Establish a Machine Learning Model between Composition and Elastic Constants C_{ij} of Multi Component Alloys.

Course Project Report

MM 719: Introduction to Ab-Initio Methods in Materials Modelling

by

Member 1: Rohit Goyal
Member 2: Prateek Kumar Singh
Member 3: Karuskar Devangkumar Dhansukhbhai

Under the Supervision of

Prof. Sumit Saxena

MEMS Department

INDIAN INSTITUTE OF TECHNOLOGY BOMBAY

Nov 2021

Abstract

This report explores application of machine learning methods for determining elastic constants and other derived mechanical properties of multi-component alloys. A number of machine learning models, including linear regression, neural network and random forest based models, are trained and tested on a dataset of binary alloys generated using density functional theory (DFT) calculations and spanning over a large number of elemental species in the periodic table. Starting with a wide range of simple and easily accessible compositionally-averaged elemental features, a correlation-based feature selection strategy was used to systematically down-select a set of most relevant features towards the prediction of the elasticity tensor components. The true predictive performance and the associated uncertainties of the models were established by testing on unseen data and bootstrapping, respectively. A single and pair-wise feature partial dependence analysis was performed to visualize the average property trends in the multi-dimensional feature space in order to further understand the achieved predictive performance. The utility of the trained model is further demonstrated by obtaining sufficiently accurate yet highly efficient approximations for bulk modulus, Young's modulus, shear modulus and Poisson's ratio for alloys beyond the binary space (i.e., two-component alloys) on which the model was originally trained. More importantly, we test and validate the predictive performance of the developed model directly against the experimentally measured elastic constants of technologically relevant multi-component alloys (such as, Ni- and Ti-based alloys). Conceptual Design is done on Random Forest model with coefficient of determination score (R^2) of 0.743. Finally, utility of such a data-enabled route is demonstrated by predicting the possible range of various elastic properties for vast composition space available within the five component Ni-Cr-Fe-Mo-W alloy system in a high-throughput manner.

Keywords:

Materials design, Data-enabled property predictions, Materials informatics

Contents

Chapter 1

Introduction

1.1 Background

In recent past, materials informatics based approaches, in particular, machine learning (ML) based methods, have aided significantly in expediting novel and improved materials design and discovery [1–7]. In contrast to the traditional approach, where materials of interest are selected in a case-by-case manner based solely on chemical intuition or trial-and-error based strategies, the data-enabled ML route offers a much more efficient path towards targeted materials design, discovery and optimization. The tools and techniques offered by the field allow one to survey complex multiscale data using well-established statistical methods and extract meaningful information from it. These advanced methods have the potential to significantly reduce the time and resources required to develop and deploy new materials for a given application, which, could take decades in the past [7]. Given the potential and already demonstrated impact, it is not surprising that *big-data* driven material discovery and development is referred to as the fourth paradigm of material science.

In the past decade, a diverse range of ML-based approaches have been developed and applied to expedite materials design. These applications include development of supervised, unsupervised and semi-supervised learning models to address materials clustering, classification and property regression problems. In addition, generative methods, inverse design, adaptive design and active learning based strategies have been devised to enable efficient and intelligent experimental design. Over the course of this period, these strategies have gradually evolved into autonomous synthesis efforts of materials with targeted properties, which are currently actively being pursued by the community. Moreover, natural language processing and related methods are beginning to show considerable potential in automated advent of the data-enabled paradigm in materials science has led to the development of an entirely new eco-system harboring

not only a range of open materials databases [10–14], but also feature generation and ML toolkits [5–7], materials software repositories [8], journals that publish curated materials datasets [9,10], and generalized materials data formats emphasizing materials metadata [9,11]. Courtesy the staggering amount of development with active participation from materials community, the nascent field of materials informatics has quickly grown into a matured discipline with enormous potential for future growth. One of the most widely employed applications of ML in materials science is the development of efficient and accurate surrogate property prediction models. The properties that have been learned using ML-based models include, but are not limited to, thermodynamic stability, electronic, magnetic, thermal, mechanical, transport and catalytic properties.

1.2 Need

Further building on the progress in the direction of ML-based surrogate model development, in this contribution we focus on learning elastic properties for multi-component alloys spanning a large chemical space. Once validated on unseen data for their predictive performance, such models are deemed highly valuable for targeted alloy design problems — a key component of overall engineering hardware design. The primary motivation of the present work is to develop an ML model that can provide reasonable estimates of key input elastic properties that can be utilized in continuum scale finite element models frequently employed for studying deformation and microstructure evolution [8,9]. More specifically, such properties include constants of elasticity, Young's modulus, hardness, fracture toughness *etc.* We note that a number of past studies have focused on learning various elastic properties of metals, alloys and crystalline materials. Furmanchuk et. al. (2016) [10] trained a random forest (RF) regressor [1] on a density functional theory (DFT) based dataset sourced from the TE Design Lab database [2]. Trained on the compounds with bulk modulus (*B*) values of up to 250 GPa the model was shown to predict *B* within a mean absolute error (MAE) of 13.58 GPa. A similar study was conducted on silica zeolites by Evans et. al. [3] to predict *B* and shear moduli (*G*) using a gradient boosting regressor [4]. The proposed model was capable of providing good estimates for *B* and *G*, closely matching the best DFT predictions. Wang et. al. (2017) [5] used several ML models, including neural networks (NNs) [6] and support vector machines [7] to predict elastic constants *Cij* of

binary alloys. In this study, a feature database of 93 materials, including 34 pure elements, was generated using DFT calculations along with experimentally measured *Cij* constants of the same materials. The model was trained on material attributes generated from DFT with experimental *Cij* constants serving as the response variable. Similarly, a study for predicting *B* and *G* values of Fe-Cr-Al ternary alloys was performed by Wang et. al. (2019) [8] to predict the properties of any alloy falling within the target chemical space at various temperatures on the phase diagram. Wen et. al. (2019) [9] formulated material design strategy to search for high entropy alloys (HEAs) with high hardness using ML surrogate models. The model was trained on the Al-Co-Cr-Cu-Fe-Ni system to map features such as chemical composition and chemistry of elements against hardness of the alloy. A utility function was formulated along with the model to find alloys with high hardness values from nearly two million possible compositions. Utilizing an active learning feedback loop with experiments, this approach enabled identification of new alloys with hardness 10% higher than the highest value present in original training dataset. Chaudry et. al. (2020) [10] have leveraged ML models trained on Al-Cu- Mg-x alloys with $x \in$ {Zn, Zr, Si, Mn, Ag, Fe, Sn, Cr, Ge} to design high performance aluminum alloys with high hardness values. The model was built on features such as composition, aging conditions, along with physical and electronic structure parameters calculated using the weighted compositional average scheme. Zhu et. al. [11] reported a similar study for designing titanium alloys. In particular, the authors studied the effect of Mo and Cr on microstructure and mechanical properties of Ti-alloys, where NN-based ML models were trained to predict microstructure characterization such as volume fraction and size of *α* phases that are known to play a significant role in determining the mechanical properties of Ti-alloys.

1.3 Problem Statement

Establish a Machine Learning Model between Composition and Elastic Constants C_{ij} of Multi Component Alloys.

1.4 Objective

Establish an ML-based mapping between a set of compositionally-averaged and easily-accessible atomic or bulk elemental properties of the constituents forming binary alloys and their DFT-computed *Cij* constants. Note that we will be using *Cij* or *Cij constants*

interchangeably to represent all the values of *Cij* constants. Unlike most of the past studies which focused on elastic property predictions using ML, we specifically focus on features that do not require DFT calculations for the feature set enumeration. DFT-computed ground state properties such as cohesive energy, formation energy, volume per atom etc. are intentionally avoided as input features in our models, even though these features are known to be effective predictors for the target elastic constants. Starting with an initial set of features, a correlation-based feature selection was used to systematically down-select a set of most relevant predictive features. After testing a number of ML-based models, the RF algorithm was selected as the most predictive learning method. The trained RF model's true predictive performance and the associated uncertainties were established by testing on unseen data and bootstrapping, respectively. After an exploratory analysis to further understand and rationalize the models performance, the utility of the trained model is demonstrated by obtaining sufficiently accurate yet highly efficient approximations for *B*, *G*, Young's modulus *E* and Poisson's ratio (*v*) for alloys beyond the binary space (i.e., two-component alloys) on which the model was originally trained. More importantly, we test and validate the predictive performance of the developed model directly against the experimentally-measured elastic constants of technologically-relevant multi-component alloys (such as, Ni- and Ti-based alloys). Finally, the usefulness of such a data-enabled route is demonstrated by predicting possible range of various elastic properties for a vast composition space available within the five-component Ni-Cr-Fe-Mo-W alloy system in a high throughput manner.

1.5 Organisation of Report

In this report, in the initial stage, chapter 1 discusses materials informatics background. Furthur, it emphasize need of materials informatics, problem statement and our objectives throughout the course project.

Chapter 2 contains a brief summary of the Data set of research work related to ML-based elastic constant prediction model has been done.

In chapter 3, the **Machine Learning Models** are applied. The reason of ML Models Selection has been discussed further. ML Model Training and Testing has been done.

Finally, results and conclusions have been discussed in chapter 4, 5.

Data Set

2.1 Training dataset details

The training data used to developed the proposed ML-based elastic constant prediction model was acquired from the Materials Project database [10] using the automation functionality available within the pymatgen library [12]. The dataset used in the present study contains only two-component alloys spanning over a large chemical space consisting of 44 metallic elements in the periodic table. An overview of the size of different crystal structure classes as well as the occurrence frequency of different elemental species forming the dataset are shown in Fig. 1(a) and (b), respectively. Note that there can be 13 independent elastic constants for monoclinic structure. However, we have considered only 9 constants for all the elements. Monoclinic structure represents only 1.9 % of the total dataset. Distribution of these components as well as their pair-wise Pearson correlation is graphically depicted in Fig. 1(c).

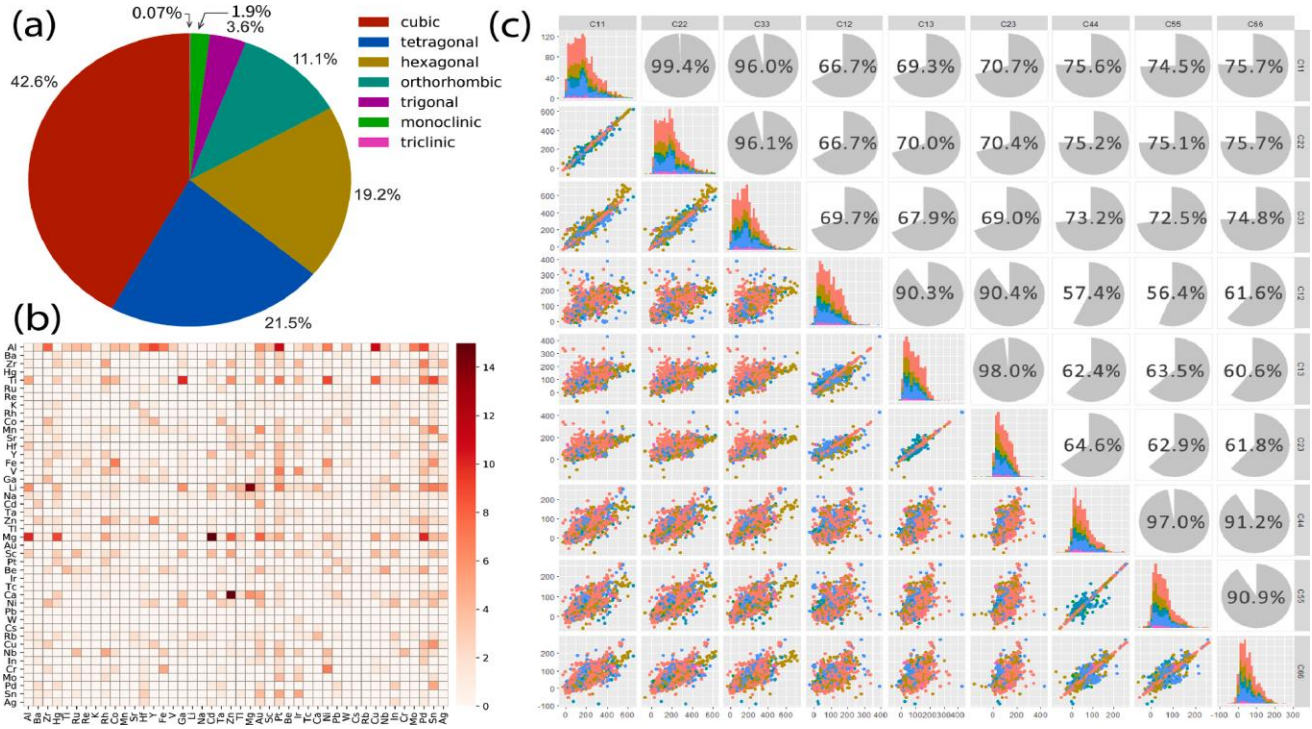

Fig. 1. Visualization of binary alloy dataset prior to preprocessing.

2.2. Data Curation and Pre Processing

Before moving on to feature selection required to numerically represent the alloy compositions in the dataset and the ML model building exercise, we carefully check the acquired dataset for internal consistency and note the presence of any outliers that do not make physical sense. The initial data curation step consisted of filtering out alloys with outliers in *Cij* and *v*. These included negative values of *Cij* and values falling beyond the range of (0–1) for *v*, which is spurious for metallic alloys. As a next step, the Voigt-Reuss-Hill [14] equations were used on *Cij* constants to obtain theoretical estimates of *B* and *G* using the relations:

$$9B_V = (C11 + C22 + C33) + 2(C12 + C23 + C31) \quad (1)$$

$$1/B_R = (S11 + S22 + S33) + 2(S12 + S23 + S31) \quad (2)$$

$$15G_V = (C11 + C22 + C33) - (C12 + C23 + C31) + 3(C44 + C55 + C66) \quad (3)$$

$$15/G_R = 4(S11 + S22 + S33) - 4(S12 + S23 + S31) + 3(S44 + S55 + S66) \quad (4)$$

The subscript *V* denotes Voigt approximation and *R* denotes Reuss approximation and *Sij* represents the corresponding elastic compliance tensor components. Using the average values of *B* (i.e., B_V+B_R 2) and *G* (i.e., G_V+G_R 2) from above equations, *v* can be obtained by,

$$v = (3B - 2G)/(6B + 2) \quad (5)$$

The estimated *B* and *v* values were compared with those computed using DFT calculations and available within the dataset. The compounds for which the DFT-calculated values showed a discrepancy of greater than ±5% from those estimated using the above equations were also considered outliers and eliminated from the ML training dataset. This curated dataset contained 1229 data points and was utilized in the subsequent model building stages.

2.3 Feature Set Selection

An enormously important step in any ML model building exercise aimed at materials properties prediction is the numerical representation of the compounds using a set of features or descriptors. These attributes should not only be efficiently computable or readily accessible but also be able to provide a unique representation for each compound in the target chemical space. Further, for the initial feature set selection physically meaningful and relevant attributes should be judiciously selected using

available domain knowledge. In the present case, we limit ourselves to either atomic or bulk solids of the elemental constituents forming the alloy. All the features considered in the present work were enumerated from Ref. [5,6], by creating a database of features for all the metallic elements. For the compounds with multiple elemental constituents, we simply take a compositional average of all the single component features. For instance, a compositionally-averaged feature *F* of an alloy *AxBy* is computed as

$$F = x \times FA + y \times FB \; x + y \quad (6)$$

where, *FA* and *FB* are the atomic or bulk elemental feature for the constituents *A* and *B*, respectively. The specific choices of features in the initial set are outlined in Table 1 and represent simple electronic, structural, elastic or thermal attributes of the alloys. Furthermore, the simple compositional averaging adopted in our feature engineering approach is trivially extended to alloys beyond the two-component.

Table 1 A list of initial feature set along with the results of feature selection strategy based on variation inflation factor (VIF) approach are presented. A VIF value greater than 10 indicates high multi-colinearity. Feature selection was performed to obtain optimum features with reduced VIF values without compromising on model performance.

Features	VIF Before feature selection	VIF After feature selection
Density	34.02	14.21
Poisson's ratio	22.82	Dropped
Bulk modulus	24.34	7.54
Melting point	37.82	Dropped
Thermal conductivity	3.88	3.09
Radius	68.81	14.41
Electronegativity	78.82	Dropped
Specific heat	2.73	2.48
Formation energy	3.31	3.17

ML models are generally considered interpolative modes in the underlying feature space. The dimensionality of the feature space then dictates the complexity of the model. As a general rule, more complex models can lead to higher accuracy but tend to be poorly generalizable beyond the specific dataset they are trained on. Furthermore, the initial feature set can often have correlated features leading to redundant information in the dataset. Therefore, as a general practice, a feature down-selection is carried out prior to building an ML model. For feature selection, we resort to a multi-collinearity analysis, which essentially aims to describe how each feature can be explained by the combined effects of all other features. It is basically measured by fitting a multi-linear regression line on each of the features while all other features are taken as the independent variables. The degree of fit of this regression line is measured by variation inflation factor (VIF) which serves as metric of multi-collinearity. A VIF value much greater than 10 indicates high multi-collinearity, which means that the corresponding feature is largely redundant and can be well explained by a linear combination of rest of the features. The list of all features before and after feature selection, along with the VIF values are shown in Table 1. Here we note that feature-selection should be carried out such that the feature down-selection step reduces the number of features involved without adversely affecting the predictive performance of the trained ML model on unseen data. This was validated by comparing the performances of the reduced 5-feature model with the 10-feature model which only differs by 1.7 GPa in terms of MAE values on unseen 10% of randomly selected test set. However, the model complexity of 5-feature model is significantly lower. Furthermore, the VIF analysis generally leads to the removal of highly correlated features. This was verified via a feature-feature pairwise Pearson's correlation analysis. The Pearson's correlation matrices for the initial and the down-selected sets are shown graphically in Fig. 2(a) and (b), respectively. It can be seen that a number of strongly correlated features present in the initial set have been omitted in the final set. For instance, in Fig. 2(a) ρ can be seen to be strongly correlated with both *EN* and *B*. Therefore, in the down-selected feature set only one of the three features, *i.e.*, ρ survives. The finalized reduced feature set consists of five features, which are subsequently used for model training and validation.

Chapter 3

Machine Learning Models

3.1 ML model selection

Similar to the feature selection step, the selection of an ML algorithm to train the model is an equally important and critical aspect of any statistical model building exercise. This selection also hinges on the underlying transparency-accuracy tradeoffs. While simple models such as linear regression are very transparent, and therefore also completely explainable and interpretable, more complex models, such as neural networks (NNs), can have a large set of model parameters leading to an improved fitting performance but little transparency and are often treated as black boxes. Tree based ensemble models, such as random forest (RF) models, on the other hand provide a balance between the model interpretability and predictive accuracy. In this work, we performed an exploratory analysis to test the predictive performance of linear, NN and RF regression models. Owing to the superior performance of RF regression, this model was selected going forward for the subsequent detailed analysis. RF-also frequently referred to as random decision trees–is a tree-based ensemble learning method. Tree-based ensemble methods combine several decision trees (weak learners) to produce better predictive performance than that obtained by utilizing a single decision tree and are generally categorized as bagging or boosting methods and RF models belong to the former category. Bagging refers to a technique to grow ensemble decision trees, where a large number of decision trees are grown via a random subset selection from the examples in the training set . RFs employ several randomly selected subsets of both the supplied training examples and the features to grow trees rather than using the entire sample set or feature set. This strategy leads to an improved prediction performance due to better variance-bias trade-offs, and makes the model inherently robust against overfitting. In the case of a random forest, the major hyperparameters that often require tuning via cross-validation include the number of decision trees in the forest and the maximum allowed tree depth considered by each tree when splitting a node.

3.2 ML model training and testing

Creating a robust ML model requires data for training, validation and testing. A 70%:20%:10% split was used in the present study, where a randomly selected 70% portion of the available dataset was used to fit the model, another 20% was used to obtain an unbiased evaluation of the trained model which also helped in tuning the model hyperparameters and the remaining 10% was used to test the performance of the final model on unseen data. After the determination of optimum hyperparameters using the validation set, training and validations splits were combined for obtaining a 90% combined-training set. This was used for the final model training. As more data was fed into the final model, it showed better accuracy for tests than the model that was trained on 70% data. To evaluate the impact of the training dataset size on the trained models prediction performance on unseen data, learning curves were also evaluated. Finally, to quantify the predictive performance of the model at model building, testing and prediction stages, we consistently use three different metrics, namely root mean squared error (RMSE), mean absolute error (MAE) and coefficient of determination (*r*2). The first two are commonly used error metrics and closely related. In fact, for a dataset with vanishing variance in the prediction error RMSE and MAE coincide. As the variance in the error increases, RMSE tends to get larger than MAE. *r*2 is another standard measure of goodness of fit capturing proportion of the variance in the dependent variable that is predictable from the independent variable.

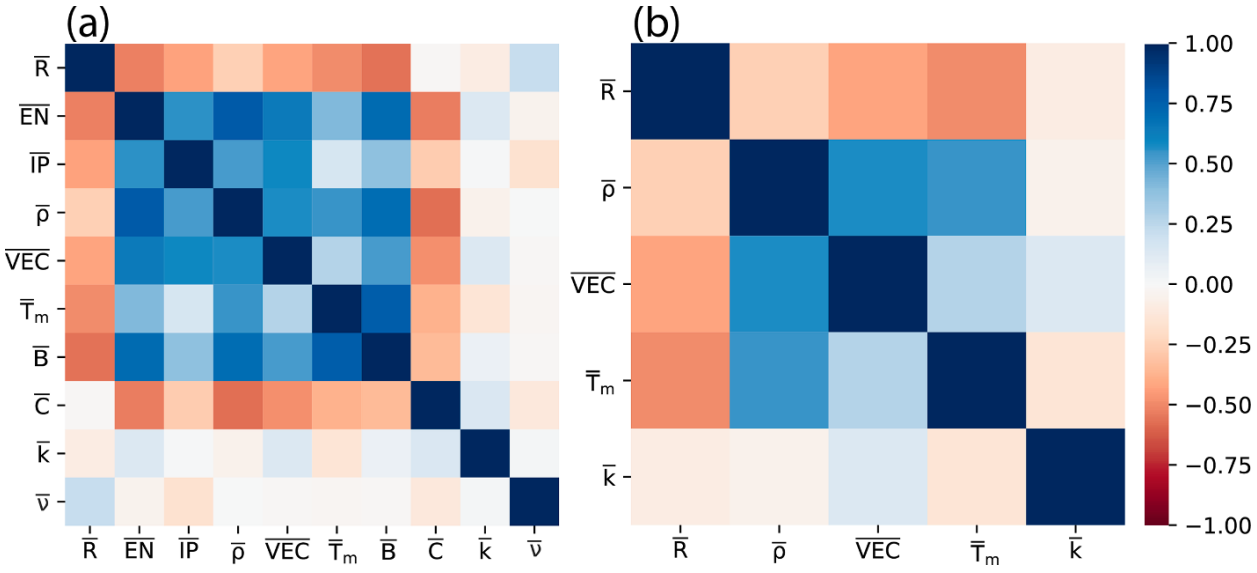

Fig.2: Pearson's correlation plot prior to and post feature selection.

Results and Discussion

With the primary goal of creating a robust predictive mapping between the down-selected compositionally-averaged alloy features and the 9 *Cij* constants, we start by evaluating reasonable choices of the two model hyperparameters, namely the *maximum allowed tree depth* and the *number of estimators/trees* for each of these models. Due to the qualitatively similar nature of the prediction problem that these models aim at addressing, we perform the hyperparameter selection using the averaged performance of all these models on the validation set. Although, tests were also performed to check that the conclusions based on the averaged predictive performance do not change when addressing each regression model separately. The hyperparameters were tuned by training the RF models using different tuples of the maximum allowed tree depth and number of estimators on a grid and then evaluating the trained models' average performance on the validation sets. The performance, evaluated in terms of RMSE and *r*2, is presented in Fig. 3(a) and (b), respectively. We find that the performance doesn't improve much beyond the number of estimators equal to 50 and the maximum allowed depth of trees equal to 10. In light of these results, we select the maximum allowed depth of trees equal to 10, but a more conservative choice of 500 for the number of estimators was made. We note that since the RFs are intrinsically robust against overfitting (also seen from the flat performance curves beyond a certain value of the hyperparameters in Fig. 3(a) and (b)), allowing additional number of estimators (or trees) can not cause a relatively poor performance on unseen data but might be useful for future situations where the current dataset might be augmented by additional data points. With the hyperparameters determined, we evaluate the impact of the training dataset size on the test set prediction performance using the learning curves. For an effective ML model, by definition the predictive performance on the model on unseen data should improve with the training dataset size. For this step, the total 1229 data points were divided into 1100 and 129. The 129 randomly selected data points were

secured for testing the model, and training was performed in increments from the other set. The number of data points for training were chosen in increments of 50 till the count reached the total size of 1100. At each increment, the model training was performed and tested 20 times, with different sets of randomly selected training and test sets each time. The size dependence of the training set on the mean training and test set prediction performance is shown in Fig. 3(c). The error bars in each case represent the standard deviations computed on the 20 different training/test splits. It can be seen that the average model performance on the test set systematically improves with the training dataset size and the confidence (quantified by the shrinking error bars on the training and testing data learning curve) in the learned model also improves. Looking at the slope of the test set performance, it is likely that the model's predictive performance can further be improved by moving on to larger size training datasets, if more data becomes available in the future.

The final performance of the RF model trained on randomly selected 90% data and tested against the remaining 10% is shown in the Table 2, as quantified in terms of RMSE, MAE and *r*2. This reported performance is averaged over the set of all *Cij* constants, while the individual model performances are graphically depicted in Fig. 4. It can be seen that the developed models can reliably predict the DFT-level elastic constants for a vast majority of the cases. In some cases, for instance in the case of *C*12, due to the limited number of training data on the higher end of the GPa scale the performance is relatively poor, which can potentially be further improved by selectively adding more training data points in that range.

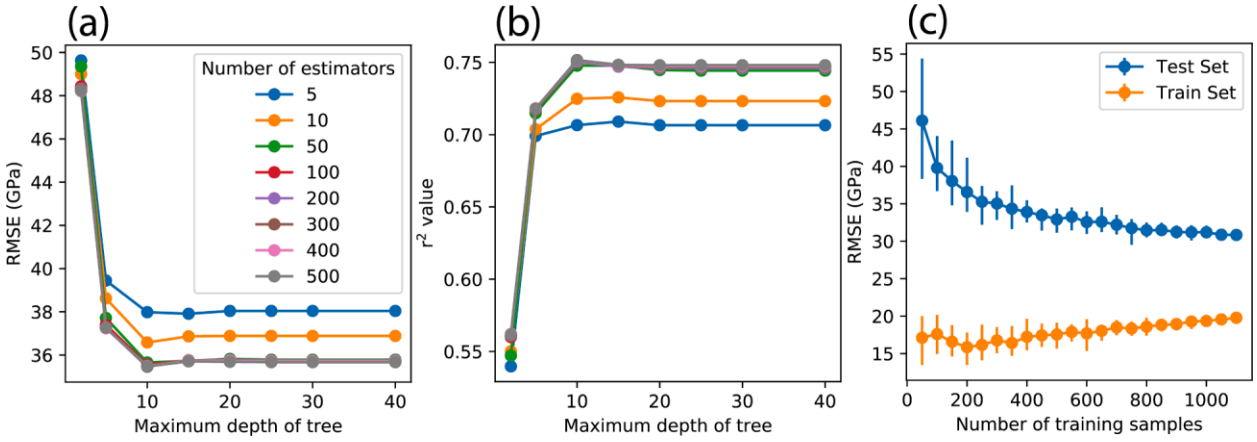

Fig. 3. Hyper-parameter optimization and learning curves for Random Forest model

To further understand the inner workings of the models beyond the usual “blackbox” treatment, we perform the relative feature importance analysis and pairwise feature-feature partial dependency plots. In decision trees forming a random forest model, at a given node, a selected feature is used to make a decision on how to divide the data set into two separate sets. In a regression problem these features are quantitatively selected using a variance reduction criterion. Therefore, one can quantify how much each feature contributes on average to the overall variance reduction in the model. The average over all trees in the forest is then used as a measure of the relative feature importance. The feature importance plot presented in Fig. 5(a) shows results of our relative feature importance analysis averaged over all the RF models. It can be seen that compositionally-averaged melting temperature of the bulk elemental solids forming the alloy, *Tm*, appears as a feature of highest importance, followed by *R* and *ρ*. The importance of *Tm* can be intuitively rationalized by noting the qualitative trend that alloys formed by high (low) melting temperature metals usually tend to be stiffer (softer) and therefore should lead to higher values of elastic constants. Similar arguments can also be made for *R* and *ρ*. Alloys formed by components with smaller atomic radii with high density configurations should translate to higher packing fraction solids with stiffer elastic constants. We further confirm the anticipated trends of the top features with *Cij* constants using pairwise feature-feature partial dependency plots. These plots essentially convey information about the dependence of one or more features on the final result of the model. Once the model is trained, single and pairwise partial dependencies for features, $p(Fi)$ and $p(Fi, Fj)$ can be computed by integrating out (or by marginalizing) one or more features, as:

$$p(f1) = \int dF2 \cdots dFN\mathbf{F}\,(F1, F2, \ldots, FN),\ p(f1, f2) = \int dF3 \cdots dFN\mathbf{F}\,(F1, F2, \ldots, FN)$$

Here, **F** represents the machine learned function. If a large number of features are involved, the cost of these calculations can become significant in a multidimensional feature space involving a fine integration grid. Therefore, it is common to replace the integration values of the features to be marginalized with their average values *Fi*. For example,

$$p(F1) = \mathbf{F}\,(F1, F2, \ldots, FN). \quad (7)$$

While the analysis was carried out for all five features and nine *Cij*, in Fig. 5(b) we present a select set of pairwise feature-feature partial dependency maps for some of the top features to highlight certain trends that were anticipated from the feature importance analysis. For the sake of illustration, the partial dependency maps are shown in form of contour plots between three pairs of top features for *C*11 and *C*66 in the top and bottom rows, respectively. Using these plots one can look at the average trends in the target variables as a function of different feature values. For example, it can be seen that a higher *Tm* consistently leads to a larger value of the elastic constants in each case. Similarly, a smaller *R* and higher *ρ* also correlate with a larger value of the target variables.

Further details on partial dependency plots for the top three single features against all *Cij* is given in SI, which further confirms that the identified trends are indeed general and consistently followed over the entire set of nine elastic constants. Such partial dependency plots are enormously useful as they not only provide a transparent explanation of the learnt rules and relationships for materials property prediction problems, but can also serve as informatics-based design maps that can guide discovery of novel materials with target properties and help to systematically navigate a high dimensional feature space. Next, to confirm that the ML-predicted *Cij* constants are consistent with the DFT-computed elastic constants such as *B* and *G*, we use the Voigt-Reuss-Hill [14] equations (i.e., Eqs. (1)–(4) respectively) to estimate *BV, BR, GV* and *GR* using the ML-predicted *Cij* constants on a randomly selected 10% test set. Subsequently, the average values from both the Voigt and Reuss equations were used to get approximations of *B* and *G*. Further, using these averages, *v* was calculated using Eq. (5). The obtained values of *B* and *v* are compared against the DFT-computed *B* and *v* in the parity plots shown in Fig. 6. The performance metrics for *B* and *v* predictions are listed in Table 2. As such, the MAEs of 6.77 GPa and 0.021 for the *B* and *v* predictions, respectively, on unseen data indicate an excellent agreement with the DFT computations taken as the ground truth.

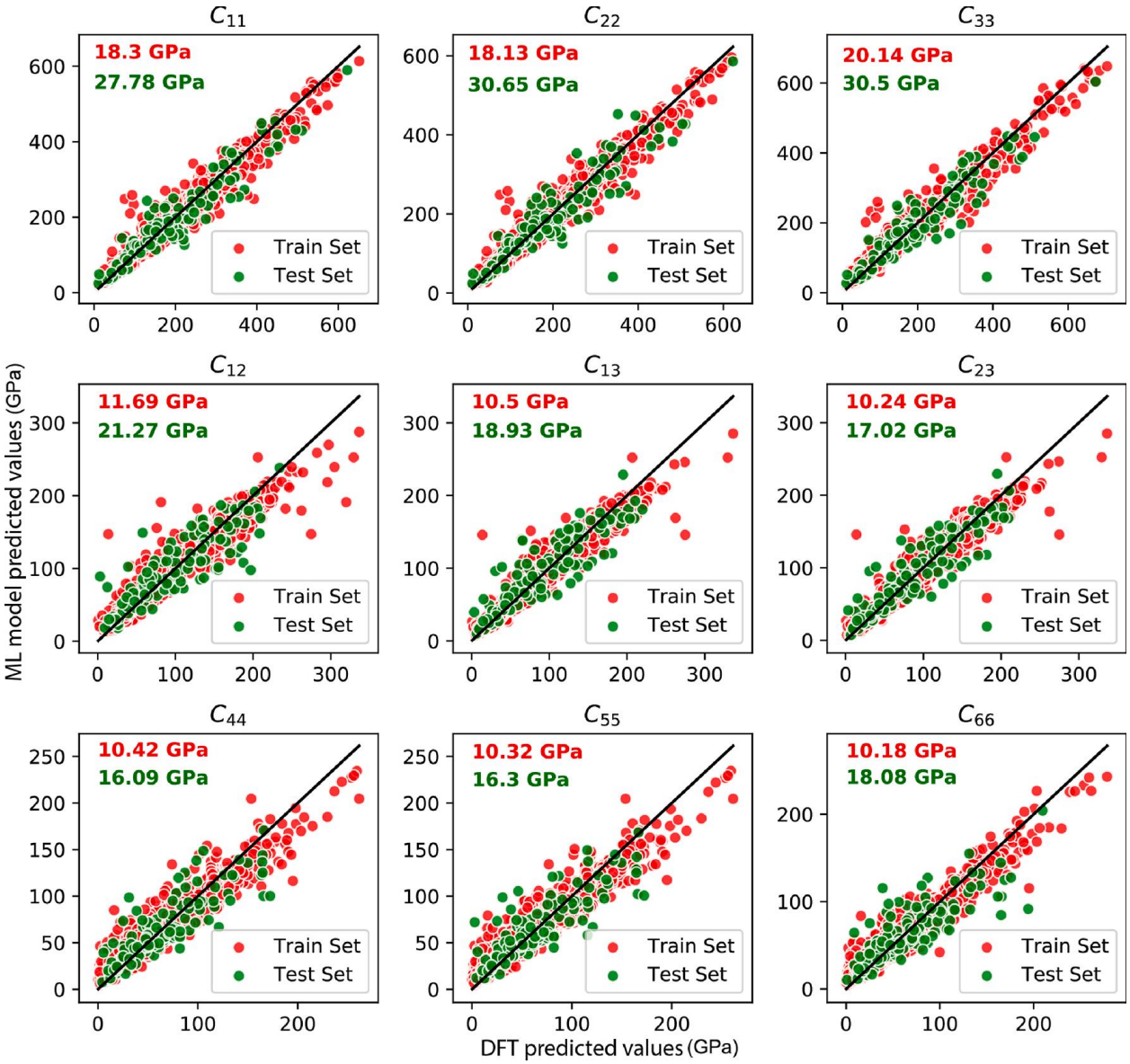

Fig. 4. Pair plot grid comparing the *Cij* constants predicted by the random forest model against the ones calculated using DFT. The training and test sets consisted of 90% and 10% of the data respectively. The numbers on the top left corner of each plot represent the mean absolute error (MAE) values for the training and test sets indicating the overall error with respect to DFT values.

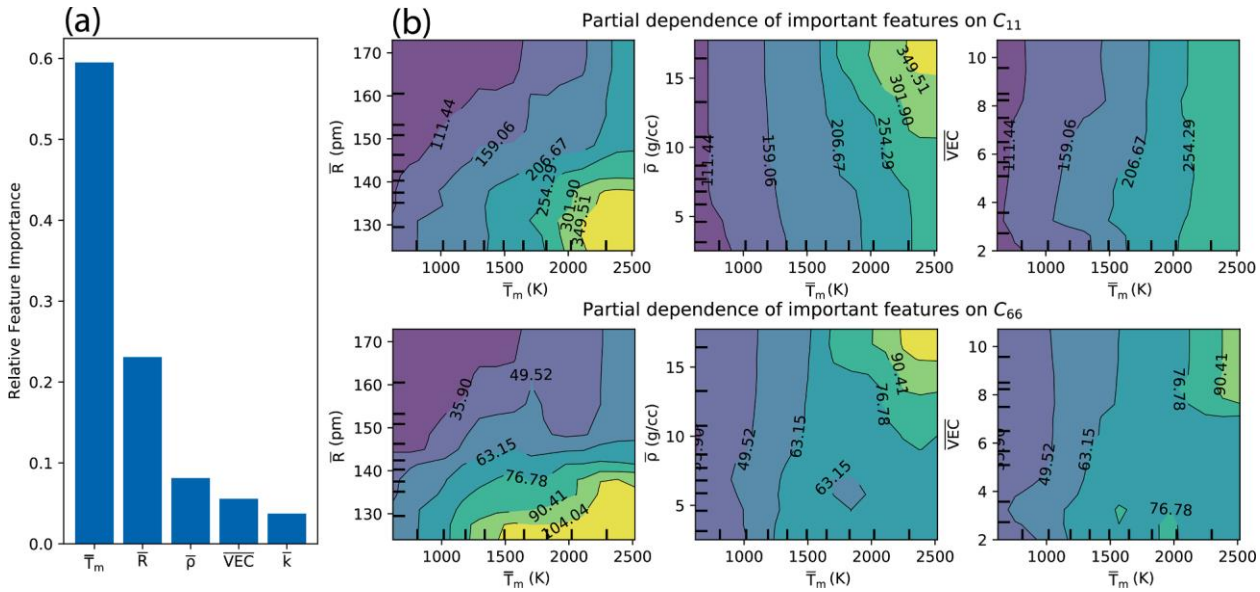

Fig. 5. Feature importance and partial dependency plots for the random forest regression model. (a) Feature importance plot indicating the rank of the five down-selected features used to build the model. (b) Selected two dimensional partial dependency plots of *C*11 (top row) and *C*66 (bottom row) as a function of melting temperature (X axis) and *R, ρ* and *VEC* on the Y axis (from left to right).

The various steps involved in the model development, validation and testing thus far have solely utilized the DFT-generated dataset. Although it is well established that DFT computations can accurately estimate the elastic constants of multi-component metallic alloys and solids in general, we explicitly test our ML models prediction performance against experimentally measured elastic constants for a diverse range of binary and superalloys that were not a part of our original training dataset. The experimental dataset used from the literature is provided in SI. Since the collected experimental dataset had several missing values, only *C*11*, C*12 and *C*44 were chosen to test the model performance. The obtained results are shown in Fig. 7. It can be seen that the model also works reasonably well for extrapolating beyond the two-component systems originally used to train the model. The predictions compare very well against the values for nickel-based superalloys such as IN792 and IN600. The *Cij* constants for titanium alloys such as Ti64, Ti6242 and Ti6246 are also predicted within 15 GPa from their true values. However there are certain alloys such as IN625 and Ti38644 for which the

elastic constants were poorly predicted by the model. This can potentially be attributed to the fact that the proposed model does not take finer details of processing conditions and microstructural parameters into consideration and simply operates under the uniform solid solution assumption. However, barring certain cases, the overall agreement with the experimental observations is comparable to that of the test set prediction performance measured on the DFT-generated data (see Table 2). Since the compositionally averaged features for the model can be created for any multi-component alloy, we can leverage the developed and validated ML models to predict the properties for any general multi-component alloy system.

To demonstrate this with a concrete example, a five-component Ni-Cr-Fe-Mo-W composition was selected. We chose Ni- Cr-Fe-Mo-W because these elemental constituents are frequently appearing key elements in superalloys – an important class of engineering materials. A combinatorially enumerated composition grid was generated with a compositional grid spacing of 0.1 between any two adjacent compositions. After a combinatorial enumeration of all possible compositions, the compositionally-averaged features were computed and used within the developed ML models for the *Cij* predictions. Subsequently, the properties such as, *B*, *G*, *v* were derived using Voigt-Reuss- Hill [14] relations and elastic equations described previously in Eqs. (2)– (5). Further, Young's modulus (*E*) was calculated from *B* and *G* using the relation: $E = 9BG$ $3B + G.$ (8) Predictions of our ML models for the entire set of the compositions considered here are provided in the SI accompanying the manuscript. Although instantaneous predictions for the individual compositions are valuable in their own right, to take a look at the average chemical trends in the predictions we further analyze chemistry-dependent analyses of the predicted properties over the entire dataset. For this, the generated compositions along with their corresponding predicted properties were grouped according to the dominant elemental constituents. If there is only a single dominant elemental constituent (that is, if a given element in an alloy composition has higher fractional occurrence than every other elemental constituent), then the composition is assigned to a group named with the symbol of corresponding element. If there are multiple dominant elements (with an equal fractional composition), then the group is named after the binary element symbols, and so on. Distributions of each of the

properties over all such possible groups are computed and presented, using a box plot, in Fig. 8. Thus the box plots indicate the range of the respective properties spanned for a particular dominant alloy composition within the Ni-Cr-Fe-Mo-W multi-component system. This will enable us to glean insights into the compositionally-average property values that one can expect for any custom composition, paving a path for targeted design. For instance, Fe and Cr-based alloys exhibit a wider variability in *B* and thus present a higher potential for tunability, while W-based alloys generally lead to higher *B, G* and *E* values, but a lower *v*. We also note that while we have used a specific example of Ni-Cr-Fe-Mo-W alloys system, the developed interpolative ML-based prediction approach is entirely flexible and can easily be extended to any other composition originally included in our training data set. However, it is anticipated that going beyond the elements in the initial training dataset would require an augmentation of the training data, followed by retraining and validation steps, before any predictions can be made with confidence.

Lastly, we close by briefly outlining some of the limitations of the developed ML-based prediction models. A consequence of using compositionally-averaged features is that a random solid solution phase is implicitly assumed for any target composition. As a result, the model is agnostic to property variations due to different ordered configurations and only applies to random solid solutions. Furthermore, effects such as phase transitions, secondary phases, precipitate hardening are not included. We utilized DFT-based training dataset, therefore the predictions are limited to only a low temperature regime. Finally, variation in elastic properties as a result of processing (hardening, tempering, annealing) are not explicitly considered in the absence of training data and effect of defects and dislocations are not accounted for. Despite these limitations, we believe that an efficient and reasonably accurate elastic property prediction scheme for multi-component alloy systems can be enormously useful in an initial round of screening of promising candidate systems for a subsequent in depth analysis.

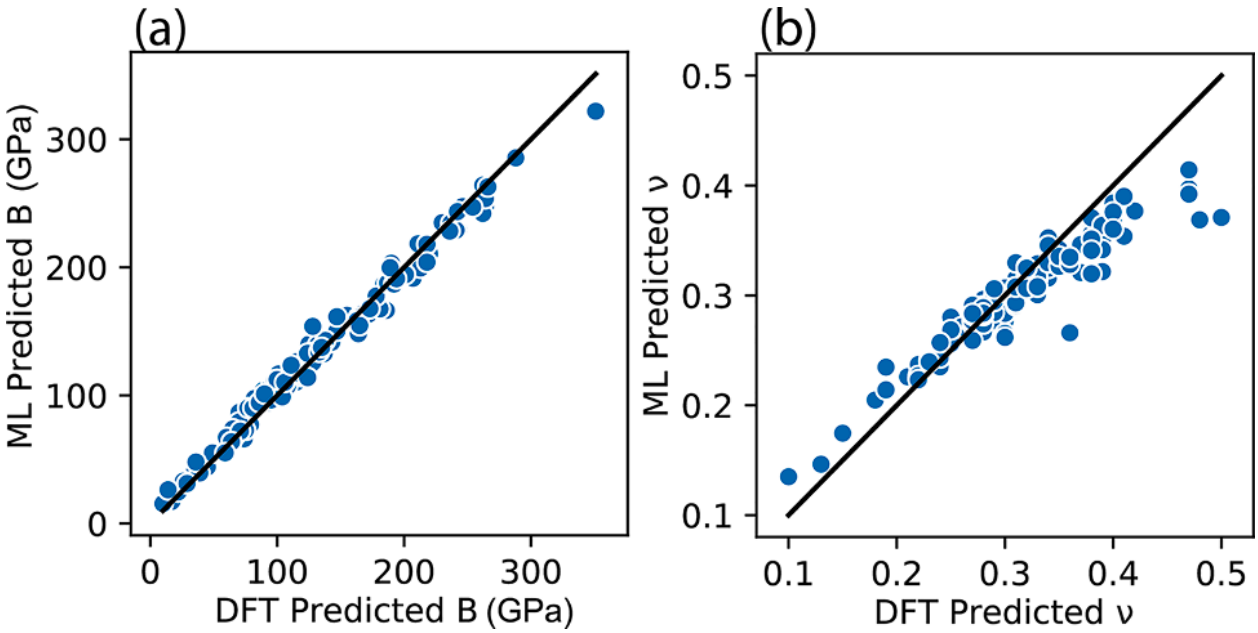

Fig. 6 Plot of true DFT predicted (a) bulk modulus (*B*) and (b) Poisson's ratio (*v*) from the unseen test dataset against the Voigt-Reuss-Hill averaged values calculated from predicted *Cij* using random forest regression model.

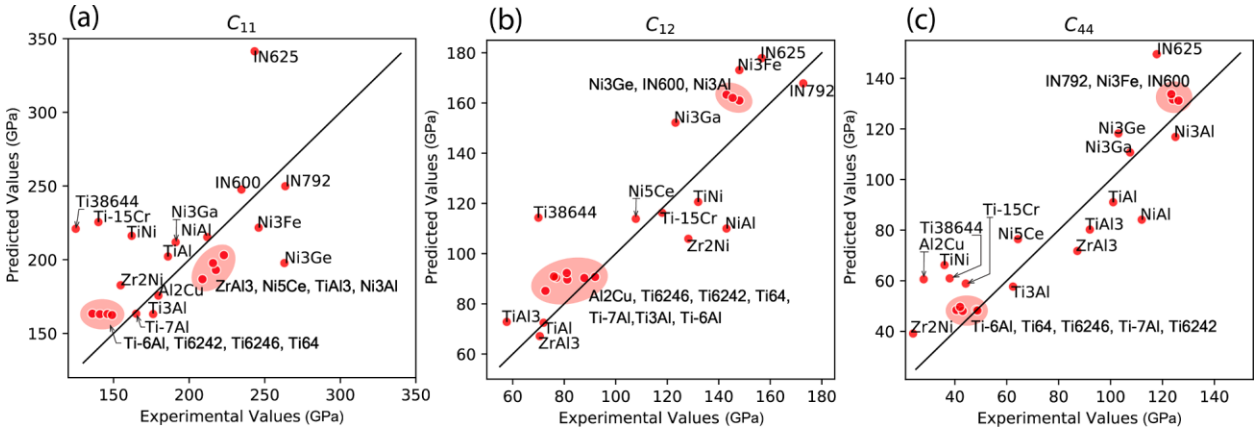

Fig. 7 Comparison of experimentally determined elastic constants (a) *C*11, (b) *C*12 and (c) *C*44 of binary alloys and super alloys collated from literature with the elastic constants predicted by the ML model.

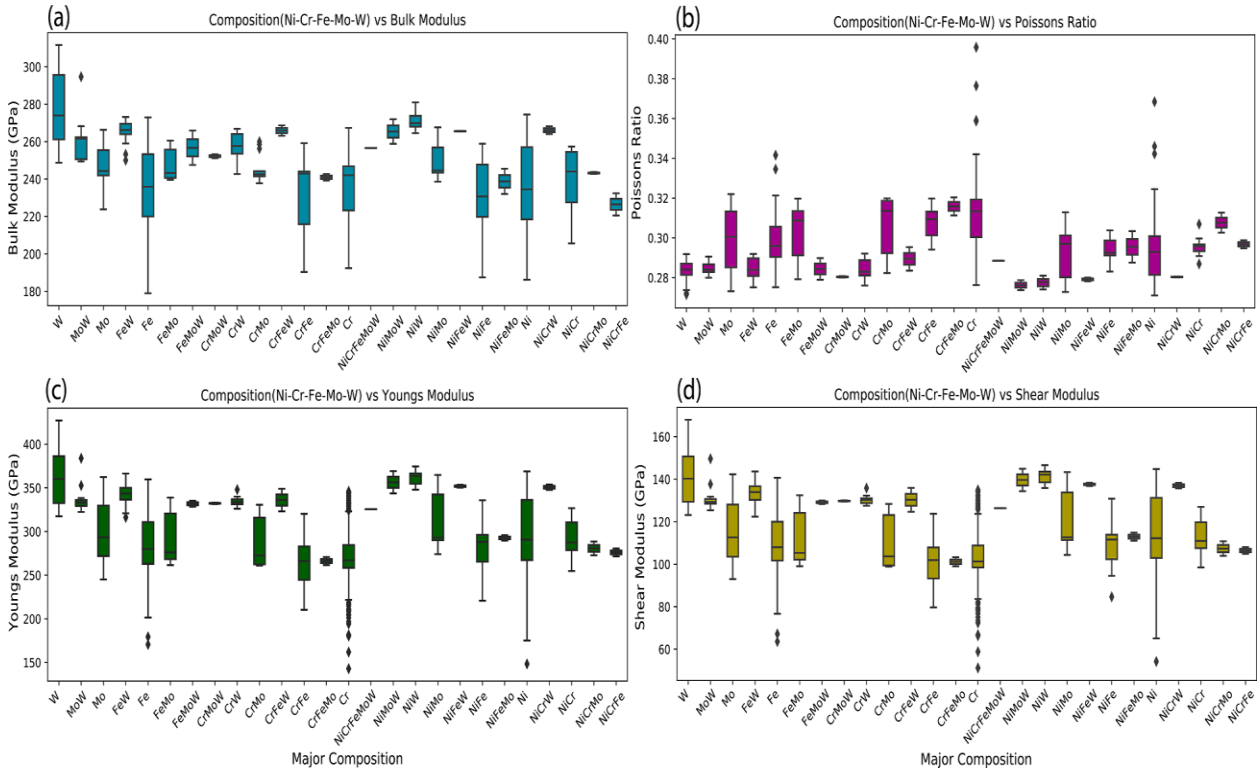

Fig. 8. Box plots comparing the elastic property predictions for a 5 element composition grid of Ni-Cr-Fe-Mo-W. The elastic property were evaluated by applying the Voigt-Reuss equations and relations on the *Cij* constants predicted by the random forest model. Alloys indicated on the X-axis are named based on the major constituent elements in the alloy. For example, W indicates that the major constituent is W by atomic percent, MoW indicates Mo and W are in equiatomic percent and all other elements are minor alloying elements. The box shows the quartiles of the predicted property while the whiskers extend to show the rest of the distribution within the group, the points beyond the whiskers denote outliers determined by a function of inter-quartile range. (a) Bulk modulus. (b) Poisson's ratio. (c) Young's modulus. (d) Shear modulus.

Chapter 5

Conclusion and Future Scope

In the present work, we used a curated database of DFT-computed elastic constants spanning a wide range of binary alloy chemistries to develop and validate ML models for efficient predictions of elastic properties. Most notably our ML models used compositionally-averaged features of atomic and elemental solids of the constituents forming an alloy and DFT-based features were intentionally avoided to ensure an efficient predictive scheme that doesn't require any computational overhead for feature set generation. After a rigorous feature down-selection and exploratory analysis that aimed at understanding the physical relevance of the selected features towards the elastic property predictions, we tested the trained models on DFT-computed data to quantify the prediction accuracy of the developed approach. More constants as a test set, we showed that the developed model is generalizable to multi-component alloys, beyond the initial training space of just the binary alloys.). Conceptual Design is done on Random Forest model with coefficient of determination score (R^2) of **0.743**.

Finally, we demonstrate the practical utility of the developed model by making high-throughput predictions on all possible compositions that can be combinatorially-enumerated within a five-component alloy system Ni-Cr-Fe-Mo-W and analyzing the resulting chemistry-dependent chemical trends.

References

[1] Krishna Rajan, Materials informatics, Materials Today 8 (10) (2005) 38–45.

[2] Gregory J. Mulholland, Sean P. Paradiso, Perspective: Materials informatics across the product lifecycle: Selection, manufacturing, and certification, APL Materials 4 (5) (2016), 053207.

[3] Dane Morgan, Ryan Jacobs, Opportunities and challenges for machine learning in materials science, Annual Review of Materials Research 50 (1) (2020) 71–103.

[4] Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, Chiho Kim, Machine learning in materials informatics: recent applications and prospects, NPJ Computational Materials 3 (1) (2017) 54.

[5] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, Aron Walsh, Machine learning for molecular and materials science, Nature 559 (7715) (2018) 547–555.

[6] Jonathan Schmidt, M´ario R.G. Marques, Silvana Botti, Miguel A.L. Marques, Recent advances and applications of machine learning in solid-state materials science, NPJ Computational Mathematics 5 (2019) 83.

[7] G. Pilania, P.V. Balachandran, J.E. Gubernatis, T. Lookman, Data-Based Methods for Materials Design and Discovery: Basic Ideas and General Methods, 2020.

[8] Ankit Agrawal, Alok Choudhary, Perspective: Materials informatics and big data: Realization of the fourth paradigm of science in materials science, APL Materials 4 (5) (2016), 053208.

[9] Claudia Draxl, Matthias Scheffler, Nomad: The fair concept for big-data-driven materials science, 2018.

[10] Logan Ward, Chris Wolverton, Atomistic calculations and materials informatics: A review, Current Opinion in Solid State and Materials Science 21 (3) (2017) 167–176. Publisher Copyright: 2016 Elsevier Ltd Copyright: Copyright 2017 Elsevier B.V., All rights reserved.

[11] Anubhav Jain, Geoffroy Hautier, Shyue Ping Ong, Kristin Persson, New opportunities for materials informatics: Resources and data mining techniques for uncovering hidden relationships, Journal of Materials Research 31 (8) (2016) 977–994.

[12] Benjamin Sanchez-Lengeling, Al´an Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, Science 361 (6400) (2018) 360–365.
[13] Lihua Chen, Ghanshyam Pilania, Rohit Batra, Tran Doan Huan, Chiho Kim, Christopher Kuenneth, Rampi Ramprasad, Polymer informatics: Current status and critical next steps, Materials Science and Engineering: R: Reports 144 (2021) 100595.
[14] Arun Mannodi-Kanakkithodi, Ghanshyam Pilania, Tran Doan Huan, Turab Lookman, Rampi Ramprasad, Machine learning strategy for accelerated design of polymer dielectrics, Scientific Reports 6 (1) (2016) 20952.
[15] Rohit Batra, Hanjun Dai, Tran Doan Huan, Lihua Chen, Chiho Kim, Will R. Gutekunst, Le Song, Rampi Ramprasad, Polymers for extreme conditions designed using syntax-directed variational autoencoders, Chemistry of Materials 32 (24) (2020) 10489–10500.