Designing optical glasses by machine learning coupled with genetic algorithms

Daniel R. Cassar, Gisele G. dos Santos, Edgar D. Zanotto

Department of Materials Engineering, Federal University of São Carlos, São Carlos, SP, Brazil

Abstract

Engineering new glass compositions has experienced a sturdy tendency to move forward from (educated) trial-and-error to data- and simulation-driven strategies. In this work, we developed a software that combines data-driven predictive models (in this case, neural networks) with a genetic algorithm aimed at designing glass compositions having desired combinations of properties. First, we induced predictive models for the glass transition temperature ($T_g$) using a dataset of 45,302 compositions with 39 different chemical elements, and for the refractive index ($n_d$) using a dataset of 41,225 compositions with 38 different chemical elements. Then, we searched for relevant glass compositions using a genetic algorithm informed by a design trend of glasses having high $n_d$ (1.7 or more) and low $T_g$ (500 °C or less). Two candidate compositions suggested by the combined algorithms were selected and produced in the laboratory. The experimental values of their properties were within the prediction uncertainty of our models. One of the glasses met all the constraints of the work, which supports the proposed framework, whereas the other missed the target by a very small margin. Therefore, this new tool can be immediately used for accelerating the design of new glasses. These results are a stepping stone in the pathway of machine learning-guided design of novel glasses for technological applications.

Keywords: oxide glasses, optical glasses, machine learning, genetic algorithms
1. Introduction

Glass science and technology are currently experiencing an "artificial intelligence renaissance". Even though many of the tools being used are not new, the interface between data and glass sciences has never seen so much interest.\textsuperscript{1–17} It is natural that the number of reports on machine learning-based property prediction of glasses has surged in the past three years due to the availability of powerful computational tools and hardware, and the recent licensing of the SciGlass database under a permissive license (https://github.com/epam/SciGlass)—which has approximately 400,000 entries on composition-properties of glasses. Moreover, the high correlation between composition and properties for inorganic non-metallic glasses makes the use of data-driven tools for these materials significantly easier than those for polycrystalline materials.

Most of the tools and ML models reported so far have been focused on predicting a chosen property given a glass composition.\textsuperscript{2,4,8,12,17–19} For new glass development, however, it is paramount to solve the inverse design problem, that is, finding possible compositions that are predicted to have a desired set of properties. To the best of our knowledge, this inverse design problem cannot be solved by traditional machine learning methods alone, but can be tackled by a combination of machine learning and optimization algorithms. In this context, Nakamura and co-authors\textsuperscript{16} recently used Bayesian optimization coupled with Gaussian process regression to search for oxide glasses with high refractive indices.

The objective of this work is to propose and test a framework to solve inverse design problems for glass development using genetic algorithms. While the proposed framework is general, here it will be tested for designing new optical glasses. The predictive models used here will be induced by neural networks.

2. Design trends in optical glasses

The growth of the smartphone market and the demand for increasingly smaller and better-defined security and car cameras have attracted significant attention and fostered optical glass research. There is enormous interest in obtaining increasingly smaller, thinner, and more efficient light transmission lenses. The recent technological advances in 4K and 8K applications and Virtual/Mixed Reality lenses are also fueling the field of optical glasses. A relevant review article by Peter Hartmann and co-authors\textsuperscript{20}, listed some of the hottest trends in optical glass research, which were: i) high refractive indices (1.7 or more), ii) a high Abbe number (60 or more), iii) high refractive indices and a high Abbe number, and iv) high refractive indices and a low Abbe number.

Glasses with high refractive indices are desired because they reduce the degree of spherical aberration and enable lens design with reduced dimensions, e.g., targeting the use in smartphones and small car cameras. Glasses with a high Abbe number (low dispersion) are used in optical systems that require a low degree of chromatic aberration. Glasses with high refractive indices and a high Abbe number could have a major impact on optical glass technology, as it would make it possible to obtain smaller and thinner lenses with less color dispersion, again targeting the smartphone market. Glasses
with high refractive indices and a low Abbe number are used for color correction in certain types of optical systems. Figure 1 shows a 2D histogram of the Abbe diagram for oxide glasses available in the SciGlass database. This figure clearly shows the current property envelope of the optical properties of oxide glasses.

**Figure 1.** 2D histogram visualization of the Abbe diagram showing the refractive indices versus the Abbe number. Approximately 24,300 oxide glass data points were used to build this plot. Each individual rectangle has horizontal sides of 1 and vertical sides of 0.01, and comprehends a density of experimental points depicted by its color (color bar on the right).

In addition to the four design trends previously discussed, there is additional interest in glasses having low glass transition temperature for optical lens production via precision molding techniques. This technique consists of applying pressure to a mold containing a glass-forming liquid, with controlled atmosphere and temperatures between the glass transition temperature ($T_g$) and the softening point. The material obtained in this way is already in its final form, without the need of expensive, time consuming additional steps, such as polishing and finishing. The most commonly used molds are made of tungsten carbide or silicon carbide with different types of coatings, and they are the most expensive parts of this process. Tungsten carbide molds, for example, are sensitive to oxidation at high temperatures, requiring the lenses to be conformed below 500 °C, ideally below 450 °C. However, the same weak intermolecular bonds that allow a glass to have a low $T_g$ often bring some disadvantages, such as low chemical stability, which makes meeting this constraint a significant challenge.
3. Materials and methods

3.1. Data collection and partition

All the data used in this work were collected from the SciGlass database, which is now licensed under the ODC Open Database License (OdbL). This database collects glass properties and their respective chemical compositions reported in scientific articles, books, and patents.

Here we collected data on glass transition temperature and refractive index ($n_d$) of oxide glasses to induce predictive models for these properties via neural networks. The definition of oxide glasses considered here is the same as what we have used in a previous work, that is: materials having an atomic fraction of oxygen of at least 0.3, and not having the chemical elements S, H, C, Pt, Au, F, Cl, N, Br, and I. Briefly, these are either elements that are too volatile or that occupy oxygen sites.

Before inducing the models, the dataset for each property was pre-processed following three steps: removal of glasses made with chemical elements having low representability, removal of glasses having properties with extremely low or high values, and replacement of duplicate entries by their median values. The first step is an iterative process where the fraction of examples containing each chemical element is computed, and then removing those glasses having chemical elements that are present in less than 1% of the examples. This process is repeated until all chemical elements are present in at least 1% of the examples. The rationale behind this choice is that each chemical element adds a new compositional dimension for the training of the model, and generalization may be compromised by having a small amount of examples.

The second step is related to examples having extreme values of the glass properties, which were also removed. Here, extreme property values are defined as those below the 0.05% percentile and above the 99.95% percentile. The rationale behind this choice is the knowledge that the SciGlass database does not curate its entries, and a significant portion of typos or mistakes are located in these extreme regions.

The third and final step is related to examples with duplicate features, i.e., entries that have the same nominal composition. These duplicate entries were grouped together into a single entry with the median value of the property. The rationale behind this choice is to avoid a problem called data leakage, where the prediction of the model is artificially improved because it had “access” to data in the reserved dataset for testing. In other words, with duplicated data it is possible that glasses having the same composition end up in different datasets, thus information in the test dataset can “leak” into training.

We computed descriptive statistics on the datasets after collection and pre-processing. After this step, each dataset was partitioned into the holdout dataset (20%) and the training and validation (80%) dataset. The holdout set was not used for training the models nor for hyperparameter tuning; its main purpose was to measure the predictive power of the models. Finally, the final model used in the genetic algorithm optimization was trained using all the pre-processed data, as this is the usual practice for inducing the final predictive model.
3.2. Property prediction using neural networks

Neural networks (NN) are a group of machine learning (ML) algorithms that are excellent at finding patterns in data. They are the most used type of ML algorithm in the field of oxide glasses, and their success is most probably due to the possibility of building a “universal regressor” when used as a supervised learning algorithm. The mathematical and statistical support for NNs are discussed in depth in the textbook by Charu Aggarwal.

In this work, we investigate shallow feedforward NNs with two hidden layers, which are good enough to predict glass properties with acceptable precision. One critical choice is the architecture of the NN, because it is known that different problems often require different architectures. Here we used a hyperparameter tuning routine to investigate some NN architectures, similarly as we did in a previous work. This process is described in detail in the Supplementary Material.

Each property of interest was investigated independently. In the end, we obtained a predictive model for each property. These models are functions for which the arguments lie in the chemical composition domain, and the output is a real number representing the predicted value of a given property. While the models can predict the properties for any glass with chemical elements within the domain of the functions, the expectation is that the prediction of compositions outside the training domain will lead to a much higher error.

3.3. Inverse design of glass compositions

Solving the inverse design problem, discussed in the introduction, requires an optimization algorithm. Many such algorithms are available; random search, Bayesian optimization, and simulated annealing are some examples. A Genetic Algorithm (GA) was chosen in this work mostly due to previous familiarity of one of the authors. GAs are heuristic algorithms inspired by the theory of evolution and natural selection, bringing concepts of individuals, population, selection, reproduction, and mutation, and using them to navigate the multi-dimensional space of a single- or multi-objective optimization problem. More information on GAs are available in Koza’s textbook for a general view, or Chakraborti’s article for a report focusing on materials design. To the best of our knowledge, the first published work to apply GA in the context of oxide liquids is that of Ojovan et al., whereas the first to apply GA in the context of oxide glasses is that of Tandia et al.

To use GA, one must first define how the “genome” of “individuals” are represented. In this work, an individual is defined as a glass with a certain chemical composition, having a genome that is represented as a row vector \( \mathbf{x} = [x_1, x_2, \ldots, x_{n}] \), where each “gene” \( x_i \) is an integer in the range \([0, 100]\) that stores the amount (in moles) of a certain chemical compound \( c_i \). Here, \( n=28 \) compounds were considered for the search space: \( \text{Al}_2\text{O}_3, \text{B}_2\text{O}_3, \text{BaO}, \text{Bi}_2\text{O}_3, \text{CaO}, \text{CdO}, \text{Gd}_2\text{O}_3, \text{GeO}_2, \text{K}_2\text{O}, \text{La}_2\text{O}_3, \text{Li}_2\text{O}, \text{MgO}, \text{Na}_2\text{O}, \text{Nb}_2\text{O}_5, \text{P}_2\text{O}_5, \text{PbO}, \text{Sb}_2\text{O}_3, \text{SiO}_2, \text{SnO}_2, \text{SrO}, \text{Ta}_2\text{O}_5, \text{TeO}_2, \text{TiO}_2, \text{WO}_3, \text{Y}_2\text{O}_3, \text{Yb}_2\text{O}_3, \text{ZnO}, \text{ZrO}_2 \).

A population \( \mathbf{P} \) with \( m \) individuals is defined as a matrix \( m \times n \), where each row holds the information of a single individual. The population size of this work was \( m=400 \) and the initial population \( \mathbf{P}_1 \) was generated by randomly sampling integers in the range of \([0, 100]\) and building a \( 400 \times 28 \) matrix. These randomly generated individuals are most probably poor solutions to the optimization problem that is
being investigated; many of them may not even form a glass or be within any constraint for which the problem is being optimized.

However, by pure chance, some randomly generated individuals of $P_1$ will be closer to a possible solution than others, even if they do not meet all the requirements of the problem. The word “possible” is emphasized because for any given inverse design problem, a solution may or may not exist, which is an issue that is not directly related to GA.

The next step is to select the individuals of $P_1$ that will “survive” to the next generation, and compose population $P_2$. To select the individuals, first their fitness score must be computed. To do so, a fitness function $f$ must be defined and computed for each individual. In this work, the fitness function was a weighed Euclidean distance in the property space, Eq. (1). The smaller the value of $f$ the better chances the individual has to survive.

$$f(x, y) = \sqrt{w_x(x - x_d)^2 + w_y(y - y_d)^2 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3}$$

In the previous equation, $x$ and $y$ are the values of two different properties of a certain individual (a glass composition, in this case); $x_d$ and $y_d$ are the desired values for these two properties, which depend on the inverse design problem that is being solved; $w_x$ and $w_y$ are the weights that each property has to compute the fitness score (1 for $T_g$ and 20 for $n_d$, in this work); and $\varepsilon_1$, $\varepsilon_2$, and $\varepsilon_3$ are penalty factors that will be discussed later on in this section. Here we studied only optimization problems with two properties, but Eq. (1) can be easily expanded for inverse design with more properties.

After calculating the fitness score for all individuals in $P_1$, the selection phase begins. There are some selection strategies available. In this paper, we used the tournament selection where 3 individuals are selected at random from the population, and the one with the lowest fitness score from this group is selected to be part of the next generation, which is $P_2$ in this example. This process continues until $P_2$ has the same number of individuals as $P_1$.

The next phase is mating, where pairs of $P_2$ individuals have a chance to exchange genetic material, which replaces the original pair (the parents) with two new individuals (the offspring). The new individuals have a uniform chance of receiving each bit of genetic material from both parents, a process called uniform crossover. The chance of mating was set to 50%.

Finally, the last step of this iteration is the mutation phase. This is a critical step as it is the only opportunity for introducing distinct genetic material that was not present in the randomly generated $P_1$. Here, each individual of $P_2$ has a 20% chance to undergo mutation. If selected, then each gene has a 5% chance of changing its value to a random integer in the range of [0, 100]. On one hand, if the mutation probabilities are too high, then the problem may not converge, as the “memory” of the best individuals are easily lost to mutations. On the other hand, if the mutation probabilities are too low, then the number of iterations required to reach a solution may become prohibitively large.

After these steps, the whole process is repeated by computing the fitness score and performing selection, mating, and mutation on $P_2$ to generate $P_3$. This iterative process was done until a solution was found, or generation 5000 was reached.
We introduced two constraints for the GA search, one related to the minimum amount of glass-formers, and the other related to the chemical domain for which the predictive models were trained. Both constraints were computed independently for each individual. The first constraint checks for the ratio $\varphi$ between the sum of the glass network-forming oxides ($\text{Al}_2\text{O}_3$, $\text{SiO}_2$, $\text{B}_2\text{O}_3$, $\text{GeO}_2$, $\text{P}_2\text{O}_5$, $\text{Sb}_2\text{O}_3$, and $\text{TeO}_2$) and the total sum of compounds. If this ratio was below 45%, then a penalty $\varepsilon_1=(100(0.45−\varphi))^2$ was computed and considered in Eq. (1), otherwise $\varepsilon_1=0$. The rationale behind this constraint is to increase the chances that a composition found by the algorithm can be made into a glass. We are aware that this procedure does not guarantee that all compositions that meet this constraint can be vitrified by laboratory melt and quench techniques, however it significantly increases the chances.

The second constraint checks if the composition is inside the chemical domain of the predictive models that are considered in the calculation of $f$ in Eq. (1). For each chemical element $i$ that is present in the individual, a distance $d_i$ is computed, which is zero if the atomic fraction of the said element is within the chemical domain of all the predictive models, or it is the absolute difference between the atomic fraction of the element and the closest atomic fraction within the domain of all predictive models. The penalty $\varepsilon_2=(100\sum_i d_i)^2$ is then computed for each individual. The rationale behind adding this constraint is that NNs trained using only the chemical composition as features are prone to higher prediction errors for compositions that are outside the domain. Sometimes, however, it may be desirable to explore chemical compositions close to the training domain, but not necessarily within in. Here, we relaxed the composition domain of each chemical element by 20%.

To have a clear difference in the fitness score between individuals that meet all constraints and individuals that do not, a final penalty $\varepsilon_3$ is computed: if $\varepsilon_1 \neq 0$ or $\varepsilon_2 \neq 0$, then $\varepsilon_3=100$, otherwise $\varepsilon_3=0$. The rationale behind adding this penalty is that we do not want to allow that individuals outside the constraints of the problem have even a small chance of winning the selection tournament against individuals that meet all the constraints.

Finally, being a heuristic algorithm, GA is not guaranteed to reach a solution even if it exists. Any solution obtained is dependent on the randomly generated first population and the various steps that are due to chance. Because of this, the GA code was run many times to obtain a diverse set of solutions. The code used in this work was written in Python using the DEAP module\textsuperscript{30}, and it is available under the GPL3 license as the GLAS module\textsuperscript{31}, which stands for Genetic Lookup for Amorphous Substances.

### 3.4. Experimental tests

The design trend that guided our research was that of optical glasses having high refractive indices (1.7 or more) and low glass transition temperature (500 °C or less). During an exploratory phase, we observed that some candidate glasses had poor chemical durability. Because of this problem, we manually reduced the search domain of the elements boron and phosphorous to [0, 0.02] and [0, 0.03] in atomic fraction, respectively. The rationale was that these two elements often decrease the chemical durability of glasses.

We obtained many composition candidates by running the GLAS software several times. From the candidate list, we selected two glasses which we deemed to be viable to melt in our laboratory (with the
available resources). Their compositions are shown in the Table 1 together with the target and predicted values of the properties of interest.

**Table 1.** Composition (mol%), target, and predicted properties of the two glasses produced in this work. † To make this glass, we did not use Nb₂O₅ as it was not available in our laboratory at the time, instead we replaced it with La₂O₃. ‡ Moreover, we did not use MnO to avoid a strong color, instead we replaced it with ZnO. * The uncertainty in the prediction is estimated by the RMSE value reported in the Table 4.

| Oxide     | Glass 1 | Glass 2 |
|-----------|---------|---------|
| SiO₂      | 66.67   | 41.75   |
| B₂O₃      | 3.03    | 0       |
| Li₂O      | 3.03    | 0       |
| CaO       | 3.03    | 1.94    |
| La₂O₃     | 0†      | 0.97    |
| Sb₂O₃     | 21.21   | 27.18   |
| Nb₂O₅     | 3.03†   | 0       |
| GeO₂      | 0       | 7.77    |
| K₂O       | 0       | 8.74    |
| Na₂O      | 0       | 3.88    |
| SnO₂      | 0       | 2.91    |
| ZnO       | 0       | 0.97‡   |
| ZrO       | 0       | 1.94    |
| MnO       | 0       | 1.94‡   |

| Target property | Glass 1 | Glass 2 |
|-----------------|---------|---------|
| Refractive index| 1.70    | 1.75    |
| Glass transition temperature (°C) | 450 | 400 |

| Predicted property | Glass 1 | Glass 2 |
|--------------------|---------|---------|
| Refractive index   | 1.73(3) | 1.76(3) |
| Glass transition temperature (°C) | 460(30) | 400(30) |

The reactants used and their respective purity are reported in the Supplementary Material. To make each glass, the chemicals were mixed, weighed, and homogenized in a rotation jar mill for 12 hours. At the end of this process, the mixture was melted in a platinum crucible at a temperature range of 1000–1200 °C in a Deltech electric furnace, then poured over a metallic surface, crushed, and remelted for homogenization. This process was repeated three times. The melt was finally poured into a 1.5 × 1.5 × 3 cm graphite mold.
The glass transition temperature was determined for small pieces of the glasses by Differential Scanning Calorimetry (DSC, NETZSCH STA 449 F3 Jupiter), with a heating rate of 10 °C/min. The refractive index was measured in 1.5 × 1.5 × 1.5 cm samples using the Na d-line (589.6 nm) of a Carl Zeiss Jena Pulfrich-refractometer PR2. Two adjacent faces of the samples (those that interacted with the light beam in the refractometer) were ground using 150–1200 mesh sandpaper and polished in velvet fabric with an aqueous cerium oxide suspension. No sign of chemical attack was observed. The refraction angle was measured and converted to refractive index by using a conversion table provided by the equipment manufacturer.

4. Results and discussion

4.1. Data analysis

Figure 2 shows the histogram for the two datasets used in this work to induce the predictive neural network models. The distribution of the refractive index values has a single mode, with a clear skew to the right. The distribution of the glass transition temperature also has a single mode, but is not visually skewed. The Table 2 shows the descriptive statistics of both distributions.

Figure 2: Distribution of the values of the (a) refractive index dataset, and the (b) glass transition temperature dataset.
Table 2: Descriptive statistics of the datasets used for inducing the predictive neural network model. Glass transition temperature in Kelvin.

| Statistic            | Refractive index | Glass transition temperature |
|----------------------|------------------|------------------------------|
| Count                | 41,225           | 45,302                       |
| Number of chemical elements | 38               | 39                           |
| Mean                 | 1.69             | 778.28                       |
| Standard deviation   | 0.18             | 150.56                       |
| Minimum              | 1.41             | 380.15                       |
| Median               | 1.64             | 773.15                       |
| Maximum              | 2.67             | 1271.15                      |
| Skewness             | 1.31             | 0.14                         |
| Kurtosis             | 2.11             | −0.33                        |

The Fig. 3 complements the analysis of the datasets by showing the distribution of examples with respect to the number of chemical elements. Glasses made with 4 chemical elements are the most common in both datasets, and multi-component glasses made with more than 10 elements are significantly less frequent than other multi-component glasses. The Table 3 shows the chemical domain of both datasets, indicating the minimum and maximum atomic fraction for each element. As already mentioned, this information is relevant during the genetic algorithm search, as candidates that fall outside the intersection of chemical domains are penalized.

Figure 3: Distribution of the number of different chemical elements that make the glasses in the (a) refractive index dataset, and (b) glass transition temperature dataset.
Table 3: Chemical domain for the refractive index and the glass transition temperature datasets in atomic fraction. Elements with a dash (–) are not present in the dataset.

| Element | Refractive index | Glass transition temperature |
|---------|------------------|------------------------------|
|         | Min | Max | Min  | Max  |
| Ag      | –   | –   | 0    | 0.421 |
| Al      | 0   | 0.38| 0    | 0.367 |
| As      | 0   | 0.4 | 0    | 0.4   |
| B       | 0   | 0.4 | 0    | 0.4   |
| Ba      | 0   | 0.326| 0   | 0.244 |
| Be      | 0   | 0.183| –   | –     |
| Bi      | 0   | 0.374| 0   | 0.376 |
| Ca      | 0   | 0.273| 0   | 0.308 |
| Cd      | 0   | 0.312| –   | –     |
| Ce      | –   | –   | 0   | 0.158 |
| Cs      | 0   | 0.4 | 0    | 0.456 |
| Cu      | –   | –   | 0   | 0.333 |
| Er      | 0   | 0.171| 0   | 0.143 |
| Fe      | 0   | 0.222| 0   | 0.316 |
| Ga      | 0   | 0.4 | 0    | 0.334 |
| Gd      | 0   | 0.4 | 0    | 0.179 |
| Ge      | 0   | 0.376| 0   | 0.385 |
| K       | 0   | 0.418| 0   | 0.497 |
| La      | 0   | 0.4 | 0    | 0.255 |
| Li      | 0   | 0.471| 0   | 0.584 |
| Mg      | 0   | 0.304| 0   | 0.228 |
| Mn      | 0   | 0.219| 0   | 0.231 |
| Mo      | –   | –   | 0   | 0.22  |
| Na      | 0   | 0.553| 0   | 0.553 |
| Nb      | 0   | 0.26| 0    | 0.263 |
| Nd      | 0   | 0.175| 0   | 0.2   |
| O       | 0.379| 0.739| 0.316| 0.745 |
| P       | 0   | 0.286| 0    | 0.286 |
| Pb      | 0   | 0.437| 0   | 0.442 |
| Sb      | 0   | 0.4 | 0    | 0.4   |
### 4.2. Predictive models

The Table 4 shows the hyperparameters used to induce the predictive neural networks. As expected, different problems often require different NN architectures, which is observed here by the architecture for predicting the refractive index that is reasonably different from the one to predict the glass transition temperature.

The Table 4 also shows some metrics for the two models computed for the holdout dataset, which was not used for training the NNs nor during the hyperparameter tuning routine. Therefore, the metrics computed with this dataset simulate how the models behave with new unseen data. However, it is important to stress that the final predictive model used in the GLAS software was trained with all the available data, as our interest is to build the best predictive model within the considered framework. We expect that this final model will have a smaller prediction error to the one trained with only 80% of the dataset, however by using the whole dataset to train the NN we lost the ability to estimate the prediction errors. In other words, the expected errors of the final model are probably lower than the ones shown in the Table 4, but we are unable to estimate them.
Table 4: Hyperparameters used to induce the predictive neural networks and metrics of the models. The hyperparameter tuning procedure is described in the Supplementary Material. † The numbers in parentheses refer to the hyperparameters of the first and second hidden layer, respectively. ‡ There are many ways to compute $R^2$; here it was computed considering a linear model without an intercept as the alternative hypothesis.

| Hyperparameter                          | Refractive index | Glass transition temperature |
|-----------------------------------------|------------------|------------------------------|
| Activation function                     | ReLU             | Sigmoid                      |
| Number of neurons †                     | (295, 115)       | (190, 290)                   |
| Dropout †                               | (11%, 27%)       | (8.2%, 25%)                  |
| Adam optimizer learning rate            | $3.6 \times 10^{-4}$ | $1.3 \times 10^{-3}$        |
| Adam optimizer epsilon                  | $7.05 \times 10^{-7}$ | $2.57 \times 10^{-5}$      |
| Patience of the early stopping routine  | 12               | 14                           |
| Batch size                              | 256              | 128                          |

| Metrics                                 | Refractive index | Glass transition temperature |
|-----------------------------------------|------------------|------------------------------|
| Coefficient of determination ‡ $(R^2)$ | 0.9997           | 0.998                        |
| Relative deviation (RD)                 | 0.9%             | 2.7%                         |
| Root mean squared error (RMSE)          | 0.031            | 34 K                         |
| Mean absolute error (MAE)               | 0.016            | 20 K                         |
| Median absolute error (MedAE)           | 0.010            | 12 K                         |

The Figs. 4 and 5 complement the analysis of the predictive models by showing a correlation plot between predicted and reported values of the properties, and the mean and standard deviation of the prediction residuals for each chemical element. All these calculations were made for the holdout dataset, again to understand how the predictive models behave in interpolating unseen data. One feature worth noting is the distribution of the prediction residuals for the refractive index, shown in the inset of Fig. 4a. This distribution is not symmetric, most probably because the distribution of $n_d$ values is not symmetric itself (see Fig. 2a).

Another relevant observation is that the prediction error depends on the chemical element present in the glasses, but does not seem to depend on the number of examples used for training in the considered framework. An example is that the standard deviation of the glass transition prediction residuals of glasses containing vanadium is significantly higher than that of glasses containing tantalum, even though the first is more frequent in the training dataset than the second. As experimentalists know, transition metals and volatile substances require special care for glass preparation, which can explain this difference. All the chemical elements used to prepare Glass 1 and Glass 2 have an “average” standard deviation of the prediction residuals, with the only exception being germanium when predicting the refractive index (only present in Glass 2), for which the standard deviation is reasonably high.
Figure 4: 2D histogram of the predicted versus reported values for (a) refractive index and (b) glass transition temperature, computed for the holdout dataset. The identity line is shown in black. The inset is the histogram of the prediction residuals, the difference between the reported and the predicted values. The vertical color bar shows the frequency of data points.
Figure 5: Mean and standard deviation of the prediction residual for each chemical element in the holdout dataset. (a) Refractive index and (b) glass transition temperature. The number in parenthesis is the number of glass compositions containing the chemical element in the holdout dataset. The prediction residuals are the difference between the reported and the predicted values. The order of the elements is from the least to the most frequent, from left to right.

As an exercise to test the induced models, we determined the predicted influence of certain oxides that are typically used to increase the refractive index of glasses, shown in the Fig. 6. These plots can give a semi-quantitative indication of the effect of each oxide on the refractive index and glass transition temperature of 4 glass-formers: SiO$_2$, B$_2$O$_3$, GeO$_2$, and TeO$_2$.
All the tested heavy oxides significantly increase the refractive index in silicate, borate, and germanate glasses. However, tellurite glasses show distinct behavior because pure TeO$_2$ already has a very high refractive index (close to 2.16). For tellurides, La$_2$O$_3$ and ZrO$_2$ decrease the index, PbO practically does not alter it, whereas Bi$_2$O$_3$, Nb$_2$O$_5$, and TiO$_2$ increase it in this general order.

Regarding the glass transition temperature, PbO and Bi$_2$O$_3$ decrease it, whereas the other 4 oxides increase it. Exceptions are found for borate systems due to the well-known boron anomaly. Another exception is for the addition of TiO$_2$ and ZrO$_2$ above about 10 mol% in silicate glasses.

This combined information about the effects of several oxides on $T_g$ and $n_d$ shows that this tool might be quite useful for designing new optical glasses.

4.3. Experimental tests

The glass compositions we selected for the experimental tests (Table 1) were those containing a balanced amount of glass formers and other elements typically found in optical glasses. To make Glass 1, we replaced niobium oxide for lanthanum oxide as the first was not available in our laboratory when the experimental phase began. To make Glass 2, we replaced manganese oxide for zinc oxide due to the variation of oxidation numbers in the former that could give strong colors to our glass. We believe that this educated manual interference and modification of the candidate compositions is still an essential part of a glass design process, as it still is a significant challenge to translate all the nuances gathered after many years of accumulated glass-making know-how into computer code.

Even with some manual selection and modifications in the suggested glass compositions, our glasses end up showing a slightly yellowish color. This is likely due to the fact that some elements, such as antimony and tin have variable valences. The viscosity of Glass 1 was considerably high, making it a
challenge to obtain a homogeneous glass free of striae. This is not surprising when dealing with optical
glasses, hence this glass required using a bar-shaped mold and a special casting technique to avoid
cords.

The DSC traces used to measure the glass transition temperature are shown in Fig. 7, for which we
obtained a value of 500(15) °C for Glass 1 and 450(13) °C for Glass 2. These analyses and
computations of the uncertainty were done using a software developed by Matthew Mancini\textsuperscript{32}. These
values are close to the predicted values of 460(30) and 400(30) °C. $T_g$ is a tricky property to predict as
it also depends on the rate of temperature change. Nevertheless, both glasses met the $T_g$ design trend
that informed this work, with Glass 2 having the lowest value of $T_g$, an advantage over Glass 1.

![Figure 7](image)

**Figure 7.** DSC traces focused on the region of the glass transition for (a) Glass 1 and (b) Glass 2. A
linear baseline was subtracted for building the plots. The dashed red line shows $T_g$ and the
dotted gray lines show the range of $T_g$ considering the uncertainty.

The measured refractive index for Glass 1 and Glass 2 were 1.686(1) and 1.749(1). The first is
smaller than the predicted value range of 1.73(3), while the second is within the predicted value range
of 1.76(3). It is important to mention that these predicted values were for the original compositions,
before the (minor) manual changes that we discussed in the first paragraph of this section. Indeed, our
simulations (Fig. 6) show that the effect of La$_2$O$_3$ on the refractive index is smaller than that of Nb$_2$O$_5$
(which was a suggested component by the GA). Therefore, it is no surprise that the measured index of
Glass 1 is smaller than the predicted value.

Despite the fact that Glass 1 does not exactly meet the value that informed this work (refractive
index of 1.7 or more), we provided a reasonable explanation for this difference. While this result could
be seen as (mildly) negative, it serves as a reminder that computer predictions are not to be taken as a
be-all and end-all solution; experimental tests continue to be necessary. All in all, these procedures,
tools, and results are compelling, and we believe are a stepping stone in the pathway of machine
learning-guided design of new glasses for technological applications.
5. Summary and Conclusion

We developed a new software that couples data-driven predictive models with a genetic algorithm to solve inverse design of new glass compositions. After training predictive models for the glass transition temperature and refractive index, we searched for relevant glass compositions guided by a design trend regarding optical glasses—high refractive index and low glass transition temperature. Two candidate compositions suggested by the combined algorithms were selected and produced in the laboratory. The experimental properties of these glasses were close to the predictions of our models, which supports the proposed framework. Therefore, this new tool can be immediately used for accelerating the design of new glasses, significantly minimizing trial-and-error. Moreover, as it reduces the quantity of resources needed, it contributes to a greener approach to glass development.

Acknowledgments

This study was financed by the São Paulo State Research Foundation support (FAPESP grant numbers 2017/12491-0 and 2013/07793-6), in part by the National Council for Scientific and Technological Development (CNPq, grant number: PQ 303886/2015-3 and 167434/2017-9), and in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. The NSG Overseas grant is greatly appreciated. We would also like to thank our colleagues Ricardo Lancelotti and Dr. Laís Dantas for helping with the refractive index and DSC measurements, respectively.

6. References

1. O. Bošák, S. Minárik, V. Labaš, Z. Ančíková, P. Košťial, O. Zimný, M. Kubliha, M. Poulain, and M.T. Soltani, Journal of Optoelectronics and Advanced Materials 18, 240 (2016).
2. J.C. Mauro, A. Tandia, K.D. Vargheese, Y.Z. Mauro, and M.M. Smedskjaer, Chem. Mater. 28, 4267 (2016).
3. N.M. Anoop Krishnan, S. Mangalathu, M.M. Smedskjaer, A. Tandia, H. Burton, and M. Bauchy, Journal of Non-Crystalline Solids 487, 37 (2018).
4. D.R. Cassar, A.C.P.L.F. de Carvalho, and E.D. Zanotto, Acta Materialia 159, 249 (2018).
5. J.C. Mauro, Current Opinion in Solid State and Materials Science 22, 58 (2018).
6. S. Bishnoi, S. Singh, R. Ravinder, M. Bauchy, N.N. Gosvami, H. Kodamana, and N.M.A. Krishnan, Journal of Non-Crystalline Solids 524, 119643 (2019).
7. H. Liu, Z. Fu, K. Yang, X. Xu, and M. Bauchy, Journal of Non-Crystalline Solids 119419 (2019).
8. A. Tandia, M.C. Onbasli, and J.C. Mauro, in Springer Handbook of Glass, edited by J.D. Musgraves, J. Hu, and L. Calvez (Springer International Publishing, Cham, 2019), pp. 1157–1192.
9. K. Yang, X. Xu, B. Yang, B. Cook, H. Ramos, N.M.A. Krishnan, M.M. Smedskjaer, C. Hoover, and M. Bauchy, Scientific Reports 9, 8739 (2019).
10. E. Alcobaça, S.M. Mastelini, T. Botari, B.A. Pimentel, D.R. Cassar, A.C.P. de L.F. de Carvalho, and E.D. Zanotto, Acta Materialia 188, 92 (2020).
11. D.R. Cassar, ArXiv:2007.03719 [Cond-Mat, Physics:Physics] (2020).
12. B. Deng, Journal of Non-Crystalline Solids **529**, 119768 (2020).
13. T. Han, N. Stone-Weiss, J. Huang, A. Goel, and A. Kumar, Acta Biomaterialia **107**, 286 (2020).
14. Y.-J. Hu, G. Zhao, M. Zhang, B. Bin, T. Del Rose, Q. Zhao, Q. Zu, Y. Chen, X. Sun, M. de Jong, and L. Qi, Npj Comput Mater **6**, 25 (2020).
15. J.N.P. Lillington, T.L. Goüt, M.T. Harrison, and I. Farnan, Journal of Non-Crystalline Solids **533**, 119852 (2020).
16. K. Nakamura, N. Otani, and T. Koike, J. Ceram. Soc. Japan **128**, 569 (2020).
17. R. Ravinder, K.H. Sridhara, S. Bishnoi, H.S. Grover, M. Bauchy, Jayadeva, H. Kodamana, and N.M.A. Krishnan, Mater. Horiz. **7**, 1819 (2020).
18. C. Dreyfus and G. Dreyfus, Journal of Non-Crystalline Solids **318**, 63 (2003).
19. D.S. Brauer, C. Rüssel, and J. Kraft, Journal of Non-Crystalline Solids **353**, 263 (2007).
20. P. Hartmann, R. Jedamzik, S. Reichel, and B. Schreder, Appl. Opt. **49**, D157 (2010).
21. S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, ACM Trans. Knowl. Discov. Data **6**, 15:1 (2012).
22. J. Ruusunen, Deep Neural Networks for Evaluating the Quality of Tempered Glass, M.Sc Dissertation, Tampere University of Technology, 2018.
23. M.C. Onbaşlı, A. Tandia, and J.C. Mauro, in *Handbook of Materials Modeling: Applications: Current and Emerging Materials*, edited by W. Andreoni and S. Yip (Springer International Publishing, Cham, 2020), pp. 1997–2019.
24. C.C. Aggarwal, *Neural Networks and Deep Learning: A Textbook* (Springer International Publishing, 2018).
25. J. Bergstra, D. Yamins, and D.D. Cox, in *Proceedings of the 12th Python in Science Conference* (2013), pp. 13–20.
26. J.S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, in *Advances in Neural Information Processing Systems* (2011), pp. 2546–2554.
27. J.R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection* (MIT Press, Cambridge, Mass, 1992).
28. N. Chakraborti, International Materials Reviews **49**, 246 (2004).
29. M.I. Ojoan, K.P. Travis, and R.J. Hand, Journal of Physics: Condensed Matter **19**, 415107 (2007).
30. F.-A. Fortin, F.-M.D. Rainville, M.-A. Gardner, M. Parizeau, and C. Gagné, Journal of Machine Learning Research **13**, 2171 (2012).
31. D.R. Cassar, *drcassar/glas: GLAS v0.1.0.dev2* (Zenodo, 2020). [https://zenodo.org/record/3991781](https://zenodo.org/record/3991781)
32. M. Mancini and M. Sendova, Under Review for the Journal of the American Ceramic Society (2020).
Supplementary material to “Designing optical glasses by machine learning coupled with genetic algorithms”

Daniel R. Cassar, Gisele G. dos Santos, Edgar D. Zanotto

Department of Materials Engineering, Federal University of São Carlos, São Carlos, SP, Brazil

S1. Chemicals used to produce the glasses

The chemical reactants used in this work are shown in the Table S.1. We used nitrates (when available) to create an oxidative atmosphere during the melting operation to control the oxidation numbers, as well as avoid chemical attacks on our platinum crucible.

Table S.1. Composition, manufacturer, and purity of the chemical reagents used in this work.

| Substance       | Manufacturer      | Purity       |
|-----------------|-------------------|--------------|
| SiO₂            | Aldrich           | >99.99%      |
| H₃BO₃           | Vetec             | 99.5%        |
| LiNO₃           | Aldrich           | 95%          |
| Ca(NO₃)₂·4H₂O   | Vetec             | 99%          |
| La₂O₃           | Alfa Aesar        | 99.99%       |
| Sb₂O₃           | Aldrich           | >99%         |
| GeO₂            | Riedel-de-Haen    | >99%         |
| KNO₃            | Aldrich           | >99%         |
| NaNO₃           | Aldrich           | >99%         |
| SnO₂            | Alfa Aesar        | 99.90%       |
| ZnO             | Riedel-de-Haen    | >99%         |
| ZrO             | Alfa Aesar        | 99.70%       |
S2. Hyperparameter tuning

Hyperparameter tuning was done using the Python module hyperopt\textsuperscript{25}. The search space of the hyperparameters are shown in the Table S.2, and it was navigated with suggestions from a Tree-structured Parzen Estimator (TPE) algorithm\textsuperscript{26}. A total of 150 hyperparameter sets were tested for each property of interest.

Before the hyperparameter tuning, the training and validation dataset was partitioned into 80\% for local training, 10\% for local validation, and 10\% for local testing. Please note that the training and validation dataset was defined in Section 3.1 of the manuscript, and it does not contain the data in the holdout dataset. For each of the 150 sets of hyperparameters tested, a neural network was trained with this local training dataset and validated on the local validation dataset after each epoch. The validation step is important as the training stops if there is no improvement in the prediction of the validation dataset for a certain number of epochs defined by the patience hyperparameter. If this early stopping routine is never met, the neural network is then trained for 500 epochs.

Each of the 150 hyperparameter sets received a score value that is the mean squared error (MSE) of the prediction of the local test dataset. Those 10 sets with the lowest MSE score were tested again, this time in a 5-fold cross-validation analysis, where the hyperparameter set with the lowest average MSE score (considering all the folds) was the one selected to induce the final models. The selected sets of hyperparameters are shown in the main manuscript in the Table 4.

Table S.2. Search space of the hyperparameters of the neural networks. ReLU is the rectifier linear unit function and ELU is the exponential linear unit.

| Hyperparameter                              | Search space            |
|---------------------------------------------|-------------------------|
| Activation function                         | ReLU, ELU, or Sigmoid   |
| Number of neurons in the first layer         | [20, 300]               |
| Number of neurons in the second layer        | [20, 300]               |
| Dropout probability of the first layer (%)  | [0, 30]                 |
| Dropout probability of the first layer (%)  | [0, 30]                 |
| Adam optimizer learning rate                | $[10^{-4}, 10^{-2}]$   |
| Adam optimizer epsilon                      | $[10^{-7}, 10^{-3}]$   |
| Patience of the early stopping routine      | [10, 14]                |
| Batch size                                  | 64, 128, or 256         |