SCIENTIFIC REPORTS

OPEN

Received: 5 February 2019
Accepted: 4 June 2019
Published online: 19 June 2019

Predicting the Young's Modulus of Silicate Glasses using High-Throughput Molecular Dynamics Simulations and Machine Learning

Kai Yang[1], Xinyi Xu[1], Benjamin Yang[1], Brian Cook[1], Herbert Ramos[1], N. M. Anoop Krishnan[2,3], Morten M. Smedskjaer[4], Christian Hoover[5] & Mathieu Bauchy[1]

The application of machine learning to predict materials' properties usually requires a large number of consistent data for training. However, experimental datasets of high quality are not always available or self-consistent. Here, as an alternative route, we combine machine learning with high-throughput molecular dynamics simulations to predict the Young's modulus of silicate glasses. We demonstrate that this combined approach offers good and reliable predictions over the entire compositional domain. By comparing the performances of select machine learning algorithms, we discuss the nature of the balance between accuracy, simplicity, and interpretability in machine learning.

Improving the mechanical properties of glasses is crucial to address major challenges in energy, communications, and infrastructure[1]. In particular, the stiffness of glass (e.g., its Young's modulus E) plays a critical role in flexible substrates and roll-to-roll processing of displays, optical fibres, architectural glazing, ultra-stiff composites, hard discs and surgery equipment, or lightweight construction materials[1–4]. Addressing these challenges requires the discovery of new glass compositions featuring tailored mechanical properties[5,6].

Although the discovery of new materials with enhanced properties is always a difficult task, glassy materials present some unique challenges. First, a glass can be made out of virtually all the elements of the periodic table if quenched fast enough from the liquid state[7]. Second, unlike crystals, glasses are out-of-equilibrium phases and, hence, do not have to obey any fixed stoichiometry[8]. These two unique properties of glass open limitless possibilities for the development of new compositions with enhanced properties—for instance, the total number of possible glass compositions[7] has been estimated to be around 10^{52}! Clearly, only a tiny portion of the compositional envelope accessible to glass has been explored thus far.

The design of new glasses for a targeted application can be formulated as an optimization problem, wherein the composition needs to be optimized to minimize or maximize a cost function (e.g., the Young's modulus) while satisfying some constraints (e.g., ensuring low cost and processability)[9]. Although the vast compositional envelop accessible to glass opens limitless possibilities for compositional tuning, optimization problems in such highly-dimensional spaces are notoriously challenging—which is known as the "curse of dimensionality." Namely, the virtually infinite number of possible glass compositions render largely inefficient traditional discovery methods based on trial-and-error Edisonian approaches[10].

To overcome this challenge, the development of predictive models relating the composition of glasses to their engineering properties is required[9]. Ideally, physics-based models should offer the most robust predictions. In the case of glass stiffness, the Makishima–Mackenzie (MM) model may be the most popular predictive model[11,12]. This approach is essentially an additive model, wherein stiffness is expressed as a linear function of the oxide concentrations. However, such additive models are intrinsically unable to capture any non-linear compositional dependence, as commonly observed for stiffness[1,5,13]. On the other hand, molecular dynamics (MD) simulations offer a powerful method to compute the stiffness of a given glass[14,15]. However, MD is a brute-force method, that

[1]Physics of AmoRphous and Inorganic Solids Laboratory (PARISlab), University of California, Los Angeles, CA, 90095, USA. [2]Department of Civil Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, 110016, India. [3]Department of Material Science and Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, 110016, India. [4]Department of Chemistry and Bioscience, Aalborg University, 9220, Aalborg, Denmark. [5]School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ 85287, USA. Correspondence and requests for materials should be addressed to M.B. (email: bauchy@ucla.edu)

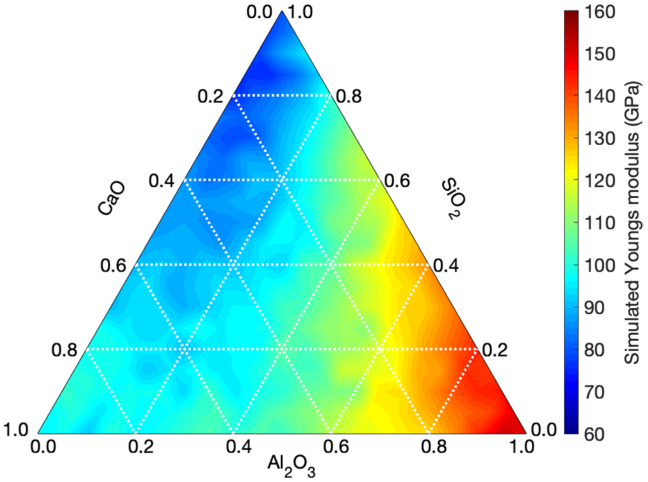

Figure 1. Ternary diagram showing the Young's modulus values predicted by high-throughput molecular dynamics simulations as a function of composition in the CaO–Al$_2$O$_3$–SiO$_2$ glass system. This database consists of 231 compositions homogeneously distributed over the entire compositional domain with 5 mol% increments in the oxide concentrations. This database is used as a basis to train the machine learning models presented herein.

is, it requires (at least) one simulation per glass composition—so that the systematic use of MD to explore the large compositional envelop accessible to glass is not a realistic option.

In turn, machine learning (ML) offers an attractive and pragmatic approach to predict glasses' properties[16]. In contrast with physics-based models, ML-based models are developed by "learning" from existing databases. Although the fact that glass composition can be tuned in a continuous fashion renders glass an ideal material for ML methods, the application of ML to this material has been rather limited thus far[16–20]. This partially comes from the fact that ML methods critically relies on the existence of "useful" data. To be useful, data must be (i) available (i.e., easily accessible), (ii) complete (i.e., with a large range of parameters), (iii) consistent (i.e., obtained with the same testing protocol), (iv) accurate (i.e., to avoid "garbage in, garbage out" models), and (v) representative (i.e., the dataset needs to provide enough information to train the models). Although some glass property databases do exist[21], some inconsistencies in the ways glasses are produced or tested among various groups may render challenging their direct use as training sets for ML methods—or would require some significant efforts in data cleaning and non-biased outlier detection.

To overcome these challenges, we present here a general method wherein high-throughput molecular dynamics simulations are coupled with machine learning methods to predict the relationship between glass composition and stiffness. Specifically, we take the example of the ternary calcium aluminosilicate (CAS) glass system—which is an archetypical model for alkali-free display glasses[22]—and focus on the prediction of their Young's modulus. We show that our method offers good and reliable predictions of the Young's modulus of CAS glasses over the entire compositional domain. By comparing the performance of select ML algorithms—polynomial regression (PR), LASSO, random forest (RF), and artificial neural network (ANN)—we show that the artificial neural network algorithm offers the highest level of accuracy. Based on these results, we discuss the balance between accuracy, complexity, and interpretability offered by each ML method.

Results

Molecular dynamics simulations.

MD simulations are first used to generate a series of 231 glasses that homogeneously span the CAS compositional ternary domain (see Methods section). The Young's modulus (E) of each glass is then computed by MD. We first focus on the compositional dependence of the Young's modulus values predicted by the MD simulations (see Fig. 1). Overall, we observe the existence of two main trends: (i) E tends to increase with increasing Al$_2$O$_3$ concentration and (ii) E tends to increase with increasing CaO concentration. However, we note that the dependence of E on composition is non-systematic and that CaO and Al$_2$O$_3$ have some coupled effects. For example, we find that E increases as the concentration of CaO increases when [Al$_2$O$_3$] = 0 mol%, whereas E decreases with increasing CaO concentration when [Al$_2$O$_3$] >40 mol%. Overall, we find that E exhibits a non-linear dependence on composition—so that one likely cannot rely on simple additive models to predict Young's modulus in the CAS system.

Relationship between composition and Young's modulus.

We now discuss the nature of the relationship between composition and Young's modulus. In general, the Young's modulus tends to increase with increasing connectivity[4]. To assess whether this trend is here satisfied (and whether it can be used to predict the linkage between composition and E), we compute based on the MD simulations the average coordination number $<r>$ of the atoms in the network for each glass composition. As shown in Fig. 2a, we find that $<r>$ increases with increasing CaO and Al$_2$O$_3$ concentrations. This arises from that fact that (i) Ca atoms have a large coordination number (around 6), while (ii) the addition of Al atoms tends to increase the degree of polymerization of the glass, i.e., by converting non-briding oxygen (NBO) into bridging oxygen (BO) atoms (we also note the formation of

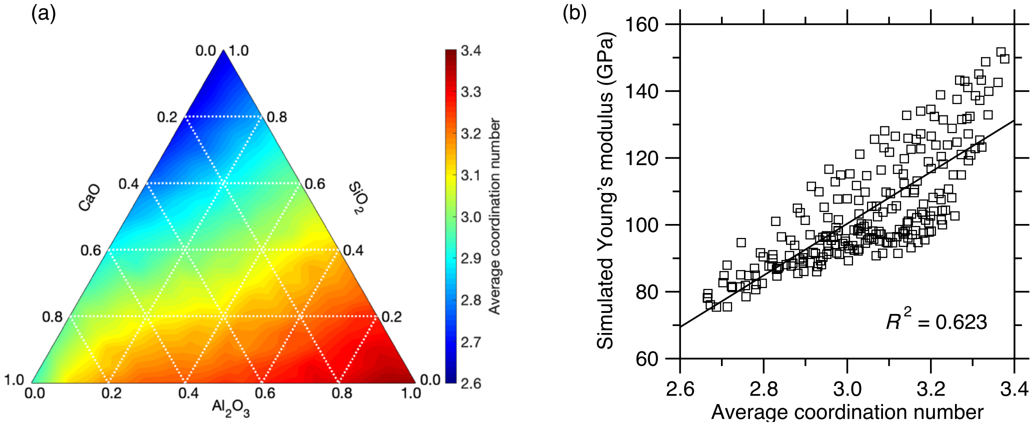

Figure 2. (**a**) Ternary diagram showing the average atomic coordination number computed by high-throughput molecular dynamics simulations as a function of composition in the CaO–Al_2O_3–SiO_2 glass system. (**b**) Young's modulus computed by molecular dynamics simulations as a function of the average atomic coordination number. The line is a linear fit. The coefficient of determination R^2 indicates the degree of linearity.

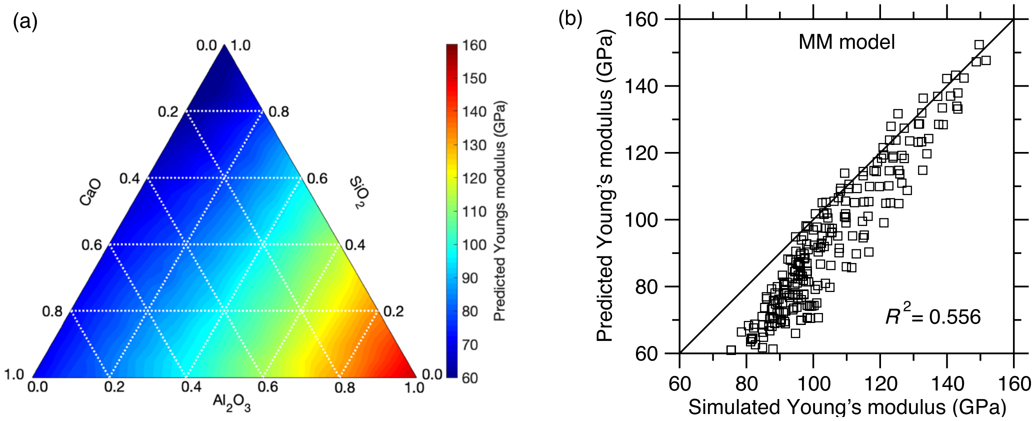

Figure 3. (**a**) Ternary diagram showing the Young's modulus values *E* predicted by the Makishima-Mackenzie (MM) model as a function of composition in the CaO–Al_2O_3–SiO_2 glass system. (**b**) Comparison between the Young's modulus values predicted by the MM model and computed by molecular dynamics simulations.

5- and 6-fold over-coordinated Al species in Al-rich glasses). Overall, we observe that the ternary plot of $<r>$ (Fig. 2a) echoes that of *E* (Fig. 1), which supports the fact that *E* increases upon increasing network connectivity. Nevertheless, as shown in Fig. 2b, we find that, although *E* and $<r>$ are indeed positively correlated with each other, the data points are widely spread and the coefficient of determination R^2 only equals 0.623. This indicates that the $<r>$ metric alone does not contain enough information to predict *E* and that other effects are not captured by simply considering the connectivity of the network—which renders challenging the development of a robust physics-based predictive model.

We now assess the ability of the popular Makishima–Mackenzie (MM) model to predict the compositional evolution of *E*[11]. The MM model relies on an additive relationship, wherein *E* is expressed as a weighted average of the dissociation energies of each oxide constituent. In details, the Young's modulus *E* is expressed as:

$$E = 83.6V_t \sum_{i=1}^{n} X_i G_i \tag{1}$$

where V_t is the overall packing density of the glass, and X_i and G_i are the concentration and volumic dissociation energy of each oxide constituent *i*, respectively. Note that the G_i terms are tabulated values, whereas V_t depends on the glass composition and is an explicit input to the model (i.e., the knowledge of the compositional dependence of V_t is a prerequisite to the MM model). To this end, we compute the packing density V_t of each glass based on the MD simulations. Figure 3a shows the ternary diagram of the *E* values predicted by the MM model as a function of composition in the CAS glass system. We observe that the MM model properly predicts the increase of *E* with increasing Al_2O_3 concentration, but fails to predict the increase in *E* upon increasing CaO concentration. This is due to fact that the dissociation energy terms associated with the CaO and SiO_2 oxides are close to each other (i.e., 15.5 and 15.4 kcal/cm^3, respectively), whereas that of Al_2O_3 (32 kcal/cm^3) is significantly higher. Overall, we observe that the MM model does not properly predict the non-linear dependence

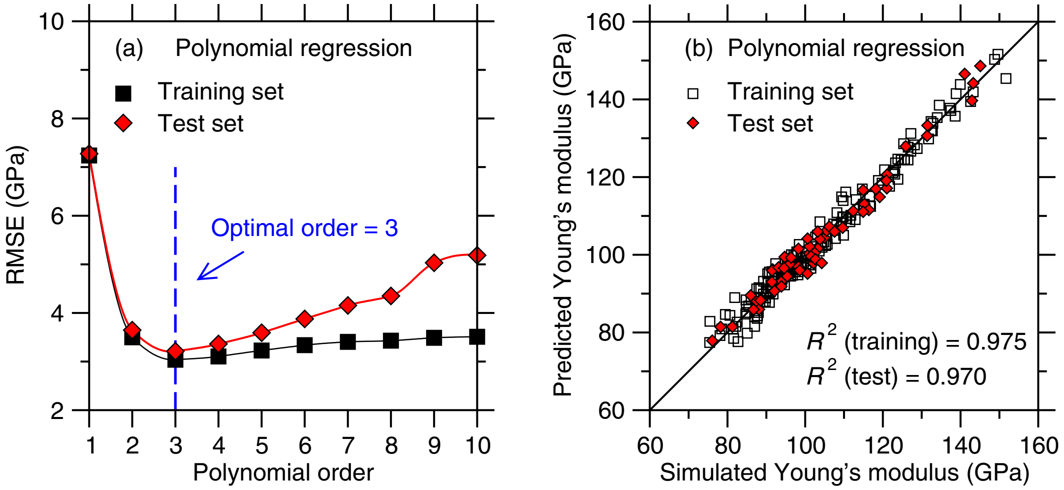

Figure 4. (**a**) Accuracy (as captured by the RMSE value) of the polynomial regression models as a function of the maximum polynomial degree considered in each model (see Sec. 2b)—as obtained for the training and test set, respectively. The optimal polynomial order is chosen as that for which the RMSE of the test set is minimum. (**b**) Comparison between the Young's modulus values predicted by polynomial regression (with a degree of 3) and computed by molecular dynamics simulations.

of *E* on composition. This is not surprising as the MM is essentially an additive model (although some level of non-linearity can exist within the V_t term). The MM model also fails to describe any coupling between the effects of CaO and Al_2O_3. Figure 3b shows a comparison between the Young's modulus values predicted by the MM model and computed by MD. Overall, we find that, although the MM model offers a fairly good prediction of *E*, the correlation remains poor (with $R^2 = 0.556$). In addition, we find that the MM model underestimates *E*, especially in the low *E* region (which corresponds to the technologically important low-Al compositional domain wherein glasses exhibit good glass-forming ability). Overall, we note that, although the MM model can be used as a rough guide to infer some compositional trends, it cannot be used to accurately predict *E* in CAS glasses.

Polynomial regression. The Young's modulus values computed by MD then serve as database to infer the relationship between glass composition (*x*, *y*) and *E* in the $(CaO)_x(Al_2O_3)_y(SiO_2)_{1-x-y}$ glass system by ML (see Methods section). In the following, we compare the performance of select ML algorithms. To this end, we adopt a nested cross-validation procedure, wherein a fraction (25%) of the data points is kept fully unknown to the models and is used as a "test set" to *a posteriori* assess the accuracy of each model, whereas the rest of the data (75%) is used as a "training set." The accuracy of the prediction is assessed by calculating the root-mean-square error (RMSE, see Methods section).

We first focus on the outcomes of polynomial regression (PR, see Supplementary Information). Figure 4a shows the RMSE offered by polynomial regression for the training and test sets as a function of the maximum polynomial degree considered in the model. As expected, we observe that the RMSE of the training set decreases upon increasing polynomial degree (i.e., increasing model complexity) and eventually plateaus. This signals that, as the model becomes more complex, it can better interpolate the training set. In contrast, we observe a significant increase in the RMSE when the polynomial degree is equal to 1 or 2—which indicates that, in this domain, the model is underfitted. This confirms again that linear models based on additive relationships are unable to properly describe the linkages between composition and Young's modulus. On the other hand, we observe that the RMSE of the test set initially decreases with increasing polynomial degree, shows a minimum for degree 3, and eventually increases with increasing degree. This demonstrates that the models incorporating some polynomial terms that are strictly larger than 3 are overfitted. This arises from the fact that, in the case of high degrees, the model starts to fit the noise of the training set rather than the "true" overall trend (see Supplementary Information). This exemplifies (i) how the training set allows identifying the minimum level of model complexity that is required to avoid underfitting and (ii) how the test set allows us to track the maximum level of model complexity before overfitting. Overall, the optimal polynomial degree (here found to be 3) manifests itself by a minimum in the RMSE of the test set.

We now focus on assessing the accuracy of the predictions of the best polynomial regression model (i.e., with a maximum polynomial degree of 3). Figure 4b shows a comparison between the Young's modulus predicted by the ML model and computed by MD. We find that the R^2 factors for the training and test sets are 0.975 and 0.970, respectively. This indicates that, even in the case of a simple algorithm like polynomial regression, ML offers a good prediction of *E* based on the simulated results.

LASSO. We now focus on the outcomes of the LASSO algorithm, which aims to reduce the complexity of the model by placing an extra weight on the model coefficients (see Methods section). Figure 5a shows the RMSE offered by LASSO for the training and test sets as a function of the degree of complexity, $-\log(\lambda)$, of the model. In contrast with the outcomes of the polynomial regression, we observe that LASSO does not yield any noticeable overfitting at high model complexity—which would manifest itself by an increase in the RMSE of the test set. This

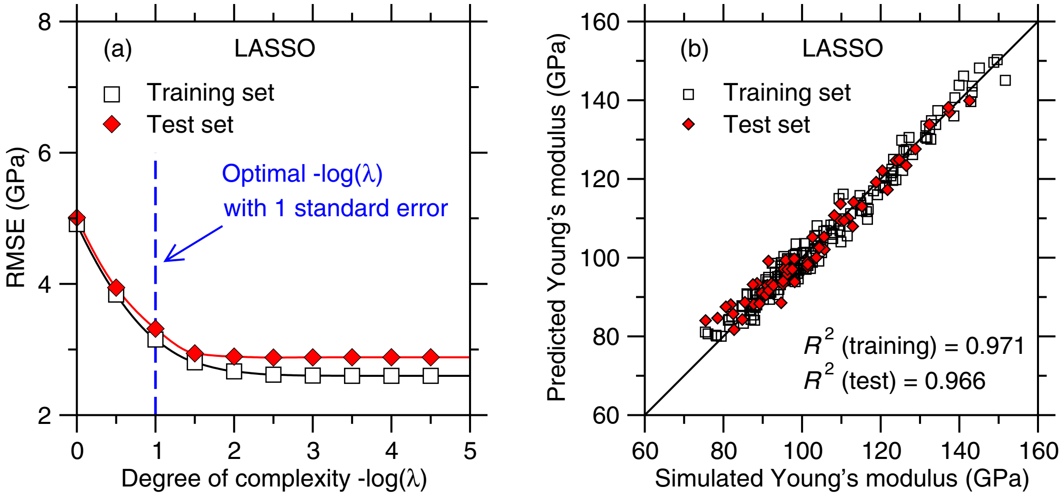

Figure 5. (**a**) Accuracy (as captured by the RMSE value) of the LASSO models as a function of the degree of complexity (see Methods section)—as obtained for the training and test set, respectively. The optimal degree of complexity is determined as the one for which the RMSE of the test set is one standard deviation away from the minimum RMSE (i.e., in the plateau regime). (**b**) Comparison between the Young's modulus values predicted by LASSO (with an optimal degree of complexity) and computed by molecular dynamics simulations.

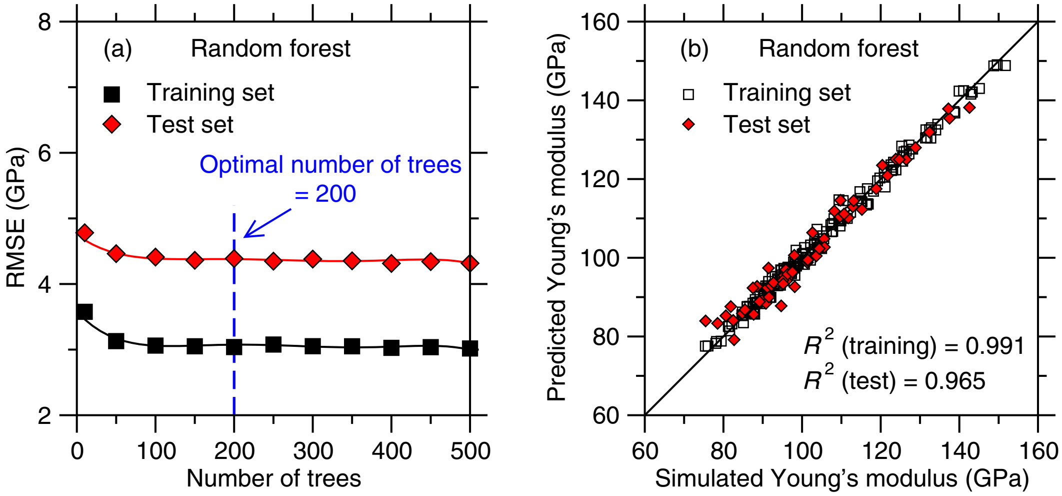

Figure 6. (**a**) Accuracy (as captured by the RMSE value) of the random forest models as a function of the number of trees considered in each model (see Sec. 2d)—as obtained for the training and test set, respectively. The optimal number of trees is taken as the threshold at which the RMSE of the test set starts to plateau. (**b**) Comparison between the Young's modulus values predicted by random forest (with 200 trees) and computed by molecular dynamics simulations.

can be understood from the fact that the LASSO algorithm specifically aims to reduce the number of polynomial terms to mitigate the risk of overfitting. Here, since the RMSE of the test set only shows a plateau with increasing $-\log(\lambda)$, we select the optimal degree of complexity as the one for which the RMSE of the test set becomes less than one standard deviation away from the minimum RMSE (i.e., in the plateau regime).

We now focus on assessing the accuracy of the predictions of the best LASSO model (i.e., with the optimal degree of complexity). Figure 5b shows a comparison between the Young's modulus predicted by the ML model and computed by MD. We find that the R^2 factors for the training and test sets are 0.971 and 0.966, respectively. As such, LASSO yields a slight decrease in accuracy as compared to polynomial regression.

Random forest. We now focus on the outcomes of the RF algorithm. Figure 6a shows the RMSE offered by RF for the training and test sets as a function of the number of trees (i.e., which characterizes the complexity of the model). As observed in the case of LASSO, we find that RF does not yield any noticeable overfitting at high model complexity, that is, the RMSE of the test set only plateaus upon increasing number of trees. Here, we select 200 as being the optimal number of trees.

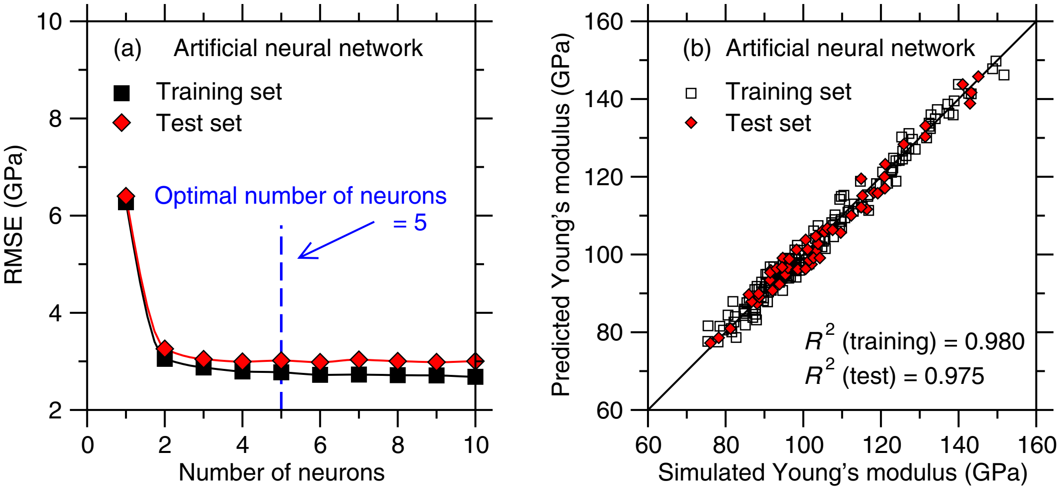

Figure 7. (**a**) Accuracy (as captured by the RMSE value) of the artificial neural network models as a function of the number of neurons considered in each model (see Methods section)—as obtained for the training and test set, respectively. The optimal number of neurons is determined as that for which the RMSE value of the test set is minimum. (**b**) Comparison between the Young's modulus values predicted by artificial neural network (with 5 neurons) and computed by molecular dynamics simulations.

We now focus on assessing the accuracy of the predictions of the best RF model (i.e., with 200 trees). Figure 6b shows a comparison between the Young's modulus predicted by the ML model and computed by MD. We find that the R^2 factors for the training and test sets are 0.991 and 0.965, respectively. This suggests that, although RF offers an excellent interpolation of the training set (i.e., with a higher R^2 value than those obtained with the other ML models), its ability to offer a good prediction of the test set is slightly lower than those of the other ML models considered herein.

Artificial neural network. Finally, we focus on the outcomes of the ANN algorithm. Herein, we adopt a multilayer perceptron (MLP) ANN, which is a class of feedforward neural network containing an input layer, a hidden layer, and an output layer. The MLP ANN model is trained using the back-propagation algorithm. We train ANN models with one hidden layer—which is found to be sufficient considering the nature of the training set. Figure 7a shows the RMSE offered by ANN for the training and test sets as a function of the number of neurons (i.e., which characterizes the complexity of the model). Overall, as previously observed in the cases of LASSO and RF, ANN does not yield any noticeable overfitting at high model complexity. Nevertheless, we note that the RMSE of the test set exhibits a slight minimum in the case of 5 neurons, which is the degree of complexity that we adopt herein.

We now focus on assessing the accuracy of the predictions of the best ANN model (with one hidden layer and five neurons). Figure 7b shows a comparison between the Young's modulus predicted by the ML model and computed by MD. We find that the R^2 factors for the training and test sets are 0.980 and 0.975, respectively. Overall, we find that the ANN algorithm offers the most accurate model among all the ML methods considered herein—as quantified in terms of the RMSE of the test set.

Discussion

We now compare the performance of the different machine learning algorithms used herein. We first focus on the level of accuracy offered by each method. To this end, Table 1 presents the coefficient of determination R^2 of each method for the training set (which characterizes the ability of the algorithm to properly interpolate the training data) and test set (which captures the accuracy of the model when predicting unknown data). We first observe that the RF algorithm offers the best interpolation on the training set (i.e., RF shows the highest R^2 for the training set). However, the RF algorithm also yields the lowest level of accuracy for the test set. This suggests that the RF algorithm presents the lowest ability to properly interpolate Young's modulus values in between two compositions of the training set and/or to offer realistic extrapolations toward the edges of the compositional domain. On the other hand, we note that the PR and LASSO algorithms offer a fairly similar level of accuracy, although the R^2 coefficient offered by PR for the test set is slightly higher than that offered by LASSO. This suggests that the slight decrease in model complexity offers by LASSO also results in a slight loss of accuracy (see Table 1). Finally, we observe that the artificial neural network algorithm clearly offers the highest level of accuracy among all the models considered herein since it yields the highest R^2 value for the test set.

To further characterize the accuracy offered by each ML algorithm, Fig. 8 shows the Young's modulus values that are predicted for two series of compositions, namely, (i) $(CaO)_x(Al_2O_3)_{40-x}(SiO_2)_{60}$, wherein the SiO_2 fraction is kept constant and equal to 60 mol% and (ii) $(CaO)_x(Al_2O_3)_x(SiO_2)_{100-2x}$, wherein the CaO/Al_2O_3 molar ratio is kept constant and equal to 1. These two series specifically aim to investigate (i) the effect of the degree of polymerization of the network (i.e., fraction of non-bridging oxygen) and (ii) the effect of network-forming atoms (i.e., Si vs. Al) at constant degree of depolymerization (i.e., in fully charge-compensated glasses). We first note that, in contrast to the other ML methods, RF yields piecewise-constant-shape results, which arises from the fact that

ML algorithms	Coefficient of determination R^2 Training set	Test set	Complexity	Interpretability
PR	0.975	0.970	Low (9)	High
LASSO	0.971	0.966	Low (8)	High
RF	0.991	0.965	High (200)	Intermediate
ANN	0.980	0.975	Intermediate (20)	Low

Table 1. Comparison between the levels of accuracy, complexity, and interpretability offered by the machine learning algorithms used herein, namely, polynomial regression (PR), LASSO, random forest (RF), artificial neural network (ANN). The level of accuracy is described by the coefficient of determination (R^2) for the training and test sets. The complexity is described in parenthesis by the number of non-zero parameters in PR and LASSO, the number of trees in RF, and the product of the number of inputs, neurons, and adjustable parameters per neuron in ANN. The "interpretability" describes the degree to which a human can understand the outcome produced by each model.

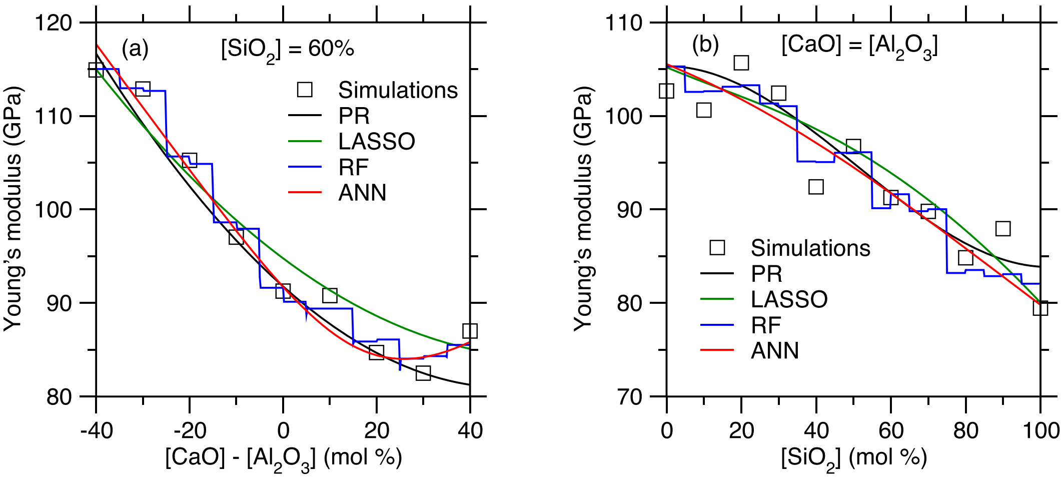

Figure 8. Comparison between the Young's modulus values computed by molecular dynamics simulations and predicted by the polynomial regression (PR), LASSO, random forest (RF), and artificial neural network (ANN) models for the series of compositions (**a**) $(CaO)_x(Al_2O_3)_{40-x}(SiO_2)_{60}$ and (**b**) $(CaO)_x(Al_2O_3)_x(SiO_2)_{100-2x}$.

the RF method is essentially based on an ensemble of decision trees. In details, the decision tree algorithm works by relying on a binary split, that is, at each node, randomly selected observations are dropped to either the left or right daughter node depending on the values and selected features. Although a single decision tree cannot capture any non-linearity within a dataset, the output of the model is eventually averaged over all its trees—so that an RF model can capture the non-linearity of a set of data by comprising enough trees. Nevertheless, we observe here that the piecewise-constant nature of single decision trees remains encoded in the outcome of this method, which yields non-smooth predictions. This feature of the RF algorithm likely explains its excellent ability to interpolate the training set while offering only a fair prediction of the test set.

We now compare the predictions offered by the PR and LASSO algorithms. Overall, although LASSO yields slightly lower R^2 values for both the training and test set as compared to PR, we note that LASSO offers an improved prediction of E at the edges of the training set (see Fig. 8a,b). For instance, we note that the PR method predicts an unrealistic slight increase in E in pure SiO_2 (see the right end of Fig. 8b). This non-monotonic evolution of E at the edges of the compositional domain suggests that the PR model might be slightly overfitted. In turn, such behaviour is mitigated by the LASSO algorithm. Finally, we find that ANN offers the best description of the non-linear nature of the data.

Besides accuracy, it is also desirable for ML-based models to be "simple" (i.e., low complexity) and "interpretable" (i.e., to avoid the use of "black box" models). Unfortunately, a higher level of accuracy often comes at the expense of higher complexity and lower interpretability. Simpler and more interpretable models are usually preferable as (i) simpler models are less likely to overfit small datasets, (ii) simpler models are usually more computationally-efficient, and (iii) more interpretable models are more likely to offer some new insights into the underlying physics governing the relationship between inputs and outputs.

We now discuss the level of complexity/interpretability of the different ML-based models developed herein. The degree of complexity of each of the trained models can be roughly captured by the number of non-zero parameters in PR and LASSO, the number of trees in RF, and the product of the number of inputs, neurons, and adjustable parameters per neuron in ANN (i.e., the number of weight coefficients and threshold terms to adjust). As presented in Table 1, we first note that RF offers a poor balance between accuracy and simplicity (as

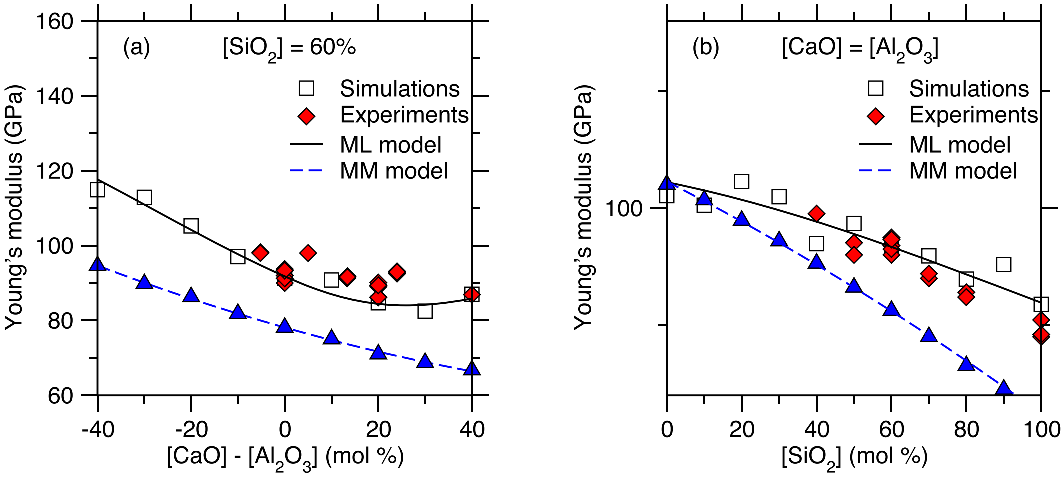

Figure 9. Comparison between the Young's modulus values computed by molecular dynamics simulations, predicted by the artificial neural network model, and predicted by the Makishima-Mackenzie (MM) model for the series of compositions (**a**) $(CaO)_x(Al_2O_3)_{40-x}(SiO_2)_{60}$ and (**b**) $(CaO)_x(Al_2O_3)_x(SiO_2)_{100-2x}$. The data are compared with select available experimental data[13,23–33].

the number of trees approaches the number of values in the training set). On the other hand, PR and LASSO clearly offer the lowest degree of complexity. The PR and LASSO algorithms also clearly yield the highest level of interpretability thanks to the analytical nature of the inputs/outputs relationship they offer. In details, we note that LASSO yields a slightly simpler analytical function—with only 8 non-zero terms, vs. 9 non-zero terms for PR. However, this slight decrease in complexity comes together with a slight decrease in accuracy. This shows that, by relying on a penalized regression method, LASSO allows us to slightly enhance the level of simplicity of the model. Finally, we note that the increased level of accuracy offered by ANN comes at the cost of higher complexity and lower interpretability, which is a common tradeoff in ML techniques.

We now compare the predictions of the most accurate ML-based model developed herein (i.e., ANN) with the simulated data (i.e., used during the training of the model) and available experimental data (see Fig. 9)[13,23–33]. We first note that the experimental data present a higher level of noise as compared to the simulation data. In the present case, these results illustrate the advantage of training ML models based on simulations rather than experimental data. Overall, we observe a good agreement between simulated data, ANN predictions, and experimental data. In contrast, we note that, as mentioned in the Results section, the MM model systematically underestimates E and does not properly capture the non-linear nature of the simulated data. Combining the results in Fig. 8 and Fig. 9, we also note that even simple algorithms, e.g., polynomial regression, can capture the non-linearity between composition and Young's modulus and yield some realistic predictions of the Young's modulus values. All the models offer a prediction that is significantly more accurate than that of the MM model. Overall, we find that the ANN model properly captures the non-linear compositional trend of *E* while filtering out the intrinsic noise of the simulation data. These results strongly support the ability of our MD + ML combined method to offer a robust prediction of the stiffness of silicate glasses.

Finally, we discuss the advantages of combining ML with high-through MD simulations—rather than directly training ML-based on available experimental data. First, we note that, although the CAS ternary system may be one of the most studied systems in glass science and engineering, the number of available experimental stiffness data available for this system is fairly limited. Further, most of the data available for this system are clustered in some small regions of the whole compositional domain (namely, pure silica, per-alkaline aluminosilicates, and calcium aluminates glasses) (see Fig. S1a in Supplementary Information). Such clustering of the data is a serious issue as, in turn, available experimental data come with a notable uncertainty—for instance, the Young's modulus of select glasses (at fixed composition) can vary by as much as 20 GPa among different references[21,32]. As such, the combination of a high level of noise and clustering of the data would not allow ML to discriminate the "true" trend of the data from the noise (see Fig. S1b in Supplementary Information). Finally, we note that conducting MD simulations is obviously faster/cheaper than synthesizing glass samples and measuring their stiffness. In turn, the results presented herein demonstrate that properly conducted MD simulations can offer a quantitative agreement with experimental data and, thereby, offer a desirable alternative to systematic experiments. However, it should be noted that the ML models developed herein necessarily reflect the nature of the data offered by the high-throughput simulations, with all their limitations. As such, the combined approach described herein critically relies on the availability of accurate interatomic forcefield to ensure the reliability of the MD simulations.

Conclusions

Overall, these results demonstrate that the combination of high-throughput molecular dynamics simulations and machine learning offers a robust approach to predict the elastic properties of silicate glasses. Further, our method clearly identifies the optimal level of complexity of each ML-based model, that is, to mitigate the risk of under- or overfitting. Based on these results, we find that the artificial neural network algorithm offers the highest level of accuracy. In contrast, the LASSO algorithm offers a model that features higher simplicity and interpretability—at the expense of a slight decrease in accuracy. The method presented herein is generic and transferable to new properties (e.g., other stiffness metrics) and new systems (e.g., other families of silicate glasses).

Methods

High-throughput molecular dynamics simulations. To establish our conclusions, we use molecular dynamics simulations to create a database consisting of the Young's modulus values of 231 glasses homogeneously covering the CAS ternary system, with 5% increments in the mol% concentration of the CaO, Al_2O_3, and SiO_2 oxide constituents. At this point, no consideration is made as to whether all these compositions would experimentally exhibit satisfactory glass-forming ability. All the simulations are conducted using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) package[34]. Each glass comprises around 3000 atoms. We adopt here the interatomic potential parametrized by Jakse—as it has been found to yield some structural and elastic properties that are in good agreement with experimental data for CAS glasses[35,36]. A cutoff of 8.0 Å is used for the short-range interactions. The Coulombic interactions are calculated by adopting the Fennell damped shifted force model with a damping parameter of 0.25 $Å^{-1}$ and a global cutoff of 8.0 Å[37]. The integration timestep is kept fixed 1.0 fs.

The glass samples are prepared with the conventional melt-quench method as described in the following[38]. First, some atoms are randomly placed in a cubic box using PACKMOL while using a distance cutoff of 2.0 Å between each atom to avoid any unrealistic overlap[39]. These initial configurations are then subjected to an energy minimization, followed by some 100 ps relaxations in the canonical (*NVT*) and isothermal-isobaric (*NPT*) ensembles at 300 K, sequentially. These samples are then fully melted at 3000 K for 100 ps in the *NVT* and, subsequently, *NPT* ensemble (at zero pressure) to ensure the loss of the memory of the initial configurations and to equilibrate the systems. Next, these liquids are cooled from 3000 to 300 K in the *NPT* ensemble at zero pressure with a cooling rate of 1 K/ps. The obtained glass samples are further relaxed at 300 K for 100 ps in the *NPT* ensemble before the stiffness computation. Note that this quenching procedure was slightly adjusted for select compositions. First, a higher initial melting temperature of 5000 K is used for the samples wherein the SiO_2 concentration is larger or equal to 95 mol%—since these glasses exhibit high glass transition temperatures. Second, a faster cooling rate of 100 K/ps is used for the samples wherein the CaO concentration is larger or equal to 90 mol%. Indeed, although the cooling rate can affect the glass stiffness, the use of a higher cooling rate here is necessary as these systems would otherwise tend to crystallize with a cooling rate of 1 K/ps.

The stiffness tensor $C_{\alpha\beta}$ of the equilibrated glasses is then computed by performing a series of 6 deformations (i.e., 3 axial and 3 shear deformations along the 3 axes) and computing the curvature of the potential energy[35,40]:

$$C_{\alpha\beta} = \frac{1}{V}\frac{\partial^2 U}{\partial e_\alpha \partial e_\beta} \tag{2}$$

where V is the volume of the glass, U is the potential energy, e is the strain, and α and β are some indexes representing each Cartesian direction. Note that all of the glass samples are found to be largely isotropic—so that the Young's modulus (E) can be calculated as:

$$E^{-1} = (S_{11} + S_{22} + S_{33})/3 \tag{3}$$

where $S = C^{-1}$ is the compliance matrix[15]. Based on previous results[24], the Jakse forcefield is found to systematically overestimate the Young's modulus of CAS glasses by about 16%—which may be a spurious effect arising from the fast cooling rate used in MD simulations or the parameterization of the forcefield. As such, the computed Young's modulus values are rescaled by this constant factor before serving as a training set for the machine learning models presented in the following.

Machine learning methodology. The 231 Young's modulus values computed by the high-throughput MD simulations serve as a database to infer the relationship between glass composition (x, y) and E in the $(CaO)_x(Al_2O_3)_y(SiO_2)_{1-x-y}$ glass system by ML. In details, we consider x and y to be the only inputs of the model (i.e., we neglect herein the effect of the thermal history of the glasses), whereas E is used as an output. Note that a similar approach can be used to predict the effect of composition on the shear modulus G, bulk modulus K, or Poisson's ratio ν. In the following, we briefly describe our overall ML strategy as well as the different ML algorithms that are considered and compared herein.

To avoid any risk of overfitting, a fraction of the data points is kept fully unknown to the models and is used as a "test set" to *a posteriori* assess the accuracy of each model. To this end, we adopt here the k-fold cross-validation (CV) technique[41]. The CV technique consists in splitting the dataset into k smaller sets, wherein the model is trained on "$k - 1$" of the folds and tested on the remaining of the data. The results are then averaged by iteratively using each of the k folds. Here, we adopt a nested two-level CV approach[42]. In detail, we first perform a 4-fold outer CV to split the dataset into the training set (75% of data) and test set (25% of data) and use the average value of the obtained scores (i.e., R^2) to compare the performance offered by different ML algorithms. Next, to obtain a proper setting for the hyperparameters of each model, we apply 10-fold inner CV within the training set. The nested CV technique allows us to avoid any arbitrary choice of test set and to partially mitigate issues arising from the limited number of data points.

For optimal predictions, ML models must achieve the best balance between accuracy and simplicity—wherein models that are too simple are usually poorly accurate (i.e., "underfitted"), whereas models that are too complex present the risk of placing too much weight on the noise of the training set and, thereby, often show poor transferability to unknown sets of data (i.e., "overfitted"). Hence, one needs to identify the optimal degree of complexity (e.g., number of terms, number of neurons, etc.) for each model. Here, we optimize the degree of complexity of each model by gradually increasing its complexity and tracking the accuracy of the model prediction for both the training and test sets. Indeed, although the accuracy of the training set prediction typically monotonically increases with increasing model complexity, overfitted models usually manifest themselves by a decrease in the accuracy of the test set prediction.

We herein adopt the polynomial regression (PR), LASSO, random forest (RF) and multilayer perceptron artificial neural network (ANN) algorithms to generate the predictive models (see Supplementary Information). This choice is motivated by the fact that these methods exhibit varying degrees of complexity and interpretability. This allows us to assess the nature of the trade-off between accuracy, simplicity, and interpretability offered by these algorithms.

Accuracy of the models. We assess the accuracy of each model (with different degrees of complexity) by computing the RMSE (root-mean-square error) and R^2 (coefficient of determination) factors. The RMSE factor measures the average Euclidian distance between the predicted and real values as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i^{'})^2} \tag{4}$$

where Y_i and $Y_i^{'}$ are the predicted and real output values, respectively. The RMSE has the property of being in the same units as the output variables and, hence, can be used to estimate the accuracy of the Young's modulus values predicted by each model (namely, lower RMSE values indicate higher accuracy). Here, we use the RMSE of the training and test sets to determine the optimal degree of complexity for each ML model.

In complement of RMSE, we compute the R^2 factor, which is the percentage of the response variable variation. This factor can be used to quantify how close the data are to the fitted line. $R^2 = 1$ indicates a perfect prediction, while smaller values indicate less accurate predictions. Here, we use the R^2 factor to compare the performances of each ML algorithm (once the degree of complexity has been optimized based on the RMSE).

Data Availability

All the data will be provided upon reasonable request.

References

1. Wondraczek, L. *et al*. Towards Ultrastrong Glasses. *Adv. Mater.* **23**, 4578–4586 (2011).
2. Rouxel, T. Designing glasses to meet specific mechanical properties. In *Challenging Glass: Conference on Architectural and Structural Applications of Glass, Faculty of Architecture, Delft University of Technology, May 2008* 39 (IOS Press, 2008).
3. Rouxel, T. Elastic properties of glasses: a multiscale approach. *Comptes Rendus Mes of gl* **334**, 743–753 (2006).
4. Rouxel, T. Elastic Properties and Short-to Medium-Range Order in Glasses. *J. Am. Ceram. Soc.* **90**, 3019–3039 (2007).
5. Mauro, J. C., Philip, C. S., Vaughn, D. J. & Pambianchi, M. S. Glass Science in the United States: Current Status and Future Directions. *Int. J. Appl. Glass Sci.* **5**, 2–15 (2014).
6. Mauro, J. C. & Zanotto, E. D. Two Centuries of Glass Research: Historical Trends, Current Status, and Grand Challenges for the Future. *Int. J. Appl. Glass Sci.* **5**, 313–327 (2014).
7. Zanotto, E. D. & Coutinho, F. A. B. How many non-crystalline solids can be made from all the elements of the periodic table? *J. Non-Cryst. Solids* **347**, 285–288 (2004).
8. Varshneya, A. K. *Fundamentals of Inorganic Glasses*. (Academic Press Inc, 1993).
9. Mauro, J. C. Decoding the glass genome. *Curr. Opin. Solid State Mater. Sci.* **22**, 58–64 (2018).
10. Liu, H., Du, T., Krishnan, N. M. A., Li, H. & Bauchy, M. Topological optimization of cementitious binders: Advances and challenges. *Cem. Concr. Compos*, https://doi.org/10.1016/j.cemconcomp.2018.08.002 (2018).
11. Makishima, A. & Mackenzie, J. D. Direct calculation of Young's moidulus of glass. *J. Non-Cryst. Solids* **12**, 35–45 (1973).
12. Makishima, A. & Mackenzie, J. D. Calculation of bulk modulus, shear modulus and Poisson's ratio of glass. *J. Non-Cryst. Solids* **17**, 147–157 (1975).
13. Eagan, R. J. & Swearekgen, J. C. Effect of Composition on the Mechanical Properties of Aluminosilicate and Borosilicate Glasses. *J. Am. Ceram. Soc.* **61**, 27–30 (1978).
14. Du, J. Challenges in Molecular Dynamics Simulations of Multicomponent Oxide Glasses. In *Molecular Dynamics Simulations of Disordered Materials: From Network Glasses to Phase-Change Memory Alloys* (eds Massobrio, C., Du, J., Bernasconi, M. & Salmon, P. S.) 157–180, https://doi.org/10.1007/978-3-319-15675-0_7 (Springer International Publishing 2015).
15. Pedone, A., Malavasi, G., Cormack, A. N., Segre, U. & Menziani, M. C. Insight into Elastic Properties of Binary Alkali Silicate Glasses; Prediction and Interpretation through Atomistic Simulation Techniques. *Chem. Mater.* **19**, 3144–3154 (2007).
16. Anoop Krishnan, N. M. *et al*. Predicting the dissolution kinetics of silicate glasses using machine learning. *J. Non-Cryst. Solids* **487**, 37–45 (2018).
17. Dreyfus, C. & Dreyfus, G. A machine learning approach to the estimation of the liquidus temperature of glass-forming oxide blends. *J. Non-Cryst. Solids* **318**, 63–78 (2003).
18. Mauro, J. C., Tandia, A., Vargheese, K. D., Mauro, Y. Z. & Smedskjaer, M. M. Accelerating the Design of Functional Glasses through Modeling. *Chem. Mater.* **28**, 4267–4277 (2016).
19. Onba. M, M. C., Tandia, A. & Mauro, J. C. Mechanical and Compositional Design of High-Strength Corning Gorilla® Glass. *Handb. Mater. Model.* 1–23, https://doi.org/10.1007/978-3-319-50257-1_100-1 (2018).
20. Cassar, D. R., de Carvalho, A. C. P. L. F. & Zanotto, E. D. Predicting glass transition temperatures using neural networks. *Acta Mater.* **159**, 249–256 (2018).
21. Priven, A. I. & Mazurin, O. V. Glass Property Databases: Their History, Present State, and Prospects for Further Development. *Adv. Mater. Res.* **39–40**, 145–150 (2008).
22. Ellison, A. & Cornejo, I. A. Glass Substrates for Liquid Crystal Displays. *Int. J. Appl. Glass Sci.* **1**, 87–103 (2010).
23. Ecolivet, C. & Verdier, P. Proprietes elastiques et indices de refraction de verres azotes. *Mater. Res. Bull.* **19**, 227–231 (1984).
24. Inaba, S., Todaka, S., Ohta, Y. & Morinaga, K. Equation for Estimating the Young&rsquo;s Modulus, Shear Modulus and Vickers Hardness of Aluminosilicate Glasses. *J. Jpn. Inst. Met.* **64**, 177–183 (2000).
25. Inaba, S., Oda, S. & Morinaga, K. Equation for Estimating the Thermal Diffusivity, Specific Heat and Thermal Conductivity of Oxide Glasses. *J. Jpn. Inst. Met.* **65**, 680–687 (2001).
26. Weigel, C. *et al*. Elastic moduli of XAlSiO4 aluminosilicate glasses: effects of charge-balancing cations. *J. Non-Cryst. Solids* **447**, 267–272 (2016).

27. Rocherulle, J., Ecolivet, C., Poulain, M., Verdier, P. & Laurent, Y. Elastic moduli of oxynitride glasses: Extension of Makishima and Mackenzie's theory. *J. Non-Cryst. Solids* **108**, 187–193 (1989).
28. Yamane, M. & Okuyama, M. Coordination number of aluminum ions in alkali-free alumino-silicate glasses. *J. Non-Cryst. Solids* **52**, 217–226 (1982).
29. Sugimura, S., Inaba, S., Abe, H. & Morinaga, K. Compositional Dependence of Mechanical Properties in Aluminosilicate, Borate and Phosphate Glasses. *J. Ceram. Soc. Jpn.* **110**, 1103–1106 (2002).
30. Gross, T. M., Tomozawa, M. & Koike, A. A glass with high crack initiation load: Role of fictive temperature-independent mechanical properties. *J. Non-Cryst. Solids* **355**, 563–568 (2009).
31. Yasui, I. & Utsuno, F. Material Design of Glasses Based on Database – INTERGLAD. In *Computer Aided Innovation of New Materials II* (eds Doyama, M., Kihara, J., Tanaka, M. & Yamamoto, R.) 1539–1544, https://doi.org/10.1016/B978-0-444-89778-7.50147-X (Elsevier 1993).
32. Hwa, L.-G., Hsieh, K.-J. & Liu, L.-C. Elastic moduli of low-silica calcium alumino-silicate glasses. *Mater. Chem. Phys.* **78**, 105–110 (2003).
33. Bansal, N. P. & Doremus, R. H. *Handbook of Glass Properties*. (Elsevier, 2013).
34. Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **117**, 1–19 (1995).
35. Bauchy, M. Structural, vibrational, and elastic properties of a calcium aluminosilicate glass from molecular dynamics simulations: The role of the potential. *J. Chem. Phys.* **141**, 024507 (2014).
36. Bouhadja, M., Jakse, N. & Pasturel, A. Structural and dynamic properties of calcium aluminosilicate melts: a molecular dynamics study. *J. Chem. Phys.* **138**, 224510 (2013).
37. Fennell, C. J. & Gezelter, J. D. Is the Ewald summation still necessary? Pairwise alternatives to the accepted standard for long-range electrostatics. *J. Chem. Phys.* **124**, 234104 (2006).
38. Li, X. *et al*. Cooling rate effects in sodium silicate glasses: Bridging the gap between molecular dynamics simulations and experiments. *J. Chem. Phys.* **147**, 074501 (2017).
39. Martínez, L., Andrade, R., Birgin, E. G. & Martínez, J. M. PACKMOL: a package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **30**, 2157–2164 (2009).
40. Liu, H. *et al*. Effects of polydispersity and disorder on the mechanical properties of hydrated silicate gels. *J. Mech. Phys. Solids* **122**, 555–565 (2019).
41. Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. *J. R. Stat. Soc. Ser. B Methodol.* **36**, 111–147 (1974).
42. Cawley, G. C. & Talbot, N. L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J Mach Learn Res* **11**, 2079–2107 (2010).

Acknowledgements

This work was supported by the National Science Foundation under Grants No. 1562066, 1762292, and 1826420. M.M.S. acknowledges support from the Independent Research Fund Denmark under Grant No. 8105-00002.

Author Contributions

M.B. conceived this study. B.C., H.R. and K.Y. performed the molecular dynamics simulations. X.X. and B.Y. conducted the machine learning analysis. X.X., B.Y., B.C. and K.Y. analyzed the results. K.Y. wrote the main manuscript and prepared the figures with the help of X.X. and B.Y. N.M.A.K., M.M.S., C.H. and M.B. provided suggestions, supervised the research, and discussed the results. All the authors reviewed and refined the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-45344-3.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.