Ceramics International 50 (2024) 3018–3025

Contents lists available at ScienceDirect

Ceramics International

journal homepage: www.elsevier.com/locate/ceramint

Accelerating design of glass substrates by machine learning using small-to-medium datasets

Jiaqian Zhu, Linfeng Ding [*], Guohao Sun [**], Lianjun Wang

State Key Laboratory for Modification of Chemical Fibers and Polymer Materials, Engineering Research Center of Advanced Glass Manufacturing Technology, Ministry of Education, Donghua University, Shanghai, 201620, China

ARTICLE INFO

Handling Editor: Dr P. Vincenzini

Keywords:
Glass substrates
Machine Learning (ML)
Composition design
Property prediction

ABSTRACT

The demand for high-performance displays is driving the development of glass substrates with extremely low strain variation during manufacturing. To achieve the necessary balance of physical properties to resist strain, the traditional compositional design methods for glass substrates like trial-and-error and classical computational techniques should be optimized. As an alternative approach, machine learning (ML) algorithms have emerged for designing new glass compositions. In this work, we conduct ML research focused on the compositional design for high-performance glass substrates. We employ three ML algorithms i.e., Random Forest (RF), Classification And Regression Tree (CART) and k-Nearest Neighbors (k-NN) to predict five physical properties that are key to the performance of alkaline-free aluminosilicate glass substrates. By using only small-to-medium dataset sizes, our model reaches a high coefficient of determination of 0.9879. Furthermore, our model achieves an accurate generalized prediction of 50 experimental data and enables the prediction and design of glass substrate compositions with advanced comprehensive properties.

1. Introduction

The demand for high-performance displays with expanded size and resolution capabilities has surged in the current era, driven by the widespread use of display applications in televisions, car dashboards, phones, and wearable devices [1]. The glass substrates are the key to the development of these cutting-edge displays, particularly in thin-film-transistor (TFT) technologies, which confront a host of challenges during the production process [1]. These glass substrates undergo significant changes in shape or size, inducing strain known as the change in total pitch (TP) throughout the typical TFT backplane process. Crucially, among the various attributes of glass substrates, the total pitch variation (TPV) assumes paramount importance, representing the deviation in glass movement within and across sheets. To ensure exemplary TPV performance, glass substrates must possess a harmonious amalgamation of physical properties, adeptly countering the diverse sources of strain, which encompass elastic distortion, stress relaxation, and compaction. Consequently, achieving the epitome of TPV performance hinges on the precise balance of the glass composition, striking an optimal equilibrium among these physical properties.

In the pursuit of excellent TPV performance, the characteristics of glasses become significant, as they are non-crystalline materials with a disordered atomic and molecular structure, possessing inherent stoichiometry, which contributes to the unpredictability of their properties [2]. However, traditional methods like trial-and-error and classical computational techniques such as *ab initio* and classical molecular dynamics simulations have proven to be time-consuming, expensive, and limited in their data output [3,4]. With the rapid development of computer science and technology, machine learning (ML) algorithms have emerged as an alternative approach for designing new glass compositions, overcoming the disadvantages of traditional methods. The primary objective of ML-based strategies is to specify desired properties and identify potential candidate compositions. Numerous oxide glass properties, including Young's Modulus [5–9], *CTE* [6,10,11], viscosity [12,13], density [14–17] and liquidus temperature [18] have been investigated using ML algorithms. In the realm of oxide glasses, Dreyfus et al. pioneered to predict the liquidus temperature for oxide glass-forming liquids [18]. Brauer et al. first used ANNs to predict the chemical durability of glasses containing P_2O_3, CaO, MgO, Na_2O, and TiO_2 [19], Ravinder et al. employed deep learning to aid design of oxide

[*] Corresponding author.
[**] Corresponding author.
E-mail addresses: Linfeng.Ding@dhu.edu.cn (L. Ding), ghsun@dhu.edu.cn (G. Sun).

https://doi.org/10.1016/j.ceramint.2023.11.048
Received 8 September 2023; Received in revised form 23 October 2023; Accepted 3 November 2023
Available online 7 November 2023

glass [20], and Cassar et al. successfully predicted glass transition temperatures using neural networks [21]. Moreover, other ML algorithms like Random Forest (RF), Classification And Regression Tree (CART) and k-Nearest Neighbors (k-NN) have also been explored to study glass properties. Notably, Cassar et al. recently used a large dataset of around 150,000 oxide glasses to investigate the predictive performance of three ML algorithms for six different glass properties and also employed SHAP (Shapley Additive exPlanations) values to interpret the models [6]. Deng proved the ability of ML on predicting the density and elastic properties of oxide glasses by leveraging a large dataset of over 30,000 data points from Corning Incorporated [22]. Nevertheless, to the best of our knowledge, none of these works focused on developing glass substrates for achieving optimal TPV performance via a balanced glass composition on medium to small datasets.

To optimize glass substrate and minimize elastic distortion during manufacturing, a high elastic modulus (i.e., higher than 80 GPa) has proven effective in reducing strain resulting from stress variations. Additionally, to address stress relaxation, it is essential to utilize a substrate with a higher effective viscosity at elevated temperatures. The elastic modulus directly reflects the glass's ability to resist deformation during the manufacturing process, while employing low density results in a lightweight glass sheet. Notably, the ratio of elastic modulus to density plays a crucial role in determining the extent of sag in the glass. A higher ratio, achieved through a higher modulus and/or lower density, leads to reduced sag and improved overall performance [3].

Considering the significance of alkali-free display glasses, our focus revolves around glass substrates composed of seven oxides, namely SiO_2, Al_2O_3, CaO, MgO, SrO, B_2O_3, and Sb_2O_3. Our research aims to explore an accurate prediction of multiple properties with small to medium datasets through an ML approach. Fig. 1 illustrates the whole process of the experiment. Firstly, we performed a comprehensive statistical analysis of the data distribution to choose appropriate methods. Subsequently, we applied three ML techniques i.e., RF, k-NN, and CART to train and predict the respective datasets. To enhance the predictive capabilities, we carefully selected the best-performing model and employed it to fill in the missing values in the original dataset. Furthermore, we manually gathered a small dataset (∼50 entries) of compositions that were unseen during the training period, to thoroughly test the generalization ability of the joint prediction model. In addition, we also utilized SHAP values to gain insights into the contribution of each oxide to the properties of interest. Our approach successfully predicted five crucial properties of alkali-free glass substrates with remarkable precision using small to medium datasets. The ML model developed in this work has the potential to greatly assist in the design and development of glass substrates with improved characteristics and performance. For the purpose of full reproducibility, the codes and datasets utilized in this research are made accessible on GitHub (https://github.com/JiaQianZhu/ML_glass).

2. Methodology

2.1. Data collection and processing

All the data used in the experiments were collected from the SciGlass database [23] via a Python module named GlassPy [24]. Besides, data cleaning was performed on the selected dataset by removing negative or missing values to ensure that the molar sum of the seven selected oxides equals 1. The composition of glass covers seven oxides: SiO_2, Al_2O_3, CaO, MgO, SrO, B_2O_3, and Sb_2O_3. Because Sb_2O_3 is serving as a fining agent during glass melting, and will not significantly influence the glass properties, we show the distribution plots of six primary oxides in Fig. 2. The property dataset includes density, Young's Modulus, *CTE*, T_{strain} and T_{soft}. The distribution plots of these five properties are shown in Fig. 3. The original dataset comprises approximately 9,400 records (around 6, 120 glass compositions), of which merely around 100 possess complete values for all five target attributes. Thus, the dataset was segregated into five distinct subsets based on the presence of non-missing values in their respective target attributes. The datasets of density and *CTE* consist of 2, 639 and 2,078 data points, respectively. While the datasets of Young's modulus, T_{strain} and T_{soft} have rather small sizes of 630, 534 and 145, respectively. Within our categorization, datasets with less than 1,000 data points are classified as small-sized, whereas datasets containing 1, 000 to 3,000 points are classified as medium-sized. In comparison to sparse datasets with only several hundred points [7,8] and large datasets with over 3,000 points [6,10,12,13,15,22], our datasets fall within the small to medium-sized categories. Through the extraction of non-missing segments from each property, we derived five distinct datasets. Each dataset is characterized by a single property and its corresponding oxide composition. Leveraging ML models, we conducted separate learning endeavors on each of these datasets. The best-performing model from this process was then preserved, serving as the key tool for predicting and filling in missing values within the dataset.

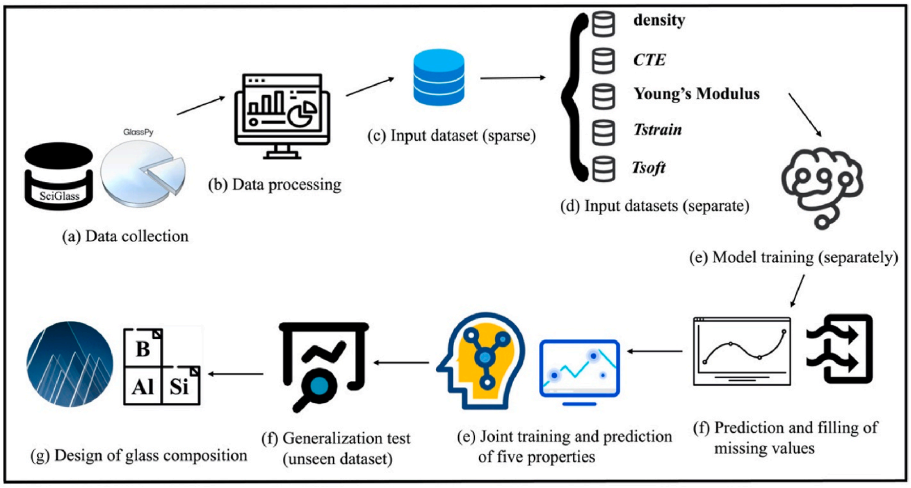
SciGlass
GlassPy
(a) Data collection
(b) Data processing
(c) Input dataset (sparse)
density
CTE
Young's Modulus
Tstrain
Tsoft
(d) Input datasets (separate)
(e) Model training (separately)
(f) Prediction and filling of missing values
(e) Joint training and prediction of five properties
(f) Generalization test (unseen dataset)
B
Al
Si
(g) Design of glass composition

Fig. 1. Illustration of the experiment's sequential stages, initiated by (a) data collection and (b) data processing. This led to the formation of a (c) sparse input dataset. Addressing the limitation of small datasets, we (d) partitioned the initial dataset into five segments and pursued (e) independent model training for each. The best-performing model was then chosen for (f) predicting and completing missing values within the initial dataset. Subsequently, (e) joint training and prediction were conducted on the original dataset. Furthermore, unseen data was employed to rigorously (f) assess the generalization capability of the joint prediction model. Conclusively, insights gleaned from this study guided the (g) design of glass substrates.

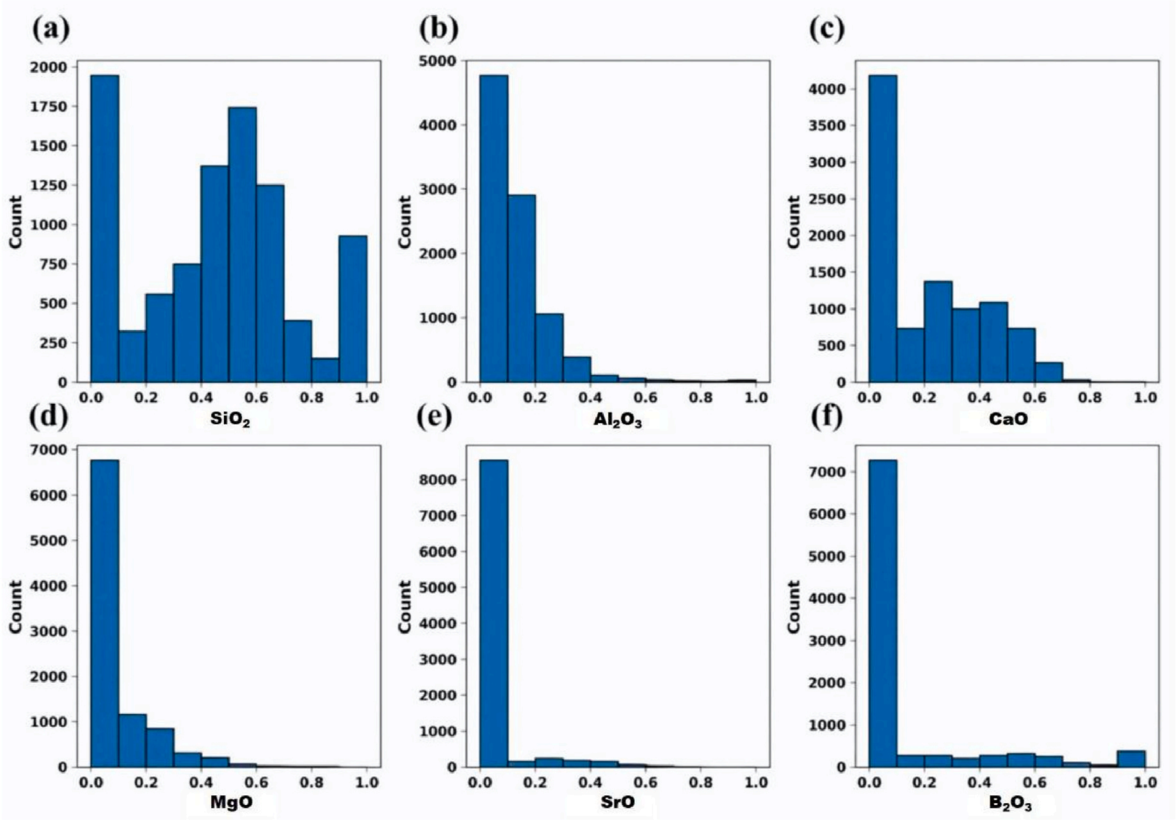

Fig. 2. The oxide composition distributions of (a) SiO_2, (b) Al_2O_3, (c) CaO, (d) MgO, (e) SrO, (f) B_2O_3.

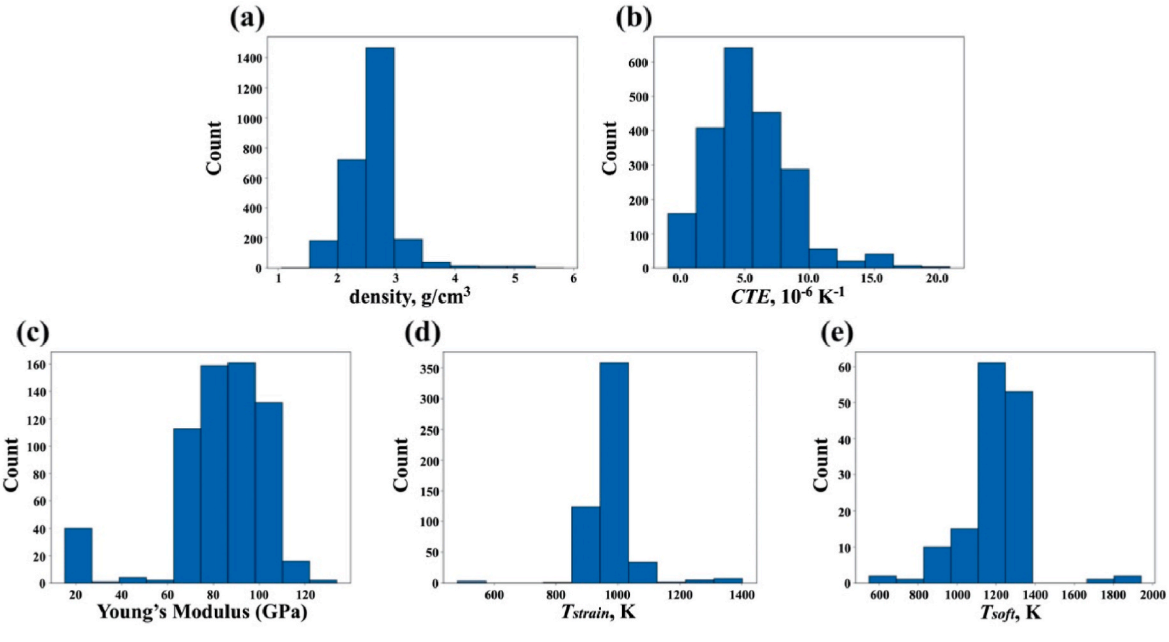

Fig. 3. The distribution of (a) density, (b) *CTE*, (c) Young's Modulus, (d) T_{strain}, (e) T_{soft}.

2.2. Machine learning algorithms

Prior to applying the ML algorithms, ten different ones were tested on density and *CTE* datasets to evaluate their performance. These algorithms included RF, k-NN, CART, Linear Regression, Ridge Regression, Gaussian Process Regression, XGBoost, Lasso Regression, Support Vector Machine, and Multilayer Perceptron. Based on the results of this preliminary analysis (shown in https://github.com/JiaQianZhu/ML_glass), we selected three ML algorithms, namely CART, k-NN, and RF, which demonstrated the best overall performance in predicting glass properties [6,25].

During the training phase, the CART algorithm constructs a decision tree for classification and regression tasks [26] using a greedy approach. The CART algorithm has a vital contribution to regression analysis, as it constructs decision trees using a greedy training approach to capture intricate non-linear relationships and patterns in the data. By selecting optimal split points, CART accommodates diverse feature variables, including continuous and categorical types. It provides interpretable and explainable results, offering importance rankings of pivotal variables to facilitate a comprehensive understanding of the data's influencing factors. Moreover, CART exhibits robustness against outliers and missing data, making it widely used in regression analysis.

The k-NN algorithm is a widely utilized non-parametric classification and regression technique [27]. It operates on the principle of proximity, where an unlabeled instance is assigned a class label based on the classes of its k nearest neighbors in the feature space. The k-NN algorithm can handle both numerical and categorical data, making it versatile in various domains. However, its computational complexity grows with the size of the training dataset. Despite this limitation, k-NN remains an effective and interpretable method for pattern recognition and data analysis tasks.

In contrast to the CART algorithm which generates a single decision tree, the RF algorithm forms a forest by inducing a collection of decision trees [28]. It constructs a multitude of decision trees through bootstrapping and feature randomization, and aggregates their predictions to produce robust and accurate results. RF handles both categorical and continuous variables and effectively addresses overfitting issues. Its ability to capture complex interactions and handle high-dimensional data makes it popular in various domains.

2.3. Model training

All the ML models were trained using a Python package named scikit-learn [29]. Based on dataset analysis, our experiment commenced with the exploration of five distinct datasets, each derived from non-null data segments. Firstly, CART, k-NN and RF were used to train and predict the properties of density, Young's Modulus, *CTE*, T_{strain} and T_{soft}. For properties with medium datasets (>1000 entries), the datasets were initially divided into 80% for training and 20% for testing. Subsequently, 10-fold Cross-Validation (CV) was conducted on the training sets. For properties with small datasets, the Leave-One-Out Cross--Validation (LOOCV) method was combined with CART, k-NN and RF algorithms. The assessment of each ML algorithm's performance involved employing widely accepted evaluation metrics. The model yielding the optimal performance was identified, saved, and stored for subsequent utilization. Upon incorporating the filled values, the dataset was partitioned into proportions of 64:20:16 for training, testing, and validation sets, respectively. Following Pareto's principle, the 80:20 split is used as a rule of thumb by researchers and has been verified to be effective empirically [30]. Thus, we divided the dataset into 80% for training and 20% for testing. Additionally, taking the advice of Pedro Domingos [31], we introduced a validation set to enhance model evaluation and mitigate overfitting risks. As a result, we performed another 80:20 split on the training data, which ultimately allocated 64% of the dataset for training, 16% for validation, and 20% for testing.

To test the generalization capability of the finalized joint prediction model, a novel dataset extracted from the US patent data released by Corning Incorporated [32] was employed as the testing set. By employing the saved model, direct predictions were generated, and the residuals between the predicted values and experimental data were also analyzed.

2.4. SHAP analysis

The SHAP values were also employed to interpret the predictions generated by the trained models. These values were estimated using the SHAP Python module [33]. The SHAP values corresponded to each feature in the models, representing the quantity of each chemical element in this particular work. They shared the same unit as the target property predicted by the models. Importantly, the SHAP values exhibited an additive property, where the sum of the SHAP values for a specific prediction, along with a base value (the mean of the target value), equated to the model's predicted value. This additivity facilitated understanding the contribution of each feature to the final prediction, which is particularly relevant to the glass research community. The SHAP values provided insights into which elements contributed to increasing or decreasing a specific property, as well as the magnitude of their influence.

2.5. Evaluation measures

To evaluate the model performance, five metrics: mean square error (MSE) [34], coefficient of determination R^2 factor [34], mean absolute error (MAE) [34], median absolute error(MedAE) [35] and relative deviation(RD) [35] were applied. MSE measures the average of the squared differences between the real value and the predicted value. R^2 is a statistical factor that measures how close the data are to the fitted line. MAE calculates the average absolute difference between the real value and the predicted value. MedAE calculates the median absolute difference between the real value and the predicted value, providing a robust measure of prediction error. RD measures the relative deviation between the real value and the predicted value as a percentage of the real value, indicating the magnitude of the prediction error relative to the actual value.

These evaluation metrics can be considered together to assess the performance of the model. Generally, lower values of MSE and MAE indicate smaller prediction errors, while a higher R^2 value indicates a better fit of the model to the data. A smaller value of MedAE suggests that the model is robust to outliers, while a smaller value of RD indicates a smaller prediction error.

3. Results and discussion

3.1. Individual prediction

Table 1 shows the performance of different models from individual predictions evaluated by MSE, MAE, MedAE, RD, and R^2. We observe that the RF algorithm demonstrates the overall best performance in predicting the missing attributes, followed by CART and k-NN. The only exception is for Young's Modulus, where k-NN has a slightly better performance. For attributes with medium-sized datasets, such as *CTE* and density, RF algorithm exhibits excellent performance with an R^2 higher than 0.95. However, for attributes with small datasets, such as Young's Modulus, T_{strain} and T_{soft}, the performance of RF is comparatively poorer with R^2 between 0.83 and 0.92. The performance of the three different ML models is consistently aligned with findings from a study utilizing larger datasets (∼150,000 oxide glasses) [6]. We save the values individually predicted by the best-performing models and fill in the corresponding missing values.

Table 2 shows the experimental results of 10-fold CV on the density and *CTE* within their respective training sets. It is evident that the RF model consistently outperforms other models when evaluated individually for density and *CTE*. Additionally, when comparing these results to those presented in Table 1, we can observe a slight decrease in performance in the 10-fold CV scenario. This indicates that our model demonstrates commendable generalization capabilities.

3.2. Joint Prediction

Based on the individual predictions detailed above, the original dataset, encompassing approximately 9,400 records, was effectively augmented. We performed a joint prediction using RF, CART, and k-NN, and the results are shown in Table 3. We found that RF remains the overall best algorithm, achieving an impressive R^2 score of 0.9879, an MSE score of 24.9990, and an MAE score of 1.2986. Although we attempted to use a neural network for improved prediction accuracy, its

Table 1
The performance of different models from individual predictions evaluated by MSE, MAE, MedAE, RD and R^2. The upward arrows indicate that the higher the metric the better, whereas the downward arrows indicate the opposite.

Metrics	density			*CTE*			Young's Modulus			T_{strain}			T_{soft}		
	RF	CART	k-NN	RF	CART	k-NN	RF	CART	k-NN	RF	CART	k-NN	RF	CART	k-NN
MSE↓	**0.0071**	0.0099	0.0189	**4.1394e-13**	5.6218e-13	4.7553e-13	32.8873	47.7488	**28.6867**	**993.9044**	1007.2885	1116.9643	**4517.2962**	5038.0845	7490.8303
MAE↓	**0.0392**	0.0461	0.0650	**4.2220e-07**	4.8139e-07	4.4517e-07	2.9207	3.3697	**2.9100**	**9.8485**	11.0842	12.0940	**31.7827**	35.1207	37.5586
MedAE↓	**0.0164**	0.0180	0.0177	**2.6437e-07**	2.9999e-07	2.4352e-07	1.5125	1.5000	**1.6400**	**3.5575**	5.0000	5.2000	**6.6058**	9.0000	9.6000
RD(%)↓	**1.4326**	1.6964	2.8255	**8.7130**	9.6459	9.0149	40.3478	40.8232	**40.3887**	**6.8310**	6.9320	6.6021	**12.7056**	13.3146	11.8268
R^2↑	**0.9614**	0.9460	0.8974	**0.9599**	0.9455	0.9539	0.9245	0.8904	**0.9342**	**0.8501**	0.8383	0.8207	**0.8304**	0.8108	0.7187

Table 2
The experimental results of 10-fold CV on the density and *CTE* within their respective training sets. The upward arrows indicate that the higher the metric the better, whereas the downward arrows indicate the opposite. Each metric's first row represents the mean values obtained from 10-fold CV, and the second row represents the corresponding standard deviation.

Metrics	density			*CTE*		
	RF	CART	k-NN	RF	CART	k-NN
MSE↓	**0.0163** ± 0.0177	0.0187 ± 0.0180	0.0174 ± 0.0175	**4.5901e-13**± 1.0904e-13	7.0233e-13± 1.4109e-13	5.0814e-13± 1.2591e-13
MAE↓	**0.0463** ± 0.0069	0.0521 ± 0.0070	0.0519 ± 0.0091	**4.2066e-07**± 2.9906e-08	5.2780e-07± 3.6453e-08	4.5045e-07± 3.3855e-08
MedAE↓	**0.0170** ± 0.0020	0.0179 ± 0.0031	0.0196 ± 0.0016	**2.2165e-07**± 2.8566e-08	2.8243e-07± 2.9424e-08	2.6119e-07± 2.0633e-08
RD(%)↓	**1.7138** ± 0.1749	1.9279 ± 0.2043	1.9514 ± 0.3956	**9.4883**± 1.8817	11.3129 ± 1.9611	9.9910± 2.1658
R^2↑	**0.9241** ± 0.0554	0.9102 ± 0.0559	0.9138 ± 0.0590	**0.9519**± 0.0143	0.9265± 0.0197	0.9471± 0.0152

Table 3
The experimental results of joint prediction. The upward arrows indicate that the higher the metric the better, whereas the downward arrows indicate the opposite.

Metrics	RF	CART	k-NN
MSE↓	**24.9990**	43.4276	110.5426
MAE↓	**1.2986**	1.5109	2.6001
MedAE↓	0.2876	**0.0161**	0.5699
RD(%)↓	**1.5205**	1.7233	1.5351
R^2↑	**0.9879**	0.9768	0.9853

performance was not as satisfactory as that of RF. Moreover, the predictive accuracy of RF has already matched the experimental precision. Considering the law of diminishing marginal returns, further algorithm improvements may not yield significant performance gains. Therefore, we are assigning the RF algorithm as the final joint prediction model and applying it to accelerate the composition design of glass substrates.

3.3. Model interpretation with Scatter Density Plot

A Scatter Density Plot combines the features of a scatter plot and a density plot, offering a visual representation of data point distribution and density within a two-dimensional space. It offers valuable insights into the relationships among variables, patterns, clusters, and outliers present within the dataset. As a result, it becomes an indispensable tool for conducting exploratory data analysis and comprehending the underlying data structure [36]. Fig. 4(a) to Fig. 4(e) provide a comprehensive view of the RF model's predictions compared to experimental measurements. In Fig. 4(a), the model excels in predicting density, achieving an excellent performance with an R^2 value of 0.9887. The predicted residuals largely fall within −0.1 to 0.1. Similarly, Fig. 4(b) shows the RF model's remarkable ability to predict *CTE*, achieving an excellent R^2 value of 0.9900. Most predicted residuals cluster between −1.0 and 1.0 (scaled by a factor of 1e-6), except for a few outliers. Moving to Fig. 4(c), the RF model performs remarkably well in predicting Young's Modulus with an R^2 value of 0.9871, despite a relatively larger number of outliers compared to *CTE*. Fig. 4(d) highlights the RF model's exceptional performance in predicting T_{strain}, as indicated by an impressive R^2 value of 0.9981. T_{strain} data reveals very few outliers, with the majority of predicted residuals closely distributed around zero.

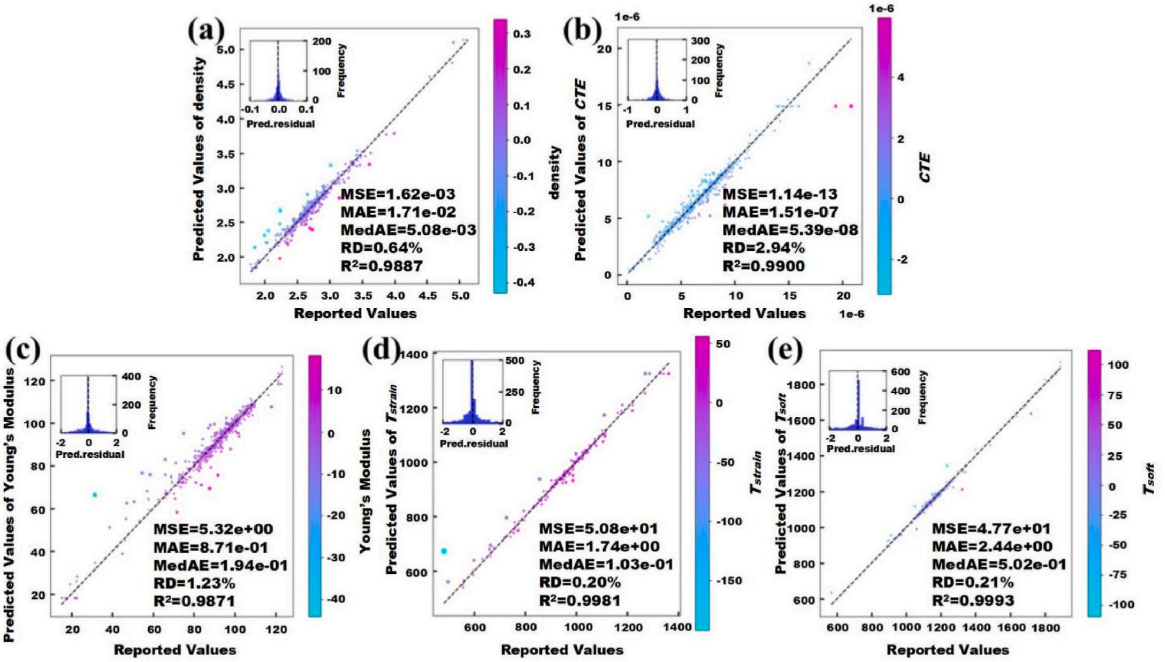

Fig. 4. Scatter Density Plots for (a) density, (b) *CTE*, (c) Young's Modulus, (d) T_{strain}, (e) $T_{\text{soft.}}$

Meanwhile, Fig. 4(e) demonstrates the RF model's excellent predictive power for T_{soft}, achieving an R^2 value of 0.9993. Despite the presence of outliers in T_{soft}, which leads to relatively larger prediction deviations, the RF model consistently demonstrates its proficiency in capturing the underlying data patterns and relationships.

3.4. Model interpretation with SHAP

By examining the SHAP plots presented in Fig. 5, we can ascertain the hierarchy of oxide importance in predicting each property. Notably, the consistently top-ranked oxides across properties are SiO_2, B_2O_3, and Al_2O_3, which are glass network formers [1]. For density (Fig. 5(a)), increased SiO_2 and B_2O_3 content results in decreased density in alkaline-earth aluminosilicate glass. The incorporation of additional network formers within the glass matrix results in the polymerization of the glass network, consequently leading to a reduction in glass density [1,37]. Fig. 5(b) reveals insights regarding *CTE*. An increase in SiO_2, B_2O_3, and Al_2O_3 content is associated with a reduction in *CTE*. Additionally, slight variations such as an increase in CaO or SrO, or a decrease in MgO, oxides acting as glass network modifiers, might result in increased *CTE* due to the mix-alkaline earth effect. This is consistent with the consensus in the field of glass research, which indicates that higher concentrations of glass network formers result in lower *CTE*. Additionally, the alkali-earth mixing effect allows for fine-tuning of *CTE* of glass [1,37]. Turning to Young's Modulus, as presented in Fig. 5(c), decreases in B_2O_3 and SiO_2 content lead to a reduction in Young's Modulus. Conversely, a decrease in Al_2O_3 content exhibits an opposite trend. The presence of B_2O_3, SiO_2, and Al_2O_3 in calcium aluminosilicate glasses has been reported to induce a non-monotonic behavior in Young's Modulus, as demonstrated in a recent study [38]. Fig. 5(d) and (e) reveal analogous patterns for T_{strain} and T_{soft}. An increase in SiO_2 content is linked to elevated viscosity of glass substrates, while an increase in B_2O_3 content corresponds to reduced glass viscosity. This can be attributed to the strong atomic bonding within the silicon tetrahedral

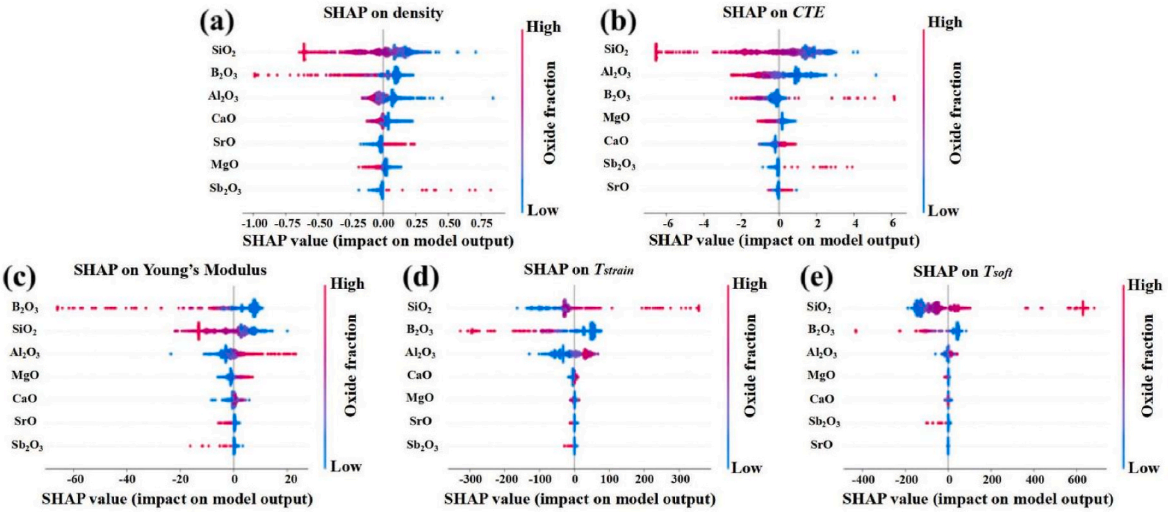

Fig. 5. SHAP values for (a) density, (b) *CTE*, (c) Young's Modulus, (d) T_{strain}, (e) $T_{\text{soft.}}$

structure, contrasted with the relatively weaker bonding of the three-coordinated boron [37]. Furthermore, Al_2O_3's impact on T_{strain} outweighs its effect on T_{soft}, which may increase or decrease the viscosity of glass due to the rich coordination variations during the glass forming process [37]. While some of the insights gleaned from the SHAP analysis are well-established within the glass community, this analysis contributes to empirical glassmaking and glass simulations. It aids in achieving targeted property adjustments through minor glass composition tweaks, thus accelerating the development of glass substrates [6,39, 40].

3.5. Generalization testing

To evaluate the generalization capability of our model, we collected 50 experimental records of newly developed glass substrates from a recent US patent released by Corning Incorporated [32]. We calculated the differences (named residuals) between the reported values and the predicted values from our model. The outcomes, tabulated in Table 4, highlight notable predictive accuracy for density and *CTE*, where the original training dataset is sufficiently abundant (medium-sized dataset). In contrast, the existence of significant missing data within the original dataset introduces a certain degree of disparity between predicted and reported values for the remaining three properties. Specifically, our analysis reveals an average overestimation of approximately 2 GPa in the predicted values of Young's Modulus compared to the corresponding experimental values. Similarly, predicted values for T_{strain} and T_{soft} display an average overestimation of approximately 3 K. It is pertinent to note that, when considering the typical experimental measurement uncertainty of ±2 GPa for Young's Modulus from nanoindentation [41] and a measurement uncertainty exceeding ±5 K for T_{strain} and T_{soft} (measured using beam-bending [42] or fiber-elongation viscometers), our model's predictions achieve a commendable level of accuracy. Moreover, we also calculated the difference between experimental and predicted mean values, and found out the differences are rather minor within 2.5%.

3.6. Towards composition design for glass substrates

After the above training and validating, our model has gained the capacity to accelerate compositional design for glass substrates, even when working with small to medium datasets. We establish a target objective for advanced glass substrates, aiming to achieve optimal TPV performance as outlined below: density ≤ 2.5 g/cm^3, *CTE* ranging from 3.3×10^{-6}/K to 3.9×10^{-6}/K, Young's modulus ≥ 80 GPa, $T_{strain} \geq 993.15$ K and $T_{soft} \geq 1273.15$ K.

Based on the design experience derived from the US patent released by Corning Incorporated [32] and the aforementioned SHAP analysis, we define the following ranges for the oxide molar ratios: SiO_2 should be within 65%–73%, Al_2O_3 between 10% and 14%, CaO ranging from 3% to 9%, MgO ranging from 2% to 6%, SrO within 0%–6%, B_2O_3 between 3% and 8%, and Sb_2O_3 within 0%–1%.

We conduct a random search within the predefined oxide ranges to identify potential composition ratios, referred to as candidate ratios. These candidate ratios are then input into the previously saved joint prediction model, generating predicted values for the target properties—referred to as predicted property values. Subsequently, we compare the predicted property values with the respective target range values for each property and retain candidate ratios that fulfill all the target criteria. By implementing this method, we compile a list of compositions that meet the specified conditions (https://github.com/JiaQianZhu/ML_glass), thereby accelerating the glass substrate development process. Moreover, the approach employed in this study also holds promise for the development of advanced glass compositions featuring optimized comprehensive properties.

Table 4
The minimum, maximum, mean and median values of the residuals and the difference from the model prediction and experiments.

Residuals	density	*CTE*	Young's Modulus	T_{strain}	T_{soft}
Min	−0.0291	−3.6930e-07	0.7283	−4.3699	−5.4231
Max	0.0171	3.9055e-07	2.9217	19.3620	25.1608
Mean	−0.0031	−4.9837e-08	2.0490	2.9473	3.3895
Median	−0.0035	−4.4100e-08	2.1388	2.0150	1.4591
Difference (%)	−0.1206	−1.5238	2.5064	0.2821	0.2596

4. Conclusions

In this study, we conducted an ML research on the compositional design for high-performance glass substrates using small to medium dataset sizes, which comprises 7 oxides (SiO_2, Al_2O_3, CaO, MgO, SrO, B_2O_3, Sb_2O_3) and 5 properties (density, Young's Modulus, *CTE*, T_{strain} and T_{soft}). First, we successfully used RF, CART and k-NN algorithms to perform both individual and joint predictions, with RF reaching the highest R^2 value of 0.9879. Then, our model achieved an accurate generalized prediction of 50 experimental data. Last, we further realized the prediction and design of glass substrate compositions with advanced comprehensive properties.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Dr. Daniel R. Cassar for helpful suggestions on data collection and valuable comments on initiate this research. We also thank Dr. Zhen Zhang for reviewing the manuscript and providing valuable comments. The authors gratefully acknowledge financial support from the National Natural Science Foundation of China (U22A20125, 52102002, 62202095), Shanghai Municipal Natural Science Foundation (22ZR1400400), Shanghai Pujiang Program (22PJD002), Fundamental Research Funds for the Central Universities (2232023G-07, 22D111210, 22D111207), and Shanghai Science and Technology Commission (22YF1401100).

References

[1] J.D. Musgraves, J. Hu, L. Calvez, Springer Handbook of Glass, first ed., Springer Nature Switzerland AG, Switzerland, 2019.

[2] D. Jiang, J. Zhang, Z. Wang, C. Feng, K. Jiao, R. Xu, A prediction model of blast furnace slag viscosity based on principal component analysis and K-nearest neighbor regression, JOM 72 (2020) 3908–3916, https://doi.org/10.1007/s11837-020-04360-9.

[3] J. Singh, S. Singh, A review on Machine learning aspect in physics and mechanics of glasses, Mater. Sci. Eng. B 284 (2022), 115858, https://doi.org/10.1016/j.mseb.2022.115858.

[4] R. Ravinder, V. Venugopal, S. Bishnoi, S. Singh, M. Zaki, H.S. Grover, M. Bauchy, M. Agarwal, N.M.A. Krishnan, Artificial intelligence and machine learning in glass science and technology: 21 challenges for the 21st century, Int. J. Appl. Glass Sci. 12 (2021) 277–292, https://doi.org/10.1111/ijag.15881.

[5] J.C. Mauro, A. Tandia, K.D. Vargheese, Y.Z. Mauro, M.M. Smedskjaer, Accelerating the design of functional glasses through modeling, Chem. Mater. 28 (2016) 4267–4277, https://doi.org/10.1021/acs.chemmater.6b01054.

[6] D.R. Cassar, S.M. Mastelini, T. Botari, E. Alcobaça, A.C.P.L.F. de Carvalho, E. D. Zanotto, Predicting and interpreting oxide glass properties by machine learning using large datasets, Ceram. Int. 47 (2021) 23958–23972, https://doi.org/10.1016/j.ceramint.2021.05.105.

[7] S. Bishnoi, S. Singh, R. Ravinder, M. Bauchy, N.N. Gosvami, H. Kodamana, N.M. A. Krishnan, Predicting Young's modulus of oxide glasses with sparse datasets

using machine learning, J. Non-Cryst. Solids 524 (2019), 119643, https://doi.org/10.1016/j.jnoncrysol.2019.119643.

[8] K. Yang, X. Xu, B. Yang, B. Cook, H. Ramos, N.M.A. Krishnan, M.M. Smedskjaer, C. Hoover, M. Bauchy, Predicting the Young's modulus of silicate glasses using high-throughput molecular dynamics simulations and machine learning, Sci. Rep. 9 (2019) 8739, https://doi.org/10.1038/s41598-019-45344-3.

[9] H. Liu, Z. Fu, K. Yang, X. Xu, M. Bauchy, Machine learning for glass science and engineering: a review, J. Non-Cryst. Solids 557 (2021), 119419, https://doi.org/10.1016/j.jnoncrysol.2019.04.039.

[10] S. Bishnoi, R. Ravinder, H.S. Grover, H. Kodamana, N.M.A. Krishnan, Scalable Gaussian processes for predicting the optical, physical, thermal, and mechanical properties of inorganic glasses with large datasets, Mater. Adv. 2 (2021) 477–487, https://doi.org/10.1039/D0MA00764A.

[11] B.M. Tripathi, A. Sinha, T. Mahata, Machine learning guided study of composition-coefficient of thermal expansion relationship in oxide glasses using a sparse dataset, Mater. Today: Proc. 67 (2022) 326–329, https://doi.org/10.1016/J.MATPR.2022.07.170.

[12] D.R. Cassar, ViscNet: neural network for predicting the fragility index and the temperature-dependency of viscosity, Acta Mater. 206 (2021), 116602, https://doi.org/10.1016/j.actamat.2020.116602.

[13] J. Hwang, Y. Tanaka, S. Ishino, S. Watanabe, Prediction of viscosity behavior in oxide glass materials using cation fingerprints with artificial neural networks, Sci. Technol. Adv. Mater. 21 (2020) 492–504, https://doi.org/10.1080/14686996.2020.1786856.

[14] Y.-J. Hu, G. Zhao, M. Zhang, B. Bin, T. Del Rose, Q. Zhao, Q. Zu, Y. Chen, X. Sun, M. de Jong, L. Qi, Predicting densities and elastic moduli of SiO2-based glasses by machine learning, npj Comput. Mater. 6 (2020) 25, https://doi.org/10.1038/s41524-020-0291-z.

[15] S.A. Ghasemi, A. Hofstetter, S. Saha, S. Goedecker, Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network, Phys. Rev. B 92 (2015), 045131, https://doi.org/10.1103/PhysRevB.92.045131.

[16] S. Gong, T. Xie, T. Zhu, S. Wang, E.R. Fadel, Y. Li, J.C. Crossman, Predicting charge density distribution of materials using a local-environment-based graph convolutional network, Phys. Rev. B 100 (2019), 184103, https://doi.org/10.1103/physrevb.100.184103.

[17] S.K. Ahmmad, N. Jabeen, S.T.U. Ahmed, S.A. Ahmed, S. Rahman, Artificial intelligence density model for oxide glasses, Ceram. Int. 47 (2021) 7946–7956, https://doi.org/10.1016/J.CERAMINT.2020.11.144.

[18] C. Dreyfus, G. Dreyfus, A machine learning approach to the estimation of the liquidus temperature of glass-forming oxide blends, J. Non-Cryst. Solids 318 (2003) 63–78, https://doi.org/10.1016/S0022-3093(02)01859-8.

[19] D.S. Brauer, C. Rüssel, J. Kraft, Solubility of glasses in the system P2O5–CaO–MgO–Na2O–TiO2: experimental and modeling using artificial neural networks, J. Non-Cryst. Solids 353 (2007) 263–270, https://doi.org/10.1016/j.jnoncrysol.2006.12.005.

[20] R. Ravinder, K.H. Sridhara, S. Bishnoi, H.S. Grover, M. Bauchy, Jayadeva, H. Kodamana, N.M.A. Krishnan, Deep learning aided rational design of oxide glasses, Mater. Horiz. 7 (2020) 1819–1827, https://doi.org/10.1039/d0mh00162g.

[21] D.R. Cassar, A.C.P.L.F. de Carvalho, E.D. Zanotto, Predicting glass transition temperatures using neural networks, Acta Mater. 159 (2018) 249–256, https://doi.org/10.1016/j.actamat.2018.08.022.

[22] B. Deng, Machine learning on density and elastic property of oxide glasses driven by large dataset, J. Non-Cryst. Solids 529 (2020), 119768, https://doi.org/10.1016/j.jnoncrysol.2019.119768.

[23] Epam/SciGlass, EPAM Systems, 2019. https://github.com/epam/SciGlass. (Accessed 6 November 2019).

[24] D.R. Cassar, GlassNet: a Multitask Deep Neural Network for Predicting Many Glass Properties, 2021, https://doi.org/10.48550/arXiv.2103.03633. ArXiv:2303.15538 [Cond-Mat, Soft]. (Accessed 5 March 2021).

[25] E. Alcobaça, S.M. Mastelini, T. Botari, B.A. Pimentel, D.R. Cassar, A.C.P.L.F. de Carvalho, E.D. Zanotto, Explainable machine learning algorithms for predicting glass transition temperatures, Acta Mater. 188 (2020) 92–100, https://doi.org/10.1016/j.actamat.2020.01.047.

[26] L. Breiman, Classification and Regression Trees, Routledge, 2017.

[27] T. Cover, T. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theor. 13 (1967) 21–27.

[28] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[30] V.R. Joseph, Optimal ratio for data splitting, Stat. Anal. Data Min. 15 (2022) 531–538, https://doi.org/10.1002/sam.11583.

[31] P. Domingos, A few useful things to know about machine learning, Commun. ACM 55 (2012) 78–87, https://doi.org/10.1145/2347736.2347755.

[32] A.J. Ellison, S. Gomez, Y. Kato, Alkali-free Boroalumino silicate glasses, U.S. Patent 0 339 468 (Jul. 2020).

[33] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (2020) 56–67, https://doi.org/10.1038/s42256-019-0138-9.

[34] D. Chicco, M.J. Warrens, G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, PeerJ Comput. Sci. 7 (2021) e623, https://doi.org/10.7717/peerj-cs.623.

[35] G. Jekabsons, J. Lavendels, V. Sitikovs, Model evaluation and selection in multiple nonlinear regression analysis, Mach. Model Anal. 12 (2007) 81–90, https://doi.org/10.3846/1392-6292.2007.12.81-90.

[36] E. Bertini, G. Santucci, Quality metrics for 2d scatterplot graphics: automatically reducing visual clutter, in: Proceedings of 4th International Symposium on SmartGraphics, May 2004.

[37] A.K. Varshneya, J.C. Mauro, Fundamentals of Inorganic Glasses, third ed., Elsevier, Amsterdam, 2019.

[38] M. Kazembeyki, K. Yang, J.C. Mauro, M.M. Smedskjaer, M. Bauchy, C.G. Hoover, Decoupling of indentation modulus and hardness in silicate glasses: evidence of a shear- to densification-dominated transition, J. Non-Cryst. Solids 553 (2021), 120518, https://doi.org/10.1016/j.jnoncrysol.2020.120518.

[39] D.R. Cassar, G.G. Santos, E.D. Zanotto, Designing optical glasses by machine learning coupled with a genetic algorithm, Ceram. Int. 47 (2021) 10555–10564, https://doi.org/10.1016/j.ceramint.2020.12.167.

[40] K. Nakamura, N. Otani, T. Koike, Multi-objective Bayesian optimization of optical glass compositions, Ceram. Int. 47 (2021) 15819–15824, https://doi.org/10.1016/j.ceramint.2021.02.155.

[41] L. Ding, Y. Xu, R. Yang, Y. Yang, R. Lu, H. Liu, H. He, Q. Zheng, J.C. Mauro, Lateral-pushing induced surface lift-up during nanoindentation of silicate glass, J. Am. Ceram. Soc. 105 (2022) 2625–2633, https://doi.org/10.1111/jace.18289.

[42] L. Ding, C. Qu, Y. Yang, C.J. Wilkinson, K.H. Lee, A.V. DeCeanne, K. Doss, J. C. Mauro, Dilatometric fragility and prediction of the viscosity curve of glass-forming liquids, J. Am. Ceram. Soc. 103 (2020) 4248–4255, https://doi.org/10.1111/jace.17125.