Decoding the genome of cement by Gaussian Process Regression

Yu Song, Yongzhe Wang, Kaixin Wang, Gaurav Sant, Mathieu Bauchy*
Department of Civil and Environmental Engineering
University of California, Los Angeles
Los Angeles, CA 90095
{yusong, wingzhewang1, kaixinwang, gsant, bauchy}@ucla.edu

Abstract

Reducing the carbon footprint in cement production is a pressing challenge faced by the construction industry. In the past few years, the world annual cement consumption is approximately at 4 billion tons, where each ton leads to 1-ton CO_2 emissions. To curb the massive environmental impact, it is pertinent to improve material performance and reduce carbon embodiment of cement. This requires an in-depth understanding of how cement strength is controlled by its chemical composition. Although this problem has been investigated for more than one hundred years, our current knowledge is still deficient for a clear decomposition of this complex composition-strength relationship. Here, we take advantage of Gaussian process regression (GPR) to decipher the fundamental compositional attributes (the cement "genome") to cement strength performance. Among all machine learning methods applied to the same dataset, our GPR model achieves the highest accuracy of predicting cement strength based on the chemical compounds. Based on the optimized GPR model, we are able to decompose the influence of each oxide on cement strength to an unprecedented level.

1 Introduction

Human life has been increasingly challenged by environmental problems associated with social developments. Among all, global warming is undoubtedly one of the top issues constituting a profound global impact. In this regard, massive CO_2 emissions are directly attributed to the manufacturing of engineering raw materials. Despite the extensive research attention, our current knowledge of many fundamentals aspects of the engineering materials is still limited, which impedes proper material design and optimization for attaining the low-carbon aim of sustainable engineering. The recent advances in machine learning open up new possibilities to material research (1; 2; 3). Different from the conventional ways to develop our cognition based on the accumulation of experience and knowledge, machine learning is especially outstanding in establishing the correlation between (seemingly) unrelated things. In particular, this data-driven nature of machine learning presents an excellent fit for advancing our understanding of the composition-property correlation for various engineering materials, as the correlation can be established without explicit knowledge of some underlying physical and chemical mechanisms remaining unknown. Here, we showcase an effort of using machine learning (Gaussian process regression) to advance our knowledge of the composition-strength relationship of cement.

Cement represents a typical engineering material produced in the construction industry that involves a strong carbon footprint. As a matter of fact, the world annual cement consumption is approximately at 4 billion tons over the past few years and this number is expected to be continuously growing in

*Corresponding author: bauchy@ucla.edu, http://www.lab-paris.com

Workshop on machine learning for engineering modeling, simulation and design @ NeurIPS 2020

the next decades (4), and producing each ton of cement results in about 1 ton of CO_2 emitted into the atmosphere (5). Currently, cement is often redundantly used in construction since the knowledge of this material is still lacking to ensure more precise control of its engineering performance (e.g., strength). In terms of the chemical composition, an ordinary cement typically consists of 67% CaO, 22% SiO_2, 5% Al_2O_3, 3% Fe_2O_3, with 3% other components (6). During cement hydration, all these components make their contributions to the material strength overtimes. The real hydration process in cement, however, involves complex and systematic interactions between its many components, leaving it rather challenging to isolate the actual effect of a particular component (7; 8; 9). In particular, with the incursion of a wide degree of variables in cement, the expected reactivity of a specific phase may not be the same in cement as if it is measured independently. In addition, accurate quantification of the physical-chemical reactions of some minor components in cement maybe even impossible to achieve experimentally. Therefore, cement research provides a good manifestation of the common difficulties faced by many studies in engineering materials, including but not limited to nonlinear, non-additive, high-dimensional, time-dependent, and practical difficulties of the experiments.

In recent years, the use of machine learning to predict material property has become an emerging trend in the field of cement and concrete research (10; 11; 12; 13; 14; 15). However, little attention has been paid to taking advantage of this approach for more in-depth interrogation of the intrinsic composition-property relationship regarding the cement hydration. To address this problem, this study aims to apply Gaussian process regression (GPR) to deconstruct the composition-strength relationship of cement. Specifically, our goal is to decipher how the physical and chemical features of a cement (i.e., the cement "genome") govern its strength development—in the same fashion as the DNA of humans control many of their characteristics.

Here, GPR is adopted because it has several advantages for studying engineering materials over the other prevailing machine learning methods. First, GPR model is a non-parametric learning algorithm (16), meaning that predictability does not rely on strong assumptions of the form of the mapping function. This greatly eases the exploration of the input-output relationship in material research, as the exact relationship is often not available and/or may vary greatly from case to case. As such, GPR has sufficient flexibility to fit completely unknown relationships without the concerns of the constraint of the parametric models. Second, GPR allows leveraging the prior knowledge about the input-output relationship into the model construction via the selection of computational kernels. For many engineering material, certain prior knowledge is available. Thus, a specific set of basic kernels can be combined to approximate the expectation. Third, the prediction made by GPR is essentially a probabilistic distribution of the expected output. From a practical perspective, the statistical nature of the GPR model prediction is of special significance for the rational design of engineering materials.

For the GPR model training, we adopt a dataset comprising of more than 2000 commercial cement samples, where the input features are the contents of bulk oxide compositions, along with particle fineness, and the output is the material strength after 28 days of hydration (i.e., the most accepted strength metric of cement). Base on the optimized GPR model, we conducted feature impact and feature effect analysis to investigate the influence of the individual input features to cement strength. Remarkably, the work presented herein provides one of the first machine learning investigations to explore the material-strength down to the fine level of individual oxides.

2 Related Works

Machine learning in cement and concrete research. Due to the initial stage of the investigations, the primary focus of the recent studies has been placed on applying various learning techniques (e.g., multiple linear regression, regression network, support vector machine, and tree-based models) to predict the macroscopic performance of cementitious materials such as strength (10; 11; 12; 13; 17) and durability (14; 18; 19), while some other topics can be found in terms of mixture optimization (20; 21) and image-based investigation (22; 23; 24). Thus far, limited effort has been paid to taking advantage of this approach for more in-depth interrogation of the intrinsic composition-property relationship regarding the cement hydration. The most relevant work to predicting the strength behavior of cement is found in (13), where the authors compared several prevailing machine learning models (e.g. KNN and several tree-based approaches) to predict the 28-day cement strength.

3 Methodology

3.1 The cement dataset

Herein, the cement dataset used for the machine learning analysis is adopted from a previous study (13). This dataset is established on an industry survey over a large number of U.S. cement testing institutes, led by the Portland Cement Association (PCA) and the National Institute of Standards and Technology (NIST), wherein all the cement samples were tested per varying ASTM standards. Based on the raw data, a preliminary screening is carried out to remove some unreasonable samples (e.g., negative strength record, replication, missing key attributes, etc.) that are misinforming. The dataset adopted herein comprises of 2060 cement samples, where each sample includes the information of ten chemical compositions (CaO, $\mathrm{SiO_2}$, $\mathrm{Al_2O_3}$, $\mathrm{Fe_2O_3}$, $\mathrm{SO_3}$, MgO, LOI (loss on ignition), $\mathrm{Free\ CaO}$, $\mathrm{K_2O}$, and $\mathrm{Na_2O}$) and Blaine fineness of the cement particles as the input features, and the ASTM C109 cement mortar strength at 28 days of hydration as the label (25). As an illustration, the direct correlations between the two major inputs (CaO and $\mathrm{SiO_2}$) and the 28-day strength are given in Fig. 1. It is apparent from these examples that their contributions to cement strength cannot capture intuitively.

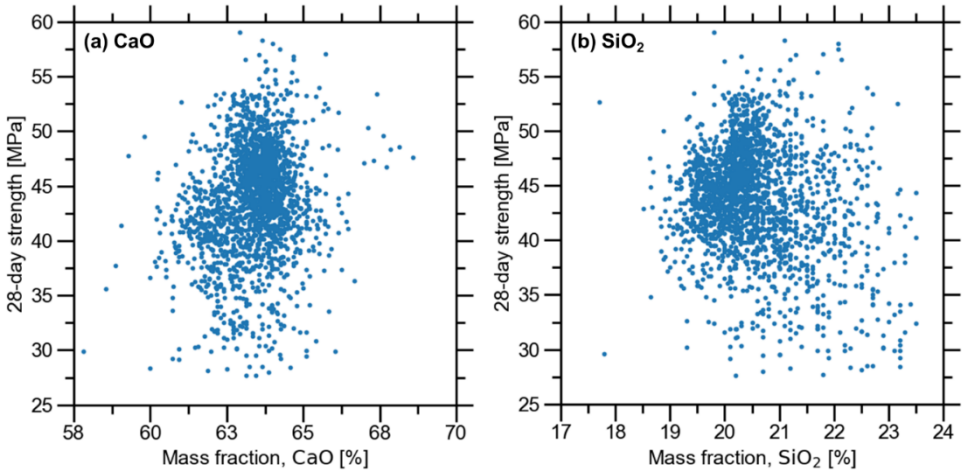

Figure 1: Direct correlations between individual features and the 28-day cement mortar strength of all samples in the cement dataset, for (a) CaO, and (b) $\mathrm{SiO_2}$.

3.2 Pipeline of the machine learning analysis

For training each machine learning model, we use 90% cement samples in the raw dataset for the model training and the remaining 10% samples for testing. This train-test split is carried out with stratification sampling to ensure that the split subsets are representative of the distribution of 28-day strength in the raw dataset (26). For optimizing the GPR model, we execute a five-fold cross-validation to tune the model hyperparameters. With respect to the evaluation of models, we predominantly adapt to the coefficient of determination (R^2), with mean absolute percentage error (MAPE) as additional references. More details are provided in Section 4.1.

After obtaining the final model for predicting cement strength, we carry out feature impact and feature effect analyses to explore the composition-strength relationship based on the optimized model. The feature impact analysis is done based on the permutation feature importance (27). Here, we repeat the permutation for each feature by 100 times to ensure a statistically stable interpretation of the feature impact. The feature effect analysis focuses on probing the effect of the individual features. In particular, we define a single centroid composition based on the distribution of the individual features, wherein each of its feature value equals to the median of the corresponding feature over the whole dataset. The centroid composition has 3.01% $\mathrm{SO_3}$, 63.75% CaO, 4.64% $\mathrm{Al_2O_3}$, 20.37% $\mathrm{SiO_2}$, 3.25% $\mathrm{Fe_2O_3}$, 2.10% MgO, 0.19% $\mathrm{Na_2O}$, 1.72% LOI, 0.87% $\mathrm{Free\ CaO}$, 0.93% $\mathrm{K_2O}$, and 391 $\mathrm{m^2/kg}$ for Blaine fineness). Next, the effect of the single feature of interest is interrogated by altering its value within its maximum distance to the extremum (min or max, whichever is further from the

median), while adjusting all the other features proportionally in correspondence such that the sum of the features remains close to 100%. During this process, we use the trained models to predict the strength at a fine interval of the feature jittering. As such, we can obtain a curve of each feature where it shows how the mortar strength varies as a function of the investigated feature.

4 Results

4.1 Prediction accuracy of the optimized GPR model

To optimized the GPR model, we select the hyperparameters by following a three-step optimization: (i) compare the model performance when using the individual candidate kernel(s), wherein the kernel parameter is left as default, (ii) fine-tune the kernel parameters of the best candidate kernel, and (iii) adjust the level of expected noise to prevent overfitting problem. For good reliability, the optimal hyperparameters are selected based on an average of 10 repetitions. Here, we adopt a combination of Linear+RBF kernels as the general composition-strength relationship is expected to be monotonic. The R^2 accuracy of this model on the test set is 0.59 ± 0.03. As a reference, a comparison between the predicted versus true 28-day cement strength is displayed in Fig. 2. We note that the main cluster aligns with the line of equality fairly reasonable, suggesting an accurate strength prediction for most of the cement samples. To our knowledge, this model achieves the highest prediction accuracy among all the machine learning methods that have been applied to this dataset, and the best R^2 accuracy reported previously is 0.51 (13).

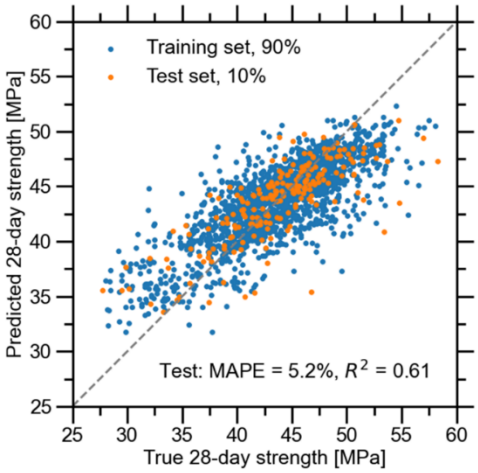

Figure 2: Predicted vs. true 28-day strength of the cement dataset considered herein. The $y = x$ dashed line indicates perfect agreement.

4.2 Feature analyses on the 28-day cement strength

Now turning to the analyses of the different features on the cement strength of hydrated cement. Based on the optimized GPR model, we first implement the permutation analysis to investigate the influence of each input feature to the 28-day strength, as displayed in Fig. 3a, and then we implement the feature effect analysis (see Section 3.2) to further check the exact effect of the individual features, as displayed in Fig. 3b. In either analysis, a clear contrast is seen between some major features, namely, $\mathrm{Fineness}$, CaO, $\mathrm{Al_2O_3}$, $\mathrm{SiO_2}$, and $\mathrm{SO_3}$. Furthermore, the slope of the feature effect curves in Fig. 3b is broadly consistent with the impact ranking Fig. 3a. Here, we first note that the overwhelming role of the fineness of the cement powder. This echoes with the common expectations that the fineness has a significant impact on cement strength, which has been long recognized in cement research (28; 29; 30). Furthermore, it also makes sense to have $\mathrm{SO_3}$, $\mathrm{Al_2O_3}$, CaO ranked high among all, as these oxides are known to play an important role in the setting behavior of cement, as well as the strength build-up (31; 32; 33).

To this point, it is clear that the presented feature analyses provide us valuable information regarding how the cement strength can be affected by the variation of its composition. Indeed, this is the first time that such an interpretation is achieved by machine learning for cement research. These results are extremely useful for the reverse design of the cement composition based on the practical need of specific strength requirements, which, in turn, helps to improve the efficiency of raw material usages and reduce the carbon footprint in the cement production.

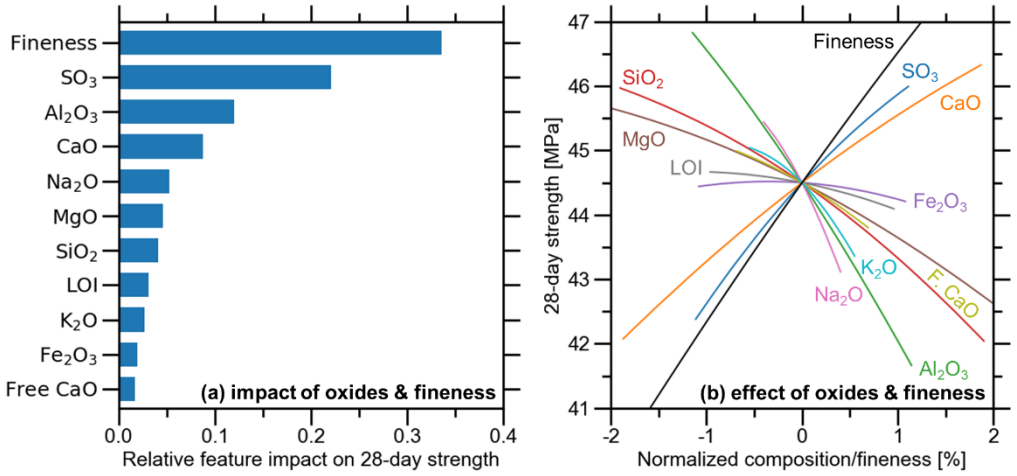

Figure 3: Feature analyses for the 28-day cement strength: (a) feature impact analysis, and (b) feature effect analysis. For (a), the individual features are normalized so that the impact values sum up to 1. For (b), each line represents altering the value of one feature from the centroid composition (see Section 3.2) while adjusting the other feature accordingly to ensure the total content remaining unchanged.

5 Discussion

We find that the results yielded from the GPR model are rather interesting and encouraging—after all, all the findings presented in this work are based on the sole inputs of the basic material attributes of the cement instances in the dataset. Most of the findings offered by this learning analysis largely echo with the existing understanding from the field of cement research (6). Admittedly, some points cannot be fully explained by the GPR-based feature analyses. However, this machine learning analysis is nevertheless agreed by the established studies in cement chemistry with respect to the major aspects of the effect of the investigated features. To a certain extent, the challenge of the verification comes from the fact that findings obtained by machine learning are not necessarily easily measurable using experiments, especially for quantifying the feature effect.

Overall, our study shows that machine learning analysis, when properly conducted, can be a powerful technique for understating the composition-property relationship of engineering materials. In particular, this study demonstrates that machine learning can be exclusively useful for disentangling the relationship between different coupled components in the material that are hard to be investigated independently. Furthermore, machine learning can significantly facilitate the quantitative comparison of feature effects to the material property of interest. While the goal of many current machine learning-related studies in this field stays on improving the prediction accuracy of the model (which is fundamental to the success of this technique), this emerging approach can indeed offer many more valuable insights for further clarifying the complex material-property relationships of materials, from a novel data-driven perspective that may not be otherwise attained.

Broader Impact

We foresee several positive impacts of the presented work. So far, many traditional carbon-heavy manufacturing industries have been awaiting breakthroughs in face of the increasing pressure from

environmental problems. In this regard, machine learning has a clear advantage of mapping the inputs (e.g. material attributes) to the output (e.g. engineering property), especially when the obscure correlation cannot be captured by our current knowledge. As such, there is an urgent need of applying the emerging machine learning techniques to augment the solutions for many pressing engineering problems. Importantly, the presented work showcases a pioneering effort of leveraging artificial intelligence (AI) to unveil new insights of the composition-property relationship of one of the most important engineering materials. The presented archetypal pipeline can be transferred to advance the understanding, design, and optimization of many other engineering materials. We envision that such AI-informed approaches will induce a paradigm shift in the pathways of engineering material research. Overall, the adoption of machine learning techniques in traditional engineering fields will spur many chances to promote the green engineering concept and mandate of a circular economy.

Acknowledgments

The authors acknowledge some financial support for this research provided by the U.S. Department of Transportation through the Federal Highway Administration (Grant: 693JJ31950021) and the U.S. National Science Foundation (DMREF: 1922167).

References

[1] S. Idowu, S. Saguna, C. Åhlund, and O. Schelén, "Applied machine learning: Forecasting heat load in district heating system," *Energy and Buildings*, vol. 133, pp. 478–488, 2016.

[2] Y. Zhang and C. Ling, "A strategy to apply machine learning to small datasets in materials science," *Npj Computational Materials*, vol. 4, no. 1, pp. 1–8, 2018.

[3] K. Hazelwood, S. Bird, D. Brooks, S. Chintala, U. Diril, D. Dzhulgakov, M. Fawzy, B. Jia, Y. Jia, A. Kalro *et al.*, "Applied machine learning at facebook: A datacenter infrastructure perspective," in *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2018, pp. 620–629.

[4] J. S. van Deventer, C. E. White, and R. J. Myers, "A roadmap for production of cement and concrete with low-co 2 emissions," *Waste and Biomass Valorization*, pp. 1–31, 2020.

[5] J. Davidovits, "Global warming impact on the cement and aggregates industries," *World resource review*, vol. 6, no. 2, pp. 263–278, 1994.

[6] J. F. Young, S. Mindess, and D. Darwin, *Concrete*. Prentice Hall, 2002.

[7] X. Li, W. Xu, S. Wang, M. Tang, and X. Shen, "Effect of so3 and mgo on portland cement clinker: Formation of clinker phases and alite polymorphism," *Construction and Building Materials*, vol. 58, pp. 182–192, 2014.

[8] Y.-M. Kim and S.-H. Hong, "Influence of minor ions on the stability and hydration rates of β-dicalcium silicate," *Journal of the American Ceramic Society*, vol. 87, no. 5, pp. 900–905, 2004.

[9] S. Bergold, F. Goetz-Neunhoeffer, and J. Neubauer, "Interaction of silicate and aluminate reaction in a synthetic cement system: Implications for the process of alite hydration," *Cement and Concrete Research*, vol. 93, pp. 32–44, 2017.

[10] J.-S. Chou, C.-F. Tsai, A.-D. Pham, and Y.-H. Lu, "Machine learning in concrete strength simulations: Multi-nation data analytics," *Construction and Building Materials*, vol. 73, pp. 771–780, 2014.

[11] J.-S. Chou, C.-K. Chiu, M. Farfoura, and I. Al-Taharwa, "Optimizing the prediction accuracy of concrete compressive strength based on a comparison of data-mining techniques," *Journal of Computing in Civil Engineering*, vol. 25, no. 3, pp. 242–253, 2011.

[12] B. A. Young, A. Hall, L. Pilon, P. Gupta, and G. Sant, "Can the compressive strength of concrete be estimated from knowledge of the mixture proportions?: New insights from statistical analysis and machine learning methods," *Cement and Concrete Research*, vol. 115, pp. 379–388, 2019.

[13] T. Oey, S. Jones, J. W. Bullard, and G. Sant, "Machine learning can predict setting behavior and strength evolution of hydrating cement systems," *Journal of the American Ceramic Society*, vol. 103, no. 1, pp. 480–490, 2020.

[14] A. K. Das, D. Suthar, and C. K. Leung, "Machine learning based crack mode classification from unlabeled acoustic emission waveform features," *Cement and Concrete Research*, vol. 121, pp. 42–57, 2019.

[15] N.-D. Hoang, C.-T. Chen, and K.-W. Liao, "Prediction of chloride diffusion in cement mortar using multi-gene genetic programming and multivariate adaptive regression splines," *Measurement*, vol. 112, pp. 141–149, 2017.

[16] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*. Springer, 2003, pp. 63–71.

[17] P. Chopra, R. K. Sharma, M. Kumar, and T. Chopra, "Comparison of machine learning techniques for the prediction of compressive strength of concrete," *Advances in Civil Engineering*, vol. 2018, 2018.

[18] Y. Okazaki, S. Okazaki, S. Asamoto, and P.-j. Chun, "Applicability of machine learning to a crack model in concrete bridges," *Computer-Aided Civil and Infrastructure Engineering*, 2020.

[19] R. Cai, T. Han, W. Liao, J. Huang, D. Li, A. Kumar, and H. Ma, "Prediction of surface chloride concentration of marine concrete using ensemble machine learning," *Cement and Concrete Research*, vol. 136, p. 106164, 2020.

[20] P. Ziolkowski and M. Niedostatkiewicz, "Machine learning techniques in concrete mix design," *Materials*, vol. 12, no. 8, p. 1256, 2019.

[21] H. Choi, G. Venkiteela, A. Gregori, and H. Najm, "Advanced quality control models for concrete admixtures," *Journal of Materials in Civil Engineering*, vol. 32, no. 2, p. 04019349, 2020.

[22] S. S. Bangaru, C. Wang, M. Hassan, H. W. Jeon, and T. Ayiluri, "Estimation of the degree of hydration of concrete through automated machine learning based microstructure analysis–a study on effect of image magnification," *Advanced Engineering Informatics*, vol. 42, p. 100975, 2019.

[23] Y. Song, Z. Huang, C. Shen, H. Shi, and D. A. Lange, "Deep learning-based automated image segmentation for concrete petrographic analysis," *Cement and Concrete Research*, vol. 135, p. 106118, 2020.

[24] A. Das, Y. Song, S. Mantellato, T. Wangler, R. J. Flatt, and D. A. Lange, "Influence of pumping/extrusion on the air-void system of 3d printed concrete," in *RILEM International Conference on Concrete and Digital Fabrication*. Springer, 2020, pp. 417–427.

[25] C. Committee *et al.*, "Test method for compressive strength of hydraulic cement mortars (using 2-in. or [50-mm] cube specimens)," *ASTM International*, 2013.

[26] K. M. Jablonka, D. Ongari, S. M. Moosavi, and B. Smit, "Big-data science in porous materials: materials genomics and machine learning," *Chemical reviews*, vol. 120, no. 16, p. 8066, 2020.

[27] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[28] S. Tsivilis, S. Tsimas, A. Benetatou, and E. Haniotakis, "Study on the contribution of the fineness on cement strength," *ZKG*, vol. 1, pp. 26–29, 1990.

[29] K. Alexander, "The relationship between strength and the composition and fineness of cement," *Cement and Concrete Research*, vol. 2, no. 6, pp. 663–680, 1972.

[30] D. P. Bentz, E. J. Garboczi, C. J. Haecker, and O. M. Jensen, "Effects of cement particle size distribution on performance properties of portland cement-based materials," *Cement and concrete research*, vol. 29, no. 10, pp. 1663–1671, 1999.

[31] D. Hobbs, "The influence of so3 content on the behaviour of portland cement mortars," *World Cement Technology*, vol. 8, no. 3, 1977.

[32] R. Sersale, R. Cioffi, G. Frigione, and F. Zenone, "Relationship between gypsum content, porosity and strength in cement. i. effect of so3 on the physical microstructure of portland cement mortars," *Cement and concrete Research*, vol. 21, no. 1, pp. 120–126, 1991.

[33] K. Tosun, "Effect of so3 content and fineness on the rate of delayed ettringite formation in heat cured portland cement mortars," *Cement and concrete composites*, vol. 28, no. 9, pp. 761–772, 2006.