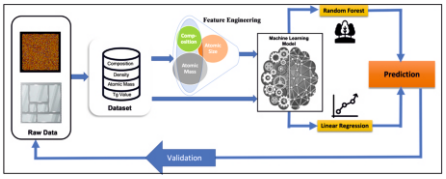
Applications of Artificial Intelligence

4

Predicting Properties of Glasses using Machine Learning Algorithms

Pooja Sahu, Vishwas Tiwari and Sk. Musharaf Ali*

Chemical Engineering Division, Bhabha Atomic Research Centre, Trombay – 400085, INDIA



Typical application of machine learning

ABSTRACT

Prediction of suitable glass composition is very important and also challenging. The glass can have the possibility of having large number of probable combination of compositions, which makes it quite challenging to apply the experimental method of trial and error. The computational tools like Ab initio and Classical Molecular Dynamics simulations have very high computational cost. Machine learning (ML) is a very promising tool for predicting the properties of glass with sufficient efficiency and at a low computational cost. The ML models were built by applying feature engineering and varying the descriptors for higher accuracy. The two main ML models implemented for investigation are Linear Regression and Random Forest. We investigated two types of glasses: Sodium Borosilicate glass and Radiation shielding window (RSW) glass. The different models were built for different properties (T_g value, density, and Young's modulus) and then optimized for higher accuracy.

KEYWORDS: *Classical Molecular Dynamics, Machine learning, Radiation shielding window, Young's modulus*

Introduction

Apart from many other important uses, glasses are used for immobilization of radioactive waste and radiation shielding. In that context, tailor made glass is necessary. Glasses are known as supercooled liquid which are non-equilibrium, non-crystalline material that spontaneously relax at room temperature [1] . The glass does not require to satisfy any rigid stoichiometric rules. Glasses can be made of any element present in the periodic table if quenched fast enough from the melt state to solid state. Zanotto et al. [2] showed that there is a possibility of forming 10^{52} different glass compositions with 80 most useful chemical elements of the periodic table when combining them in 1.0 mol %. The experimental method of trial and error to make the glass with desired properties is expensive as well as time consuming. Computational tools like ab initio and classical molecular dynamics simulations can be used as an alternative to reduce the experimental expenses, but these computational tools have certain limitations like this can be precisely applied only for simple glass compositions containing maximum 5-6 elements. At least one simulation is required for each composition of the glass which increases the computational cost because there are a large number of compositions available.

The Machine Learning (ML) model can be used to address the complex problem in the material science by using the existing information about the glasses [3]. The ML models are built on the concept of learning from the available database. In order to use the ML algorithms to predict the properties of the new glasses, it is essential that the developed model should have high predicting accuracy. The model accuracy mostly relies on the existence of the useful data which are accurate, consistent and complete. There are many ML algorithm which can be utilized to build the model. Previous study by Cassar et al. [4] reported a successful application of a Multilayer Perceptron (MLP) artificial neural network (ANN) to predict the glass transition temperature (T_g) of multicomponent oxide glasses containing over 46 chemical elements. Further, Alcobaca et al. [5] showed application of different ML models to predict the T_g of glasses containing over 65 chemical elements, where the random forest (RF) and k-nearest neighbors algorithm (K-NN) model had the highest accuracy as compared to other models. In these studies, the relative deviation (RD) at extreme T_g (450K<; 1150K >) was shown to be higher as compared to the intermediate T_g (459K<= T_g>= 1150K) range. The prediction of Young's modulus values was done by Yang et al. [6] using high-throughput molecular dynamics simulations and ML models to infer the relationship between glass composition and Young's modulus.

This letter reports the different ML models to predict the T_g value, Young's modulus and density of the glasses, with the help of different features (% of compositions, atomic mass, atomic size and density) as input. Among the used ML models i.e. linear Regression [7] (LR) and random forest (RF) [8] , the RF model was found to have better accuracy.

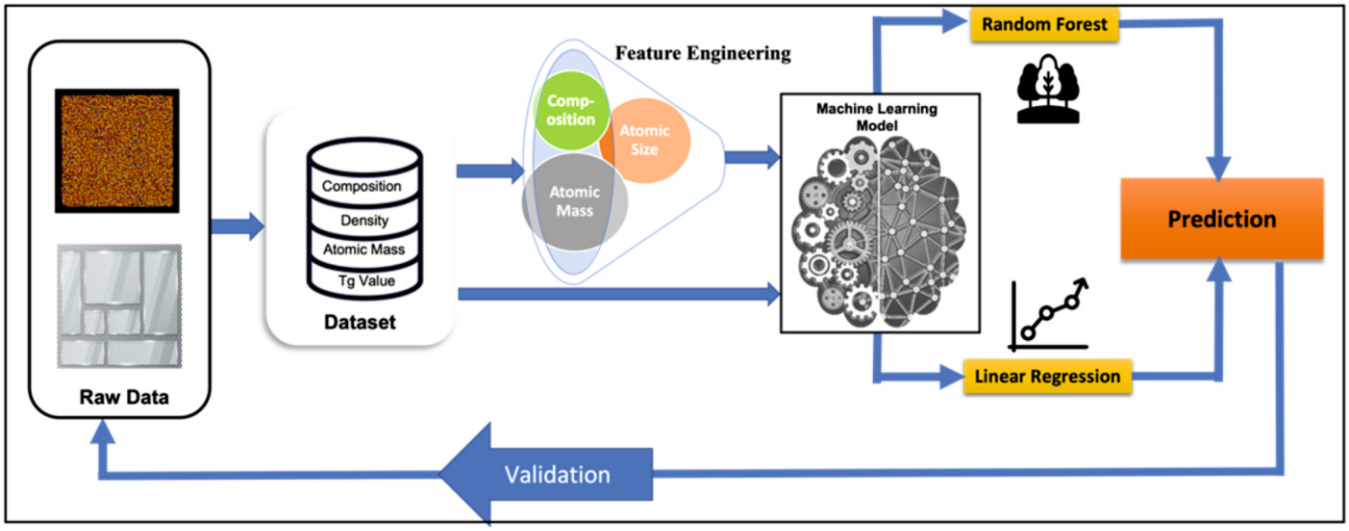
Computational Methodology

A schematic for ML application for predicting the glass properties is shown in Scheme 1.

Dataset

For both the sodium borosilicate and radiation shielding window (RSW) glasses, the dataset used was collected from the SciGlass database, which contains 4,20,000 glasses and 2,68,000 oxide glasses. We limited our dataset to different compositions of SiO_2, K_2O, Na_2O, PbO, LiO, BaO and As_2O_3. We excluded all the compositions containing any other chemical element. In the selected dataset, the range of T_g was varied from 673K-1173K, the density was ranged from 1.8 g /cm^3 - 9g/cm^3 and Young's modulus was ranged from 40GPa to

*Author for Correspondence: Dr. Sk. Musharaf Ali
E-mail: musharaf@barc.gov.in



Scheme 1. Illustration of typical application of machine learning.

90GPa. The dataset after cleaning was reduced to greater extent T_g – 1349, Young's Modulus – 659, Density – 1445. The duplicate data was removed, and the median was selected for the clean dataset. The elemental compositions are presented in Fig. 1.

Machine Learning Algorithms and Evaluation

The ML algorithms used in present studies are linear regression and random forest. In most of the studies, the RF model has been used because it gives good predication that can be understood easily, and it is one of the few models which can perform both regression and classification tasks. The performance of the used ML models was checked by evaluating the coefficient R^2.

Training and Evaluation Setup

The Feature Engineering [9] is an essential phase of developing machine learning models and is performed on the dataset as per the requirement. A feature is defined as a unique attribute or variable in a dataset. Feature engineering helps to improve the performance of machine learning model by selecting the right features for the model and preparing the features in a way that is suitable for the machine learning model. The step used in feature engineering were data cleansing, data transformation, feature extraction, feature selection.

The dataset was randomly divided into 80:20 ratio. The 80% of data was taken for the training the models and the 20 % for the testing the models. Then the models were built by using default Machine Learning Algorithms and then were modified to improve the accuracy of the models. The modification were made by changing the number of descriptors [10] as input data and adjusting the number of decision tree [11] in the models. Then the model was saved and used to predict the unknown data to check the accuracy of the model for unknown

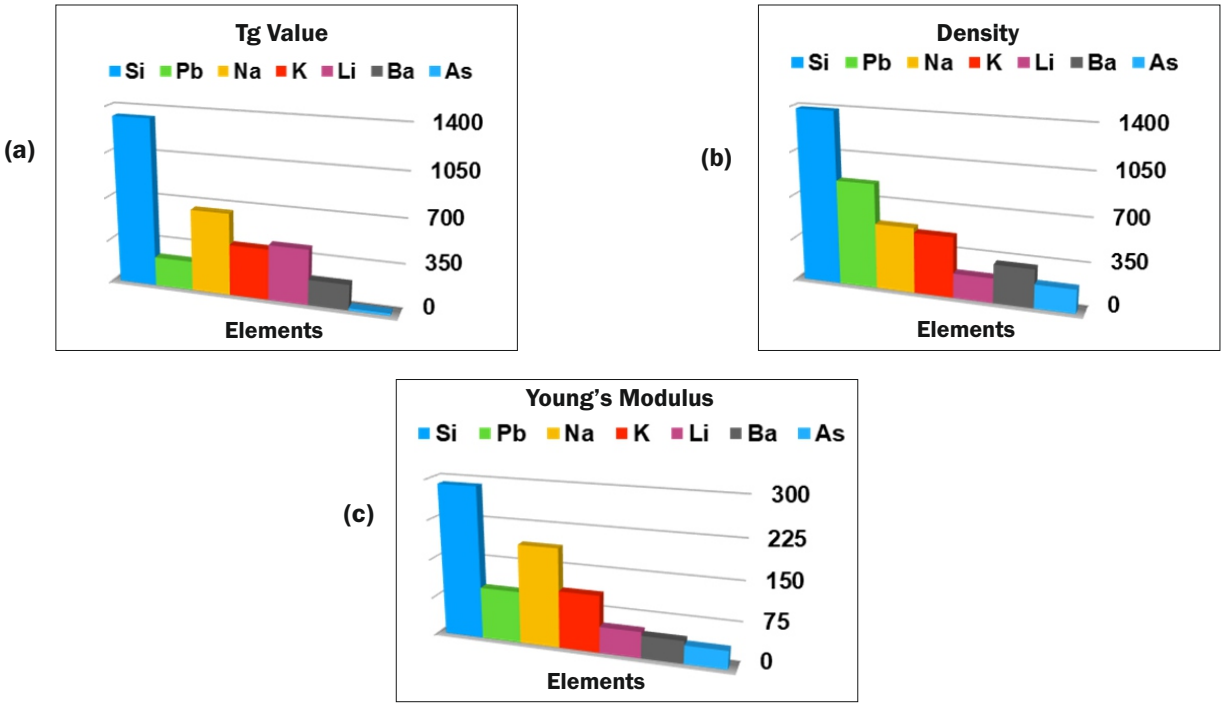


Fig.1: Number of compositions containing each elements in the clean dataset (a) Density, (b) T_g and (c) Young's Modulus. (Note: Oxygen is present in all glasses of this dataset).

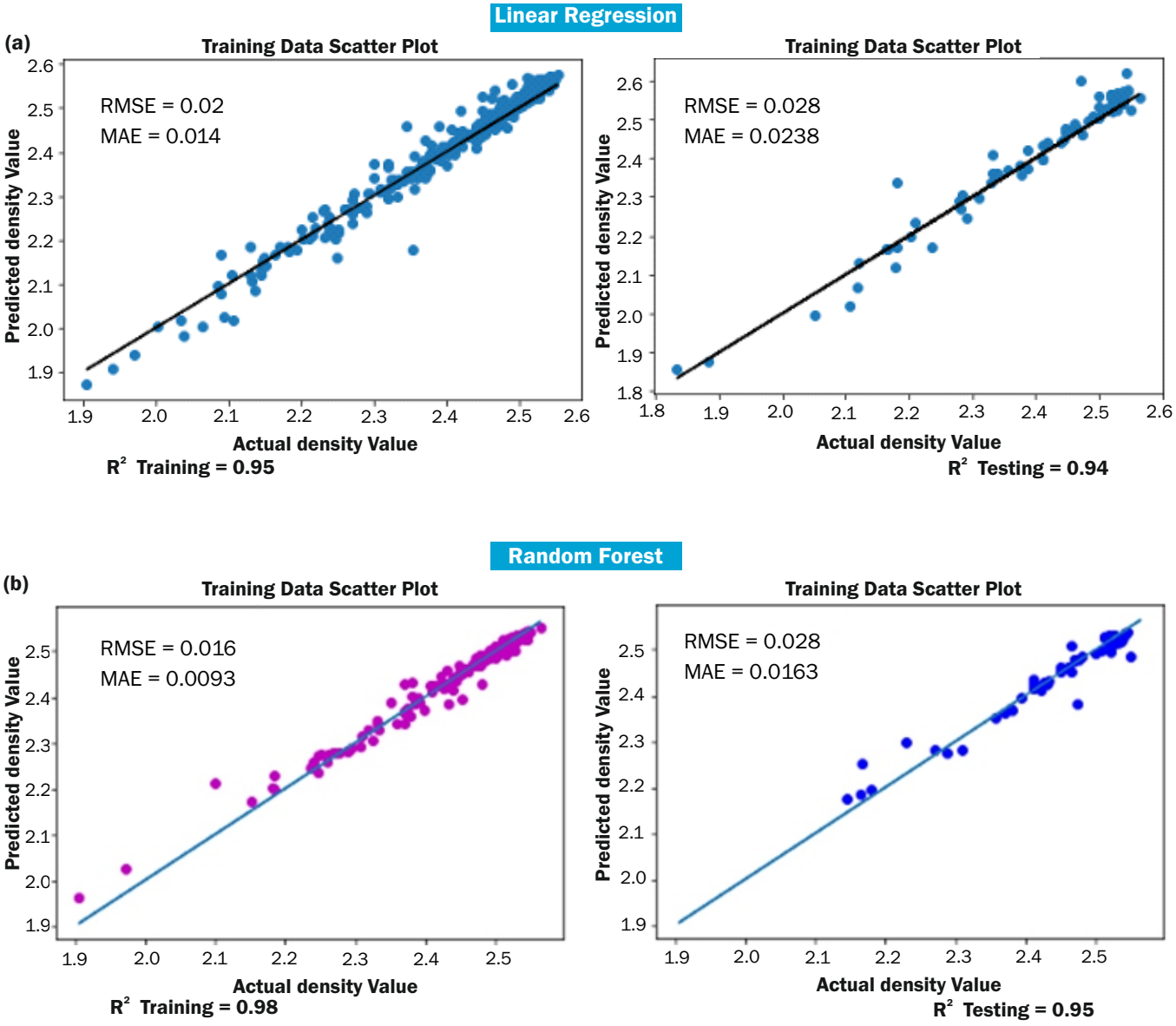


Fig.2: Scattering plot of actual Density vs. predicted density of training and the testing data. (a) Linear Regression algorithm. (b) Random Forest Algorithm.

composition. Then we selected the best ML model for each properties that were calculated. All the studies were performed using the Python programming language with the sklearn, numb, pandas, matplot, and pickle libraries available in Anaconda-3.0.

Results & Discussion

Sodium Borosilicate Glasses

Density

The density prediction was done using glass composition as input feature and no additional descriptors were considered. Both Linear regression and the Random Forest model had given accurate results with R^2 Training and R^2 Testing of 0.95 and 0.94 values for linear Regression and 0.98 and 0.95 for Random Forest model (see Fig.2). The Random Forest method gives better R^2 as compared to Linear Regression Model with percentage error of 0.7. Very close values of glass densities from MD simulation and from ML model in Table 1 for unseen glass compositions by ML model (either for training or testing), show the accuracy of RF model for prediction of glass density of sodium borosilicate glasses.

Table 1: Comparison of density from MD simulation and ML model for sodium borosilicate glasses.

Sr. No.	Glass composition			MD estimated density (g/cm³)	ML predicted density (g/cm³)
	B_2O_3	Na_2O	SiO_2		
1	10	10	80	2.39	2.39
2	12.5	12.5	75	2.44	2.43
3	15	15	70	2.48	2.48
4	50	10	40	2.17	2.20

Glass transition temperature (T_g)

The linear regression model was not suitable for T_g as R^2 value was less than 0.8. Therefore, Random Forest model was used for the prediction of the T_g value. The RF model was first tested with single input descriptor i.e. composition and then by using two descriptors - composition and density. The RF model with single descriptor has R^2 training and testing equal to 0.98 and 0.96 respectively as displayed in Fig.3. This model was

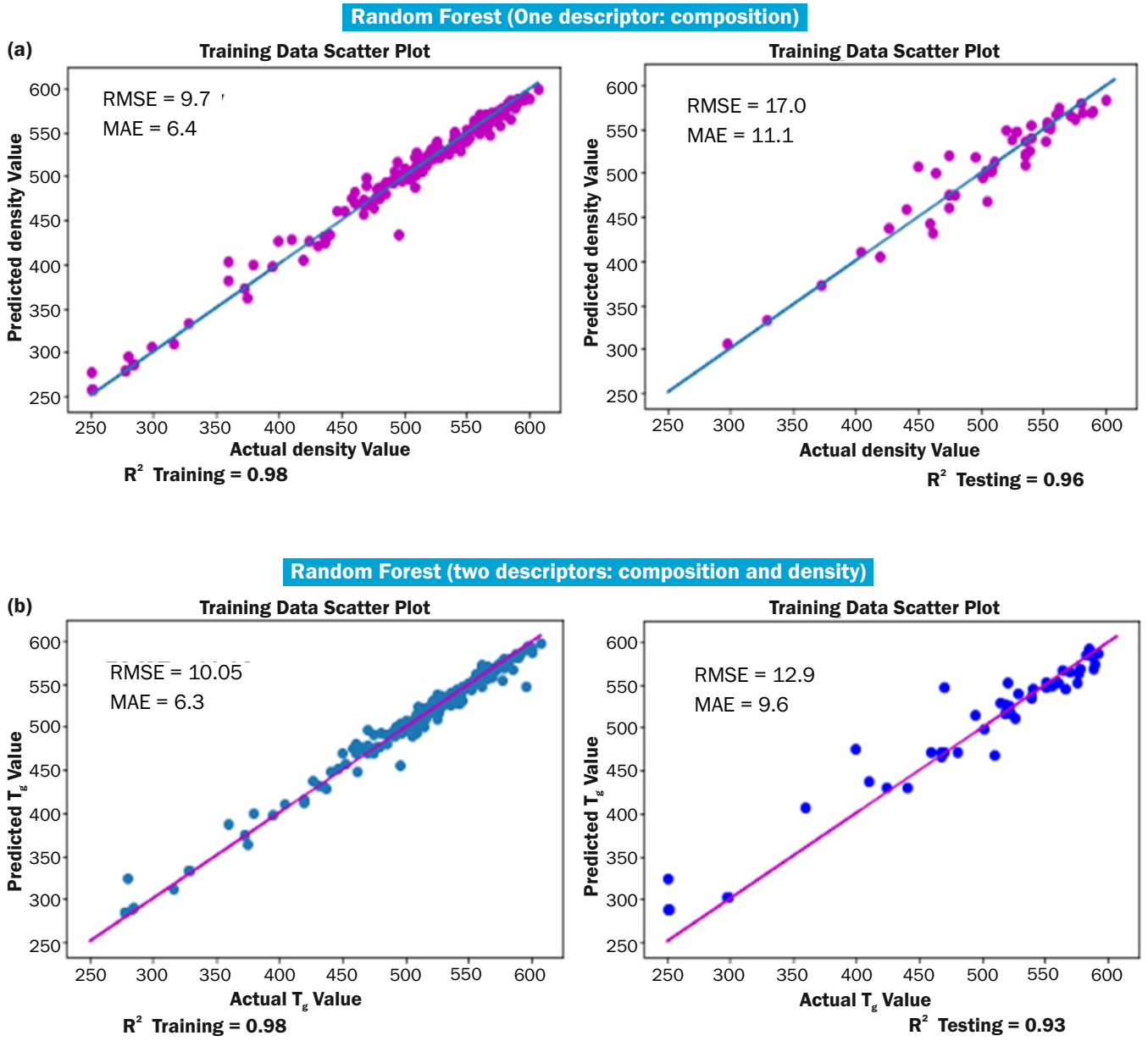


Fig.3: Plot of actual vs predicted T_g of training and testing using Random Forest Model. (a) Model with only one descriptor (composition), (b) Model with two descriptors (composition and density).

able to predict the T_g of unknown compositions with error of 20 ± 0.01 K. The second model, with two descriptors has R^2 training and testing equal to 0.98 and 0.93 respectively. In spite of lower testing R^2, the second model was able to predict the unknown data with lesser error of 15.8 ± 0.01K (data shown in Table 2). Hereby, it can be remarked that the accuracy of model would increase significantly with the addition of the density descriptor as input feature.

Radiation Shielding Window (RSW) Glass

Density

The density model for radiation shielding window (RSW) glass build with one descriptor i.e. composition, which gave R^2 training and testing of 0.99 and 0.93 respectively as shown in Fig.4. The predicted accuracy was 96% for unknown glass compositions, data reported in Table 3. The RF was seen to optimize at 50 trees as shown in Fig.5. The estimated RMSE and MAE are within acceptable range for density.

Table 2: Comparison of T_g from MD simulation and ML model for sodium borosilicate glass.

Sr. No.	Glass composition			MD estimated T_g (°C)	ML predicted T_g (°C)
	B_2O_3	Na_2O	SiO_2		
1	30	14	56	530	541
2	25	10	65	511	506
3	20	12	68	570	567
4	15	15	70	590	586

Glass transition temperature (T_g)

To predict T_g, RF model was used with three different number of descriptors. The first one with single descriptor of composition, showed R^2 training and testing equal to 0.89 and 0.64 respectively as shown in the Fig.6(a). Further, we added

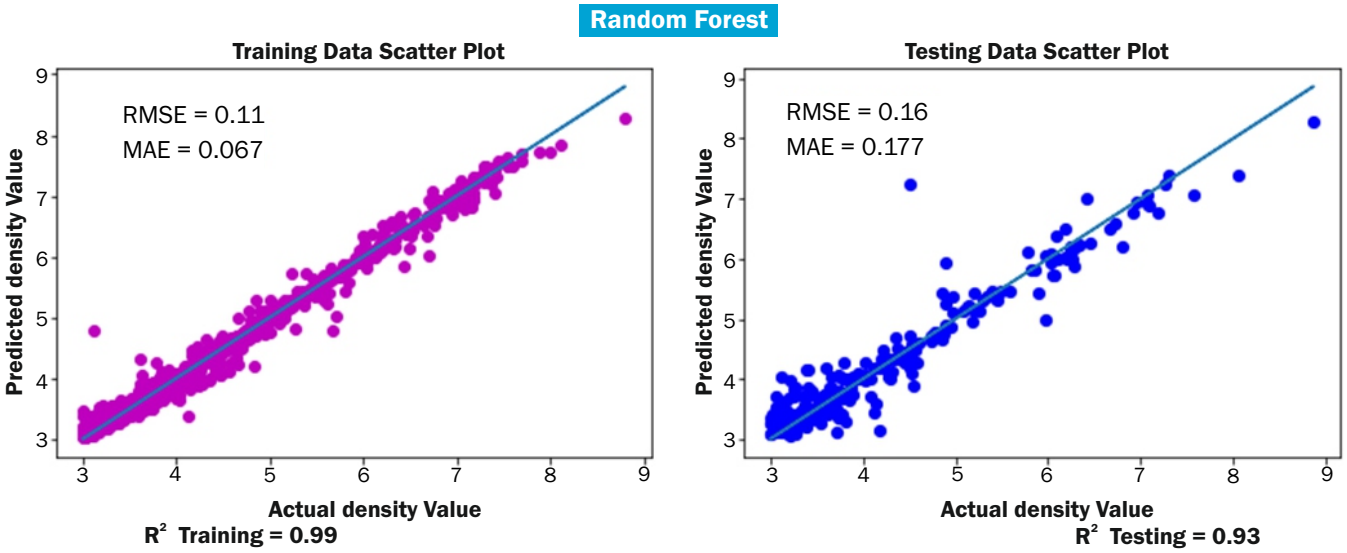


Fig.4: Scattering plot of actual density value vs predicted density value of training and the testing using Random Forest Model. Note – RMSE (Root Mean Square Error) and MAE (Mean Absolute Error).

Table 3: Comparison of density from MD simulation and ML model for RSW glass.

Sr. No.	Glass composition				MD estimated density (g/cm³)	ML predicted density (g/cm³)
	SiO_2	K_2O	Na_2O	PbO		
1	84	16	0	0	3.32	3.32
2	67	0	33	0	3.07	3.10
3	51	0	0	49	5.91	5.88
4	64.8	1.8	0	33.4	4.60	4.50

Comparison of density from MD simulation and ML model [using RF method with composition as descriptor] for RSW glass.

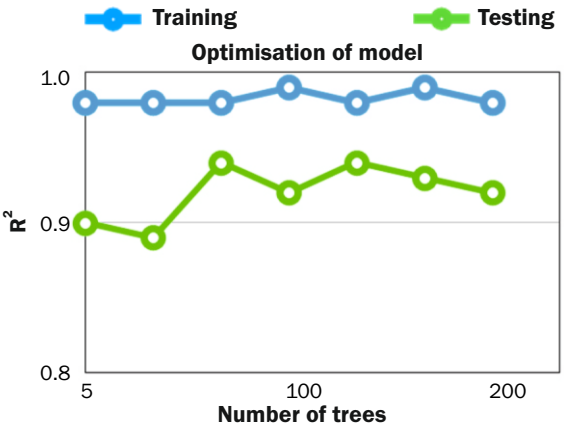


Fig.5: The accuracy (as captured by the R^2 value) of the Random Forest model as a function of number of trees considered in each model as obtained for the training and testing set respectively.

atomic mass descriptor to composition and found that the R^2 training and testing was decreased to 0.88 and 0.61 respectively (see Fig.6(b)). R^2 testing was furthermore reduced to 0.42 (with R^2 training as 0.92), while including 3rd descriptor of atomic size as shown in the Fig.6(c). Nevertheless, the prediction of T_g for unknown dataset made using RF model with three descriptors showed sufficient close to the MD estimated T_g. (see Table 4). Notably, the estimated RMSE and MAE values are higher in case of RSW glass compared to NBS-Glass due to higher data range [5]. It might be noteworthy to mention that the sole-observation of the mean error values (RMSE, MAE) and R^2 values can't give enough evidence for the statistical accuracy of model for prediction than its competitors with higher values of mean error bar and R^2. Similar to studies of Alobaca et al. [5], we found that in spite of higher mean error and low R^2 testing, our model predicts good match of T_g with MD estimated data as shown in Table 4. The R^2 training and testing as a function of number of trees (for 3rd model) is shown in Fig.7, where model seem to be optimized with 5 trees for training and 150 trees for testing.

Young's Modulus

Similar to previous case of T_g, three RF models with different descriptors; (i) composition, (ii) composition and atomic mass, and (iii) composition, atomic mass and atomic size; were used for prediction of Young's modulus. The overlapping of predicted data with actual data and the R^2 training and testing for last two models (with two and three descriptors) is shown in Fig.8. The RF Model with three descriptors was found to have highest accuracy with R^2 training and testing equal to 0.88 and 0.89 respectively. This model was optimized with 150 number of RF tress (data shown in Fig.9. The accuracy of 3rd model can be noted from nearly similar values of ML predicted and MD estimated Young's modulus values in Table 5 for unknown data which was not used either for training or testing.

Conclusion

In this work, we carried out large number of experiments evaluating two popular ML algorithms: Linear Regression and Random Forest to analyze data sets of sodium borosilicate glasses and window shielding glasses and their respective density, glass transition temperature T_g and Young's Modulus. We investigated the performance of these algorithm when used for the prediction with default and featured engineered models. We also investigated the effect of descriptors on the different property prediction and the R^2 value of the models. The impact of descriptor on the different properties of glass was noted to be different. The density had direct relation with composition and addition of any other descriptor doesn't have

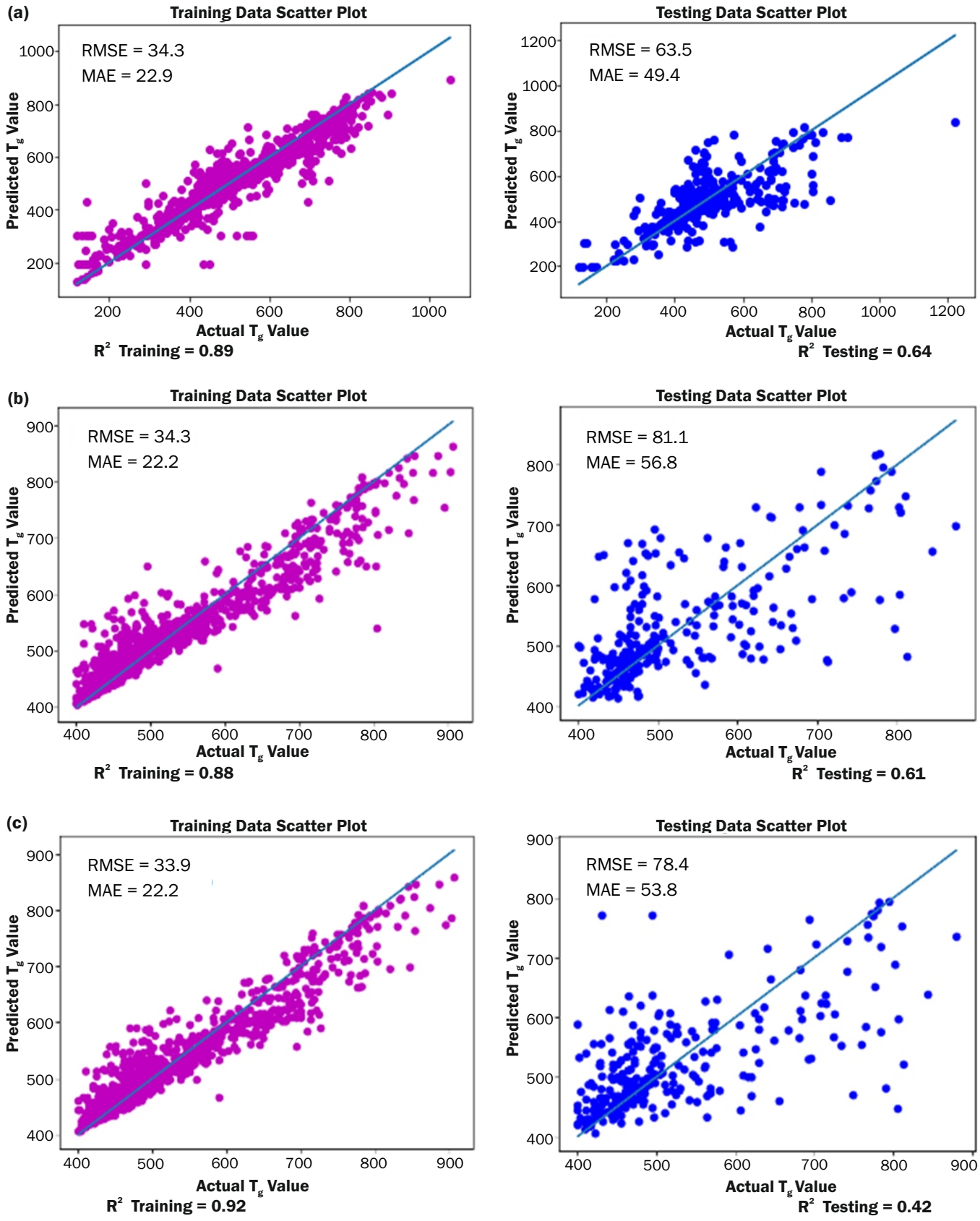


Fig.6: Scattering plot of actual T_g value vs predicted T_g value of training and the testing using Random Forest Model. (a) Model with one descriptor (composition), (b) Model with two descriptors (composition and Atomic Mass), (c) Model with three descriptors (composition, Atomic Mass and Atomic size).

any significant effect on the model accuracy. On the other hand, Young's modulus not only depends on the composition but also greatly depends on other descriptors like atomic mass and atomic size. While T_g mostly depends on the compositions and density. Importantly, with the prediction accuracy of these models for unknown data, it was shown that the sole-observation of the mean error values (RMSE, MAE) and R^2 values can't give enough evidence for the statistical accuracy of model than its competitors with higher values of mean error bar and R^2.

Table 4: Comparison of T_g from MD simulation and ML model for RSW glass.

Sr. No.	Glass composition						T_g (°C) [MD]	T_g (°C) [ML]
	SiO_2	K_2O	Na_2O	PbO	Li_2O	As_2O_3		
1	66.7	16.6	16.7	0	0	0	431.2	420
2	75.8	6.8	7.6	9.8	0	0	422.4	413
3	90	0	5	0	5	0	452.0	467
4	77.3	2.8	4	14.4	1.4	0.04	444.4	445

Comparison of Tg from MD simulation and ML model[using RF method with three descriptors: composition, atomic mass and atomic size] for RSW glass.

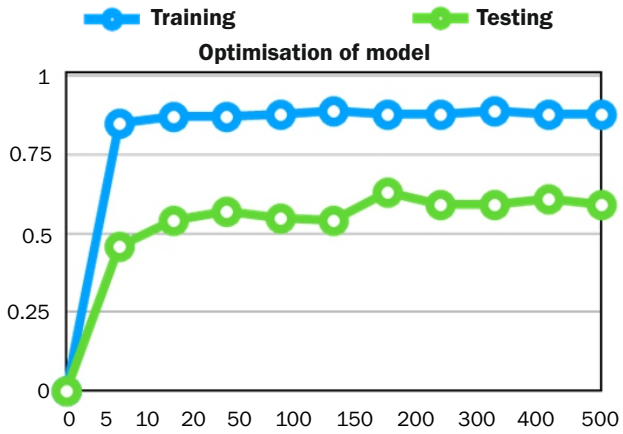


Fig.7: The accuracy (as captured by the R^2 value) of the Random Forest model as a function of number of trees considered in each model as obtained for the training and testing set respectively.

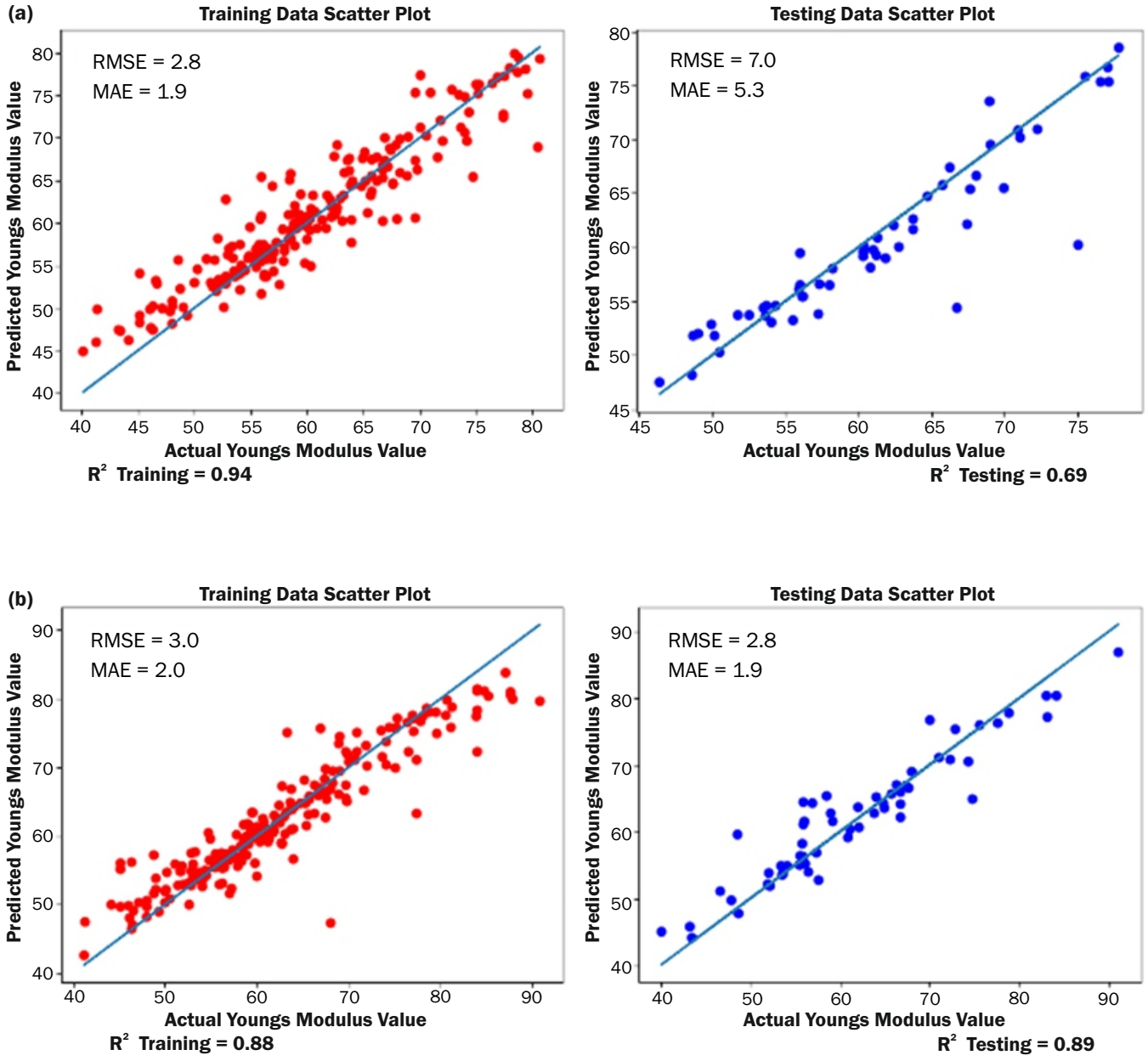


Fig.8: Scattering of actual Young's Modulus vs predicted Young's Modulus of training and the testing using Random Forest Model. (a) Model with two descriptors (composition and atomic mass), (b) Model with three descriptors (composition, atomic mass and atomic size).

Table 5: Predicted Young's modulus and actual Young's modulus from 3rd RF model.

Sr. No.	Glass composition							Y [GPa] [MD]	Y [GPa] [ML]
	SiO_2	K_2O	Na_2O	PbO	Li_2O	BaO	As_2O_3		
1	66.7	14.5	0	19	0	0	0.24	50.50	50.28
2	77.3	7.6	5.8	0	0	9.3	0	71.00	71.15
3	75	5	10	0	0	10	0	64.92	64.36
4	52	8	20	0	20	0	0	55.90	55.99

The presented study can be easily expanded to predict other properties such as thermal expansion coefficient, elastic modulus, and hardness to successfully replace empirical approves for developing novel glasses with useful properties and applications. The Machine Learning can also be used to construct an algorithm which would be able to predict more than one property at once.

Acknowledgement

Authors sincerely acknowledge Director, Chemical Engineering Group, BARC and Head, Chemical Engineering Division, BARC for their and support and encouragement.

References

[1] Zanotto, E. D.; Mauro, J. C. The Glassy State of Matter: Its Definition and Ultimate Fate. Journal of Non-Crystalline Solids2017, 471, 490–495. https://doi.org/10.1016/j.jnoncrysol.2017.05.019.

[2] Zanotto, E. D.; Coutinho, F. A. B. How Many Non-Crystalline Solids Can Be Made from All the Elements of the Periodic Table? Journal of Non-Crystalline Solids2004, 347 (1), 285–288. https://doi.org/10.1016/j.jnoncrysol.2004.07.081.

[3] Liu, H.; Fu, Z.; Yang, K.; Xu, X.; Bauchy, M. Machine Learning for Glass Science and Engineering: A Review. Journal of Non-Crystalline Solids: X2019, 4, 100036. https://doi.org/10.1016/j.nocx.2019.100036.

[4] Cassar, D. R.; de Carvalho, A. C. P. L. F.; Zanotto, E. D. Predicting Glass Transition Temperatures Using Neural Networks. Acta Materialia2018, 159, 249–256. https://doi.org/10.1016/j.actamat.2018.08.022.

[5] Alcobaça, E.; Mastelini, S. M.; Botari, T.; Pimentel, B. A.; Cassar, D. R.; de Carvalho, A. C. P. de L. F.; Zanotto, E. D. Explainable Machine Learning Algorithms For Predicting Glass Transition Temperatures. Acta Materialia2020, 188, 92–100. https://doi.org/10.1016/j.actamat.2020.01.047.

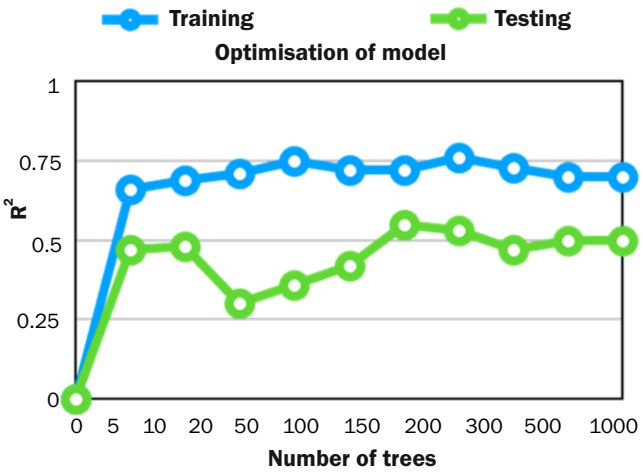


Fig.9: The accuracy (as captured by the R^2 value) of the Random Forest model as a function of number of trees considered in each model as obtained for the training and testing set respectively.

[6] Yang, K.; Xu, X.; Yang, B.; Cook, B.; Ramos, H.; Krishnan, N. M. A.; Smedskjaer, M. M.; Hoover, C.; Bauchy, M. Predicting the Young's Modulus of Silicate Glasses Using High-Throughput Molecular Dynamics Simulations and Machine Learning. Sci Rep 2019, 9 (1), 8739. https://doi.org/10.1038/s41598-019-45344-3.

[7] Breiman, L.; Friedman, J. H. Predicting Multivariate Responses in Multiple Linear Regression. Journal of the Royal Statistical Society: Series B (Methodological)1997, 59 (1), 3–54. https://doi.org/10.1111/1467-9868.00054.

[8] Breiman, L. Random Forests. Machine Learning2001, 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

[9] Feature Engineering for Machine Learning [Book]. https://www.oreilly.com/library/view/feature-engineering-for/9781491953235/ (accessed 2023-06-07).

[10] Jain, A.; Bligaard, T. Atomic-Position Independent Descriptor for Machine Learning of Material Properties. Phys. Rev. B2018, 98 (21), 214112. https://doi.org/10.1103/PhysRevB.98.214112.

[11] Liu, H.; Zhang, T.; Anoop Krishnan, N. M.; Smedskjaer, M. M.; Ryan, J. V.; Gin, S.; Bauchy, M. Predicting the Dissolution Kinetics of Silicate Glasses by Topology-Informed Machine Learning. npj Mater Degrad2019, 3 (1), 1–12. https://doi.org/10.1038/s41529-019-0094-1.