RILEM Technical Letters (2021) 6: 140-149
https://doi.org/10.21809/rilemtechlett.2021.147

Advancing cement-based materials design through data science approaches

Renee T. Rios[1], Christopher M. Childs[2], Scott H. Smith[1], Newell R. Washburn[2]*, Kimberly E. Kurtis[1]*

[1] School of Civil and Environmental Engineering, Georgia Institute of Technology, USA
[2] Department of Chemistry, Carnegie Mellon University, USA

Received: 26 July 2021 / Accepted: 14 December 2021 / Published online: 30 December 2021

Abstract

The massive scale of concrete construction constrains the raw materials' feedstocks that can be considered – requiring both universal abundance but also economical and energy-efficient processing. While significant improvements– from more efficient cement and concrete production to increased service life – have been realized over the past decades through traditional research paradigms, non-incremental innovations are necessary now to meet increasingly urgent needs, at a time when innovations in materials create even greater complexity. Data science is revolutionizing the rate of discovery and accelerating the rate of innovation for material systems. This review addresses machine learning and other data analytical techniques which utilize various forms of variable representation for cementitious systems. These techniques include those guided by physicochemical and cheminformatics approaches to chemical admixture design, use of materials informatics to develop process-structure-property linkages for quantifying increased service life, and change-point detection for assessing pozzolanicity in candidate supplementary cementitious materials (SCMs). These latent variables, coupled with approaches to dimensionality reduction driven both algorithmically as well as through domain knowledge, provide robust feature representation for cement-based materials and allow for more accurate models and greater generalization capability, resulting in a powerful design tool for infrastructure materials.

Keywords: Cement; Pozzolan; Machine learning; Statistical analysis; Chemical admixture

1 Introduction

The increasing rate of concrete placement in recent decades – now at nearly 3 tonnes/per person/year – has intensified pressure to further reduce the environmental impacts of this essential material while providing growth in global infrastructure necessary to meet the needs of a growing and increasingly affluent population [1]. The massive scale of concrete construction constrains the raw materials' feedstocks that can be considered – requiring both universal abundance but also economical and energy-efficient processing. This ubiquity and the necessity of concrete infrastructure prompts the need for increasing innovation to address the global challenge of meeting societal needs in the most sustainable and economical ways possible. While significant improvements– from more efficient cement and concrete production to increased service life – have been realized over the past decades through traditional research paradigms, non-incremental innovations are necessary now to meet increasingly urgent needs, at a time when innovations in materials create even greater complexity.

Transition from approaches based on experience and intuition to data-guided approaches in advancing engineering systems is critical to accelerating the pace of innovation. Along with reducing costly and time-consuming iterative testing, data science can resolve economic and sustainability constraints. Data-driven engineering can also optimize conditions to address challenges in engineering such as producing highly sustainable materials and improvements in urban infrastructure [2]. Machine learning (ML) represents a diverse collection of powerful algorithms utilized to identify relationships in data, allowing for modeling and optimization of complex systems.

Early ML approaches in cement and concrete research came through the utilization of artificial neural networks (ANNs), a "black box" algorithm that uses statistical inference in the form of a series of layers with weight-optimized nodes to establish relationships between composition (input features) and material properties (objectives). In 1998, Yeh et al. [3] applied an ANN on a collection of 1030 concrete samples from 17 data sources, where the compositional proportions of cement, supplementary cementitious materials (SCMs), aggregate, water, age, and admixture were features used to predict compressive strength. With a coefficient of determination (R^2) slightly higher than 0.90 on both the training and testing sets, the model outperformed traditional

* Corresponding authors: Kimberly E. Kurtis, E-mail: kkurtis@gatech.edu, Newell R. Washburn, E-mail: washburn@andrew.cmu.edu

regression analysis. Numerous ML algorithms have since been applied to this published dataset comparing how well their algorithm can predict compressive strength [4].

In 2001, Haj-Ali et. al [5] applied an ANN to predict sulfate-induced concrete expansion as a function of water to cement ratio (w/c), cement C_3A content, and time on a historical dataset of over 8,000 points from 51 different mixtures, tested over 40 years. Although the model improved upon existing analytical equations, it predicted expansion values as much as 75% lower than the actual expansion values when applied to the test set. While an approach to improving ML algorithms is to increase the sample size, in the case of long-term or historical data, such approaches are impractical. This shows the limitations of data-centric approaches, where a model can only perform well if it is tasked with predicting behavior already displayed in the trained data set. The emergence of new materials, which lack historical performance data, highlights the importance for these models to incorporate physical and chemical features outside the traditional compositional space.

The commonality among these approaches is that the input variables are all based on the compositional space of the cement-based materials. Training on the compositional space hinders both interpretability and generalizability of the model. For example, while training on the Yeh dataset, Dutta et. al [6] found a Gaussian process regression (GPR) model had a coefficient of correlation (R) on the test set of 0.95, a root mean square error (RMSE) of 0.06375, and a mean absolute error (MAE) of 0.04292. While these are typical values of a high-quality model, sensitivity analysis indicated that 'cement content' was the most important factor in determining compressive strength, which was already well-known. This observation, however, provides no microstructural or chemical insight into the development of strength, and the model could not be used to predict changes due to different cement sources. Further, current research remains focused on predicting concrete strength, even though the new information gained from these studies is marginal. Ouyang et. al [7] showed that the sample size of a dataset reaches a threshold after which there are marginal gains to the accuracy of the model for predicting compressive strength. While it is an encouraging find to discover that tens of thousands of samples are not needed to develop a more accurate model, their analysis found a maximum R^2 of only 0.62, a minimum mean absolute percentage error (MAPE) of 8.74%, and a minimum RMSE of 4.37 MPa in four models studied.

Apart from predicting compressive strength on Portland cement mix designs, others have modeled the effect that SCMs, particularly fly ash, have on hardened cementitious properties, such as electrical resistivity, compressive strength, chloride resistance, expansion caused by alkali-silica reaction (ASR), and ion diffusivity [8-15]. ML has also been used to attempt to screen the reactivity of fly ashes based on their network topology [16]. However, the supply of high-quality, ready to use fly ash is diminishing as energy supply shifts from coal-fired power plants. As a result, the usefulness of these models is limited to however long these fly ashes remain a main source of SCMs in construction. Cement-based materials can greatly change their properties due to slight changes in the feature space, such as when a new SCM is introduced. These models are inherently unable to generalize to other classes of SCMs, such as clay-containing materials, which are one of the only SCMs available to meet the long-term demands of SCMs in the concrete industry [17]. Because of this, models that incorporate a wide range of potential SCMs are needed to truly be useful in the future design of cementitious structures.

A central goal in cementitious materials is the ability to design them to meet a diverse – and often competing - set of performance criteria using mineral feedstocks characterized by specific composition, particle size, and reactivity that is of function of these parameters but also a function of processing, as well as their combined use and the availability of water and other reactants. Algorithms parameterized by these variables would be specific to the materials in the training set as would the predictions. Design approaches that are based on machine learning but generalizable across a disparate range of feedstocks are critically needed. While traditional ML can successfully model datasets consisting of 10,000's of unique samples, smaller datasets require embedded domain knowledge to improve ML modeling [18]. Although many ML models for cement-based materials have developed and trialed new algorithms for property prediction, one of the most important factors in developing a successful ML algorithm is domain-specific feature engineering [19]. This paper reviews emerging ML and statistical approaches for predicting performance of cementitious systems relying on smaller experimental data sets and feature representation informed by domain knowledge of underlying chemistry, physics, and engineering, which represent a significant advantage to training over existing approaches.

Domain-specific feature engineering can be used to represent cement-based materials, as shown in Figure 1. To demonstrate the range of applications of this approach, four example investigations are reviewed here. First, an Hierarchical Machine Learning (HML) representation for workability allows for the design of a superplasticizer specifically for metakaolin-modified cement. Second, a cheminformatics representation of chemical admixtures for calcium sulfoaluminate (CSA) allows for the virtual screening and identification of new set retarders. Third, linkages between property, structure, and performance (PSP) are utilized to develop microstructural understanding of diffusivity in cement pastes, relating composition to durability. Finally, change-point detection is used to statistically determine if a candidate material undergoes a pozzolanic reaction.

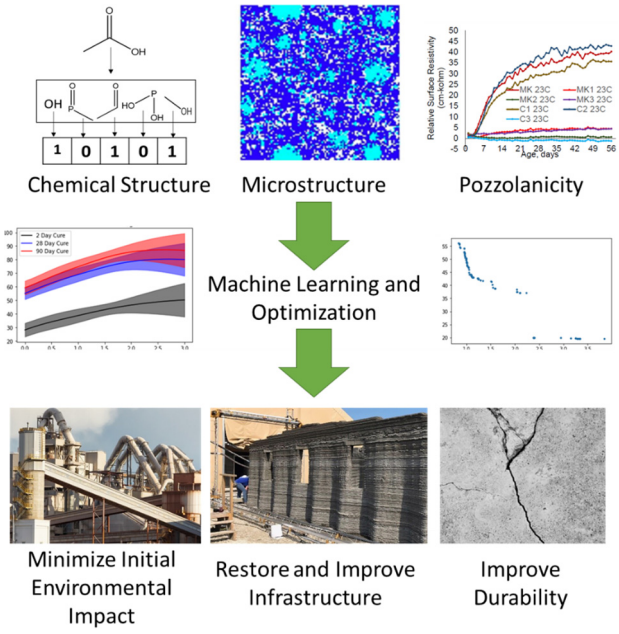

Figure 1. Latent variable representation of cement-based materials (top) can be modeled and optimized through data science techniques (middle) to develop non-incremental innovations and meet global goals for sustainable development (bottom).

2 Latent variables using HML

In HML, compositional parameters are represented in the bottom layer of the model and are related to a middle layer of physics-informed, latent features. This embedding of physicochemical equations allows HML to learn on system properties as opposed to extensive testing over broad compositional spaces, thus greatly reducing data requirements compared to conventional ML [20, 21]. For example, Bone et. al [22] utilized HML to physically relate ink concentration and print parameters to viscosity, shear rate, and proportionality in a model to predict and optimize print fidelity in 3D printed biopolymers. Similarly, Menon et. al [23] predicted the Young's modulus of polyurethanes through relating the molecular composition to a middle layer of physicochemical properties utilizing stochastic simulation and molecular modeling.

A major challenge in incorporating minimally processed minerals, such as calcined clays, in cementitious binders is understanding and addressing their tendency to decrease workability. Clay behavior in suspension is sharply affected by solution concentration [24], such as the pore solution in cement-based materials. As a result of their physical and chemical characteristics, calcined clays produce sharp reductions in flow, which can alter not just rheology but also hydration kinetics, microstructure, and durability [25]. Each are, in turn, affected by chemical admixtures; if those admixtures are not tuned to interact with calcined clays – i.e., direct application of admixtures developed for ordinary Portland cement (OPC) – unintended admixture-mineral interactions occur, including adsorption and absorption within the clay structure, leading to unpredictable variations in important qualities like mineral dispersion, kinetics of cement hydration and pozzolanic reaction, set time and strength development [26, 27].

To bridge the relationships between experimental variables and system properties, solution-based forces (viscosity η and osmolality π) and particle-based forces (electrostatic ζ and electrosteric s) as well as the coupling between solution and particle forces, are determined by the extent of adsorption θ and the polymer chemistry. These form the middle layer. Note that in this study [28], the cement, mineral, and water variables were all held fixed and the only free variables were those of polymer chemistry, as shown in Figure 2.

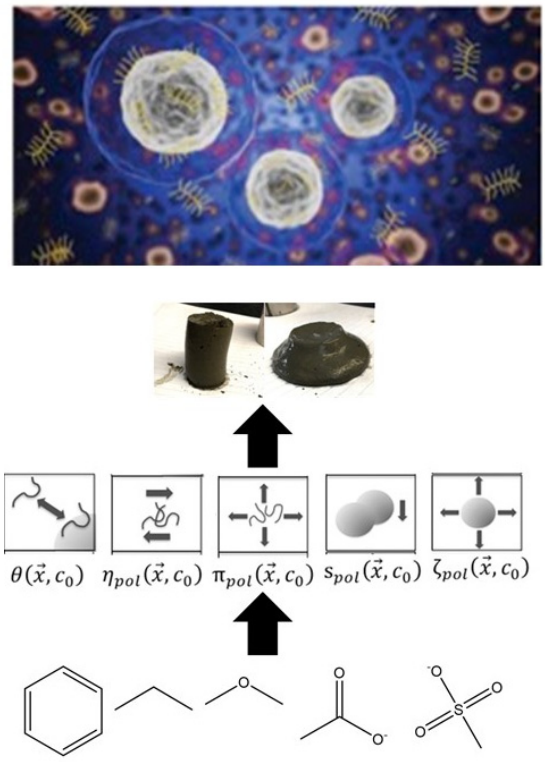

Figure 2. Illustration of interactions of superplasticizers in cement paste depicting both adsorbed and free polymers (top). HML representation of how dispersant chemistry determines the flow, or workability, of cement paste, including effects on both particle and solution characteristics (bottom). Note that in this preliminary study, the cement, mineral, and water variables were all held constant so the final model is only parameterized by polymer chemistry. The HML algorithm requires knowledge of all forces and surrogate physical measurements to provide estimates of their values. This allows it to make accurate predictions based on extremely small training sets but can make it challenging to implement for highly complex materials.

The approach in HML is to treat these underlying forces as latent variables but estimate their effects through additional experimental measurements or theoretical models. Based on estimates of these forces in polymer solutions or dilute suspensions, the HML algorithm employs statistical learning to determine which forces and combinations of forces are responsible for the variations in material properties. As a proof-of-concept, this design approach was attempted for superplasticizing chemical admixtures for blends of relatively

pure calcined clay (i.e., commercially sourced metakaolin) and OPC. A training set composed of behavior of the interaction of pastes and slurries with just seven commercial water-reducing admixtures was used to train the algorithm. Polymers in this training set were the solid polymer products after dialyzation and lyophilization of commercial superplasticizers or dispersants synthesized via radical polymerization. The flow (*S*) of the paste was expressed as a 2nd-order polynomial in the variables representing the underlying forces using the Least Average Shrinkage and Selection Operator (Lasso):

$$S_{MK-PC} = 1.11\eta - 0.55\zeta + 0.36\zeta s_{MK} + 0.12\eta s_{PC} \quad (1)$$

Interpretability of the predictions is an important feature of HML, which is accomplished through analysis of the parameterization of the system properties in terms of the underlying forces. Here, the term with the largest coefficient, which the models predicted as dominating the response, was the effects of the polymer on solution viscosity η, with the positive sign indicates that increasing the viscosity would result in an increase in the flow of the paste, a result that appears counterintuitive. (Other important forces were the effects on particle zeta potential ζ and the electrosteric interactions involving metakaolin s_{MK} and Portland cement s_{PC}.) However, when these forces were represented in terms of their composition dependences, a global maximum in the flow of the paste was predicted to occur with a polymer having a complex combination of chemical functional groups, which could then be synthesized and tested.

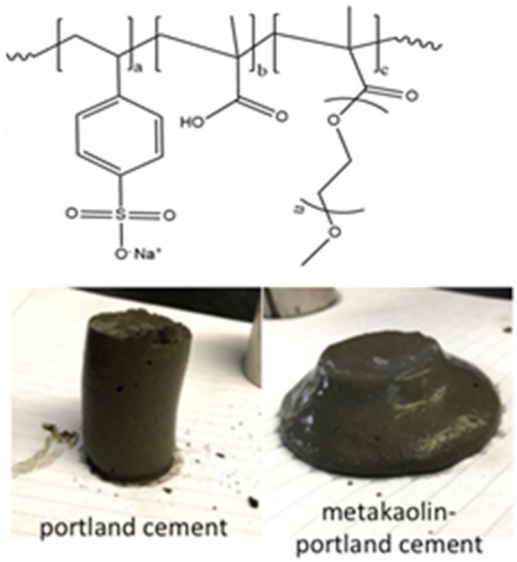

Figure 3. The random copolymer based on SS, MAA, and PEGMA predicted by HML to maximize the slump of MK-OPC cement paste (top). This superplasticizer was found to be specific for MK-OPC blends and did not significantly plasticize OPC paste (bottom), unlike other PCE superplasticizers whose effects displayed the opposite trends from OPC to MK-OPC.

The outcome of this work is that the trained HML algorithm predictions were used to guide the synthesis of the random copolymer with a molar composition of 50% styrene sulfonate (SS), 25% methacrylic acid (MAA), and 25% poly(ethylene glycol) methacrylate (PMA). The results are depicted in Figure 3 [28]. It had three unexpected attributes:

1. It was specific for metakaolin-Portland cement blends, improving their flowability, but had no effect on the workability of ordinary cement pastes.
2. Only 8-10% of the polymer adsorbed to the particle surfaces, indicating that its mechanism was not based on tuning particle-particle interactions.
3. It had an intrinsic viscosity that was nearly 20× greater than PCE, suggesting its mechanism of plasticization was through solution-based forces, in contrast to conventional water-reducers developed for OPC systems.

This application demonstrates that while HML can provide novel predictions and physical insight in optimizing complex systems based on small datasets, it requires analytical representation of all latent variables that drive system responses and surrogate physical measurements to estimate them for the model to be effective.

3 Cheminformatics approaches

Through efficiently encoding the molecular architecture of admixtures, cheminformatics allows for the comparison of similarity in molecular structure to function, and, as demonstrated in this application, this approach can be used to streamline chemical admixture development. Cheminformatics is a methodology to represent chemical structure as a vector, which in turn can be related to function, facilitating design of molecules with intended functionality. Cheminformatics approaches have been utilized for such tasks as predicting the glassy transition temperature of polymers [29, 30], drug discovery [31], and improving quantum mechanical calculations for molecules [32]. A typical representation is shown in Figure 4, where each group on the molecule is assigned a binary integer to represent if the group is present or not. Extensive folding (where the encoded information on each bit is increased) and filtering (an algorithmic reduction in the feature space, such as through the utilization of Lasso Regression) methodologies can be applied in order to reduce the feature space to be less than the number of tested samples(i.e., 23 distinct admixtures) [33].

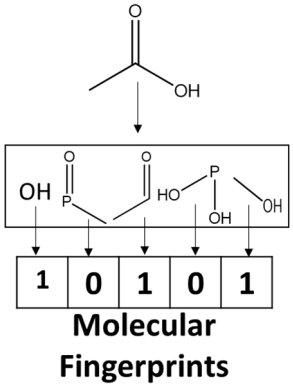

Figure 4. Representation of a binary fingerprint for utilization in cheminformatics.

While OPC has been the most widely used cement historically, alternative binder chemistries are increasingly important in finding innovative solutions for sustainable development and new admixtures will be needed to broaden their practical use. Calcium sulfoaluminate (CSA) cements contain no tricalcium silicates in clinker, reducing the embodied CO_2 compared to OPC, but leading to rapid reactivity, including very short set time. To prolong CSA hydration, set-retarding admixtures are commonly introduced to the formulation. Set-retarding admixtures include a diverse range of chemistries which require innovative methodologies to handle both the feature representation of these molecules, along with availability of only sparse datasets. The leading retarder for CSA cements is citric acid, which here extended the set time 165 minutes at a dose of 1% by weight of cement, but its cost has prevented broader adoption. Identification of alternative retarding admixtures by chemiformatics exemplifies the utility of this approach in accelerating innovation in a complex, and incompletely understood, cementitious system.

In this application, a library of small-molecule retarders that featured carboxylate, sulfonate, and phosphono chemistries was tested to assess their effect on hydration kinetics and set time. In the training set, it was observed that species with combinations of anionic and polar functional groups could impart retardation times similar to that of citric acid. Molecular fingerprints of the compounds in the training set were used to train predictive models, and it was found that the extended connectivity fingerprint (ECFP) provided the most accurate predictions with an R^2 of 0.98 and an RMSE of 26 min. While not commercially available, the model predicted that (3-(formylhydroxyamino)-1-propenyl) phosphonic acid would have a set time of 183 minutes. The commercial compound glyphosate was also in the library and predicted to have a set time of 61 +/- 26 min. Experimentally, glyphosate imparted a set time of 55 minutes at a cost that is competitive with citric acid.

This cheminformatics methodology can either virtually screen existing molecules with constraints on cost (or other criteria) or guide the design of novel structures to produce more efficient admixtures. Although the research presented here was designed for a specific CSA system for only carboxylate, sulfonate, and phosphono chemistries, the same principles can be extended to other types of cements, blended cements and other admixtures. For example, cheminformatics has been applied in the molecular design of shrinkage reducing admixtures for Portland cement [34], and can be extended to superplasticizers, viscosity modifiers, or corrosion inhibitors, among many others, including in a broader range of binder compositions.

4 Process-structure-property linkages

To connect materials science to engineering properties, process-structure-property (PSP) linkages can be used. Material informatics workflows allow for the identification and quantification of microstructural latent variables which influence the material properties of interest [35]. Once identified and quantified, the most salient material parameters that correlated with material properties can be used in the creation of PSP linkages. Similar to the current state of modeling cement-based materials, advanced structural metals had been represented as inputs through elemental composition and phase volume fractions. However, these 1-point statistics do not account for the significant effects of the surrounding local microstructures. Multiphase systems are now increasingly accounted for through 2-point, or n-point, statistics to allow for PSP linkages [36].

The methodology in creating these process-structure-property linkages is shown in Figure 5. As explained by the single void dissolution kinetics (SVDK) model [37], the diffusivity property in cement pastes governs the rate of air void saturation. The SVDK model is used to describe the saturation kinetics of the air void and the significance of transport mechanisms in the microstructure. In particular, it can determine how water enters air voids for surrounding capillaries based on the diffusivity. As a result, diffusivity becomes the property of interest in the process-structure-property linkage approach in predicting time to critical saturation. Critical saturation is an important concept in understanding likelihood of damage due to freezing and thawing cycles in porous, brittle materials [38]. To accomplish the modeling of the diffusivity property, a database consisting of 349 samples of various water-to-cement ratios and hydration times as the processing parameters was modeled using NIST's Virtual Cement and Concrete Testing Laboratory Consortium (VCCTL) software [39]. Corresponding to each combination of the processing parameters, a diffusivity and a structure were simulated and collected [40]. The structure was evaluated using 2-point statistics and then the dimensionality was reduced using principal component analysis. The principal components were then used in a Gaussian process regression (GPR) model to infer the diffusivity. A GPR model was selected because the model form was initially unknown, the dataset was small, and a lack of availability of prior time-series information to inform the model [41].

The simulated microstructures were first simplified using tri-phase segmentation. As shown in Figure 5, instead of the 30+ phases present in the initial database structures, three phases were chosen to represent the microstructures: water/pores, hydration products, and unhydrated cement grains. Two-point statistics were used to determine spatial features in the microstructure that correlate strongly with the diffusivity property and is used to describe the structure in high dimensions. The 2-point statistics representation of the various microstructures output many more features than there were data points, creating a sparsity problem. To overcome this, the microstructural feature space was reduced using principal component analysis. The first three principal components were chosen to describe the 2-point statistics because they were able to retain 99.9% of the variance [41].

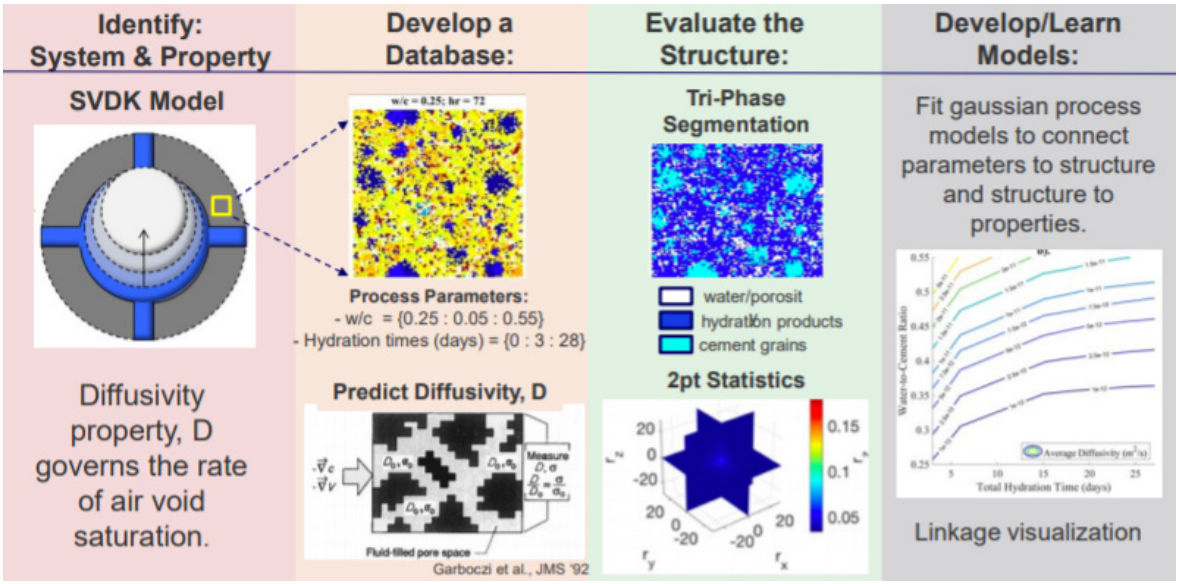

Figure 5. Workflow of the process-structure-property linkages to model time to critical saturation in cement pastes [41].

The first three principal components are plotted along with the two processing parameters in Figure 6 [41]. As seen in the figure, the three principal components that describe the microstructures are easily separable and distinguishable as a function of the processing parameters: water-to-cement ratio and hydration time.

Finally, the process-structure-property linkages were made using GPR models. The process parameters of water-to-cement ratio and hydration time were used to predict each of the three principal components. The three principal components were then used to predict the diffusivity property. As previously discussed, the predicted diffusivity property can be implemented in the SVDK model to give time to dissolution and then correlated with the controllable process parameters. Figure 7 [41] shows the outcome of the combination of these models where the time to full dissolution can be described as a function of the processing parameters. In particular, Figure 7 shows the crucial finding that for hydration times above 15 days, the time to saturation doubles for every 0.05 decrease in the water-to-cement ratio over the range of 0.25 to 0.55.

Process-structure-property relations can predict properties, and as a result, long-term material behavior. This research demonstrates how PSP linkages can be applied to the design of mixtures against freeze-thaw damage and how data driven approach can also be applied toward other concrete properties. Specifically, this work allows for the design of materials from the processing stage to be able to manufacture a desired property, which would prove extremely beneficial for the design of novel materials and mixtures durable in a range of environments.

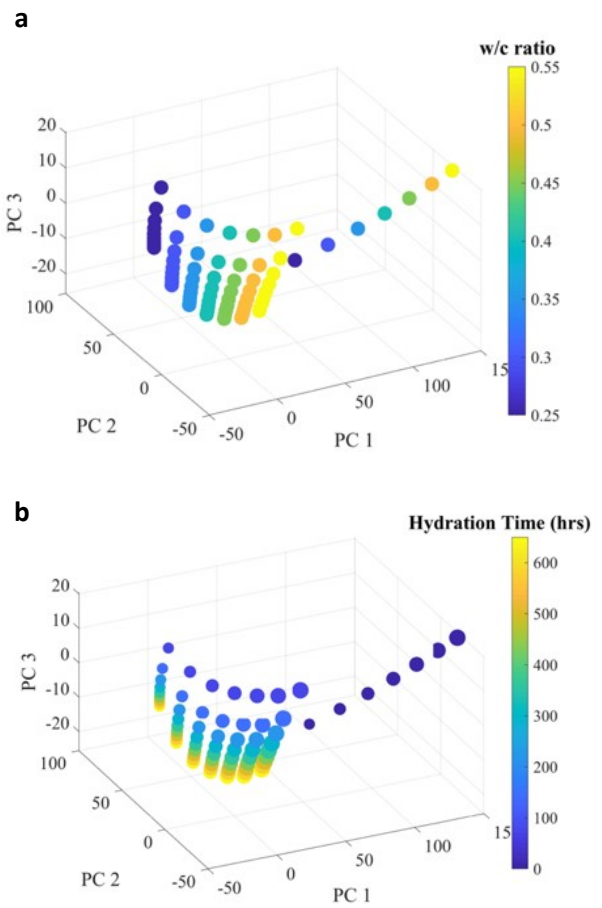

Figure 6. 3D visualization of the first three principal components to distinguish between process parameters (a) water-to-cement ratio and (b) total hydration time [41].

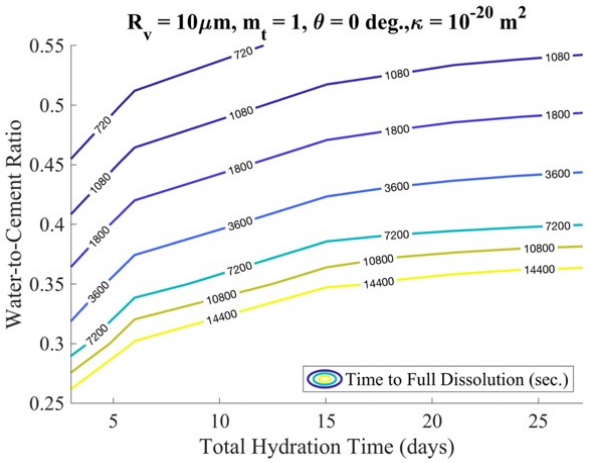

Figure 7. The influence of processing parameters on time to full dissolution using the SVDK model which shows that for hydration times above 15 days, the time to saturation doubles for every 0.05 decrease in the w/c ratio. The processing parameters used to describe the time to full dissolution are w/c ratio and total hydration time at a fixed radius of air void (R_v) = 10 um, a normalized trapped mass of gas (m_t) = 1, contact angle (θ) = 0 deg., and intrinsic permeability of hydrated cement paste (κ) = 10^{-20} m^2 [41].

5 Microstructure and Durability Representations

A novel method for rapidly and rigorously identifying reactive materials is through surface resistivity measurements, which in a previous study by Nadelman et. al [42] showed this approach to capture how variations in binder compositions using traditional pozzolans affect the microstructural development in concrete over time. This work was extended [43] to include ASTM C618 non-conforming materials and to both non-accelerated curing and accelerated curing, at 23°C and 38°C respectively. The elevated curing temperature was chosen to investigate if the onset of the pozzolanic reaction in Class F fly ashes, known to begin at around 28 days, could be accelerated. Figure 8 shows results for four concretes of the same mix design but with varying binder composition: one neat OPC mixture (OPC) and three binary blends with 20% cement replacement by mass by each of three SCMs (F, Y1, BH1) were tested. F is an ASTM C618 Class F fly ash, and Y1 and BH1 are 'off-spec' fly ashes that are reclaimed from ash ponds and do not meet C618 specifications. The premise is SR time-series data could be used to distinguish between intrinsic matrix densification caused by the presence of inert fine particles from densification resulting from pozzolanic reaction with portlandite, which varies over time depending on the type of pozzolan, by a change in the time-series' slopes. Figure 8 presents a dashed line approximating when this change in slope occurs, but to determine an unbiased and absolute assessment of the age at which the microstructure changes due to a pozzolanic reaction, the statistical technique change-point detection was applied to the time series data.

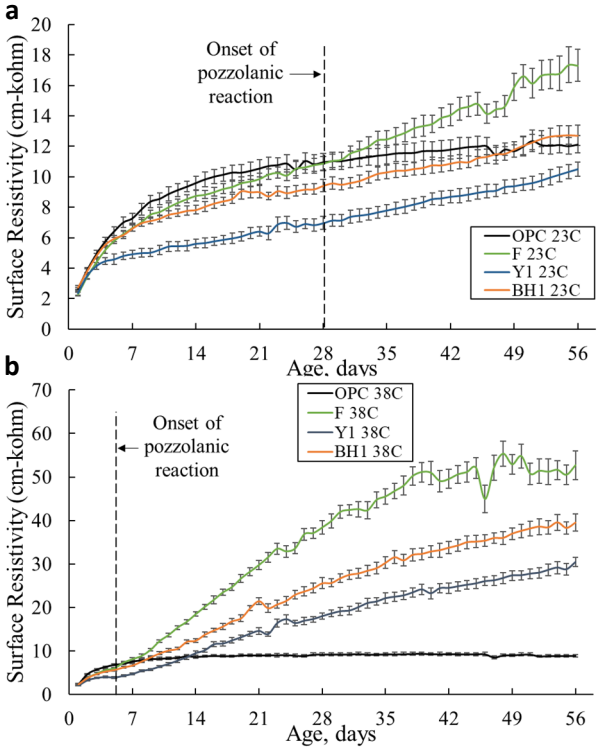

Figure 8. Time series surface resistivity data for four concrete blends at two curing regimes: (a) 23°C and (b) 38°C with estimated onsets of pozzolanic reactions. The blends had a 20% cement weight replacement by the candidate fly ashes. F is an ASTM C618 Class F fly ash, and Y1 and BH1 are off-spec fly ashes that are reclaimed from ash ponds.

The determination of the onset of the pozzolanic reaction in surface resistivity measurements as shown in Figure 9 has a subjective component, lacking any statistical analysis, which makes it challenging to implement in practice to identify new SCM sources. To overcome this, an approach known as slope change-point detection [44, 45] was applied on the processed data, shown in Figure 9a. The goal of slope change-point detection is to see if a stochastic process or time-series has changed, usually using measurable parameters such as the mean or variance. In this approach, surface resistivity for a concrete mixture was measured from 1 to 56 days, and the difference from that of the OPC mixture was calculated. The resulting vector $\tilde{y}_t = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_{56})$ denotes the residual surface resistivity values between OPC and a mixture's value in kohm-cm. This is shown for the concrete mixture containing F in Figure 9b.

The presence of a pozzolan is associated with the SR becoming greater than the SR of the OPC mix, which indicates reduced carrier concentration or mobility and can be associated with improved durability. In this analysis, to ensure that only positive change-points are considered, the residual surface resistivity vector $\tilde{y}_t$ is manipulated so that the resulting vector is $\tilde{y}_t^+ = \max(0, \tilde{y}_t)$, which is shown graphically for the concrete mixture containing F in Figure 9c.

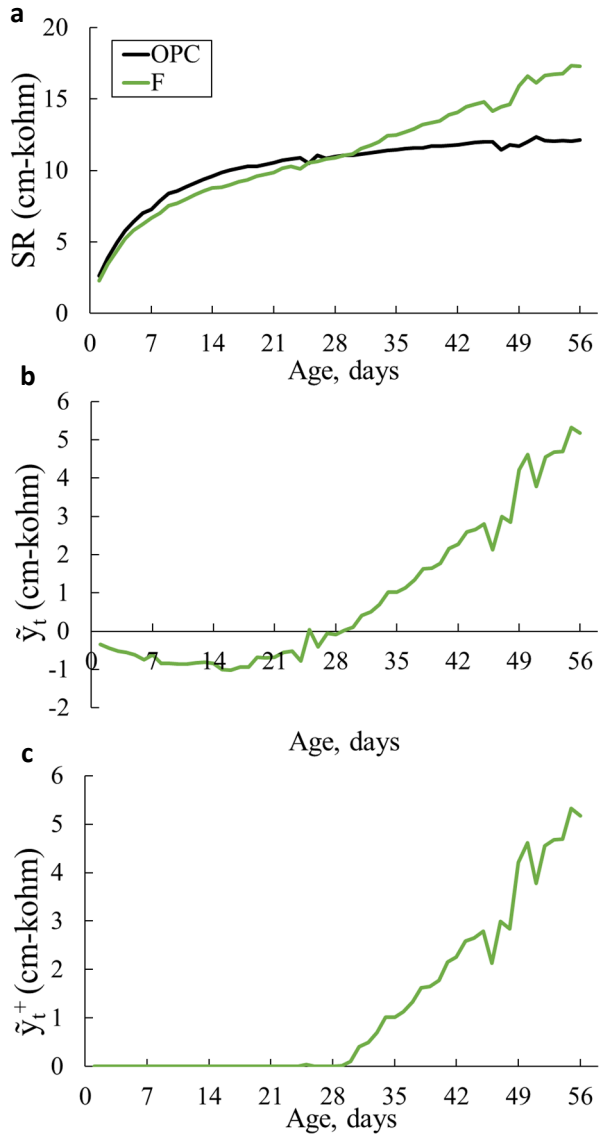

Figure 9. (a) Unprocessed surface resistivity time-series results for mixtures OPC and F, cured at 23°C. (b) Processed residual surface resistivity values between mixtures OPC and F. (c) Processed residual surface resistivity values from 1b to ensure that only positive change-points are considered for mixture F in change-point detection analysis [32].

The vector $\tilde{y}_t^+$ for each material is then used in the slope change-point detection procedure. Using this pre-processing procedure for the OPC mixture, $\tilde{y}_{t,PC}^+$ would be a 56-length vector consisting of zeros. With the pre-processed sequences $\tilde{y}_t^+$ for each material, the problem is reduced to determining a slope change from the sequence for the OPC mixture, $\tilde{y}_{t,PC}^+$, which would be the x-axis with slope equal to 0.

For each data point in each $\tilde{y}_t^+$ for each candidate material, the generalized likelihood ratio test (GLRT) was conducted. Using the GLRT approach, from one successive point to the next, the maximum likelihood ratio, $U_{k,t}$, was calculated as shown in Equation 2.

$$U_{k,t} = \frac{\left[\sum_{i=k+1}^{t}(i-k)\tilde{y}_i\right]^2}{\sum_{i=k+1}^{t}(i-k)^2} \quad (2)$$

A change-point has occurred if for a given sequence, $\max_{0<k<t} U_{k,t}$ is larger than a pre-set threshold value, b, which is determined through Monte Carlo simulations of the test statistic $U_{k,t}$. The t corresponding to the $\max_{0<k<t} U_{k,t} > b$ is the time, in days, of when the slope change-point has occurred. Using this procedure, the results of slope change-point detection are shown in Table 1.

Table 1. Results of slope change-point detection on processed SR data at 23°C and 38°C. None indicates that no slope change-point was detected.

Mix ID	CP detected at 23°C (days)	CP detected at 38°C (days)
OPC	None	None
F	30	2
Y1	None	10
BH1	None	6

Table 1 provides determination of pozzolanicity of both ASTM C618 and off-spec materials. At elevated curing conditions, a material's pozzolanicity can be determined in as quickly as two days, an enormous time advantage over the current standard, ASTM C618, which requires 28 days of testing before a determination can be made.

This data-driven approach on time-series experimental data [43] allows the rigorous determination of changes in material behavior in a quicker time frame than is the current industry standard. This approach is transferrable to additional surface resistivity data of emerging materials, which would promote the use of novel materials in concrete, as well as to other time-series measurements in civil engineering, such as those related to durability.

6 Future directions and conclusion

Machine learning and other data analysis approaches provide powerful frameworks for integrating physicochemical and materials-engineering models of cementitious materials with statistical learning. A diversity of material properties ranging from rheology to microstructure and durability can be predicted from material constituents. Here, through example applications, machine learning and statistical techniques were shown to enable the identification and design of chemical admixtures, to link the composition of cement-based materials to its resistance to saturation and freeze/thaw damage, and to effectively screen new SCM sources.

More broadly, these examples demonstrate the potential of such data-driven approaches to transform cement-based materials research, design, and specification. In research, curation of and open access to data and metadata will be critical in accelerating the translation of new knowledge from the laboratory to practice. Platforms and data standards to facilitate the integrated analysis of data from multiple sources are desperately needed in this field. With depletion of traditional materials feedstocks and with increasing availability of alternative materials sources, a need exists to reduce risk when designing concrete with emerging materials. By combining existing and emerging knowledge of materials performance, uncertainty in workability, strength

development, and durability can be mitigated, increasing the rate at which new, more sustainable materials can be translated into practice. Finally, as materials and design specifications transition from those based on prescription to those based on performance, machine learning can be beneficial. Cost of testing, along with managing risk and responsibility, have been major impediments to increased adoption of performance-based specification. Facilitated by the use of standard test methods, machine learning can use these existing vast data sets – linking materials compositional parameters to performance – to reduce the amount and duration of testing, resulting not only in cost and time savings, but also increasing the industry's ability to innovate while mitigating risk.

Acknowledgements

The authors gratefully acknowledge the input of Dr. Liyan Xie and Prof. Yao Xie from the Georgia Institute of Technology (Atlanta, USA) for discussions about change-point detection and Prof. Surya Kalidindi from the Georgia Institute of Technology (Atlanta, USA) for discussions regarding process-structure-property relationships. The authors also gratefully acknowledge ARPA-E for partial support of this work under DE-AR0001138. The first author acknowledges support by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1650044. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Authorship statement (CRediT)

Renee T. Rios: Conceptualization; Methodology; Formal analysis; Investigation; Visualization; Writing - original draft, Writing - Review and editing; Funding acquisition.

Christopher M. Childs: Conceptualization; Methodology; Formal analysis; Investigation; Visualization; Writing - original draft; Writing - review and editing.

Scott H. Smith: Conceptualization; Methodology; Formal analysis; Investigation; Visualization; Writing - original draft.

Newell R. Washburn: Conceptualization; Methodology; Resources; Supervision; Writing - review and editing; Funding acquisition.

Kimberly E. Kurtis: Conceptualization; Resources; Supervision; Writing - original draft; Writing - review and editing; Funding acquisition.

References

[1] K.E. Kurtis, Innovations in cement-based materials: Addressing sustainability in structural and infrastructure applications, MRS Bulletin, (2015) 40: 1102-1109. https://doi.org/10.1557/mrs.2015.279

[2] W. Perry, A. Broers, F. El-Baz, W. Harris, B. Healy, W.D. Hillis, C. Juma, D. Kamen, R. Kurzweil, R. Langer, J. Lerner, B. Lohani, J. Lubchenco, M. Molina, L. Page, R. Socolow, J.C. Venter, J. Ying, NAE GRAND CHALLENGES FOR ENGINEERING, Washington, D.C., 2017.

[3] I.-C. Yeh, Modeling of strength of high-performance concrete using artificial neural networks, Cem Concr Res (1998) 28: 1797-1808. https://doi.org/10.1016/S0008-8846(98)00165-3

[4] J.-S. Chou, C.-K. Chiu, M. Farfoura, I. Al-Taharwa, Optimizing the prediction accuracy of concrete compressive strength based on a comparison of data-mining techniques, J Comp Civ Eng (2011) 25: 242-253. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000088

[5] R.M. Haj-Ali, K.E. Kurtis, A.R. Sthapit, Neural network modeling of concrete expansion during long-term sulfate exposure, Mater J (2001) 98: 36-43. https://doi.org/10.14359/10158

[6] S. Dutta, P. Samui, D. Kim, Comparison of machine learning techniques to predict compressive strength of concrete, Computers and Concrete (2018) 21: 463-470.

[7] [7] B. Ouyang, Y. Song, Y. Li, F. Wu, H. Yu, Y. Wang, G. Sant, M. Bauchy, Predicting Concrete's Strength by Machine Learning: Balance between Accuracy and Complexity of Algorithms, ACI Mater J (2020) 117 . https://doi.org/10.14359/51728128

[8] M. Jalal, Experimental Assessment, Optimization, and Multi-scale Modeling of Alkali-Silica Reaction (ASR) Through Machine-Learning Techniques, 2020.

[9] M. Marks, M.A. Glinicki, K. Gibas, Prediction of the chloride resistance of concrete modified with high calcium fly ash using machine learning, Materials (2015) 8: 8714-8727. https://doi.org/10.3390/ma8125483

[10] H. Naseri, H. Jahanbakhsh, F. Moghadas Nejad, A. Golroo, Developing a novel machine learning method to predict the compressive strength of fly ash concrete in different ages, AUT J Civ Eng (2020) 4: 3-3.

[11] C. Zheng, X. Liu, P. Yan, Y. Zhang, Y. Wang, K. Qiu, X. Gao, Measurement and prediction of fly ash resistivity over a wide range of temperature, Fuel (2018) 216: 673-680. https://doi.org/10.1016/j.fuel.2017.12.047

[12] S. Kang, M.T. Ley, A. Behravan, Predicting ion diffusion in fly ash cement paste through particle analysis, Constr Build Mater (2021) 272: 121934. https://doi.org/10.1016/j.conbuildmat.2020.121934

[13] S. Kang, M.T. Ley, Z. Lloyd, T. Kim, Using the Particle Model to predict electrical resistivity performance of fly ash in concrete, Constr Build Mater (2020) 261: 119975. https://doi.org/10.1016/j.conbuildmat.2020.119975

[14] S. Kang, Z. Lloyd, T. Kim, M.T. Ley, Predicting the compressive strength of fly ash concrete with the Particle Model, Cem Concr Res (2020) 137: 106218. https://doi.org/10.1016/j.cemconres.2020.106218

[15] T. Kim, M.T. Ley, S. Kang, J.M. Davis, S. Kim, P. Amrollahi, Using particle composition of fly ash to predict concrete strength and electrical resistivity, Cem Concr Compos (2020) 107: 103493. https://doi.org/10.1016/j.cemconcomp.2019.103493

[16] Y. Song, K. Yang, J. Chen, K. Wang, G. Sant, M. Bauchy, Machine Learning Enables Rapid Screening of Reactive Fly Ashes Based on Their Network Topology, ACS Sust Chem Eng (2021) 9: 2639-2650. https://doi.org/10.1021/acssuschemeng.0c06978

[17] K. Scrivener, F. Martirena, S. Bishnoi, S. Maity, Calcined clay limestone cements (LC3), Cem Concr Res (2018) 114: 49-56. https://doi.org/10.1016/j.cemconres.2017.08.017

[18] C.M. Childs, N.R. Washburn, Embedding domain knowledge for machine learning of complex material systems, MRS Communications (2019) 9: 806-820. https://doi.org/10.1557/mrc.2019.90

[19] P. Domingos, A few useful things to know about machine learning, Communications of the ACM (2012) 55: 78-87. https://doi.org/10.1145/2347736.2347755

[20] N. Washburn, A. Menon, C. Childs, B. Poczos, K. Kurtis, Machine learning approaches to admixture design for clay-based cements, Calcined Clays for Sustainable Concrete, Springer2018, 488-493. https://doi.org/10.1007/978-94-024-1207-9_78

[21] A. Menon, C. Gupta, K.M. Perkins, B.L. DeCost, N. Budwal, R.T. Rios, K. Zhang, B. Póczos, N.R. Washburn, Elucidating multi-physics interactions in suspensions for the design of polymeric dispersants: a hierarchical machine learning approach, Mol Syst Des Eng (2017) 2: 263-273. https://doi.org/10.1039/C7ME00027H

[22] J.M. Bone, C.M. Childs, A. Menon, B. Poczos, A.W. Feinberg, P.R. LeDuc, N.R. Washburn, Hierarchical Machine Learning for High-Fidelity 3D Printed Biopolymers, ACS Biomater Sci Eng (2020) 6: 7021-7031. https://doi.org/10.1021/acsbiomaterials.0c00755

[23] A. Menon, J.A. Thompson-Colón, N.R. Washburn, Hierarchical machine learning model for mechanical property predictions of polyurethane elastomers from small datasets, Frontiers in Materials (2019) 6: 87. https://doi.org/10.3389/fmats.2019.00087

[24] H. Arora, N. Coleman, The influence of electrolyte concentration on flocculation of clay suspensions, Soil Science (1979) 127: 134-139. https://doi.org/10.1097/00010694-197903000-00002

[25] K. Scrivener, F. Avet, H. Maraghechi, F. Zunino, J. Ston, W. Hanpongpun, A. Favier, Impacting factors and properties of limestone calcined clay cements (LC^3), Green Materials (2018) 7: 3-14. https://doi.org/10.1680/jgrma.18.00029
[26] H. Tan, B. Gu, B. Ma, X. Li, C. Lin, X. Li, Mechanism of intercalation of polycarboxylate superplasticizer into montmorillonite, Appl Clay Sci, (2016) 129: 40-46. https://doi.org/10.1016/j.clay.2016.04.020
[27] L. Lei, J. Plank, A concept for a polycarboxylate superplasticizer possessing enhanced clay tolerance, Cem Concr Res (2012) 42: 1299-1306. https://doi.org/10.1016/j.cemconres.2012.07.001
[28] A. Menon, C.M. Childs, B. Poczós, N.R. Washburn, K.E. Kurtis, Molecular engineering of superplasticizers for metakaolin ‐ portland cement blends with hierarchical machine learning, Adv Theor Simul (2019) 2: 1800164. https://doi.org/10.1002/adts.201800164
[29] J.A. Pugar, C.M. Childs, C. Huang, K.W. Haider, N.R. Washburn, Elucidating the Physicochemical Basis of the Glass Transition Temperature in Linear Polyurethane Elastomers with Machine Learning, J Phys Chem B (2020) 124: 9722-9733. https://doi.org/10.1021/acs.jpcb.0c06439
[30] C. Kim, A. Chandrasekaran, T.D. Huan, D. Das, R. Ramprasad, Polymer genome: a data-powered polymer informatics platform for property predictions, J Phys Chem C (2018) 122: 17575-17585. https://doi.org/10.1021/acs.jpcc.8b02913
[31] Y.-C. Lo, S.E. Rensi, W. Torng, R.B. Altman, Machine learning in chemoinformatics and drug discovery, Drug Discovery Today (2018) 23: 1538-1546. https://doi.org/10.1016/j.drudis.2018.05.010
[32] F.A. Faber, L. Hutchison, B. Huang, J. Gilmer, S.S. Schoenholz, G.E. Dahl, O. Vinyals, S. Kearnes, P.F. Riley, O.A. Von Lilienfeld, Prediction errors of molecular machine learning models lower than hybrid DFT error, J Chem Theor Comput (2017) 13: 5255-5264. https://doi.org/10.1021/acs.jctc.7b00577
[33] M. Gütlein, S. Kramer, Filtered circular fingerprints improve either prediction or runtime performance while retaining interpretability, J Cheminformatics (2016) 8: 1-16. https://doi.org/10.1186/s13321-016-0173-z
[34] H.M. Kayello, N.K. Tadisina, N. Shlonimskaya, J.J. Biernacki, D.P. Visco Jr, An Application of Computer ‐ Aided Molecular Design (CAMD) using the signature molecular descriptor-part 1. Identification of surface tension reducing agents and the search for shrinkage reducing admixtures, J Am Chem Soc (2014) 97: 365-377. https://doi.org/10.1111/jace.12453
[35] A. Cecen, H. Dai, Y.C. Yabansu, S.R. Kalidindi, L. Song, Material structure-property linkages using three-dimensional convolutional neural networks, Acta Materialia (2018) 146: 76-84. https://doi.org/10.1016/j.actamat.2017.11.053
[36] A. Khosravani, A. Cecen, S.R. Kalidindi, Development of high throughput assays for establishing process-structure-property linkages in multiphase polycrystalline metals: Application to dual-phase steels, Acta Materialia (2017) 123: 55-69. https://doi.org/10.1016/j.actamat.2016.10.033
[37] S.H. Smith, M. Vandamme, K.E. Kurtis, Dissolution kinetics of trapped air in a spherical void: Modeling the long-term saturation of cementitious materials, Cem Concr Res (2020) 130: 105996. https://doi.org/10.1016/j.cemconres.2020.105996
[38] S.H. Smith, C. Qiao, P. Suraneni, K.E. Kurtis, W.J. Weiss, Service-life of concrete in freeze-thaw environments: Critical degree of saturation and calcium oxychloride formation, Cem Concr Res (2019) 122: 93-106. https://doi.org/10.1016/j.cemconres.2019.04.014
[39] J.W. Bullard, Virtual cement and concrete testing laboratory: Version 9.5 user guide, (2014). https://doi.org/10.6028/NIST.SP.1173
[40] E. Garboczi, D. Bentz, Computer simulation of the diffusivity of cement-based materials, J Mater Sci (1992) 27: 2083-2092. https://doi.org/10.1007/BF01117921
[41] S.H. Smith, Toward the Service-Life Design of Cementitious Materials In Freeze-Thaw Environments: Novel Models, Specifications, and Evaluation Methods, Georgia Institute of Technology, 2019.
[42] E.I. Nadelman, K.E. Kurtis, A resistivity-based approach to optimizing concrete performance, Concrete international, 36 (2014) 50-54.
[43] R.T. Rios, F. Lolli, L. Xie, Y. Xie, K.E. Kurtis, Screening candidate supplementary cementitious materials under standard and accelerated curing through time-series surface resistivity measurements and change-point detection, Cem Concr Res (2021) 148: 106538. https://doi.org/10.1016/j.cemconres.2021.106538
[44] Y. Cao, Y. Xie, N. Gebraeel, Multi-sensor slope change detection, Annals of Operations Research (2018) 263: 163-189. https://doi.org/10.1007/s10479-016-2185-5
[45] Y. Xie, D. Siegmund, Sequential multi-sensor change-point detection, 2013 Information Theory and Applications Workshop (ITA), IEEE, 2013, 1-20. https://doi.org/10.1109/ITA.2013.6502987