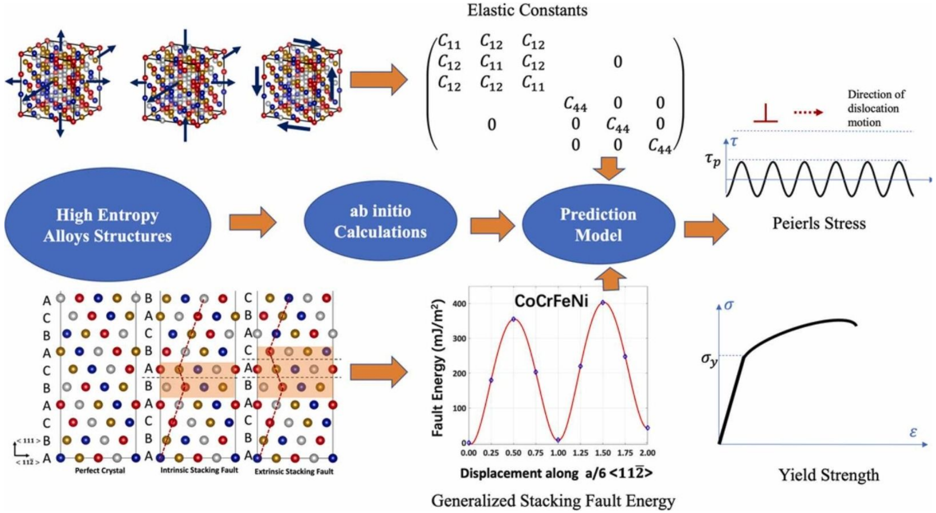
Property Prediction of HEAs using ML



Reference

Endsem ME793

Shiv Modi
19D100011

Background of the Problem

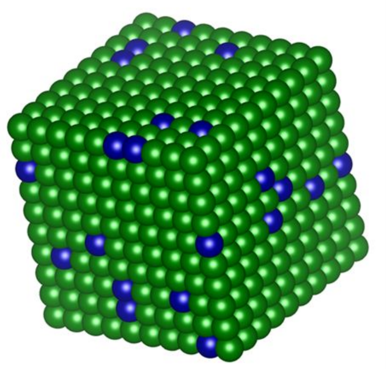
[Science Direct](#) - [Yield strength prediction of high-entropy alloys using machine learning](#).

Why is it important ?

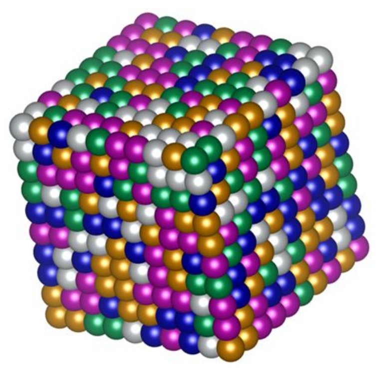
Predicting the properties of HEAs is important for several reasons:

1. Material Design and Optimization
2. Performance Assessment
3. Cost Reduction
4. HEAs tend to have 'good' properties
5. Understanding Structure-Property Relationships

Overall, accurate prediction of properties of high-entropy alloys is essential for advancing the development and application of promising materials in various industries.



Conventional alloys with one principal elemental component and small amounts of additional element.



High entropy alloys with high (equal) amounts of 5 or more different elements

Fig1: https://www.isis.stfc.ac.uk/Pages/SH22_HEAs.aspx

Motivation & Objectives

Motivation:

- HEAs have many ‘good’ properties
- Traditional trial and error method is costly and time consuming
- Thus, a good ML model can help us in this pursuit
- Particularly, finding the relationship between alloy properties is of great scientific importance
- Further, a model that can predict alloy composition for a given set of requirements is highly desirable
- There are vast datasets available for HEAs

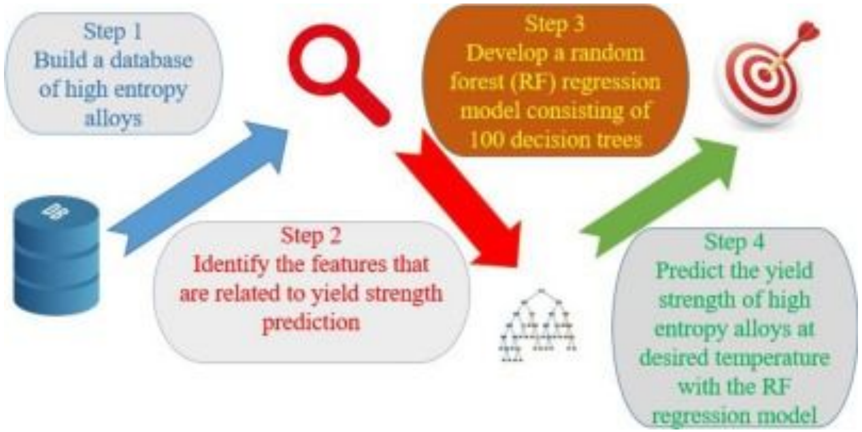


Fig2: Key Steps [Reference]

Objectives:

- Utilizing a Random Forest (RF) regressor model to predict yield strengths of MoNbTaTiW and HfMoNbTaTiZr at 800°C, 1200°C, and 1500°C. [1]
- Incorporating features like mixing of entropy, bulk modulus, valence electron concentration, and melting temperature in the ML model.
- Testing other algorithms like SVM[2], GP[3], KNN[4], NN[5] to predict other properties like UTS, phase type, Melting point, etc.

Data Analysis

- On left side, calculated the F-values and P-Values of all (103) features with respect to Bulk Modulus and D Bulk.
- The result was that there are about 50 key properties correlated with Bulk Modulus, and about 30 key properties correlated with D_Bulk.
- Rest of the properties had P values more than 10^{-3}
- The bottom values commonly correspond to elements

	features	f_values	p_values
0	a (Å)	1885.752281	7.714943e-241
1	Elec_nega	1235.266947	4.480027e-182
2	Tm (K)	651.885366	2.586338e-113
3	VEC	305.140318	1.680523e-60
4	SSsp	243.778620	8.544190e-50

	features	f_values	p_values
0	D_elec_nega	320.565804	4.061945e-63
1	D_Tm (K)	216.025241	8.687131e-45
2	Zr	199.002622	1.158671e-41
3	FCCp	187.965045	1.296955e-39
4	delta	133.417913	3.272282e-29

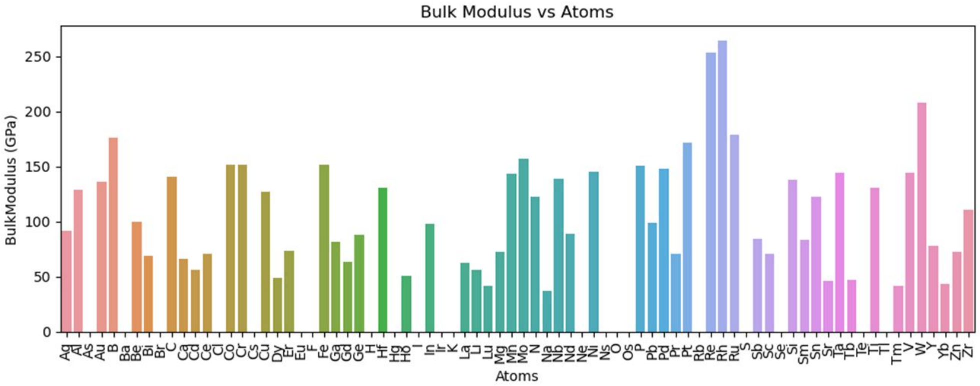
	features	f_values	p_values
93	K	0.0	1.0
94	Ne	0.0	1.0
95	Ns	0.0	1.0
96	O	0.0	1.0
97	Os	0.0	1.0
98	Rb	0.0	1.0
99	S	0.0	1.0
100	Se	0.0	1.0
101	Te	0.0	1.0
102	Tl	0.0	1.0

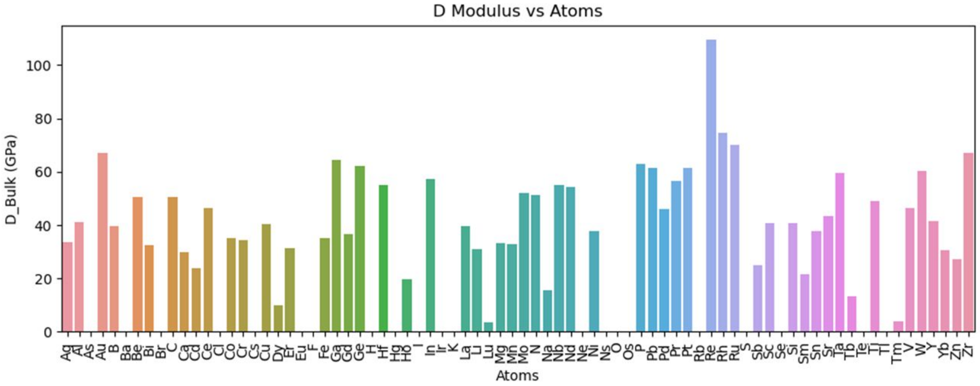
	features	f_values	p_values
93	K	0.0	1.0
94	Ne	0.0	1.0
95	Ns	0.0	1.0
96	O	0.0	1.0
97	Os	0.0	1.0
98	Rb	0.0	1.0
99	S	0.0	1.0
100	Se	0.0	1.0
101	Te	0.0	1.0
102	Tl	0.0	1.0

Top 5 and Bottom 10 values for Bulk Modulus (left)
D_Bulk (right)

Data Analysis (Contd.)

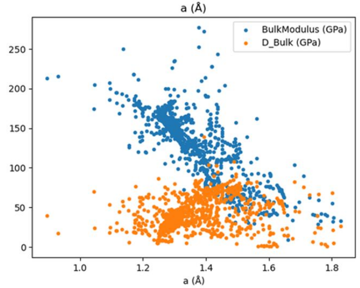
- On the left, there are two bar plots- the top plot displays the average bulk modulus of HEAs that have a particular element, say Ag present in them Similarly, the bottom plot corresponds to the D_bulk parameter.
- The gaps in the plot are because of presence of no HEAs of the given elements, and this has been taken care of while cleaning the data
- This plot gives a clear direction about which elements to choose for a particular range of bulk modulus

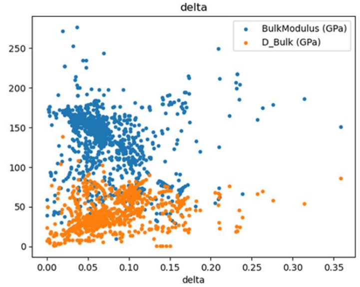


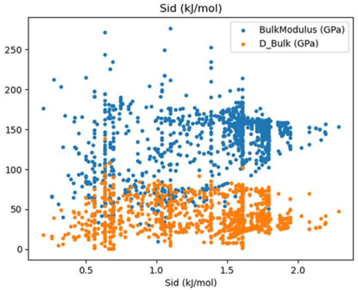


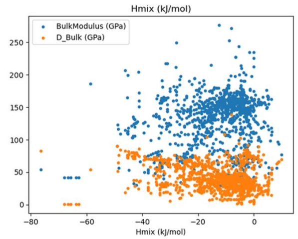
Data Analysis (Contd.)

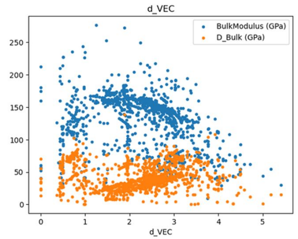
- These plots indicate the trends for the Bulk Modulus and its standard deviation with other property variables.
- Often, we can observe linear relationships for Bulk Modulus, but the D_Bulk (standard deviation) parameter often displays non-linear behavior
- Thus, it is logical to use a multilinear regression for the Bulk Modulus model.
- The exploratory data analysis (EDA) indicates that equivalent performance for D_bulk cannot be anticipated.

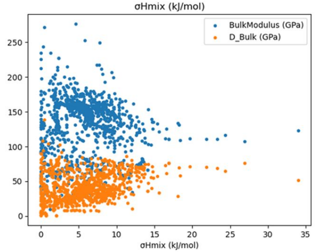


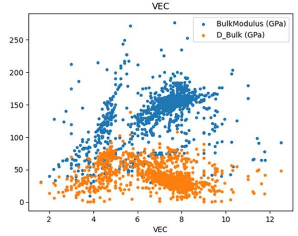


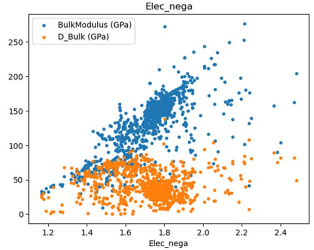


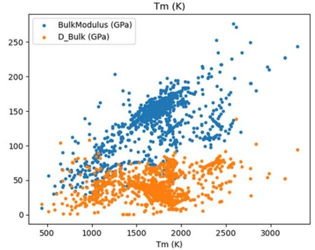


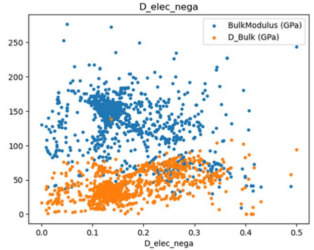


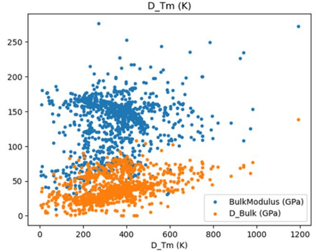








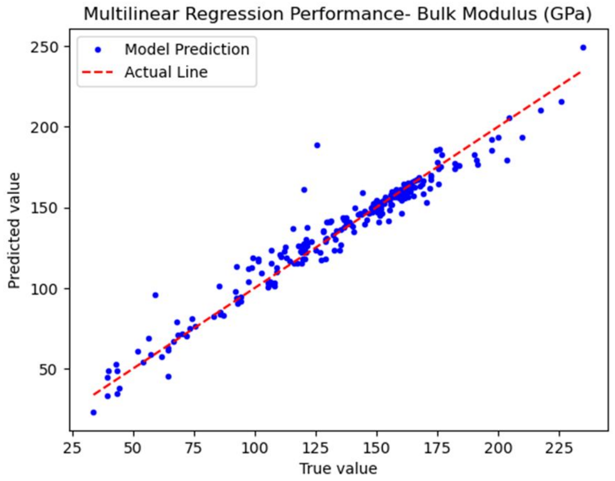


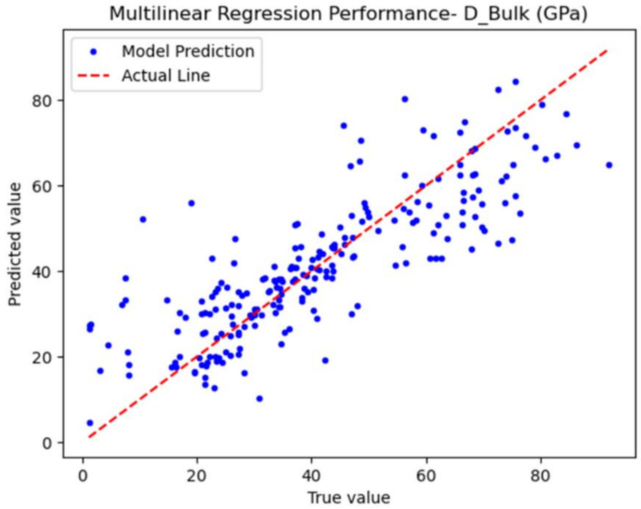


Results & Discussion

MULTILINEAR REGRESSION

- Employed a relatively simple model- multilinear regression using Linear Regression function.
- The results were good- r^2 value of 0.94 for a train-test split of 4:1.
- Performance for D_bulk was worse (as expected)- a r^2 score of only 0.69
- Still, proper feature selection needs to be done in this aspect as the model performance dipped to as low as r^2 = 0.84 when we tried to restrict features.



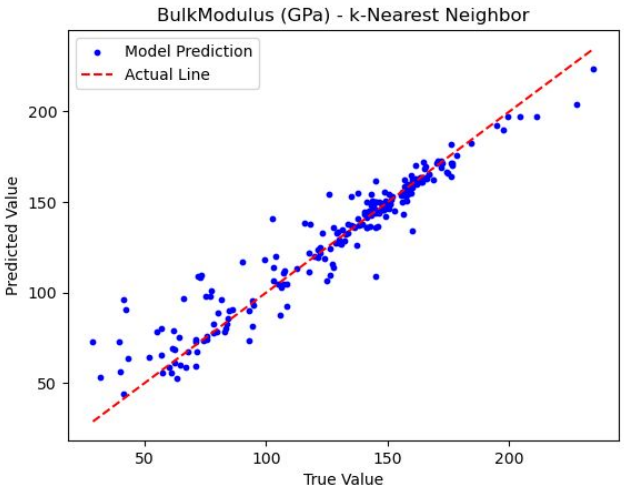


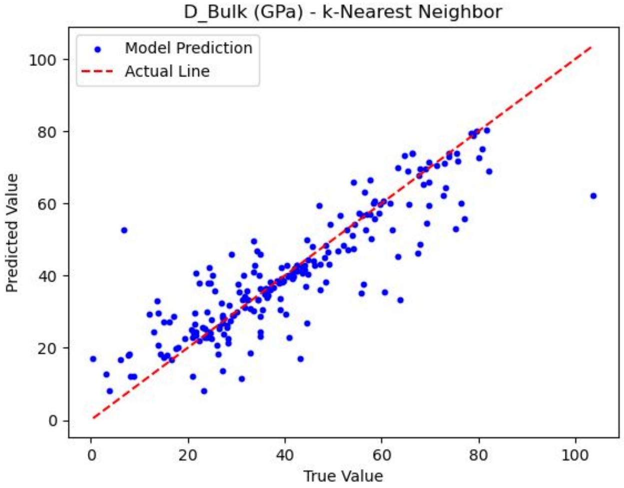
Results & Discussion

k-NEAREST NEIGHBOURS

- The key insight from employing the K-Nearest Neighbors (KNN) model lies in confirming the presence of non-linear relationships among variables.
- Utilized the KNeighborsRegressor function, leveraging its default settings.
- Achieved a test accuracy of 0.84, notably surpassing the performance of the multilinear regression model.

Test accuracy: 0.84
KNN MSE: 115.18
KNN MAE: 6.69
KNN RMSE: 10.73



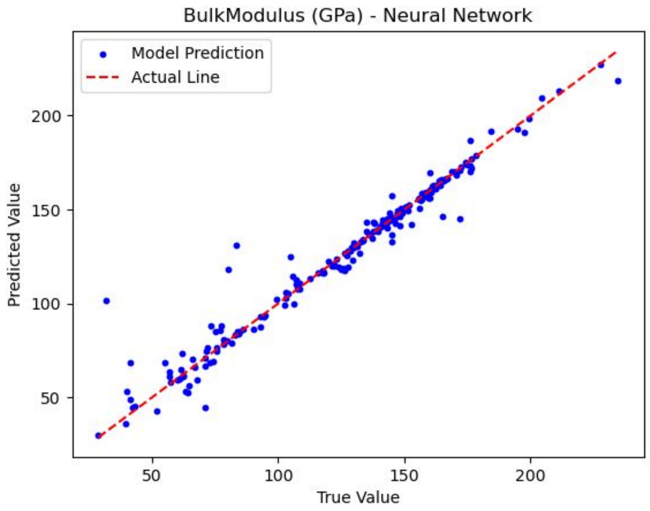


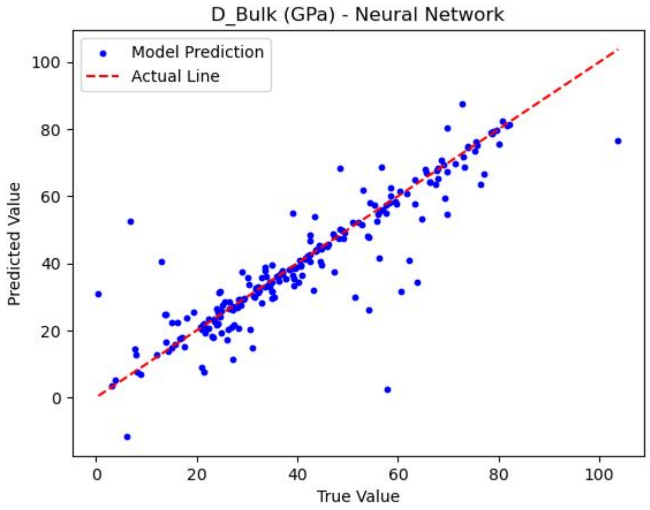
Results & Discussion

NEURAL NETWORK

- The important takeaway from this model is the verification of the non-linear relationship of variables.
- Used the MLPRegressor function, whose default activation function is the ReLu function.
- The test accuracy for this model was 0.88, significantly higher compared to the k nearest neighbour and multilinear regression model
- The RMSE, MSE and MAE values are also significant lower than previous model

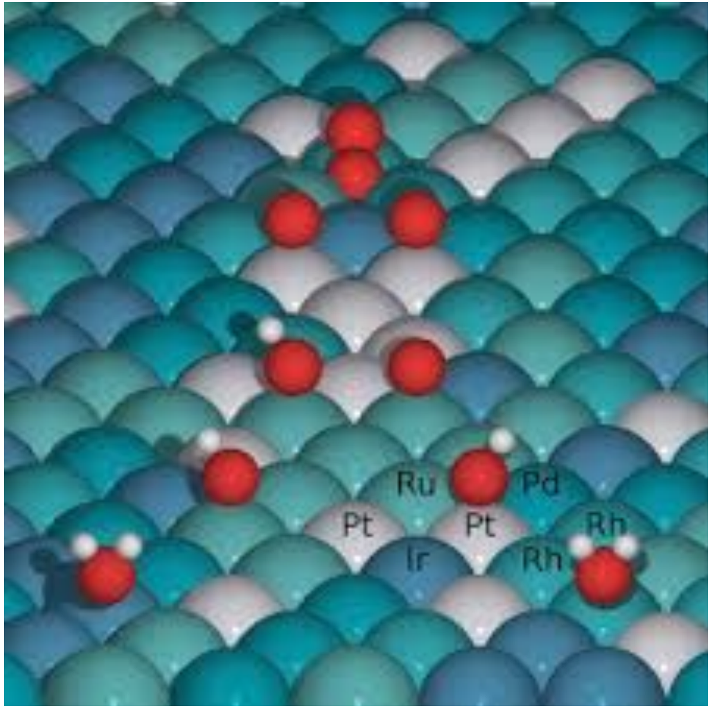
Test accuracy: 0.88
NN MSE: 72.16
NN MAE: 4.12
NN RMSE: 8.49





Conclusion

- Observation of plots for model choice and correlation for feature selection has made the explanation of model differences possible.
- Multilinear regression, K-Nearest Neighbors (KNN), and Neural Networks display promising outcomes in predicting Bulk Modulus, outperforming their performance on D_Bulk.
- Notably, the neural network exhibits superior performance overall.
- The neural network demonstrates notable efficacy with D_bulk, attributed to its capability to capture non-linear relationships.
- Exploratory data analysis (EDA) proves instrumental in guiding model selection.



https://www.cell.com/joule/fulltext/S2542-4351(18)30621-4

Thank you!

Have a nice day!