Predicting boron coordination in multicomponent borate and borosilicate glasses using analytical models and machine learning

Xiaonan Lu[1], Lu Deng[2], Jincheng Du[2], John D. Vienna[1,*]

[1] Energy and Environment Directorate, Pacific Northwest National Laboratory, Richland, WA 99352, USA

[2] Department of Materials Science and Engineering, University of North Texas, Denton, TX 76203, USA

[*]Corresponding Author: John.Vienna@pnnl.gov

Abstract

Accurate prediction of boron coordination in multicomponent glasses is critical in glass science and technology as it strongly affects the properties of borate and borosilicate glasses. We have collected a dataset containing 657 glasses from literature with boron coordination values and developed models using analytical functions based on the well accepted Dell, Xiao and Bray model. Good prediction of boron coordination with a R^2 value higher than 0.8 was obtained. The large variation of boron coordination from experiments, originated from sample preparations and characterizations, led to difficulties in obtaining models with better prediction performance. Various machine learning (ML) algorithms were evaluated and slightly better prediction performance was observed; however, interpretation of the ML models is less straight forward. This study developed various models capable of providing quantitative boron coordination predictions, providing insights into its structural roles in multi-component glasses, and suggesting fruitful areas for future research.

Keywords: boron coordination, model prediction, multicomponent glasses, machine learning, K-nearest neighbor, artificial neural network, Gaussian process regression

1. Introduction

Borosilicate glasses have a wide range of applications ranging from labware and cookware, to electronic display devices, biomedical implants and nuclear waste disposals, owing to their desirable physical and chemical properties. Accurate prediction of boron coordination in these glasses is critical as many physical and chemical properties are directly related to the coordination state of boron. The well-known boron anomaly in borate or borosilicate glasses is such an example and maybe one of the first direct evidences of structure-property correlations in glasses [1–4]. For most components, there is a unique

coordination environment irrespective of glass composition. For example, Si and Zr in glasses are mostly found to be 4 and 6 coordinated with oxygen, respectively [5]. However, boron coordination can change between trigonal planar ($^{[3]}$B) and tetrahedral ($^{[4]}$B), depending on overall glass composition and the thermal history. In pure boron oxide glass, boron coordination is 3. As alkali oxide is added to the glass matrix some $^{[3]}$B are converted to $^{[4]}$B with the alkali ions balancing the charge of the $[BO_4]^-$ units. As the alkali concentration continue to increase, the $N_4 = {}^{[4]}B/({}^{[3]}B+{}^{[4]}B)$ reaches a maximum and begins to decrease with additional alkali ions. This coordination change of boron as a function of composition is known as the origin of the "boron anomaly" phenomenon observed in these glasses [1–4].

Yun and Bray [6] first proposed a model to calculate N_4 based on two ratios: $R = Na_2O/B_2O_3$ and $K = SiO_2/B_2O_3$ from fitting nuclear magnetic resonance (NMR) results of three-component Na_2O-B_2O_3-SiO_2 glasses. The calculation of the N_4 values is divided into three regions: 1) $R < R_{max} = K/16+0.5$; 2) $R \geq R_{max}$ and $K \leq 8$; 3) $R \geq R_{max}$ and $K > 8$. A few years later, this model was further refined for $K \leq 8$ and commonly known as the Dell and Bray or Dell, Xiao and Bray model [7]. Built upon these results and models, Du and Stebbins [8] proposed a modified 'Dell and Bray's model for boroaluminosilicate glasses by adding Al_2O_3 in a modified R and K ratio, where $R' = Na_2O/(B_2O_3+Al_2O_3)$ and $K' = SiO_2/(B_2O_3+Al_2O_3)$. In this model, aluminum is assumed to play the same structural role as 4-fold coordinated boron.

In addition to the empirical models mentioned above, several thermodynamic based models have also been proposed. Araujo [9] proposed a statistical mechanical model to calculate N_4 in alkali aluminoborosilicate glasses. More recently, Smedskjaer et al. [2] developed a two-state statistical mechanical model, which can describe the relationship between the boron coordination change and the thermal history. This model accounts for the enthalpy difference between conversion of $^{[3]}$B to $^{[4]}$B and non-bridging oxygen (NBO) formation energy in silicate network when modifiers are added [2], which can capture the boron coordination of studied soda-lime borosilicate glasses. However, this model does not apply to compositions with high modifier content, and needs input of experimental glass transition temperature (T_g) value [2].

The quantitative-structure-property relationships (QSPR) analysis has evolved to be a promising approach to find structure-property correlations in glasses by using well designed structural descriptors [10]. The impacts of glass composition, atomic structure, and chemical

bonding environments are incorporated in the descriptor definition, which was used to predict glass properties [11–16]. As $^{[3]}$B and $^{[4]}$B have significant effects on glass properties, the inability to calculate N_4 limits the ability to apply QSPR and related analyses to multi-component alkali-borosilicate glasses. A new tool to estimate N_4 of both borate and borosilicate glasses is also necessary to calculate other structural parameters, including fraction of tetrahedral species [17], non-bridging oxygen concentrations [17], topological constraint theory [2], network connectivity [18]; each of which have been used for predicting glass properties. Additionally, accurate calculation of N_4 from glass compositions helps to develop transferable empirical potential parameters (e.g., composition-dependent parameters [19,20]) for studying glass structures using molecular dynamics (MD) simulations. However, to the best of the authors knowledge, no one has yet developed a model to predict N_4 values in multicomponent glasses such as nuclear waste glasses. Therefore, to improve the ability to predict properties of multicomponent glasses as a function of compositions, a more systematic and accurate prediction of compositional effects on B coordination is needed.

The objective of this work is to estimate the effects of glass composition on the N_4 values and develop various models to calculate N_4 in multicomponent borate and borosilicate glasses. In this paper, we fitted models with analytical functional forms including ones based on Du and Stebbins' (DS) model [8] and Bernstein equation [21,22]. In addition, machine learning (ML) approaches were applied to show the ability of different ML algorithms to predict N_4 values in multicomponent glasses. Even though ML has been recently used in predicting glass properties such as glass transition temperature [23], density [24], elastic modulus [24,25] and dissolution [26–29], it has not been used in prediction of glass structural properties to the best of our knowledge.

2. Methodology

2.1. Data collections

Data used in this paper were collected from the literature, and a compiled data table with references can be found in supplementary materials. A dataset containing 657 borate and borosilicate glasses was collected along with the sample preparation details (e.g., cooling rate, pressure). Table 1 shows the component (mol%) ranges of the dataset.

Table 1. Component (mol%) ranges of the dataset.

Component	Min	Max	Component	Min	Max
Al_2O_3	0.00	30.00	**ZrO_2**	0.00	8.00
B_2O_3	1.25	100.00	**BaO**	0.00	50.40
CaO	0.00	50.00	**Bi_2O_3**	0.00	50.00
Fe_2O_3	0.00	9.94	**Cs_2O**	0.00	58.33
K_2O	0.00	60.00	**HfO_2**	0.00	45.10
Li_2O	0.00	60.00	**La_2O_3**	0.00	47.00
MgO	0.00	40.00	**PbO**	0.00	80.00
Na_2O	0.00	69.97	**SrO**	0.00	12.50
P_2O_5	0.00	36.00	**Rb_2O**	0.00	44.99
SiO_2	0.00	87.30	**Y_2O_3**	0.00	43.00
ZnO	0.00	40.00			

Minor components include Cr_2O_3 (<0.03 mol%), UO_3 (<5 mol%), SO_3 (<0.1 mol%) and TiO_2 (<2.5 mol%) are contained in a few glasses of the data collected; however, these oxides were not considered during model fittings.

2.2. Models based on novel functional forms

Built upon the Dell, Bray and Xiao (DBX) model for sodium borosilicate glasses, Du and Stebbins proposed a model of boron coordination change with composition for boroaluminosilicate glasses based on NMR studies [8]. In the Du and Stebbins' (DS) model [8], Al was assumed to be 4-fold coordinated and has the same structural role as B and requires a charge compensator to balance the charge on $[AlO_4]^-$ units. Although for some compositions, especially those with high alumina content, Al can also be 5- and 6-coordinated in a few glass systems [30–33]. The DS model was successfully used to predict the N_4 in four-component alkali/alkali earth aluminoborosilicate glasses in certain composition regions (see [34,35] for examples). However, studies have observed that N_4 significantly decreases (from ~62% to ~13%) with Al_2O_3 replacing ~20 mol% B_2O_3 in alkali boroaluminosilicate glasses [34,36], where the original DS model cannot capture this change of N_4 accurately. Additionally, alkali and alkaline earth may have different effects on N_4 values [31,37–39], and it was found that higher modifier cation field strength promotes the formation of NBO instead of tetrahedral boron [40]. When the original DS model was used to calculate N_4 of the whole dataset, it was found to poorly predict with an R^2 value of 0.204 without removing any outliers. R^2 value increases to 0.75 after removing 100 data with a squared residual higher than 0.07. Therefore, we proposed to fit models with coefficients added for different oxides when calculating R and K ratios.

To enable calculations of boron N_4 in multicomponent glasses, modified R and K ratios are defined as follows (equation (1) and (2)) to add other components:

$$R = \frac{Na_2O + \sum a_i M_i}{B_2O_3 + \sum b_i NF_i} \quad (1)$$

$$K = \frac{SiO_2}{B_2O_3 + \sum b_i NF_i} \quad (2)$$

where:

- a_i are the coefficients of the i^{th} glass modifier. M_i are the concentrations (mol%) of the i^{th} glass modifier, including alkali metal oxides, alkaline earth metal oxides, ZnO, PbO, La_2O_3, Y_2O_3 and Bi_2O_3.
- b_i are the coefficients of the i^{th} glass network former or intermediate (NF). NF_i are the concentrations (mol%) of the i^{th} NF, including Al_2O_3, P_2O_5, Fe_2O_3, TiO_2, ZrO_2 and HfO_2.

Following the DBX and DS models, R_{max} (composition regions with maximum N_4) and R_D (composition regions that N_4 starts decreasing with increasing R) were refitted using equation (3) and (4):

$$R_{max} = C_1 K + C_2 \quad (3)$$

$$R_D = D_1 K + D_2 \quad (4)$$

where:

- C_1 and C_2 are the fitted slope and constant for R_{max}, respectively.
- D_1 and D_2 are the fitted slope and constant for R_D, respectively.

N_4 values were calculated using two different functional forms including one based on DS model (referred to as modified DS model hereafter) shown in equation (5), and the other one based on the Bernstein equation. [21] (referred to as modified Bernstein model hereafter) shown in equation (6):

$$N_4 = \begin{cases} R, & R < R_{max} \\ R_{max}, & R_{max} < R < R_D \\ R_{max} - \dfrac{(R - R_D)(8 + K)}{12(2 + K)}, & R > R_D \end{cases} \quad (5)$$

$$N_4 = \begin{cases} R, & R < R_{max} \\ a \times (b + R) \times (c - \dfrac{R}{d + e \times K})^5, & R \geq R_{max} \end{cases} \quad (6)$$

where $a, b, c, d\ and\ e$ are fitting parameters for Bernstein equation.

Functional-form models were fitted using Microsoft Excel™ Solver with a GRG Nonlinear method to maximize R^2.

2.3. Machine learning algorithms

Several ML algorithms including k-nearest neighbor (KNN), Gaussian process regression (GPR), and artificial neural network (ANN) have been used in this work and below is a brief summary of these techniques.

KNN is a classical and relatively simple method for pattern classification, which was developed from the need to perform discriminant analysis [41]. KNN uses simple distances such as Euclidean distance ($\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$) to measure the dissimilarities between samples represented as vector inputs [42]. In KNN regression, the prediction can be calculated based on the mean of its k-nearest neighbors [43].

GPR is an algorithm based on the Bayes' theorem, which estimates and evaluates the probability of a given outcome based on prior knowledge [27]. The process is updated iteratively throughout the training process, and the outcome with the highest probability is used as the final prediction [27]. GPR accommodates a wide class of nonlinear regression functions, and addresses problems with multidimensional functional covariates [44].

ANN algorithm was inspired by the behavior of biological neurons, which contains input, hidden, and output layers. These layers are made of artificial neurons that are interconnected by artificial synapses [45]. The hidden layers include a given number of neurons that take inputs from the input layer, apply mathematical functions and provide outputs to the output layer [26]. Each neuron weights the received input values by its associated synapse, where desired outputs are obtained by modifying these weights [45].

ANN and GPR models were fitted to the dataset using JMP™ version 14.0.0 (SAS Institute, Cary, NC). KNN regression model was fitted to the dataset using Python programming language with *sklearn* [43] and *pandas* [46] packages. 75% of the dataset was

randomly picked for training the ML models, and the remaining 25% of the dataset was used for testing the prediction performance of the models. This splitting is commonly used, which permits adequate optimizations of model parameters and predictions of model performance [27].

3. Results

3.1. Data curation and source of discrepancy

It is known that glass sample preparation details (e.g., thermal history, pressure, composition uncertainty) can significantly affect the boron coordination. For instance, it was found that higher cooling rate (lower fictive temperature) results in lower boron N_4 in sodium borosilicate glass [47], E-glass [48], and aluminoborosilicate glasses [49,50], while annealing samples increases N_4 [36].

Table 2 shows an example of different preparation methods and thermal history on boron N_4 of a sodium borosilicate glass. N_4 obtained by NMR for a $25B_2O_3$-$25Na_2O$-$50SiO_2$ glass can range between 58% to 85% depending on the thermal history. Even when studies used the same air-quenching methods, there is a 6% difference on N_4 obtained. In a study by Martens et al. [51], samples with the same compositions were prepared several times independently in order to check the quality of the sample and results. The difference of the N_4 for samples (quenched on a copper mould cooled with water) with the same composition can be up to 25 absolute% (e.g., N_4 obtained was 30% or 55% for a $22.2Na_2O$-$26.0B_2O_3$-51.8 SiO_2 glass in mol%) [51]. On the other hand, higher pressure leads to a higher boron N_4, as found in calcium aluminoborosilicate glasses [31] and sodium borosilicate glass [47]. It was also demonstrated that heavy ion irradiation can depolymerize the borosilicate glass network and results in a lower boron coordination number [52]. Additionally, sample form has an impact on boron N_4 determined by X-ray absorption fine structure, where using glass powder samples yields a higher N_4 as compared to using glass coupons [53].

In order to better understand the discrepancy in the collected dataset, mean squared pure error (MSPE) and relative standard deviation (%RSD) were calculated for 57 replicate sets with a total of 133 glasses. Sum of squares due to pure error (SSPE) is considered as a best attainable measurement error since it represents error among replicate glasses. MSPE, degrees of freedom for pure error (DFPE) and SSPE can be calculated using the following equations:

$$MSPE = \frac{SSPE}{DFPE} \tag{7}$$

$$DFPE = \sum_{j} (n_{rep\ set\ j} - 1) \tag{8}$$

$$SSPE = \sum_{j} \sum_{i} (y_{i,rep\ set\ j} - \bar{y}_{rep\ set\ j})^2 \tag{9}$$

where $n_{rep\ set\ j}$ is the number of glasses in replicate set *j*, $y_{i,rep\ set\ j}$ is the experimental N_4 value for glass *i* in replicate set *j,* and $\bar{y}_{rep\ set\ j}$ is the average experimental N_4 in replicate set *j.*

For the collected dataset, MSPE and SSPE are 0.006 and 0.485, respectively. %RSD for each replicate set was calculated by dividing standard deviation of glasses in the replicate set by their average. The average %RSD of all the replicate sets was found out to be 14.3%, where it can go up to 69% in the most extreme case. The pooled standard deviation (a weighted average of standard deviations) is 4.5 absolute%. The results show the challenges to predict boron coordination in silicate glasses.

Table 2. Reported boron N_4 values from ^{11}B NMR for $25B_2O_3$-$25Na_2O$-$50SiO_2$ with different thermal history (cooling rate and annealing) from literature.

Ref	N_4 (%)	Cooling rate	Annealing
[36]	85	Slow cooled (10 °C/min)	No
[6]	58	Air-quenched	No
[54]	60	Air-quenched	No
[55]	64	Air-quenched	No
[56]	67	Water-quenched	Annealed at T_g for 3 h
[51]	70 and 73	Water-quenched	No

3.2. Models using functional forms

Firstly, the coefficient for each oxide (defined in equation (1) and (2)) was fitted to obtain highest coefficient of determination (R^2), and then the coefficients were simplified to ease the future calculations. R^2 slightly decreased (less than 0.05) after simplifying the coefficients. In general, the coefficient of a modifier is found to be the inverse of its valence. For instance, the coefficients for Li_2O, CaO and La_2O_3 are 1, 1/2 and 1/3, respectively. The coefficients for P_2O_5, Fe_2O_3, TiO_2, HfO_2, Cr_2O_3, UO_3 and SO_3 are simplified to 0, where the fitted coefficients of these oxides are less than 0.5. Coefficients for Al_2O_3 and ZrO_2 are 4 and 3, respectively. A series of different scaling factors between component concentrations and their coefficients were attempted without success including cation field strength (values can be found in supplementary materials), optical basicity coefficient [57], and single bond energy [58].

Table 3 lists the fitting parameters of the modified DS models and modified Bernstein models based on the definitions of equation (1) and (2) for the whole dataset (borosilicate and borate glasses) and sub-dataset (only borosilicate glasses). Data with a squared residual, $(predicted\ N_4 - measured\ N_4)^2$, higher than 0.07 were removed as outliers. R^2, RMSE (root mean square error), data size (before removing outliers) and number of outliers removed for each functional-form model are listed in Table 3. For the modified Bernstein equation, it was found that correlations exist between the fitting parameters, which are indicated in the parenthesis. In addition, a beta function was unsuccessfully attempted to describe the effects of R and K on N_4.

Table 3. Fitting parameters, R^2, RMSE, data size and number of outliners removed of the modified DS models and modified Bernstein models for the whole dataset (borosilicate and borate glasses) and only borosilicate glasses.

Coefficient/ parameters	**Modified DS model (whole dataset)**	**Modified DS model (borosilicate)**	**Coefficient/ parameters**	**Modified Bernstein model (whole dataset)**	**Modified Bernstein model (borosilicate)**
R_{max}	0.019K+0.590	0.022K+0.580	R_{max}	0.06K+0.43	0.06K+0.43
R_D	0.502K+0.096	0.459K+0.256	a	0.25	0.71

			b	0.29	0.40
			c	1.30	1.03
			d	3.78	5.47 (5.63/*c*)
			e	1.00 (*d*/3.78)	1.19 (1.16×*c*)
R^2	0.816	0.825		0.830	0.836
RMSE	0.0883	0.0887		0.0836	0.0854
Data size	657	451		657	451
Number of outliers removed	11	11		19	17

Figure 1 shows the measured versus predicted N_4 plot using the modified DS model for the whole dataset and the sub-dataset. R^2 of the modified DS model for the whole and subset data are 0.816 and 0.825, respectively. Figure 2 shows the measured versus predicted N_4 plot using modified Bernstein models for the whole dataset and the subset with only borosilicate glasses. R^2 of the modified Bernstein model for the whole and subset data are 0.830 and 0.836, respectively. Figure 3 shows the N_4 obtained from the modified DS and Bernstein models of the whole dataset in comparison with experimental values as a function of R at different K values. As shown in Figure 3, both models provide relatively good predictions when $R < \text{and} >> R_{max}$. Modified Bernstein model provides better predictions around R_{max} compared to modified DS model.

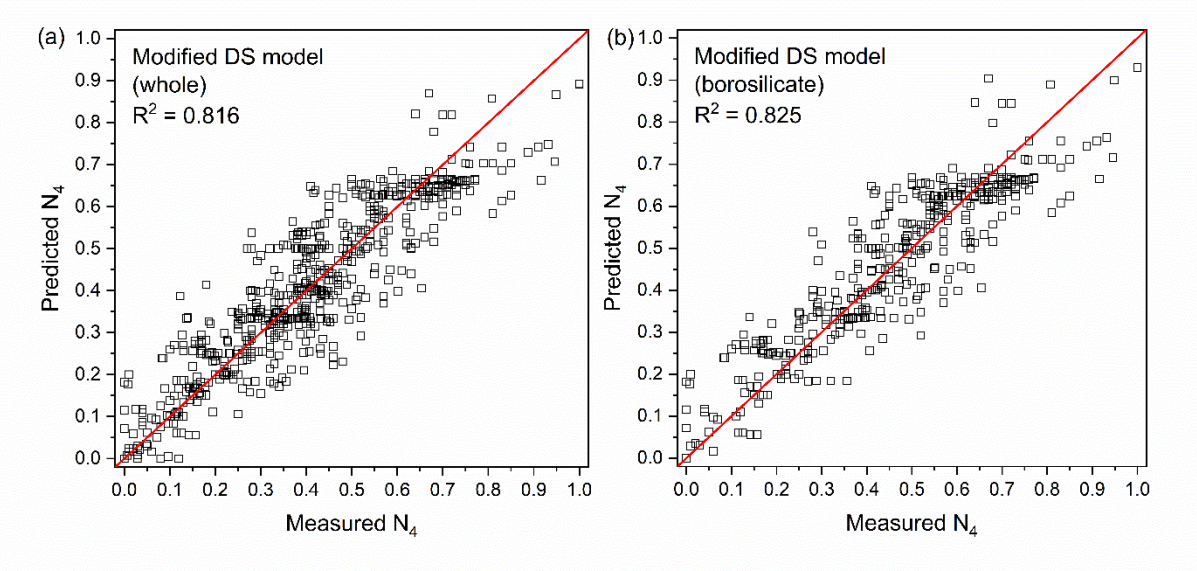

Figure 1. Measured versus predicted N_4 plot using modified DS models for (a) the whole dataset and (b) the sub-dataset with only borosilicate glasses.

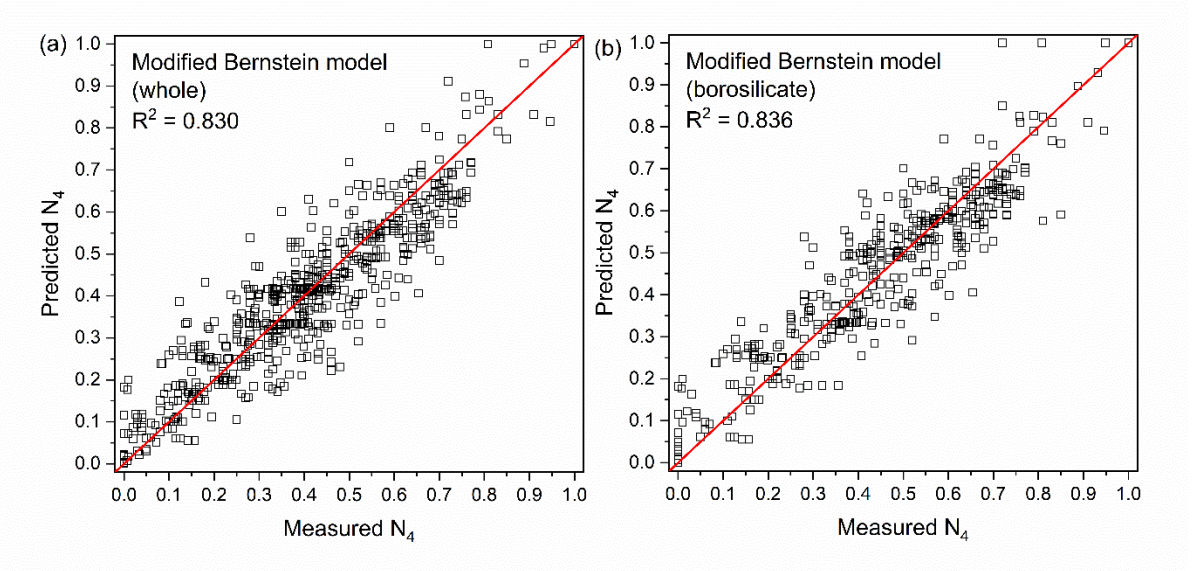

Figure 2. Measured versus predicted N_4 plot using modified Bernstein models for (a) the whole dataset and (b) the subset with only borosilicate glasses.

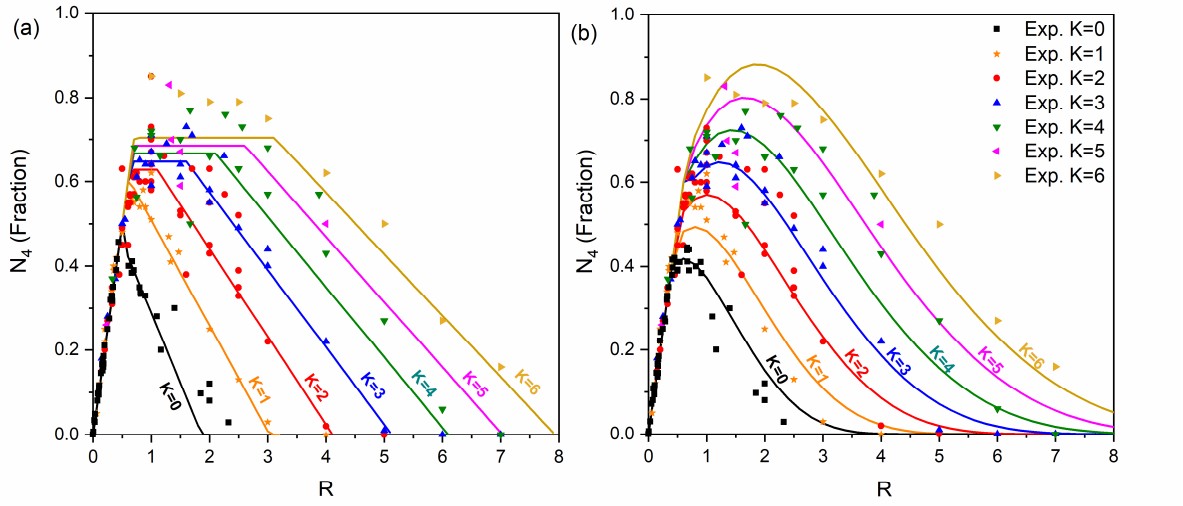

Figure 3. N_4 obtained from the modified DS model (a) and the modified Bernstein model (b) (solid lines) of the whole dataset in comparison with experimental values (dots) as a function of R at different K values.

3.3. Models from machine learning approaches

When fitting KNN models, uniform weights and standard Euclidean metric with varying the number of nearest neighbors were used. Figure 4 (a) shows the R^2 of training and testing as a function of the number of neighbors used during the model fitting. When three nearest neighbors are used, it provides the highest R^2 (0.814) for the testing dataset and a relatively high R^2 (0.923) for the training dataset. RMSE for the training and testing dataset are 0.058 and 0.090, respectively. Figure 4 (b) shows the measured vs. predicted N_4 using KNN regression (number of neighbors is 3). No data were removed as outliers, mostly because of that the outliers are dependent on the separation of training and testing dataset. However, this might also result in over fitting the model to the outliers. In this KNN model, all data in the training set have a squared residual less than 0.07 and three glasses in the testing set have a squared residual higher than 0.07 (up to 0.081).

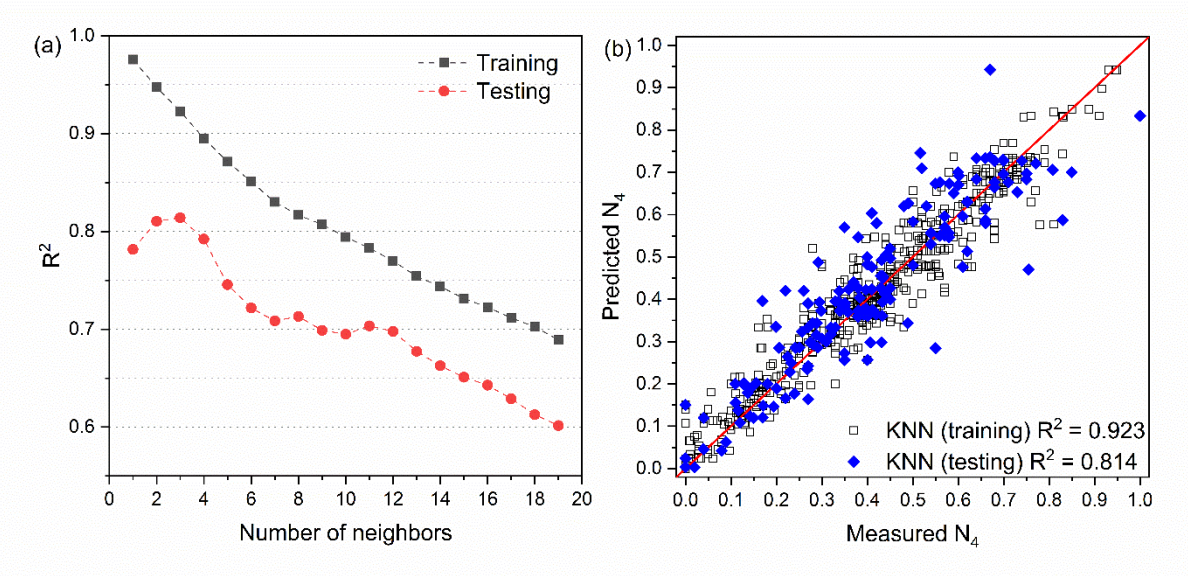

Figure 4. (a) R^2 of training and testing as a function of the number of neighbors. (b) Measured versus predicted N_4 plot using K-nearest neighbor (number of neighbors is 3) for the whole dataset. 75% of data in the whole dataset was used for training model parameters and the rest of 25% was using for testing model performance.

ANN fitting was performed with the whole dataset. The number of nodes was varied from 2 through 5 using a TanH activation function (hyperbolic tangent function, which is a sigmoid function) with one hidden layer. A 3-node model was found to give the best results (R^2_{test} closest to R^2_{train}). Figure 5 shows the measured versus predicted N_4 plot using ANN. However, the results highly depend on the random seed number, where the R^2 values vary between ~0.6 to ~0.85. The results of ANN model shown in Figure 5 were obtained using a random seed of 14. No data were removed as outliers, while seven data have a squared residual higher than 0.07.

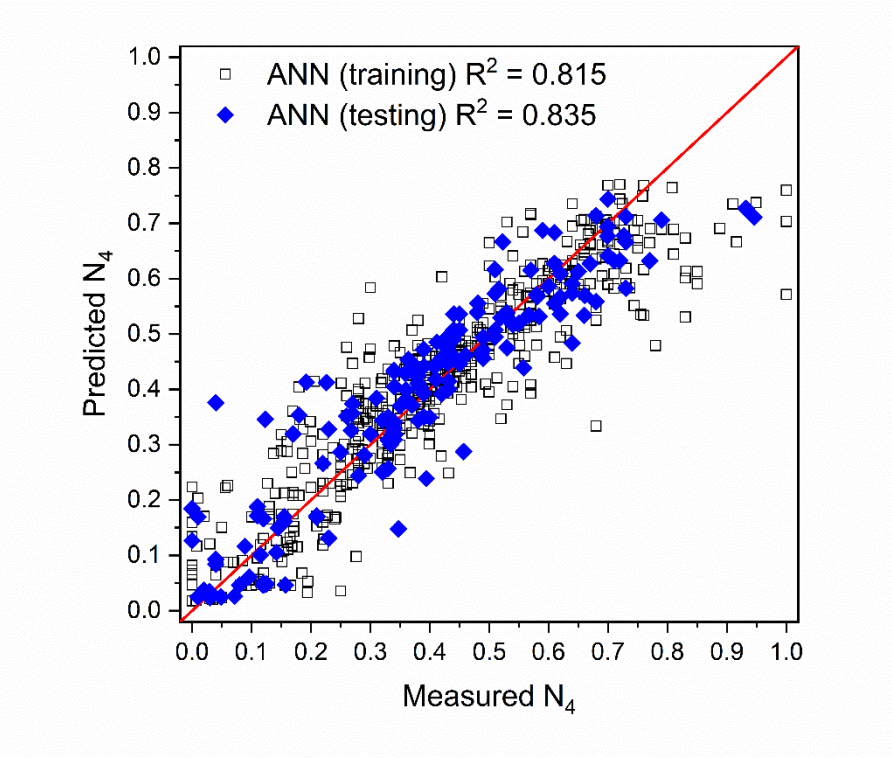

Figure 5. Measured versus predicted N_4 plot using artificial neural network for the whole dataset. 75% of data in the whole dataset was used for training model parameters and the rest of 25% was using for testing model performance.

Figure 6 shows the measured versus predicted N_4 plot using Gaussian process regression, where the model parameters were fitted via maximum likelihood. The final GPR model has a Gaussian process mean (μ) of 7.7902, variance (σ^2) of 3820 and a Nugget parameter of 0.01. This model provides the highest R^2 (0.926 for training and 0.911 for testing) in this study. RMSE of the GPR model for the training and testing dataset are 0.057 and 0.062, respectively. No data were removed as outliers, while only one glass have a squared residual higher than 0.07. In addition, supporting vector machine (SVM) and random forest regression using the *sklearn* package [43] were tested, where R^2 values are less than 0.5 for both regression methods hence the results were not reported in this paper.

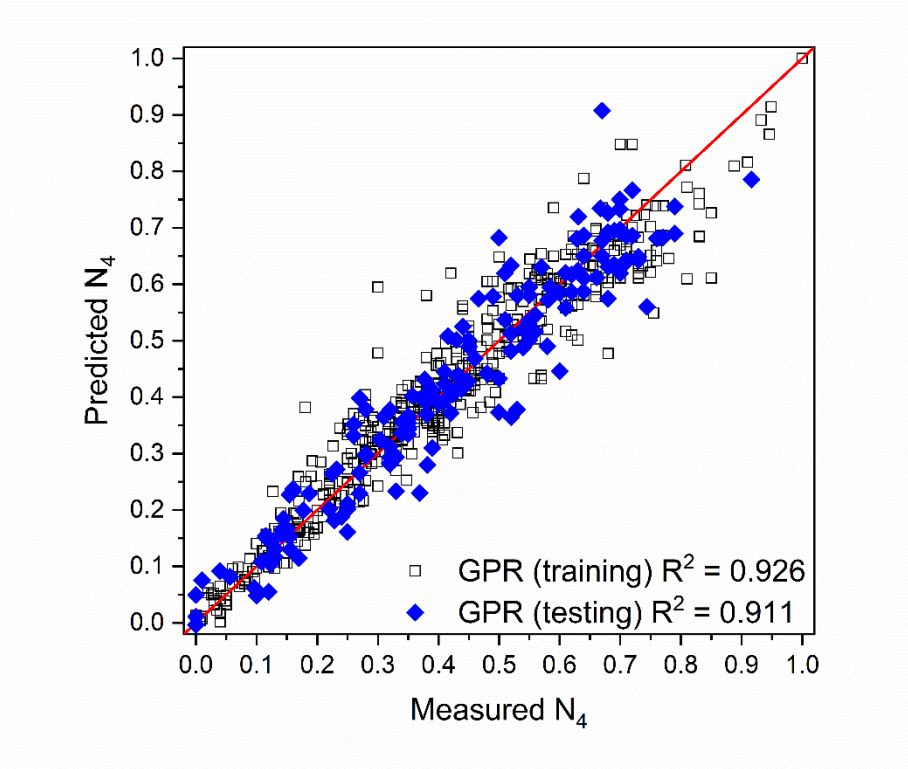

Figure 6. Measured versus predicted N_4 plot using Gaussian process regression for the whole dataset.

4. Discussion

For the first time, models have been developed to predict N_4 of multi-component borate and borosilicate glasses with over 600 experimental compositions using analytical functions and machine learning approaches. These models range in RMSE from 0.06 to 0.09 and R^2 from 0.81 to 0.93, as shown in Table 4. Modified Bernstein models provide a slightly higher R^2 compared to modified DS models; on the other hand, GPR model shows the best model performance among all the ML models tested.

Table 4. Summary of the R^2 and RMSE of functional form and ML models.

Models	R^2	RMSE
Modified DS (whole dataset)	0.816	0.0883
Modified DS (borosilicate)	0.825	0.0887
Modified Bernstein (whole dataset)	0.830	0.0836
Modified Bernstein (borosilicate)	0.836	0.0854
KNN (training)	0.923	0.0579

KNN (testing)	0.814	0.0898
GPR (training)	0.926	0.0568
GPR (testing)	0.911	0.0620
ANN (training)	0.815	0.0900
ANN (testing)	0.835	0.0829

Models using functional forms can be extrapolated to a wide range of compositions and give a reasonable estimation of the N_4; however, many assumptions were made during fitting such as structural role of glass components. For example, Fe_2O_3 was assumed to be glass network former during the model fitting; whereas, the structural role of Fe ions depends on its charge state and coordination in glasses and redox ratio of iron oxide was not always reported in the studies [59]. Likewise, other components may take on multiple different structural roles across the breadth of compositions tested (e.g., Al^{3+} may be found in 4, 5, or 6 coordination [30–33]). Also, alkali which are treated the same in this model have previously been shown to effect N_4 differently [31,37–39]. Cooling rates of the data are different from study to study and even within studies. Previous studies showed higher N_4 for more slowly cooled glasses [47–50]. An attempt to correct for this cooling rate effect was made. Models were fit with after separating data by air-quenched or water-quenched glasses; however, this caused a slight decrease in R^2 value, indicating that the difference caused by thermal history is within experimental uncertainty (e.g., from NMR characterization). It was thought that the composition region may be too large to develop a single model covering the entire space. Therefore, models were fitted using only borosilicate glasses and removing the borate-based glasses. The resulting models have a slightly higher R^2 in comparison with using the whole dataset. However, not sufficiently higher (e.g., 0.836 vs. 0.830 for the modified Bernstein model) to reduce the ability to predict over broader composition regions with a single model. The large discrepancy (e.g., average %RSD is about 14%) in the dataset caused by experimental uncertainty (discussed in section 3.1) makes it challenging to obtain models with very high R^2 values. A more consistent measurement protocol or standard for boron N_4 measurements would greatly benefit further development of theoretical calculations of N_4 from compositions.

In addition, boxplots and scatterplots were generated to study the composition ranges with different prediction performances, as well as the outliers removed for the four functional form models. The boxplots and scatterplots can be found in supplementary materials. After

comparing composition ranges of all data and removed outliers (squared residuals higher than 0.07) in each model, it was found that glasses with high Al_2O_3 and SiO_2, and low B_2O_3 are slightly more likely to be outliers in the models fitted using the whole dataset. This trend was not observed for models fitted using sub-dataset (borosilicate glasses only subset). No significant composition difference was observed when comparing between data with squared residual less than 0.02 and data with square residual higher than 0.02 (less than 0.07).

Machine learning studies show promising results and potentials for future improvement of the N_4 models. Even though KNN regression is one of the simplest algorithms in ML, it provides good prediction performance, which was also observed when predicting T_g [23] and glass dissolution [29]. It should be noted that KNN model performance can be different if different values of random state were selected during the separation of dataset into training and testing using *sklearn*. Random state of 40 was used in this study and the results are presented in section 3.3. From a test of 1000 random values, R^2 varies between ~0.7 to ~0.9. In addition, another approach that hyperparameters (e.g., number of neighbors and weights) of KNN models were tuned using GridSearchCV from *sklearn* with 5-fold cross-validation on training dataset was conducted; however, R^2 of testing dataset still varies between ~0.7 to ~0.9 with 1000 random state values. ANN provides promising results as well, while the results also highly depend on the random seed number and JMP version/installation used, where the R^2 values change between ~0.6 to ~0.85 with the same model configuration regardless of using 5-fold cross-validation or just one set of training/testing. Future works on study stability and reproducibility of ML models are thus needed. GPR provides the highest R^2 among all the models tested in this work; however, it is more computing time consuming than the other ML approaches. Some other ML algorithms such as SVM and random forest regression didn't seem to be able to fit the N_4 data with meaningful R^2 values. The SVM model (default parameters in *sklearn*) failed to predict N_4 in this study. This might be because no hyperplanes can be found to separate data, suggesting that the data are not linear separable. It was found that KNN and SVM work well if the data points are heterogeneously distributed, while for homogenous data adding a kernel into the SVM might improve the model performance [60]. Future tests with different kernels (e.g., polynomial and Gaussian) are needed. For random forest regression, further toning of hyperparameters (e.g., number of trees, maximum depth of the tree, minimum number of samples, etc.) would also be needed to evaluate the ability of random forest regression to predict N_4.

Overall, ML algorithms can be applied to multicomponent glasses without knowing the structural role of each glass component. However, these models do not provide physical interpretation of the fitting or any insights of the relationships between compositions and N_4. Many ML algorithms such as ANN and SVM have been referred to as black-box models, resulting in difficulties interpreting results [23,61]. Even though ANN and SVM are often considered as reliable models, they provide less accurate predictions for data with highly nonlinear features [27]. They are also prone to over fitting which is difficult to quantify. The models with structural based functional forms (modified DS models, and modified Bernstein function) provide similar level of fitting but allow for evaluation of impacts of new and untested/fitted glass components based on some understanding of their structural roles.

5. Conclusions

In this study, various models to predict boron N_4 for multicomponent glasses from compositions were developed and evaluated using a dataset with 657 borate and borosilicate glasses collected from literature. Models using functional forms such as Du and Stebbins (DS) models and the Bernstein function were fitted using a modified definition of the R and K values with additional fitting coefficient to each of the glass components. Even though several assumptions (e.g., structural role of glass components) have been made during model fitting, these models provide a relatively robust prediction of N_4 with a R^2 value higher than 0.830. Interestingly, models fitted using only borosilicate glasses have similar but a slightly higher R^2 value (increased by ~0.1) as compared to those using the whole dataset. Furthermore, different ML algorithms (e.g., k-nearest neighbor, artificial neural network and Gaussian process regression) to find correlation of boron N_4 and composition relations were evaluated. A R^2 value as high as 0.91 was obtained for the model using the Gaussian process regression. ML algorithms have the advantage of a slightly better prediction performance as compared to the models using functional forms; however, the relationships between composition and N_4 are less straight forward. More sophisticated ML approaches may improve fitting and provide insights of the composition effect. However, the large variance of experimental data originated from sample preparations and characterizations leads to difficulties in obtaining models with significantly better prediction performance than those models reported here. A more consistent experimental dataset with consistent N_4 measurement protocols would greatly benefit further development of the theoretical models to predict boron N_4 values in multicomponent oxide glasses.

Acknowledgements

This work was supported jointly by the Center for Performance and Design of Nuclear Waste Forms and Containers (WastePD), an Energy Frontier Research Center funded by the U.S. Department of Energy (DOE), Office of Science, Basic Energy Sciences under Award # DESC0016584, DOE Office of River Protection Waste Treatment and Immobilization Plant (WTP) Project, and US National Science Foundation DMR ceramics program (project # 1508001); who we thank for their generous support.

Data collection, tabulation, checking, and review were performed under WTP support. Model form definition and fitting was performed under WastePD support. Manuscript writing, review, and submission were joint efforts.

We would like to thank Scott Kroeker (University of Manitoba) for sending data of his previous publications, Sebastien Kerisit (PNNL) and Benjamin Parruzot (PNNL) for helping with data collection, Scott Cooley (PNNL) for helping with analyzing data discrepancy, Tongan Jin (PNNL) and Daniel Prather (PNNL) for reviewing the dataset, Dong-Sang Kim (PNNL) for reviewing the models and manuscript, and Renee Russell (PNNL) for reviewing the manuscript.

Pacific Northwest National Laboratory is a multi-program national laboratory operated for the U.S. Department of Energy by Battelle Memorial Institute under Contract DE-AC06-76RL01830.

References

[1] A.K. Varshneya, Fundamentals of inorganic glasses, Society of Glass Technology, Sheffield, 2006.

[2] M.M. Smedskjaer, J.C. Mauro, R.E. Youngman, C.L. Hogue, M. Potuzak, Y. Yue, Topological principles of borosilicate glass chemistry, J. Phys. Chem. B. 115 (2011) 12930–12946.

[3] S.S. Sørensen, H. Johra, J.C. Mauro, M. Bauchy, M.M. Smedskjaer, Boron anomaly in the thermal conductivity of lithium borate glasses, Phys. Rev. Mater. 3 (2019) 075601.

[4] F. Muñoz, L. Montagne, L. Pascual, A. Durán, Composition and structure dependence of the properties of lithium borophosphate glasses showing boron anomaly, J. Non. Cryst. Solids. 355 (2009) 2571–2577.

[5] G. Calas, L. Galoisy, L. Cormier, G. Ferlat, G. Lelong, The structural properties of cations in nuclear glasses, Procedia Mater. Sci. 7 (2014) 23–31.

[6] Y.H. Yun, P.J. Bray, Nuclear magnetic resonance studies of the glasses in the system Na_2O-B_2O_3-SiO_2, J. Non. Cryst. Solids. 27 (1978) 363–380.

[7] W.J. Dell, P.J. Bray, S.Z. Xiao, ^{11}B NMR studies and structural modeling of NaO-B_2O_3-SiO_2 glasses of high soda content, J. Non. Cryst. Solids. 58 (1983) 1–16.

[8] L.-S. Du, J.F. Stebbins, Network connectivity in aluminoborosilicate glasses: A high-resolution 11B, 27Al and 17O NMR study, J. Non. Cryst. Solids. 351 (2005) 3508–3520.

[9] R.J. Araujo, Statistical mechanical model of boron coordination, J. Non. Cryst. Solids. 42 (1980) 209–229.

[10] A. Pedone, M.C. Menziani, Computational modeling of silicate glasses: A quantitative structure-property relationship perspective, Springer Ser. Mater. Sci. 215 (2015) 113–135.

[11] L. Linati, G. Lusvardi, G. Malavasi, L. Menabue, M. Menziani, P. Mustarelli, U. Segre, Qualtitative and Quantitative Structure - Property Relationship Analysis of Multicomponent Potential Bioglasses, J. Phys. Chem. B. 109 (2005) 4989–4998.

[12] G. Malavasi, A. Pedone, M.C. Menziani, Towards a quantitative rationalization of multicomponent glass properties by means of molecular dynamics simulations, Mol. Simul. 32 (2006) 1045–1055.

[13] G. Lusvardi, G. Malavasi, L. Menabue, M.C. Menziani, A. Pedone, U. Segre, Density of multicomponent silica-based potential bioglasses: Quantitative structure-property relationships (QSPR) analysis, J. Eur. Ceram. Soc. 27 (2007) 499–504.

[14] A. Pedone, X. Chen, R.G. Hill, N. Karpukhina, Molecular Dynamics Investigation of Halide-Containing Phospho-Silicate Bioactive Glasses, J. Phys. Chem. B. 122 (2018) 2940–2948.

[15] X. Lu, L. Deng, S. Gin, J. Du, Quantitative Structure–Property Relationship (QSPR) Analysis of ZrO2-Containing Soda-Lime Borosilicate Glasses, J. Phys. Chem. B. 123 (2019) 1412–1422.

[16] X. Lu, J. Du, Quantitative structure-property relationship (QSPR) analysis of calcium

aluminosilicate glasses based on molecular dynamics simulations, J. Non. Cryst. Solids. 530 (2020) 119772.

[17] J.D. Vienna, J.J. Neeway, J. V. Ryan, S.N. Kerisit, Impacts of glass composition, pH, and temperature on glass forward dissolution rate, Npj Mater. Degrad. 2 (2018) 22.

[18] M. Edén, The split network analysis for exploring composition–structure correlations in multi-component glasses: I. Rationalizing bioactivity-composition trends of bioglasses, J. Non. Cryst. Solids. 357 (2011) 1595–1602.

[19] L. Deng, J. Du, Development of effective empirical potentials for molecular dynamics simulations of the structures and properties and boroaluminosilicate glasses, J. Non. Cryst. Solids. 453 (2016) 177–194.

[20] L. Deng, J. Du, Development of boron oxide potentials for computer simulations of multicomponent oxide glasses, J. Am. Ceram. Soc. 102 (2019) 2482–2505.

[21] L. Deng, Molecular dynamics simulations of the structure and properties of boron containing oxide glasses: empirical potential development and applications, 2017.

[22] S. Bernstein, Démo istration du th'eorème de Weierstrass fondée sur le calcul des probabilités, Comm. Soc. Math. Kharkov. 13 (1912) 1–2.

[23] E. Alcobaça, S.M. Mastelini, T. Botari, B.A. Pimentel, D.R. Cassar, A.C.P. de L.F. de Carvalho, E.D. Zanotto, Explainable Machine Learning Algorithms For Predicting Glass Transition Temperatures, Acta Mater. 188 (2020) 92–100.

[24] Y.-J.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) 1–13.

[25] 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.

[26] N.M. Anoop Krishnan, S. Mangalathu, M.M. Smedskjaer, A. Tandia, H. Burton, M. Bauchy, Predicting the dissolution kinetics of silicate glasses using machine learning, J. Non. Cryst. Solids. 487 (2018) 37–45.

[27] T. Han, N. Stone-Weiss, J. Huang, A. Goel, A. Kumar, Machine learning as a tool to

design glasses with controlled dissolution for healthcare applications, Acta Biomater. 107 (2020) 286–298.

[28] H. Liu, T. Zhang, N.M. Anoop Krishnan, M.M. Smedskjaer, J. V. Ryan, S. Gin, M. Bauchy, Predicting the dissolution kinetics of silicate glasses by topology-informed machine learning, Npj Mater. Degrad. 3 (2019) 32.

[29] J.N.P. Lillington, T.L. Goût, M.T. Harrison, I. Farnan, Predicting radioactive waste glass dissolution with machine learning, J. Non. Cryst. Solids. 533 (2020) 119852.

[30] N. Jakse, M. Bouhadja, J. Kozaily, J.W.E. Drewitt, L. Hennet, D.R. Neuville, H.E. Fischer, V. Cristiglio, A. Pasturel, Interplay between non-bridging oxygen, triclusters, and fivefold Al coordination in low silica content calcium aluminosilicate melts, Appl. Phys. Lett. 101 (2012) 201903.

[31] S. Bista, J.F. Stebbins, J. Wu, T.M. Gross, Structural changes in calcium aluminoborosilicate glasses recovered from pressures of 1.5 to 3 GPa: Interactions of two network species with coordination number increases, J. Non. Cryst. Solids. 478 (2017) 50–57.

[32] E.I. Morin, J.F. Stebbins, Separating the effects of composition and fictive temperature on Al and B coordination in Ca, La, Y aluminosilicate, aluminoborosilicate and aluminoborate glasses, J. Non. Cryst. Solids. 432 (2016) 384–392.

[33] L.M. Thompson, J.F. Stebbins, Non-stoichiometric non-bridging oxygens and five-coordinated aluminum in alkaline earth aluminosilicate glasses: Effect of modifier cation size, J. Non. Cryst. Solids. 358 (2012) 1783–1789.

[34] E.M. Pierce, L.R. Reed, W.J. Shaw, B.P. McGrail, J.P. Icenhower, C.F. Windisch, E.A. Cordova, J. Broady, Experimental determination of the effect of the ratio of B/Al on glass dissolution along the nepheline (NaAlSiO4)-malinkoite (NaBSiO4) join, Geochim. Cosmochim. Acta. 74 (2010) 2634–2654.

[35] Q. Zheng, R.E. Youngman, C.L. Hogue, J.C. Mauro, M. Potuzak, M.M. Smedskjaer, Y.Z. Yue, Structure of boroaluminosilicate glasses: impact of [Al2O3]/[SiO2] ratio on the structural role of sodium, Phys. Rev. B. 86 (2012) 054203.

[36] J. Marcial, M. Saleh, D. Watson, S.W. Martin, C.L. Crawford, J.S. McCloy, Boron-speciation and aluminosilicate crystallization in alkali boroaluminosilicate glasses along the NaAl1-xBxSiO4 and LiAl1-xBxSiO4 joins, J. Non. Cryst. Solids. 506

(2019) 58–67.

[37] J. Wu, J.F. Stebbins, Effects of cation field strength on the structure of aluminoborosilicate glasses: High-resolution11B,27Al and23Na MAS NMR, J. Non. Cryst. Solids. 355 (2009) 556–562.

[38] N. Ollier, T. Charpentier, B. Boizot, G. Wallez, D. Ghaleb, A Raman and MAS NMR study of mixed alkali Na-K and Na-Li aluminoborosilicate glasses, J. Non. Cryst. Solids. 341 (2004) 26–34.

[39] J.M. Tainio, D.A.A. Salazar, A. Nommeots-nomm, C. Roiland, B. Bureau, D.R. Neuville, D.S. Brauer, J. Massera, Structure and in vitro dissolution of Mg and Sr containing borosilicate bioactive glasses for bone tissue engineering, J. Non. Cryst. Solids. 533 (2020) 119893.

[40] J. Wu, J.F. Stebbins, T. Rouxel, Cation Field Strength Effects on Boron Coordination in Binary Borate Glasses, J. Am. Ceram. Soc. 97 (2014) 2794–2801.

[41] L.E. Peterson, K-nearest neighbor, Scholarpedia. 4 (2009) 1883.

[42] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (2009) 207–244.

[43] 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 Fabian, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[44] J. Shi, T. Choi, Gaussian Process Regression Analysis for Functional Data, 2011.

[45] 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.

[46] W. McKinney, Data structures for statistical computing in python, Proc. 9th Python Sci. Conf. (SCIPY 2010). 445 (2010) 51–56.

[47] L. Wondraczek, S. Sen, H. Behrens, R.E. Youngman, Structure-energy map of alkali borosilicate glasses: Effects of pressure and temperature, Phys. Rev. B - Condens. Matter Mater. Phys. 76 (2007) 1–8.

[48] T.J. Kiczenski, L.S. Du, J. Stebbins, The effect of fictive temperature on the structure of E-glass: A high resolution, multinuclear NMR study, J. Non. Cryst. Solids. 351

(2005) 3571–3578.

[49] J. Wu, J.F. Stebbins, Temperature and modifier cation field strength effects on aluminoborosilicate glass network structure, J. Non. Cryst. Solids. 362 (2013) 73–81.

[50] J. Wu, J.F. Stebbins, Quench rate and temperature effects on boron coordination in aluminoborosilicate melts, J. Non. Cryst. Solids. 356 (2010) 2097–2108.

[51] R. Martens, W. Müller-Warmuth, Structural groups and their mixing in borosilicate glasses of various compositions - an NMR study, J. Non. Cryst. Solids. 265 (2000) 167–175.

[52] C. Mendoza, S. Peuget, T. Charpentier, M. Moskura, R. Caraballo, O. Bouty, A.H. Mir, I. Monnet, C. Grygiel, C. Jegou, Oxide glass structure evolution under swift heavy ion irradiation, Nucl. Instruments Methods Phys. Res. Sect. B. 325 (2014) 54–65.

[53] J.P. Icenhower, B.P. McGrail, W.J. Shaw, E.M. Pierce, P. Nachimuthu, D.K. Shuh, E.A. Rodriguez, J.L. Steele, Experimentally determined dissolution kinetics of Na-rich borosilicate glass at far from equilibrium conditions: Implications for Transition State Theory, Geochim. Cosmochim. Acta. 72 (2008) 2767–2788.

[54] T. Nanba, M. Nishimura, Y. Miura, A theoretical interpretation of the chemical shift of 29Si NMR peaks in alkali borosilicate glasses, Geochim. Cosmochim. Acta. 68 (2004) 5103–5111.

[55] A.A. Osipov, V.E. Eremyashev, L.M. Osipova, Structure of Ca-Sr-Ba Sodium-Borosilicate Glasses according to 11B and 28Si NMR spectroscopy, Glas. Phys. Chem. 44 (2018) 71–77.

[56] J.M. Roderick, D. Holland, A.P. Howes, C.R. Scales, Density-structure relations in mixed-alkali borosilicate glasses by 29Si and 11B MAS-NMR, J. Non. Cryst. Solids. 293–295 (2001) 746–751.

[57] J.S. McCloy, M.J. Schweiger, C.P. Rodriguez, J.D. Vienna, Nepheline Crystallization in Nuclear Waste Glasses: Progress Toward Acceptance of High-Alumina Formulations, Int. J. Appl. Glas. Sci. 2 (2011) 201–214.

[58] K.H. Sun, Fundamental condition of glass formation, J. Am. Ceram. Soc. 30 (1947) 277–281.

[59] B. Cochain, D.R. Neuville, G.S. Henderson, C.A. McCammon, O. Pinet, P. Richet, Effects of the iron content and redox state on the structure of sodium borosilicate glasses: A Raman, Mössbauer and boron k-edge xanes spectroscopy study, J. Am. Ceram. Soc. 95 (2012) 962–971.

[60] A. Goel, S. Mahajan, Comparison: KNN & SVM Algorithm, Int. J. Res. Appl. Sci. Eng. Technol. 887 (2017) 2321–9653.

[61] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, Proc. - 2018 IEEE 5th Int. Conf. Data Sci. Adv. Anal. DSAA 2018. (2019) 80–89.