Predicting the dissolution kinetics of silicate glasses by topology-informed machine learning

Liu, Han; Zhang, Tony; Krishnan, N. M. Anoop; Smedskjær, Morten Mattrup; Ryan, Joseph V.; Gin, Stephane; Bauchy, Mathieu

Published in:
npj Materials Degradation

DOI (link to publication from Publisher):
10.1038/s41529-019-0094-1

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Liu, H., Zhang, T., Krishnan, N. M. A., Smedskjær, M. M., Ryan, J. V., Gin, S., & Bauchy, M. (2019). Predicting the dissolution kinetics of silicate glasses by topology-informed machine learning. *npj Materials Degradation*, *3*, [32]. https://doi.org/10.1038/s41529-019-0094-1

npj | Materials Degradation

ARTICLE OPEN

Predicting the dissolution kinetics of silicate glasses by topology-informed machine learning

Han Liu[1], Tony Zhang[1], N. M. Anoop Krishnan[1,2,3], Morten M. Smedskjaer[4], Joseph V. Ryan[5], Stéṕhane Gin[6] and Mathieu Bauchy[1]

Machine learning (ML) regression methods are promising tools to develop models predicting the properties of materials by learning from existing databases. However, although ML models are usually good at interpolating data, they often do not offer reliable extrapolations and can violate the laws of physics. Here, to address the limitations of traditional ML, we introduce a "topology-informed ML" paradigm—wherein some features of the network topology (rather than traditional descriptors) are used as fingerprint for ML models—and apply this method to predict the forward (stage I) dissolution rate of a series of silicate glasses. We demonstrate that relying on a topological description of the atomic network (i) increases the accuracy of the predictions, (ii) enhances the simplicity and interpretability of the predictive models, (iii) reduces the need for large training sets, and (iv) improves the ability of the models to extrapolate predictions far from their training sets. As such, topology-informed ML can overcome the limitations facing traditional ML (e.g., accuracy vs. simplicity tradeoff) and offers a promising route to predict the properties of materials in a robust fashion.

npj Materials Degradation (2019)3:32 ; https://doi.org/10.1038/s41529-019-0094-1

INTRODUCTION

Machine learning (ML)—a subfield of artificial intelligence—offers a promising route to predict the properties of silicate glasses as a function of their composition.[1–7] Indeed, by "learning" from existing data set, ML algorithm can infer some complex patterns within the data that would otherwise remain hidden to human eyes.[8–10] As such, ML has previously been used with great success to predict the compositional dependence of the liquidus temperature,[1] solubility,[2] glass transition temperature,[3] stiffness,[4] and dissolution kinetics[5] of oxide glasses.

However, data-driven models present several limitations and challenges. (i) The use of ML requires the existence of large, accurate, and consistent data sets (wherein a consistent data set should comprise data that are measured by the same operation, including the same equipment, operator, protocol, data processing scheme, and environmental conditions), which are not always available.[8,11] (ii) Data-driven models are usually good at "interpolating" data, but typically fail to "extrapolate" data far from the training set.[5,10,12] This is a serious issue as it implies that ML cannot reliably be used to investigate presently unexplored compositional domains that are not explicitly considered during the training phase. This limits the potential of ML for the discovery of novel glasses with significantly improved properties. (iii) Data-driven models do not embed any mechanistic knowledge and, as such, can violate physical laws.[8,12] (iv) Finally, ML-based models are usually complex and hardly interpretable (i.e., they act as "black boxes"). Hence, they usually do not offer any new physical insights.[3,5,8] These issues are challenging to mitigate within traditional ML frameworks—wherein traditional descriptors (e.g., glass composition, interatomic bond energy, etc.) ignore underlying physical and chemical mechanisms and may not properly exhibit a simple and direct relationship with the predicted properties. More generally, when the linkages between the descriptors and the mechanism governing the target property of interest is unclear, the causality of the learned descriptor–property relation is uncertain.[13]

Here, to address the challenges facing traditional "*blind machine learning*" (i.e., which does not embed any topological information), we introduce a "*topology-informed machine learning*" paradigm—wherein some features of the network topology are used as descriptors—and apply it to predict the stage I dissolution kinetics (i.e., forward rate, far from saturation) of sodium aluminosilicate glasses.[14–16] Indeed, no universal physics-based model is presently available to predict the dissolution kinetics of silicate glasses (even in stage I). This arises from (i) a lack of knowledge regarding the rate-controlling mechanism of dissolution,[14,17–19] (ii) the large number of potential intrinsic (e.g., glass composition) and extrinsic (e.g., temperature, pH, etc.) parameters,[5,14,20] and (iii) an incomplete knowledge of the complex, disordered structure of silicate glasses.[21–25] In the present contribution, we show that, by embedding some degree of physics and chemistry, our approach yields a predictive model that is simple (linear), accurate, and transferable to untrained glass compositions.

RESULTS

Nature of the data set

To establish our conclusions, we rely on the database developed by Hamilton et al.,[24,26–28] which comprises the forward dissolution rate (DR) of a series of aluminosilicate glasses with varying

[1]Physics of AmoRphous and Inorganic Solids Laboratory (PARISlab), Department of Civil and Environmental Engineering, 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, Aalborg, Denmark; [5]Energy and Environment Directorate, Pacific Northwest National Laboratory, Richland, WA 99352, USA and [6]CEA Marcoule, DE2D SEVT, F-30207 Bagnols-sur-Ceze, France
Correspondence: Mathieu Bauchy (bauchy@ucla.edu)

Received: 8 April 2019 Accepted: 7 August 2019
Published online: 29 August 2019

npj nature partner journals

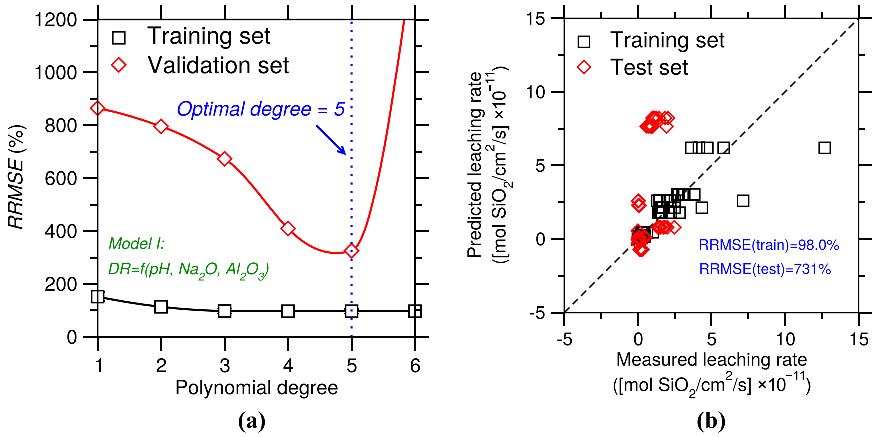


Fig. 1 Predictions from “blind” machine learning (“Model I”). **a** Evolution of the relative root square mean square error (RRMSE) of the training and validation sets with respect to the polynomial degree p. The minimum in the RRMSE of the validation set indicates that $p = 5$ is the optimal polynomial degree. **b** Predicted dissolution rate for $p = 5$ as a function of the measured dissolution rate

compositions under varying pH conditions. In details, the database comprises dissolution data for two families of glasses, namely, (i) the “Glasses A” series $(Na_2O)_{25}(Al_2O_3)_y(SiO_2)_{75-y}$, with $y = 5\%$, 10%, 15%, 20%, and 25% and (ii) the “Glasses B” series $(Na_2O)_x(Al_2O_3)_x(SiO_2)_{100-2x}$, with $x = 12.5\%$, 16.7%, and 25%. As such, the glass compositions cover both the tectosilicate and peralkaline domains, with varying fractions of non-bridging oxygen (BO) atoms. The dissolution kinetics of these glasses is systematically studied in unsaturated aqueous solutions over a wide domain of pH, ranging from pH 1 to 12. The DR is here quantified in terms of the SiO_2-leaching rate (expressed in units of mol/cm^2/s). In total, the database comprises 200 data points.[26] More details can be found in the Methods section. Note that simple metrics (e.g., the fraction of non-BO atoms) do not offer any good correlation with the DR (see ref. [5]). In particular, all the glasses from the series B are fully charge-compensated and, hence, present a theoretical zero fraction of non-BO atoms and yet exhibit varying DRs. This justifies the use of more complex descriptors as presented in the following.

Blind ML

We first assess the ability of “blind ML”[8,10,12] (that is, which does not embed any physics/chemistry about the dissolution process) to offer realistic prediction of the dissolution kinetics of the aluminosilicate glasses considered herein. To this end, we first consider as inputs the glass composition (i.e., the molar fractions of Na_2O and Al_2O_3) and the solution pH, whereas the DR is used as output. We then adopt the polynomial regression technique to infer the relationship between inputs and output.[9,10] Indeed, although our previous work on the same DR data set has shown that more complex ML algorithms (e.g., artificial neural network) offer improved performance,[5] such complex algorithms do not yield any analytical, easily usable function relating the inputs and output of the model and are poorly interpretable. In contrast, the polynomial regression method eventually yields an analytical model expressing the DR as a polynomial function of the inputs:

$$\text{Model I}: \quad \text{DR} = f(\text{pH}, Na_2O, Al_2O_3) \tag{1}$$

In the following, we refer to this model as “Model I.” To avoid any overfitting, we divide the database into (i) a training set, which is used to train the model, (ii) a validation set (10% of the data points of the database generated by the cross-validation method[9,10]), which is used to validate the performance of the model and identify the optimal polynomial degree, and (iii) a test set, that is, some data that are kept fully invisible to the model and that are used to assess its ability to predict unknown data. The test is here chosen by randomly selecting 30% of the data points from the database. The accuracy of the prediction is assessed by calculating the relative-root-mean-square-error (RRMSE,[29] see Methods section). More details about the ML methodology can be found in the Methods section.

We first consider the evolution of RRMSE of the training and validation sets with respect to the maximum polynomial degree (p) of the model (see Fig. 1(a)). As expected, the RRMSE of the training set decreases monotonically with increasing polynomial degree and eventually plateaus. This arises from the fact that, as complexity increases, the model necessarily offers an improved interpolation of the training set. In contrast, the RRMSE of the validation set initially decreases upon increasing polynomial degree, shows a minimum at $p = 5$, and finally increases with increasing model complexity. This can be understood from the fact that, when $p < 5$, the polynomial model is too simple to properly interpolate the training set and to predict the validation set (i.e., underfitting). In turn, when $p > 5$, the model starts to fit the “noise” of the training set and fails to capture the “true” overall trend (i.e., overfitting). These results exemplify how the evolution of RRMSE vs. polynomial degree allows us to identify the optimal model complexity to avoid either underfitting or overfitting. Overall, the optimal polynomial degree (here found to be 5) manifests itself by a minimum in the RRMSE of the validation set.

Figure 1b shows the dissolution rate values predicted by this model with $p = 5$ for the training and test sets. Overall, we find that blind polynomial regression (Model I) does not accurately capture the relationship between glass composition, pH, and dissolution rate. The RRMSE of the training set is found to be very high (98%), which indicates that the model does not properly interpolate the data used during its training. In turn, the RRMSE of the test set (731%) highlights the fact that this model is largely unable to properly predict the dissolution rate of glasses/pH for which it has not explicitly been trained for. This likely arises from the fact that the relationship between inputs and output is here largely nonlinear and, hence, cannot be properly captured by a linear model—in agreement with our previous findings.[5] Note that, considering the low performance of the present model, no effort is here made to understand why the dissolution rates of certain glasses are well predicted, whereas others are not.

Strategy for topology-informed ML

Figure 2 illustrates the main idea of “topology-informed” ML and how it compares to traditional “blind” ML. By being blind to the nature of the mechanism governing the property of interest, traditional blind ML ignores (i) which descriptors govern the

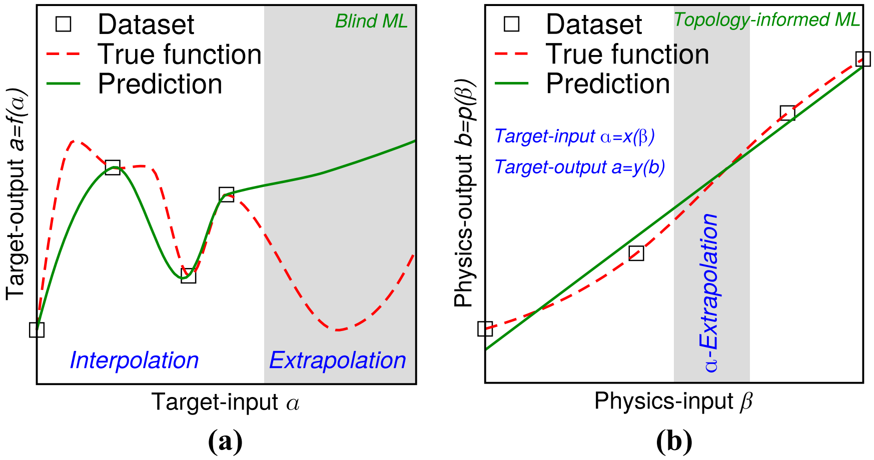


Fig. 2 Schematic illustrating the ability (or inability) to extrapolate predictions far from the training set of **a** traditional blind machine learning (trained based on arbitrary descriptors *α*) and **b** topology-informed machine learning (trained based on topological descriptors *β*). In both panels, the dashed red curve represents the true function relating the inputs to the targeted output. The squares indicate the known points from the training set. The solid green curve represents the "guessed" function interpolated by the ML model. The gray window indicates a range of systems (i.e., specific values of descriptors *α*) that is not represented within the training set and for the predictions from the ML models are tested. Note that this window is outside the training set in **a**, but not in **b**—since several systems with different descriptors *β* may present the same topology

output property and (ii) the analytical form of the input-output relationship. As illustrated in Fig. 2a, a poor choice of descriptors can result in a complex, highly nonlinear function. Although complex regression algorithms can properly interpolate such nonlinear data sets, they are unlikely to offer realistic predictions extrapolated far from the training set. In contrast, topology-informed ML models are expected to address these limitations by: (i) reducing the dimensionality of the problem (as several glasses with varying compositions can present the same topology and, hence, similar dissolution kinetics), (ii) simplifying the trained models (as the number of descriptors is decreased), and (iii) linearizing the relationship between inputs and output. As illustrated in Fig. 2b, relying on a topological fingerprint (rather than traditional descriptors) is expected to facilitate extrapolations far from the training set.

In detail, to address the intrinsic limitations of blind ML highlighted in Figs 1 and 2, we adopt the following strategy. (i) First, we focus on the polynomial regression method as more complex ML algorithms (e.g., artificial neural network or random forest[9,10]) offer poor interpretability.[8] Rather, the polynomial regression yields an analytical function, which, in turn, can serve to infer some of the underlying physics of the dissolution mechanism. (ii) Second, we attempt to "linearize" the relationship between inputs and output based on our physical understanding of the dissolution process. This is based on the idea that linear models are expected to be more likely to offer a good transferability to unknown inputs and to potentially yield some useful physical insights.[8,10,12] (iii) Third, we attempt to identify some relevant reduced-dimensionality descriptors capturing the effect of the atomic structure of the glass on dissolution rate that can be used as inputs. This is based on the idea that, although the dissolution kinetics of glasses is controlled by their composition (at fixed thermal history) for a given set of environment conditions (*T*, pH, and solution composition[30–33]), the knowledge of the structure of the atomic network makes it possible to decipher the relationship between composition and dissolution rate—so that it should be easier for ML algorithms to infer the relationship between "structure and dissolution rate" than between "composition and dissolution rate." In the following, we present how these topology-informed ingredients allow us to derive less complex, yet more accurate predictive models.

Linearization of the inputs/output relationship

In an attempt to linearize the relationship between the inputs and output of the model, we first note that, in general, the dissolution rate is an exponential (rather than linear) function of pH and composition. This can be illustrated from the fact that, based on transition state theory, the Aagaard-Helgeson model expresses the forward dissolution rate in terms of (i) the activity of H^+ ions, which, in turn, is an exponential function of pH,[34] and (ii) an Arrhenius term $\exp(-E_a/RT)$, wherein the activation energy has recently be suggested to be a function of the number of topological constraints per atom in the network, which, in turn, is often a linear function of composition.[21,30,31] Based on this fact, it follows that one can increase the degree of linearity of the relationship between inputs and output by predicting the logarithm of the dissolution rather than the dissolution rate itself (referred to as "Model II" thereafter):

$$\text{Model II}: \quad \log(\text{DR}) = f(\text{pH}, \text{Na}_2\text{O}, \text{Al}_2\text{O}_3) \tag{2}$$

We find that, by using Model II, the prediction accuracy is significantly improved when the polynomial degree *p* decreases to 3 (see Supplementary Information). To further enhance the degree of linearity of the inputs/output relationship, we now consider the dependence of the dissolution on pH. As illustrated in Fig. 3, the dissolution rate exhibits a fairly bilinear V-shape dependence on pH, with a minimum in neutral condition (pH 7).[30,32] This is an issue as the description of a bilinear function in terms of a sum of polynomials requires the use of high degrees to account for the break in slope. As an alternative route, we define two new input variables, namely, pH_{acid} and pH_{base}, which are defined as $\text{pH}_{\text{acid}} = \max(0; 7\text{–pH})$ and $\text{pH}_{\text{base}} = \max(0; \text{pH–}7)$. Note that these inputs contain the same information of the pH variable but allow us to describe the linear evolution of the dissolution rate with respect to pH_{acid} and pH_{base} for $\text{pH} < 7$ and $\text{pH} > 7$, respectively, rather than the bilinear evolution of the dissolution with respect to pH (see Fig. 3). Note that the variables pH_{acid} and pH_{basic} are equal to 0 for $\text{pH} > 7$ and $\text{pH} < 7$, respectively, so that only one of these variables at a time is non-zero. Model III expresses the logarithm of the dissolution rate in terms of the glass composition and these two new variables:

$$\text{Model III}: \quad \log(\text{DR}) = f(\text{pH}_{\text{acid}}, \text{pH}_{\text{base}}, \text{Na}_2\text{O}, \text{Al}_2\text{O}_3) \tag{3}$$

Figure 4a shows the RRMSE of the training and validation sets as a function of the maximum polynomial degree p for Model III. Importantly, we find that the explicit description of the bilinear dependence of the dissolution rate on pH allows us to further reduce the complexity of the model since the RRMSE of the validation set shows a minimum for $p = 1$. This indicates that Model III can express the dissolution rate through a simple, linear relationship. In addition to decreasing the complexity of the model, Model III also offers an increased degree of accuracy since the RRMSE of the test set is found to be 3.76% (as compared with 731% for Model I, see Fig. 4b). These results illustrate how the linearization of the relationship between inputs and output based on our physical/chemical understanding of the dissolution process can results in the training of a less complex, yet more accurate model.

Topology-informed reduced-dimensionality descriptors

We now attempt to further increase the accuracy of the model by identifying a structural "fingerprint" of the structure of the atomic network—which is based on the idea that the structure of the atomic network of a glass has a first order effect on its dissolution kinetics. To this end, we adopt the framework of topological constraint theory (TCT), which describes complex disordered network as mechanical trusses, wherein some nodes (the atoms) are connected to each other by some topological constraints (the chemical bonds).[21,35–37] Based on this framework, the number of topological constraints per atom (n_c) has been shown to offer a useful reduced-dimensionality descriptor that captures the connectivity of the atomic network and, hence, can be used to predict various glass properties, e.g., hardness, stiffness, fracture toughness, glass transition temperature, fragility, etc.[21,38–43] Importantly, the effective activation energy of dissolution for a fixed pH has recently been suggested to be proportional to n_c.[31,33,44–50] Based on these findings, we compute the number of topological constraints of the rigid aluminosilicate network n_c for each glass (see Methods section) and use it as a descriptor of the atomic structure. As shown in Fig. 5, we observe that, at fixed pH, the dissolution rate is indeed largely correlated to n_c, which supports the use of this metric as an input to the model. We then define Model IV, which expresses the logarithm of the dissolution rate in terms of pH, n_c, and the fraction of network modifiers (i.e., Na_2O)—as the network modifiers are not explicitly accounted for in the number of topological constraints of the rigid aluminosilicate network (see Methods):[47]

$$\text{Model IV}: \quad \log(\mathrm{DR}) = f(\mathrm{pH}_{\mathrm{acid}}, \mathrm{pH}_{\mathrm{base}}, n_c, \mathrm{Na_2O}) \tag{4}$$

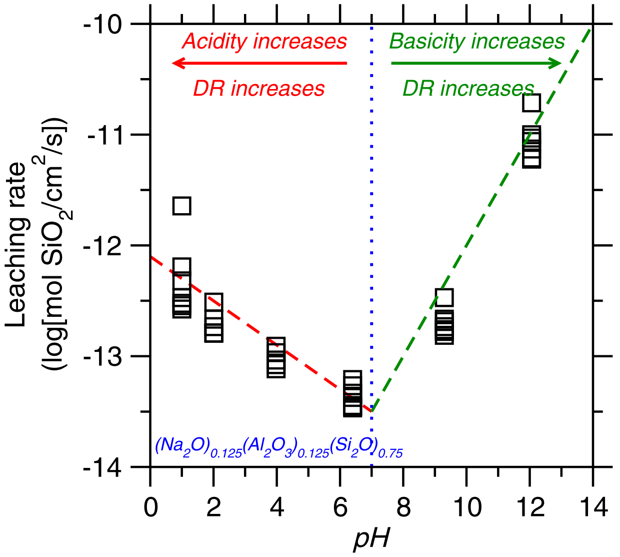
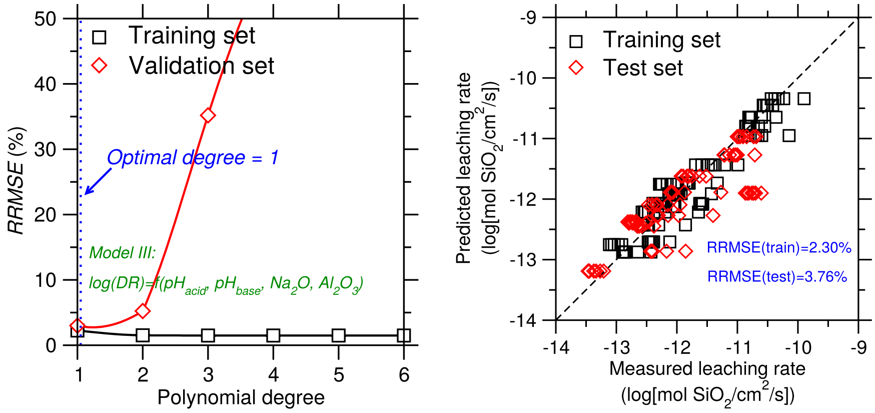


Fig. 3 Measured dissolution rate of a $(Na_2O)_{0.125}(Al_2O_3)_{0.125}(SiO_2)_{0.75}$ glass as a function of pH[26]

Figure 6a shows the RRMSE of the training and validation sets as a function of the maximum polynomial degree p for Model IV. Like Model III, we note that a linear model (i.e., $p = 1$) offers the best performance. As shown in Fig. 6b, Model IV is able to (i) properly interpolate the training set and (ii) predict realistic values for the test set. Nevertheless, we note that the overall degree of accuracy remains comparable to that offered by Model III. In particular, select points appear to systematically act as outliners in all the models considered herein and, hence, might be experimental artefacts.

Overcoming the tradeoff between accuracy and simplicity in ML

ML-based models usually exhibit a tradeoff between accuracy and simplicity.[8–10] Indeed, simple models (e.g., polynomial regression) are less complex but tend to exhibit limited accuracy, whereas more advanced models (e.g., random forest or artificial neural network) are often more accurate but, in turn, exhibit higher complexity and lower interpretability.[5,10,51] In general, simpler and



(a) (b)

Fig. 4 Predictions from machine learning while explicitly accounting for the exponential dependence of the dissolution rate on the inputs and capturing the distinct acidic and caustic regimes ("Model III"). **a** Evolution of the relative root square mean square error (RRMSE) of the training and validation sets with respect to the polynomial degree p. The minimum in the RRMSE of the validation set indicates $p = 1$ as an optimal polynomial degree (i.e., linear model). **b** Predicted dissolution rate for $p = 1$ as a function of the measured dissolution rate

more interpretable models are desirable. On the one hand, adopting a simple model reduces the risk of overfitting small data sets and is usually more computationally efficient. On the other hand, simpler models are more likely to properly capture the underlying physics governing the relationship between inputs and outputs. Figure 7 shows the complexity (captured by the optimal polynomial degree) and accuracy (captured by the RRMSE) of the different models considered herein. Overall, we find that embedding topological descriptors yields models that are less complex and more accurate. This establishes topology-informed ML as a promising route to overcome the tradeoff between accuracy and simplicity, which are otherwise often mutually exclusive.[5,10,51]

DISCUSSION

We now discuss the interest of using topology-informed reduced-dimensionality descriptors as inputs to the ML model. As shown in Fig. 5, the number of constraints per atom n_c offers a powerful reduced-dimensionality since it allows us to describe the evolution of the dissolution rate in terms of one variable (i.e., n_c) instead of two (that is, the molar fractions of Na_2O and Al_2O_3). However, as shown in Fig. 7, we find that Model III (which is blind to the topology of the atomic network) offers a level of accuracy that is comparable to that offered by Model IV (which embeds n_c as an explicit input). To further understand this point, we now assess whether Model III is able to "learn" by itself that the dissolution rate can be described by the reduced-dimensionality parameter n_c. To this end, we analyze the coefficients of the final linear function yielded by Model III, which relates –log(DR) to the pH and the molar fractions of Na_2O and Al_2O_3. This model can be expressed as:

$$\mathrm{DR} = F(\mathrm{pH})\exp(a[\mathrm{Na_2O}] + b[\mathrm{Al_2O_3}]) \tag{5}$$

where F(pH) is a function that depends only on the pH of the solution and a and b are some coefficients of the model. On the other hand, ref. [31] suggests that the dissolution rate can be expressed as:

$$\mathrm{DR} = \mathrm{DR_0}(\mathrm{pH})\exp\left[\frac{-n_c E_0}{RT}\right] \tag{6}$$

where $\mathrm{DR_0}$(pH) is the dissolution rate when $n_c = 0$, E_0 is activation energy needed to break a unit constraint per atom, R is the perfect gas constant, and T is the temperature.

A comparison between Eqs. (5) and (6)—i.e., by setting equal their respective exponent terms—allows us to determine the number of topological constraints per atom n_c^{guessed} that is "guessed" by Model III as a function of the glass composition (see Supplementary Information for more details). As shown in Fig. 8, we find that Model III is able to infer how the number of constraints depends on the glass composition (see Methods section), which explains why Model III and Model IV eventually offer the same level of accuracy. This demonstrates that, in the present case, ML is able to learn by itself some chemical rules governing the number of topological constraints created by each atom. Note that the number of constraints per atom (n_c) depends not only on glass composition, but also on some "chemical knowledge" of the system, that is, (i) the coordination number of each atom, (ii) the energy of each chemical bond, which can be active or thermally-broken, and (iii) the directionality of each interatomic bond (i.e., covalent vs. ionic), which governs the existence of BB constraints. In that sense, it is notable that the ML model is able to properly "guess" all these chemical features and how they govern the dissolution rate. As discussed below, this is

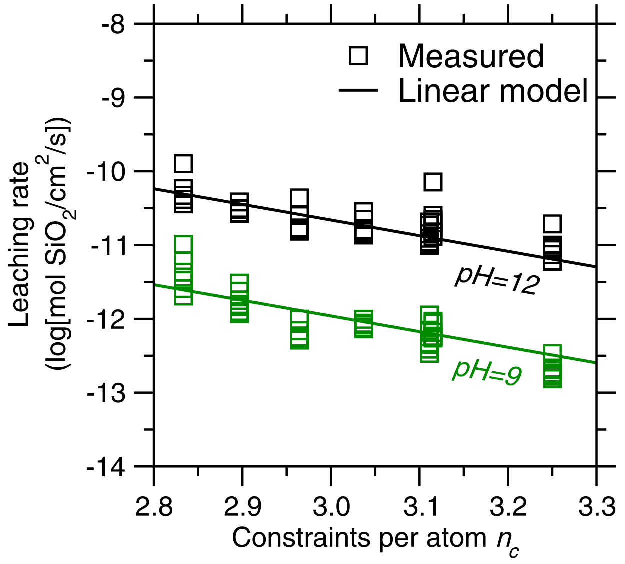


Fig. 5 Dissolution rate of the silicate glasses considered herein as a function of the number of topological constraints per atom for pH = 9 and 12

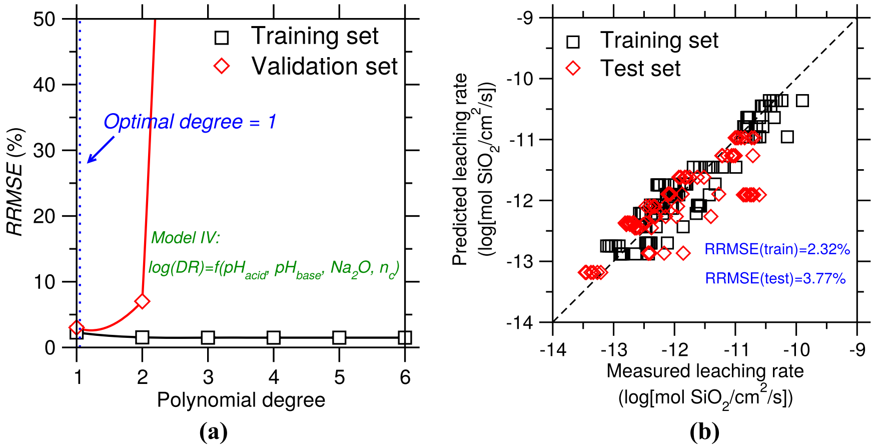


Fig. 6 Predictions from "topology-informed" machine learning, that is, by explicitly accounting for the exponential dependence of the dissolution rate on the inputs, capturing the distinct acidic and caustic regimes, and describing the glass structure in terms of the number of topological constraints per atom n_c ("Model IV"). **a** Evolution of the relative root square mean square error (RRMSE) of the training and validation sets with respect to the polynomial degree p. The minimum in the RRMSE of the validation set indicates $p = 1$ as an optimal polynomial degree (i.e., linear model). **b** Predicted dissolution rate for $p = 1$ as a function of the measured dissolution rate

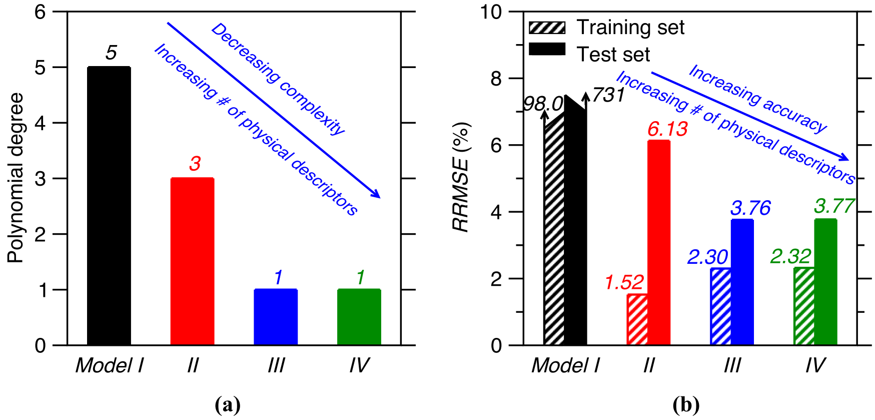


Fig. 7 **a** Complexity (as captured by the polynomial degree) and **b** accuracy (as captured by the relative root square mean square error, RRMSE) of the "blind" and "topology-informed" machine learning models described herein

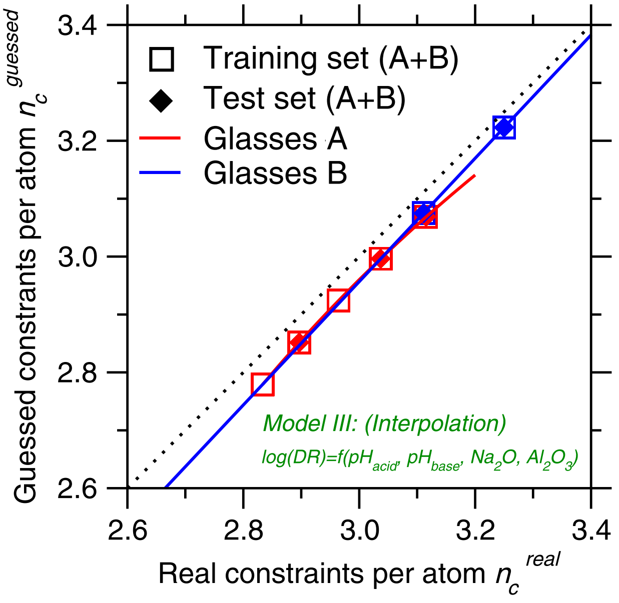


Fig. 8 Number of topological constraints per atom n_c "guessed" by Model III (which is blind to the topology of the atomic network) as a function of the real value of n_c—wherein the training set randomly covers the whole range of glass composition and solution pH. The red and blue lines indicate the guessed n_c values for the two families of glasses considered herein, namely, $(Na_2O)_{0.25}(Al_2O_3)_x(SiO_2)_{0.75-x}$ (Glasses A) (red color) and $(Na_2O)_x(Al_2O_3)_x(SiO_2)_{1-2x}$ (Glasses B) (blue color). Note that, the symbol shape (square or diamond) represents "training set" or "test set", whereas the color (red or blue) denotes the glass family, namely, "Glasses A or B"

permitted by the fact that, here, the training set homogeneously covers all the range of the possible glass compositions. More generally, these results exemplify how an interpretable ML model can offer some physical insights into the relationship between inputs and output—which would not be possible with a less interpretable model (e.g., artificial neural network).

We now assess whether the models considered herein can be used to extrapolate predictions, that is, to predict the dissolution rate of glasses with compositions that are different from those used during the training of the model. To this end, rather than randomly choosing data from database to serve as a test set, we purposely select the data from the Glasses A series as a training set and those from the Glasses B series as a test set. In other words, (i) we train our models based on the dissolution rate data of the first series of glasses with varying Na/Al molar ratios, namely, $(Na_2O)_{25}(Al_2O_3)_y(SiO_2)_{75-y}$ and (ii) we test the ability of the models to predict the dissolution rate of the second series of fully charge-compensated glasses with varying fractions of Na_2O, namely, $(Na_2O)_x(Al_2O_3)_x(SiO_2)_{100-2x}$. In this scenario, the training set does not homogeneously sample the range of glass composition, which allows us to determine whether the models are able to extrapolate predictions from their training sets. Note that these two families of glasses exhibit significantly different structures, namely, (i) Glasses A exhibit varying degrees of polymerization and present some non-bridging oxygen (NBO) atoms, whereas (ii) Glasses B are fully-compensated and theoretically do not comprise any NBO. In addition, the training set (Glasses A) presents a fixed fraction of Na_2O, so that the test set (Glasses B, with varying fractions of Na_2O) is truly unknown to the model.

Figure 9 shows the dissolution rate data predicted by Model III ("topology-blind") and Model IV ("topology-informed") based on the above-mentioned training scenarios. In both cases, the prediction error distribution of the training set is centered ~ 0 with a standard deviation that is close to experimental uncertainty (i.e., ±0.2 log[mol $SiO_2/cm^2/s$]) (see Fig. 9c). This indicates that both models are able to properly interpolate the training set (i.e., Glasses A). In contrast, we find that the test set RRMSE of Model IV is lower than that offered by Model III. In addition, we note that the prediction error distribution is ~ 0 in Model IV, but shows a systematic deviation from 0 in Model III (see Fig. 9c). This signals that the topology-informed Model IV shows an enhanced ability to extrapolate predictions far from the training set.

To further understand how explicitly using the number of constraints per atom n_c as a reduced-dimensionality input enhances the extrapolability of Model IV, we assess whether Model III is still able to "guess" by itself the compositional dependence of the number of constraints per atom when the training set does not homogeneously sample the range of glass compositions. Figure 10 shows the number of constraints per atom "guessed" by Model III. We find that, here, Model III fails to properly infer the compositional evolution of n_c. This arises from the fact that, in this case, the training set does not homogeneously sample the whole domain of glass compositions—so that it is unable to properly capture how the glass composition governs the number of constraints per atom over the entire compositional domain.

Overall, the fact that training the ML model explicitly based on the number of constraints per atom n_c rather than based on the glass composition enhances the potential for extrapolation can be understood as follows. To offer accurate predictions, topology-blind models (e.g., Model III) have to infer how each elementary oxide (e.g., NaO_2 and Al_2O_3) governs the dissolution rate. This

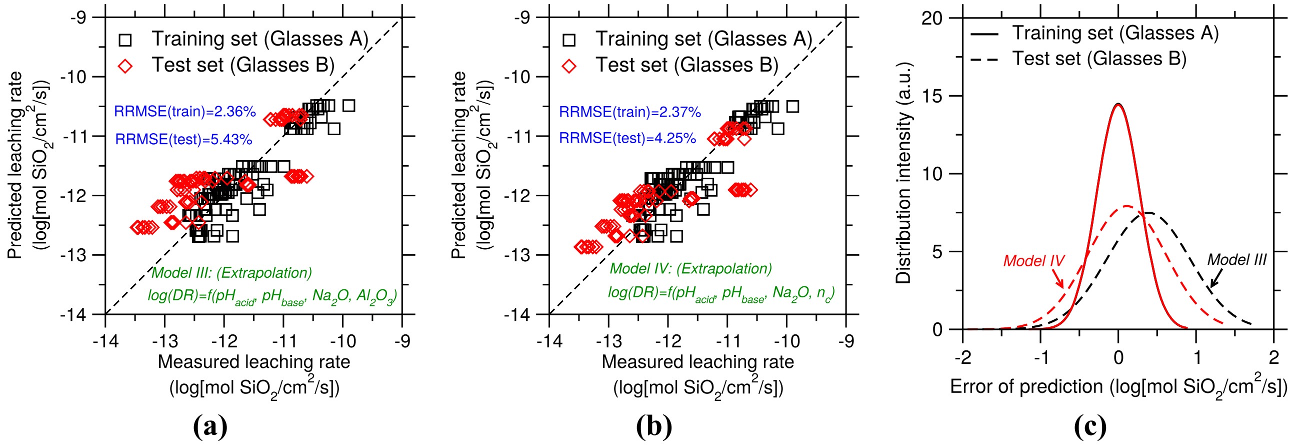


Fig. 9 Dissolution rate predicted by **a** "topology-blind" machine learning (Model III) and **b** "topology-informed" machine learning (Model IV) as a function of the measured dissolution rate—wherein the dissolution data of Glasses A ($(Na_2O)_{0.25}(Al_2O_3)_x(SiO_2)_{0.75-x}$, training set) are used as a training set to predict the dissolution kinetics of Glasses B ($(Na_2O)_x(Al_2O_3)_x(SiO_2)_{1-2x}$, test set). **c** Distribution of prediction error for the training (solid line) and test sets (dash line) offered by Models III (black) and IV (red), respectively. The error is defined as the difference between predicted and measured dissolution rate

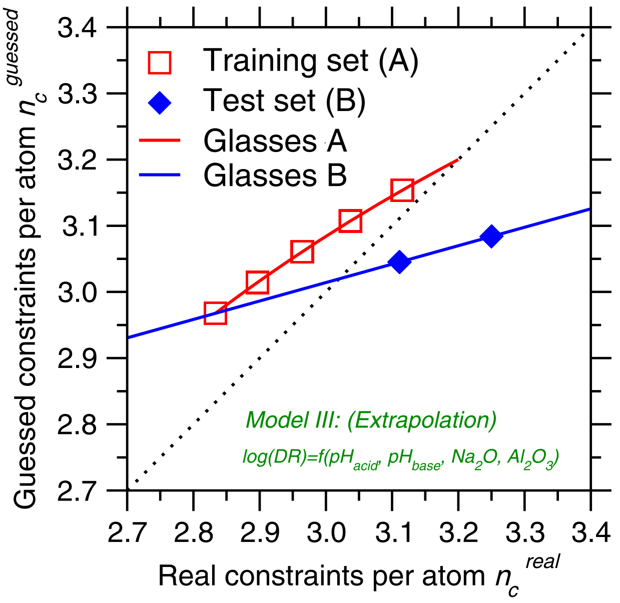


Fig. 10 Number of topological constraints per atom n_c "guessed" by Model III (which is blind to the topology of the atomic network) as a function of the real value of n_c. The red and blue lines indicate the guessed n_c values for the two families of glasses considered herein, namely, $(Na_2O)_{0.25}(Al_2O_3)_x(SiO_2)_{0.75-x}$ (Glasses A) (red color) and $(Na_2O)_x(Al_2O_3)_x(SiO_2)_{1-2x}$ (Glasses B) (blue color), respectively. Here, the dissolution data of Glasses A are used as a training set to predict the dissolution kinetics of Glasses B (test set). Note that, the symbol shape (square or diamond) represents "training set" or "test set", whereas the color (red or blue) denotes the glass family, namely, "Glasses A or B"

requires the use of a large training set that homogeneously sample all the possible glass compositions. In contrast, topology-informed models (Model IV) only have to infer the relationship between the n_c and the dissolution rate. It follows that, once the relationship between n_c and the dissolution rate is properly parameterized, the model will be able to properly predict the dissolution rate of new unknown glass compositions, provided that their number of constraints n_c is similar to that of some glasses of the training set—based on the idea that two glasses with different composition but similar n_c values will exhibit a comparable dissolution rate. As such, topology-informed models only need to be trained with a relatively small training set comprising different glasses with varying n_c values to be able to properly predict the dissolution rate of new glasses with compositions that are unknown to the model. This is illustrated by Fig. 11, which shows that here, some of the glasses of the B series have a number of constraints per atom n_c that is similar to some of glasses of the A series—so that Model IV (topology-informed) succeeds in predicting their dissolution rate while Model III (topology-blind) does not. This suggests that the use of topological inputs capturing into a single metric (n_c) some details of the glass structure makes it possible to reduce the dimensionality of the problem and, thereby, to train predictive models based on limited data sets.

We now further assess the degree of transferability of our topology-informed ML model by testing its ability to predict the dissolution rate of pure glassy silica (SiO_2, taken from ref. [31]). It is worth mentioning that, although the composition of this glass may look similar the those of the training set (i.e., Glasses A), pure glassy silica often exhibits unique, anomalous behaviors. For instance, it is notable that the dissolution rate of SiO_2 (or logarithm thereof) cannot be predicted as a linear extrapolation of those of Glasses B $(Na_2O)_x(Al_2O_3)_x(SiO_2)_{100-2x}$ toward $x \to 0$. As shown in Fig. 12, we find that our topology-informed ML model offers an excellent prediction of the dissolution rate of glassy silica (with RRMSE = 1.66%). It is notable that, although it is trained for glasses comprising a fixed fraction (25%) of Na_2O, our model is able to accurately predict the dissolution rate of pure silica. These results exemplify how adopting topological descriptors enables extrapolations far from the training set—although it will certainly be desirable in the future to test the predictions of this model to some additional families of silicate glasses (e.g., borosilicate, phosphosilicate, etc.).

Note that traditional ML approaches typically rely on a large number of descriptors (e.g., molar masses, bond energy, atomic charges, field strength, etc.), which can be a posteriori be filtered out to reduce the complexity of the model (e.g., using LASSO[52]). Although using a large number of descriptors can increase the ability of the model to interpolate complex data, this comes with several challenges, namely, (i) the computational burden required to filter out irrelevant descriptors is increased, (ii) certain descriptors can appear as being insignificant when taken individually, but may become very useful when combined with each other, (iii) models relying on a large number of descriptors typically require large training sets, (iv) a larger number of descriptors usually increase the complexity of the model, (v) a

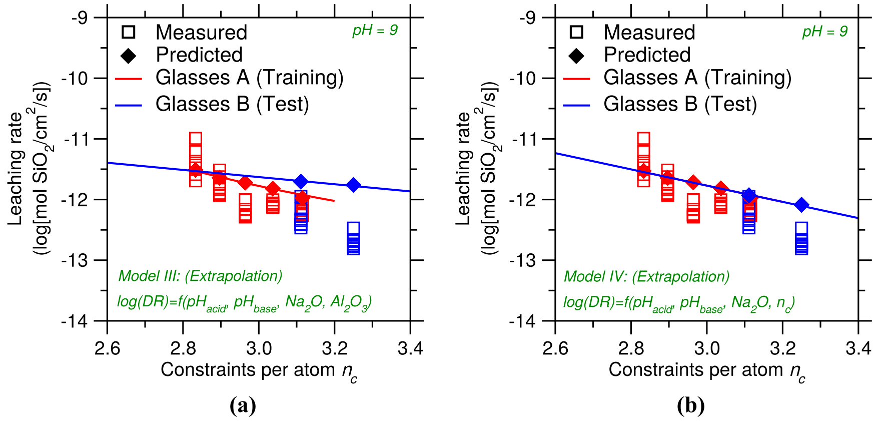


Fig. 11 Dissolution rate predicted by **a** "topology-blind" machine learning (Model III) and "topology-informed" machine learning (Model IV) as a function of the number of topological constraints per atom n_c for pH 9—wherein the dissolution data of Glasses A ($(Na_2O)_{0.25}(Al_2O_3)_x(SiO_2)_{0.75-x}$, training set) (red color) are used to predict the dissolution kinetics of Glasses B ($(Na_2O)_x(Al_2O_3)_x(SiO_2)_{1-2x}$, test set) (blue color). The measured dissolution rates are added for comparison. Note that, the symbol shape (square or diamond) represents the "predicted" or "measured" values, whereas the color (red or blue) denotes the glass family, namely, "Glasses A or B"

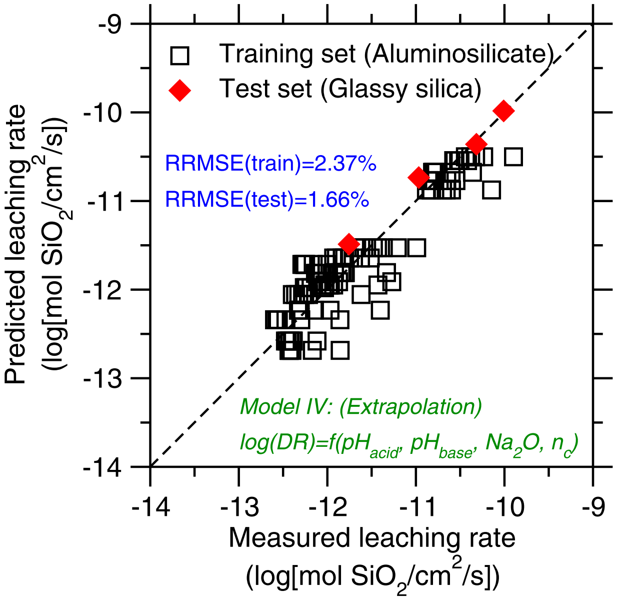


Fig. 12 Dissolution rate predicted by "topology-informed" machine learning (Model IV) as a function of the measured dissolution rate—wherein the dissolution data of sodium aluminosilicate Glasses A ($(Na_2O)_{0.25}(Al_2O_3)_x(SiO_2)_{0.75-x}$, training set) are used as a training set to predict the dissolution kinetics of glassy silica (SiO_2, test set)

larger number of descriptors usually decrease the interpretability of the model, and (vi) the use of a large number of descriptors can result in some degree of overfitting, which, in turn, tends to decrease the extrapolability of the model. In contrast, adopting a topological fingerprint of the atomic network filters out some of the structural details. As such, the use of topological descriptors only may not fully capture some of the fine details of the relationship between composition and dissolution kinetics, but, nevertheless, we find here this level of simplification/filtering to be key in enhancing the extrapolability of the trained models.

Finally, we discuss in terms of (i) model accuracy and (ii) interpretability the choice of using polynomial regression (rather than more complex regression techniques) within the topology-informed ML framework presented herein. To this end, we consider the artificial neural network (ANN[53]) and Gaussian process regression (GPR[54]) techniques. The ANN model used herein is a multilayer perceptron model including one hidden layer made up of 20 neurons, whereas the GPR model used herein is a nonparametric regression model adopting the Matern-type kernel, the noise level of data set being set as 0.01 (see Supplementary Information). Both of these models are trained with topological descriptors (model IV). We assess their potential for extrapolation by training them based on Glasses A and testing their ability to predict the dissolution rates of Glasses B (see above). As expected, we find that both the ANN and GPR models can very accurately interpolate the training set. In both cases, the RRMSE of the training set is below 2%, which is smaller than that offered by polynomial regression (2.4%). We note that the distribution of the prediction error is centered ~0 and is sharper than that offered by polynomial regression (see Fig. 13c). This arises from the fact that, as compared with polynomial regression, both the ANN and GPR models exhibit higher complexities, that is, higher numbers of adjustable parameters. This complexity provides them with more flexibility to interpolate fine details of the training set.

However, we find that both the ANN and GPR models do not offer satisfactory predictions for the test set (see Fig. 13a, b). In detail, the RRMSE of the test set offered by ANN and GPR is 5.62% and 4.51%, respectively, which are both higher than that offered by polynomial regress (i.e., 4.25%, see Fig. 9b). Notably, a visual inspection of Fig. 13a, b and the analysis of the distribution of the prediction error (see Fig. 13c) reveals that both ANN and GPR exhibit a systematic error when predicting the test set—especially for slowly-dissolving glasses, whose dissolution rate tends to be overpredicted. This poor extrapolability can be understood from the fact that both ANN and GPR are intrinsically nonlinear and, hence, do not capture the linear dependence of the logarithm of the dissolution rate on the number of constraints per atom. Such nonlinearity can clearly be observed in Fig. 13a, b. In contrast, polynomial regression intrinsically relies on a linear formulation and, as such, offers more realistic predictions far from the training set. These results exemplify that, in addition of informing the choice of the descriptors, our physical understanding of the underlying mechanism can also guide the choice of the regression technique.

As a notable advantage over more complex regression techniques, polynomial regression offers a high degree of interpretability, which, in turn, can offer some physical insights into the nature of the relationship between inputs and outputs. To illustrate this point, we further expand the number of topological descriptors and use our ML model to assess their weight in governing the dissolution kinetics. To this end, we construct two new "topology-informed" models (referred to as Model IV-a and

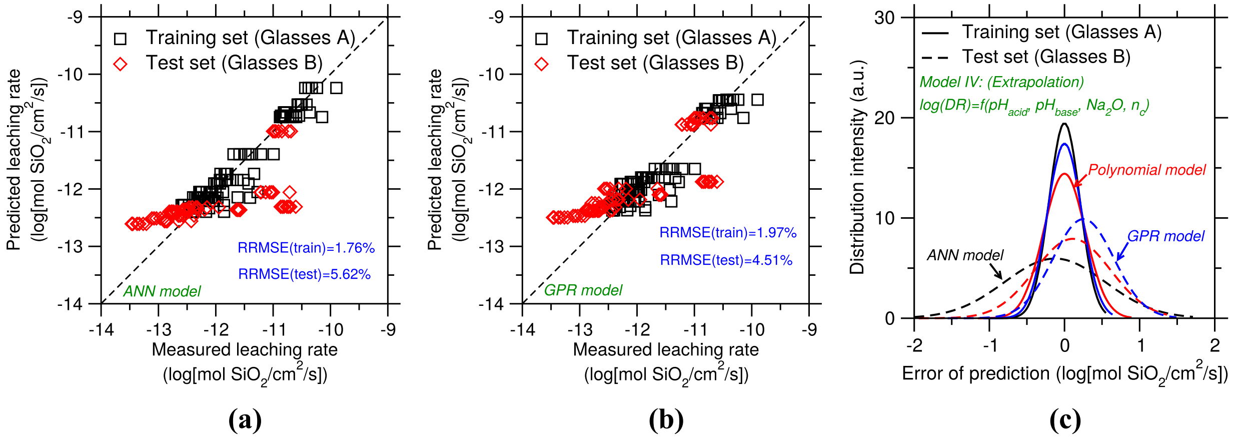


Fig. 13 Dissolution rate predicted by "topology-informed" machine learning (Model IV) using **a** Artificial Neural Network (ANN) and **b** Gaussian Process Regression (GPR) as a function of the measured dissolution rate—wherein the dissolution data of Glasses A ($(Na_2O)_{0.25}(Al_2O_3)_x(SiO_2)_{0.75-x}$, training set) are used as a training set to predict the dissolution kinetics of Glasses B ($(Na_2O)_x(Al_2O_3)_x(SiO_2)_{1-2x}$, test set). **c** Distribution of the prediction error for the training (solid line) and test set (dash line) by using the ANN (black) and GPR models (blue), respectively. The results offered by polynomial regression are added for reference. The error is here defined as the difference between predicted and measured dissolution rates

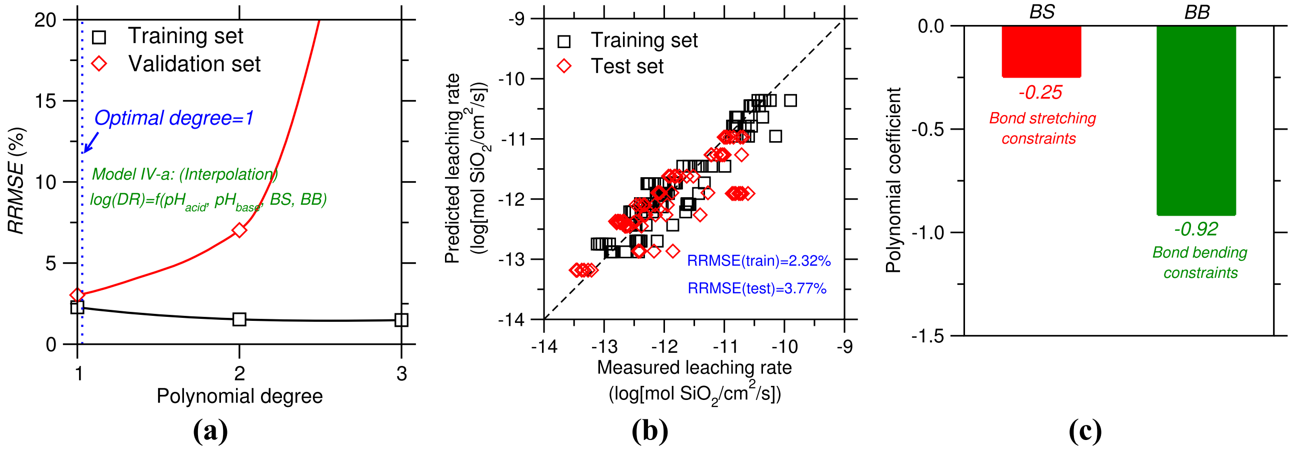


Fig. 14 Outcomes of the "topology-informed" machine learning (Model IV-a) using as inputs the numbers of bond-stretching constraints per atom (BS) and bond-bending constraints per atom (BB). **a** Evolution of the relative root square mean square error (RRMSE) of the training and validation sets with respect to the polynomial degree p. The minimum in the RRMSE of the validation set indicates $p = 1$ as an optimal polynomial degree (i.e., linear model). **b** Predicted dissolution rate (for $p = 1$) as a function of the measured dissolution rate. **c** Coefficients of the polynomial model associated with the BS and BB inputs. Note that the BS and BB input values are normalized in the training process to ensure that the model coefficients reflect the contribution of each input to the dissolution rate

IV-b) by decomposing the term "constraints per atom (n_c)" into several contributions:

$$\text{Model IV} - \text{a}: \quad \log(\text{DR}) = f(\text{pH}_{\text{acid}}, \text{pH}_{\text{base}}, \text{BS}, \text{BB}) \tag{7}$$

$$\text{Model IV} - \text{b}: \quad \log(\text{DR}) = f(\text{pH}_{\text{acid}}, \text{pH}_{\text{base}}, n_c^{\text{Si}}, n_c^{\text{Al}}) \tag{8}$$

In detail, Model IV-a investigates the relative weights of the radial bond-stretching (BS) and angular bond-bending (BB) constraints, whereas Model IV-b investigates the relative weights of the constraints created by Si and Al atoms (n_c^{Si} and n_c^{Al}, respectively). Note $n_c = \text{BS} + \text{BB}$ (see Methods section), so that the original Model IV assumes that radial and angular constraints have the same weight, and so do the constraints created by different elements.

Figures 14 and 15 show the outcomes of Models IV-a and IV-b. First, we find that both models present an optimal degree of 1 (see Figs 14a and 15a). This highlights that the relationship between network topology and the logarithm of the dissolution rate is intrinsically linear. Second, we observe that both models properly interpolate the data set, with a level of accuracy that is comparable to that offered by the original Model IV (see Fig. 14b and 15b). The coefficients of the polynomial regression models can then be interpreted as the weight of each type of constraints in governing the dissolution kinetics. We first note that all the coefficients are negative (see Fig. 14c and 15c), which confirms that all the topological constraints, whatever their nature, tend to decrease the dissolution rate. Interestingly, we find that the angular BB constraints present a larger weight than the linear BS constraints (see Fig. 14c). This finding is confirmed by the fact that the topological constraints created by Si atoms have a larger weight than those created by Al atoms (see Fig. 15c)—as Al atoms do not create any angular constraints (see Methods section).[55] Overall, these results signal that BB constraints have more influence than radial ones on the dissolution kinetics. This suggests that the dissolution kinetics is strongly affected by the directionality of the interatomic bonds. We note that insights of this nature would be challenging to obtain from more complex, less interpretable "black-box" ML models (e.g., ANN). Finally, it is worth to point out that certain glasses are observed to exhibit the same predicted dissolution rate while have different measured

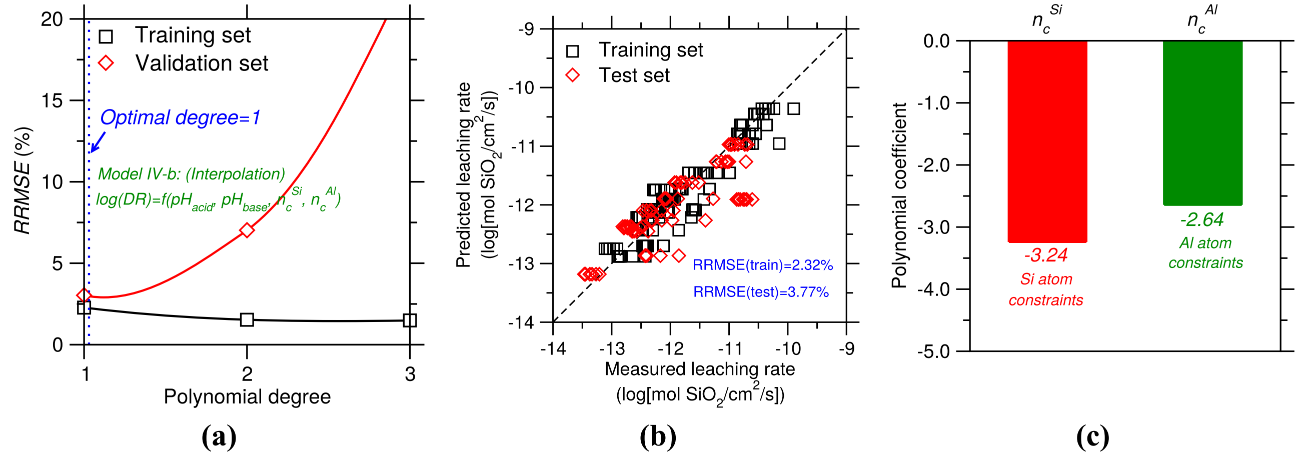


Fig. 15 Outcomes of the "topology-informed" machine learning (Model IV-b) using as inputs the number of constraints per atom created by silicon (n_c^{Si}) and aluminum (n_c^{Al}). **a** Evolution of the relative root square mean square error (RRMSE) of the training and validation sets with respect to the polynomial degree p. The minimum in the RRMSE of the validation set indicates $p = 1$ as an optimal polynomial degree (i.e., linear model). **b** Predicted dissolution rate (for $p = 1$) as a function of the measured dissolution rate. **c** Coefficients of the polynomial model associated with the n_c^{Si} and n_c^{Al} inputs. Note that, the n_c^{Si} and n_c^{Al} input values are normalized in the training process to ensure that the model coefficients reflect the contribution of each input to the dissolution rate

dissolution rate ("flat" pattern in Fig. 14b and 15b). This signals that certain second-order glass features (e.g., powder particle size, surface roughness, thermal history, etc.), if they were available, could further enhance our predictive model.

Overall, these results show that embedding some physical and chemical descriptors within ML models can increase the degree of linearity of the input/output relationship and reduce the dimensionality of the model. This establishes topology-informed ML as a promising route to address some of the limitations of traditional blind ML, namely, by (i) reducing the complexity and increasing the interpretability of the trained models, (ii) limiting the need for large training sets, and (iii) enhancing the ability of the models to extrapolate predictions far from their training sets.

METHODS

Experimental dissolution rate data

For each glass composition and pH, the dissolution tests conducted by Hamilton et al. were carried out on glass powders comprising grain sizes ranging from 74 to 149 µm. These experiments were conducted under static conditions at a surface area to solution volume ratio (SA/V) of ~ 1.4–2.0 cm^{-1}.[26] For each pH, the extent of dissolution was assessed from the concentration of leached SiO_2 (as measured by ICP-AES and ICP-MS) in solution at 5-to-7 regular intervals (for example, 24, 49, 96, 168, and 336 h) of solvent contact. In each case, the pH was recorded before any dissolution and at the time of the dissolution measurement. All the experiments were conducted at 25 °C and ambient pressure. The experimental data present an uncertainty of ± 1.5% of the logarithm of the dissolution rates—as estimated from the standard deviation of the dissolution rate data associated with different measurements conducted on the same glass and at constant pH. More details about the measurements can be found in ref. [26]

ML method

The data points from the training set are first divided into a training and test set (which comprises 30% of the data points). The test set is created by randomly selecting some data points within the training set, while ensuring that the data from the test set are truly unknown to the training set (that is, the pH/compositions combinations used in the test set are not present in the training set). Polynomial regression is then used as a regression method to infer the relationship between inputs and output.[9,10] The least square optimization method is used during the training process of the regression models. We then adopt the 10-fold cross-validation technique[9,10] to optimize the complexity of the model, that is, to identify the maximum polynomial degree of the model. This is accomplished by dividing the initial training set into 10 folds, training the model based on nine of the folds, and using the remaining fold for validation. This procedure is then repeated 10 times until each of the folds has been used as a validation set. The accuracy of the model (for a given maximum polynomial degree) is then determined by averaging the accuracy of the prediction over all the 10 validation folds. The accuracy of the final model (i.e., with optimal complexity) is then assessed by computing the relative root square mean square error by comparing the measured and predicted dissolution rate values DR_i present in the test set:[29]

$$\mathrm{RRMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(\mathrm{DR}_i^{\mathrm{predicted}} - \mathrm{DR}_i^{\mathrm{measured}}\right)^2}{n}} \Bigg/ \left|\frac{\sum_{i=1}^{n}\mathrm{DR}_i^{\mathrm{measured}}}{n}\right| \tag{9}$$

The intrinsic uncertainty of the dissolution data is here directly embedded within the ML framework by incorporating in the training set all the dissolution data obtained for the same glass composition and solution pH (rather than only their average value). This imposes a lower bound of RRMSE = 1.5%, which corresponds to the intrinsic degree of uncertainty of the DR data set measured in experiments.

Topological constraints enumeration

TCT describes the disordered network of glasses as a mechanical truss wherein the atoms are connected to each other via some constraints.[21,35,36] TCT considers two kinds of constraints, namely, (i) the radial BS constraints that keep the interatomic bond length fixed around their average values and (ii) the angular BB constraints that fix the average values of the interatomic angles. A previous study recently suggested that the dissolution rate is related to the number of constraints per atom in the "skeleton" network (that is, that formed by the network-forming species, i.e., Si and O here) rather than to the number of constraints per atom in the whole network (that is, including the network-modifying species, i.e., Na here).[47] Based on this, we enumerate the number of constraints per atom in $(Na_2O)_x(Al_2O_3)_y(SiO_2)_{1-x-y}$ as follows. (i) Each Si creates four BS constraints with its four surrounding O neighbors and five BB constraints (i.e., the number of independent angles that needs to be fixed to define the tetrahedral angular environment of Si atoms). Note that, here, the BS constraints are fully attributed to the cations—so that we do not attribute any BS constraint to the O atoms. (ii) Each tetrahedral Al creates four BS constraints with its four oxygen neighbors. However, based on previous findings,[45] Al atoms do not create any BB constraints—in agreement with the fact that the angular environment of Al atoms is not as well defined as that of Si atoms.[55] (iii) BO atoms (i.e., surrounded by two network-forming cations) form one BB constraint. The number of constraints per atom n_c is then calculated by summing the number of constraints created by each element and dividing by the total number of atoms in the skeleton

Table 1. Table summarizing the fraction, coordination number (CN), number of bond-stretching (BS), and number of bond-bending (BB) constraints created by each atomic species in $(Na_2O)_x(Al_2O_3)_y(SiO_2)_{1-x-y}$ glasses

Atom	Fraction	CN	BS	BB	BS + BB
Si	$1-x-y$	4	4	5	9
Al	$2y$	4	4	0	4
Na	$2x$	—	—	—	—
O	$2-x+y$	—	—	-	—
NBO	$2x-2y$	1	—	0	0
BO	$2-3x+3y$	2	—	1	1

Note that $y \geqslant x$ in all glasses, so that all the Al atoms are assumed to be in tetrahedral configuration[56]

network, namely, Si, Al, BO, NBO (NBO atoms), but excluding Na. The constraints enumeration is summarized in Table 1. It follows that:

$$n_c = \frac{11-12x+2y}{3-2x+2y} \tag{10}$$

This metric (n_c) is used as an input (in lieu of x and y) in Model IV.

Similarly, the number of BS constraints per atom BS is calculated by summing all BS constraints created by each element and dividing by the total number of atoms in the skeleton network:

$$BS = \frac{4-4x+4y}{3-2x+2y} \tag{11}$$

The number of BB constraints per atom BB is calculated by summing all BB constraints created by each element and dividing by the total number of atoms in the skeleton network:

$$BB = \frac{7-8x-2y}{3-2x+2y} \tag{12}$$

The silicon-dominated constraints per atom n_c^{Si} is calculated by summing the number of constraints created by silicon atoms and dividing by the total number of atoms in the skeleton network:

$$n_c^{Si} = \frac{9-9x-9y}{3-2x+2y} \tag{13}$$

The aluminum-dominated constraints per atom n_c^{Al} is calculated by summing the number of constraints created by aluminum atoms and dividing by the total number of atoms in the skeleton network:

$$n_c^{Al} = \frac{8y}{3-2x+2y} \tag{14}$$

In all cases, each input X (i.e., BS, BB, n_c^{Si}, and n_c^{Al}) is transformed into a normalized variable $0 < X' < 1$ as:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{15}$$

where X_{min} and X_{max} are the minimum and maximum values of X, respectively.

DATA AVAILABILITY

All dissolution data that support the findings of this study are already published. The parameters of the models developed herein are given in Supplementary Material.

ACKNOWLEDGEMENTS

This research was performed using funding received from the DOE Office of Nuclear Energy's Nuclear Energy University Programs (Award #18-15020) for the modeling and analysis, and from the Center for Performance and Design of Nuclear Waste Forms and Containers, an Energy Frontier Research Center funded by the US DOE, Office of Science, Basic Energy Sciences under Award #DESC0016584 for subsequent data analysis and preparation of this manuscript. We also acknowledge some financial support provided by the National Science Foundation (Grant No. 1562066).

AUTHOR CONTRIBUTIONS

M.B. designed the research. H.L., T.Z., and N.M.A.K. conducted the ML analysis. H.L. wrote the paper. M.B., M.M.S., J.V.R., and S.G. supervised the research and contributed to the manuscript revision. All authors approved the final version of the manuscript.

ADDITIONAL INFORMATION

Supplementary Information accompanies the paper on the *npj Materials Degradation* website (https://doi.org/10.1038/s41529-019-0094-1).

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.

REFERENCES

1. 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).
2. Brauer, D. S., Rüssel, C. & Kraft, J. Solubility of glasses in the system P2O5–CaO–MgO–Na2O–TiO2: experimental and modeling using artificial neural networks. *J. Non-Cryst. Solids* **353**, 263–270 (2007).
3. 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).
4. Yang, K. et al. Prediction of silicate glasses' stiffness by high-throughput molecular dynamics simulations and machine learning. *arXiv:1901.09323* [cond-mat, physics:physics] (2019).
5. Anoop Krishnan, N. M. et al. Predicting the dissolution kinetics of silicate glasses using machine learning. *J. Non-Cryst. Solids* **487**, 37–45 (2018).
6. Onbaşlı, M. C., Tandia, A. & Mauro, J. C. Mechanical and Compositional Design of High-Strength Corning Gorilla® Glass. in *Handbook of Materials Modeling: Applications: Current and Emerging Materials* (eds Andreoni, W. & Yip, S.) 1–23 (Springer International Publishing, 2018).
7. Cubuk, E. D. et al. Structure-property relationships from universal signatures of plasticity in disordered solids. *Science* **358**, 1033–1037 (2017).
8. Lookman, T., Alexander, F. & Rajan, K. *Information science for materials discovery and design*. (Springer, Berlin, Heidelberg, 2015).
9. Bishop, C. M. *Pattern Recognition and Machine Learning*. (Springer, New York, 2006).
10. Alpaydin, E. *Introduction to Machine Learning*. (MIT Press, 2014).
11. Gubernatis, J. E. & Lookman, T. Machine learning in materials design and discovery: examples from the present and suggestions for the future. *Phys. Rev. Mater.* **2**, 120301 (2018).
12. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *Npj Comput. Mater.* **3**, 54 (2017).
13. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
14. Vienna, J. D., Ryan, J. V., Gin, S. & Inagaki, Y. Current understanding and remaining challenges in modeling long-term degradation of borosilicate nuclear waste glasses. *Int. J. Appl. Glass Sci.* **4**, 283–294 (2013).
15. Grambow, B. Nuclear waste glasses - how durable? *Elements* **2**, 357–364 (2006).
16. Jantzen, C. M., Brown, K. G. & Pickett, J. B. Durable glass for thousands of years. *Int. J. Appl. Glass Sci.* **1**, 38–62 (2010).
17. Collin, M. et al. Structure of international simple glass and properties of passivating layer formed in circumneutral pH conditions. *Npj Mater. Degrad.* **2**, 4 (2018).
18. Helgeson, H. C., Murphy, W. M. & Aagaard, P. Thermodynamic and kinetic constraints on reaction rates among minerals and aqueous solutions. II. Rate constants, effective surface area, and the hydrolysis of feldspar. *Geochim. Cosmochim. Acta* **48**, 2405–2432 (1984).
19. Doremus, R. H. Diffusion-controlled reaction of water with glass. *J. Non-Cryst. Solids* **55**, 143–147 (1983).
20. Christie, J. K., Ainsworth, R. I. & de Leeuw, N. H. Investigating structural features which control the dissolution of bioactive phosphate glasses: beyond the network connectivity. *J. Non-Cryst. Solids* **432**, 321–34.
21. Bauchy, M. Deciphering the atomic genome of glasses by topological constraint theory and molecular dynamics: sa review. *Comput. Mater. Sci.* **159**, 95–102 (2019).
22. Mauro, J. C. Decoding the glass genome. *Curr. Opin. Solid St. Mater. Sci.* **22**, 58–64 (2018).
23. Varshneya, A. K. *Fundamentals of Inorganic Glasses* (Academic Press Inc, 1993).

24. Hamilton, J. P. & Pantano, C. G. Effects of glass structure on the corrosion behavior of sodium-aluminosilicate glasses. *J. Non-Cryst. Solids* **222**, 167–174 (1997).
25. Mysen, B. O. & Richet, P. *Silicate Glasses and Melts: Properties and Structure* (Elsevier, 2005).
26. Hamilton, J. P. *Corrosion behavior of sodium aluminosilicate glasses and crystals* (1999).
27. Hamilton, J. P., Pantano, C. G. & Brantley, S. L. Dissolution of albite glass and crystal. *Geochim. Cosmochim. Acta.* **64**, 2603–2615 (2000).
28. Hamilton, J. P., Brantley, S. L., Pantano, C. G., Criscenti, L. J. & Kubicki, J. D. Dissolution of nepheline, jadeite and albite glasses: toward better models for aluminosilicate dissolution. *Geochim. Cosmochim. Acta.* **65**, 3683–3702 (2001).
29. Li, M.-F., Tang, X.-P., Wu, W. & Liu, H.-B. General models for estimating daily global solar radiation for different solar radiation zones in mainland China. *Energy Convers. Manage.* **70**, 139–148 (2013).
30. Vienna, J. D., Neeway, J. J., Ryan, J. V. & Kerisit, S. N. Impacts of glass composition, pH, and temperature on glass forward dissolution rate. *Npj Mater. Degrad.* **2**, 22 (2018).
31. Pignatelli, I., Kumar, A., Bauchy, M. & Sant, G. Topological control on silicates' dissolution kinetics. *Langmuir* **32**, 4434–4439 (2016).
32. Pierce, E. M., Rodriguez, E. A., Calligan, L. J., Shaw, W. J. & Pete McGrail, B. An experimental study of the dissolution rates of simulated aluminoborosilicate waste glasses as a function of pH and temperature under dilute conditions. *Appl. Geochem.* **23**, 2559–2573 (2008).
33. Mascaraque, N. et al. Dissolution kinetics of hot compressed oxide glasses. *J. Phys. Chem. B* **121**, 9063–9072 (2017).
34. Aagaard, P. & Helgeson, H. C. Thermodynamic and kinetic constraints on reaction rates among minerals and aqueous solutions; I,theoretical considerations. *Am. J. Sci.* **282**, 237–285 (1982).
35. Mauro, J. C. Topological constraint theory of glass. *Am. Ceram. Soc. Bull.* **90**, 7 (2011).
36. Phillips, J. C. Topology of covalent non-crystalline solids .1. Short-range order in chalcogenide alloys. *J. Non-Cryst. Solids* **34**, 153–181 (1979).
37. Phillips, J. C. Topology of covalent non-crystalline solids II: medium-range order in chalcogenide alloys and As-Si-Ge. *J. Non-Cryst. Solids* **43**, 37–77 (1981).
38. Smedskjaer, M. M., Mauro, J. C. & Yue, Y. Prediction of glass hardness using temperature-dependent constraint theory. *Phys. Rev. Lett.* **105**, 115503 (2010).
39. Bauchy, M. et al. Fracture toughness anomalies: viewpoint of topological constraint theory. *Acta Mater.* **121**, 234–239 (2016).
40. Bauchy, M. et al. Topological control on the structural relaxation of atomic networks under stress. *Phys. Rev. Lett.* **119**, 035502 (2017).
41. Gupta, P. K. & Mauro, J. C. Composition dependence of glass transition temperature and fragility. I. A topological model incorporating temperature-dependent constraints. *J. Chem. Phys.* **130**, 094503-094503–094503-094508 (2009).
42. Mauro, J. C., Gupta, P. K. & Loucks, R. J. Composition dependence of glass transition temperature and fragility. II. A topological model of alkali borate liquids. *J. Chem. Phys.* **130**, 234503-234503–234503-234508 (2009).
43. Yang, K. et al. Prediction of the Young's modulus of silicate glasses by topological constraint theory. *J. Non-Cryst. Solids* **514**, 15–19 (2019).
44. Pignatelli, I. et al. Direct experimental evidence for differing reactivity alterations of minerals following irradiation: the case of calcite and wquartz. *Sci. Rep.* **6**, 20155 (2016).
45. Oey, T. et al. Topological controls on the dissolution kinetics of glassy aluminosilicates. *J. Am. Ceram. Soc.* **100**, 5521–5527 (2017).
46. Oey, T. et al. Rate controls on silicate dissolution in cementitious environments. *RILEM Tech. Lett.* **2**, 67–73 (2017).
47. Oey, T. et al. The role of the network-modifier's field-strength in the chemical durability of aluminoborate glasses. *J. Non-Cryst. Solids* **505**, 279–285 (2019).
48. Hsiao, Y.-H. et al. Effects of irradiation on Albite's chemical durability. *J. Phys. Chem. A* **121**, 7835–7845 (2017).
49. Mascaraque, N., Bauchy, M. & Smedskjaer, M. M. Correlating the network topology of oxide glasses with their chemical durability. *J. Phys. Chem. B* **121**, 1139–1147 (2017).
50. Hsiao, Y.-H. et al. Role of electrochemical surface potential and irradiation on garnet-type almandine's dissolution kinetics. *J. Phys. Chem. C.* **122**, 17268–17277 (2018).
51. Aragones, E., Gilboa, I., Postlewaite, A. & Schmeidler, D. Accuracy vs. s implicity: sa complex trade-off. *SSRN Electron. J.* https://doi.org/10.2139/ssrn.332382 (2002).
52. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
53. Gardner, M. W. & Dorling, S. R. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmos. Environ.* **32**, 2627–2636 (1998).
54. Rasmussen, C. E. & Williams, C. K. I. *Gaussian processes for machine learning*. (MIT Press, 2008).
55. 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).
56. Zheng, Q. J. et al. Structure of boroaluminosilicate glasses: Impact of [Al_2 O_3]/[SiO_2] ratio on the structural role of sodium. *Phys. Rev. B* **86**, 054203 (2012).