Predicting Tc in Y123 based superconductors using Machine Learning

Supplementary

In this material a detailed record has been penned about modelling of Tc in Y123 based superconductors. All the required datasets can be found at https://github.com/Abhii3000/Y123.

1. Parameters

Parameters are a critical aspect of training ML models. Based on thorough research, 8 predictors have been selected as mentioned in the manuscript: -

Table 1 Parameters to train ML Models

Parameter	Formula	Description
Chemical Mass	$\Sigma M_i * Coeff_i$	The electron phonon interaction depends on the compound mass
Oxygen Deviation	(7 – Coeff of Oxygen)/7 * 100	Oxygen deviation may change the lattice parameters and length across the c axis which may affect the apical distance affecting the critical temperature. The deviation is calculated keeping 7 as standard because the modeling is being done for the Y123 configuration. The formula is subject to change in future works for more generalized models. Also, it introduces holes, whose density can be a critical parameter.
Avg Ionic Diameter	$\Sigma IonicRadii_i * Coeff_I / S$	To observe the effect of lattice variation due to changes in the size of constituent atoms.
Edited Avg Electronegativity	Σ Electronegativity i * Coeff i/S*	-
Edited Avg Number of Valence electrons	Σ Number of Valence electrons$_i$ * Coeff$_I$/S*	-
Avg Specific-Heat	$\Sigma Specific\ Heat_i * Coeff_I / S$	-
Edited Avg First Ionization	$\Sigma Ionization\ Potential_i * Coeff_I / S$*	-

Layers	n (number of CuO2 layers present)	As CuO2 is the conducting layer, the number of these layers might play an important role. But in the current model, it cannot be incorporated as all cuprates have 2 layers and it can't be a differentiating parameter. But in more generalized models it absolutely will be an important parameter.
%Crystalline character of material of each class (cubic etc.)	$\frac{(\Sigma \text{ (Coeff of ith element if it belongs to a given class a given structure))} * 100}{\Sigma \text{Coeff}_i}$	Intuitional parameter
Structure score	ΣW_c*(%Crystalline character of material of each class)	To minimize the dimensionality of the model

where, M is atomic mass of the element

S = ΣCoeff_i, S=ΣCoeff_i (except Oxygen and Copper)

i signifies the ith element present in the compound

Coeff is the coefficient related to the element in the compound

c determines each crystal class (cubic, hexagonal, monoclinic,

orthorhombic, tetrahedral, trigonal and triclinic)

W is importance of each class of crystal obtained from model1

*The parameters where Edited Avg have been taken, it means that all elements in the composition have been considered except oxygen and copper

These parameters can be direct or indirect. To create those datasets following sources were used

- Superconducting Material database with their Tc openly available by khamidieh
- Atomic and Ionic radii from crystalmaker
- Electronegativity, First Ionization Potential, Specific Heat and Number of Valence electrons from GoodmanSciences
- Crystal structure of elements from Periodic Table

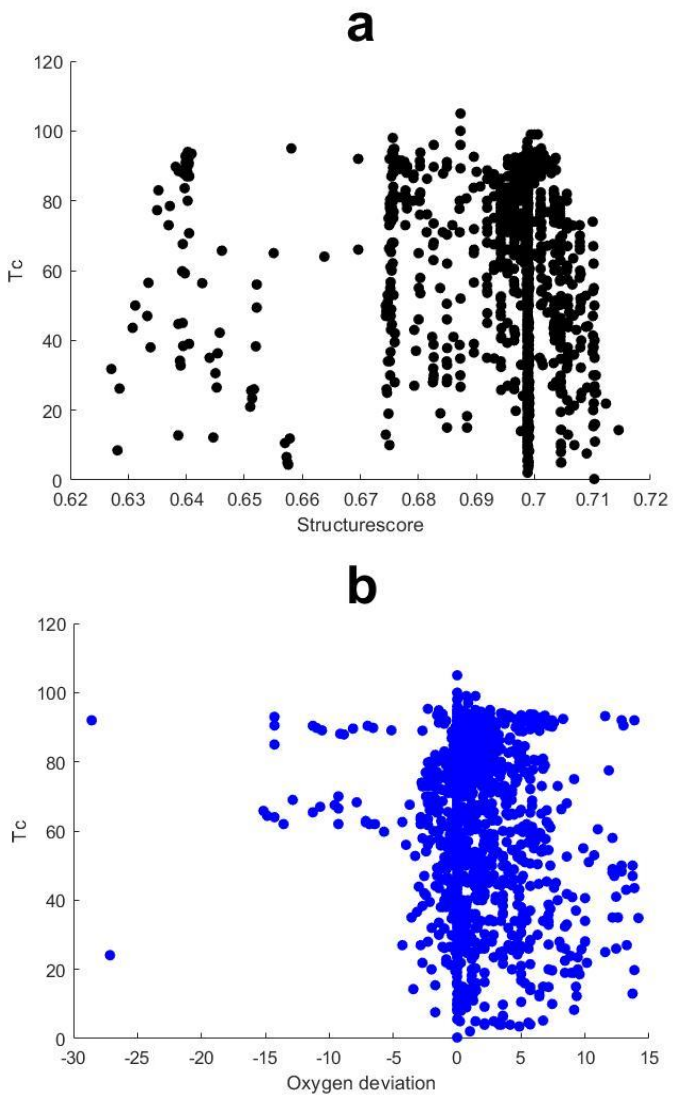
While creating the valency parameter it was found that in dataset values for number of electrons were missing for transition elements and actinide series. So they were added based on the explanation showed in https://www.youtube.com/watch?v=0JNIYDtfXzw and has been tabulated in Table 2.

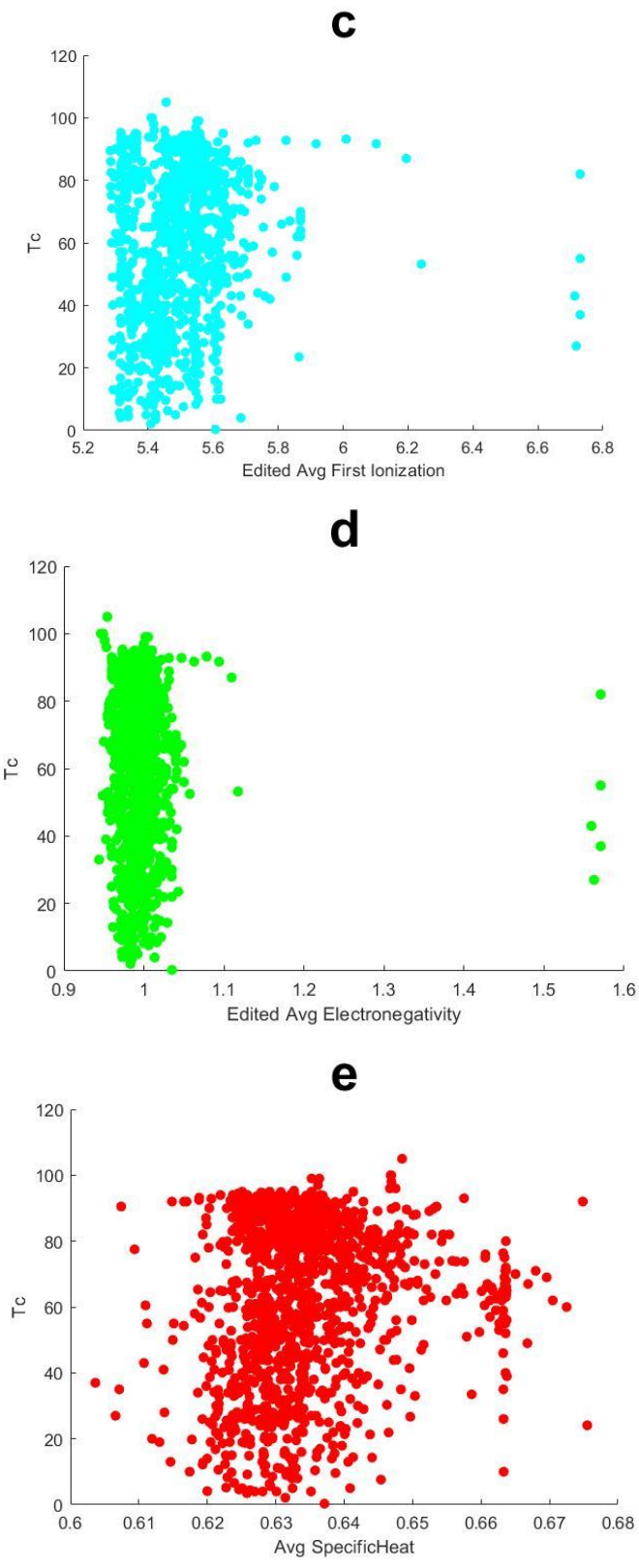
Table 2 Number of valence electrons considered for modelling

Element	Number of Valence electrons
Hydrogen	1
Helium	2
Lithium	1
Beryllium	2
Boron	3
Carbon	4
Nitrogen	5
Oxygen	6
Fluorine	7
Neon	8
Sodium	1
Magnesium	2
Aluminum	3
Silicon	4
Phosphorus	5
Sulfur	6
Chlorine	7
Argon	8
Potassium	1
Calcium	2
Scandium	3
Titanium	4
Vanadium	5
Chromium	6
Manganese	7
Iron	8
Cobalt	9
Nickel	10
Copper	11
Zinc	12
Gallium	3
Germanium	4
Arsenic	5
Selenium	6
Bromine	7
Krypton	8
Rubidium	1
Strontium	2
Yttrium	3
Zirconium	4

Niobium	5
Molybdenum	6
Technetium	7
Ruthenium	8
Rhodium	9
Palladium	10
Silver	11
Cadmium	12
Indium	3
Tin	4
Antimony	5
Tellurium	6
Iodine	7
Xenon	8
Cesium	1
Barium	2
Lanthanum	3
Cerium	4
Praseodymium	5
Neodymium	6
Promethium	7
Samarium	8
Europium	9
Gadolinium	10
Terbium	11
Dysprosium	12
Holmium	13
Erbium	14
Thulium	15
Ytterbium	2
Lutetium	3
Hafnium	4
Tantalum	5
Wolfram	6
Rhenium	7
Osmium	8
Iridium	9
Platinum	10
Gold	11
Mercury	12
Thallium	3
Lead	4
Bismuth	5

Polonium	6
Astatine	7
Radon	8




c
Tc
Edited Avg First Ionization
d
Tc
Edited Avg Electronegativity
e
Tc
Avg SpecificHeat

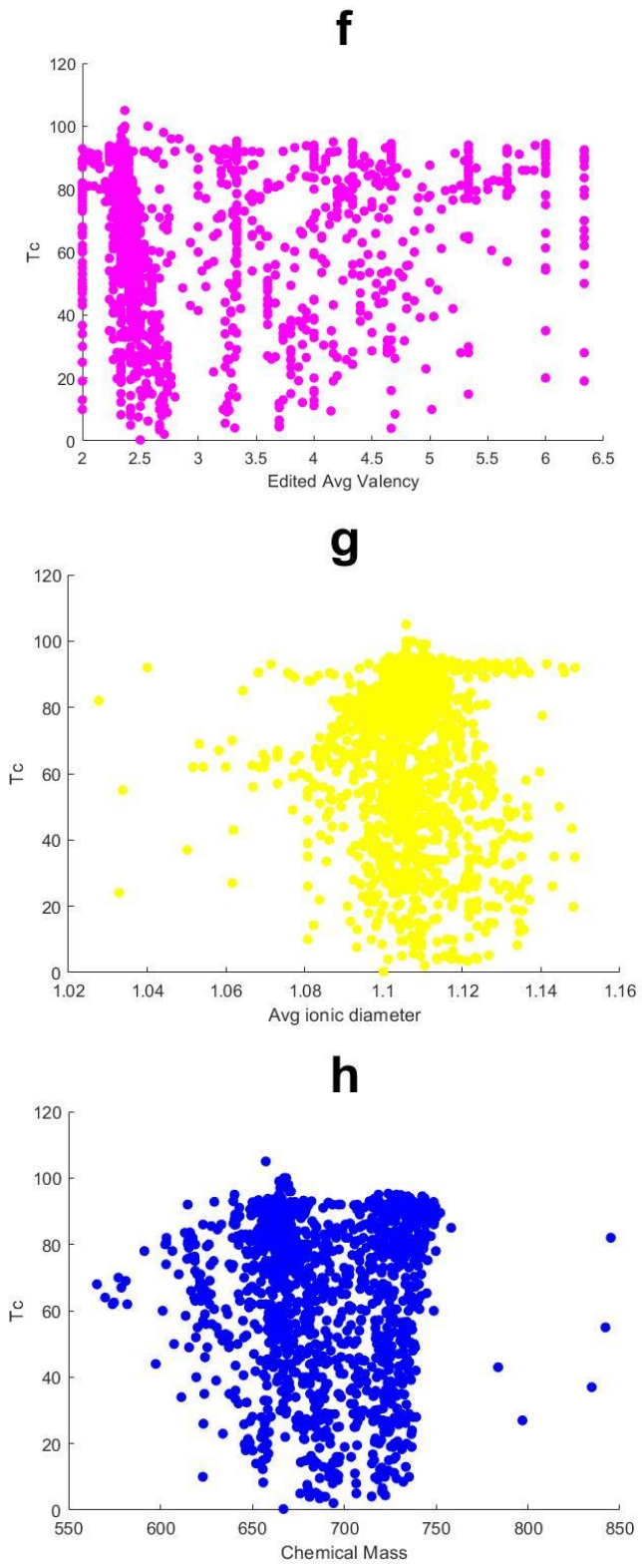


Fig. 1 Predictors vs Tc: **a** Structurescore **b** Oxygen deviation **c** Edited Avg First Ionization **d** Edited Avg Electrnegativity **e** Avg SpecificHeat **f** Edited Avg Valency **g** Avg Ionic Diameter **h** Chemical Mass

2. Workflow

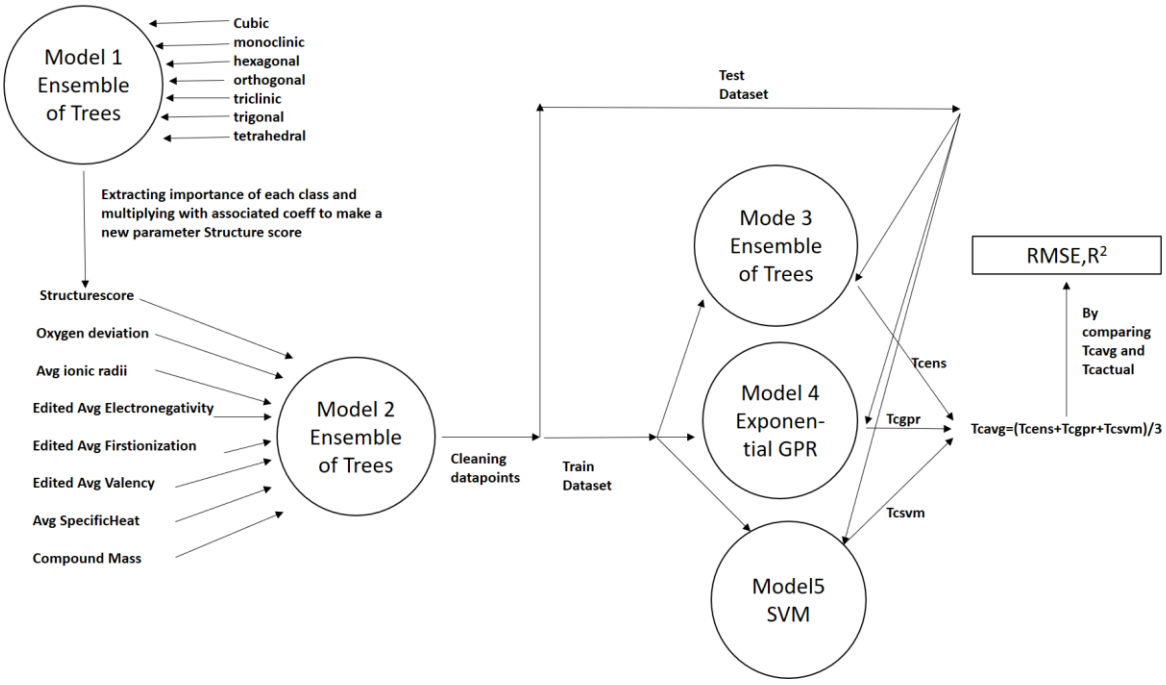


Fig. 2 Flowchart of training of the Model

Once the predictors were prepared, it was time to train the model. MATLAB Regression Learning Toolbox was used to train the model. The steps are as follows

1. Initially, only the 7 crystal classes were used to train a bagged ensemble tree from where the coefficients of the importance of the parameters were taken. Then a new parameter structurescore was designed.
2. Now structurescore along with 9 other parameters was used to train a bagged ensemble tree with "5" fold cross validation.
3. The training data may contain some wrong datapoint due to errors during experimentation. So a cleaning process was done.
4. The newly formed dataset was split into training and test sets in a 17:3 ratio.
5. The training set was then trained with 3 models (Bagged ensemble tree, SVM, and GPR) with "5" fold cross validation.
6. Tc was calculated for all three models in the Test set.
7. Then a new Tc was calculated by taking avg of predicted Tc of all the three models

   $$Tc\_avg=(Tc\_ens+Tc\_gpr+Tc\_svm)/3$$
8. R^2 was calculated for predicted Tc(Bagged ensembled tree, SVM, GPR and Avg).
9. The importance of all predictors was noted down from the ensemble model.

3. Training

While training **Model 1**, The dataset was named **Y123**. Associated predictor value of different crystal classes were calculated. The model was ensemble of tree based (bagged). Minimum leaf size was set to **1** and Number of leaves were set to **30**. R^2 was found to be 0.27.

Table 3 Importance of crystal classes obtained from training Model 1

Crystal Class(Predictors)	Importance
Hexagonal	0.95572
Monoclinic	0.692642
Cubic	0.656
Trigonal	0.197797
Orthogonal	0
Triclinic	0
Tetrahedral	0

As mentioned in the workflow, new predictor Structurescore was created. Along with it other 7 parameters were added.

Table 4 Alias used for parameters used in ML models 1,2,3,4 and 5

Parameter	Alias
Chemical Mass	Mass
Oxygen Deviation	Odev
Avg Ionic Diameter	
Edited Avg Electronegativity	Electronegativity
Edited Avg Number of Valence electrons	VE1
Avg Specific-Heat	SpecificHeat
Edited Avg First Ionization	FirstIonization
%Crystalline character of material of each class (cubic etc.)	cubic monoclinic hexagonal orthogonal triclinic trigonal tetrahedral
Structure score	structurescore

With the above parameters Y123 was used to train Model which was ensemble of tree based (bagged). Minimum leaf size was set to **1** and Number of leaves were set to **30.** With Model 2 Tc was predicted for Y123 dataset and Tc actual vs Tc predicted graph was plotted.

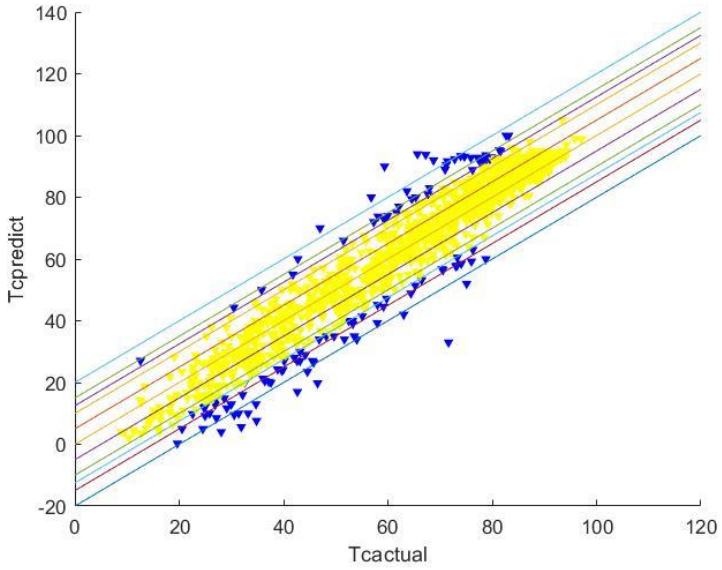


Fig 3. Tc actual vs Tc predict for dataset Y123

It was observed within ΔTc (Tc actual – Tcavg) of +/- 12.5K most of the data points were present(points marked in yellow in Fig). These points were then screened as a new dataset New as a cleaning process and points marked in blue were considered outliers.

Dataset New was then split in 2 sets Test and Train in 3:17 ratio. The Train set was then used to train Model3,4 and 5. Model 3 is an ensemble of trees(bagged) model with minimum leaf size **1** and number of leaves were **30**. Model 4 is an exponential GPR model with parametrics as shown in Fig 4.

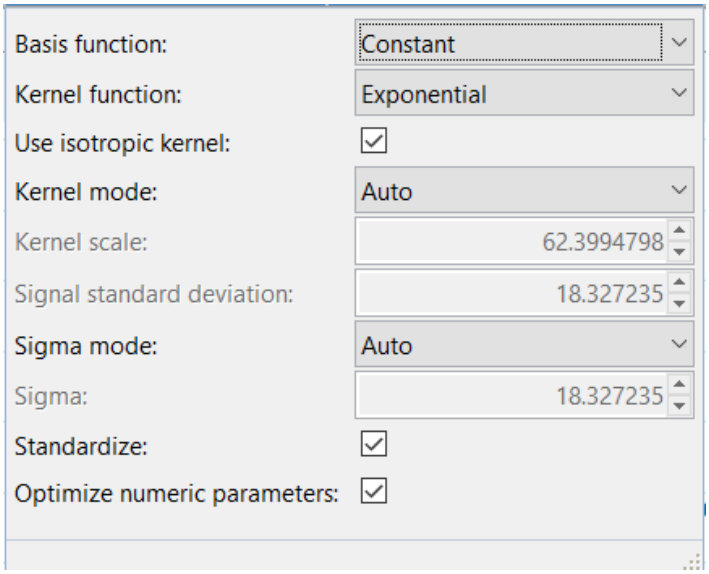


Fig. 4 Parameters in Model 4

Model 5 ias a Fine Gaussian SVM based model with parameters as shown in Fig

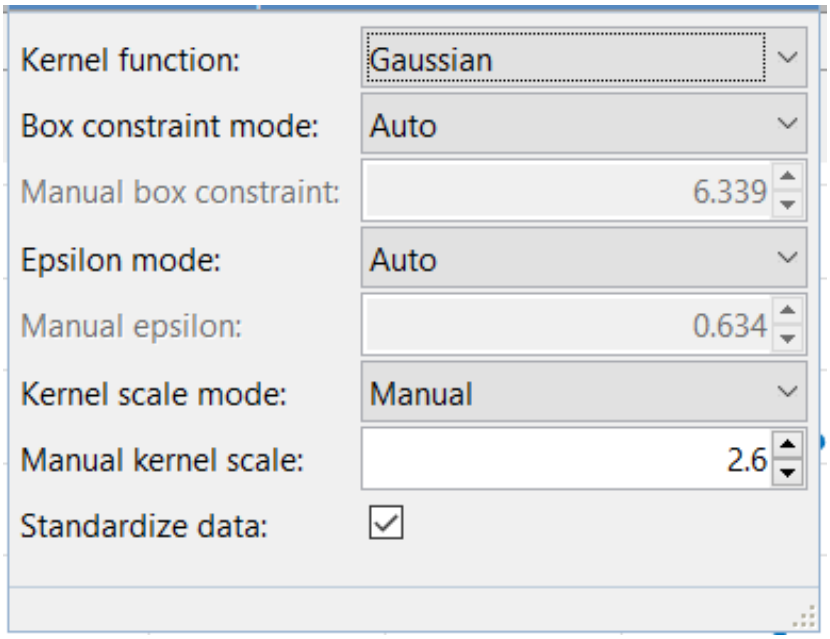


Fig. 5 Parameters in Model 5

Table 5 Importance of predictors obtained from training Model 3

Predictors	Importance
Edited Avg First Ionization	0.4533
Edited Avg Valency	0.4091
Structurescore	0.2975
Avg SpecificHeat	0.2793
Chemical Mass	0.248
Avg Ionic Diameter	0.2256
Oxygen deviation	0.2159
Edited Avg Electronegativity	0.1789

The R^2 for the Tcavg as obtained when applied on Test dataset was found to be 0.8365 and the RMSE was found to be 9.5K. The decision to obtain an avg temperature seems to have paid off as the avg helps to balance the biases as the metrics of individual models on Training set were poorer. They have been tabulated in table 6-

Table 6 R^2 and RMSE obtained from Train set for Models 3,4,5

Model	R^2	RMSE (K)
Ensemble of Trees(Model 3)	0.75	11.6
SVM (MODEL 4)	0.78	10.49
GPR (MODEL 5)	0.80	9.53

Table 7 Tc ctual, Tcavg, Tc predicted by Model 3,4,5 for few compositions from Test dataset

Material	Tc actual (K)	Tcgpr (K) GPR	Tcsvm(K) SVM	Tcens (K) Ensemble of Trees	Tcavg (K)
Y0.5Pr0.5Ba2Cu3O6.965	10	18.04804	20.79692	22.16209	20.33568
La0.75Dy0.25Ba1Ca1Cu3O7	56	64.86275	63.02022	57.61713	61.83337
Er0.85Hf0.15Ba2Cu3O6.813	79.9	81.48778	82.02405	80.86239	81.45807
Y0.7Pr0.3Ba2Cu3O6.95	62.5	58.64661	58.47753	54.40683	57.17699
Lu0.8Ca0.2Ba2Cu3O6.68	78	64.73108	67.77049	63.93945	65.48034

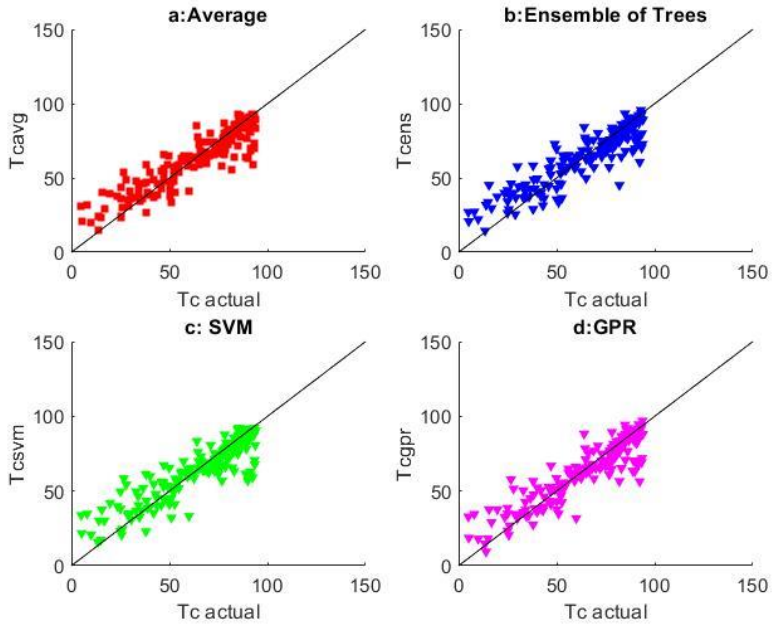


Fig. 6 Tc actual vs Tc predicted: **a** Average **b** Ensemble of Trees **c** SVM **d** GPR

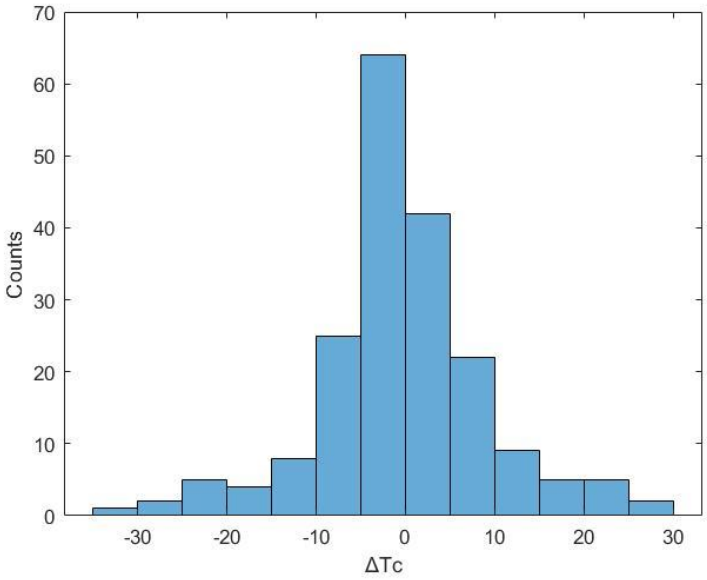


Fig. 7 Histogram for ΔTc (Tc actual – Tcavg) for Test dataset

Table 8 Tcavg, Tc predicted by Model 3,4,5 for random compositions with dataset exp

Material	Tcgpr (K) GPR	Tcsvm(K) SVM	Tcens (K) Ensemble of Trees	Tcavg (K)
Y0.2Pr0.2Ca0.2Nd0.2Sm0.2Ba2Cu3O7	73.65973	61.85139	54.53992	63.35035
La0.2Pr0.3Ca0.25Gd0.25Ba2Cu3O7	74.19799	61.77571	57.32989	64.43453
Nd0.2Pr0.3Ca0.3Gd0.2Ba2Cu3O7	75.73984	61.79782	67.50894	68.34887
Y0.5Gd0.25Tc0.25Ba1.975La0.025Cu3O6.95	68.87614	61.77975	72.3492	67.66836
Y0.2Sm0.2Nd0.2Eu0.2Dy0.2Ba1.95La0.05Cu3O6.9	67.21644	63.07869	51.27799	60.52437