Physics-based Data-Augmented Deep Learning for Enhanced Autogenous Shrinkage Prediction on Experimental Dataset

Vishu Gupta
Northwestern University
Evanston, Illinois, USA
vishugupta2020@u.northwestern.edu

Yuhui Lyu
Northwestern University
Evanston, Illinois, USA
yuhuilyu2022@u.northwestern.edu

Derick Suarez
Northwestern University
Evanston, Illinois, USA
dericksuarez2024@u.northwestern.edu

Yuwei Mao
Northwestern University
Evanston, Illinois, USA
yuweimao2019@u.northwestern.edu

Wei-keng Liao
Northwestern University
Evanston, Illinois, USA
wkliao@eecs.northwestern.edu

Alok Choudhary
Northwestern University
Evanston, Illinois, USA
choudhar@eecs.northwestern.edu

Wing Kam Liu
Northwestern University
Evanston, Illinois, USA
w-liu@northwestern.edu

Gianluca Cusatis
Northwestern University
Evanston, Illinois, USA
g-cusatis@northwestern.edu

Ankit Agrawal
Northwestern University
Evanston, Illinois, USA
ankitag@eecs.northwestern.edu

ABSTRACT

Prediction of the autogenous shrinkage referred to as the reduction of apparent volume of concrete under seal and isothermal conditions is of great significance in the service life analysis and design of durable concrete structures, especially with the increasing use of concrete with low water-to-cement ratios. However, due to the highly complex mechanism of autogenous shrinkage, it is hard to design accurate mechanistic models for it. Existing state-of-the-art models for autogenous shrinkage do not perform well for several reasons such as not being able to capture faster shrinkage change at early ages (swelling), coefficients used are derived using statistical optimization methods to fit certain databases only, and mechanism to identify the most influencing factors on autogenous shrinkage is not present. Moreover, it is also challenging to deploy a machine learning framework directly to perform predictive analysis due to the sparse and noisy nature of the available experimental dataset. In this paper, we study and propose a method to combine the physics-based knowledge and the predictive ability of deep regression neural networks to mitigate the shortcomings of the existing models. We introduce a novel data augmentation technique that utilizes physics based knowledge to improve the accuracy while maintaining the characteristics of autogenous shrinkage in its predictions simultaneously. Using state-of-the-art B4 model, a genetic algorithm, and a deep neural network trained using raw data for comparison, we show that the proposed methods help improve the accuracy of the model as compared to other methods. We also observe that the proposed method is able to successfully learn and predict the swelling component of the shrinkage strain curve as well, which cannot be predicted using the existing state-of-the-art models.

IC3 2023, August 3–5, 2023, Noida, India

ACM ISBN 979-8-4007-0022-4/23/08…$15.00
https://doi.org/10.1145/3607947.3607980

CCS CONCEPTS

• **Applied computing** → **Chemistry**; • **Computing methodologies** → **Neural networks**.

KEYWORDS

Physics Based Data Augmentation, Deep Learning, Deep Regression, Predictive Modeling, Autogenous Shrinkage

ACM Reference Format:
Vishu Gupta, Yuhui Lyu, Derick Suarez, Yuwei Mao, Wei-keng Liao, Alok Choudhary, Wing Kam Liu, Gianluca Cusatis, and Ankit Agrawal. 2023. Physics-based Data-Augmented Deep Learning for Enhanced Autogenous Shrinkage Prediction on Experimental Dataset. In *2023 Fifteenth International Conference on Contemporary Computing (IC3-2023) (IC3 2023), August 3–5, 2023, Noida, India.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3607947.3607980

1 INTRODUCTION

While the inception of concrete dates back thousands of years, it is still one of the most highly used materials in the construction industry. Concrete shrinkage is of particular interest to researchers as it plays a major role in the sustainability, durability and serviceability of concrete structures [22]. Concrete shrinkage refers to the reduction in volume through time. It is the time-dependent deformation of concrete, which is caused by water movement within a concrete's porous structure and chemical reactions. It can lead to cracks that directly affect durability and serviceability. There are many types of shrinkage that affect concrete such as drying shrinkage, autogenous shrinkage, chemical shrinkage, carbonation shrinkage, thermal shrinkage, and plastic shrinkage [41]. In this work, we mainly focus on the autogenous shrinkage [28] which refers to reduction of apparent volume of concrete under seal and isothermal conditions [25, 38]. Autogenous shrinkage has always been neglected as it causes relatively minor deformations in conventional concrete

compared to those caused by other types of shrinkage [8]. However, with the increasing use of high-performance concrete (HPC), autogenous shrinkage is becoming more and more important [25, 32] since the internal moisture for HPC is insufficient to fully hydrate cement particles due to a low water-to-cement ratio and hence the probability of autogenous shrinkage may increase [37, 39]. In addition to autogenous shrinkage, another phenomenon called as swelling can also occur in the cement. Swelling [6, 37] is the increase in volume of concrete caused by the hydration reaction, usually occurring at short time scales.

Although autogenous shrinkage has comparatively less visibility in the field of concrete shrinkage, several studies proposing models to predict autogenous shrinkage have been published in this field [7, 11, 16, 17, 23, 33]. Among the models proposed to predict autogenous shrinkage (ACI, B3, GL00, MC10, MC99, B4), the recalibrated B4 model [23], a statistically obtained model using the available datasets is shown to perform the best. However, none of them agree well with all the experimental observed phenomena. The main reasons are 1) the mechanism of autogenous shrinkage is not well understood, so we do not have a clear theory on how the autogenous shrinkage mechanisms work in cementitious materials exactly or how it can be correlated with microstructural characteristics, 2) some factors have not been taken into account such as temperature and the type of supplementary cementitious materials, 3) some models cannot exactly capture the interaction among different factors and often ignore the connection between various shrinkage mechanisms [39]. It is also challenging to deploy a machine learning framework directly to perform predictive analysis [2, 3, 10, 18, 19, 26, 30] due to the sparse and noisy nature of the available experimental dataset.

Our goal in this work is to design a deep learning framework that can maximize the predictive ability of the model where the prediction follows the physics behind the nature of the shrinkage strain curve by using only compositional features as model input. We introduce the idea of training the model using data augmented from physics-based knowledge for deep regression networks. Several works have proposed various data augmentation methods to improve the performance of the model [14, 15, 24, 36]. Work in [36] uses the theory of normal variance-mean mixtures to derive a data-augmentation scheme for a class of common regularization problems for non-Gaussian regression. Discacciati et. al. [15] proposes a command, penlogit for approximate bayesian logistic regression using penalized likelihood estimation via data augmentation where the proposed command automatically adds specific prior-data records to a dataset. The method employed in [14] proposes a time series augmentation method, using generative models and checks the viability of augmenting multivariate time series with exogenous inputs. [24] deploys a framework called MixRL, where a data augmentation meta learning framework for regression learns how many nearest neighbors it should be mixed with for each example to maximize the model performance using a small validation set which is achieved using Monte Carlo policy gradient reinforcement learning.

There has also been a major focus to incorporate physics-based knowledge during different phases of training models to improve the performance [4, 29, 31, 40]. Work in [4] used a comprehensive physics based mathematical model based on an unsteady, two dimensional solid fuel and the gas domains to predict the regression rate in solid fuels for hybrid propulation. Majda et. al. [29] introduced a new class of physics constrained multi-level quadratic regression models to build reduced stochastic models from data of nonlinear systems which has advantages such as incorporating memory effects in time and the nonlinear noise from energy conserving nonlinear interactions. Physics-informed CoKriging proposed in [40] is a GPR-based multi-fidelity method, a modified version of the recently developed physics-informed Kriging (PhIK) which integrates simulation results and observation data efficiently. [31] developed a diagnostic tool for operations and maintenance cost reduction application where physics-based diagnostic models are used for reactor feed pumps and motors which uses data augmented from real-time plant data for model training.

There are also some works that study the effect of both the data augmentation and physics-based model training [5, 12, 13, 34]. Work in [12] leverages sensor physics along with large amounts of readily available background data by inserting observations of target signatures under clutter-free conditions into a cluttered scene in a way consistent with the physics governing the sensor to improve discrimination performance. Omigbodun et. al. [34] introduces a new physics-based data augmentation for false-positive reduction in the automatic detection of lung nodules, which can emulate new computed tomography data acquisition protocols in two forms. Semantic physics-based data augmentation method proposed in [5] performs segmentation on the esophagus in both planning CT and cone beam CT using 3D convolutional neural networks that can generalize well across modalities thus improving the accuracy of treatment setup and response analysis. [13] uses a novel physics-based data augmentation strategy by synthesizing a large dataset of perfectly/inherently registered pCT and synthetic-CBCT pairs for locally advanced lung cancer patient cohort, which are then used in a multitask three-dimensional (3D) deep learning framework to simultaneously segment and translate real weekly CBCT images to high-quality pCT-like images. Although data augmentation and physics-based model training has been widely used with data that are images/continuous in nature or work that have access to physics-based instruments or computational technique to perform acquisition protocols different from the training set to acquire large amount of dataset, to the best of our knowledge, no previous work investigates physics-based data augmentation method using only the limited dataset with experimental noises for building deep regression networks composed of fully connected layers for numerical vector inputs.

In this paper, we analyze and propose a efficient model training method performed using data augmented from physics-based knowledge for deep regression networks composed of fully connected layers using numerical vectors as inputs. We propose a novel physics-based data augmentation method where we utilize a scientific formula based on domain knowledge and understanding. We compare the proposed model trained using data augmented from physics-based knowledge against state-of-the-art B4 model, statistical equation obtained using genetic algorithm, and deep learning model trained using the raw experimental dataset. First, we focus on the design problem of predicting the autogenous shrinkage from an input vector composed of compositional features with NU database [22] containing experimental shrinkage strain test curves that

does not have swelling component. We also perform a study where we observe the impact of number of data augmented per curve to perform the physics-based data augmentation of the training dataset and deploy genetic algorithm to perform symbolic regression and derive a mathematical equation to predict autogenous shrinkage to compare against our proposed method.

Our proposed physics-based data augmentation method achieves significantly better result in terms of test error than the state-of-the-art B4 model, statistical equation obtained using genetic algorithm, and deep learning model trained using the raw data. We also perform a stringent test by evaluating the performance of the proposed method on the full dataset by including the test curves which show swelling and found that our proposed approach consistently outperforms the B4 model and deep learning model trained using the raw data on the prediction tasks. Finally, we use feature importance function of multiple machine learning algorithms to find out the most influential compositional factors of autogenous shrinkage. Overall, the proposed physics-based data augmentation method provided more accurate model as compared to the state-of-the-art B4 model and deep learning model trained using the raw data on experimental datasets, and is expected to be widely useful for fast and accurate predictive modeling on small experimental datasets which are sparse and noisy in nature.

2 PROBLEM STATEMENT

2.1 Autogenous Shrinkage

Autogenous shrinkage [28] refers to reduction of apparent volume of concrete under seal and isothermal conditions [25, 38]. It is dependent on a complex array of factors including the mix design of concrete, aggregate content, cement content, curing temperature, and the type and percentage of supplementary cementitious materials. Additionally, some of these parameters are highly related and interdependent, which causes the complexity of autogenous shrinkage. In addition to autogenous shrinkage, another phenomenon called as swelling can also occur in a cement. Swelling [6, 37] is the increase in volume of concrete caused by the hydration reaction, usually occurring at short time scales. Autogenous shrinkage curves can contain both the typical decrease in strain over time with a small increase in strain at short time scales due to swelling.

2.2 Experimental Data

Since the mechanism of autogenous shrinkage is not well understood, obtaining simulation data is not possible. Hence, in order to study and obtain a predictive model for shrinkage response which is affected by various parameters such as compositional parameters, curing/processing conditions, etc., one must solely rely on experimental data. Also, due to the nature of a shrinkage experiment (which involves maintaining a sealed sample for a long period of time), there is only limited data available in literature, with few parameters reported making it harder to use predictive modelling directly to solve the problem. The most common and reliable reported parameters are the compositional parameters that are water to cement ratio (w/c), aggregate to cement ratio (a/c), specific binder percentages ($SiO2$, $flyash$, $slag$, $filler$, etc.) and cement content/type.

Table 1: List of different types of input features used to create model to predict shrinkage response.

Type	Notation	Description
Compositional Feature	w/c	Water to cement ratio
	a/c	Aggregate to cement ratio
	c	Cement mass
	SiO_2, $Slag$	Binder weight percentage
	k_Y	Cement type
Alternate Compositional Feature	wb	Water to binder ratio = water/(cement+fly ash+SiO_2)
	ab	Aggregate to binder ratio = aggregate/(cement+fly ash+SiO_2)
Time	t	Shrinkage response time

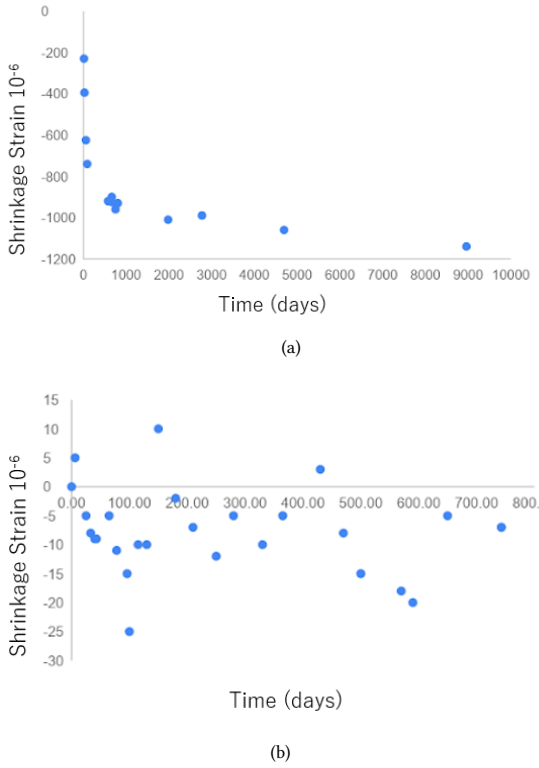


Figure 1: The figures shows what typical autogenous shrinkage curves look like. (a) shows an example of neat curve and (b) shows an example of noisy curve.

Table 1 depicts the list of different types of compositional parameters that are used to predict autogenous shrinkage (note that various dimensionaless numbers can be made from composition parameters). In this study, we use a database with experimental shrinkage curves (NU database) [22].

Figure 1 depicts what typical autogenous shrinkage curves in the database look like. Ideally, there is a slight increase in the strain at the early stage which is contributed by swelling (may not be seen for every autogenous shrinkage curve) and then a sharp decrease in strain followed by a plateau which is contributed by shrinkage.

Although there are some neat curves which are available in the NU database (as shown in Figure 1 (a)), due to various factors related to experiment error, many of the data are very noisy (as shown in Figure 1 (b))

Figure 2 shows the summary of the NU database used for our study. From the 319 unique test curves plotted in Figure 2 (a), we can observe that while a large majority of the curves follow the typical pattern, some (that goes above the zero mark on the y-axis) exhibit swelling. Figure 2 (b) shows the histogram of the number of data points for each autogenous shrinkage curve, and on average, we observe 24 points per test curve. Figure 2 (c) shows the histogram of the final time of experimental recording for each autogenous shrinkage curve, and the average time span during which the shrinkage was recorded was 262 days, with few tests that span between 1500 and 3000 days. Careful inspection of NU database also shows that some of the curves does not have compositional parameter data available or are too noisy and unreliable to use directly.

2.3 Existing Predictive Model

Autogenous shrinkage is a highly complex mechanism and is hard to model accurately. Several models have been proposed by domain scientists to predict autogenous shrinkage such as B4 model [23]. However, they do not perform well in all cases for several reasons. First, these models cannot capture the faster and short scale shrinkage change at early ages which happens due to the contribution of swelling. Second, existing models do not have a mechanism to identify the most influential factors on autogenous shrinkage making it harder for the statistical models to give accurate prediction for a wide range of cases. Third, these models were generally proposed based on empirical experiences and domain knowledge, and their coefficients are often derived by using statistical optimization methods to fit from certain databases.

From Figure 3 which shows the comparison between prediction of B4 model and experimental data of autogenous shrinkage for test curves without swelling, we see that they agree relatively well for the given time span. However, when we compare the prediction of B4 model and experimental data for those test curve which shows swelling (which has the opposite sign to that of pure autogenous shrinkage) as done in Figure 4, we can observe that the prediction of B4 model of total autogenous shrinkage (sum of pure autogenous shrinkage and swelling) is not accurate. Hence, we propose a model which not only improves the predictive accuracy but also produces results which follows the nature of the autogenous shrinkage curve by incorporating the data augmented using physics-based knowledge during the model training process.

3 METHOD

We next describe how we build deep regression models, composed of multiple fully connected layers, for autogenous shrinkage prediction with composition-based numerical vectors as inputs. We first introduce state-of-the-art B4 model, a statistical model obtained via optimum fitting of the available experimental dataset. Next, we introduce two types of the deep learning models used in our work based on the inputs used to train the models. First, we describe straight network (SNet) which uses raw experimental data as the input. Second, we describe our novel B5 Network (B5Net), in which physics guided augmented data is used to train the network. We use the B4 model and SNet as baseline models for comparison against the B5Net.

Table 2: Cement type factor

Cement Type (CEB equivalent)	N (R)	H (RS)	M (N)	L (SL)
k_γ	1.0	1.2	0.85	0.4

3.1 B4 model

Based on the available experimental datasets, Rasoolinejad et al [37] proposed the B4 model, which characterizes the autogenous shrinkage by a power function:

$$\epsilon_{au} = k_\gamma k_s C \left(\frac{t}{1day}\right)^n$$

Here t is the time (in days) (measured from the moment of set), C and n are empirical dimensionless parameters, k_γ is a cement factor listed in Table 1, and k_s accounts for the effect of additives such as slag and SiO_2.

$$C = \frac{100}{\left(\frac{w}{c}\right)^{2.5} + \left(\frac{\frac{a}{c}}{10}\right)^{1.5}} \tag{1}$$

$$n = p + q \ln C \quad where \quad p = 1.2 - 0.1\left(\frac{a}{c}\right) \quad \& \quad q = -0.14 + 0.005\left(\frac{a}{c}\right) \tag{2}$$

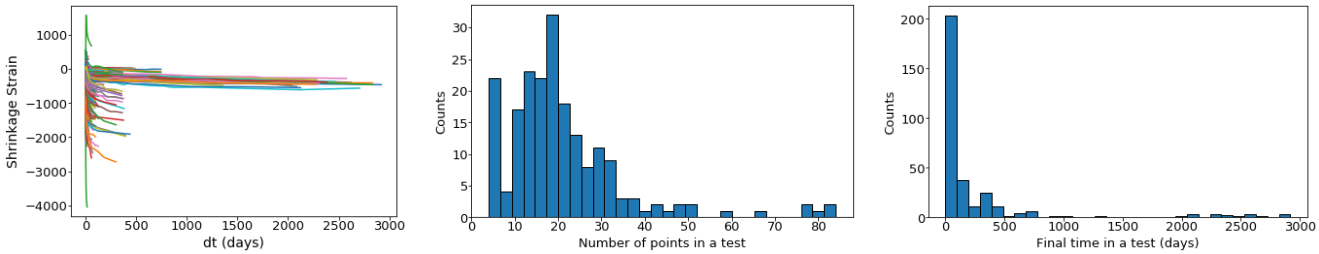
$$k_s = \left\{1 + 3\left(\frac{SiO_2}{Cement}\right)\right\}\left\{1 + 2\left(\frac{Slag}{Cement}\right)\right\} \tag{3}$$

Due to the limited understanding of the mechanisms by which concrete's main compositions affect autogenous shrinkage, the authors defined, based on empirical experiences, Equations (1), (2), and (3) to characterize the relationships between the above parameters (i.e., k_γ, k_s, C and n) and concrete composition properties (i.e.,wc, ac, $slag$ and $cement$). The equation was calibrated by fitting the NU database on experimental shrinkage curves with statistical optimization methods. The B4 model is used as a baseline to which we will compare our data-driven methods of prediction.

3.2 Deep learning model

We use two types of input to train our deep learning model i.e. using raw data and data augmentation. The base model architecture used in this work is formed by putting together a series of seven stacks, each composed of one or more sequences of two basic components with the same configuration. Since the input is a numerical vector, the model uses a fully connected layer as the initial layer in each sequence. Next, ReLU [1] is used as the activation function after the fully connected layer. The detailed architecture for the network is illustrated in Figure 5.

3.2.1 Raw data. The neural network architecture uses raw experimental data as the input (wb, ab and t) and thus learns the experimental error and noise that comes along with the limited dataset. We refer to this network as a straight network (SNet).



(a) (b) (c)

Figure 2: The figures shows the summary of the NU database used for our study. In (a) we plot all the autogenous shrinkage curves, (b) shows histogram of the the number of data points for each autogenous shrinkage curve and (c) shows the histogram of the final time of experimental recording for each autogenous shrinkage curve.

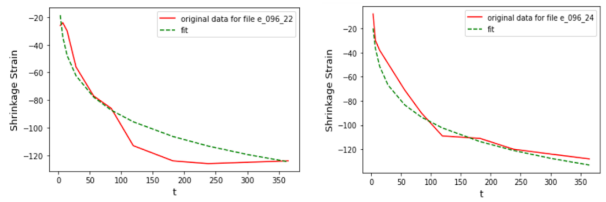


Figure 3: The figures shows the actual vs predicted shrinkage using B4 model on test curves without swelling

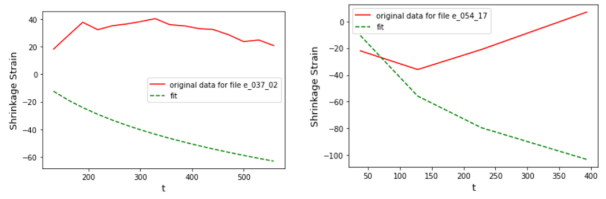


Figure 4: The figures shows the actual vs predicted shrinkage using B4 model on test curves with swelling

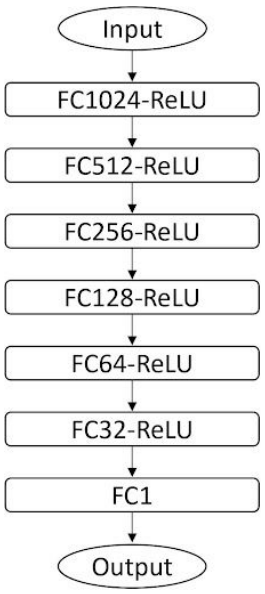


Figure 5: The base model architecture used in this work. Each "layer" is a fully-connected neural network layer with size depicted in each of the blocks followed by ReLU.

3.2.2 Data augmentation. Due to limited size and the presence of possible experimental error in the autogenous shrinkage database, it is challenging to deploy deep neural networks and create a model that successfully captures the nature of the shrinkage and swelling. The existing models used for autogenous shrinkage prediction only takes shrinkage into account, which limits its general use. To solve this issue, we introduce a novel technique of using B5 model-based augmented data to train the neural network instead of raw experimental data. B5 model, which superposed swelling function to predict the autogenous shrinkage ϵ_{auto} is defined as follows:

$$\epsilon_{auto} = \varepsilon_{s\infty}\left(\frac{wb}{1+ab}\right)\left(\left(\frac{wb}{4}\right)\frac{1}{1+\left(\frac{\tau_{sw}}{t}\right)^m} - \left(1-\frac{wb}{4}\right)\frac{1}{1+\left(\frac{\tau_{au}}{t}\right)^n}\right)$$

where $\varepsilon_{s\infty}$ is the ultimate shrinkage value, τ_{sw} is swelling halftime , τ_{au} is autogenous shrinkage halftime, m and n are constants, wb is water to binder ratio, and ab is aggregate to binder ratio. wb and ab are relevant to concrete main compositions, which are known a priori. However, for other parameters, $\varepsilon_{s\infty}$, τ_{sw}, τ_{au}, m, n in the above equation, need to be determined based on the NU database. Hence, we use the raw experimental data on the B5 model to perform curve fitting for each shrinkage set and calculate the value of parameters for the B5 model. The average R^2 for the curve fitting using B5 is 0.83 (some curves with very bad fits were not included). Note that the B5 model cannot be directly used for autogenous shrinkage prediction as most of the parameters are not determined for general use purposes.

Figures 6 and 7 show the B5 predictions for some tests without swelling and with swelling data respectively. Note that the five parameters fitted are unique to each curve. We then perform data augmentation using the B5 model and parameters for each shrinkage set, and by default create 100 data points for each test curve.

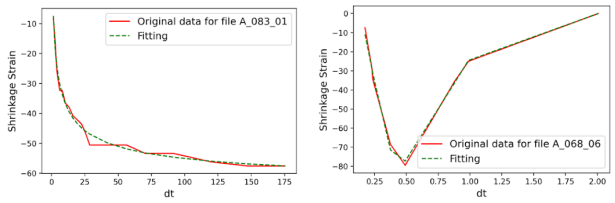


Figure 6: The figures shows the actual vs predicted shrinkage using B5 model on test curves without swelling

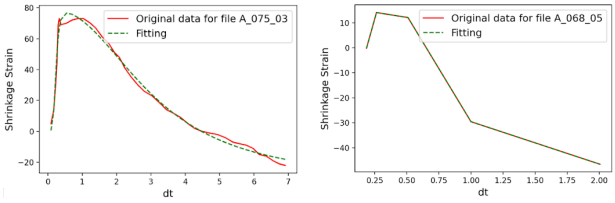


Figure 7: The figures shows the actual vs predicted shrinkage using B5 model on test curves with swelling

Finally, we use the augmented data to train the network. This way, we can make the neural network learn the nature of the shrinkage strain curve for autogenous shrinkage by bypassing sparse and noisy nature of the dataset.

4 EMPIRICAL EVALUATION

In this section, we present a detailed analysis and evaluation of the proposed deep learning (DL) model trained using data augmentation from physics-based knowledge. We will proceed in several steps. First, we perform our evaluation of the proposed model for the design problem and compare its performance against B4 model, statistical equation obtained using genetic algorithm and DL model trained using raw data when applied to the subset of the experimental shrinkage curves database by excluding the curves in which swelling is observed. Next, we perform an stringent test by evaluating the performance of the proposed method on the full experimental shrinkage curves database. Finally, we use the feature importance function of the machine learning algorithm to find the most influencing compositional factors of autogenous shrinkage. Before presenting the results, we discuss the experimental settings and datasets used in this work.

4.0.1 Experimental Settings. We implement deep learning models with Python using Keras [9] framework. In this study, we implement a 7-layered neural network as the deep neural network model architecture to perform model training as shown in Figure 5. The number of neurons in each layers are fixed by referring to [20, 21, 27].For genetic algorithm we use an extension of scikit-learn [35] called gplearn to perform symbolic regression. We used mean absolute error (MAE) as loss function and the error metric for all the results. We used early stopping with patience of 200 which stops the training if the validation loss does not improve for 200 epochs. We use B4 model for comparison against our proposed model since it demonstrates a powerful ability to predict the autogenous shrinkage from compositional inputs alone.

Table 3: Prediction performance benchmarking for the prediction task in "Non-Swelling Dataset Analysis" for the design problem. The table shows the test MAE for all the model used in the analysis.

Model	Test MAE
B4 model	115.23
SNet	88.96
B5Net	**79.86**

4.0.2 Datasets. In this work, we create and use two datasets from the raw experimental data [22]. First a dataset that contains both swelling and non-swelling data, called the full dataset, and second, a dataset that only contains non-swelling data, non-swelling data set. For both of these datasets a 85:15 training:testing split (approximately) is performed based on the number of curves. This results in the full dataset having 218 curves for training (5368 data points) and 41 curves for testing (916 data points). For the non-swelling dataset, there are 135 curves for training (2640 data points) and 26 curves for testing (547 data points).

4.1 Design Problem

First, we analyze the impact of different choices of model and data used by evaluating the proposed methods on the design problem. We perform autogenous prediction from an input vector composed of compositional features.

4.1.1 Non-Swelling Dataset Analysis. First, we evaluate the performance of our proposed method on the non-swelling dataset, a subset of the dataset which does not contain curves with swelling. We perform autogenous shrinkage prediction using input vector composed of compositional features. For physics-based data augmentation, we create 100 data points for each of the curves in this analysis.

In Table 3, we show the test MAE for all the model used in the analysis. When performing the predictive analysis for the non-swelling dataset, the results shows that the deep learning model performs better as compared to the B4 models. Among the deep learning models, B5Net which uses the data augmented using the physics-based knowledge as input for the model training performs better as compared to the SNet which uses raw experimental data as input for the model training.

As the accuracy improvement gained from introducing the data augmented using the physics-based knowledge does not seem to be significant just by looking at the MAE of the two models, we evaluate its significance by plotting the predicted curves obtained by using all the models used in the analysis for some of the tests cases in Figure 8. From Figure 8 we can clearly see the benefit of using the proposed B5Net model as compared to B4 model and SNet. As the B4 model does not take into account the swelling component of the autogenous shrinkage, the prediction is quite off for the whole curve. For SNet, as the model learns the noises and experimental error from the raw data during the training phase, the predicted curves also

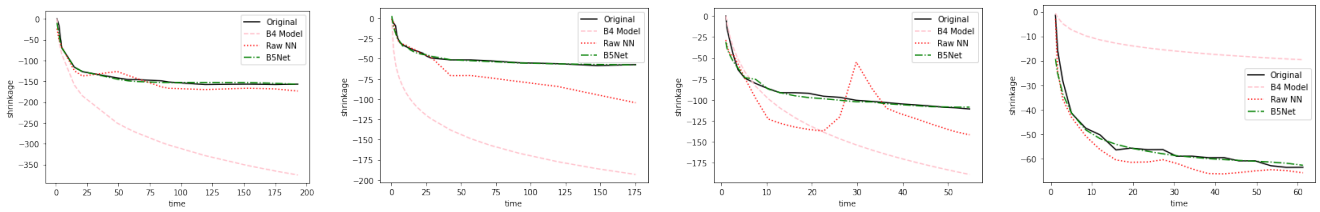


Figure 8: Predicted curves for some of the test cases for "Non-Swelling Dataset Analysis"

Table 4: Prediction performance benchmarking for the prediction task in "Impact of Data Augmentation" for the design problem. The table shows the test MAE obtained using different number of data augmented after training B5Net used in the analysis.

No. of Data Augmented Per Curve	Test MAE
25 data points	85.70
50 data points	83.37
75 data points	82.18
100 data points	**79.86**
125 data points	80.73
150 data points	81.60
175 data points	80.25
200 data points	82.72

reflect that and shows prediction which does not follow the nature of the shrinkage strain curve for autogenous shrinkage making is not useful for further analysis (although quantatively there is not much difference in the MAE as compared to B5Net). These results demonstrate that deep learning, along with our proposed B5Net model, can help construct a robust model for predicting autogenous shrinkage which can enforce the prediction to follow the nature of the shrinkage strain curve for autogenous shrinkage.

4.1.2 Impact of Data Augmentation. So far, we created 100 data points for physics-based data augmentation to train the deep learning model. Here, we first perform the data augmentation using different values ranging from 25 to 200 data points for each of the curve in the non-swelling dataset. We then perform the model training of a deep learning model using the data augmented using physics-based knowledge for autogenous shrinkage to see how the performance of the proposed method is effected by the choice of number of data augmentation on the accuracy of the model.

Table 4 shows the test MAE obtained using different number of data augmented after training B5Net. From Table 4 we can see that initially the test MAE decreases as we increase the number of data points augmented using physics-based knowledge till a certain number (100 data points for this dataset). Then we see a stagnation in the performance of the autogenous shrinkage prediction. After this initial exploration, we use 100 data points as the number of data points used to performed data augmentation from physics-based knowledge for the rest of the analysis.

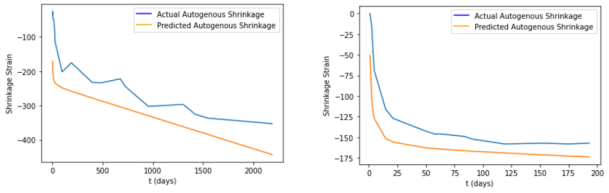


Figure 9: The figures shows the actual (blue) vs predicted (orange) shrinkage using equation (4) on test curves without swelling

4.1.3 Against Genetic Algorithm. Finally, we deploy genetic algorithm to perform symbolic regression and derive an statistical equation to predict autogenous shrinkage to compare against our proposed physics-based data augmentation method. After performing symbolic regression on the training data and simplifying the result we get the following equation:

$$\epsilon_{gen} = 0.512\left(2.718^{2.718^{\frac{X_0}{X_1}}} \cdot 2.718^{2.718^{X_0} - 0.594X_1} \cdot Y - 0.245X_0 \cdot Z\right) \tag{4}$$

where $Y = 2.718^{-0.594X_1 - \frac{2.230}{X_2^2}\left(-46.875 - \frac{0.909}{X_0}\right) - 15.625X_1}$, $Z = X_2 + 2.718^{\frac{-0.214}{X_1}}$ and X_0, X_1, X_2 are the features to perform symbolic regression. The average R^2 for the curve fitting using equation (4) is 0.7. Some of the examples of curve fitting performed using equation (4) is shown in Figure 9.

After performing predictive analysis using the equation obtained using genetic algorithm we obtain an MAE of 96.39 which shows that although genetic algorithm provides a close-form prediction and is able to outperform B4 model, it is still not able to perform better compared to deep learning models SNet and B5Net (from Table 3). Moreover, as the raw data gets more noisy, several problems arise for the genetic algorithm, such as increase in the complexity of the resulting equation, need to perform symbolic regression every time the variation in the training set changes and having no physics-based background behind the equation obtained. Also, as we know that model trained using data augmented from physics-based knowledge (i.e. B5 model) performs better for predicting autogenous shrinkage, we will not use the equation obtained from genetic algorithm for the rest of the work to perform predictive analysis.

Table 5: Prediction performance benchmarking for the prediction task in "Full Dataset Analysis". The table shows the test MAE for all the model used in the analysis.

Model	Test MAE
B4 model	202.23
SNet	85.30
B5Net	**82.05**

4.1.4 Summary of Insights. We derive the following insights from our experiments by training different models for performing deep regression from numerical vector inputs for autogenous shrinkage on the design problem.

- **Model Training** From all the experiments performed, we observe that it is better to use deep learning models as compared to state-of-the-art B4 model and statistical equation obtained using genetic algorithm when performing regression for predicting autogenous shrinkage on small experimental dataset. Among the deep learning models, it is better to train the model from data augmented using the physics-based knowledge as compared to the raw data as the model is able to automatically mitigate the sparse and noisy nature of the experimental datasets and capture the underlying nature of the shrinkage strain curve accurately.
- **Data Augmentation** After performing model training using different number of data points for data augmentation using the physics-based knowledge, we find that initially test MAE decreases as we increase the number of data augmented until a certain point (100 data points for our dataset) and then stagnates after that.
- **Genetic Algorithm** Genetic algorithm can be used to perform predictive analysis in absence of any physics-based knowledge such as B5 model for our work as the model trained using data augmented from physics-based knowledge obtained using the domain expertise have shown better performance in predicting autogenous shrinkage.

We believe that the proposed method can help build more accurate and robust predictive models than the traditional models derived based on empirical and domain knowledge. The proposed method can also be easily adapted to perform classification by modifying the architecture, i.e., using softmax as the activation of the last layer and cross-entropy as the loss function.

4.2 Full Dataset Analysis

For the analysis of the full dataset we first evaluate the performance of our proposed method by predicting autogenous shrinkage using the same input vector composed of compositional features on the full dataset containing swelling experimental shrinkage curves as well.

Table 5 shows the test MAE for all the models used in the analysis. When performing the predictive analysis for the full dataset, the results show that the deep learning model still performs better as compared to the statistical models which is traditionally used to predict autogenous shrinkage. Even among the deep learning models, B5Net which uses the data augmented using the physics-based knowledge as input for the model training performs better as compared to the SNet which uses raw experimental data as input for the model training.

Table 6: Feature Importance value for four compositional parameters associated with autogenous shrinkage using multiple machine learning algorithms.

Model	Input Parameter			
	w/c	a/c	c	t_{dry}
Linear Regression	-457.528	20.775	-0.808	-0.001
CART	0.056	0.628	0.256	0.060
Random Forest	0.160	0.517	0.259	0.064
XGBoost	0.152	0.2563	0.525	0.066

As we can see that the test MAE between SNet and B5Net are still close to each other, we evaluate the significance of introducing the data augmented using the physics-based knowledge by plotting the predicted curves obtained by using all the models used in the analysis for some of the tests cases in Figure 10, which shows that B5Net is able to predict the autogenous shrinkage curve more accurately as compred to B4 model and SNet by following the nature of the shrinkage strain curve for autogenous shrinkage. These results demonstrate that deep learning along with physics-based data augmentation, as encapsulated by our proposed B5Net model, can help construct a robust model for predicting autogenous shrinkage even with a complex and noisy dataset.

4.3 Feature Importance Analysis

Finally, we use "feature importance" function of multiple machine learning algorithms (Linear Regression, Classification And Regression Trees (CART), Random Forest and XGBoost) to find out the most influencing compositional factors of autogenous shrinkage among water to cement ratio (w/c), aggregate to cement ratio (a/c), cement content/type (c), and time it takes to remove the water content in the cement to prepare for experiment (t_{dry}), which are four compositional parameters associated with autogenous shrinkage.

From Table 6, we can see that the the most influencing compositional factors of autogenous shrinkage are the aggregate to cement ratio and cement content/type such as blast-furnace slag and silica fume. Among them aggregate to cement ratio is shown as the most influencing compositional factor by three out of four algorithms. The next important factor is water to cement ratio. This knowledge can help domain scientist better understand the functioning of the autogenous shrinkage and hence advance the field. We believe that the proposed method can also be used to better understand other small experimental datasets which can show noises related to experimental errors.

5 CONCLUSION AND FUTURE WORK

In this work, we have studied how to predict autogenous shrinkage, which is an unavoidable volume reduction in sealed concrete specimens because of the self-desiccation of concrete. Using the NU database, we have shown that a new model is necessary to account for swelling as the B4 model is only applicable for non-swelling data. We have also studied the autogenous shrinkage prediction accuracy

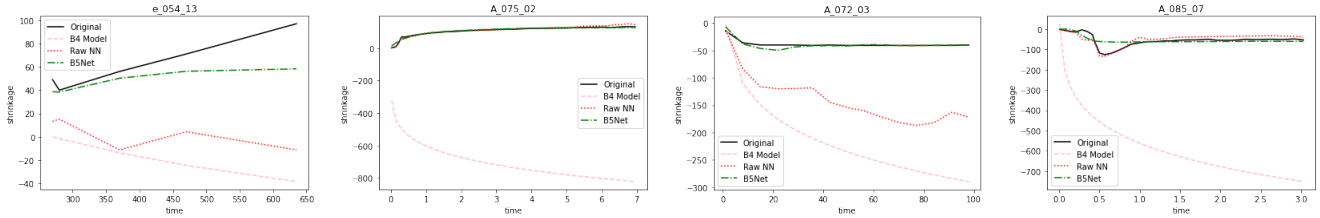


Figure 10: Predicted curves for some of the test cases for "Full Dataset Analysis"

of a new analytical model, B5 model, which aims to predict the total autogenous shrinkage with swelling. However, the empirical parameters in the B5 model are difficult to decide due to the limited theories. To solve this problem, we present B5Net, a model trained from the data augmented using the physics-based knowledge for the predictions of autogenous shrinkage and show that it can enhance the accuracy of the model and perform prediction based on the nature of the shrinkage strain curve. Additionally, we use a function of machine learning algorithm "feature importance" to find out the most influencing compositional factors of autogenous shrinkage. The insights obtained from this work can help in building predictive models for other with applications small experimental datasets containing noises derived from experimental error with numerical vector inputs. The code, data, and models developed in this work are publicly available at https://github.com/GuptaVishu2002/B5Net to the community to facilitate reproducibility and further building upon this work.

In future, we plan to explore the effect of the proposed model training from data augmentation using physics-based knowledge on other data mining problems such as classification, and on problems in other domain with sparse and noisy dataset. We also plan to formulate a method to establish relationship between B5 coefficients and compositional parameters to derive an equation between the coefficients and compositional parameters.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Zdenek P Bazant and Dr. Ahmet Abdullah Dönmez for helpful discussions. This work is supported in part by the following grants: National Institute of Standards and Technology (NIST) award 70NANB19H005; Predictive Science and Engineering Design Cluster (PS&ED, Northwestern University); Department of Energy (DOE) awards DE-SC0019358, DE-SC0021399; NSF award CMMI-2053929, and Northwestern Center for Nanocombinatorics.

REFERENCES

[1] Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375* (2018).

[2] Ankit Agrawal and Alok Choudhary. 2016. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Materials* 4, 5 (2016), 053208.

[3] Ankit Agrawal and Alok Choudhary. 2019. Deep materials informatics: Applications of deep learning in materials science. *MRS Communications* 9, 3 (2019), 779–792.

[4] K Akyuzlu, Antonis Antoniou, and Michael Martin. 2002. Determination of regression rate in an ablating hybrid rocket solid fuel using a physics based comprehensive mathematical model. In *38th AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit*. 3577.

[5] Sadegh R Alam, Tianfang Li, Pengpeng Zhang, Si-Yuan Zhang, and Saad Nadeem. 2021. Generalizable cone beam CT esophagus segmentation using physics-based data augmentation. *Physics in Medicine & Biology* 66, 6 (2021), 065008.

[6] ZP Bažant, A Donmez, E Masoero, and S Rahimi Aghdam. 2015. Interaction of concrete creep, shrinkage and swelling with water, hydration, and damage: Nano-macro-chemo. In *CONCREEP 10*. 1–12.

[7] Zdenêk P Bažant and Sandeep Baweja. 1995. Justification and refinements of model B3 for concrete creep and shrinkage 1. statistics and sensitivity. *Materials and structures* 28, 7 (1995), 415–430.

[8] Zdenek Pavel Bazant, MIlan Jirasek, MH Hubler, and Ignacio Carol. 2015. RILEM draft recommendation: TC-242-MDC multi-decade creep and shrinkage of concrete: material model and structural analysis. Model B4 for creep, drying shrinkage and autogenous shrinkage of normal and high-strength concretes with multi-decade applicability. *Materials and structures* 48, 4 (2015), 753–770.

[9] François Chollet et al. 2015. Keras. https://github.com/fchollet/keras.

[10] Kamal Choudhary, Daniel Wines, Kangming Li, Kevin F Garrity, Vishu Gupta, Aldo H Romero, Jaron T Krogel, Kayahan Saritas, Addis Fuhr, Panchapakesan Ganesh, et al. 2023. Large Scale Benchmark of Materials Design Methods. *arXiv preprint arXiv:2306.11688* (2023).

[11] CEB-FIP Model Code. 2013. Fib Model Code for Concrete Structures 2010; Ernst & Sohn. *Wiley: Berlin, Germany* (2013).

[12] Miles Crosskey, Patrick Wang, Rayn Sakaguchi, and Kenneth D Morton Jr. 2018. Physics-based data augmentation for high frequency 3D radar systems. In *Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXIII*, Vol. 10628. SPIE, 430–440.

[13] Navdeep Dahiya, Sadegh R Alam, Pengpeng Zhang, Si-Yuan Zhang, Tianfang Li, Anthony Yezzi, and Saad Nadeem. 2021. Multitask 3D CBCT-to-CT translation and organs-at-risk segmentation using physics-based data augmentation. *Medical physics* 48, 9 (2021), 5130–5141.

[14] Sumeyra Demir, Krystof Mincev, Koen Kok, and Nikolaos G Paterakis. 2021. Data augmentation for time series regression: Applying transformations, autoencoders and adversarial networks to electricity price forecasting. *Applied Energy* 304 (2021), 117695.

[15] Andrea Discacciati, Nicola Orsini, and Sander Greenland. 2015. Approximate Bayesian logistic regression via penalized likelihood by data augmentation. *The Stata Journal* 15, 3 (2015), 712–736.

[16] Fédération Internationale du Béton. 1999. Structural Concrete: Textbook on Behaviour, Design and Performance: Updated Knowledge of the CEB-FIP Model Code 1990. *Bulletin* 2 (1999), 35–52.

[17] BS EN et al. 2004. Design of concrete structures. Part 1-1: General rules and rules for buildings (Eurocode 2). *European Committee for Standardization (CEN), Brussels, Belgium* (2004).

[18] Vishu Gupta, Kamal Choudhary, Yuwei Mao, Kewei Wang, Francesca Tavazza, Carelyn Campbell, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. 2023. MPpredictor: An Artificial Intelligence-Driven Web Tool for Composition-Based Material Property Prediction. *Journal of Chemical Information and Modeling* (2023).

[19] Vishu Gupta, Kamal Choudhary, Francesca Tavazza, Carelyn Campbell, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. 2021. Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. *Nature communications* 12, 1 (2021), 1–10.

[20] Vishu Gupta, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. 2022. BRNet: Branched Residual Network for Fast and Accurate Predictive Modeling of Materials Properties. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*. SIAM, 343–351.

[21] Vishu Gupta, Alec Peltekian, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. 2023. Improving deep learning model performance under parametric constraints for materials informatics applications. *Scientific Reports* 13, 1 (2023), 9128.

[22] Mija H Hubler, Roman Wendner, and Zdeneěk P Bažant. 2015. Comprehensive Database for Concrete Creep and Shrinkage: Analysis and Recommendations for Testing and Recording. *ACI Materials Journal* 112, 4 (2015).

[23] Mija H Hubler, Roman Wendner, and Zdeněk P Bažant. 2015. Statistical justification of Model B4 for drying and autogenous shrinkage of concrete and comparisons to other models. *Materials and Structures* 48, 4 (2015), 797–814.
[24] Seong-Hyeon Hwang and Steven Euijong Whang. 2021. MixRL: Data mixing augmentation for regression using reinforcement learning. *arXiv preprint arXiv:2106.03374* (2021).
[25] Ole Mejlhede Jensen and Per Freiesleben Hansen. 2001. Autogenous deformation and RH-change in perspective. *Cement and Concrete Research* 31, 12 (2001), 1859–1865.
[26] Dipendra Jha, Vishu Gupta, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. 2022. Moving closer to experimental level materials property prediction using AI. *Scientific reports* 12, 1 (2022), 1–9.
[27] Dipendra Jha, Vishu Gupta, Logan Ward, Zijiang Yang, Christopher Wolverton, Ian Foster, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. 2021. Enabling deeper learning on big data for materials informatics applications. *Scientific reports* 11, 1 (2021), 1–12.
[28] Charles Garner Lynam. 1934. *Growth and movement in Portland cement concrete.* Oxford University Press.
[29] Andrew J Majda and John Harlim. 2012. Physics constrained nonlinear regression models for time series. *Nonlinearity* 26, 1 (2012), 201.
[30] Yuwei Mao, Mahmudul Hasan, Arindam Paul, Vishu Gupta, Kamal Choudhary, Francesca Tavazza, Wei-keng Liao, Alok Choudhary, Pinar Acar, and Ankit Agrawal. 2023. An AI-driven microstructure optimization framework for elastic properties of titanium beyond cubic crystal systems. *npj Computational Materials* 9, 1 (2023), 111.
[31] Tat Nghia Nguyen, Roberto Ponciroli, Timothy Kibler, Marc Anderson, Molly J Strasser, and Richard B Vilim. 2022. A physics-based parametric regression approach for feedwater pump system diagnosis. *Annals of Nuclear Energy* 166 (2022), 108692.
[32] Minehiro Nishiyama. 2009. Mechanical properties of concrete and reinforcement state-of-the-art report on HSC and HSS in Japan. *Journal of Advanced Concrete Technology* 7, 2 (2009), 157–182.
[33] Japan Society of Civil Engineers (JSCE). 2007. Standard specification for concrete structures. *Structural performance verification* (2007).
[34] Akinyinka O Omigbodun, Frederic Noo, Michael McNitt-Gray, William Hsu, and Scott S Hsieh. 2019. The effects of physics-based data augmentation on the generalizability of deep neural networks: Demonstration on nodule false-positive reduction. *Medical physics* 46, 10 (2019), 4563–4574.
[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[36] Nicholas G Polson and James G Scott. 2013. Data augmentation for non-Gaussian regression models using variance-mean mixtures. *Biometrika* 100, 2 (2013), 459–471.
[37] Mohammad Rasoolinejad, Saeed Rahimi-Aghdam, and Zdeněk P Bažant. 2019. Prediction of autogenous shrinkage in concrete from material composition or strength calibrated by a large database, as update to model B4. *Materials and Structures* 52, 2 (2019), 1–17.
[38] Ei-ichi Tazawa. 1999. *Autogenous shrinkage of concrete.* CRC Press.
[39] Linmei Wu, Nima Farzadnia, Caijun Shi, Zuhua Zhang, and Hao Wang. 2017. Autogenous shrinkage of high performance concrete: A review. *Construction and Building Materials* 149 (2017), 62–75.
[40] Xiu Yang, David Barajas-Solano, Guzel Tartakovsky, and Alexandre M Tartakovsky. 2019. Physics-informed CoKriging: A Gaussian-process-regression-based multifidelity method for data-model convergence. *J. Comput. Phys.* 395 (2019), 410–431.
[41] Hailong Ye and Aleksandra Radlińska. 2016. A review and comparative study of existing shrinkage prediction models for portland and non-portland cementitious materials. *Advances in Materials Science and Engineering* 2016 (2016).