Acta Materialia 240 (2022) 118302

Contents lists available at ScienceDirect

Acta Materialia

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

Full length article

# Machine learning unveils composition-property relationships in chalcogenide glasses

Saulo Martiello Mastelini[a], Daniel R. Cassar[b,c,*], Edesio Alcobaça[a], Tiago Botari[a], André C.P.L.F. de Carvalho[a], Edgar D. Zanotto[c]

[a] Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos, Brazil
[b] Ilum School of Science, Brazilian Center for Research in Energy and Materials (CNPEM), Campinas, Brazil
[c] Department of Materials Engineering, Federal University of São Carlos, São Carlos, Brazil

ARTICLE INFO

*Article history:*
Received 14 June 2021
Revised 15 August 2022
Accepted 23 August 2022
Available online 26 August 2022

*Keywords:*
Chalcogenide glasses
Machine learning
Property prediction

ABSTRACT

Due to their unique optical and electronic functionalities, chalcogenide glasses are materials of choice for numerous microelectronic and photonic devices. However, to extend the range of compositions and applications, profound knowledge about *composition-property* relationships is necessary. To this end, we collected a large quantity of composition-property data on chalcogenide glasses from the SciGlass database regarding glass transition temperature ($T_g$), coefficient of thermal expansion (CTE), and refractive index ($n_D$). With these data, we induced predictive models using four machine learning algorithms: Random Forest, K-nearest Neighbors, Neural Network (Multilayer Perceptron), and Classification and Regression Trees. Finally, the induced models were interpreted by computing the SHapley Additive exPlanations (SHAP) values of the chemical features, which revealed the key elements that significantly impacted the tested properties and quantified their impact. For instance, Ge and Ga increase $T_g$ and decrease CTE (two properties that depend on bond strength), whereas Se has the opposite effect. Te, As, Tl, and Sb increase $n_D$ (which strongly depends on polarizability), whereas S, Ge, and P diminish it. The SHAP interaction analysis indicated two-element pairs that are likely to exhibit the mixed-former effect: arsenic-germanium and sulfur-selenium. Knowledge about the role of each element on the glass properties is precious for semi-empirical compositional development trials or simulation-driven formulations. The induced models can be used to design novel chalcogenide glasses with the required combinations of properties.



## 1. Introduction

Chalcogenide glasses contain one or more chalcogens (sulfur, selenium, and tellurium) and no oxygen. Their relatively small band gaps ($E_g$ = 1–3 eV) lead to optical and electrical properties very different from those of oxide glasses ($E_g$ = 2.5–5 eV). This feature allows several high-technology applications, especially in far-infrared transmission, which are not possible with other glass types. The unique functionalities of chalcogenide glasses make them the selected materials for microelectronic and photonic devices. They can be made as thin and thick films, molded into lenses, or drawn into fibers. They have been used in commercial applications, such as infrared cameras, fibers, laser waveguides for optical switching, and chemical and temperature sensors [1].

Chalcogenide compounds, such as AgInSbTe and GeSbTe, are also applied in re-writable optical disks and phase-change memory devices. They are fragile glass formers according to Angell's classification [2]; by controlling heating and cooling, they can rapidly switch between non-crystalline and crystalline states, thereby significantly changing their optical and electrical properties and allowing information storage [1].

Chalcogenide glasses are traditionally composed of at least one chalcogen (Se, Te, and S) combined with Ge, As, Sb, Si, P, B, Pb, La, Al, or other neighboring atoms on the periodic table. A key characteristic of chalcogens provides chalcogenide glasses with unique properties: they generate low-energy phonons within the non-crystalline network and confer wide optical transparency, extending far into the infrared. This property is a defining characteristic and has been the source of much research on infrared optics applications.

* Corresponding author at: Ilum School of Science, Brazilian Center for Research in Energy and Materials (CNPEM), Rua Lauro Vanucci 1020, CEP 13087-548, Campinas, Brazil.
*E-mail address:* daniel.cassar@ilum.cnpem.br (D.R. Cassar).

https://doi.org/10.1016/j.actamat.2022.118302
1359-6454/

Chalcogenide glasses are glassy semiconductors. There is relatively firm knowledge about their short-range structure, which covers the coordination number, the bond length, and the bond angle. Also, knowledge of structural dependence on atomic composition, which is practically possible in covalent glasses, has added valuable insights into the chalcogenide glass science [1].

The classical chalcogenide glasses (mainly sulfur-based, such as As-S or Ge-S) are reasonable glass-formers; however, their glass-forming abilities significantly decrease with increasing the molar weight of their constituent elements, i.e., S > Se > Te. Most of the formulations available are far worse glass formers than the typical oxide compositions, and this is a critical issue in this glass family [1]. More recently, the glass research community started digging deeper into the crystallization behavior [3] and development of chalcogenide glass-ceramics, while keeping their optoelectronic properties and showing improved mechanical behavior [4].

A Scopus search made in May 30, 2022 with the keywords chalco* and glass* showed that from 4100 (article title) to 12,100 (title, abstract, or keyword) articles addressing chalcogenide glasses have been published since the pioneering article by Kolomiets and Pishlo in 1963 [5]; the current rate is approximately one article per day. Due to incomplete structural knowledge, especially about medium-range structures, density fluctuations, and defects, the chalcogenide glass science is far behind those constructed for single-crystalline semiconductors or oxide glasses. Also, despite the substantial research conducted in the past 50 years, the understanding of composition-property relationships for chalcogenide glasses is still behind the accumulated knowledge about oxide glasses, which have been systematically studied by many researchers for approximately two centuries. Therefore, to extend the range of available compositions and applications of chalcogenide glasses, there is a pressing need for more profound knowledge about the *composition-structure-property* relationships.

While the number of machine learning (ML) papers addressing oxide glasses has upsurged in the past five years [6–19], to the best of our knowledge, there are only three publications on ML research in chalcogenide glasses [20,21,38].

The first study [20] reports on a multivariate linear regression (MLR) capable of predicting the glass transition temperature of the $As_xSe_{1-x}$ binary system. The obtained MLR model ($T_g = 2464 + 597.3\langle r\rangle - 6755.3\nu - 301.61$ K $+ 4.9257\ U_{0ex} + 0.50313\ KU_{0ex}$) agreed with experimental values for this particular binary system and was based on physical and chemical properties, such as the average coordination number $\langle r\rangle$, the Poisson ratio $\nu$, the bulk modulus $K$, and the mean experimental atomic bonding energy $U_{0ex}$.

The second study [21] reports on recognizing mid-gap states in chalcogenide glasses. To avoid a formidable computational task, Xu et al. [21] adopted a machine learning procedure to understand and predict mid-gap states (MGS) in typical Ovonic Threshold Switching (OTS) materials; selectors are used to suppress current leakage in high-density memory chips. They built hundreds of chalcogenide glass models and collected major structural features from both short-range order (SRO) and medium-range order (MRO) of the amorphous cells. After training an artificial neural network using these features, the induced model recognized MGS in new glasses with 95% accuracy. By analyzing the synaptic weights of the input structural features, they discovered that the bonding and coordination environments from SRO, and particularly MRO, are closely related to MGS. The resulting model could be used in several other OTS chalcogenides after minor modification. The authors concluded that the machine learning technique allowed them to understand the OTS mechanism from a vast amount of structural data without heavy computational tasks, providing a new strategy to design functional amorphous materials from first principles.

Finally, the third study [38] reports on a Hopfield neural network capable of predicting the partial radial distribution function of a two-component glassy chalcogenide, namely $GeSe_3$.

The incentive of researching ML algorithms applied to chalcogenides was pointed out as an opportunity in the field by Tandia et al. [7]. Meeting this incentive is the main objective of this work. Here we use a different approach from that of Refs. [20,21,38]. We aim to *induce* ML models referring to *composition-property* relationships and interpret them to find the effect of each element on the glass properties. Also, we will deal with much more complex compositions, containing up to six elements rather than a single binary system. To this end, we collected published data regarding properties of chalcogenide glasses: glass transition temperature ($T_g$), coefficient of thermal expansion (CTE), and refractive index ($n_D$), and used ML-based approaches to *generate predictive models* for these properties.

The following ML algorithms were investigated in this study: Classification and Regression Trees (CART) [34], *k*-Nearest Neighbors (*k*-NN) [35], Multilayer Perceptron (MLP, a type of Neural Network) [36], and Random Forest (RF) [37], which were chosen because our previous work on oxide glasses indicated that these are the top performers among six ML algorithms [8].

Our first objective is to obtain the best performer model produced using the investigated ML algorithms. Then, we generated the explanation and so the interpretation of the best obtained model by computing the SHapley Additive exPlanations (SHAP) values of the features [22–24]. The produced explanations allow us to obtain the chemical elements' role in each investigated property. We also computed the SHAP interaction values, which allow the investigation of possible non-trivial correlations between the chemical elements that affect the glass properties. We expect that the results of this work will help to understand the role played by the chemical elements in the chalcogenide glasses and aid the design of new glasses with desired combinations of properties.

## 2. Methodology

### 2.1. Data collection

The data on chalcogenide glasses used in this work were collected from the SciGlass database (https://github.com/epam/SciGlass). In this work, we considered glasses part of the chalcogenide family if they have in their composition a non-zero amount of sulfur, selenium, or tellurium; and have no oxygen, fluorine, chlorine, bromine, or iodine. After this filtering procedure, we only collected entries with one or more investigated properties, i.e., glass transition temperature, coefficient of thermal expansion, or refractive index. Additionally, we also investigated the Young's Modulus of chalcogenide glasses; however, the performance of the predictive models was subpar; these results are reported in the supplementary material for completeness. We also considered studying the Abbe number of chalcogenide glasses; however, only approximately fifty examples were available, which would not be enough to use ML algorithms properly.

After the data collection step, we performed a data cleaning step to remove extremely low or high property values that likely refer to typos or gross measurement errors. The strategy used was similar to that employed in our previous publication [16]: we removed the extreme values for each property and the duplicate entries by taking the median value of the property. We defined all values below the 0.05% percentile or above the 99.95% percentile as extreme. Descriptive statistics on the collected dataset are shown in Table 1.

**Table 1**
Descriptive statistics of the used datasets for each property.

| | $T_g$ (K) | $\log_{10}$(CTE) | $n_D$ |
|---|---|---|---|
| **Count** | 7620 | 942 | 456 |
| **Mean** | 476.4 | −4.69 | 2.61 |
| **Std Dev** | 107.9 | 0.20 | 0.41 |
| **Min** | 266.2 | −5.17 | 1.97 |
| **50%** | 453.2 | −4.73 | 2.50 |
| **Max** | 877.2 | −4.02 | 4.34 |
| **Skewness** | 0.78 | 0.62 | 1.04 |
| **Kurtosis** | 0.35 | 0.01 | 0.91 |

**Table 2**
Values of the performance metrics for the three properties obtained using the tuned RF algorithm. The up arrow indicates that the higher the metric, the better; the down arrow indicates the opposite.

| Metric | $T_g$ (K) | $\log_{10}$(CTE) | $n_D$ |
|---|---|---|---|
| RD (↓) | 3.4 ± 0.1 | 1.2 ± 0.2 | 3.4 ± 0.7 |
| R2 (↑) | 0.93 ± 0.01 | 0.76 ± 0.09 | 0.87 ± 0.06 |
| RMSE (↓) | 28 ± 2 | 0.10 ± 0.02 | 0.15 ± 0.05 |
| RRMSE (↓) | 0.26 ± 0.02 | 0.49 ± 0.08 | 0.37 ± 0.09 |

### 2.2. Machine learning experiments

We followed the same ML-base strategy of our recent work on oxide glasses [8]. We considered three ML algorithms that performed well in a previous analysis, namely, CART, *k*-NN, MLP, and RF. Detailed explanations of how they induce predictive models are available in the supplementary material of Ref. [8].

The predictive models were induced using the scikit-learn Python package [25]; a hyperparameter tuning routine was also employed. We adopted a nested cross-validation routine considering an outer-fold of 10 for testing and an inner-fold of 5 for validation. The tuning strategy was the use of random search, testing 500 sets of hyperparameters for each outer fold. Moreover, we used the same search space adopted in Ref. [16]. For experimental reproducibility, we make available the code used on GitHub (https://github.com/ealcobaca/mlglass).

We used nested cross-validation to avoid overfitting and to select the models [26]. This approach can be considered overzealous for most practical applications [27]. Thus, we split the original dataset into training and test sets using 10-outer-fold, creating ten disjoint sets [26,27]. For the internal validation of the hyperparameter optimization procedure, we used, for each nine training folds, 5-inner-fold, creating five disjoint validation sets for the training data [26,27]. Table 2 (Results section) shows the average results from the 10-outer-fold test sets, which were *not* used in the training and validation subsets.

### 2.3. Interpreting the induced models through SHAP analysis

Models induced by ML algorithms can hold a significant amount of information, depending on the used algorithm, that may not be easily interpreted by humans. A new and powerful data analysis tool called SHAP [22–24], distributed as a Python module (https://github.com/slundberg/shap), is a model-agnostic approach to interpreting any predictive function and extracting/visualizing meaningful information in a human-readable fashion. The approach used by SHAP is the computation of the Shapley values [28], which are based on game theory and inform how much a given prediction is affected by the input features concerning a given base value. Detailed information on this procedure is reported by the creators of this method [22].

One viable way to visualize the results of the SHAP analysis is via beeswarm plots. These plots can be thought of as horizontal violin plots, with features sorted by decreasing order of importance. In this case, the feature importance is measured by the absolute value of the SHAP values, which indicates the features that have a higher impact on the predicted value of the model. The SHAP values have the same units of the property being predicted and convey how much a given feature (amount of chemical elements in the glass, in this case) impacts the property in relation to a base value, which is taken as the mean value of the property (see Table 1).

## 3. Results and discussion

### 3.1. Analysis of the datasets used in this study

Table 1 shows the descriptive statistics of the glass compositions collected from the SciGlass database. The smallest dataset was labeled with the refractive index with 456 unique compositions, whereas the largest dataset was labeled with the glass transition temperature with 7620 unique compositions. While these numbers are much smaller than those used in the previous ML works on oxide glasses [9–13,16,17], they are still significant and can be used by ML algorithms to extract composition-property relations. It is relevant to note that current chalcogenide formulations comprise 58 elements, with only 1 to 6 different elements in each glass.

Fig. 1 shows the histogram of the number of chemical elements in the glasses for each property, which varies from 1 to 6. These relatively “simple” compositions contrast with those of the widely studied oxide glasses, for which multi-component glasses with more than 20 elements are reported. Hopefully, this work could guide researchers in formulating novel multi-component chalcogenide glasses, as discussed further in this communication. Similarly, Fig. 2 shows the histogram for the property values, for which the minimum and maximum values can be found in Table 1. All studied properties have an asymmetric distribution, which is evidenced by the non-zero value for their skewness (Table 1).

### 3.2. Predictive performance measures

Table 2 shows the predictive performance metrics for the three properties obtained by the RF algorithm. Additional tables, with CART, MLP, and *k*-NN results, can be found in the Appendix. In general, the predictive performance values obtained by RF and *k*-NN outperformed those obtained by CART and MLP. However, if we compared the produced models for the chalcogenide glasses with those obtained for oxide glasses [16], we observed a decrease in the predictive performance. This decrease in performance is related to the lower number of training instances available for the chalcogenides compared with the instances available for oxide glasses. The number of examples (composition-property points) used in the training procedures was much larger for the oxide glasses, about 20,000 to 50,000, while for the chalcogenides, the available number was in the range of 450–7500. As expected, the uncertainty decreased with the number of examples used in the training procedure; for instance, $R^2$ is 0.87 for $n_D$ (456 examples) versus 0.93 for $T_g$ (7620 examples).

Fig. 3 shows the main results of the relative deviation of the $T_g$ prediction for the four ML algorithms used in the experiments. Again, as reported in previous communications [6,8,17], the uncertainty in the extremes of low and high $T_g$ is higher than in the intermediate range. This behavior is similar to those from other

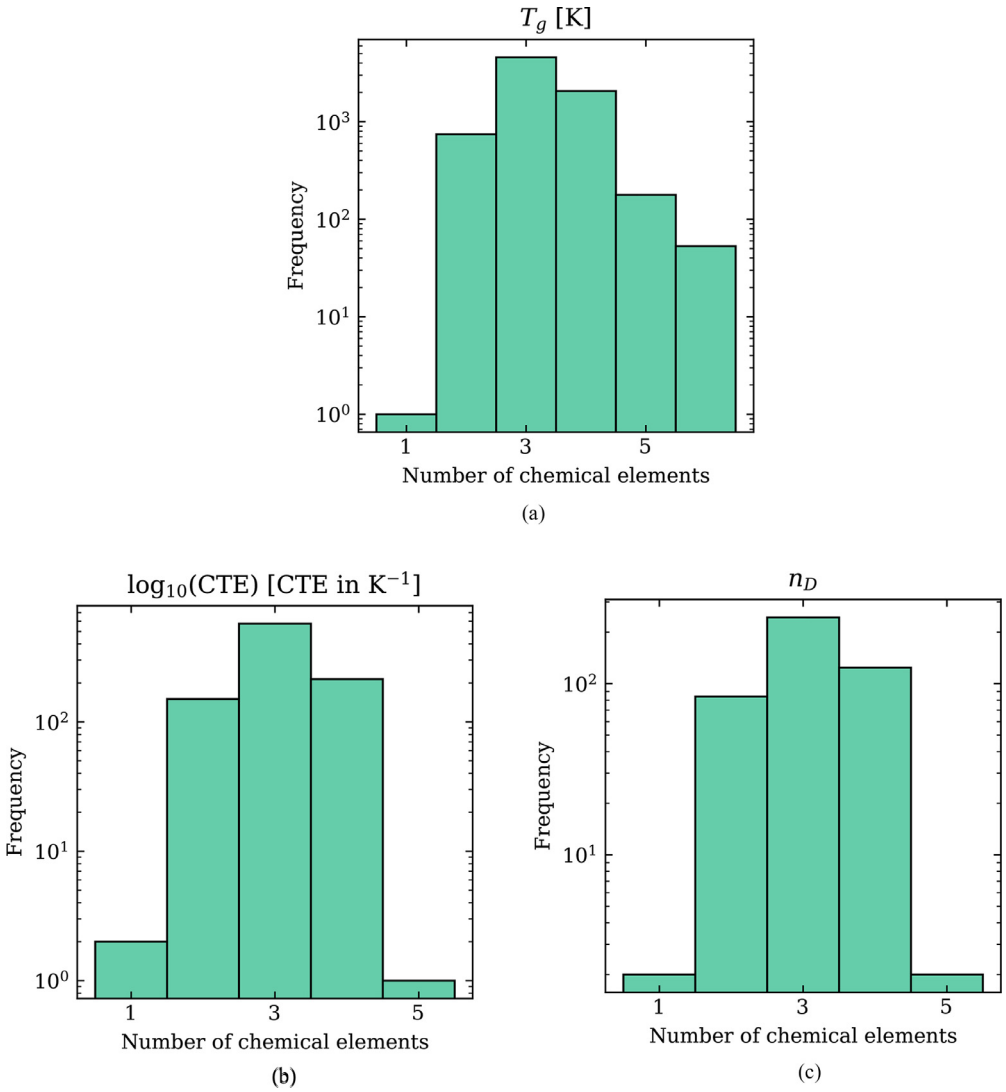



**Fig. 1.** Frequency versus the number of chemical elements in each composition for three properties of chalcogenide glasses.

studied properties and reflects the small number of examples in the extreme regions. The plots for the other properties are reported in the Appendix.

Fig. 4 shows the mean and standard deviation of the residual prediction (reported minus predicted values) of the $T_g$, $\log_{10}$(CTE), and $n_D$ models induced by RF for each chemical element in the glass, that is, for glasses having these elements in their composition. The upper region of these plots shows how many examples were available with each of the chemical elements considered. Again, this result is similar to those previously reported for oxide glasses [16]. Elements that are included in a large number of glass compositions tend to have a mean residual prediction close to zero, whereas most of the others show much larger errors.

The induced predictive models can be used for the computer-aided design of new chalcogenide glasses for the desired combinations of properties. However, due to the limited dataset used for training these models, unsatisfactory predictions will likely result from searching for chemical compositions that contain certain elements that are present in a small number of compositions, such as Co, U, Mg, Sm, Tm, Y, Ce, Cs, H and a few others shown in Fig. 4. The same restriction applies to new formulations that are far away from those present in the training dataset. To mitigate this problem, we would have to significantly extend the dataset.

In the following section, we further explored the RF-induced models in an attempt to extract useful information regarding the effect of each chemical element on the investigated properties. To this end, we will use the SHAP analysis discussed in the methodology section.

### 3.3. Interpreting the induced models

By computing the SHAP values, we obtained the plots shown in Fig. 5 for the three studied properties. Although the SHAP method still presents some problems [29,30], these beeswarm plots provide valuable insights for designing new chalcogenide glasses. Each dot in these plots represents a glass having the chemical element shown in the respective left label (note that the dots can stack vertically, conveying the message that many glasses have the same SHAP value). The *x*-axis shows the SHAP value, which has the same units as the target property and quantifies the impact of the feature (chemical element) on the property. Finally, each dot has a color representing the atomic fraction of the element in the glass (increasing from purple to yellow).

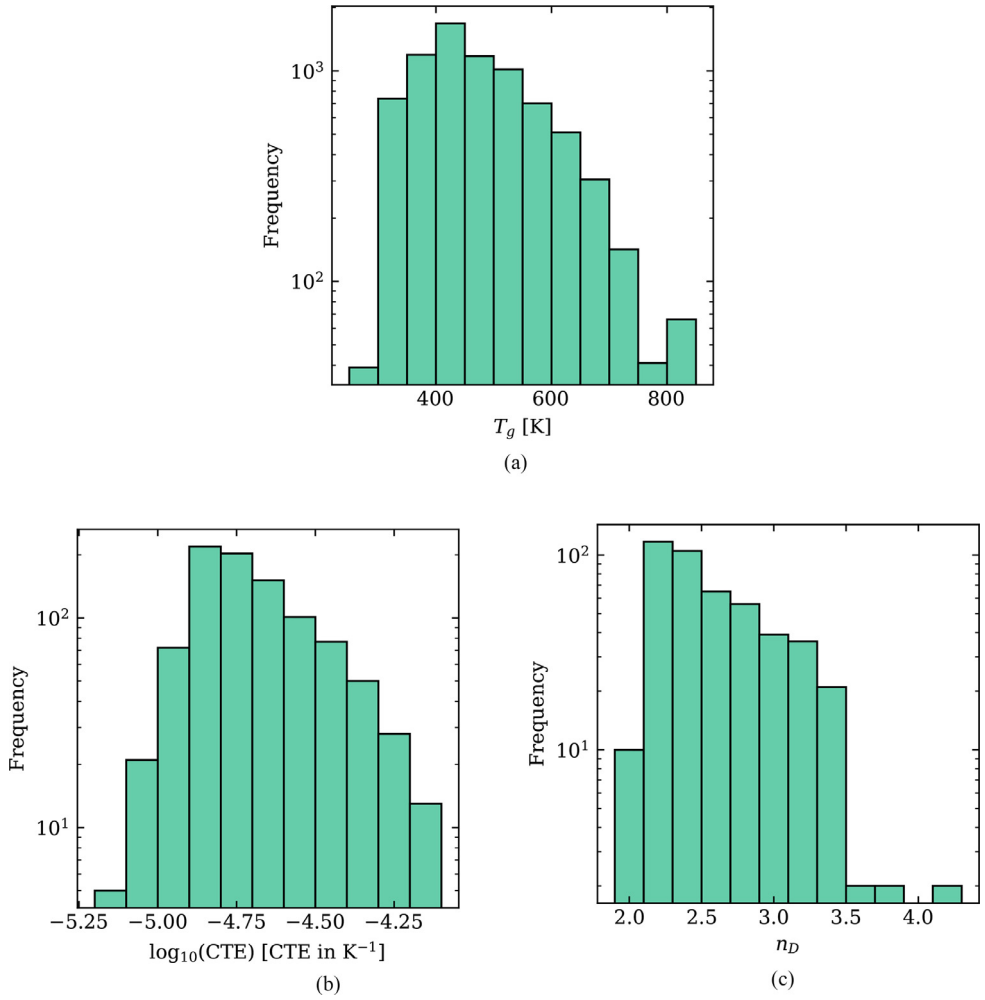



**Fig. 2.** Frequency versus value for three properties of chalcogenide glasses.

Fig. 5a shows that large amounts of silicon, gallium, germanium, lanthanum, and barium contribute to increasing the $T_g$ of chalcogenide glasses, whereas thallium, selenium, and tellurium contribute to decreasing it. Sulfur, bismuth, indium, copper, and arsenic have a mixed effect; they can either increase or decrease this property, suggesting that these elements interact within the glass network in a complex way.

Fig. 5b shows that sulfur, selenium, and thallium increase CTE, while sodium, germanium, silicon, gallium, and antimony contribute to decreasing it. A mixed effect is observed for arsenic, tellurium, and phosphorus for this property.

The properties discussed ($T_g$ and CTE) are related to the chemical bond energy. The analysis of the refractive index (a property that is not directly related to the chemical bond energies, but the polarizability of the elements) is shown in Fig. 5c. Here, tellurion, arsenic, thallium, antimony, lead, silicon, and indium are elements that may increase this property, while sulfur, phosphorus, and germanium may decrease it. No clear mixed effect was observed by only looking at Fig. 5c.

Now, we will look in more detail at the magnitude of the SHAP values, which, as already mentioned, quantifies the impact of the elements on the final prediction of the model. Starting with Fig. 5a, we see that silicon (6%), gadolinium (42%), and germanium (88%) can rise $T_g$ the most, up to about 170 K. The numbers in parentheses refer to the percentage of high $T_g$ glasses (above the 80% percentile) in the dataset containing these chemicals. As it can be seen, by simply looking at the number of reported chalcogenide glasses having high $T_g$, one could miss the significant impact of silicon on this property.

Similarly, thallium (22.3%), selenium (80.7%), and tellurium (42.6%) can decrease $T_g$ the most, down by about 100 K in the extreme case. The numbers in parentheses refer to the percentage of low $T_g$ glasses (below the 20% percentile) in the dataset containing these chemicals. As previously mentioned, these analyses provide us with rich information to empirically design new chalcogenide glasses. The following paragraphs explore the other two properties.

Fig. 5b shows that sulfur, zinc, sodium, germanium, silicon, and selenium are the elements with the most significant impact on increasing CTE, which can amount to 0.25 in the base-10 logarithm scale for the most extreme case. Interestingly, germanium only increases CTE when present in small quantities, but even so, it has a significant impact on this property. Germanium, silicon, and gallium play the most significant role in decreasing CTE, reaching up to 0.2 in the base-10 logarithm scale.

Finally, Fig. 5c shows that tellurium, arsenic, and thallium can significantly increase $n_D$, the first reaching an impressive impact of 0.4 on this property. Meanwhile, sulfur and phosphorus can decrease this property by more than 0.2.

### 3.4. SHAP interaction analysis

Another way to interpret the induced models is by computing the SHAP interaction values. According to Lundberg et al., “SHAP

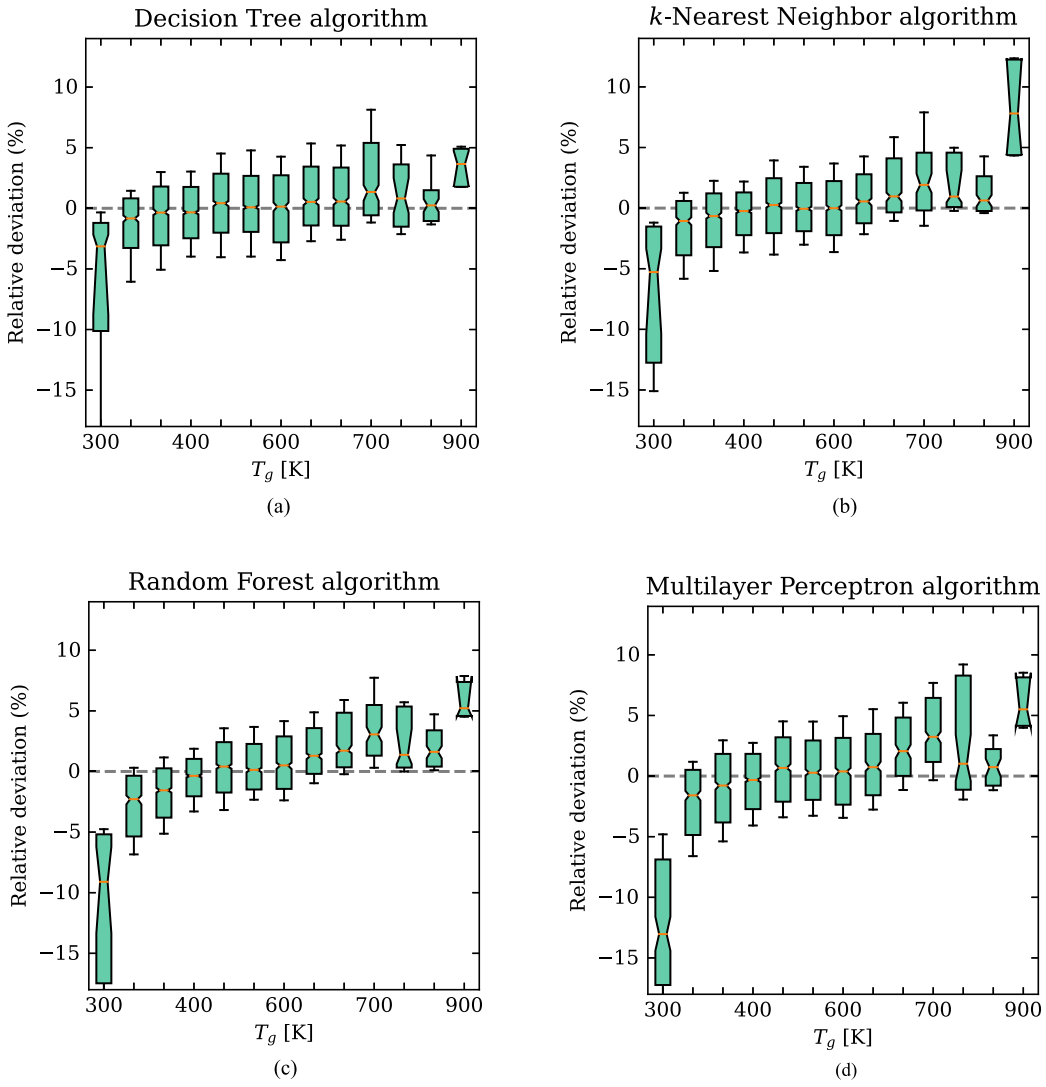



**Fig. 3.** Boxplot of residuals for the prediction of $T_g$ for the tuned models. The boxes are bounded by the first and third quartiles, while the error bars comprehend 66% of the data. The mean is shown by a horizontal orange line and the notch represents its confidence interval.

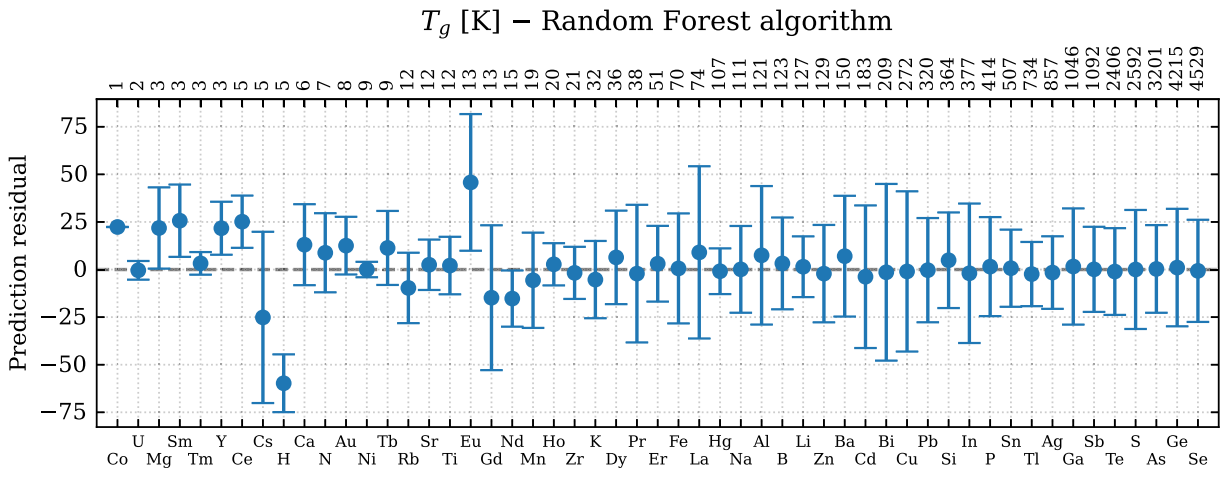
interaction values can be interpreted as the difference between the SHAP values for feature $i$ when feature $j$ is present and the SHAP values for feature $i$ when feature $j$ is absent" [24]. Thus, if the chemical elements A and B have a null SHAP interaction value for a given property, then the contribution of element A to the property is independent of the presence of element B and vice-versa.

What is particularly interesting for glass science and glass composition design is when the SHAP interaction value is significantly higher than zero. A famous case of interaction between chemical elements in oxide glasses (as recently pointed out by Ravinder et al. [12]) is the well-known boron anomaly [31–33]. This anomaly is explained by the change in the number of the boron network bridging oxygens with the increase of network modifier elements, such as the alkali and alkali-earth. Thus, any property that depends on the connectivity of the glass network will be affected by an interaction of boron with the modifiers, thus yielding a higher SHAP interaction value.

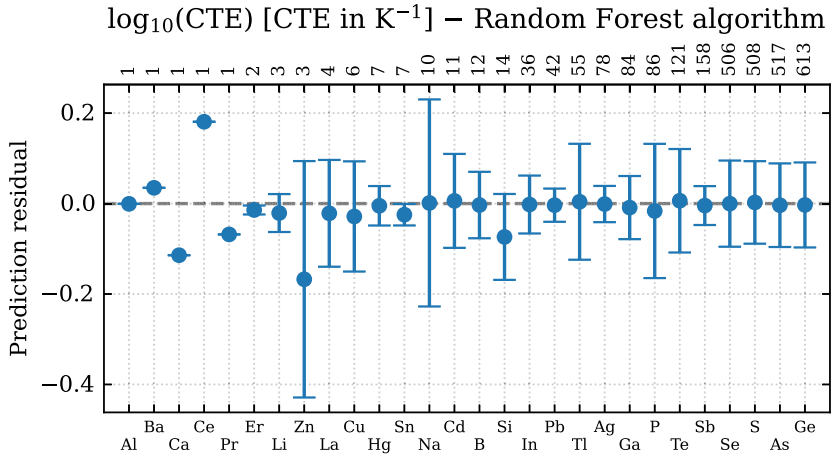
Fig. 6 shows the SHAP interaction values for the three studied properties for the most relevant pairs of chemical elements. Fig. 6a shows that the higher SHAP interaction values for $T_g$ occur for pairs containing germanium, sulfur, arsenic, selenium, tellurium, and gallium. The pairs arsenic-germanium and sulfur-selenium are particularly interesting, as they have the *highest* SHAP interaction values. All these elements can form glass networks; hence this figure indicates the mixed-former effect on the glass transition temperature of chalcogenide glasses.

Similarly, Fig. 6b and c show that the higher SHAP interaction values for $\log_{10}$(CTE) and $n_D$ occur for pairs containing a subset of the elements mentioned above for $T_g$. For $\log_{10}$(CTE) the elements with higher SHAP interaction values are germanium, arsenic, sulfur, and selenium, and for $n_D$, the key elements are sulfur, selenium, arsenic, tellurium, and germanium. Interestingly, the arsenic-germanium pair shows the highest SHAP interaction values for $\log_{10}$(CTE), whereas the sulfur-selenium pair shows the highest value for $n_D$. These two pairs also show the highest SHAP interaction values in the $T_g$ analysis.

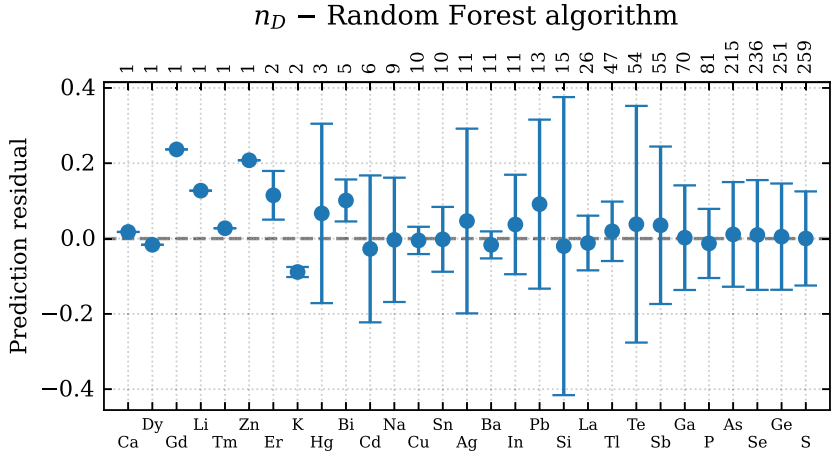
The above discussions show that the SHAP analysis not only reveals the individual effect of the chemical elements on the glass properties and their respective magnitudes but also gives clues on non-trivial, useful *interactions* between element pairs.




(a)

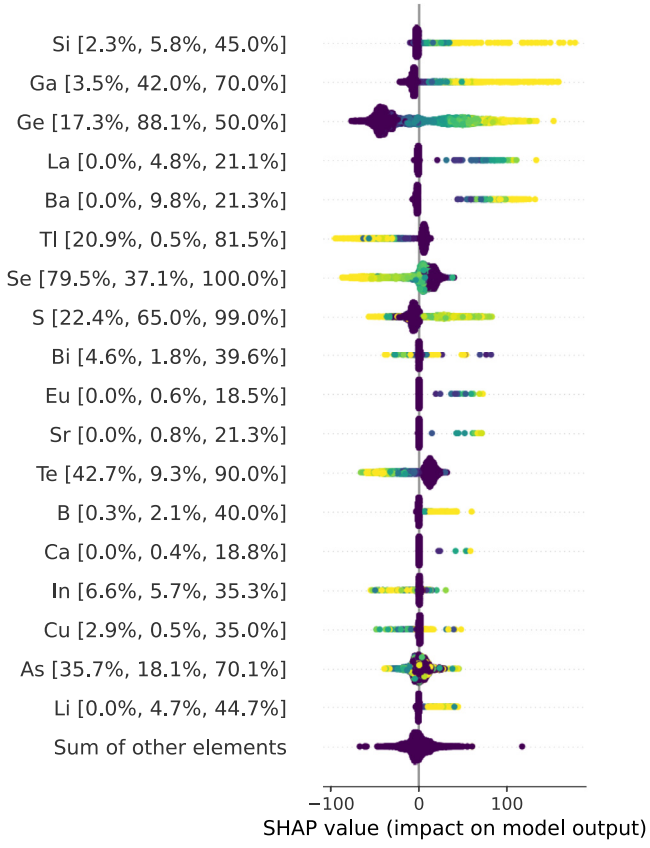



(b)

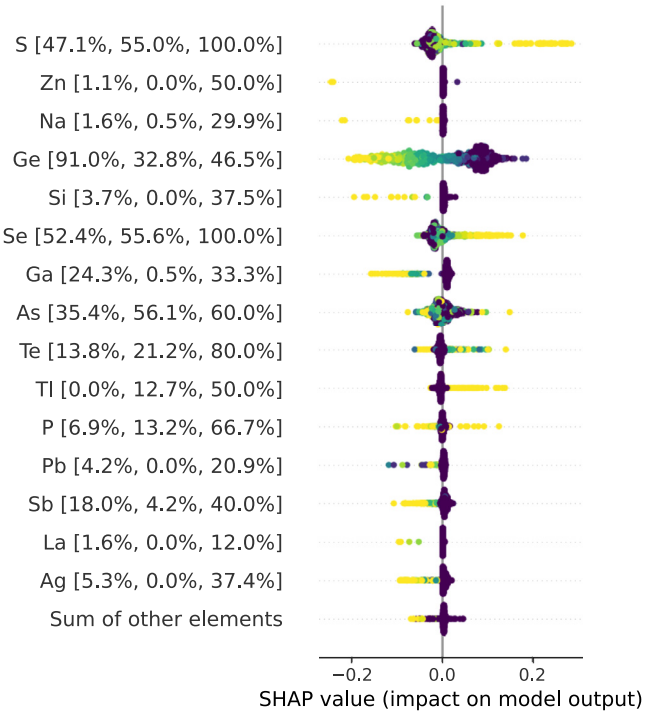



(c)

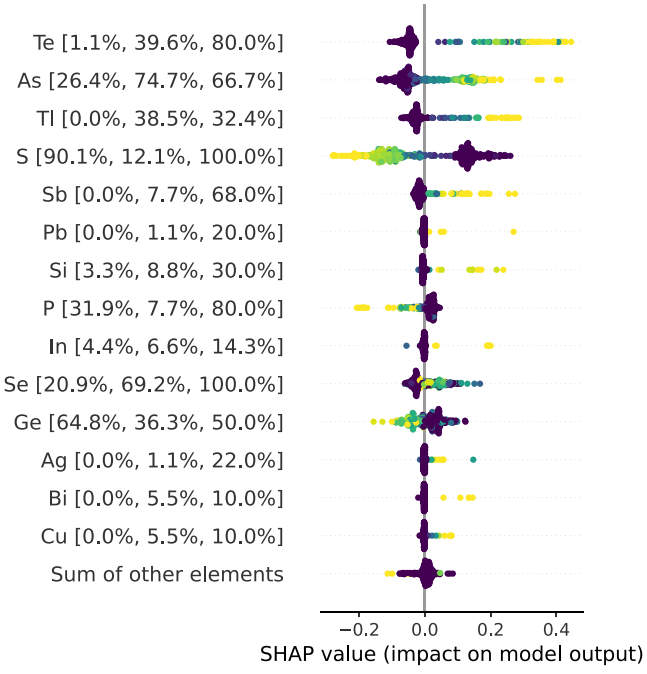
**Fig. 4.** Mean and standard deviation of the RF prediction residual of (**a**) $T_g$, (**b**) $\log_{10}$(CTE), and (**c**) $n_D$ for glasses having the chemical element shown on the *x-axis*. The figures on the top are the number of examples (glass compositions) having the corresponding element in the dataset.




(a)

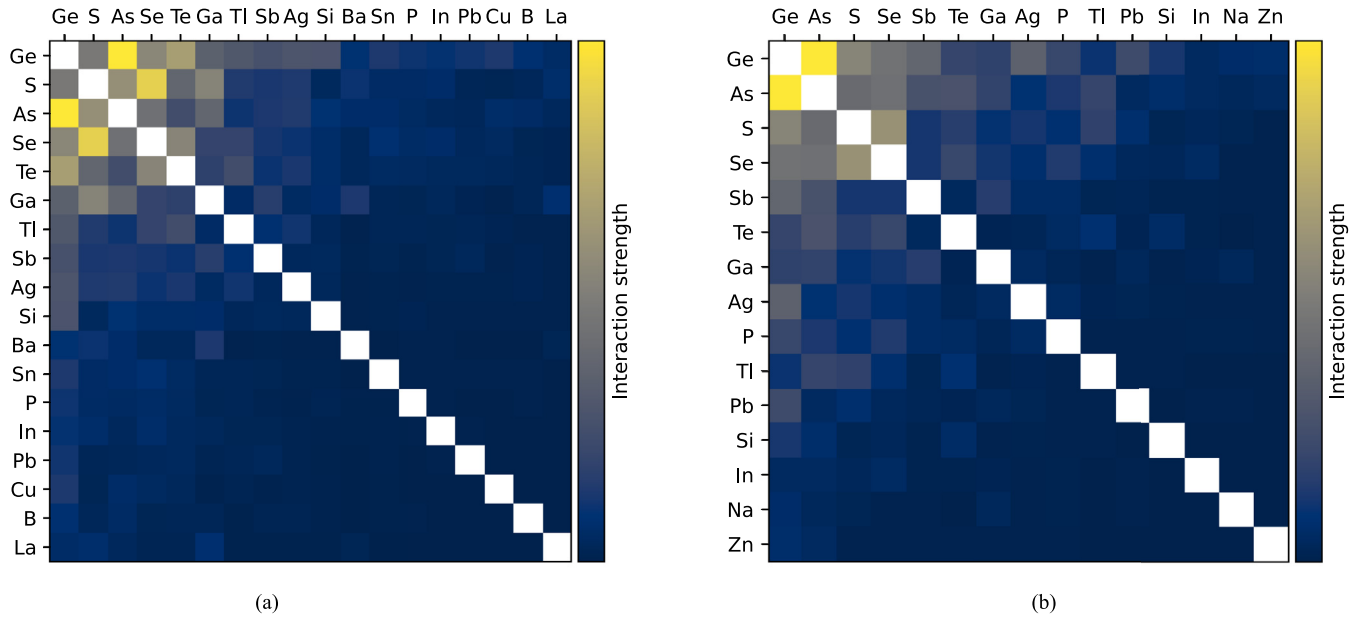



(b)




(c)

**Fig. 5.** Beeswarm plot of the SHAP values obtained from the RF predictive model of (**a**) $T_g$, (**b**) $\log_{10}$(CTE), and (**c**) $n_D$. The numbers within brackets beside the chemical element labels represent, respectively, the percentage of examples that contain the said element in the low range of the property (lower than the 20% percentile), the percentage of examples that contain the said element in the high range of the property (higher than the 80% percentile), and the maximum atomic fraction of the element in one of the glasses in the dataset. Each dot represents a glass and its color represents the atomic fraction of the element in the glass (increasing from purple to yellow).



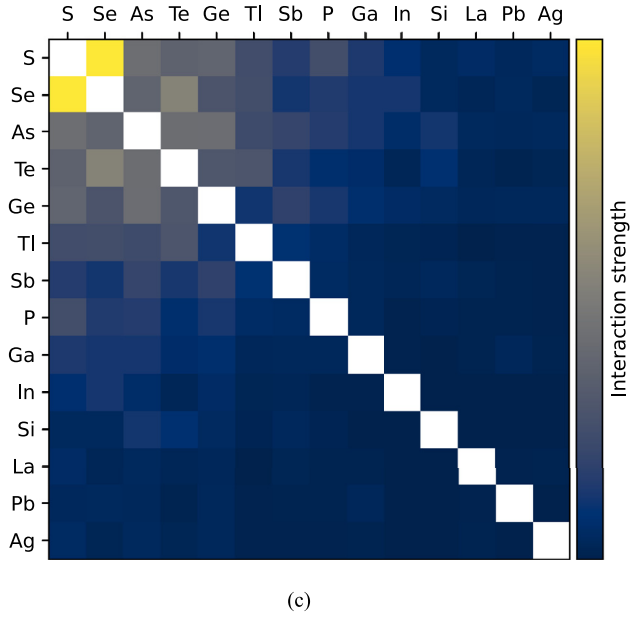




**Fig. 6.** 2D histogram of the SHAP interaction values for (**a**) $T_g$, (**b**) $\log_{10}$(CTE), and (**c**) $n_D$. The diagonal (where the interaction strength is zero) was removed.

## 4. Summary and conclusions

In this study, we collected over nine thousand composition-property sets for three properties of chalcogenide glasses. Current chalcogenide formulations comprise 58 chemical elements, with 1 to 6 elements in each glass. We used these data to train and test four different ML algorithms to compute composition-property relationships for this important glass family, for the first time. The RF and *k*-NN algorithms outperformed the MLP and CART algorithms in predictive performance, confirming previous results for oxide glasses.

A SHAP analysis of the RF models indicated the key elements that significantly increase or decrease the value of the tested properties and their maximum possible variation. For instance: germanium, silicon, and gallium increase $T_g$ and decrease CTE. This occurs likely because these elements rise the interatomic bond strength of these covalent glasses. Selenium has the opposite effect on these properties. Tellurium, arsenic, thallium, and antimony increase $n_D$, which depends mostly on polarizability, whereas sulfur and phosphorus diminish it.

A SHAP interaction analysis revealed some element pairs that potentially exhibit the mixed-former effect: arsenic-germanium and sulfur-selenium.

This knowledge about the effect of each element on properties can be precious for semi-empirical compositional development trials of chalcogenide glasses. Besides, the induced predictive models can be used for the computer-aided design of new chalcogenide glasses having desired combinations of properties. However, due to the limited dataset used for training these models, unsatisfactory predictions will likely result in searching for chemical compositions that are too far away from those present in the training dataset. The same restriction applies to other substances, such

as oxide, metallic, and organic glasses. One solution to mitigate this problem is to significantly extend the available composition-property dataset.

## Declaration of Competing Interest

The authors declare no competing financial or non-financial interests.

## CRediT authorship contribution statement

**Saulo Martiello Mastelini:** Methodology, Software, Investigation, Data curation, Writing – review & editing. **Daniel R. Cassar:** Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Edesio Alcobaça:** Methodology, Software, Data curation, Writing – review & editing, Visualization. **Tiago Botari:** Methodology, Software, Data curation, Writing – review & editing. **André C.P.L.F. de Carvalho:** Resources, Data curation, Writing – review & editing, Supervision, Funding acquisition. **Edgar D. Zanotto:** Conceptualization, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition.

## Acknowledgments

The São Paulo Research Foundation (FAPESP) financed this study through grants 2017/12491-0, 2018/07319-6, 2017/06161-7, 2018/14819-5, 2013/07375-0 (CEPID), and 2013/07793-6 (CEPID).

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.actamat.2022.118302.

## Appendix

Table A.1 shows the hyperparameter tuning space and the best values obtained after tuning.

**Table A.2**
Values of the performance metrics for the four properties obtained using the tuned CART algorithm. The up arrow indicates that the higher the metric, the better; the down arrow indicates the opposite.

| Metric | $T_g$ (K) | $\log_{10}$(CTE) | $n_D$ |
|---|---|---|---|
| RD (↓) | 4.4 ± 0.2 | 1.7 ± 0.2 | 5 ± 1 |
| R2 (↑) | 0.88 ± 0.02 | 0.6 ± 0.1 | 0.8 ± 0.1 |
| RMSE (↓) | 38 ± 3 | 0.12 ± 0.02 | 0.20 ± 0.06 |
| RRMSE (↓) | 0.35 ± 0.02 | 0.6 ± 0.1 | 0.5 ± 0.1 |

**Table A.3**
Values of the performance metrics for the four properties obtained using the tuned *k*-NN algorithm. The up arrow indicates that the higher the metric, the better; the down arrow indicates the opposite.

| Metric | $T_g$ (K) | $\log_{10}$(CTE) | $n_D$ |
|---|---|---|---|
| RD (↓) | 3.7 ± 0.1 | 1.3 ± 0.2 | 3.3 ± 0.6 |
| R2 (↑) | 0.92 ± 0.01 | 0.76 ± 0.08 | 0.87 ± 0.05 |
| RMSE (↓) | 30 ± 2 | 0.10 ± 0.02 | 0.15 ± 0.05 |
| RRMSE (↓) | 0.28 ± 0.02 | 0.49 ± 0.09 | 0.36 ± 0.09 |

**Table A.4**
Values of the performance metrics for the four properties obtained using the tuned MLP algorithm. The up arrow indicates that the higher the metric, the better; the down arrow indicates the opposite.

| Metric | $T_g$ (K) | $\log_{10}$(CTE) | $n_D$ |
|---|---|---|---|
| RD (↓) | 4.1 ± 0.5 | 1.3 ± 0.1 | 3.5 ± 0.6 |
| R2 (↑) | 0.92 ± 0.02 | 0.76 ± 0.08 | 0.87 ± 0.08 |
| RMSE (↓) | 31 ± 4 | 0.10 ± 0.02 | 0.15 ± 0.06 |
| RRMSE (↓) | 0.29 ± 0.03 | 0.5 ± 0.09 | 0.4 ± 0.1 |

Tables A.2–A.4 show the performance measures for the CART, *k*-NN, and MLP algorithms. Tables A.4–A.7 show the metrics for the induced models for the four properties studied in this work. Finally, Figs. A.1 and A.2 show the boxplots and the residual plots vs. chemical elements for $\log_{10}$(CTE) and $n_D$.

**Table A.1**
Hyperparameter search space and best values. For more information about the hyperparameters, please check the scikit-learn user guide at https://scikit-learn.org/stable/user_guide.html.

| Algorithm | Hyperparameter | Range | $T_g$ | $\log_{10}$(CTE) | $n_D$ |
|---|---|---|---|---|---|
| **CART** | criterion | {mse, friedman_mse} | mse | friedman_mse | friedman_mse |
| | min_impurity_decrease | [0, 0.1] | 0.0952 | 0.0238 | 0.0737 |
| **k-NN** | n_neighbors | [1, 1000] | 6 | 4 | 4 |
| | weights | {uniform, distance} | distance | distance | distance |
| **MLP** | hidden_layer_sizes | [0, 100] | (71, 48, 88) | (36, 64, 94) | (94, 3, 52) |
| | solver | {lbfgs, sgd, adam} | lbfgs | lbfgs | lbfgs |
| | activation | {logistic, tanh, relu} | logistic | relu | relu |
| | alpha | $\{10^{-5}, 10^{-4}, 10^{-3}\}$ | $10^{-3}$ | $10^{-4}$ | $10^{-3}$ |
| | learning_rate | {constant, adaptive} | constant | constant | constant |
| | learning_rate_init | [0.001, 0.1] | 0.489 | 0.0117 | 0.0906 |
| | batch_size | {200, 500, 1000} | 1000 | 500 | 1000 |
| | max_iter | [200, 1000] | 701 | 748 | 338 |
| | momentum | [0, 1] | 0.0667 | 0.2089 | 0.7697 |
| **RF** | n_estimators | [500, 1000] | 417 | 274 | 120 |
| | max_features | {auto, sqrt, log2} | sqrt | log2 | log2 |

**Table A.5**
Experimental results for *Tg*.

| Metric | Cart | | *k*-NN | | MLP | | RF | |
|---|---|---|---|---|---|---|---|---|
| | Default | Tuning | Default | Tuning | Default | Tuning | Default | Tuning |
| RD | 4.4 ± 0.2 | 4.4 ± 0.2 | 3.9 ± 0.1 | 3.7 ± 0.1 | 7.2 ± 0.4 | 4.1 ± 0.5 | 3.5 ± 0.1 | 3.4 ± 0.1 |
| $R^2$ | 0.88 ± 0.01 | 0.88 ± 0.02 | 0.92 ± 0.01 | 0.92 ± 0.01 | 0.80 ± 0.01 | 0.92 ± 0.02 | 0.92 ± 0.01 | 0.93 ± 0.01 |
| RMSE | 38 ± 2 | 38 ± 3 | 31 ± 2 | 30 ± 2 | 49 ± 3 | 31 ± 4 | 30 ± 2 | 28 ± 2 |
| RRMSE | 0.35 ± 0.02 | 0.35 ± 0.02 | 0.29 ± 0.02 | 0.28 ± 0.02 | 0.45 ± 0.01 | 0.29 ± 0.03 | 0.27 ± 0.02 | 0.26 ± 0.02 |

**Table A.6**
Experimental results for $\log_{10}$(CTE).

| Metric | Cart | | *k*-NN | | MLP | | RF | |
|---|---|---|---|---|---|---|---|---|
| | Default | Tuning | Default | Tuning | Default | Tuning | Default | Tuning |
| RD | 1.6 ± 0.3 | 1.7 ± 0.2 | 1.3 ± 0.2 | 1.3 ± 0.2 | 2.4 ± 0.3 | 1.3 ± 0.1 | 1.2 ± 0.2 | 1.2 ± 0.2 |
| $R^2$ | 0.6 ± 0.1 | 0.6 ± 0.1 | 0.76 ± 0.08 | 0.76 ± 0.08 | 0.46 ± 0.08 | 0.76 ± 0.08 | 0.75 ± 0.09 | 0.76 ± 0.09 |
| RMSE | 0.12 ± 0.03 | 0.12 ± 0.02 | 0.10 ± 0.02 | 0.10 ± 0.02 | 0.15 ± 0.02 | 0.10 ± 0.02 | 0.10 ± 0.02 | 0.10 ± 0.02 |
| RRMSE | 0.6 ± 0.1 | 0.6 ± 0.1 | 0.49 ± 0.08 | 0.49 ± 0.09 | 0.78 ± 0.06 | 0.50 ± 0.09 | 0.50 ± 0.09 | 0.49 ± 0.08 |

**Table A.7**
Experimental results for $n_D$.

| Metric | Cart | | *k*-NN | | MLP | | RF | |
|---|---|---|---|---|---|---|---|---|
| | Default | Tuning | Default | Tuning | Default | Tuning | Default | Tuning |
| RD | 5 ± 1 | 5 ± 1 | 3.6 ± 0.8 | 3.3 ± 0.6 | 4.9 ± 0.8 | 3.5 ± 0.6 | 3.4 ± 0.6 | 3.4 ± 0.7 |
| $R^2$ | 0.7 ± 0.1 | 0.8 ± 0.1 | 0.86 ± 0.07 | 0.87 ± 0.05 | 0.79 ± 0.08 | 0.87 ± 0.08 | 0.86 ± 0.05 | 0.87 ± 0.06 |
| RMSE | 0.22 ± 0.07 | 0.20 ± 0.06 | 0.15 ± 0.06 | 0.15 ± 0.05 | 0.19 ± 0.06 | 0.15 ± 0.06 | 0.16 ± 0.05 | 0.15 ± 0.05 |
| RRMSE | 0.6 ± 0.2 | 0.5 ± 0.2 | 0.4 ± 0.1 | 0.36 ± 0.09 | 0.47 ± 0.09 | 0.4 ± 0.1 | 0.39 ± 0.09 | 0.37 ± 0.09 |

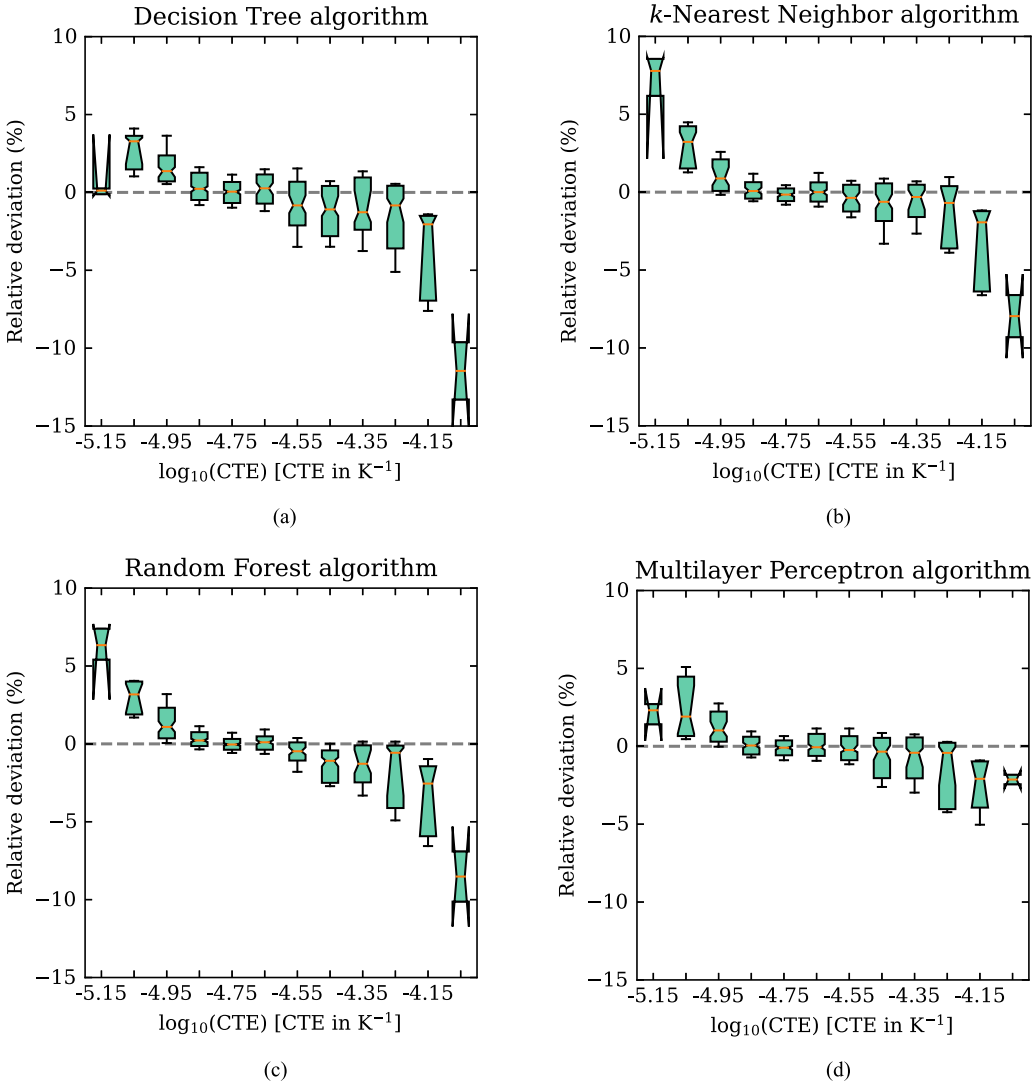



**Fig. A.1.** Boxplot of residuals for the prediction of $\log_{10}$(CTE) for the tuned models. The boxes are bounded by the first and third quartiles, while the error bars comprehend 66% of the data. The mean is shown by a horizontal orange line and the notch represents its confidence interval.

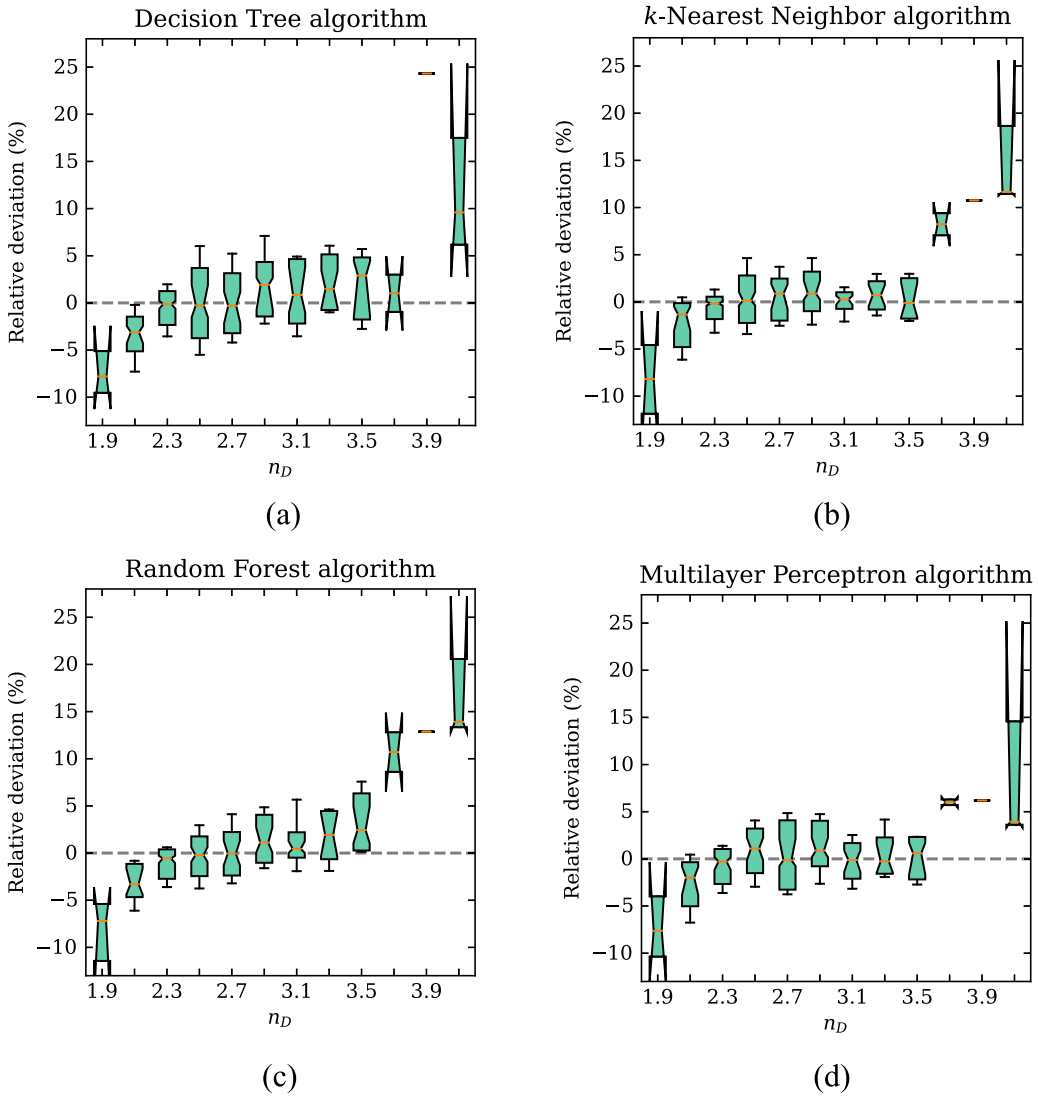



**Fig. A.2.** Boxplot of residuals for the prediction of $n_D$ for the tuned models. The boxes are bounded by the first and third quartiles, while the error bars comprehend 66% of the data. The mean is shown by a horizontal orange line and the notch represents its confidence interval.

## References


[1] J.L. Adam, X. Zhang, Chalcogenide Glasses: Preparation, Properties and Applications, WP, Woodhead Publishing, Oxford, 2014.

[2] C.A. Angell, K.L. Ngai, G.B. Wright, Strong and fragile liquids, in: Relaxation in Complex Systems, Naval Research Laboratory, Springfield, 1985, pp. 3–12.

[3] J. Orava, A.L. Greer, Classical-nucleation-theory analysis of priming in chalcogenide phase-change memory, Acta Mater. 139 (2017) 226–235, doi:10.1016/j.actamat.2017.08.013.

[4] C. Lin, C. Rüssel, S. Dai, Chalcogenide glass-ceramics: functional design and crystallization mechanism, Prog. Mater. Sci. 93 (2018) 1–44, doi:10.1016/j.pmatsci.2017.11.001.

[5] B.T. Kolomiets, V.P. Pishlo, Softening temperatures of some chalcogenide glasses, Glass Ceram. 20 (1963) 413–415.

[6] 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, doi:10.1016/j.actamat.2018.08.022.

[7] A. Tandia, M.C. Onbasli, J.C. Mauro, J.D. Musgraves, J. Hu, L. Calvez, Machine learning for glass modeling, in: Springer Handbook of Glass, Springer International Publishing, Cham, 2019, pp. 1157–1192.

[8] 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, doi:10.1016/j.actamat.2020.01.047.

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

[10] 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, doi:10.1039/D0MH00162G.

[11] M. Zaki, V. Venugopal, R. Bhattoo, S. Bishnoi, S.K. Singh, A.R. Allu, Jayadeva, N.M.A. Krishnan, Interpreting the optical properties of oxide glasses with machine learning and Shapely additive explanations, Journal of the American Ceramic Society 105 (2022) 4046–4057, doi:10.1111/jace.18345.

[12] R. Ravinder, S. Bishnoi, M. Zaki, N.M.A. Krishnan, Understanding the compositional control on electrical, mechanical, optical, and physical properties of inorganic glasses with interpretable machine learning, Social Science Research Network, Rochester, NY, 2022. 10.2139/ssrn.4075602 is preprint from Acta Materialia first look (the paper is still in the review stage) available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4075602.

[13] 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, doi:10.1039/D0MA00764A.

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

[15] 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, doi:10.1016/j.ceramint.2021.05.105.

[16] 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, doi:10.1016/j.ceramint.2020.12.167.

[17] C.J. Wilkinson, C. Trivelpiece, R. Hust, R.S. Welch, S.A. Feller, J.C. Mauro, Hybrid machine learning/physics-based approach for predicting oxide glass-forming ability, Acta Mater. (2021) 117432, doi:10.1016/j.actamat.2021.117432.
[18] K. Nakamura, N. Otani, T. Koike, Multi-objective Bayesian optimization of optical glass compositions, Ceram. Int. (2021), doi:10.1016/j.ceramint.2021.02.155.
[19] Y. Zhang, X. Xu, Predicting $As_x$ $Se_{1-x}$ glass transition onset temperature, Int. J. Thermophys. 41 (2020) 149, doi:10.1007/s10765-020-02734-4.
[20] M. Xu, M. Xu, X. Miao, Deep machine learning unravels the structural origin of mid-gap states in chalcogenide glass for high-density memory integration, InfoMat 4 (2022) e12315, doi:10.1002/inf2.12315.
[21] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017) 4765–4774.
[22] S.M. Lundberg, B. Nair, M.S. Vavilala, M. Horibe, M.J. Eisses, T. Adams, D.E. Liston, D.K.W. Low, S.F. Newman, J. Kim, S.I. Lee, Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nat. Biomed. Eng. 2 (2018) 749–760, doi:10.1038/s41551-018-0304-0.
[23] 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, doi:10.1038/s42256-019-0138-9.
[24] 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.
[25] D. Krstajic, L.J. Buturovic, D.E. Leahy, S. Thomas, Cross-validation pitfalls when selecting and assessing regression and classification models, J. Cheminform. 6 (2014) 10, doi:10.1186/1758-2946-6-10.
[26] J. Wainer, G. Cawley, Nested cross-validation when selecting classifiers is overzealous for most practical applications, Expert Syst. Appl. 182 (2021) 115222, doi:10.1016/j.eswa.2021.115222.
[27] L.S. Shapley, in: H.W. Kuhn, A.W. Tucker (Eds.), A value for n-person games, Contributions to the Theory of Games, 1953, pp. 307–317.
[28] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo, Benchmarking and survey of explanation methods for black box models, ArXiv:2102.13076 [Cs]. (2021). arxiv.org/abs/2102.13076 (accessed May 13, 2021).
[29] D. Slack, S. Hilgard, E Jia, S. Singh, H. Lakkaraju, Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, 2020, pp. 180–186. doi: 10.1145/3375627.3375830.
[30] J. Krogh-Moe, On the structure of boron oxide and alkali borate glasses, Phys. Chem. Glas. 1 (1960) 26.
[31] J. Krogh-Moe, New evidence on the boron coordination in alkali borate glasses, Phys. Chem. Glas. 3 (1962) 1–6.
[32] A.K. Varshneya, J.C. Mauro, Fundamentals of Inorganic Glasses, 3rd ed., Elsevier, 2019.
[33] L. Breiman, Classification and Regression Trees, Routledge, 2017.
[34] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (2009) 207–244.
[35] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Netw. 2 (1989) 359–366, doi:10.1016/0893-6080(89)90020-8.
[36] L. Breiman, Random forests, Mach Learn. 45 (2001) 5–32.
[37] F.S. Carvalho, J.P. Braga, Partial radial distribution functions for a two-component glassy solid, $GeSe_3$, from scattering experimental data using an artificial intelligence framework, J. Mol. Model. 28 (2022) 99, doi:10.1007/s00894-022-05055-5.