Literature mining for alternative cementitious precursors and dissolution rate modeling of glassy phases

DR HUGO JAKE UVEGI (Orcid ID : 0000-0002-2846-6078)
MR. BRIAN TRAYNOR (Orcid ID : 0000-0003-2193-3902)

Article type : Article

Literature mining for alternative cementitious precursors and dissolution rate modeling of glassy phases

Hugo Uvegi[1], Zach Jensen[1], Trong Nghia Hoang[2], Brian Traynor[1], Tunahan Aytaş[1], Richard T. Goodwin[3], Elsa A. Olivetti[1*]

[1]Department of Materials Science and Engineering, MIT, Cambridge, USA 02139

[2]MIT-IBM Watson AI Lab, IBM Research, Cambridge, MA, USA 02139

[3]IBM T. J. Watson Research Center, Hawthorne, NY 10532

Hugo Uvegi (huvegi@mit.edu)

Zach Jensen (zjensen@mit.edu)

Trong Nghia Hoang (nghiaht@ibm.com)

Brian Traynor (btraynor@mit.edu)

Tunahan Aytaş (tunahan@mit.edu)

Richard T. Goodwin (rgoodwin@us.ibm.com)

Elsa A. Olivetti (elsao@mit.edu), +16172530877

Abstract

Efforts to reduce the carbon footprint associated with cement and concrete production have resulted in a number of promising lower-emissions alternatives. Still, research has emphasized a small subset of potentially useful precursor materials. With the goal of expanding the precursor pool, this work presents results of parallel literature mining and rate modeling activities. As a result of literature mining, materials with appropriate SiO_2, Al_2O_3, and CaO concentrations were assembled into a comprehensive, representative ternary diagram. 23,000+ materials were extracted from 7,000 DOIs, and 7,500 materials from 6,000 DOIs with $80 \leq SiO_2 + Al_2O_3 + CaO \leq 105$ wt% automatically classified. Both supervised and semi-supervised models were used for dissolution rate prediction of glassy materials with all models pulling from a single data set (n = 802 reported dissolution rates from 105 different glasses). Supervised modeling utilized linear and decision tree regressions to determine features most predictive of dissolution rate, resulting in log-linear relationships between rate and pH, inverse temperature (1/K), and non-bridging oxygens per tetrahedron (NBO/T). Semi-supervised modeling was observed to be more robust to broader feature inclusion, providing similar predictive ability with a relatively larger set of descriptive features. Most importantly, results indicated that models trained on data from disparate scientific communities was adequately predictive (RMSE ≈ 1), particularly under pH ≥ 7 conditions relevant to the cement and alkali activation communities.

1. Introduction

With annual production volumes of 4.1 billion metric tons, ordinary Portland cement (OPC) is purportedly responsible for 5-11%[1] of annual global greenhouse gas emissions.[1–5] Separately, annual red brick production in India alone is estimated at 250 billion bricks, resulting in topsoil degradation and CO_2 emissions to the tune of 40 million metrics tons per year.[6] The search for alternatives, therefore, necessitates not only comparable structural properties and similarly

[1] Previous reports estimate 0.6-1.1 tons CO_2 emitted per ton of cement produced. With 4.1 billion metric tons (BMT) of cement produced, CO_2 emissions from cement are in the range of 2.46-4.51 BMT. Global Carbon Project data estimate emissions of 42.5 ± 3.3 BMT CO_2 in 2018.[4] Given these values, CO_2 emissions from cement production represent 5-11% of annual global emissions, in line with previously published estimates.

large quantities of available resources,[2] but also lower-emission pathways to production. In this way, when used as a replacement for OPC or as primary precursors in alkali activated materials, alternative cementitious materials (ACMs) synthesized from industrial byproducts present a solution to the emissions problem associated with conventional building materials.

Research into three specific ACM precursors—metakaolin, blast furnace slag, and coal fly ash—has dominated the literature. Observed and projected supply shortages necessitate expansion into under-studied silicate and aluminosilicate alternatives.[7, 8] In their 2016 review, Bernal, *et al.* meticulously examined many of the individual findings from across the ACM literature.[9] Still, it is important to further comprehend relationships between material physicochemical properties and reactivity to effectively expand the set of useful precursors. Computational methods offer an opportunity to more broadly survey the literature and expound such connections.

A rather vague term itself, reactivity has been investigated experimentally through a number of distinct methodologies. Tests such as the saturated lime test and the more-recently developed R^3 (rapid, relevant, and reliable) method have focused on pozzolanicity with varied reports of success.[10–13] Procedures based on calorimetry and thermogravimetric analysis have shown promise in differentiating between pozzolanicity and hydraulicity.[14, 15] Still other approaches based on comprehensive material characterization—employing spectroscopic and diffraction-based techniques, among others—have been useful for investigating in-situ reaction extent.[16–27]

Given the aqueous nature of cement chemistry, material dissolution is vital to precursor reactivity,[28] and direct tracking of dissolution has gained popularity in cement science in recent years.[29–40] Studies, such as those by Snellings,[30, 41] Schöler,[35] and Oey[39] have been instrumental

[2] While not the focus of this report, the authors stress the oft-ignored topic of resource availability—either in absolute global quantity or local. Resource reactivity is important, but only insofar as material supply can satisfy demand. For a more comprehensive review of supplementary cementitious material-availability, the readers are directed to Snellings' 2016 work.[7]

in applying concepts and methods long-used in geochemistry,[42–56] glass science,[57, 58] and nuclear waste containment[59–74] to cement science. Still, there is a surprising dearth of dissolution-focused work in reactivity characterization, due in part to the experimental challenges of separating material dissolution from subsequent reaction product formation. Building upon this, in the present work we understand aqueous reactivity by proxy of dissolution rate. This understanding, which has been used previously by the authors,[36, 37, 75] is vital to expanding the pool of useful precursor materials. With this interpretation, materials that readily yield vital elemental species (*e.g.*, Si, Al) to solution can prove useful as cementitious precursors, while inert materials, which do not undergo even surface reactions, can promptly be disregarded.

Quantitative theoretical and empirical relationships describing dissolution in terms of experimental features have been developed over the years. The following is a brief review of this development. For a more in-depth review, the reader is directed to Strachan (2017)[76] and Palandri and Kharaka (2004).[77]

Initially developed by Aagaard and Helgeson (1982),[42] expanded by Grambow (1985),[59] and rederived by Oelkers (2001),[78] Equation (1) has been suggested to describe the dissolution rate of mineral and glass species as a function of distance from equilibrium.

$$r = \vec{r}\left(1 - \exp\left(-\frac{A}{\sigma \mathrm{RT}}\right)\right) \quad (1)$$

$$A = -\mathrm{RT}\ln\left(\frac{\mathrm{Q}}{\mathrm{K}}\right) \quad (2)$$

Here, $\vec{r}\ [\mathrm{mol\ cm^{-2}\ s^{-1}}]$ signifies the forward rate of reaction—implicitly incorporating any kinetically important pre-factors—and is multiplied by a thermodynamic Arrhenius expression, with chemical affinity, $A\ [\mathrm{J\ mol^{-1}}]$, Temkin's average stoichiometric number, σ (unitless), gas constant, $R\ [\mathrm{J\ K^{-1}\ mol^{-1}}]$, and absolute temperature, $T\ [\mathrm{K}]$. Chemical affinity is further described in Equation (2) as a function of ion activity product, Q, and equilibrium constant, K.

As discussed by Strachan, it is important to note that for glasses—thermodynamically metastable phases at best—the equilibrium constant, K, cannot be explicitly defined.[76] This is discussed further in Section 2.4 and Supplementary Information Section S.3.

This rate equation has been cited in many studies, including those recently focused on the dissolution of silicate glasses.[30, 44] At far from equilibrium conditions (*i.e.*, $\mathrm{Q/K} \ll 1$), the overall dissolution rate is equivalent to the forward rate, which is itself a function of pH and activation energy, as shown in Equation (3)[42];

$$r = \vec{r} = \vec{k} a_{\mathrm{H}^+}^{\eta} \exp\left(-\frac{E_a}{\mathrm{RT}}\right) \quad (3)$$

where $\vec{k}\ [\mathrm{mol\ cm^{-2}\ s^{-1}}]$ is the forward rate constant, a_{H^+} (unitless) is the hydrogen ion activity raised to an empirical factor, η (unitless), $E_a\ [\mathrm{J\ mol^{-1}}]$ is activation energy, and other variables are as defined above.

In glasses, for which molar chemistries are not universally defined, explicit stoichiometric factors are often expressed relative to an element or oxide set at one mole per mole glass (*e.g.*, one mole Si or SiO_2 per mole glass for silicate glasses). Furthermore, there is disagreement over the exact role aqueous elemental species play in dissolution processes, but their involvement is generally observed. Specifically, a number of studies have investigated the effect of initial aqueous aluminum species on aluminosilicate dissolution.[47, 55, 76, 79, 80] Still, at low concentration, Strachan[76] showed general dependence only on H^+ and OH^-, describing the dissolution rate as in Equation (4);

$$r = \overrightarrow{k_i}\left[\exp\left(\frac{-E_{a\mathrm{H}^+}}{\mathrm{RT}}\right) a_{\mathrm{H}}^{\eta_{\mathrm{H}}} + \exp\left(\frac{-E_{a\mathrm{H_2O}}}{\mathrm{RT}}\right) + \exp\left(\frac{-E_{a\mathrm{OH}^-}}{\mathrm{RT}}\right) a_{\mathrm{OH}}^{\eta_{\mathrm{OH}}}\right] \quad (4)$$

wherein activities of H^+ and OH^- dominate in their respective regions of the pH scale, and a third term is introduced to describe the rate dependence in the near-neutral region. This follows a similar argument made by Palandri and Kharaka.[77]

Taking the logarithm of Equation (3), log-linear relationships between rate and the hydrogen ion activity exponent, η, pH (*i.e.,* $\log\left(a_{\mathrm{H}^+}\right)$) and the inverse temperature, $1/\mathrm{T}$, are obtained as shown in Equation (5);

$$\log\left(\vec{r}\right) = \log\left(\vec{k}\right) + \eta\log\left(a_{\mathrm{H}^+}\right) - \frac{1}{2.303\mathrm{RT}}\frac{E_a}{} \qquad (5)$$

Throughout the literature, in order to measure direct relationships between each variable and rate, experiments are often designed to quantify specific variables while holding others constant. This is most frequently done to measure pH dependence[75] (*i.e.*, incorporating the E_a term into $\vec{k}$ in Equations (3) and (5) through constant temperature experiments).

In what follows, we explore two parallel methods of examining potentially useful precursors for the synthesis of cementitious materials.

First, we explore the compositional landscape of relevant materials through automated extraction of literature-reported calcium aluminosilicate chemistries. This section explores the compositions of crystalline, glassy, and heterogeneous materials without specific regard for discipline or direct discussion of associated reactivity. The data is visualized by plotting 7,500 labeled samples (and an additional 15,500 yet unlabeled samples) on a ternary SiO_2-Al_2O_3-CaO diagram. While such ternary plots are familiar within the cement science literature, they are typically schematic in nature. To the authors' knowledge, this is the largest set of literature-extracted data directly plotted as such, with data pulled from thousands of DOIs

Subsequently, we focus specifically on the dissolution of glassy materials. In order to elucidate specific feature-dissolution rate relationships, we use machine learning methods to model

reported dissolution rates as a function of chemical, physical, and experimental features. Reported rates (n = 802 rates from 105 distinct glasses) are extracted and aggregated from 33 different papers—pulled from cement,[29, 30, 33–36] geochemistry,[44–48, 50, 52–56] and nuclear waste and glass science[57, 60–74] literature. While other studies have analyzed published data to discover trends and ascertain relationships between driving factors and dissolution rates[32, 81] or other material properties,[82–84] as far as the authors are aware, this is both the largest data set specifically focused on glass dissolution and the first time such a study has tested cross-discipline relationships.

With ample evidence—both experimental and theoretical—supporting the claim that amorphous materials have a higher affinity for reaction than their crystalline analogues,[38, 85–89] this study aims to elucidate why certain glasses react (*i.e.,* dissolve) more readily than others. Whether through metrics such as non-bridging oxygens per tetrahedron (NBO/T),[31, 35, 89] optical basicity,[90, 91] or number of constraints per atom,[32, 38, 39, 92, 93] previous studies have observed dissolution rate dependence on glass network connectivity. That being said, such metrics are often applied in a siloed manner.[3] Though beyond the scope of the current work, it is also important to note that thermal[94–98] and mechanical[99–104] histories of a given material will impact associated aqueous dissolution rates and reactivity in cementitious systems. While only limited studies have investigated the effect of thermal history on glass reactivity,[94] previous work has shown the effect of thermal history on internal stress[96] and glass structure,[97] which are important determinants of dissolution propensity. More effort has focused on mechanical activation, surface roughness, and their effect on material reactivity in cementitious systems. These studies have demonstrated that milling and grinding not only increase reactivity by increasing particle surface area,[101, 102] but they may also reveal more reactive surface chemistry.[99, 100, 103, 104]

[3] While the number of constraints per atom, described in the context of topological constraint theory,[92] has shown recent promise in its ability to represent material connectivity and reactivity,[32, 38, 39, 93] it is not investigated in the current study. Here, only NBO/T and optical basicity are considered due to the ability to directly calculate their values from material oxide composition, where topological constraint calculations rely on molecular dynamics simulations for accurate valuation.

Experimental variables, such as temperature, precursor concentration—as understood by liquid-to-solid (L/S), surface area-to-solution volume (SA/V), or surface area-to-flow rate (SA/FR) ratios—and solution composition (*e.g.*, ionic strength, alkali concentration, pH), also play distinct roles in controlling such reactivity by dictating distance from equilibrium and influencing reaction kinetics. Furthermore, as most experimental dissolution studies operate under extremely dilute conditions (*i.e.*, high L/S, low SA/V, or low SA/FR), results often do not directly translate to reaction kinetics of more concentrated systems. More explicitly, while dissolution experiments may involve milligrams to grams per liter of solution (*i.e.*, L/S >> 1), real systems often have L/S < 1. In this way, real systems quickly become diffusion-limited as surfaces impinge, while dissolution studies are designed to be surface-limited to investigate the intrinsic reactivity of a given material. Such dissolution studies typically also normalize by metrics such as specific surface area, material mass, and solution volume to further discount exogenous parameters. As described below, data aggregated for this study was re-normalized where necessary.

2. Methods

2.1. Data extraction

From a database of over 2.5 million papers—accumulated as part of ongoing work towards a streamlined data extraction pipeline[105–107]—papers were analyzed for the presence of keywords as described in Supplementary Information Section S.1. and detailed in Table S1. Table-reported data from these papers were then extracted via text-mining techniques (*e.g.*, regular expression), utilizing a comprehensive custom dictionary to ensure data were collected and stored in an accessible fashion. Data engineering tasks involved identifying table orientation, handling tables with multiple label columns/rows, and ensuring units were consistent. This resulted in a preliminary data set of 44,000 samples from 9,900 DOIs.

Once extracted, samples were filtered to those with $80 \leq SiO_2 + Al_2O_3 + CaO \leq 105$ wt%—of which there were 23,000 samples from 7,000 DOIs.[4] Normalized data ($SiO_2 + Al_2O_3 + CaO = 100$ wt%) were plotted on a ternary diagram spanning the SiO_2-Al_2O_3-CaO composition space. Further utilizing table-mined sample information, 7,500 samples from 6,000 unique DOIs were automatically labeled as belonging to one of eleven categories (in order of decreasing prevalence): cement, silica species, slag, fly ash, clay, glass, other ash, metakaolin, granite, alumina species, or lime species. While most calcined clays are included within the clay category, metakaolin samples were segregated due to their relative frequency in the literature.

2.2. Data engineering and normalization for dissolution rate analysis

In order to ensure inter-experiment comparisons were appropriate, only experimental results involving dissolution of silica-based glasses and where results were described based on Si-extraction were used. Furthermore, solutions containing initial concentrations of species other than pH-influencing alkalis were excluded (*i.e.*, experiments measuring dissolution in NaOH, KOH, HCl, NaCl, and other similar solutions were included, while those with initial Si, Al, or Ca concentrations were excluded due to their influence on dissolution and precipitation of secondary products). Finally, early-age steady state and initial rates of dissolution were included, while late-age steady state dissolution rates (often referred to as "residual" rates) were not. While informative, rates at very late age (on the order of months to years), are understood to reflect different phenomena than those at earlier age (hours to days), and thus were excluded. We note that while included data was chosen in a deliberate manner, there is no agreed upon methodology for measuring dissolution, nor do all materials dissolve in a completely congruent manner. With that in mind, we have chosen to include Si-based dissolution due to the role of silicon as the major network-former in most of the included glass species.

[4] The authors note that materials outside of this range (*i.e.*, $SiO_2 + Al_2O_3 + CaO < 80$ wt%) can also serve as useful precursors for cementitious binders, and, in fact, may also fall into one of the categories listed. This range was chosen to limit the total number of plotted samples and ensure the validity of the ternary diagram. Future work will expand this search to include broader chemistries.

To effectively compare between different dissolution studies, a standard normalization scheme was established, according to Equations (6)-(8). While similar to equations previously reported in the literature,[30, 53] the authors note that these were developed in order to clarify comparison between samples from disparate fields. Specifically, these equations use commonly reported chemistry measurements from XRF (oxide wt%) and make straightforward the conversion between data reported as [g m^{-2} d^{-1}] and [mol cm^{-2} s^{-1}], respectively common in the nuclear waste glass and geochemistry literature.

$$\vec{r}_{glass,\,mol}\left[\frac{\mathrm{mol_{ox\ in\ glass}}}{\mathrm{cm^2 \cdot s}}\right] = \frac{\vec{r}_i}{f_i} = \frac{\Delta(C_{i,mol})}{\Delta t}\frac{V_{soln}}{m_{glass}A_{glass}}\frac{1}{f_i} \quad (6)$$

$$\vec{r}_{glass,\,g}\left[\frac{\mathrm{g_{glass}}}{\mathrm{cm^2 \cdot s}}\right] = \frac{\vec{r}_i}{X_i} = \frac{\Delta(C_{i,g})}{\Delta t}\frac{V_{soln}}{m_{glass}A_{glass}}\frac{1}{X_i} \quad (7)$$

$$\vec{r}_{glass,mol} = \vec{r}_{glass,g} * \frac{X_i}{f_i} * \frac{1}{M_i} \quad (8)$$

Here, $\vec{r}_i$ represents the forward dissolution rate of oxide i, f_i and X_i respectively represent the mole and mass fractions of oxide i, $C_{i,mol}$ $[\mathrm{mol\ L^{-1}}]$ and $C_{i,g}$ $[\mathrm{g\ L^{-1}}]$ respectively represent the concentration of oxide i in solution at time t [s], V_{soln} [L] represents the initial solution volume, m_{glass} [g] represents the initial mass of glass introduced, A_{glass} $[\mathrm{cm^2\ g^{-1}}]$ represents the initial BET specific surface area of the glass, and M_i $[\mathrm{g\ mol^{-1}}]$ represents the molar mass of oxide i. Applying these equations to experimental dissolution results enables further probing the intrinsic dissolution rates, and thereby reactivity, of the glasses in question. Full dimensional analysis is included in Supplementary Information Section S.2.

2.3. Dissolution rate modeling

Data were analyzed via a number of machine learning-based techniques in order to probe different potential relationships between material chemistry, experimental parameters, and dissolution rate. Features explored included pH, temperature, oxide compositions, chemistry-based connectivity metrics (NBO/T and optical basicity—both of which were calculated directly from reported sample compositions), experimental design ratios (*e.g.*, liquid-to-solid), and others. The full list of investigated features is included in the Supplementary Information in

Table S3. Additional information on the calculation of NBO/T and optical basicity is also included in Supplementary Information Section S.7.

The sample space was initially split into train (80%) and test (20%) sets, and the training set was further split into 5 cross-validation sets of approximately equal sample size used to optimize hyperparameters. Once optimized, the entire training set was used to train each model, then used to predict on the test set. To safeguard against prediction on correlated samples (*i.e.*, samples of identical or similar origin)—which could yield deceptively effective model prediction—data were partitioned ensuring all samples from a single DOI were contained within a single set. Furthermore, we manually checked for feature overlap between DOIs in train and test sets in order to minimize potential bias. While there were few instances of composition overlap, there were no instances of overlap for the complete feature vectors (*i.e.*, even where composition was identical, other features, such as pH and temperature differed across sets). Full train/test splits are detailed in Supplementary Information Tables S4 and S5.

Base 10 logarithm of dissolution rate (log(rate)) was modeled as a function of different subsets of the feature space to test potential predictive relationships. The use of log(rate) was chosen over absolute rates given previously reported log-linear relationships as described in the introduction (Equation (5)). All models were optimized to achieve highest R-squared (R^2) and lowest root mean square error (RMSE) scores when comparing predicted and true $\log_{10}$ rates of dissolution. RMSE scores were deemed particularly relevant given they represent uncertainty in the units of the target (*i.e.*, RMSE = 1 indicates predictions are correct to within 1 log(rate) unit or 1 order of magnitude in real rates). Modeling was completed in two manners:

1) Supervised Model: Log-linear regression models were optimized using the sci-kit learn package in Python. Feature importance was determined via linear and decision tree regression analyses on a manually curated and fully labeled dataset. Features were filtered to optimize for lowest R^2 / highest RMSE of predicted vs. true log(rate).

2) Semi-Supervised Model: Generative embedding models based on Kingma (2014)[108] were optimized (in an unsupervised manner) to learn a latent feature representation directly from a much larger set of incomplete and unlabeled data. The learned representation could then be leveraged as a set of high-level features to analyze the labeled data by means of supervised linear and non-linear models which map these features to a prediction (*i.e.*, log rate of dissolution).

Linear and decision tree regression models were initially fit using the full feature space in order to determine features with the greatest predictive relationships with log(rate). These models were chosen due to the interpretable nature of their results (*i.e.*, predictive features could be easily identified). R^2 and RMSE scores were observed to respectively increase and decrease upon lowering feature dimensionality. Features lacking predictive capacity were determined by exploring the modeled feature weights (*i.e.*, features found to have heavier weight were determined to be more influential) and noting features absent from the produced decision trees. Those with low weights or found deep in the decision tree (*i.e.*, uninfluential) were removed and models were again trained and fit until optimal R^2 and RMSE scores were achieved. Similar analyses were also conducted by removing explicitly correlated features and calculating model scores to avoid over-emphasizing certain features. Prior to linear and decision tree regressions, all data was standardized to ensure proper comparison between feature weights. Additionally, for decision tree regression, the "max depth" hyperparameter was optimized by testing increasing maximum tree depths (from 2-10) and similarly optimizing for R^2 and RMSE scores.

Semi-supervised learning analysis on the other hand was conducted using an end-to-end computation pipeline that comprises (1) an unsupervised neural embedding model to project both labeled and unlabeled data onto a latent space;[109] and (2) a supervised prediction model which makes prediction based on the latent representation of data.[108] Both models can then be optimized simultaneously such that the prediction loss of the supervised model on labeled data (e.g., its RMSE) is used as a learning signal (along with an additional unsupervised information

summarization loss on unlabeled data) to refine the latent projection learned by the neural embedding model. The learned distributions over the latent values of any missing features can then be marginalized out (*i.e.*, weight-averaged out) during prediction. Interested readers are referred to Nazábal, *et al.* (2020)[110] for an in-depth discussion on marginalizing out missing features. For better clarity, a schematic diagram of this entire computation pipeline is also provided below in Figure 1. While this semi-supervised model reduces result interpretability, the latent projection enables the inclusion of unlabeled and potentially incomplete data in a way that would be impossible using standard supervised techniques.

2.4. Thermodynamic considerations

As expressed in Equations (1) and (2), dissolution rate is a function of species undersaturation in solution (Q/K). For small values of Q/K (*i.e.*, $Q/K < 0.05$), conditions are far from equilibrium and dissolution is thus independent of Q/K .[30, 42] In this study, we assume the surface of the glass to be hydrated and its equilibrium constant to be a weighted sum of the equilibrium constants for the dissolution reactions of SiO_2(am), $Al(OH)_3$(am), FeO(OH), $Ca(OH)_2$, and $Mg(OH)_2$. This assumption implies that these hydroxides are rate limiting at the surface of the dissolving glass and also neglects other network forming elements, such as boron. Despite these limitations and assumptions, the model has been applied in previous literature with success[30, 56, 59, 111] and approximates distance from equilibrium during dissolution. Further discussion of thermodynamic considerations and calculations involved can be found in Supplementary Information Section S.3.

3. Results and discussion

3.1. Expanding the precursor pool: Ternary diagram

As introduced in Section 2.1. above, automated data processing resulted in the extraction of over 23,000 sample chemistries, of which 7,500 have been labeled as cement, silica species, slag, fly ash, clay, glass, other ash, metakaolin, granite, alumina species, or lime species, as shown in Figure 2.

While such a ternary diagram is useful in its own right, delving into the compositional distributions within each material class sheds important light on the differences in material variability between industrial products, such as cement, and byproducts/wastes, such as coal fly ash. Figure 3A depicts the cumulative density functions for SiO_2, Al_2O_3, and CaO content (wt%) (normalized to $SiO_2 + Al_2O_3 + CaO = 100$ wt%) across all samples included in the ternary diagram. These data indicate the following trends:

- 58% of samples contain $\leq$ 15 wt% CaO, there is a stable increase in total samples containing 15-70 wt% CaO (26% of samples), and a sharp jump from 70-74 wt% CaO (14% of samples), capturing 98% of samples,
- Only 6% of samples contain $\leq$ 21 wt% SiO_2, 16% of samples contain 21-25% SiO_2, and there is a relatively stable increase in samples containing 25-100% SiO_2 (78% of samples), and
- Almost all (95%) of samples contain $\leq$ 45 wt% Al_2O_3.

Figures 3B and 3C respectively show distributions for cement and fly ash, and other distributions are included in Figure S2. The variance across each oxide dimension is significantly larger for the fly ash than the cement, epitomizing the differences in compositional variability between industrial goods (cement) and byproducts (fly ashes). While industrial grade products are produced to stringent specifications, inconsistencies in byproduct production yields significant sample-to-sample variability, demanding additional considerations for use as ACMs.

Still, category-level information—such as whether a material is a certain type of slag or ash—can be useful in determining expected phase composition and, thus, the potential for reactivity in a cementitious system. While this work prioritizes glassy phases, byproduct heterogeneities necessitate attention to major phases present in any attractive materials. Two examples of byproduct phase estimation based on material category and processing have been demonstrated previously by the authors—one relating to the dissolution and reactivity of mixed-feedstock biomass ash,[36] and another for steel and copper slag mineral phases[75]—demonstrating the utility of this technique in both glassy and crystalline systems.

3.2. Dissolution rate modeling

3.2.1. Supervised model

As described in Section 2.3, a combination of linear and decision tree regression analysis was used to comprehend the data. First pass modeling to determine important features displayed clear trend segregation between dissolution in high and low pH, as expected from previous studies[32, 76] and clearly visible when simply plotting log rate as a function of pH for all samples explored, presented in Figure 4A. Figure 4B displays the same data with thermal information overlaid, demonstrating the trend towards faster dissolution with increasing temperature.

That said, given the correlated nature of many included features and anomalies in some included data, using the full feature space (as defined in Table S3) and all samples in linear regression analysis resulted in poor predictive capacity. In contrast, even with all features present, decision tree regression resulted in surprisingly good prediction (RMSE = 1.24), likely due to its non-linear handling of the data.

Examining the decision tree with all features as shown in Figure 5, we see several interesting relationships appear. First, the root node split at T = 331.15 K (1/T = 0.003 1/K), indicating the importance of temperature as a predictor of rate. Further inspection revealed that most (74%) examined experiments were carried out at temperatures lower than 331.15 K. Furthermore, no clear relationship between temperature and log(rate) was apparent at moderate temperatures (273-331 K), and the lower quantity of experiments performed at higher temperature (>331.15 K) showed more obvious correlation between temperature and log(rate).

Moving to the second layer, we observe an expected split between low and high pH regions for the high temperature branch. For the moderate temperature branch, we initially observed roughness factor and surface area-to-flow rate (SA/FR) ratio splits on the lower temperature branch. For mixed flow reactors, the saturation state of the dissolution reaction is controlled by SA/FR. Given that the dependence of dissolution rate on the saturation state of the reaction has

been a subject of interest,[65, 67, 69, 71, 72, 112, 113] further investigation into the saturation state of the included glass studies was carried out, as described in Section 2.4. Approximating the saturation state of the dissolution reactions enabled us to cross-check these computational findings with thermodynamic domain knowledge. It was possible to calculate the saturation state of 726 of the 802 total data points, and of these 726 samples, only 30 data points had values of $Q/K > 0.05$. In other words, the bulk of these experiments were carried out too far from equilibrium to comment on the effect of SA/FR or other such experimental variables (*e.g.*, L/S, SA/V) on dissolution rate. Therefore, the presence of these variables in the decision tree represented spurious relationships brought on by the presence of such data for only a subset of the samples.

Removing these features revealed a split over NBO/T for the moderate temperature branch. Examining the relationship between log(rate) and NBO/T first revealed an overarching positive correlation between NBO/T and log(rate). However, while the model output indicates a split at NBO/T = 0.074 influences rate prediction, such delineation was not visibly apparent in a plot of log(rate) vs. NBO/T. Returning to the pH split observed at high temperature, we see similar partitions in the third layer of the moderate temperature branch, all reminiscent of the distinct trends in low and high pH environments as observed in Figure 4A, yet all splitting at higher-than-expected pH.

Given that linear regression indicated the same three features as important (*i.e.*, such features exhibited large average feature weight and low standard deviation), the supervised modeling resulted in best fit models of log(rate) based only on pH, inverse absolute temperature (1/K), and NBO/T. While NBO/T is admittedly limited in its ability to differentiate between glasses with varied composition yet equivalent NBO/T values, the modeling here indicated its superiority to any other compositional features investigated. Therefore, further modeling was carried out with only these three features.[5]

[5] Activation energy of dissolution, while expected to be predictive of rate as per Equations (3)-(5), was not reported broadly enough to conclusively determine its influence. It should be noted, however, that when modeling

Decision tree regression at low dimension, not shown here, produced similar fit statistics as determined by RMSE and R^2. This low-dimensional model also revealed an additional split along pH in the fourth layer of the tree (not shown), where we observe not only a clear delineation between rates at acidic and basic pH, but also a mid-range dependence reminiscent of the near neutral term in Equation (5).[76, 77]

Testing linear regression models on the three features and partitioning by pH resulted in significantly better predictive ability. In Figure 6, we compare prediction using only the three features on all samples (Figure 6A) and only on samples at pH ≥ 7 (Figure 6B), focusing for the rest of this work on the latter due to its relevance to cement science. While the former resulted in an RMSE of 1.53, fitting only pH ≥ 7 samples improved the RMSE to 0.97.

Focusing on experiments conducted at pH ≥ 7, we probed additional relationships, including the influence of data from different scientific communities, different experimental set-ups, and presence of B_2O_3, given its relative prevalence in certain subsections of the scientific literature (*i.e.*, nuclear waste glass science is significantly more interested in borosilicate glasses than cement scientists). While dissolution studies focused on industrial waste materials used as precursors for alkali-activated binders have only recently garnered attention, such studies in geochemistry, nuclear waste, and glass science have been carried out for decades. Given the relative history of dissolution studies in the listed scientific communities as compared to the cement and alkali-activation communities and the transferability of results as described below, it is important that we continue exploring fields not traditionally associated with cement. In this way, we can broaden our understanding of material reactivity and utility in cementitious and alkali activated binder systems.

In Figures 7A and 7B, data for samples at pH ≥ 7 is re-segregated by subject area:

---

was conducted only on samples for which activation energies were originally reported, the inclusion of the activation energy term only served to maintain or decrease the predictive capability of the model.

1) Cement and alkali activation[29, 30, 33–36, 57]
2) Nuclear waste and glass science[60–74]
3) Geochemistry[44–48, 50, 52–56]

while data are segmented by B_2O_3-presence and experimental set-up (batch and flow) in Figures 7C and 7D, respectively.

Data in Figure 7A are trained as labeled and tested on the other two subject areas, while the opposite is true in Figure 7B. While in Figure 7A, predictions are biased to faster log(rate) at low true log(rate) and to slower log(rate) at high true log(rate), in Figure 7B, these biases disappear. This is most apparent in the cement-literature trained model, and is likely due to both the relatively smaller sample set (n = 53) and more homogeneous experimental parameters in the cement literature-based samples. Comparing predictions based on the other two subject areas in Figure 7A and all predictions in Figure 7B with those above in Figure 6B, it is clear that cross-field learning results in adequate predictive capabilities when larger and more varied data sets are employed, as all are able to predict dissolution rates to within approximately 1 log(rate) unit.

Similarly, in Figures 7C and 7D, we observe that segregating by either B_2O_3 content or experimental set-up result in RMSE ≤ 1. Biases observed in Figure 7C can be explained as resulting from slower mean log dissolution rates for B_2O_3-absent glasses than B_2O_3-containing glasses, respectively yielding underestimation of rates for the model trained on the former and overestimation when train on the latter. Full statistical parameters are included in the Supplementary Information in Table S6. While there is also some apparent bias at low log(rate) for experiments split on experimental set-up (Figure 7D), the ability of batch and flow experiments to reliably cross-predict is pleasantly surprising. By separating the data in these ways, we observe that these models based on pH, inverse temperature, and NBO/T are indeed predictive across both literature and chemistry.

3.2.2 Semi-supervised model

Semi-supervised machine learning analysis was conducted using the same data as input into the supervised linear and decision tree regression analyses discussed above. The primary goal was to observe the predictive performance of a black-box model in comparison to the supervised model.

In Figure 8, we present RMSE as a function of training epochs and size of training set, as labeled. In Figure 8A, we present results for a model trained on the full feature set, while in Figure 8B, results are for a model trained only on pH, inverse temperature, and NBO/T (comparable with Figure 6A). Finally, in Figure 8C, we compare predictive ability of the models trained on all samples with those trained on experiments at pH $\geq$ 7 (comparable with Figure 6B).

As discussed, while linear modeling did not yield predictive capability with all features included, in Figure 8A, we observe an RMSE of ~2, indicating that, similar to the decision tree regression, this model was able to handle the presence of less-predictive features well. Additionally, comparing Figures 8B and 6A, we observe similar predictive capacity for models trained on only the three most relevant features, yielding performance improvements over full-feature models. Finally, in comparing Figures 8C and 6B, we again see additional performance improvement for models focused only on pH $\geq$ 7 samples.

Examining model performance in its own right, we observe that models are fully trained after only a few hundred epochs. Additionally, with identical test sets, model performance is consistent down to 20% of the training set (n ≈ 128), with minor performance degradation down to 5% of the training set (n ≈ 32), revealing the utility of such models. This is particularly relevant to our situation, as dissolution rates for most materials in the ternary diagram (discussed in Sections 2.1 and 3.1) are unknown. We note that for the model trained on the least data and full feature set, there is a potential for overfitting after many training epochs, as is visible in Figure 8A where model performance degrades in some cases. Figure 8B, however, shows that when only 3 features are used, performance on the test set remains stable or

continues to fall even for large numbers of epochs. This demonstrates that, given the small data set, restricting the feature set adequately reduces the potential for model overfitting.

4. Conclusions

Through this work, we demonstrated the ability of computational methods to both (1) extract data for potential cementitious and alkali activation precursors reported in the literature and (2) learn dissolution models based on data from separate scientific disciplines, with various levels of detail and often distinct methodological eccentricities. These efforts resulted in the largest sets of literature-extracted sample data and dissolution rate modeling data compiled and analyzed to date. As methodologies and data reporting conventions vary significantly between fields, substantial effort was necessary to ensure data and units were compatible.

Log-linear models were useful in confirming previously reported relationships between log(dissolution rate) and pH, inverse temperature, and NBO/T over the input data, and machine learning models proved more robust to anomalous or incomplete data, arriving at similar predictive capability. The results described herein indicate that while simpler linear methods perform better when only the most relevant attributes are provided as input features to the model, non-linear embeddings and data standardization are capable of learning from sparse and irregular data. This is quite promising given the involved and potentially bias-introducing process of manually reducing feature dimensionality. The trade-off, however, is in reduced model interpretability. Proposed enhancements to the model include implementing a (1) an attention-based model (2) a relaxed decision tree model and (3) a mixture of experts model, all of which would improve the discovery of important feature-rate relationships by segmenting the data and would increase model interpretability.

Furthermore, while these models perform well for glasses, further work is necessary to incorporate complex, heterogeneous materials into this model. Many of the industrial byproducts of interest as supplementary and alternative cementitious materials are heterogeneous in nature, containing multiple phases known to dissolve and react according to

distinct mechanisms and at different rates. Equation (9) depicts our initial hypothesis of how understanding the aqueous reactivity of such heterogeneous systems could be understood;

$$r_{\text{material}} = \sum_i f_{\text{phase } i}\, r_{\text{phase } i} \quad (9)$$

where r_{material} represents the overall dissolution rate of a given material, $f_{\text{phase } i}$ represents the fraction of phase i in said material, and $r_{\text{phase } i}$ represents the dissolution rate of phase i. In this way, dissolution of heterogeneous materials can be modeled as a weighted sum of dissolution rates of each component material. A similar concept was introduced by Gudbrandsson, *et al*. (2011) in their discussion of heterogeneous crystalline basalt rock dissolution.[114] Here, we suggest that this idea can similarly be applied to a broader array of heterogeneous materials. Further work will be necessary to confirm this hypothesis.

5. Acknowledgments

We would like to acknowledge funding support from the MIT-IBM Watson AI Lab. We would also like to acknowledge partial funding from the National Science Foundation DMREF Awards 1922311, 1922372, and 1922090; the Office of Naval Research (ONR) under contract N00014-20-1-2280; and the MIT Energy Initiative.

References

1. Gartner E. Industrially interesting approaches to “low-CO2” cements. *Cem Concr Res*. 2004;34(9):1489–1498. https://doi.org/10.1016/j.cemconres.2004.01.021
2. Habert G, Billard C, Rossi P, Chen C, Roussel N. Cement production technology improvement compared to factor 4 objectives. *Cem Concr Res*. 2010;40(5):820–826. https://doi.org/10.1016/j.cemconres.2009.09.031
3. Olivetti EA, Cullen JM. Toward a sustainable materials system. *Science (80-)*. 2018;360(6396):1396–1398. https://doi.org/10.1126/science.aat6821
4. Friedlingstein P, Jones MW, O’Sullivan M, *et al.* Global carbon budget 2019. *Earth Syst Sci Data*. 2019;11(4):1783–1838. https://doi.org/10.5194/essd-11-1783-2019
5. U.S. Geological Survey. Mineral commodity summaries 2020. 2020 https://doi.org/10.3133/mcs2020
6. Nath AJ, Lal R, Das AK. Fired Bricks: CO2 Emission and Food Insecurity. *Glob Challenges*. 2018;2(4):1700115. https://doi.org/10.1002/gch2.201700115
7. Snellings R. Assessing, understanding and unlocking supplementary cementitious materials. *RILEM Tech*

Lett. 2016;1:50–55. https://doi.org/10.21809/rilemtechlett.2016.12

8. Juenger MCG, Snellings R, Bernal SA. Supplementary cementitious materials: New sources, characterization, and performance insights. *Cem Concr Res*. 2019;122:257–273. https://doi.org/10.1016/j.cemconres.2019.05.008
9. Bernal SA, Rodríguez ED, Kirchheim AP, Provis JL. Management and valorisation of wastes through use in producing alkali-activated cement materials. *J Chem Technol Biotechnol*. 2016;91(9):2365–2388. https://doi.org/10.1002/jctb.4927
10. Donatello S, Tyrer M, Cheeseman CR. Comparison of test methods to assess pozzolanic activity. *Cem Concr Compos*. 2010;32(2):121–127. https://doi.org/10.1016/j.cemconcomp.2009.10.008
11. Snellings R, Scrivener KL. Rapid screening tests for supplementary cementitious materials: past and future. *Mater Struct*. 2016;49(8):3265–3279. https://doi.org/10.1617/s11527-015-0718-z
12. Avet F, Snellings R, Alujas Diaz A, Ben Haha M, Scrivener K. Development of a new rapid, relevant and reliable (R3) test method to evaluate the pozzolanic reactivity of calcined kaolinitic clays. *Cem Concr Res*. 2016;85:1–11. https://doi.org/10.1016/j.cemconres.2016.02.015
13. Suraneni P, Weiss J. Examining the pozzolanicity of supplementary cementitious materials using isothermal calorimetry and thermogravimetric analysis. *Cem Concr Compos*. 2017;83:273–278. https://doi.org/10.1016/j.cemconcomp.2017.07.009
14. Suraneni P, Hajibabaee A, Ramanathan S, Wang Y, Weiss J. New insights from reactivity testing of supplementary cementitious materials. *Cem Concr Compos*. 2019;103:331–338. https://doi.org/10.1016/j.cemconcomp.2019.05.017
15. Wang Y, Suraneni P. Experimental methods to determine the feasibility of steel slags as supplementary cementitious materials. *Constr Build Mater*. 2019;204:458–467. https://doi.org/10.1016/j.conbuildmat.2019.01.196
16. Fernández-Jiménez A, Puertas F. Alkali-activated slag cements: Kinetic studies. *Cem Concr Res*. 1997;27(3):359–368. https://doi.org/10.1016/S0008-8846(97)00040-9
17. Fernández-Jiménez A, de la Torre AG, Palomo A, López-Olmo G, Alonso MM, Aranda MAG. Quantitative determination of phases in the alkaline activation of fly ash. Part II: Degree of reaction. *Fuel*. 2006;85(14–15):1960–1969. https://doi.org/10.1016/j.fuel.2006.04.006
18. Skibsted J, Snellings R. Reactivity of supplementary cementitious materials (SCMs) in cement blends. *Cem Concr Res*. 2019;124:105799. https://doi.org/10.1016/j.cemconres.2019.105799
19. Kucharczyk S, Zajac M, Stabler C, *et al.* Structure and reactivity of synthetic CaO-Al2O3-SiO2 glasses. *Cem Concr Res*. 2019;120:77–91. https://doi.org/10.1016/j.cemconres.2019.03.004
20. Rees CA, Provis JL, Lukey GC, van Deventer JSJ. In situ ATR-FTIR study of the early stages of fly ash geopolymer gel formation. *Langmuir*. 2007;23(17):9076–82. https://doi.org/10.1021/la701185g
21. Provis JL, van Deventer JSJ. Geopolymerisation kinetics. 1. In situ energy-dispersive X-ray diffractometry.

Chem Eng Sci. 2007;62(9):2309–2317. https://doi.org/10.1016/j.ces.2007.01.027

22. Haha M Ben, De Weerdt K, Lothenbach B. Quantification of the degree of reaction of fly ash. *Cem Concr Res*. 2010;40(11):1620–1629. https://doi.org/10.1016/j.cemconres.2010.07.004
23. Zhang Z, Wang H, Provis JL. Quantitative study of the reactivity of fly ash in geopolymerization by FTIR. *J Sustain Cem Mater*. 2012;1(4):154–166. https://doi.org/10.1080/21650373.2012.752620
24. White CE, Provis JL, Bloomer B, Henson NJ, Page K. In situ X-ray pair distribution function analysis of geopolymer gel nanostructure formation kinetics. *Phys Chem Chem Phys*. 2013;15(22):8573–82. https://doi.org/10.1039/c3cp44342f
25. Provis JL, Hajimohammadi A, White CE, *et al.* Nanostructural characterization of geopolymers by advanced beamline techniques. *Cem Concr Compos*. 2013;36:56–64. https://doi.org/10.1016/j.cemconcomp.2012.07.003
26. Durdziński PT, Dunant CF, Haha M Ben, Scrivener KL. A new quantification method based on SEM-EDS to assess fly ash composition and study the reaction of its individual components in hydrating cement paste. *Cem Concr Res*. 2015;73:111–122. https://doi.org/10.1016/j.cemconres.2015.02.008
27. Chaunsali P, Uvegi H, Osmundsen R, *et al.* Mineralogical and microstructural characterization of biomass ash binder. *Cem Concr Compos*. 2018;89:41–51. https://doi.org/10.1016/j.cemconcomp.2018.02.011
28. Gartner EM, Macphee DE. A physico-chemical basis for novel cementitious binders. *Cem Concr Res*. 2011;41(7):736–749. https://doi.org/10.1016/j.cemconres.2011.03.006
29. Li C, Li Y, Sun H, Li L. The composition of fly ash glass phase and its dissolution properties applying to geopolymeric materials. *J Am Ceram Soc*. 2011;94(6):1773–1778. https://doi.org/10.1111/j.1551-2916.2010.04337.x
30. Snellings R. Solution-controlled dissolution of supplementary cementitious material glasses at pH 13: The effect of solution composition on glass dissolution rates. *J Am Ceram Soc*. 2013;96(8):2467–2475. https://doi.org/10.1111/jace.12480
31. Newlands KC, Foss M, Matchei T, Skibsted J, Macphee DE. Early stage dissolution characteristics of aluminosilicate glasses with blast furnace slag- and fly-ash-like compositions. *J Am Ceram Soc*. 2017;100(5):1941–1955. https://doi.org/10.1111/jace.14716
32. Liu H, Zhang T, Anoop Krishnan NM, *et al.* Predicting the dissolution kinetics of silicate glasses by topology-informed machine learning. *npj Mater Degrad*. 2019;3(1):32. https://doi.org/10.1038/s41529-019-0094-1
33. Durdziński PT, Snellings R, Dunant CF, Haha M Ben, Scrivener KL. Fly ash as an assemblage of model Ca-Mg-Na-aluminosilicate glasses. *Cem Concr Res*. 2015;78:263–272. https://doi.org/10.1016/j.cemconres.2015.08.005
34. Maraghechi H, Rajabipour F, Pantano CG, Burgos WD. Effect of calcium on dissolution and precipitation reactions of amorphous silica at high alkalinity. *Cem Concr Res*. 2016;87:1–13. https://doi.org/10.1016/j.cemconres.2016.05.004

35. Schöler A, Winnefeld F, Haha M Ben, Lothenbach B. The effect of glass composition on the reactivity of synthetic glasses. *J Am Ceram Soc*. 2017;100(6):2553–2567. https://doi.org/10.1111/jace.14759
36. Uvegi H, Chaunsali P, Traynor B, Olivetti E. Reactivity of industrial wastes as measured through ICP-OES: A case study on siliceous Indian biomass ash. *J Am Ceram Soc*. 2019;102(12):7678–7688. https://doi.org/10.1111/jace.16628
37. Uvegi H, Traynor B, Chaunsali P, Olivetti E. Determining viability of industrial byproducts in alkali activated systems. *15th Int. Congr. Chem. Cem.* Prague, Czech Republic: Guarant International spol s r.o.; 2019
38. Pignatelli I, Kumar A, Bauchy M, Sant G. Topological Control on Silicates' Dissolution Kinetics. *Langmuir*. 2016;32(18):4434–4439. https://doi.org/10.1021/acs.langmuir.6b00359
39. Oey T, Kumar A, Pignatelli I, *et al.* Topological controls on the dissolution kinetics of glassy aluminosilicates. *J Am Ceram Soc*. 2017;100(12):5521–5527. https://doi.org/10.1111/jace.15122
40. Oey T, Timmons J, Stutzman P, *et al.* An improved basis for characterizing the suitability of fly ash as a cement replacement agent. *J Am Ceram Soc*. 2017;100(10):4785–4800. https://doi.org/10.1111/jace.14974
41. Snellings R. Surface chemistry of calcium aluminosilicate glasses. *J Am Ceram Soc*. 2015;98(1):303–314. https://doi.org/10.1111/jace.13263
42. Aagaard P, Helgeson HC. Thermodynamic and kinetic constraints on reaction rates among minerals and aqueous solutions; I, Theoretical considerations. *Am J Sci*. 1982;282(3):237–285. https://doi.org/10.2475/ajs.282.3.237
43. Lasaga AC. Chemical kinetics of water-rock interactions. *J Geophys Res Solid Earth*. 1984;89(B6):4009–4025. https://doi.org/10.1029/JB089iB06p04009
44. Wolff-Boenisch D, Gislason SR, Oelkers EH, Putnis C V. The dissolution rates of natural glasses as a function of their composition at pH 4 and 10.6, and temperatures from 25 to 74°C. *Geochim Cosmochim Acta*. 2004;68(23):4843–4858. https://doi.org/10.1016/j.gca.2004.05.027
45. Fraysse F, Pokrovsky OS, Schott J, Meunier J-D. Surface properties, solubility and dissolution kinetics of bamboo phytoliths. *Geochim Cosmochim Acta*. 2006;70(8):1939–1951. https://doi.org/10.1016/j.gca.2005.12.025
46. Pierce EM, Reed LR, Shaw WJ, *et al.* Experimental determination of the effect of the ratio of B/Al on glass dissolution along the nepheline (NaAlSiO4)–malinkoite (NaBSiO4) join. *Geochim Cosmochim Acta*. 2010;74(9):2634–2654. https://doi.org/10.1016/j.gca.2009.09.006
47. Declercq J, Diedrich T, Perrot M, Gislason SR, Oelkers EH. Experimental determination of rhyolitic glass dissolution rates at 40–200°C and 2<pH<10.1. *Geochim Cosmochim Acta*. 2013;100:251–263. https://doi.org/10.1016/j.gca.2012.10.006
48. Rimstidt JD, Zhang Y, Zhu C. Rate equations for sodium catalyzed amorphous silica dissolution. *Geochim Cosmochim Acta*. 2016;195:120–125. https://doi.org/10.1016/j.gca.2016.09.020

49. Schott J, Pokrovsky OS, Oelkers EH. The link between mineral dissolution/precipitation kinetics and solution chemistry. *Rev Mineral Geochemistry*. 2009;70(1):207–258. https://doi.org/10.2138/rmg.2009.70.6
50. White AF. Surface chemistry and dissolution kinetics of glassy rocks at 25°C. *Geochim Cosmochim Acta*. 1983;47(4):805–815. https://doi.org/10.1016/0016-7037(83)90114-X
51. Hamilton JP, Pantano CG. Effects of glass structure on the corrosion behavior of sodium-aluminosilicate glasses. *J Non Cryst Solids*. 1997;222(15):167–174. https://doi.org/10.1016/S0022-3093(97)90110-1
52. Hamilton JP, Pantano CG, Brantley SL. Dissolution of albite glass and crystal. *Geochim Cosmochim Acta*. 2000;64(15):2603–2615. https://doi.org/10.1016/S0016-7037(00)00388-4
53. Hamilton JP, Brantley SL, Pantano CG, Criscenti LJ, Kubicki JD. Dissolution of nepheline, jadeite and albite glasses: toward better models for aluminosilicate dissolution. *Geochim Cosmochim Acta*. 2001;65(21):3683–3702. https://doi.org/10.1016/S0016-7037(01)00724-4
54. Hamilton JP. Corrosion behavior of sodium aluminosilicate glasses and crystals. 1999
55. Oelkers EH, Gislason SR. The mechanism, rates and consequences of basaltic glass dissolution: I. An experimental study of the dissolution rates of basaltic glass as a function of aqueous Al, Si and oxalic acid concentration at 25°C and pH = 3 and 11. *Geochim Cosmochim Acta*. 2001;65(21):3671–3681. https://doi.org/10.1016/S0016-7037(01)00664-0
56. Gislason SR, Oelkers EH. Mechanism, rates, and consequences of basaltic glass dissolution: II. An experimental study of the dissolution rates of basaltic glass as a function of pH and temperature. *Geochim Cosmochim Acta*. 2003;67(20):3817–3832. https://doi.org/10.1016/S0016-7037(03)00176-5
57. Wirth G., Gieskes J. The initial kinetics of the dissolution of vitreous silica in aqueous media. *J Colloid Interface Sci*. 1979;68(3):492–500. https://doi.org/10.1016/0021-9797(79)90307-2
58. Conradt R. Chemical durability of oxide glasses in aqueous solutions: A review. *J Am Ceram Soc*. 2008;91(3):728–735. https://doi.org/10.1111/j.1551-2916.2007.02101.x
59. Grambow B. A general rate equation for nuclear waste glass corrosion. *MRS Proc*. 1985;44:15–27. https://doi.org/10.1557/PROC-44-15
60. Knauss KG, Bourcier WL, McKeegan KD, *et al.* Dissolution kinetics of a simple analogue nuclear waste glass as a function of pH, time and temperature. *MRS Proc*. 1989;176:371–381. https://doi.org/10.1557/PROC-176-176
61. Liu S, Ferrand K, Lemmens K. Transport- and surface reaction-controlled SON68 glass dissolution at 30°C and 70°C and pH=13.7. *Appl Geochemistry*. 2015;61:302–311. https://doi.org/10.1016/j.apgeochem.2015.06.014
62. Cassingham N, Corkhill CL, Backhouse DJ, *et al.* The initial dissolution rates of simulated UK Magnox–ThORP blend nuclear waste glass as a function of pH, temperature and waste loading. *Mineral Mag*. 2015;79(6):1529–1542. https://doi.org/10.1180/minmag.2015.079.6.28

63. Icenhower JP, Steefel CI. Dissolution rate of borosilicate glass SON68: A method of quantification based upon interferometry and implications for experimental and natural weathering rates of glass. *Geochim Cosmochim Acta*. 2015;157:147–163. https://doi.org/10.1016/j.gca.2015.02.037
64. Fournier M, Ull A, Nicoleau E, *et al.* Glass dissolution rate measurement and calculation revisited. *J Nucl Mater*. 2016;476:140–154. https://doi.org/10.1016/j.jnucmat.2016.04.028
65. Iwalewa TM, Qu T, Farnan I. Investigation of the maximum dissolution rates and temperature dependence of a simulated UK nuclear waste glass in circum-neutral media at 40 and 90 ° C in a dynamic system. *Appl Geochemistry*. 2017;82:177–190. https://doi.org/10.1016/j.apgeochem.2017.05.018
66. Neeway JJ, Rieke PC, Parruzot BP, Ryan J V., Asmussen RM. The dissolution behavior of borosilicate glasses in far-from equilibrium conditions. *Geochim Cosmochim Acta*. 2018;226:132–148. https://doi.org/10.1016/j.gca.2018.02.001
67. Abraitis PK, McGrail BP, Trivedi DP, Livens FR, Vaughan DJ. Single-pass flow-through experiments on a simulated waste glass in alkaline media at 40°C. I. Experiments conducted at variable solution flow rate to glass surface area ratio. *J Nucl Mater*. 2000;280(2):196–205. https://doi.org/10.1016/S0022-3115(00)00041-6
68. Abraitis PK, Livens FR, Monteith JE, *et al.* The kinetics and mechanisms of simulated British Magnox waste glass dissolution as a function of pH, silicic acid activity and time in low temperature aqueous systems. *Appl Geochemistry*. 2000;15(9):1399–1416. https://doi.org/10.1016/S0883-2927(99)00118-3
69. Jeong S-Y, Ebert WL. Glass dissolution rates from static and flow-through tests. Reno, NV: 2002
70. Pierce EM, Rodriguez EA, Calligan LJ, Shaw WJ, 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 Geochemistry*. 2008;23(9):2559–2573. https://doi.org/10.1016/j.apgeochem.2008.05.006
71. Pierce EM, Richards EL, Davis AM, Reed LR, Rodriguez EA. Aluminoborosilicate waste glass dissolution under alkaline conditions at 40°C: implications for a chemical affinity-based rate equation. *Environ Chem*. 2008;5(1):73. https://doi.org/10.1071/EN07058
72. Icenhower JP, McGrail BP, Shaw WJ, *et al.* Experimentally determined dissolution kinetics of Na-rich borosilicate glass at far from equilibrium conditions: Implications for Transition State Theory. *Geochim Cosmochim Acta*. 2008;72(12):2767–2788. https://doi.org/10.1016/j.gca.2008.02.026
73. Gin S, Beaudoux X, Angéli F, Jégou C, Godon N. Effect of composition on the short-term and long-term dissolution rates of ten borosilicate glasses of increasing complexity from 3 to 30 oxides. *J Non Cryst Solids*. 2012;358(18–19):2559–2570. https://doi.org/10.1016/j.jnoncrysol.2012.05.024
74. Icenhower JP, Steefel CI. Experimentally determined dissolution kinetics of SON68 glass at 90°C over a silica saturation interval: Evidence against a linear rate law. *J Nucl Mater*. 2013;439(1–3):137–147. https://doi.org/10.1016/j.jnucmat.2013.04.008
75. Traynor B, Mulcahy C, Uvegi H, Aytas T, Chanut N, Olivetti EA. Dissolution of olivines from steel and copper

slags in basic solution. *Cem Concr Res*. 2020;133:106065. https://doi.org/10.1016/j.cemconres.2020.106065

76. Strachan D. Glass dissolution as a function of pH and its implications for understanding mechanisms and future experiments. *Geochim Cosmochim Acta*. 2017;219:111–123. https://doi.org/10.1016/j.gca.2017.09.008
77. Palandri JL, Kharaka YK. A compilation of rate parameters of water-mineral interaction kinetics for application to geochemical modeling. Menlo Park, CA: 2004
78. Oelkers EH. General kinetic description of multioxide silicate mineral and glass dissolution. *Geochim Cosmochim Acta*. 2001;65(21):3703–3719. https://doi.org/10.1016/S0016-7037(01)00710-4
79. Chappex T, Scrivener KL. The effect of aluminum in solution on the dissolution of amorphous silica and its relation to cementitious systems. *J Am Ceram Soc*. 2013;96(2):592–597. https://doi.org/10.1111/jace.12098
80. McGrail BP, Bacon DH, Icenhower JP, *et al.* Near-field performance assessment for a low-activity waste glass disposal system: Laboratory testing to modeling results. *J Nucl Mater*. 2001;298(1–2):95–111. https://doi.org/10.1016/S0022-3115(01)00576-1
81. Anoop Krishnan NM, Mangalathu S, Smedskjaer MM, Tandia A, Burton H, Bauchy M. Predicting the dissolution kinetics of silicate glasses using machine learning. *J Non Cryst Solids*. 2018;487:37–45. https://doi.org/10.1016/j.jnoncrysol.2018.02.023
82. Cassar DR, de Carvalho ACPLF, Zanotto ED. Predicting glass transition temperatures using neural networks. *Acta Mater*. 2018;159:249–256. https://doi.org/10.1016/j.actamat.2018.08.022
83. Ravinder R, Sridhara KH, Bishnoi S, *et al.* Deep learning aided rational design of oxide glasses. *Mater Horizons*. 2020;15(iii). https://doi.org/10.1039/D0MH00162G
84. Young BA, Hall A, Pilon L, Gupta P, Sant G. Can the compressive strength of concrete be estimated from knowledge of the mixture proportions?: New insights from statistical analysis and machine learning methods. *Cem Concr Res*. 2019;115:379–388. https://doi.org/10.1016/j.cemconres.2018.09.006
85. Rimstidt JD, Barnes HL. The kinetics of silica-water reactions. *Geochim Cosmochim Acta*. 1980;44(11):1683–1699. https://doi.org/10.1016/0016-7037(80)90220-3
86. Douglass I, Harrowell P. Kinetics of dissolution of an amorphous solid. *J Phys Chem B*. 2018;122(8):2425–2433. https://doi.org/10.1021/acs.jpcb.7b12243
87. Wolff-Boenisch D, Gislason SR, Oelkers EH. The effect of crystallinity on dissolution rates and CO2 consumption capacity of silicates. *Geochim Cosmochim Acta*. 2006;70(4):858–870. https://doi.org/10.1016/j.gca.2005.10.016
88. Perez A, Daval D, Fournier M, Vital M, Delaye J-M, Gin S. Comparing the reactivity of glasses with their crystalline equivalents: The case study of plagioclase feldspar. *Geochim Cosmochim Acta*. 2019;254:122–141. https://doi.org/10.1016/j.gca.2019.03.030

89. Criscenti LJ, Kubicki JD, Brantley SL. Silicate Glass and Mineral Dissolution: Calculated Reaction Paths and Activation Energies for Hydrolysis of a Q 3 Si by H 3 O + Using Ab Initio Methods. *J Phys Chem A*. 2006;110(1):198–206. https://doi.org/10.1021/jp044360a

90. Duffy JA, Ingram MD. An interpretation of glass chemistry in terms of the optical basicity concept. *J Non Cryst Solids*. 1976;21(3):373–410. https://doi.org/10.1016/0022-3093(76)90027-2

91. Duffy JA. Optical basicity: A practical acid-base theory for oxides and oxyanions. *J Chem Educ*. 1996;73(12):1138. https://doi.org/10.1021/ed073p1138

92. Mauro JC. Topological constraint theory of glass. *Am Ceram Soc Bull*. 2011;90(4):31–37.

93. La Plante EC, Oey T, Hsiao Y-H, Perry L, Bullard JW, Sant G. Enhancing silicate dissolution kinetics in hyperalkaline environments. *J Phys Chem C*. 2019;123(6):3687–3695. https://doi.org/10.1021/acs.jpcc.8b12076

94. Ehrenberg A, Romero Sarcos N, Hart D, Bornhöft H, Deubener J. Influence of the Thermal History of Granulated Blast Furnace Slags on Their Latent Hydraulic Reactivity in Cementitious Systems. *J Sustain Metall*. 2020;6(2):207–215. https://doi.org/10.1007/s40831-020-00269-4

95. Richet P, Bottinga Y. Thermochemical properties of silicate glasses and liquids: A review. *Rev Geophys*. 1986;24(1):1. https://doi.org/10.1029/RG024i001p00001

96. Shen J, Green DJ, Tressler RE, Shelleman DL. Stress relaxation of a soda lime silicate glass below the glass transition temperature. *J Non Cryst Solids*. 2003;324(3):277–288. https://doi.org/10.1016/S0022-3093(03)00260-6

97. Angeli F, Villain O, Schuller S, *et al.* Effect of temperature and thermal history on borosilicate glass structure. *Phys Rev B*. 2012;85(5):054110. https://doi.org/10.1103/PhysRevB.85.054110

98. Guo X, Potuzak M, Mauro JC, Allan DC, Kiczenski TJ, Yue Y. Unified approach for determining the enthalpic fictive temperature of glasses with arbitrary thermal history. *J Non Cryst Solids*. 2011;357(16–17):3230–3236. https://doi.org/10.1016/j.jnoncrysol.2011.05.014

99. Tsuyuki N, Koizumi K. Granularity and surface structure of ground granulated blast-furnace slags. *J Am Ceram Soc*. 1999;82(8):2188–2192.

100. Kleiv RA, Thornhill M. Mechanical activation of olivine. *Miner Eng*. 2006;19(4):340–347. https://doi.org/10.1016/j.mineng.2005.08.008

101. Kumar S, Kumar R, Bandopadhyay A, *et al.* Mechanical activation of granulated blast furnace slag and its effect on the properties and structure of portland slag cement. *Cem Concr Compos*. 2008;30(8):679–685. https://doi.org/10.1016/j.cemconcomp.2008.05.005

102. Kumar S, Kumar R. Mechanical activation of fly ash: Effect on reaction, structure and properties of resulting geopolymer. *Ceram Int*. 2011;37(2):533–541. https://doi.org/10.1016/j.ceramint.2010.09.038

103. Kriskova L, Pontikes Y, Cizer Ö, *et al.* Effect of mechanical activation on the hydraulic properties of stainless steel slags. *Cem Concr Res*. 2012;42(6):778–788. https://doi.org/10.1016/j.cemconres.2012.02.016

104. Kriskova L, Pontikes Y, Zhang F, *et al.* Influence of mechanical and chemical activation on the hydraulic properties of gamma dicalcium silicate. *Cem Concr Res*. 2014;55:59–68. https://doi.org/10.1016/j.cemconres.2013.10.004

105. Kim E, Huang K, Saunders A, McCallum A, Ceder G, Olivetti E. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem Mater*. 2017;29(21):9436–9444. https://doi.org/10.1021/acs.chemmater.7b03500

106. Kim E, Huang K, Tomala A, *et al.* Machine-learned and codified synthesis parameters of oxide materials. *Sci Data*. 2017;4(1):170127. https://doi.org/10.1038/sdata.2017.127

107. Jensen Z, Kim E, Kwon S, *et al.* A machine learning approach to zeolite synthesis enabled by automatic literature data extraction. *ACS Cent Sci*. 2019;acscentsci.9b00193. https://doi.org/10.1021/acscentsci.9b00193

108. Kingma DP, Rezende DJ, Mohamed S, Welling M. Semi-supervised learning with deep generative models. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, eds. *Adv. Neural Inf. Process. Syst. 27*. Vol. 4. Curran Associates, Inc.; 2014:3581–3589.

109. Kingma DP, Welling M. Auto-Encoding Variational Bayes. *arXiv.org*. 2013.

110. Nazábal A, Olmos PM, Ghahramani Z, Valera I. Handling incomplete heterogeneous data using VAEs. *Pattern Recognit*. 2020;107:107501. https://doi.org/10.1016/j.patcog.2020.107501

111. Daux V, Guy C, Advocat T, Crovisier J-L, Stille P. Kinetic aspects of basaltic glass dissolution at 90°C: role of aqueous silicon and aluminum. *Chem Geol*. 1997;142:109–126. https://doi.org/10.1016/S0009-2541(97)00079-X

112. Jollivet P, Gin S, Schumacher S. Forward dissolution rate of silicate glasses of nuclear interest in clay-equilibrated groundwater. *Chem Geol*. 2012;330–331:207–217. https://doi.org/10.1016/j.chemgeo.2012.09.012

113. Ebert WL. The effects of the glass surface area/solution volume ratio on glass corrosion: A critical review. Los Alamos, NM: 1995 https://doi.org/10.2172/67461

114. Gudbrandsson S, Wolff-Boenisch D, Gislason SR, Oelkers EH. An experimental study of crystalline basalt dissolution from 2⩽pH⩽11 and temperatures from 5 to 75°C. *Geochim Cosmochim Acta*. 2011;75(19):5496–5509. https://doi.org/10.1016/j.gca.2011.06.035

Figure Captions:

FIGURE 1 Schematic diagram of our semi-supervised learning analysis that spans multiple domains of data. Unlabeled and labeled data (X) were first entered into a probabilistic encoder Q(Z|X) that maps to a space of latent feature representation (Z), which were re-directed to a probabilistic decoder P(X|Z) that generates a standardized

data summarization (X*). The distributional discrepancy between X and X* (conditioned on Z) is then used as an unsupervised loss on information summarization while the latent representation Z that corresponds to the labeled data is used to optimized for a latent predictor that maps from Z to the corresponding outcome Y. All component parameterizations in the above pipeline are customized following the guiding principles in Kingma (2014),[108] which can be optimized end-to-end via gradient backpropagation.

FIGURE 2 Ternary SiO_2-Al_2O_3-CaO diagram depicting different material categories of interest to the cement community. All chemistry and category data extracted from tables in the literature. The numbers below each category label reflect the number of samples over the number of unique DOIs (*i.e.*, Samples/DOIs). Data points shown in yellow are as yet unlabeled, reflecting an additional 15,500 samples.

FIGURE 3 (A) Cumulative density functions of normalized oxide content (SiO_2, Al_2O_3, and CaO) for all extracted samples; (B)-(C) Frequency density of normalized oxide content of cement species (n = 3,133 from 2,592 DOIs) and fly ash species (n = 725 from 496 DOIs), respectively. The sum of the three oxides (SiO_2, Al_2O_3, and CaO) was normalized to 100%.

FIGURE 4 (A) Plot of log rate as a function of pH for all samples extracted from the literature (n = 802). Different colors represent distinct DOIs. (B) Same plot with overlaid thermal information (low temperature = darker, high temperature = lighter). Trendlines are schematic to show opposing trends in acidic and basic regions.

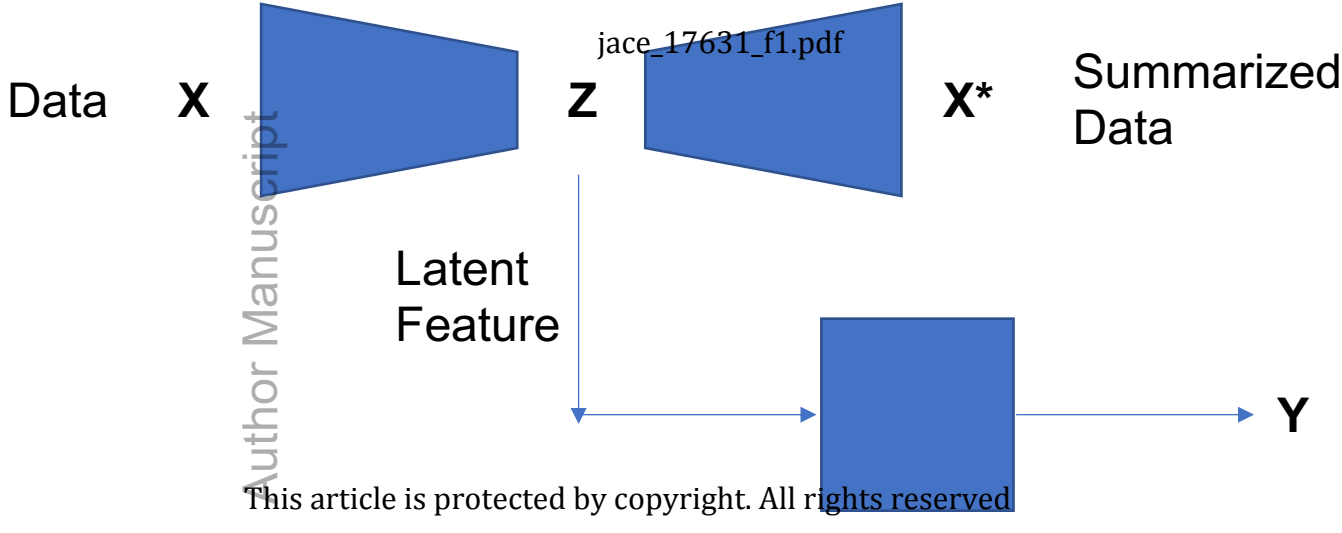
FIGURE 5 Decision tree regression trained on samples from [29, 30, 33, 34, 36, 44-47, 52–54, 57, 60, 61, 63, 65, 66, 68–74] (n = 636) and all features as listed in Tables S3 (excluding roughness factor and SA/FR, as described in text). Testing the regression on samples from [35, 48, 50, 55, 56, 62, 64, 67] (n = 166) with a max depth of 3 (optimized) yielded RMSE = 1.24 and R^2 = 0.40.

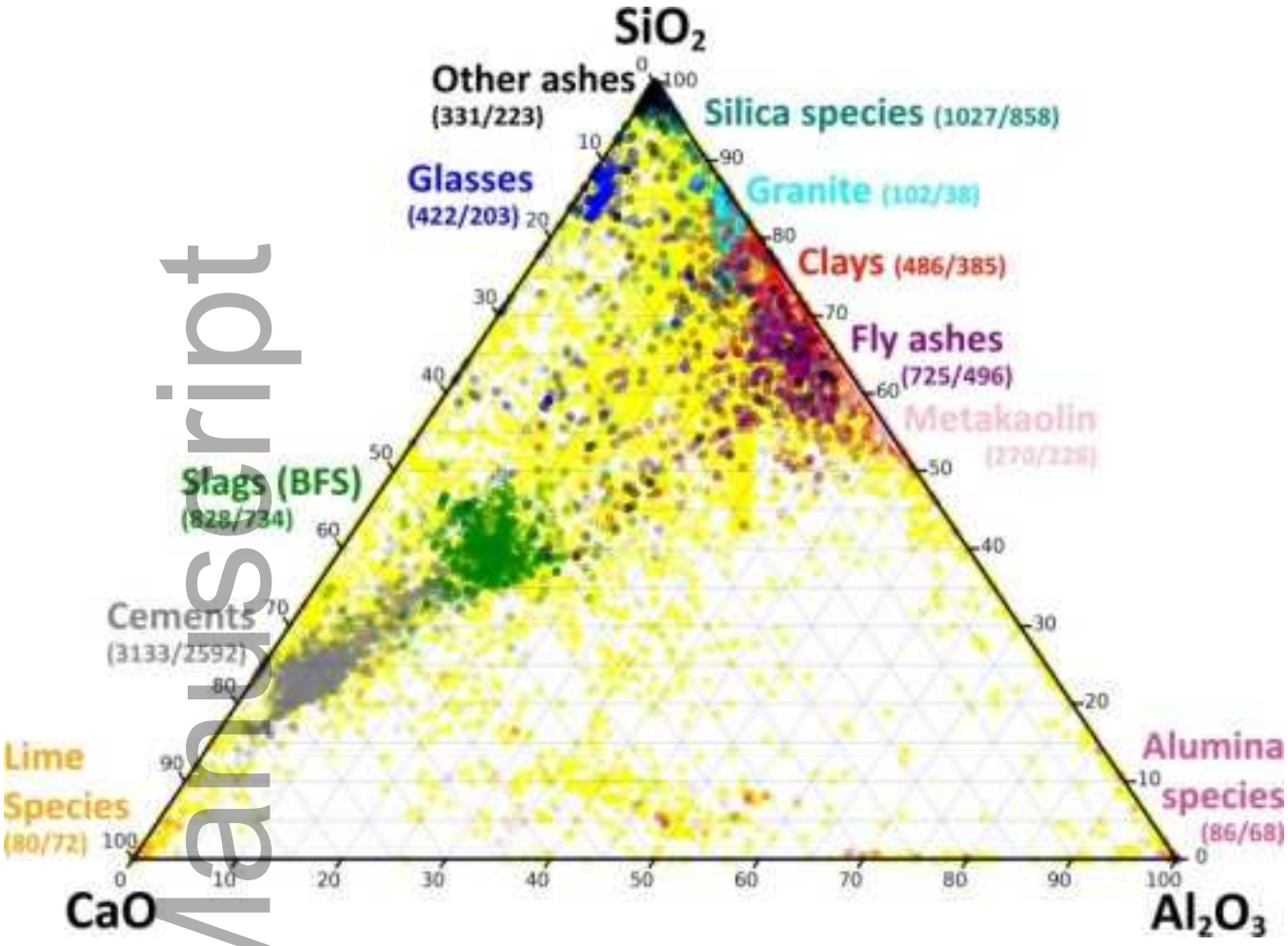
FIGURE 6 Linear regression trained on all samples from [29, 30, 33, 34, 36, 44-47, 52–54, 57, 60, 61, 63, 65, 66, 68–74] and tested on samples from [35, 48, 50, 55, 56, 62, 64, 67]. (A) Train (n = 636) and test (n = 166) sets included all samples and resulted in respective RMSE values of 0.95 and 1.53; (B) Train (n = 442) and test (n = 116) sets included samples from experiments with pH $\geq$ 7 and resulted in respective RMSE values of 0.66 and 0.97.

FIGURE 7 Graphs of predicted vs true values of $\log_{10}$(dissolution rate) for data with pH $\geq$ 7. In (A) the model is trained on a single data set as listed (cement, nuclear, or geochem), while in (B) the opposite is true—the model is trained on 2/3 sets and tested on the third set as listed. In (C), the model is trained and tested on B_2O_3-containing

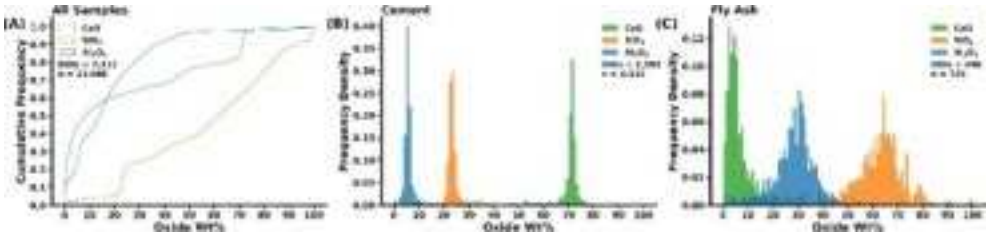
or -absent glasses as labeled. In (D) the model is trained and tested on batch or flow experiments as labeled. Number of samples in training set (n) and RMSE scores of prediction on test sets are listed on each plot.

FIGURE 8 (A) RMSE of machine learning models as a function of number of training epochs. Input consisted of full feature set as defined by Table S3, with size of training set varied as labeled. (B) RMSE of machine learning models, where input consisted of only 3 key features (pH, 1/temperature (1/K), and NBO/T) and size of training set varied as labeled. (C) Comparison between RMSE of four input sets: (1) All features, all samples, (2) 3 features, all samples, (3) All features, samples with pH $\geq$ 7, and (4) 3 features, samples with pH $\geq$ 7

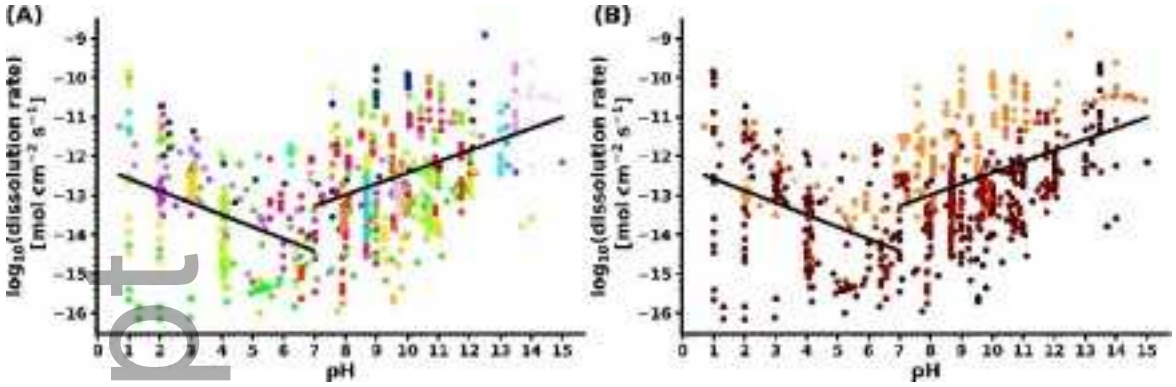




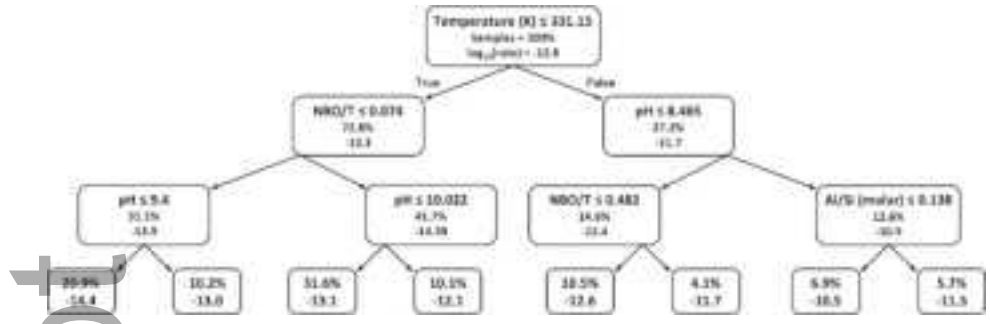
jace_17631_f2.png



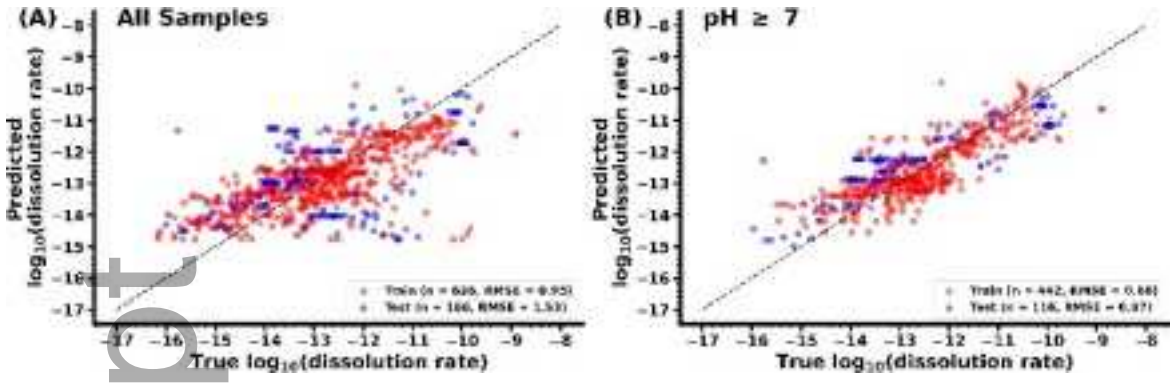
jace_17631_f3.png



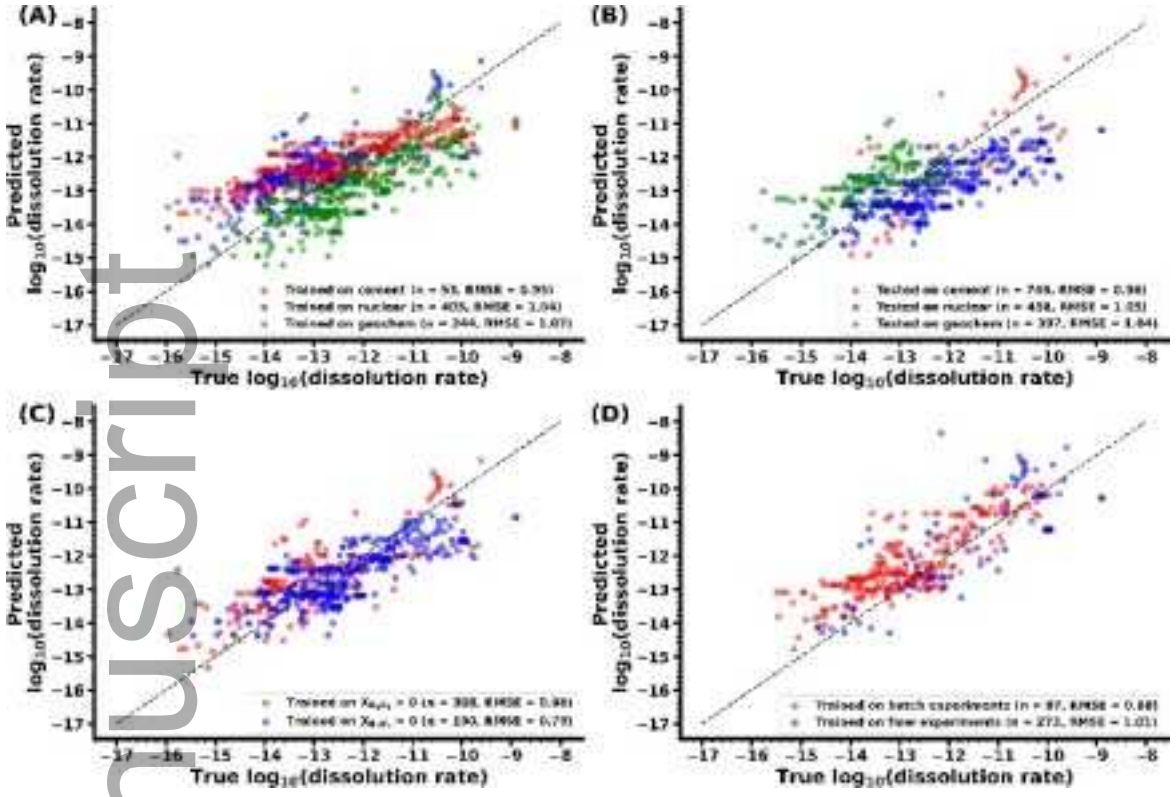
jace_17631_f4.png



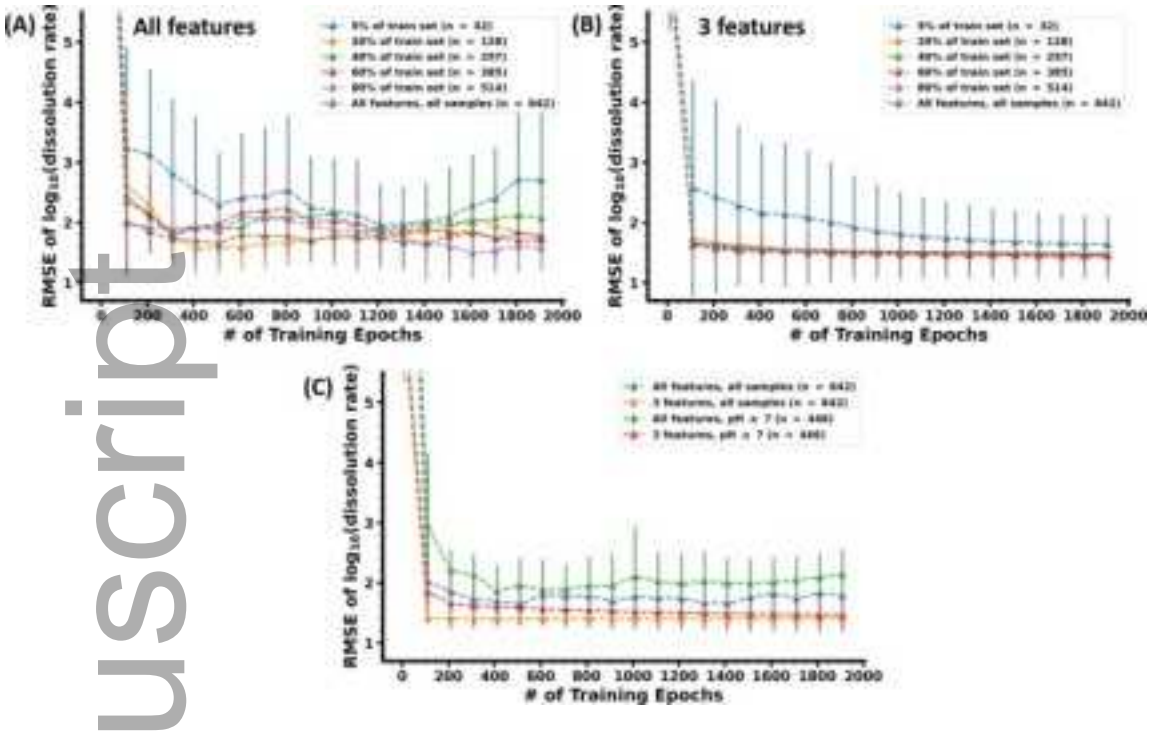
jace_17631_f5.png



jace_17631_f6.png



jace_17631_f7.png



jace_17631_f8.png