selected on the other side of the lattice structure. Rest of the surfaces were frictionless.

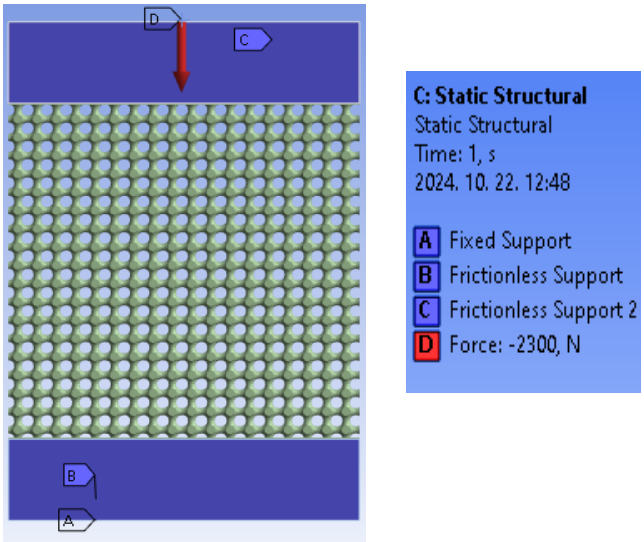

Figure 2 Demonstrates the boundary conditions applied on the the 3d infill lattice cube in the Mechanical solver on Ansys Workbench 2025 R2

3.1.3 Meshing Strategy

Meshing was carried out by using tetrahedral element and patch independent method, as patch independent method can generate high quality mesh of complex geometries like lattices

without the need for surface base topology, where the minimum size limit was always the respective strut thickness of the lattice structure. Linear element type was selected to ensure the balance between computational efficiency and acceptable accuracy

3.1.4 Solution Process

In the post processing part, a user defined result was created to extract the total deformation of the lattice long the direction of the compression. "-uy" expression was used so that the solution gives the negative displacement along the y direction (direction of the loading). Then the obtained displacement value was used as an input of a sequence of calculations for which the output is young's modulus.

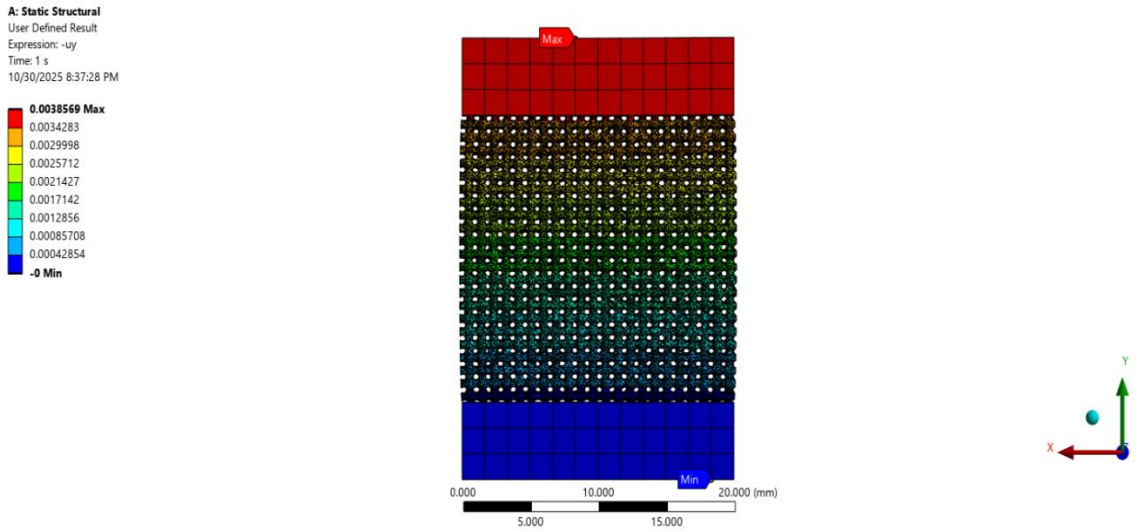

Figure 3 Shows an example of User Defined Result solution of a lattice structure with 0.5mm strut length and 0.4mm strut thickness

3.1.5 Dataset Preparation

As mentioned above, the data set comprised of 25 observations with 2 input variables, strut length and thickness, and one output feature, young's modulus. Strut lengths of {0.7,0.65,0.6,0.55,0.5}mm and strut thicknesses of {0.3,0.35,0.4,0.45,0.5}mm produced a comprehensive input parameter space of 5*5 matrix of 25 combinations.

- $Stress(MPa) = Force/Area, Area = 20mm \times 20mm$
- $Strain = \frac{Displacement}{20mm}$
- Young's modulus (MPa) $E = Stress/Strain$

Table 1. Calculated Young's Modulus of the respective geometric inputs

Strut Length(mm)	Strut Thickness(mm)	User defined result (mm)	Stress (MPa)	Strain	Youngs modulus (MPa)
0.7	0.3	0.0073255	5.75	0.000366275	15698.58713
0.7	0.35	0.0062363	5.75	0.000311815	18440.4214
0.7	0.4	0.0054876	5.75	0.00027438	20956.33793
0.7	0.45	0.0048885	5.75	0.000244425	23524.59855
0.7	0.5	0.0041339	5.75	0.000206695	27818.76678
0.65	0.3	0.0068549	5.75	0.000342745	16776.32059
0.65	0.35	0.0058579	5.75	0.000292895	19631.6086
0.65	0.4	0.0050393	5.75	0.000251965	22820.62985
0.65	0.45	0.0044648	5.75	0.00022324	25757.03279
0.65	0.5	0.0040143	5.75	0.000200715	28647.58488
0.6	0.3	0.0064032	5.75	0.00032016	17959.77011
0.6	0.35	0.00525	5.75	0.0002625	21904.7619
0.6	0.4	0.0046038	5.75	0.00023019	24979.36487
0.6	0.45	0.0040515	5.75	0.000202575	28384.54893
0.6	0.5	0.0037512	5.75	0.00018756	30656.85647
0.55	0.3	0.0056204	5.75	0.00028102	20461.17714
0.55	0.35	0.0048975	5.75	0.000244875	23481.36804
0.55	0.4	0.0044902	5.75	0.00022451	25611.33134
0.55	0.45	0.0037191	5.75	0.000185955	30921.45949
0.55	0.5	0.0033883	5.75	0.000169415	33940.32406
0.5	0.3	0.0050581	5.75	0.000252905	22735.80989
0.5	0.35	0.0043019	5.75	0.000215095	26732.37407
0.5	0.4	0.0038569	5.75	0.000192845	29816.69216
0.5	0.45	0.0034091	5.75	0.000170455	33733.24338
0.5	0.5	0.0031029	5.75	0.000155145	37062.10319

3.1.6 Data Normalisation and Splitting

Data collected then were normalised to ensure the efficiency and accuracy of the models as it prevents any outlying numerical range from disproportionately influencing the model's learning process. All the input features have a mean of zero and standard deviation of one and the standardisation equation that was followed:

$$X' = \frac{(x-u)}{\sigma}$$

Data was divided into training set (80%) and the testing set (20%) so that the models could learn from a large share of the data and the testing set can be used to assess the generalization ability. K-fold cross validation method was used (where K=5) where the data was split into 5 subsets and 4 of them were used to train models and 1 was tested on them. This method rotated between the 4 folds. This division reduces the overfitting in the models. Also, to ensure the reproducibility of the results, rng(1) was in the beginning of MATLAB code to maintain the consistency of the random partitions.

3.2 Regression Model Development

3.2.1 Model selection and Justification

Tool Used: MATLAB

4 machine learning regression models were investigated for the predictive analysis of effective young's modulus and they are: LR(Linear Regression), PR(polynomial regression), SVR(Support vector regression), GPR(Gaussian process regression). LR is responsible for capturing the linear relationship[25], PR captures nonlinear relationship through polynomial features[25] , SVR uses kernel to solve non linear relationship[25], [36], [37] and GPR uses probality prediction[36], [37], and estimation to solve complex non linear relationship, which is the common characteristic between the strut length, thickness and the effective youngs modulus of lattice geometry.

3.2.2 Model Formulation

The mathematical formulation of each models are given below

LR: $y = \beta_o + \sum_{i=1}^{n} \beta_i x_i + \epsilon$ [25]

PR: degree=2; $y = \beta_o + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \in$ [25]

Where, $y = dependent\ variable$, $x = independent\ variable$,

$\beta_o, \beta_1, .., \beta_n = regression\ coefficient,\ \in = error\ term$ [25]

SVR: Uses RBF kernel function and minimises the ∈-insensitive loss function.

GPR: Uses joint Gaussian to predict and assume distribution over the functions[36], [37].

LR and PR were trained using MATLAB's $fitlm()$ function, SVR was trained using fitrsvm() function and GPR was trained using $fitrgp()$ function.

3.2.3 Training and Validation Approach

As mentioned in the data preparation, all models were using 5-fold cross validation method where in each fold 80% of the data used to train the model and 20% of data was used to test the model to compare its predictive ability with the actual values to ensure robustness against limited data. The function $cvpartition()$ was used in MATLAB to employ the 5-fold cross validation method.

3.2.4 Performance metrics

RMSE (Root Mean Square Error), R^2 (Coefficient of Determination), MAE (Mean absolute Error) metrics were averaged across folds of the validation method of each regression models for the comparative analysis of their predictive ability. These metrics tell us the accuracy level of the predictions done by the models trained with the 25 sets of data provided.

3.2.5 Method Robustness

To ensure the robustness and reliability of the of the comparison the 5-fold cross validation procedure was repeated total 5 times with different random number genetaor seeds, $rng()$ values. Although there were minor variations in the RMSE values of the models among the seeds $Rng(1), rng(43), rng(97), rng(167)$ and $rng(987)$, the overall ranking of predictive accuracy among the models based on the rmse values remained the same. This consistency tells that the results are sensitive to the random initialisation of data and thus confirming the robustness and the reproducibility of the investigation.

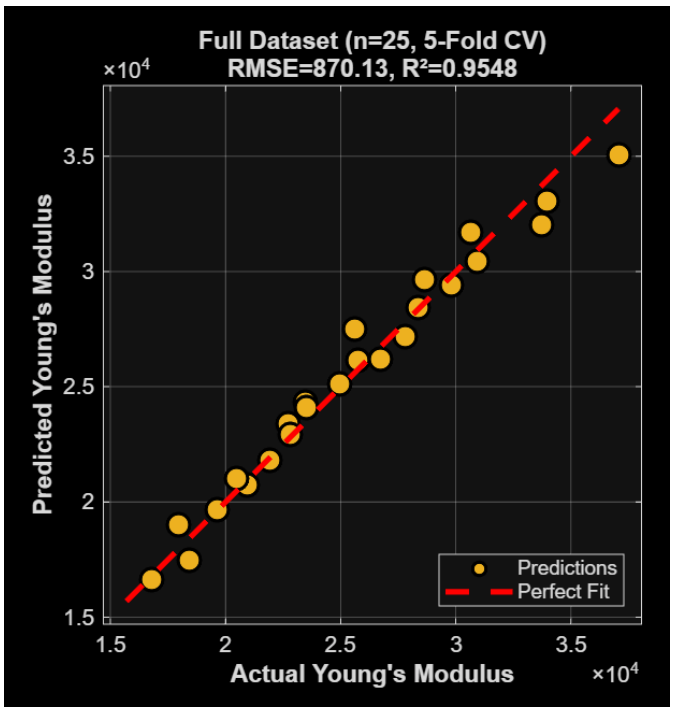

Figure 4 Predictive ability of LR at full training size. It shows the predicted points across plotted across a straight accurate line.

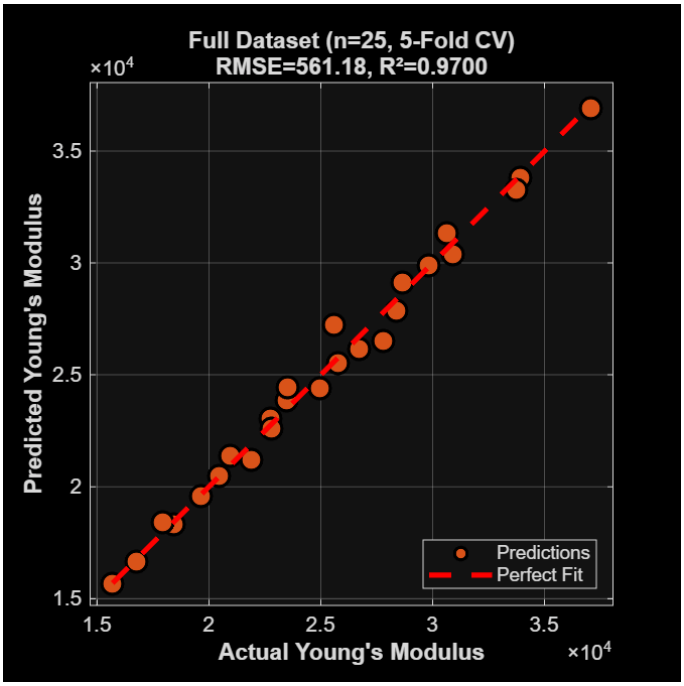

Figure 5 Predictive ability of PR at full training size. It shows the predicted points across plotted across a straight accurate line.

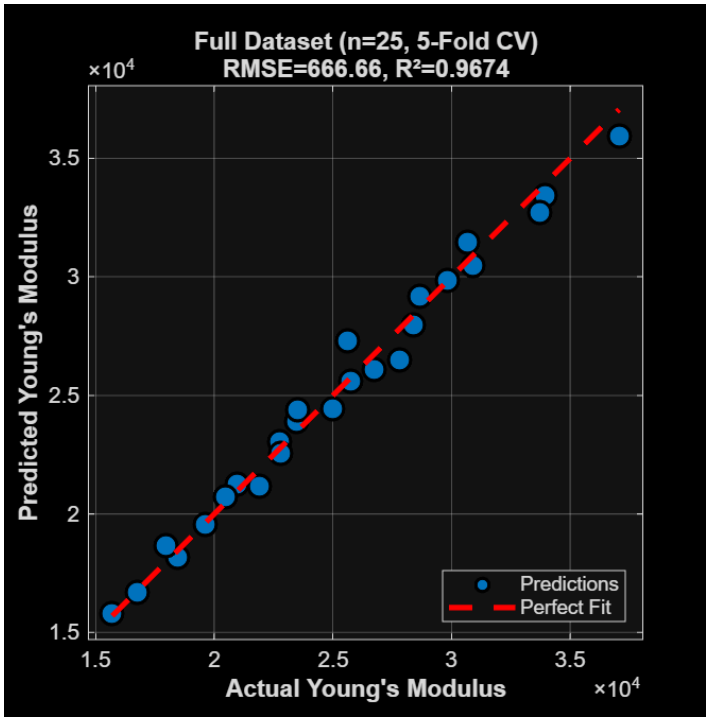

Figure 6 Predictive ability of GPR at full training size. It shows the predicted points across plotted across a straight accurate line

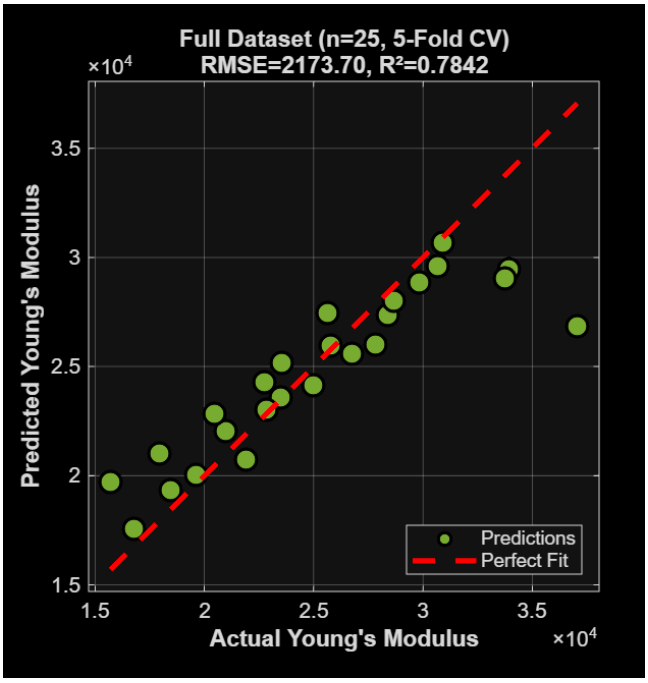

Figure 7 Predictive ability of LR at full training size. It shows the predicted points across plotted across a straight accurate line

3.3 Finding Optimal Training Size

The next step of methodology was carried out with the purpose of finding out the minimum number of training sample beyond which the performances of the 4 models do not improve significantly meaning how many learning points are enough to train the models for predictive analysis. This part of methodology shows how k-fold cross validation method and multiple random seeds were used to establish optimal size and its robustness.

3.3.1 Input Data and Configuration

The variables used were represented by x and y where x is the input matrix comprised of $[strut\_length, strut\_thickness]$ and y is the output effective young's modulus obtained from the FEA process. All the 25 data obtained were used as 25 the learning points.

Just like in the regression model development process, multiple random seeds were evaluated for the robust analysis of optimal number of learning point. Seeds tested were 1, 25, 42, 89, 123, 359, 456, 685, 789, 1478, 2024, 2567, 3141, 4999, 5678, 8354, 9999, 11099, 12345, 13456. Thus, 20 different random initialisation was carried out with the function $seeds\_to\_test$ and the results for all the seeds were stored to compare and analyse the optimal RMSE and R^2 among each seed. Because each seed changes the training split each time and thus helps estimating variability.

For each seed, 5 fold cross validation, a type of k fold cross validation method, was implemented, where in each fold the dataset was divided into 80%:20% or 4:1 and 4 subsets were used for training and 1 for testing, recurrently. Performance metrices named RMSE and

R^2 were calculated for each and averaged to account for predictive accuracy and error variance.

3.3.2 Incremental Training Size Evaluation

Each model (GPR, PR, SVR, LR) was trained with increasing number of learning points from 5 to 24 and for each size the mean RMSE, standard RMSE and mean R^2 were recorded across folds and seeds. The RMSE values collected were used to generate the learning curve for all models where the curve showed the relationship between RMSE and the training sizes. Then a convergence threshold for RMSE was set as 3% ($threshold = 0.03$) with the $required\_consecutive = 3$, needing 3 consecutive stable points to decide on the optimal number, on the learning curve to detect at which size the cross-validation error plateaus. Once the improvement between the training sizes was less than 3% the preceding size was considered to be the optimal number of learning points needed to train the respective model.

The MATLAB code used:

*abs_change_pct = abs((rmse_vals(i + j + 1) - rmse_vals(i + j)) / rmse_vals(i + j)) * 100*

3.3.3 Aggregation and Statistical Analysis

Using the $for\_seed\_in\_range(20)$ function, the whole optimal learning point number process was repeated 20 times, where each iteration investigated all possible training sizes and selected the first size from each seed where the RMSE improvement value between two consecutive data sizes was less than the threshold value (3%). Therefore, each seed gave out one optimal number of learning points to train that particular regression model. These stored optimal numbers are used to carry out statistical analysis like generating learning curves with the help of several computational metrics to determine the final optimal number. In the Matlab code the metrics, to quantify the robustness of the investigation, used were:

- Mean optimal points: The average of the 20 optimal sizes found. $\text{Mean} = \text{np.mean(optimal\_points)}$
- Standard deviation (SD): measures the dispersion of the optimal number across random seeds
- Median optimal points: Used as the final representative of the values as its is less sensitive to outliers.
- Mode optimal points give which optimal size number occurred the most.
- Range of optimal points: The difference between the maximum and minimum optimal size number across the 20 seeds.
- Coefficient of variation (CV): it is a normalised measure of variability which calculates the ratio of SD to mean in percentage. $CV = (SD/Mean) * 100$

Meaning, a lower CV indicated higher consistency and statistical stability across seeds.

3.3.4 RMSE and R^2 at Optimal Points

The MATLAB code stored the RMSE and R^2 value corresponding to each seed's optimal number of learning point and calculated the mean of RMSE and R^2 with standard deviation of them. Mean RMSE ± SD and Mean R^2 ± SD, where lower RMSE value mean better prediction accuracy and Rsquare closer to 1 shows closer association between the model's

predicted and actual values. Smaller SD means higher consistency of performance across seeds.

3.3.5 Statistical Interpretation

The optimal number of learning points for each model taken to be the point from where RMSE and R^2 learning curves and the Aggregated learning curve plateaued. SD and CV are considered to ensure the robustness and accuracy of the decision.

3.3.6 Model validation

To ensure the consistency with the models' initial comparison of predicted vs accurate results, a final 5-fold cross validation was performed after finding the optimal number and recorded. The collected optimal number of learning needed to train the models accurately were then used to investigate on finding the location and distribution of the n number of learning points which can accurately train the respective regression models, where n is the optimal number of learning points recorded for each model.

3.4 Location of Optimal Number of Learning Points

After finding out the optimal number of learning points across input data set, the next logical step is to analyse which are the specific learning points that contribute most to the train of models for accurate prediction. Finding out the location of these points will reduce the redundancy of data and increase data efficiency and thus increasing model generalisation.

3.4.1 Data and Setup

In this process, the input data were the 25 sets of strut length, strut thickness and their young's modulus and the number of optimal training sizes for each model recorded from the previous step. This process also used the same 20 multiple random seeds (seeds tested: 1, 25, 42, 89, 123, 359, 456, 685, 789, 1478, 2024, 2567, 3141, 4999, 5678, 8354, 9999, 11099, 12345, 13456) for the data and test consistency and reproducibility of the results. Each of these seeds had random split of training and testing points ensuring the robustness and remove any stochastic variability of this process.

3.4.2 Multi Seed Training Procedure

The code used uses the ($Seeds\_to\_test$) function to investigate how different initialisation affects the selection of learning points locations where each seed permuted completely random data. Each seed chooses the training points based on the number of learning points previously determined given to them ($optimal\_n\_train$) and fits the regression models e.g. $fitgrp()$ for GPR, $fitrsvm()$ for SVR, $fitlm()$ for PR and LR followed by finding the RMSE on the remaining test data. The code uses 100 iterations($n\_iteration = 100$) so the seed carries out median performance selection and chooses the iteration which gave out the median RMSE to reduce the impact of outliers and overestimating the performance of the respective models due to one lucky random split. The training points locations from each seed is then stored for further frequency and statistical analysis to observe how each data points contribute across multiple model realisations.

3.4.3 Frequency Analysis of Data point Selection

In this step the variable used was $selection\_frequency$ to observe how many times a particular data point was selected across the 20 seeds and

$$Selection\ Frequency = (frequency/number\ of\ seeds) * 100\%$$

was calculated to provide a quantitative measure of appearance of the learning points. The data points were then labelled into following way:

- Aways Selected: Selection frequency 100%. These data points were selected across all the 20 seeds.
- Often selected: Selection frequency ≥ 60%. Considered highly informative data points.
- Sometimes selected: Selection frequency ≤ 60%. Considered less informative and useful.
- Never selected: Selection frequency 0%. They were not selected across any single seed.

Learning points with high frequency were considered the representative sample across the input feature space. On the other hand, data points with low selection frequency were considered redundant for the training of the models. The data informativeness were analysed and stable learning points were determined from this step. The code produced graphs and chart including threshold lines for the frequency analysis for better visual understanding and analysis.

3.4.4 Distribution Consistency

The data points selected for each seed had specific strut length and thickness which were recorded to see how the strut length and strut thickness varied across the seeds. If the any of the strut length or thickness value appear more frequently than others, then that means those regions of feature space are critical for prediction. For strut length and strut thickness the mean ± SD was computed to get the solution for $length\_distribution$ and $thickness\_distribution$. If the distribution varies strongly across the seeds then the model sensitivity to initial sample is considered higher.

3.4.5 Spatial And Performance Analysis

The visualisation of the spatial selection pattern is showcased through the 2D scatter plot of strut length and thickness coloured with frequency. RMSE across seeds are plotted to check for the generalisation consistency performance variability across runs.

3.4.6 Final Recommendation Training Set and Verification

Top $optimal\_n\_train$ points with highest selection frequency was selected as the final training set. Coverage analysis was done on the final training subset and rest of the data set was used for the testing ensuring both representativeness and efficient learning of the models. For the final validation of the process, the models were retrained with the training subset and generalisation was checked with the testing data and their RMSE and R square was calculated to compare their original performances against performance with small training data set.

4 Result and Discussion

4.1 FEA Dataset

As expected, the value for young's modulus increased as the strut thickness increases as the volume fraction increased and thus improved the load transferability as opposed to strut length where the young's modulus decreased as the length increased as the flexibility in the strut structure increased. This direct correlation suggests that the mechanical behaviour of the lattice structure is heavily dependent on the geometry dependent relative density. From the figure, the trend surface is slightly non-planar, smooth but small curvature noticed, justifying the investigation on both nonlinear (GPR, SVR, PR) and linear regression (PR) models. These data points formed on the basis of training and testing models.

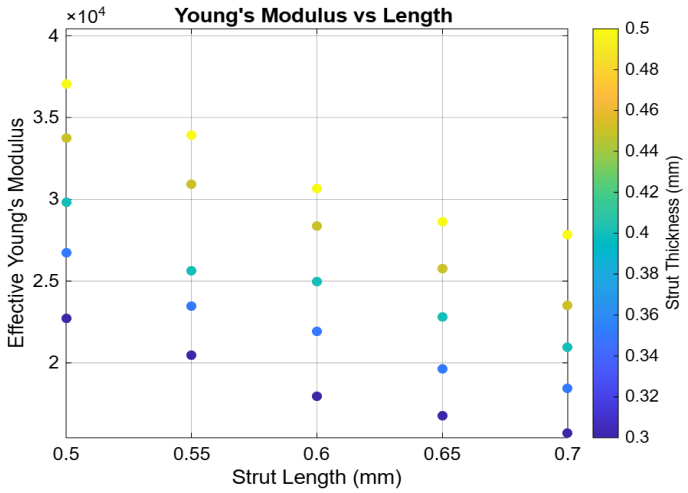

Figure 8 Illustrates how effective young's modulus changes when strut length is kept constant. Young's Modulus decreases with the increase of strut length.

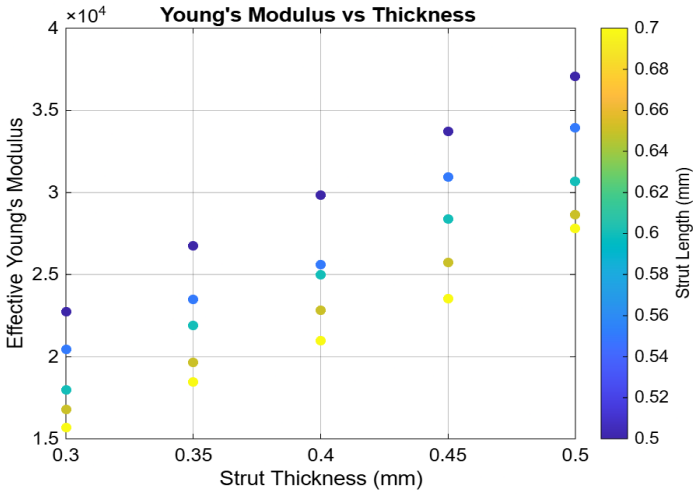

Figure 9 Illustrates how effective young's modulus changes when strut thickness is kept constant. Young's Modulus increases with the increase of strut thickness.

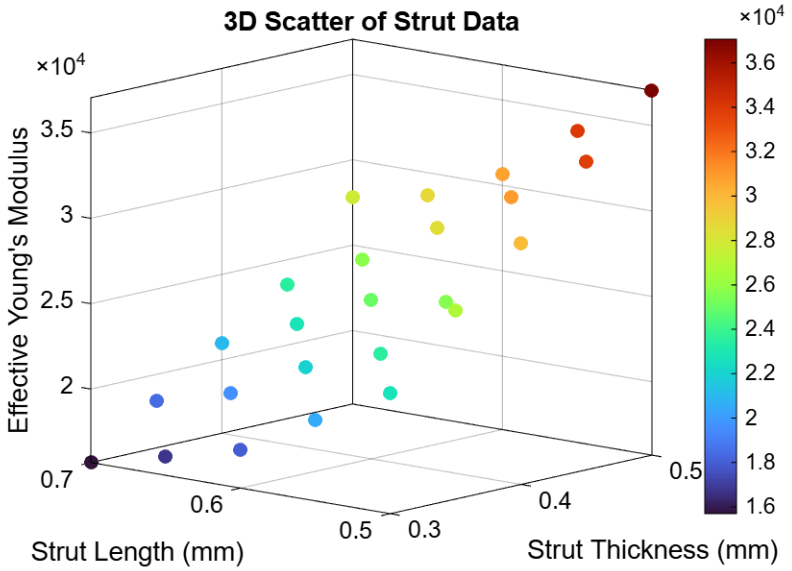

Figure 10 Shows there is an tendency of linear correlation between Young's Modulus and strut inputs but at outer strut configurations the trend is not linear anymore.

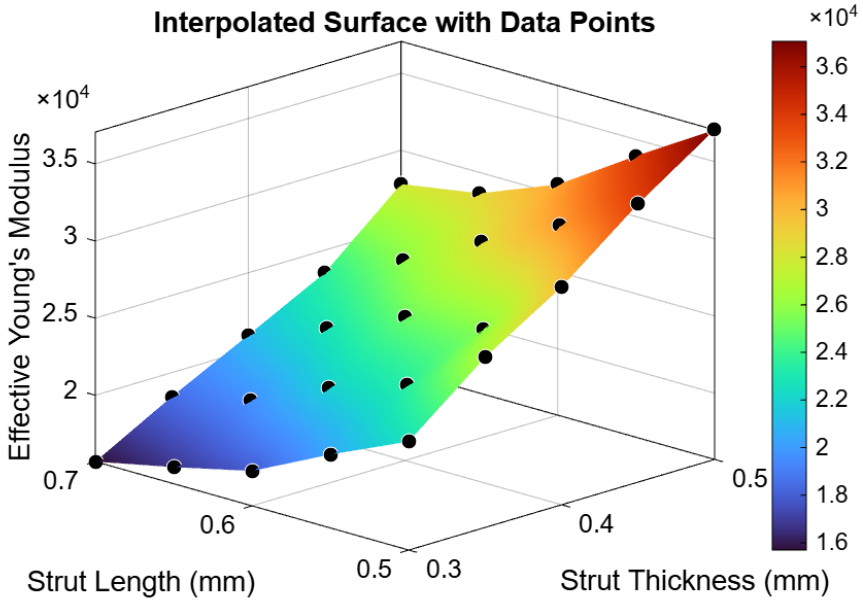

Figure 11 Surface Plot of the data points.Slightly planar relationship but small bumps can be visible in the relationship between Young's Modulus and strut inputs

4.2 Optimal Number of Learning Points (L.P.)

Recommended optimal number of training points needed to train the regression models which were computed by the MATLAB was displayed in the comment section along the mean, median, mode and coefficient of variation. This automated detection was verified visually using the plots and graphs generated from MATLAB.

4.2.1 Optimal Number of Linear Regression (LR)

Result from MATLAB
Recommended optimal training size:9

Supporting Statistics:
- Aggregated curve optimal: 9 points
- Median across 20 seeds: 11 points
- Mean ± SD: 10.9 ± 3.5 points
- Coefficient of variation: 31.56% (Moderate variability)

Performance at Recommended Size (n=9):
- RMSE: 894.20 ± 81.26
- R^2: 0.9722

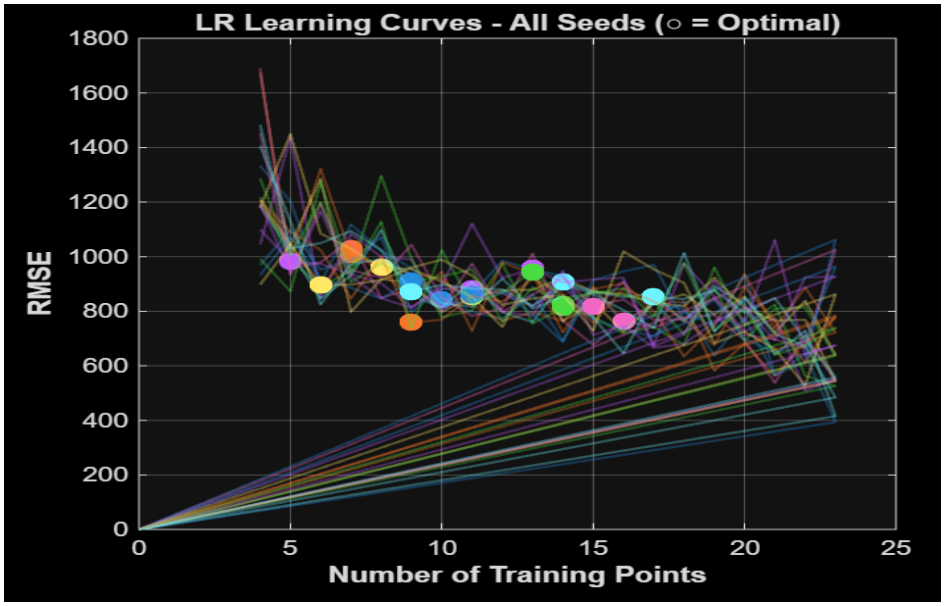

Figure 12 RMSE vs Number of training size graphs across 20 seeds of LR

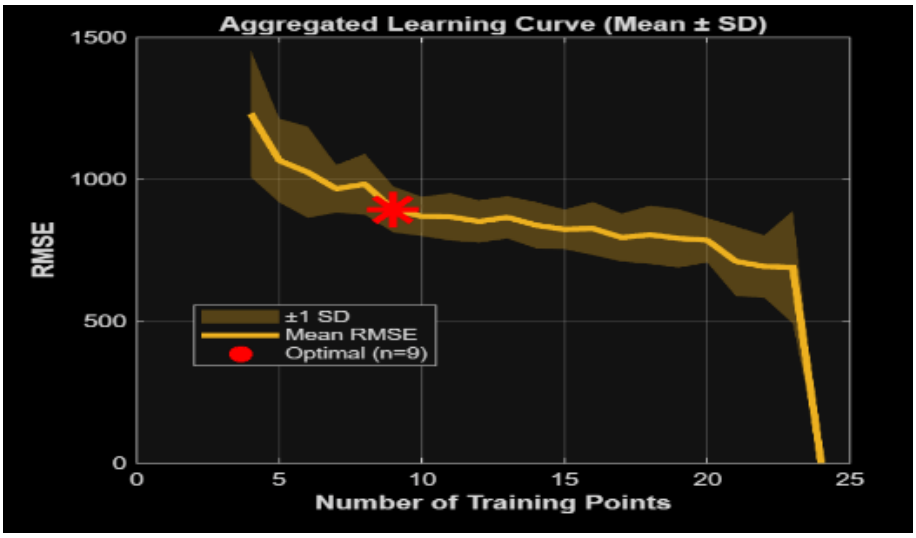

Figure 13 Aggregated RMSE vs Number of Training Size of LR

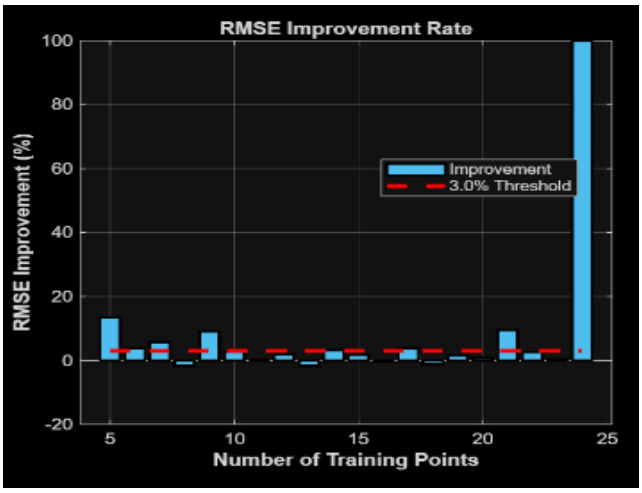

Figure 14 RMSE improvement rate across training size for LR

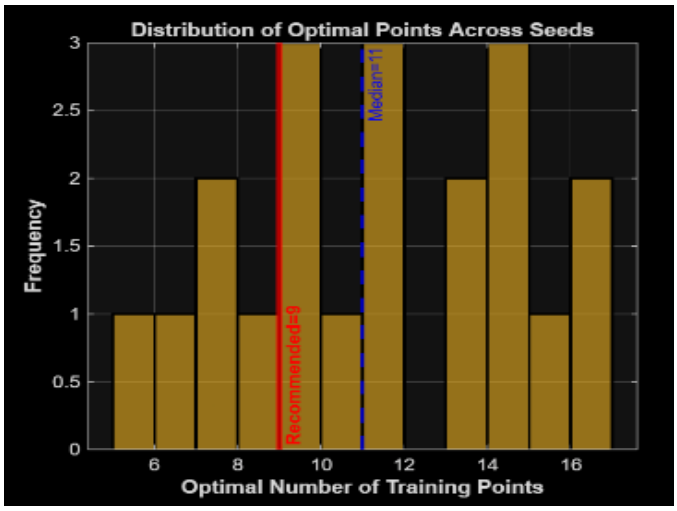

Figure 15 Frequency of a number chosen by LR across 20 seeds.

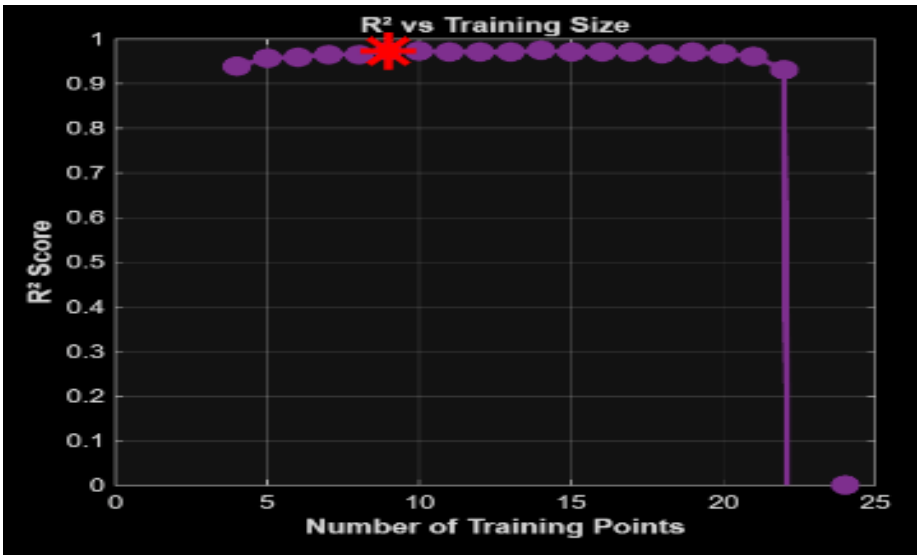

Figure 16 The graphs shows how R^2 of LR changes with the number of

There is an early saturation of learning performance in the figure. The curved across different seeds behaved in a similar manner with the presence of mild variations explaining the CV value. The aggregated curved nearly flattened near 9 points meaning the RMSE improvement rate slowed in this region but the standard deviation range of LR is broader than of GPR and PR. The R^2 value stabilised at around 0.9722 after the 9 training points suggesting that the predictability of LR did not improve much from 9 points. The histogram showed highest frequency in 3 training sizes including 9 being the lowest out of the 3 optimal numbers. After the training size 9, the RMSE improvement rate remained under the threshold as training size increased shown in figure 10.

Overall, LR captured the general with minimal training data but lacked depth of predictability at higher strut inputs and had noticeable variation, resulting in moderate accuracy of the trained model.

4.2.2 Optimal Number of Polynomial Regression (PR)

Result for MATLAB

RECOMMENDED OPTIMAL TRAINING SIZE: 12 points

Supporting Statistics:
- Aggregated curve optimal: 12 points
- Median across 20 seeds: 13 points
- Mean ± SD: 13.8 ± 3.3 points
- Coefficient of variation: 23.57% (Moderate variability)

Performance at Recommended Size (n=12):
- RMSE: 647.27 ± 82.82
- R^2: 0.9832

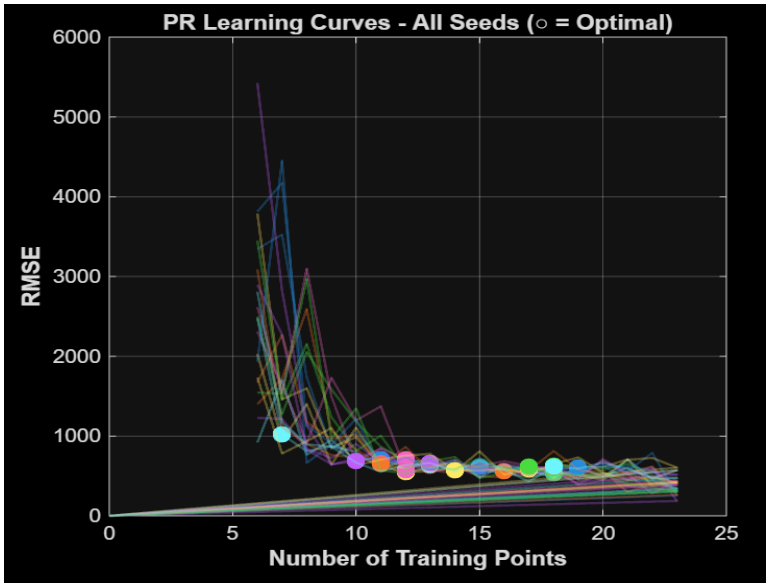

Figure 17 RMSE vs Number of training size graphs across 20 seeds of PR

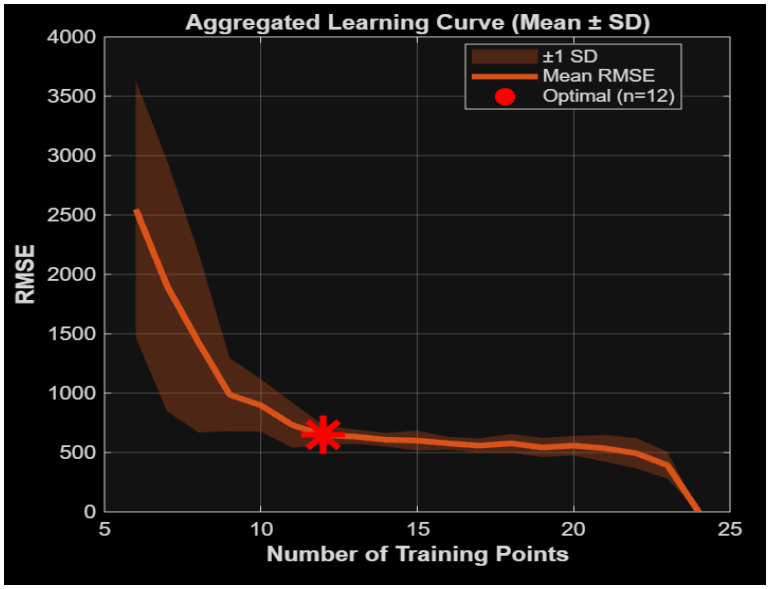

Figure 18 Aggregated RMSE vs Number of Training Size of PR

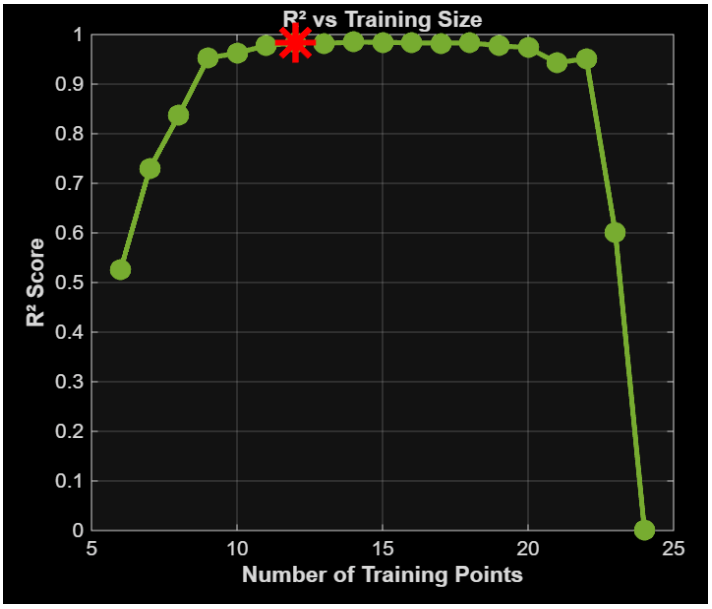

Figure 19 The graphs shows how $R^2$of PR changes with the number of training points

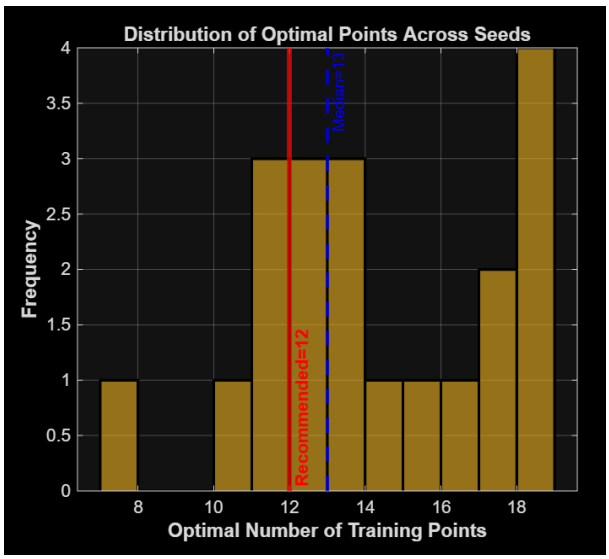

Figure 21 Frequency of a number chosen by PR across 20 seeds.

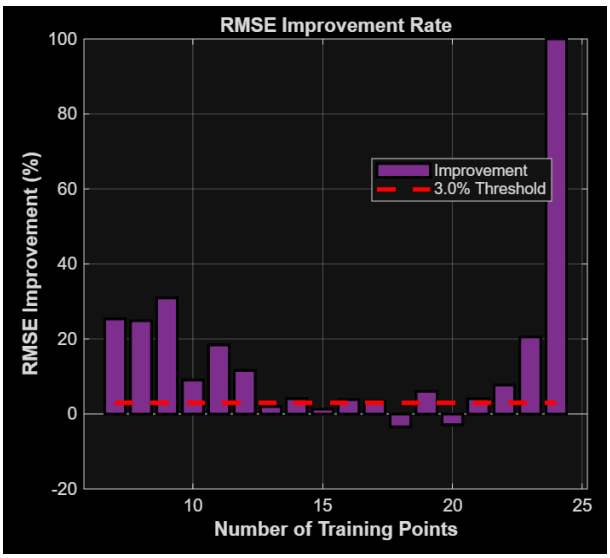

Figure 20 RMSE improvement rate across training size for PR

In figure 13, RMSE value fell sharply and started facing stability after 10 points across all 20 seeds. The variation across the seeds became visibly insignificant from training size 12. The aggregated curve behaved in a similar manner with the range standard deviation becoming the thinnest after point 12 and flattening from that point showcasing the reduction in error being less than 3%. The R^2 in figure 15 improved steadily and reached highest among all models before becoming flat line. The improvement rate of RMSE fell under the threshold line from the training data set 12. The histogram cantered around 12 which is a very stable model. In

the accurate vs predicted graph, the points closely followed the diagonal. With RMSE at its optimal training being the lowest among all models and CV being around 23%, it tells us that PR has high accuracy and robustness.

4.2.3 Optimal Number of Gaussian Process Regression (GPR)

Result from MATLAB

RECOMMENDED OPTIMAL TRAINING SIZE: 13 points

Supporting Statistics:
- Aggregated curve optimal: 13 points
- Median across 20 seeds: 14 points
- Mean ± SD: 13.4 ± 4.2 points
- Coefficient of variation: 31.61% (Moderate variability)

Performance at Recommended Size (n=13):
- RMSE: 655.90 ± 53.76
- R^2: 0.9831

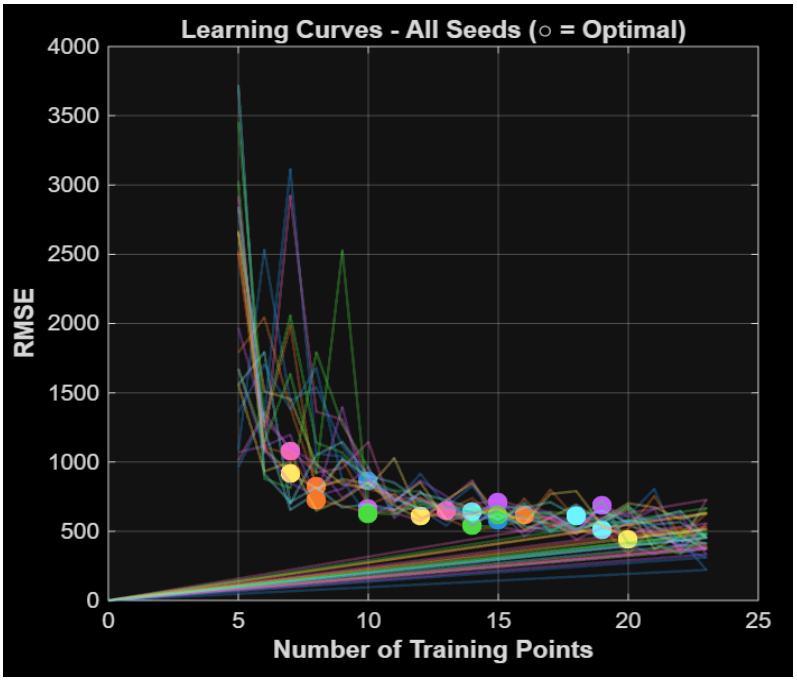

Figure 22 RMSE vs Number of training size graphs across 20 seeds of GPR

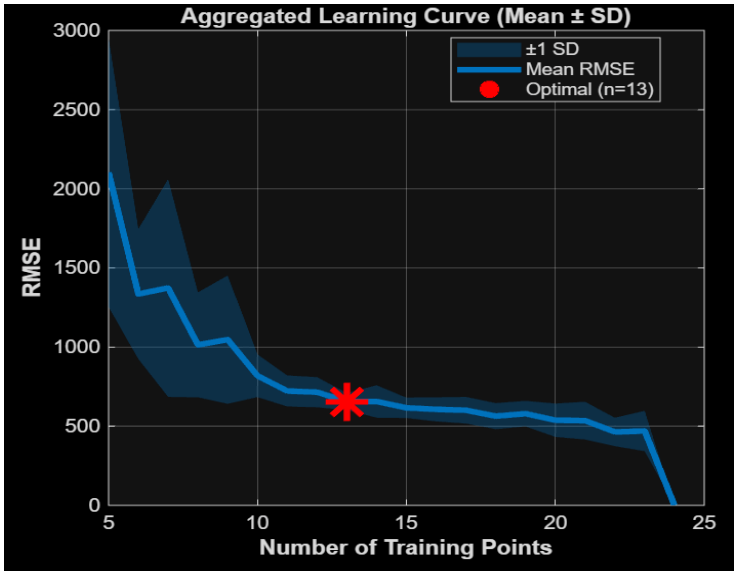

Figure 23 Aggregated RMSE vs Number of Training Size of GPR

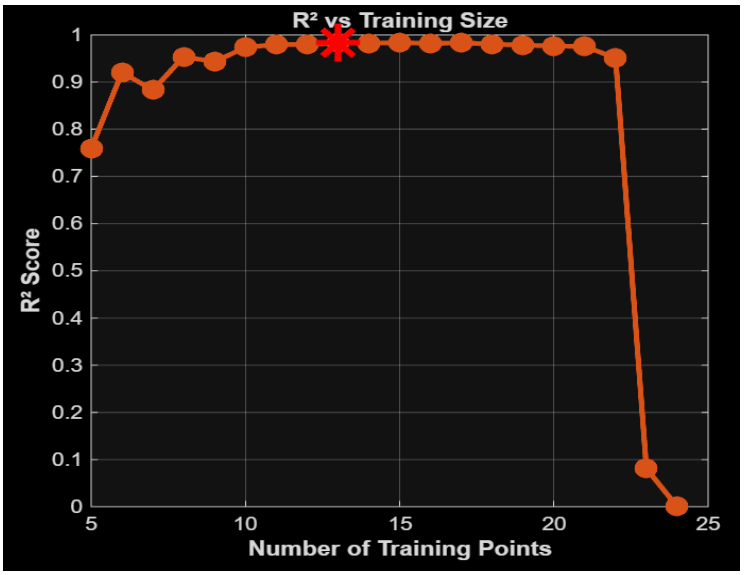

Figure 24 The graphs shows how R square of GPR changes with the number of training points

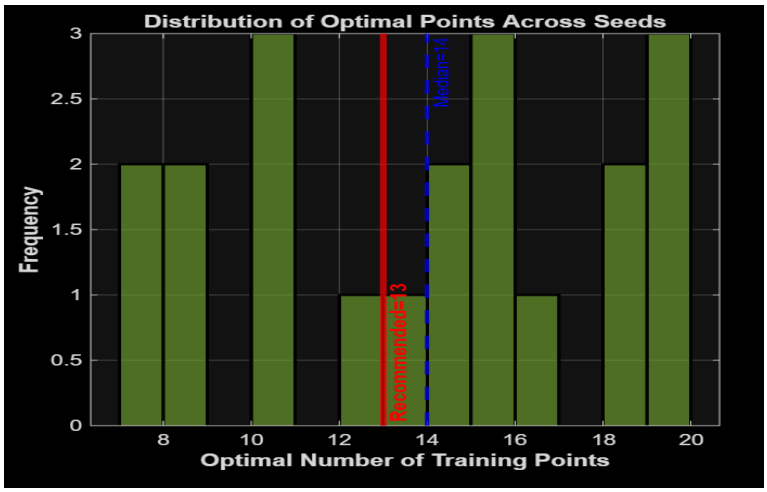

Figure 25 Frequency of a number chosen by GPR across 20 seeds

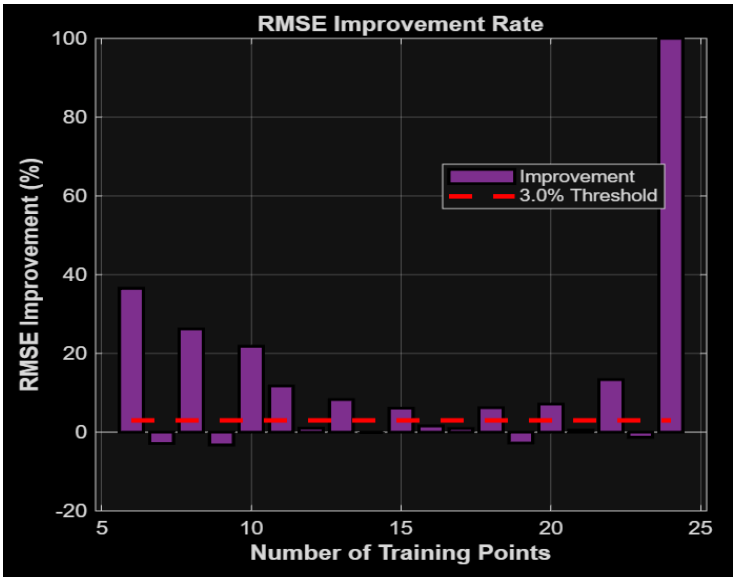

Figure 26 RMSE improvement rate across training size for GPR

The RMSE for GPR across all seeds decreased sharply till the training 10 and then started to face steadiness from point 13 with slightly more variations than PR but comparative less than LR. Also, even though the selected optimal number faced a range across 20 seeds, point 13 had a denser amount of selected optimal number. Similarly, the standard deviation thickness can be comprehended as slightly broader than PR but lesser than LR and SVR. The aggregated curve stabilised after 13 points as well as the R^2 curve. RMSE improvement rate became closer and under the threshold from training size 13 and onward. Considering the RMSE at optimal number of learning points and its CV, it can be determined that GPR's accuracy is good but has a higher variability than PR.

4.2.4 Optimal Number of Support Vector Regression (SVR)

Result from MATLAB

RECOMMENDED OPTIMAL TRAINING SIZE: 12 points

Supporting Statistics:
- Aggregated curve optimal: 12 points
- Median across 20 seeds: 10 points
- Mean ± SD: 10.5 ± 4.5 points
- Coefficient of variation: 42.98% (Moderate variability)

Performance at Recommended Size (n=12):
- RMSE: 2922.95 ± 605.46
- R^2: 0.6511

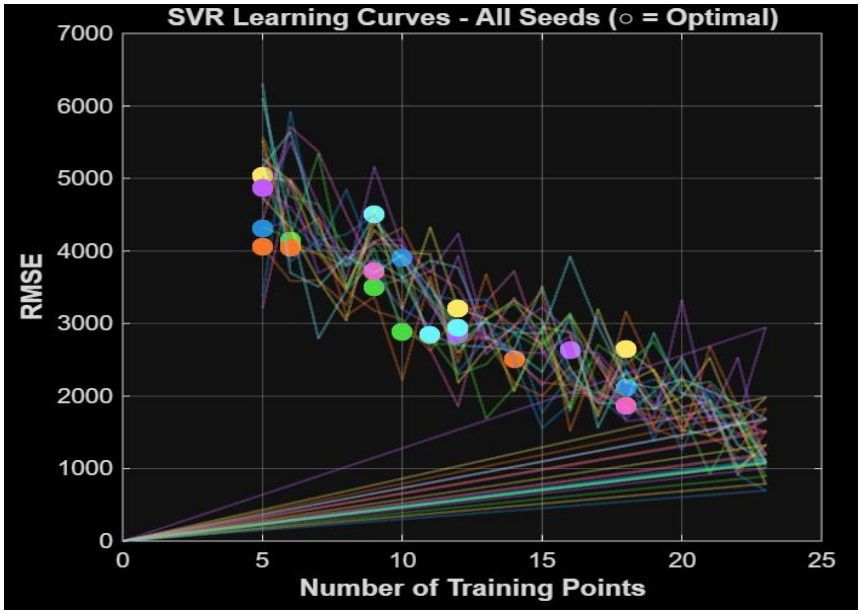

Figure 27 RMSE vs Number of training size graphs across 20 seeds of SVR

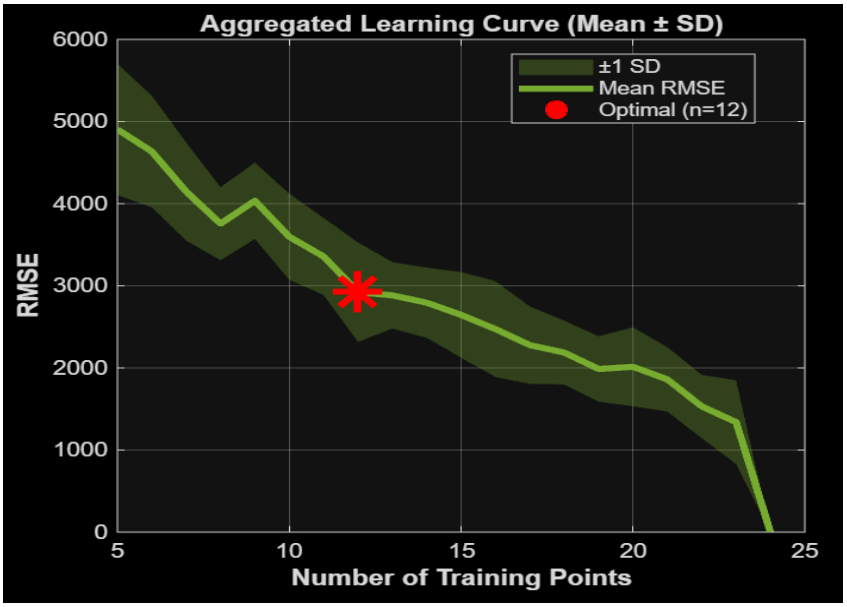

Figure 28 Aggregated RMSE vs Number of Training Size of SVR

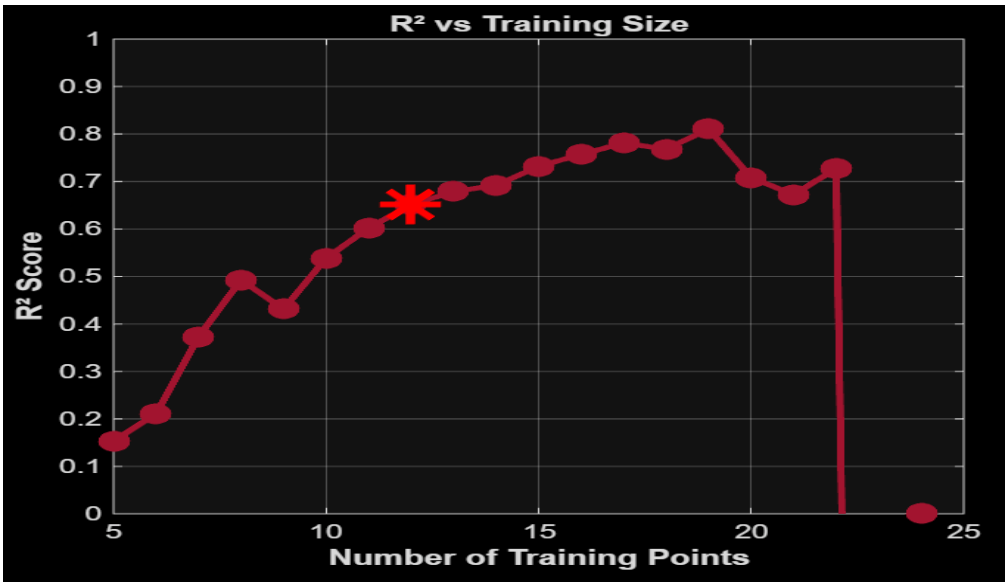

Figure 29 The graphs shows how R^2 of SVR changes with the number of training points

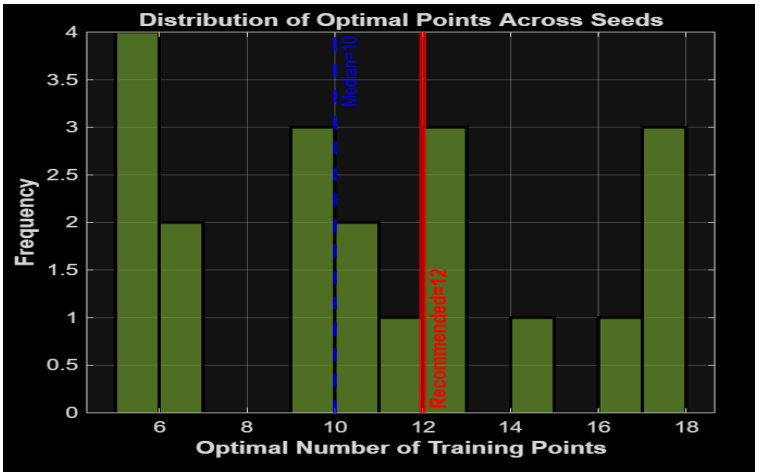

Figure 30 Frequency of a number chosen by SVR across 20 seeds

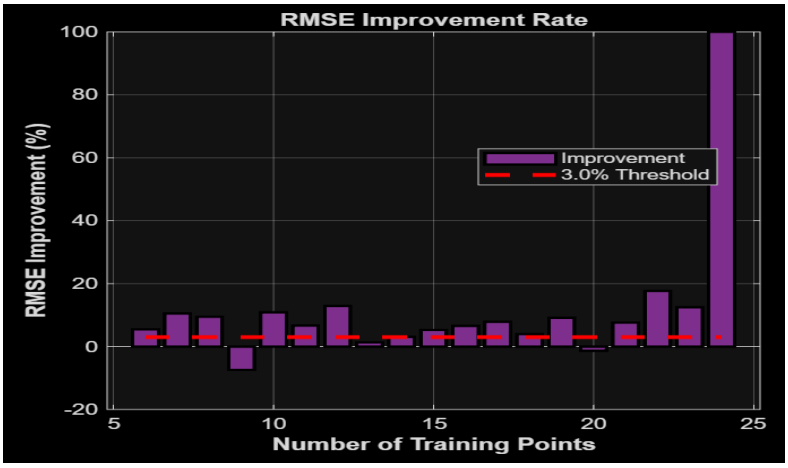

Figure 31 RMSE improvement rate across training size for SVR

The RMSE curve never really stabilised across all the 20 seeds for SVR and MATLAB chose optimal point based where the RMSE change was lowest. The learning curve varied widely across all seeds explaining the highest CV. The aggregated curve shows the RMSE decreased gradually and followed a consistent decreasing rate, never reaching a plateau. The R^2 value also did not flatten and rather curved upward. Except for 2 optimal numbers which are nor consecutive the improvement rate did not get below 3%. The RMSE at optimal number of training point was the highest, the predictive accuracy is lowest among all the models trained.

4.3 Location Of Optimal Number of L.P.

4.3.1 Location of Training Points Selected by LR

MATLAB result:

Recommended training indices based on multi-seed analysis (LR)

$$train\_indices = [1\ 4\ 5\ 8\ 13\ 14\ 17\ 20\ 23]$$

--- Point Categories ---
Always selected (100%): 0 points
Often selected (≥60%): 2 points
Sometimes selected (<60%): 23 points
Never selected: 0 points

PERFORMANCE ACROSS SEEDS

RMSE Statistics:
Mean: 881.34
Median: 879.77
Std: 18.28
Range: 853.49 to 921.60
CV: 2.07%

--- Coverage Analysis ---
Strut Length Coverage: 100.0% (0.50 to 0.70)
Strut Thickness Coverage: 100.0% (0.30 to 0.50)

--- Verification with Linear Regression Model --- Test RMSE: 820.72, Test R^2: 0.9803

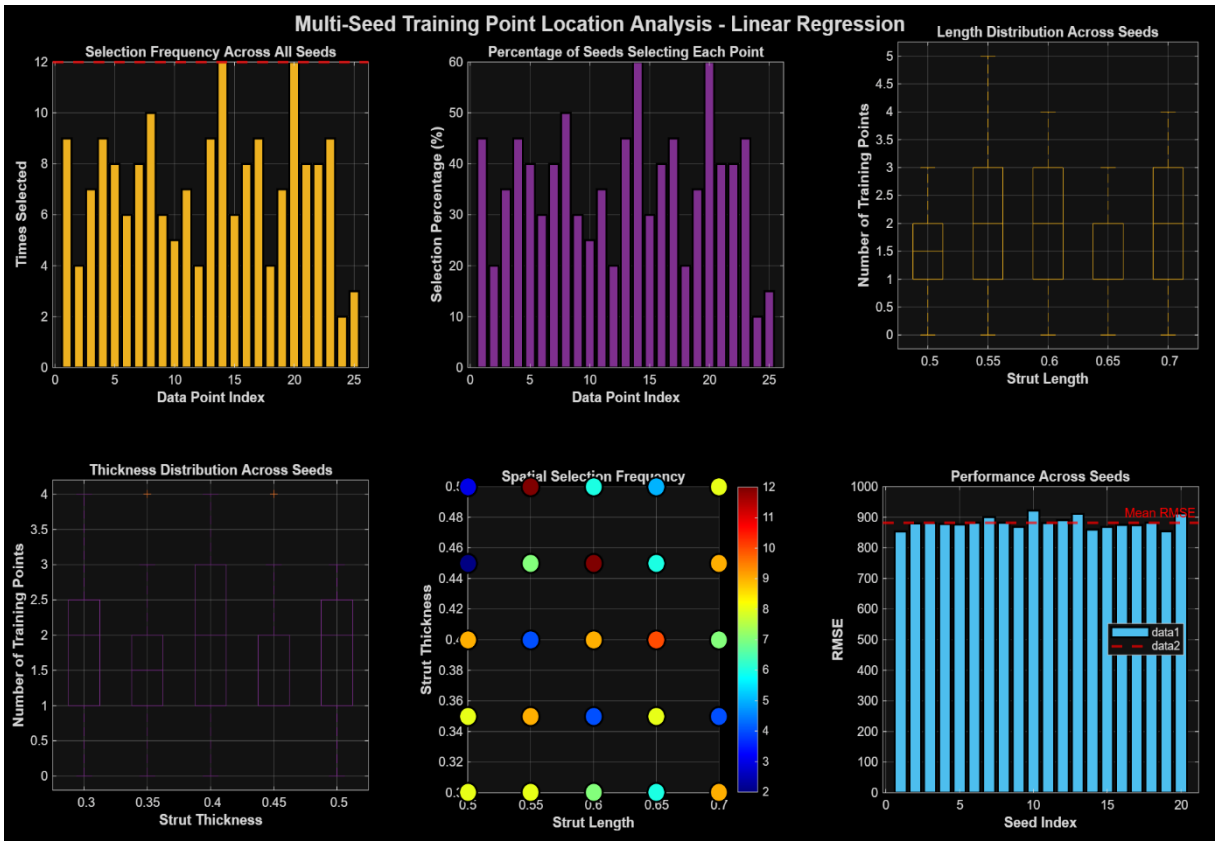

Figure 32 A compilation of 6 figures demonstrating how the location of the learning points affected the accuracy of the model's predictability and based on that which points were selected by LR

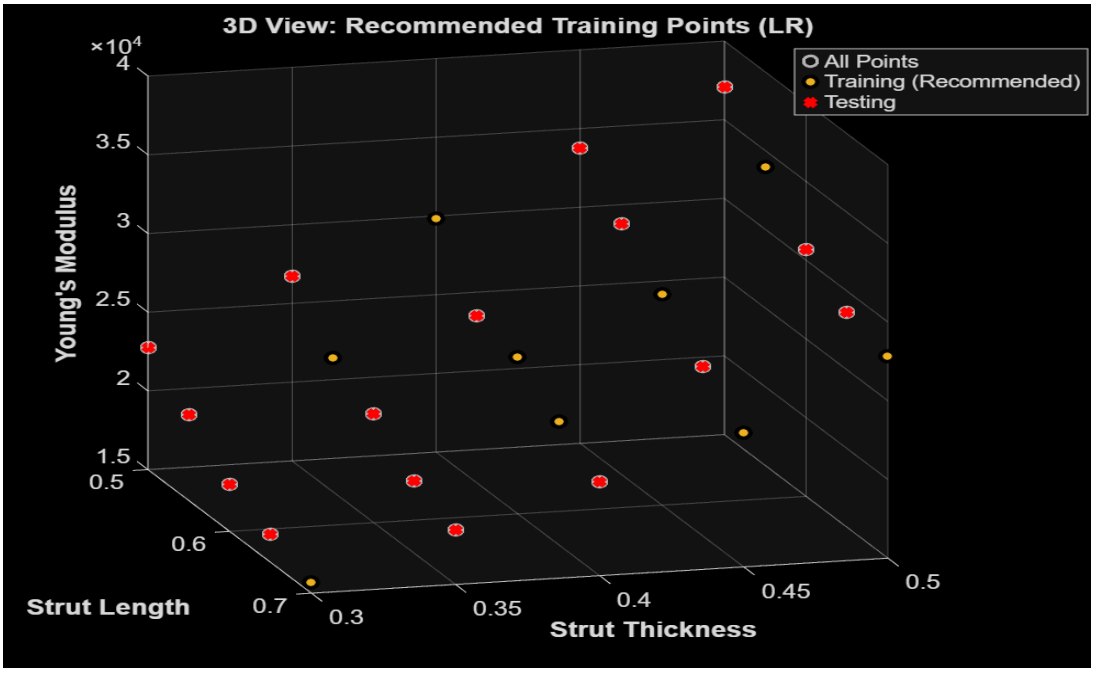

Figure 33 Distribution of the recommended learning points for LR

Selection frequency across all seeds showed 2 points exceeding 60% frequency threshold, 6 points crossing 40% threshold and 5 points touching the 40% selection frequency. The length and thickness distribution plots shows that it has moderate coverage of the whole input space but shows a slight bias towards the mid range geometry. 5 points were selected from the outer line of input space and 4 selected are clustered in the centre. Performance across seeds shows that the RMSE across seeds remains relatively stable proving the solution's robustness. The arrangement of the selected points formed a nearly planar cluster reflecting the linear nature of the model.

4.3.2 Location of Training Points Selected by PR

MATLAB result

Recommended training indices based on multi-seed analysis (PR)

$$train\_indices = [3\ 5\ 7\ 9\ 10\ 17\ 19\ 20\ 21\ 22\ 23\ 24]$$

--- Point Categories ---
Always selected (100%): 0 points
Often selected (≥60%): 6 points
Sometimes selected (<60%): 19 points
Never selected: 0 points

PERFORMANCE ACROSS SEEDSRMSE Statistics:
Mean: 644.40
Median: 641.23
Std: 24.85
Range: 596.14 to 715.12
CV: 3.86%--- Coverage Analysis ---
Strut Length Coverage: 100.0% (0.50 to 0.70)
Strut Thickness Coverage: 100.0% (0.30 to 0.50)

--- Verification with Polynomial Regression Model ---
Test RMSE: 643.30
Test R^2: 0.9878

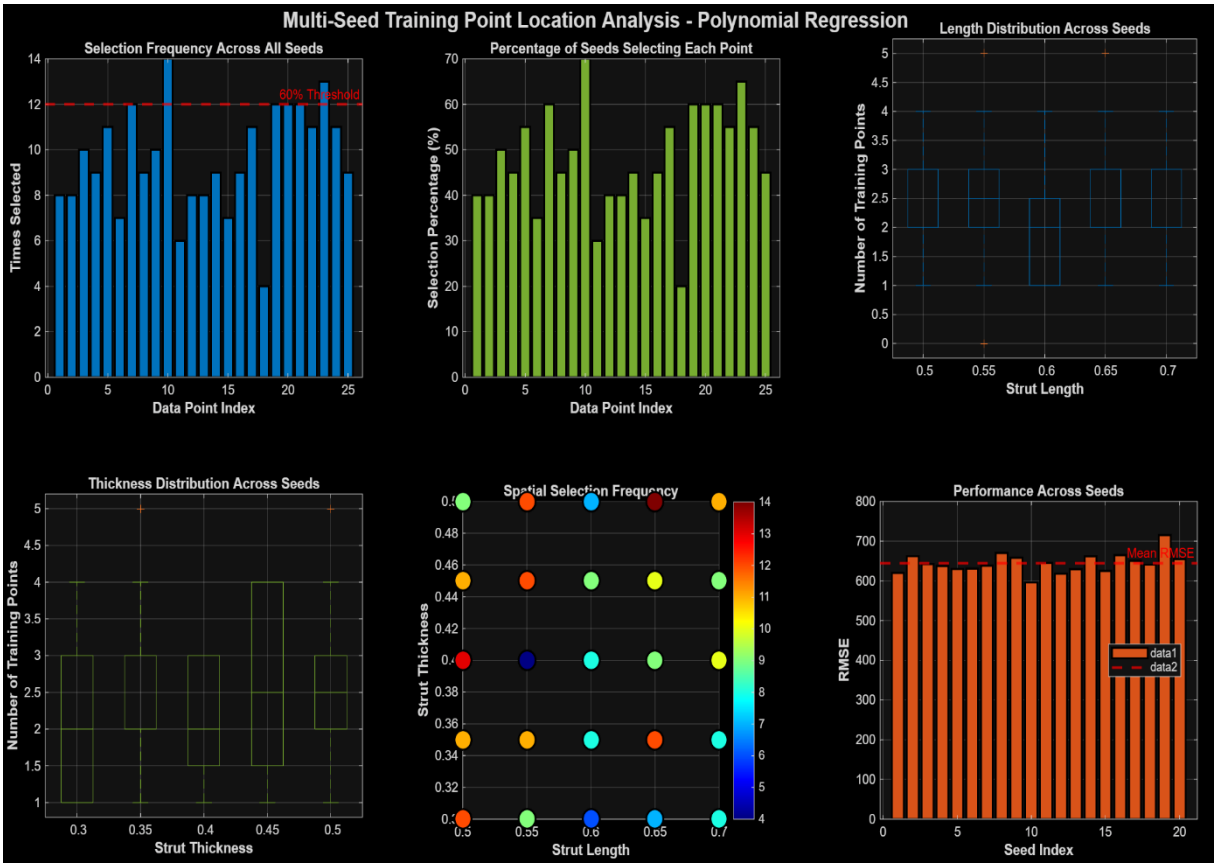

Figure 34 A compilation of 6 figures demonstrating how the location of the learning points affected the accuracy of the model's predictability and based on that which points were selected by PR

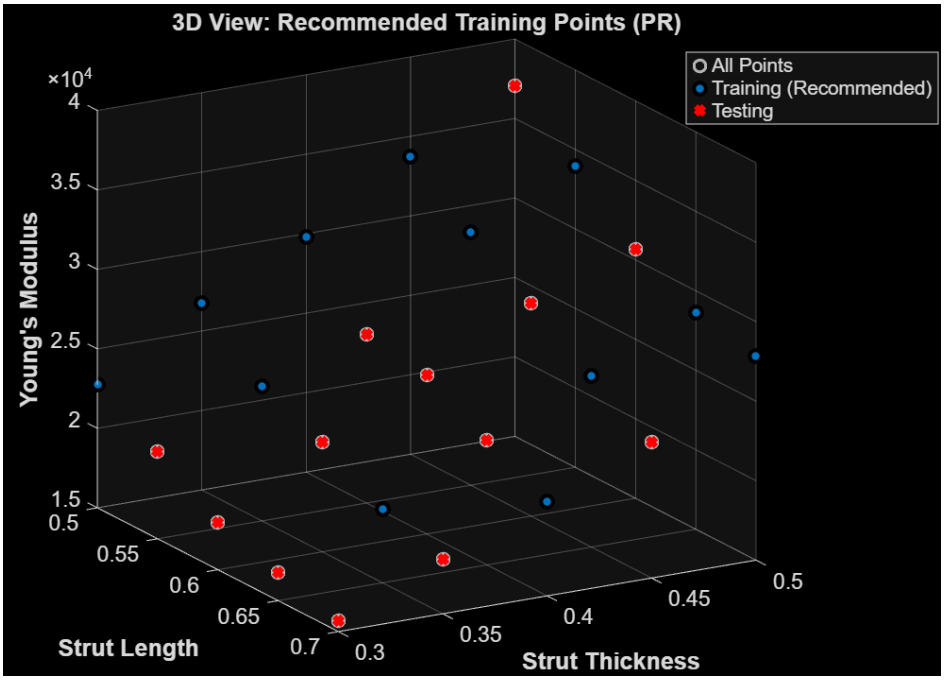

Figure 35 Distribution of the recommended learning points for PR

Length distribution and thickness distribution histogram shows that were more uniform than of LR, showing that the model learns from both higher and lower input space. The spatial selection frequency shows several clusters across the design space meaning it predicts nonlinear relationship well. RMSE remained consistently low across the seeds, indicating robust, accurate learning. The selected points span across the entire stiffness range

4.3.3 Location of Training Points Selected by GPR

MATLAB result

Recommended training indices based on multi-seed analysis

$$train\_indices = [1\ 2\ 3\ 5\ 7\ 9\ 11\ 13\ 16\ 19\ 20\ 22\ 23]$$

--- Point Categories ---
Always selected (100%): 0 points
Often selected (≥60%): 9 points
Sometimes selected (<60%): 16 points
Never selected: 0 points

Performance across seeds
RMSE Statistics:
Mean: 647.98
Median: 651.66
Std: 23.63
Range: 588.45 to 692.97
CV: 3.65%

Coverage Analysis ---
Strut Length Coverage: 100.0% (0.50 to 0.70)
Strut Thickness Coverage: 100.0% (0.30 to 0.50)

--- Verification with GPR Model ---
Test RMSE: 694.26
Test R^2: 0.9835

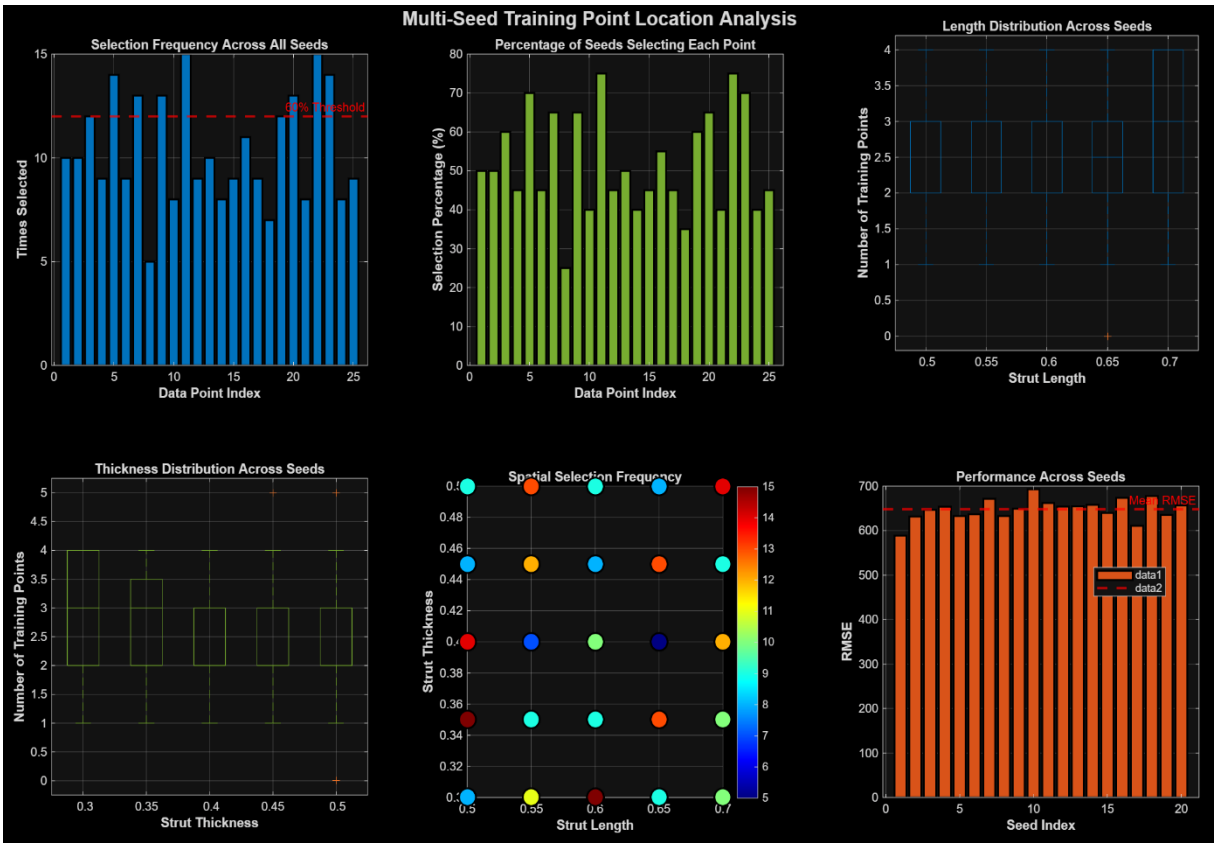

Figure 36 A compilation of 6 figures demonstrating how the location of the learning points affected the accuracy of the model's predictability and based on that which points were selected by GPR

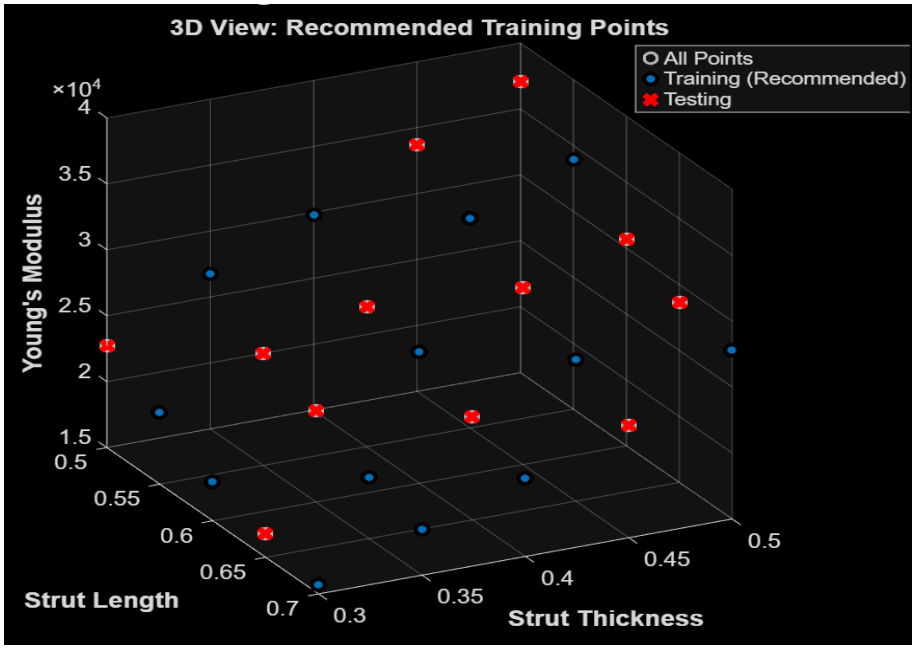

Figure 37 Distribution of the recommended learning points for GPR

Like PR, the points selected showed high frequency distributed evenly over the design space. This shows that GPR utilises information at both small and large geometries. The length and thickness distribution showed that GPR has good generalisation across the domain sampling relatively uniform data. GPR as well good robustness as it has small RMSE changes across the 20 different seeds. The 3d view showed evenly spread of selected learning points.

4.3.4 SVR

MATLAV result

Recommended training indices based on multi-seed analysis (SVR)

$$train\_indices = [3\ 5\ 6\ 8\ 9\ 11\ 13\ 14\ 18\ 20\ 23\ 24]$$

--- Point Categories ---
Always selected (100%): 0 points
Often selected (≥60%): 6 points
Sometimes selected (<60%): 19 points
Never selected: 0 points

PERFORMANCE ACROSS SEEDS

RMSE Statistics:
Mean: 2935.12
Median: 2949.68
Std: 187.76
Range: 2504.40 to 3208.74
CV: 6.40%

Coverage Analysis ---
Strut Length Coverage: 100.0% (0.50 to 0.70)
Strut Thickness Coverage: 100.0% (0.30 to 0.50)

--- Verification with SVR Model ---
Test RMSE: 2590.20
Test R^2: 0.7948

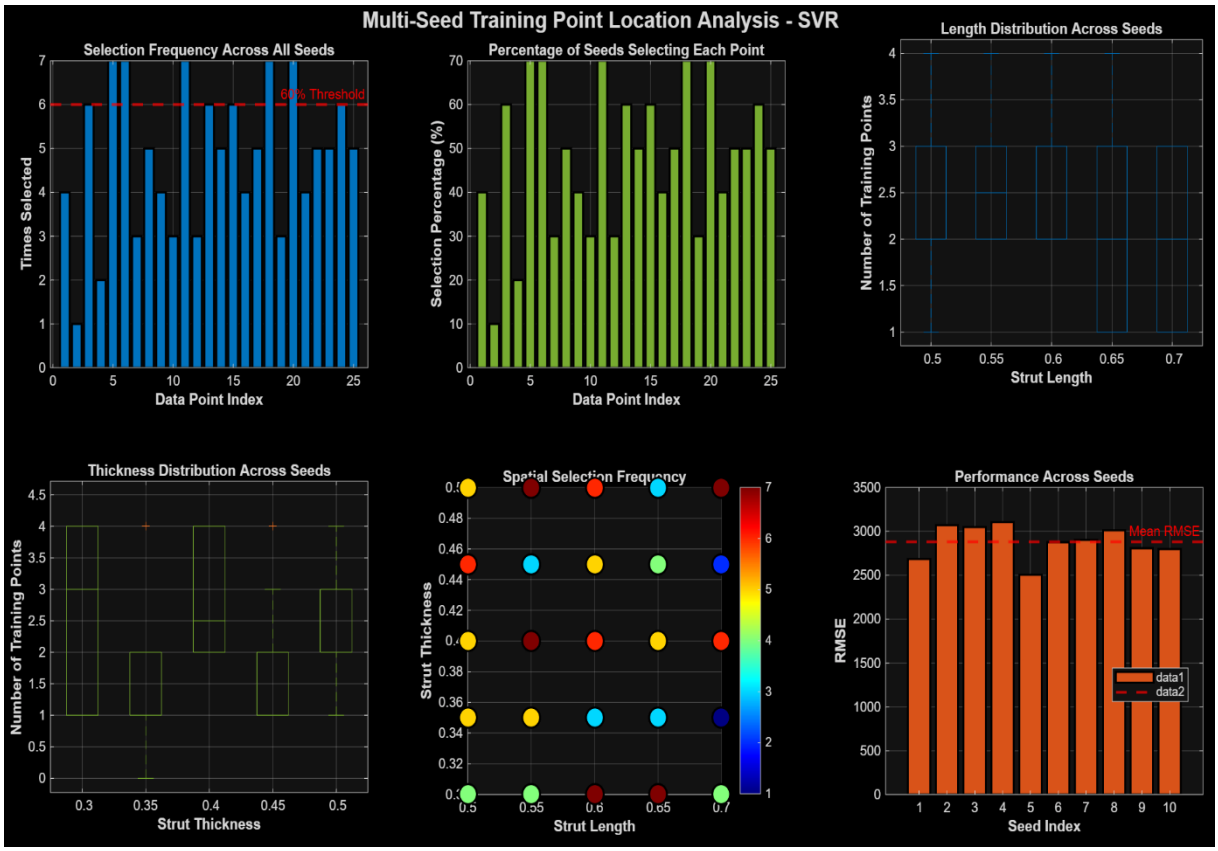

Figure 38 A compilation of 6 figures demonstrating how the location of the learning points affected the accuracy of the model's predictability and based on that which points were selected by SVR

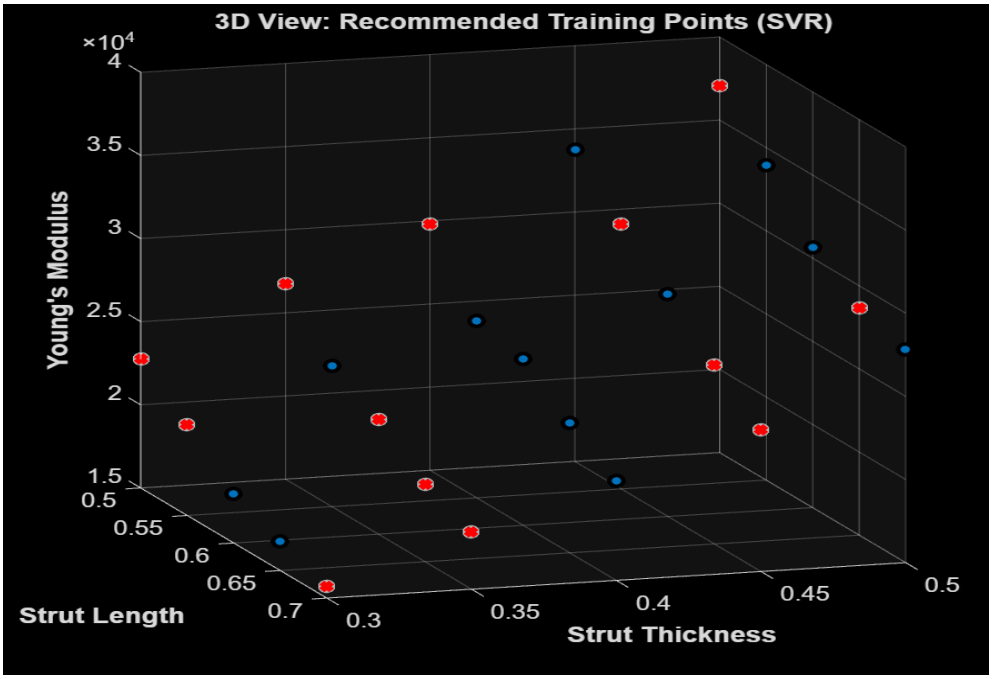

Figure 39 Distribution of the recommended learning points for SVR

Even though several points cross the 60% threshold, there were uneven representation of data. Length and thickness distribution shows that SVR is sensitive to data position and depends critically on boundary points. The spatial selection frequency was scattered indicating inconsistency at recognising data points for learning. The performance across seeds shows noticeable variations, indicating the absence of robustness.

5 Consequences

5.1.1 Discussion of Findings

The FEA process successfully demonstrated the relationship between Strut inputs (Thickness, strut length) and the effective young's modulus of the 3d infill lattice structures. Out of all the models (LR, PR, GPR, SVR) trained using these data, PR and GPR performed the best with highest accuracy, robustness and accuracy, with GPR having slightly less performance values. LR performed well, although less accurate, as the maximum part of the design space of strut input and young's modulus had mostly linear correlation. SVR had the worst performance, lowest accuracy and highest variability, among all the 4 regression models.

Table 2. Comparison of Optimal Numbers Across Models

Model	**Rec. Optimal no. (RON)**	**Median**	**Mean** ± SD	**RMSE at RON**	R^2 at RON	**CV(%)**	**Remark**
(LR)	9	11	10.9 ± 3.5	894.20 ± 81.26	0.9722	31.56%	Moderate accuracy, moderately unstable
(PR, deg=2)	12	13	13.8 ± 3.3	647.27 ± 82.82	0.9832	23.57%	Best Performer
(GPR)	13	14	13.4 ± 4.2	655.90 ± 53.76	0.9831	31.61%	Good performance but little less than PR
(SVR)	12	10	10.5 ± 4.5	2922.95±605.46	0.6511	42.98%	poor prediction, Unsuitable

Comparison of the optimal number of training points deduced from each models shows that 9 to 13 learning points are needed to train the regression models in the most data efficient and computationally cost saving way. That means about 36% to 52% of the total data set provided was actually needed to train the regression models for the prediction of effective young's modulus of lattice structures. PR had the best performance with the optimal number of training size whereas GPR had the second best performance. Overall, PR captured the predictability in the most data efficient way with lowest variability, making PR the best choice among the four models. The accuracy and data efficiency of GPR is comparable to PR but it faced slight instability making it the 2nd best choice. Even though LR was less accurate, it still showed consistent and reliable prediction for linear trend.

Table 3. Comparison of Selected Optimal Numbers Across Models

Model	Train indices	**Mean RMSE ± SD**	**RMSE**	R^2	**CV**	**Remark**
(LR)	1 4 5 8 13 14 17 20 23	881.34±18.28	820.72	0.9803	2.07%	Stable, Moderate accuracy
(PR, deg=2)	3 5 7 9 10 17 19 20 21 22 23 24	644.40±24.85	643.30	0.9878	3.86%	Most balanced, Low error & stable
(GPR)	1 2 3 5 7 9 11 13 16 19 20 22 23	647.98± 23.63	694.26	0.9835	3.65%	High accuracy and consistency
(SVR)	3 5 6 8 10 13 18 19 20 21 23 24	2944.28±193.25	2623.34	0.7855	6.56%	Least efficient

About the location and distribution of selected learning points, PR and GPR had good accuracy and robustness when trained with the selected data points. LR had lower accuracy but similar stability like PR and GPR. Where SVR had lowest accuracy and stability, proving the invalidity of the selected learning points. All the regression models' selected learning points covered both the points with high young's modulus and points with low young's modulus values. Training indices 5, 20, 23 were selected across all models meaning they have high informational value in training the models about mechanical behaviour of lattice structure. PR focused on indices from 19 to 24 where the stiffness increased sharply with strut input variation. GPR behaved the similar way. LR did not focus on any particular region.

5.1.2 Practical Implications

The result investigates on the performance, data efficiency, computational cost, reduction of data redundancy of the four regression machine learning models. PR and GPR shows that the compressive behaviour of the Ti6Al4V lattice structures can be accurately predicted with the help of FEA training data set in a much more data and computation efficient way. This can enable faster data generation and contribute significantly where data generation is costly. Moreover, the findings also tell us that the combination of FEA generated training dataset and machine learning models for the prediction of other materials and structures can have significant progress in the world of biomedical implants.

6 Conclusion

6.1.1 Summary

The study investigated the compressive behaviour of the Ti6Al4V lattice structures through combining finite element analysis (FEA) with machine learning models. Effective young's modulus, an indicator of mechanical performance, and strut geometries (strut length and strut thickness) had showed clear relationship through the study.

Among the regression machine learning models, Polynomial Regression (degree=2) showed best performance overall. It had the highest accuracy,lowest error and most stability. Gaussian Process Regression showed high accuracy and consistency coming 2nd to PR. Linear Regression gave stable but low accuracy results. Support Vector Regression model was found to be with the most unaccuracy and high variability. Thus proving it self to be the least suitable.

The results confirm that only 36% - 52% of the total dataset is required to achieve the reliable prediction showed in the study. Tis significantly reduces the dataset and lowers the computational cost. The study illustrates the effectiveness of combining machine learning with the FEA derived data for the efficient modelling of lattice structures.

6.1.2 Future Work

Future studies can investigate more into the kernel function of SVR and analyse how can SVR's reliability and accuracy be improved under small sample condition and whether the accuracy was low because of a slight linear pattern visible in the dataset.

Also, future studies can expand the data set with more geometric inputs like lattice topology and unit cell type to further improve the models' generalisation. In addition, integrating machine learning and artificial intelligence can fully eliminate the need for repeated FEA data generation.

For the full MATLAB code: https://github.com/MalihaKuchu/RegressionModelsMatlab

List of references/Bibliography

[1] E. Marin and A. Lanzutti, "Biomedical Applications of Titanium Alloys: A Comprehensive Review," Jan. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/ma17010114.

[2] Y. W. Cui, L. Wang, and L. C. Zhang, "Towards load-bearing biomedical titanium-based alloys: From essential requirements to future developments," Nov. 2024, *Elsevier Ltd*. doi: 10.1016/j.pmatsci.2024.101277.

[3] R. Alkentar, F. Máté, and T. Mankovits, "Investigation of the Performance of Ti6Al4V Lattice Structures Designed for Biomedical Implants Using the Finite Element Method," *Materials*, vol. 15, no. 18, Nov. 2022, doi: 10.3390/ma15186335.

[4] U. O. Agwu, K. Wang, C. Singh, C. Leemhuis, S. Yamakawa, and K. Shimada, "Assessing Tetrahedral Lattice Parameters for Engineering Applications through Finite Element Analysis," *3D Print Addit Manuf*, vol. 8, no. 4, pp. 238–252, Nov. 2021, doi: 10.1089/3dp.2020.0222.

[5] J. Yang, D. Yang, Y. Tao, and J. Shi, "Machine learning assisted prediction and analysis of in-plane elastic modulus of hybrid hierarchical square honeycombs," *Thin-Walled Structures*, vol. 198, Nov. 2024, doi: 10.1016/j.tws.2024.111736.

[6] J. Bai, M. Li, and J. Shen, "Prediction of Mechanical Properties of Lattice Structures: An Application of Artificial Neural Networks Algorithms," *Materials*, vol. 17, no. 17, Nov. 2024, doi: 10.3390/ma17174222.

[7] P. Xu, X. Ji, M. Li, and W. Lu, "Small data machine learning in materials science," Nov. 2023, *Nature Research*. doi: 10.1038/s41524-023-01000-z.

[8] L. Zhang, B. Song, S. K. Choi, and Y. Shi, "A topology strategy to reduce stress shielding of additively manufactured porous metallic biomaterials," *Int J Mech Sci*, vol. 197, Nov. 2021, doi: 10.1016/j.ijmecsci.2021.106331.

[9] D. Savio and A. Bagno, "When the Total Hip Replacement Fails: A Review on the Stress-Shielding Effect," Nov. 2022, *MDPI*. doi: 10.3390/pr10030612.

[10] O. Bittredge *et al.*, "Fabrication and Optimisation of Ti-6Al-4V Lattice-Structured Total Shoulder Implants Using Laser Additive Manufacturing," *Materials*, vol. 15, no. 9, Nov. 2022, doi: 10.3390/ma15093095.

[11] D. P. Papazoglou, L. Hobbs, Y. Sun, and A. Neidhard-Doll, "In Vitro Proliferation of MG-63 Cells in Additively Manufactured Ti-6Al-4V Biomimetic Lattice Structures with Varying Strut Geometry and Porosity," *Materials*, vol. 17, no. 18, Nov. 2024, doi: 10.3390/ma17184608.

[12] K. Ráž, Z. Chval, and M. Pereira, "Lattice Structures—Mechanical Description with Respect to Additive Manufacturing," *Materials*, vol. 17, no. 21, Nov. 2024, doi: 10.3390/ma17215298.

[13] M. H. K. Aljaberi, M. M. Aghdam, T. Goudarzi, and M. Al-Waily, "Compressive Behavior of Novel Additively Manufactured Ti-6Al-4V Lattice Structures:

Experimental and Numerical Studies," *Materials*, vol. 17, no. 15, Nov. 2024, doi: 10.3390/ma17153691.

[14] N. Koju, S. Niraula, and B. Fotovvati, "Additively Manufactured Porous Ti6Al4V for Bone Implants: A Review," Nov. 2022, *MDPI*. doi: 10.3390/met12040687.

[15] H. R. Sichani, M. Atapour, F. Ashrafizadeh, M. Galati, and A. Saboori, "Mechanical, electrochemical and permeability behaviour of Ti6Al–4V scaffolds fabricated by electron beam powder bed fusion for orthopedic implant applications: The role of cell type and cell size," *Journal of Materials Research and Technology*, vol. 28, pp. 3240–3257, Nov. 2024, doi: 10.1016/j.jmrt.2023.12.260.

[16] N. Wang, G. K. Meenashisundaram, D. Kandilya, J. Y. H. Fuh, S. T. Dheen, and A. S. Kumar, "A biomechanical evaluation on Cubic, Octet, and TPMS gyroid Ti6Al4V lattice structures fabricated by selective laser melting and the effects of their debris on human osteoblast-like cells," *Biomaterials Advances*, vol. 137, Nov. 2022, doi: 10.1016/j.bioadv.2022.212829.

[17] K. Monkova, S. Braut, P. P. Monka, A. Skoblar, and M. Pollák, "Numerical and Experimental Modal Analysis of a Gyroid Inconel 718 Structure for Stiffness Specification in the Design of Load-Bearing Components," *Materials*, vol. 17, no. 14, Nov. 2024, doi: 10.3390/ma17143595.

[18] C. P. Karri and V. Kambagowni, "Finite Element Analysis Approach for Optimal Design and Mechanical Performance Prediction of Additive Manufactured Sandwich Lattice Structures," *Journal of The Institution of Engineers (India): Series D*, 2024, doi: 10.1007/s40033-024-00650-7.

[19] A. Alfares, Y. A. Sha'aban, and A. Alhumoud, "Machine learning -driven predictions of lattice constants in ABX3 Perovskite Materials," *Eng Appl Artif Intell*, vol. 141, Nov. 2025, doi: 10.1016/j.engappai.2024.109747.

[20] A. Mishra, "LatticeML: A data-driven application for predicting the effective Young Modulus of high temperature graph based architected materials."

[21] K. Berladir, K. Antosz, V. Ivanov, and Z. Mitaľová, "Machine Learning-Driven Prediction of Composite Materials Properties Based on Experimental Testing Data," *Polymers (Basel)*, vol. 17, no. 5, Nov. 2025, doi: 10.3390/polym17050694.

[22] R. C. F. da Paixão, R. E. K. Penido, A. A. Cury, and J. C. Mendes, "Comparison of machine learning techniques to predict the compressive strength of concrete and considerations on model generalization," *Revista IBRACON de Estruturas e Materiais*, vol. 15, no. 5, 2022, doi: 10.1590/S1983-41952022000500003.

[23] N. Liu, Y. Sun, J. Wang, Z. Wang, A. Rastegarnia, and J. Qajar, "Estimation of static Young's modulus of sandstone types: effective machine learning and statistical models," *Earth Sci Inform*, vol. 17, no. 5, pp. 4339–4359, Nov. 2024, doi: 10.1007/s12145-024-01392-6.

[24] M. J. Hooshmand, C. Sakib-Uz-Zaman, and M. A. H. Khondoker, "Machine Learning Algorithms for Predicting Mechanical Stiffness of Lattice Structure-Based Polymer Foam," *Materials*, vol. 16, no. 22, Nov. 2023, doi: 10.3390/ma16227173.

[25] A. V Tatachar, "Comparative Assessment of Regression Models Based On Model Evaluation Metrics," *International Research Journal of Engineering and Technology*, 2021, [Online]. Available: www.irjet.net

[26] Y. Liu, F. Sun, M. Chen, J. Xiao, J. Li, and B. Wu, "Prediction of Equivalent Elastic Modulus for Metal-Coated Lattice Based on Machine Learning," *Applied Composite Materials*, vol. 30, no. 4, pp. 1207–1229, Nov. 2023, doi: 10.1007/s10443-022-10061-0.

[27] S. Thirupathi, K. U. Reddy, A. C. U. Rao, and P. V. Reddy, "Machine learning-based yield strength prediction in 3D printed Ti6Al4V lattice structures: A combined simulation and experimental approach," *Next Materials*, vol. 9, Nov. 2025, doi: 10.1016/j.nxmate.2025.101190.

[28] C. B. Pande *et al.*, "Forecasting of monthly air quality index and understanding the air pollution in the urban city, India based on machine learning models and cross-validation," *J Atmos Chem*, vol. 82, no. 1, Nov. 2025, doi: 10.1007/s10874-024-09466-x.

[29] D. L. Naik and R. kiran, "A novel sensitivity-based method for feature selection," *J Big Data*, vol. 8, no. 1, Nov. 2021, doi: 10.1186/s40537-021-00515-w.

[30] D. Ibarra-Hoyos, Q. Simmons, and S. J. Poon, "Comparing Machine Learning Models for Strength and Ductility in High-Entropy Alloys," *High Entropy Alloys and Materials*, vol. 3, no. 1, pp. 101–114, Nov. 2025, doi: 10.1007/s44210-024-00049-9.

[31] S. Banik, K. Balasubramanian, S. Manna, S. Derrible, and S. K. Sankaranarayananan, "Machine Learning for Elastic Properties of Materials: A predictive benchmarking study in a domain-segmented feature Space."

[32] G. Liu *et al.*, "Precise multi-factor immediate implant placement decision models based on machine learning," *Sci Rep*, vol. 15, no. 1, Nov. 2025, doi: 10.1038/s41598-025-89814-3.

[33] S. S. Sorour, C. A. Saleh, and M. Shazly, "A review on machine learning implementation for predicting and optimizing the mechanical behaviour of laminated fiber-reinforced polymer composites," Nov. 2024, *Elsevier Ltd*. doi: 10.1016/j.heliyon.2024.e33681.

[34] Q. Tao *et al.*, "Machine learning strategies for small sample size in materials science," Nov. 2025, *Science China Press*. doi: 10.1007/s40843-024-3204-5.

[35] Y. Liu *et al.*, "REVIEW MATERIALS SCIENCE Data quantity governance for machine learning in materials science", doi: 10.1093/nsr/nwad125/7147579.

[36] H. Liu, B. Yucel, B. Ganapathysubramanian, S. R. Kalidindi, D. Wheeler, and O. Wodo, "Active learning for regression of structure-property mapping: the importance of sampling and representation," *Digital Discovery*, vol. 3, no. 10, pp. 1997–2009, Aug. 2024, doi: 10.1039/d4dd00073k.

[37] A. Shmuel, O. Glickman, and T. Lazebnik, "A comprehensive benchmark of machine and deep learning models on structured data for regression and classification," *Neurocomputing*, vol. 655, Nov. 2025, doi: 10.1016/j.neucom.2025.131337.

[38] S. Bishnoi *et al.*, "Predicting Young's modulus of oxide glasses with sparse datasets using machine learning," *J Non Cryst Solids*, vol. 524, Nov. 2019, doi: 10.1016/j.jnoncrysol.2019.119643.

[39] A. Shmuel, O. Glickman, and T. Lazebnik, "A comprehensive benchmark of machine and deep learning models on structured data for regression and classification," *Neurocomputing*, vol. 655, Nov. 2025, doi: 10.1016/j.neucom.2025.131337.

[40] D. Rajput, W. J. Wang, and C. C. Chen, "Evaluation of a decided sample size in machine learning applications," *BMC Bioinformatics*, vol. 24, no. 1, Nov. 2023, doi: 10.1186/s12859-023-05156-9.

[41] K. P. Kutzner, T. Freitag, S. Donner, M. P. Kovacevic, and R. Bieger, "Outcome of extensive varus and valgus stem alignment in short-stem THA: clinical and radiological analysis using EBRA-FCA," *Arch Orthop Trauma Surg*, vol. 137, no. 3, pp. 431–439, Mar. 2017, doi: 10.1007/s00402-017-2640-z.