Machine learning and micromechanics as allies to establish composition-property correlations in cement pastes

Tulio Honorio, Sofiane Ait Hamadouche, Amélie Fau

OVERLAY

Journal of Theoretical, Computational and Applied Mechanics

JTCAM

DIAMOND OPEN ACCESS

Identifiers
DOI 10.46298/jtcam.9830
OAI hal-03723418v3

History
Received Jul 23, 2022
Accepted Dec 10, 2022
Published Jan 26, 2023

Associate Editor
Anna Pandolfi

Reviewers
Giovanni Di Luzio
Enrico Masoero

Open Review
OAI hal-03936241

Supplementary Material
See last page

Machine learning and micromechanics as allies to establish composition-property correlations in cement pastes

Túlio Honório, Sofiane Ait Hamadouche, and Amélie Fau

Université Paris-Saclay, CentraleSupélec, ENS Paris-Saclay, CNRS, LMPS – Laboratoire de Mécanique Paris-Saclay, 91190 Gif-sur-Yvette, France

Composition–property correlations are fundamental to understand cement-based materials' behavior and optimize their formulation. Modeling based on fundamental material components constitutes a reliable tool to establish these correlations with the advantage of better exploring the formulation space when compared to the often adopted experimental trial-and-error approaches. In this context, Machine Learning (ML) and Micromechanics-Based (MB) methods have been concurrently used for property prediction from the material composition. We show that these techniques can be allies for establishing composition–property correlations. We focus on predicting the elastic properties of Ordinary Portland Cement pastes but the outlined strategy can be extended to other cement systems. Various microstructural representations are considered in MB estimates including multiscale representations possibly with ellipsoidal inclusions. In contrast, ML predictions do not need any a priori assumption on the material microstructure. Predictions using ML and MB yield similar accuracy when compared against test datasets but ML performs much better regarding the error estimated in training datasets. Working as allies, ML can be deployed to evaluate the (lack of) knowledge over the multi-dimensional parametric domains while micromechanics provides a theoretical background for property data curation and is a tool to make up for missing data in databases.

Keywords: machine learning; micromechanics; ordinary Portland cement paste; elastic properties; data science; early ages.

1 Introduction

Establishing Processing–Composition–(Micro) Structure–Property–Performance correlations is the central paradigm for understanding material behavior in a bottom-up perspective as well as to conceive and optimize materials for tailored applications (Olson 1997). Such correlations are important for cement-based materials since, on the *composition* side, the key ingredients vary largely according to the local availability of resources, *processing* spans lower and higher technology contexts (Wangler et al. 2019), and the design of cement components and concrete structures relies on *property and performance* requirements. Material property prediction having as input the composition is therefore critical to optimize the use of cement-based materials.

MB modeling has been successfully used to unveil Composition–Property correlations for various properties in cement-based materials, including mechanical (Wyrzykowski et al. 2017; Pichler and Hellmich 2011; Sanahuja et al. 2007; Königsberger et al. 2021), transport and thermal (Bary and Béjaoui 2006; Patel et al. 2016; Honório et al. 2018a), and electromagnetic (Guihard et al. 2019; Honório et al. 2020b; Honório et al. 2020a) properties, as well as coupling properties in the thermo-poro-mechanical framework (Ulm et al. 2004; Ghabezloo et al. 2009; Honório et al. 2018a; Honório et al. 2018b). An advantage of MB modeling is the simplicity of computations, which enables assessing various scenarios of interest regarding the composition, uncertainty on phase properties (Honório et al. 2020b), and morphology of phases in a heterogeneous material. However, one may legitimately dispute the pertinence of representing the microstructure of

Nomenclature

Acronyms

DOH	Degree of Hydration
HD	high-density
HS	Hashin-Shtrikman
KHP	Königsberger-Hellmich-Pichler
LD	low-density
LOOCV	Leave-One-Out Cross-Validation
MB	Micromechanics-based
MRE	Mean Relative Error
MT	Mori-Tanaka
OPC	Ordinary Portland Cement
REV	representative elementary volume
RMSE	Root Mean Square Error
SC	Self-Consistent
SCM	Supplementary Cementitious Materials
TJ	Tennis and Jennings

Methods

ANN	Artificial Neural Network
CART	Classification and Regression Trees
DT	Decision Tree
GBT	Gradient Boosted Trees
LR	Linear Regression
ML	Machine Learning
NN	Nearest Neighbors
RF	Random Forest

cement paste under the usual assumptions adopted in analytical homogenization approaches. These assumptions include

(i) a random microstructure (there is evidence that some correlation between phases volume distribution has been quantified using microstructural hydration model (Hlobil 2020)),

(ii) phases being often represented by spherical (or ellipsoidal) inclusions (experimental evidence shows that crystalline phases cement paste are not generally spherical or ellipsoidal),

(iii) perfect interfaces among phases (while some defects may exist), and

(iv) separability of scale (especially considering that heterogeneity size, for example of cement particles, may span various magnitudes).

Numerical homogenization is not immune to the same questioning. In this context, ML arises as a promising tool to directly establish Composition–Property correlations without *a priori* assumptions on the microstructure characteristics (Agrawal and Choudhary 2016). The huge amount of experimental data produced on cement-based materials in the last century can be used to build databases that can be interrogated by ML. As highlighted by Bullard et al. (2019), a "systematic development of structure-property relationships" based on both the "curation of fundamental material component data" and "validated modeling based on fundamental scientific principles" may "revolutionize" the design of cement-based materials. However, as recognized by the authors, such an approach was given comparatively little attention in the concrete research community when compared to the "increasingly laborious trial-and-error exploration of the design space and mixture qualification process". In cement-based materials research, ML has been deployed since the 90's to predict compressive strength (Kasperkiewicz et al. 1995; Yeh 1998; Yeh and Lien 2009; Duan et al. 2013; Young et al. 2019) using frequently ANN. Other methods include support vector machines (Yan and Shi 2010), decision trees (Behnood et al. 2015), evolutionary algorithms (Golafshani and Behnood 2018). Elastic properties have also been extensively studied using ML (Ben Chaabene et al. 2020) with a strong focus on the impact of using recycled aggregates. As input variables, the composition in terms of cement and water content, as well as SCM and admixture mass or volume, are often adopted (Ben Chaabene et al. 2020). The effects of the mineralogical composition of cement and the effects of age (and property development, especially at early ages) are generally omitted.

In this work, a multi-technique modeling approach combining ML and MB methods is proposed to link cement system composition and DOH to the material elastic properties. We tackle specifically the predictions of OPC pastes elastic properties from the composition of the cement (in terms of clinker composition and gypsum fraction w/c) and age, but the outlined strategy can be extended to other cement systems and scales. Since OPC systems are simpler and better experimentally characterized than other cement systems, they are ideal candidates for testing our approach and for demonstrating its feasibility. We explore paths in which ML and MB techniques can be allies, notably in the analysis of experimental databases to evaluate existing experiments and lack of experiments and by providing missing data. Our results contribute to the development of multiscale modeling of cement-based materials informed by the cement composition variability and enhanced by blending data from different research projects. This framework can be used to improve the comprehension of correlations among the composition, microstructure, and properties of cement-based materials.

2 Machine Learning approach and database construction for predicting elastic properties

Knowledge about cement paste and behavior is fundamentally offered through experimental observations. A direct approach for exploiting the large literature is collecting a wide range of published experimental results and using ML methods to predict properties for new compositions based on the training dataset.

2.1 A database construction for cement pastes linking composition and elastic constants

Based on experimental data from the literature (Helmuth and Turk 1966; Haecker et al. 2005; Boumiz 1996; Tamtsia et al. 2004; Wang and Subramaniam 2011; Constantinides and Ulm 2004; Lura et al. 2003; Chamrova 2010; Sun et al. 2007; Maruyama and Igarashi 2014) a dataset with 376 entries is built, which will be used for training and validation. Details on database construction are given in Appendix A. Input in the datasets are cement composition (in terms of clinker minerals and gypsum contents), water-cement ratio, age, and DOH. Other input of interest for the formulation, such as admixtures, curing conditions (including temperature), etc., are not considered because of the lack of full data and the inadequacy of MB methods to date to take into account these factors properly. Of course, future work might focus on introducing these effects as input in ML-based strategies.

The outputs are the elastic constants: E Young, K bulk and G shear moduli, and ν Poisson's ratio. Note that the dimensionality of the manifold can be reduced considering that the elastic constants are linked through simple relations in the case of isotropic materials: $E = 9KG/(3K+G)$; $\nu = (3K - 2G)/(2(3K + G))$; $K = E/(3(1 - 2\nu))$ and $G = E/(2(1 + \nu))$. Also, the age and the DOH can be related using a bijection e.g., a sigmoid function.

Table 1 shows the statistical parameters associated with the training dataset. In the various ML applications for mechanical properties of cement-based materials, the dataset size spans from 74 (Ben Chaabene et al. 2020) up to more than 10,000 (Young et al. 2019) observations, most of the cases with data size in the range 100 to 1,000 observations (Ben Chaabene et al. 2020). The size of the dataset provided here has, therefore, an intermediary size. It can already provide sufficient support for learning but could surely be improved with complementary data in future works.

Data	Variable		Min.	Max.	Mean	St. Dev.	Exceed Kurtosis*	Skewness*
Input	age	[days]	0.12	720	49	124	13.9	3.6
	DOH	[-]	0.03	1	0.5	0.3	−0.7	0.6
	w/c	[-]	0.25	0.8	0.44	0.1	−0.2	0.64
	m_{C_3S}	[%]	24.5	100	60.2	13.8	2.1	−0.7
	m_{C_2S}	[%]	0	61.3	16.6	15.2	2.4	1.6
	m_{C_3A}	[%]	0	12.7	8.1	3.4	−0.7	−0.6
	m_{C_4AF}	[%]	0	12.7	5.8	4.2	−1.5	−0.2
	m_{gypsum}	[%]	0	6.8	2.9	2.9	−1.7	0.3
Output	E	[GPa]	0.22	37.2	11.2	7.8	0.3	0.8
	ν	[-]	0.07	0.49	0.3	0.07	0.88	0.58
	K	[GPa]	0.15	32.2	9.3	5.6	2	1.3
	G	[GPa]	0.07	14.6	4.4	3.1	0.3	0.8

Table 1 Statistical analysis of the cement paste dataset of 365 observations used for training. *dimensionless.

2.2 Machine Learning methods

For *prediction* purposes, the following algorithms are employed:

Linear Regression The output is predicted using a linear combination of the numerical features vector. The conditional probability is computed using a parameter vector estimated from the minimization of a loss function.

Decision Tree A decision tree (i.e., a flow chart structure in which the internal nodes correspond to a test on a feature, while the branches correspond to an outcome of the test) is built using the CART algorithm (Breiman et al. 2017).

Gradient Boosted Trees A prediction model is constructed in the form of an ensemble of trees which is trained sequentially in order to enhance the capability of the previous trees. The implementation adopted is based on the LightGBM algorithm (Ke et al. 2017).

Nearest Neighbors This instance-based learning technique predicts a value by analyzing the nearest neighbors in the feature space.

Artificial Neural Network A neural network is constituted of stacked layers, each associated with simple computation. The information is processed layer by layer, starting at the input layer until the output layer. The neural network is trained in order to minimize a loss function on the training set. A gradient descent method is used to perform this minimization.

Random Forest Various decision trees are constructed and the prediction is made by taking the mean value of the tree predictions based on the bootstrap aggregating algorithm (Breiman 1996) where each decision tree is trained using only a random subset of the features.

Gaussian Process Predictions are made using Bayesian inference on the Gaussian process conditioned to the training data (Williams and Rasmussen 1996). The underlying assumption of the method is that the prediction function can be associated with a Gaussian process defined by its kernel or covariance function. The training phase consists of estimating the parameters of the kernel.

We use Mathematica 13.0 (Wolfram 2021) in which these algorithms are built-in. The numerical cost associated with the creation of the predictor functions is detailed in Appendix B. Methods like ANN, LR, and GP produce a smooth predictors, whilst DT, NN, and RF produce discrete prediction values. The implementation of the various methods in Mathematica leaves the user the possibility to impose the associated parameters or use built-in optimized procedures to determine these parameters. The last option seemed more appropriate for us because it reduced the number of cases to be tested. For reproducibility reasons, we provide the information used by each method as supplementary material. For example, ANN uses two layers for E and G predictions and eight layers for K and ν predictions; DT uses between 23-27 nodes and 12-14 leaves. The specific number of nodes and leaves is provided to GBT and RF.

2.3 Validation process

The performances of the ML methods are estimated using a k-fold cross-validation technique (Bengio and Grandvalet 2004). The training dataset is divided in k folds, i.e., subsets $\mathcal{D}_i$, in which elements are randomly sampled from the dataset. In each fold construction, care is taken so that a given element is not chosen more than once (in order to ensure that the intersection set of all folds is the empty set: $\mathcal{D}_i \cap \mathcal{D}_j = \emptyset, \forall (i,j) \in [1;k]^2$ if $i \neq j$). The predictor is trained on $k-1$ folds and then is used to predict the values in the remaining fold. This operation is repeated k times so that all folds are used for validation. Here, we use k-fold method with $k = 5$ and 10 folds, as usually done in the literature (Nematzadeh et al. 2015; Rodriguez et al. 2010).

3 Estimation of elastic properties from micromechanics

MB approaches have been proven useful to get accurate predictions of properties of cement systems at various scales (Wyrzykowski et al. 2017), sometimes with error not even exceeding 3 % (Königsberger et al. 2021). In the literature, various propositions of representation of the cement paste microstructure exist based on different number of scales, system morphology, and on the use of different models to describe the volume fraction of the constituents in the system. To fully explore the relevant microstructure representations mostly adopted, here we consider sixteen representations (see details in Section 3.2). Each representation combines different assumptions regarding the number of scales to be considered, the shape of the constituent phases, or the model to describe phase assemblage. These representations are based on previous studies on the upscaling of different physical properties of cement-based materials (Sanahuja et al. 2007; Honório et al. 2016a; Honório et al. 2018a; Königsberger et al. 2021).

The Powers (Powers and Brownyard 1946), KHP (Königsberger et al. 2016), and TJ (Tennis and Jennings 2000) models are considered to evaluate the phases evolution with the DOH in OPC pastes. The former is the earliest and one of the simplest strategies. The latter is one of the most detailed descriptions of phase assemblage in OPC systems before resorting to thermodynamics modeling. The KHP model updates the Powers model by introducing C-S-H densification. In the following, we detail how we obtain the input for micromechanics estimations (i.e., volume fractions of phases as a function of the age or DOH). Then, the formulation of the homogenization schemes is recalled.

3.1 Phase assemblage approximation from hydration models for Ordinary Portland Cement pastes

The Powers hydration model (Powers and Brownyard 1946) considers only three phases, as listed in Table 2. It has been coupled with micromechanics strategies to study early-age property development of cement-based materials in (Sanahuja et al. 2007; Pichler et al. 2013). This model has the advantage of simplicity but does not account for a variety of phases that can be present in OPC systems. The KHP model extends the Powers model by considering C-S-H densification (in agreement with NMR data) and by providing the volume fraction of portlandite.

For comparison, a more elaborate model, the Tennis and Jennings (2000) model is explored. It describes the chemical rearrangement due to the hydration process by stoichiometric relationships based on a more detailed separation of phases, i.e., the evolution of clinker minerals and gypsum fractions as well as the main hydrates separately as a function of the DOH, as listed in Table 2. It even allows us to distinguish LD and HD C-S-H. More details about formulations of the Powers and TJ models are given in Appendix C.

The elastic properties of the constituent phases are given in Table 2. A Poisson's ratio of 0.5 is the typical value for fluids with zero shear rigidity. Adopting $K = 2.18$ GPa for the porosity presupposes that the pores contribute to the mechanical response and are saturated with liquid water, i.e., the *active porosity* assumption. We have also tested the *inactive porosity* assumption, that is $K = 0$ for the porosity, and the results of MB methods exhibited slightly large deviations from the experimental data. Since the references consulted for the database do not provide, in all cases, details about curing conditions, we decided to adopt hereon the *active porosity* assumption.

Hyd. model	Phase	E [GPa]	ν [GPa]	G [GPa]	K [GPa]	Reference
Powers/KHP	clinker	140	0.3	53.8	116.7	Acker (2001)
	hydrates	22.06	0.24	11.8	18.7	Pichler and Hellmich (2011)
	pores	0	0.5	0	2.18	Lide (1997)
TJ	C_3S	135±7	0.3	51.9	112.5	Velez et al. (2001)
	C_2S	130±20	0.3	50	108.3	Velez et al. (2001)
	C_3A	145±10	0.3	55.8	120.8	Velez et al. (2001)
	C_4AF	125±25	0.3	48.1	104.2	Velez et al. (2001)
	CSH_2 **	45.7	0.33	17.2	44.8	Aller et al. (1996)
	HD C-S-H	29.4±2.4	0.24	11.8	18.8	Constantinides and Ulm (2004)
	LD C-S-H	21.7±2.2	0.24	8.8	13.9	Constantinides and Ulm (2004)
	CH	42	0.315	16	37.8	Monteiro and Chang (1995)
	AFt	25±2	0.34±0.02	9.3	26	Speziale et al. (2008)
	AFm*	24.5	0.34	9.1	25.5	Honório et al. (2020c)[a]
	C_4AH_{13}	25	0.34	9.3	26	Speziale et al. (2008)
	hydrogarnet	55.5	0.35	20.6	61.7	Manzano (2009)[a]
	pores	0	0.5	0	2.18	Lide (1997)

Table 2 Elastic constants of phases. *Monosulfoaluminate. **Dihydrate. [a]Molecular simulations.

3.2 Representations of the microstructure

Sixteen representations of the microstructure of the cement paste are considered here, each one combining different assumptions regarding the number of scales to be considered, the shape of

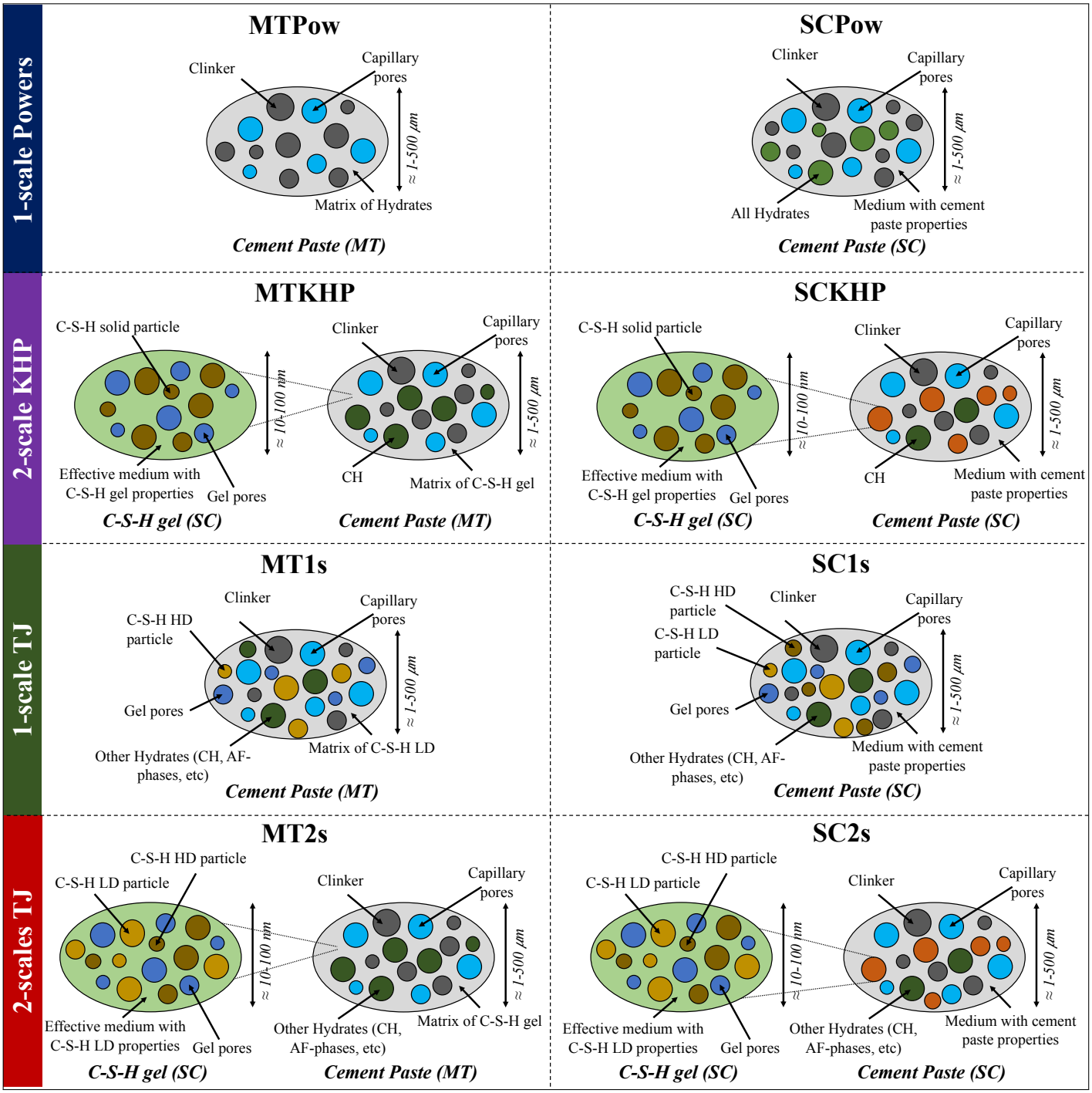


Figure 1 Cement paste microstructure considered for micromechanics upscaling schemes: Input volume fractions are obtained from Powers or TJ models. SC or MT are deployed to upscale cement paste elastic properties. For 2-scale representations, C-S-H gel effective properties are upscaled using SC scheme. All the other phases, including pores, are considered as spherical inclusions. SC scheme is deployed to upscale cement paste elastic properties. For the representations with ellipsoidal inclusions: with Powers model, hydrates are considered as prolate particles $a_r = 10$; with TJ model, C-S-H LD and HD are considered as prolate particles $a_r = 10$, and CH and AFm as oblate particles $a_r = 0.2$.

the constituent phases, or the model to describe phase assemblage.

Figure 1 shows the eight representations of the microstructure of the cement paste tested in the case of spherical inclusions. The other eight representations refer to the adoption of ellipsoidal inclusions to represent some phases. For the cases with ellipsoidal inclusions, similar representations are adopted with the following modifications: (i) C-S-H (when TJ model is used) or hydrates (when Powers model is used) are modeled as elongated inclusions with an aspect ratio of 10; and (ii) AF-phases and CH (when TJ model is used) are modeled as oblate particles with an

aspect ratio of 0.2. All the other phases, including pores, are considered as spherical inclusions.

Using the description of phases by the *Powers model*, the Mori-Tanaka, with hydrates functioning as the matrix, and Self-Consistent schemes are concurrently considered, which leads to the MTPow and SCPow macroscopic behaviors, respectively.

Using the TJ model, in addition to the flexibility offered by the two upscaling schemes, the microstructure can be constructed with different perspectives. All hydrates, anhydrates, and pores can be treated at the same scale, which gives MT1s and SC1s corresponding to MT with LD C-S-H as matrix and self-consistent schemes, respectively. Or, C-S-H gel can be handled at a smaller scale comprising LD and HD C-S-H domains and gel porosity. First, the effective properties of C-S-H gel are obtained using the SC scheme on a heterogeneous material. The effective properties of C-S-H gel are then used in parallel with the properties of other hydrates and clinker inclusions at the cement paste scale as input for the second stage of homogenization, which can be processed using MT with C-S-H gel as the hosting matrix or self-consistent schemes to give MT2s and SC2s effective properties.

Using the description of phases by the KHP model, a two-scale representation is considered with the C-S-H gel scale and a cement paste scale *per se* at the higher level.

3.3 Analytical homogenization of the elastic properties of micro and macro isotropic heterogeneous materials

We deploy the MT and SC homogenization schemes for micro and macro-isotropic heterogeneous materials with ellipsoidal inclusions randomly distributed in a REV. According to these schemes, the effective stiffness tensor $\mathbf{C}_{\text{est}}$, the subscript 'est' designating the MT or SC estimate, of a heterogeneous material is given by e.g., (Zaoui 2002)

$$\mathbf{C}_{\text{est}} = \left(\sum_{r=1}^{N} f_r \mathbf{C}_r : [\mathbf{I} + \mathbf{P}^0 : (\mathbf{C}_r - \mathbf{C}^0)]^{-1}\right) : \left(\sum_{r=1}^{N} f_r [\mathbf{I} + \mathbf{P}^0 : (\mathbf{C}_r - \mathbf{C}^0)]^{-1}\right)^{-1} \tag{1}$$

where f_r is the volume fraction of the phase r, $\mathbf{C}_r$ is the stiffness tensor of phase r; $\mathbf{P}^0 = \mathbf{S}_H^0 : \mathbf{C}^0$ is the Hill tensor obtained from the Eshelby tensor $\mathbf{S}_H^0$ (which depends only on the properties of the reference medium, see (Mura 1987) for the expressions of Eshelby tensors including the case of ellipsoidal inclusions) and the stiffness tensor of the reference medium $\mathbf{C}_0$, which is defined according to the scheme chosen:

- $\mathbf{C}^0 = \mathbf{C}_0$ where $\mathbf{C}_0$ refers to the matrix stiffness tensor (subscript 0 stands for matrix properties).
- $\mathbf{C}^0 = \mathbf{C}^{\text{SC}}$ for the SC scheme, i.e., the reference medium is the effective medium itself.

An important input for estimations using non-spherical particles is the aspect ratio of the particles. We adopt an aspect ratio of $a_r = 10$ (prolate particle) for C-S-H needles and $a_r = 0.2$ (oblate particle) for crystalline hydrates such as CH and AFm.

In the case of spherical isotropic inclusions, Equation (1) simplifies into the forms described below.

- For an $(N+1)$-phase heterogeneous material with a matrix/inclusion morphology constituted of N isotropic spherical inclusions randomly distributed in a matrix (percolating phase), the MT estimates of the effective bulk K^{MT} and shear G^{MT} moduli are, respectively, obtained from (Torquato 2002)

$$\frac{K^{\text{MT}} - K_0}{K^{\text{MT}} + \frac{4}{3}G_0} = \sum_{r=1}^{N} f_r \frac{K_r - K_0}{K_r + \frac{4}{3}G_0} \quad \text{and} \quad \frac{G^{\text{MT}} - G_0}{G^{\text{MT}} + \frac{4}{3}H_0} = \sum_{r=1}^{N} f_r \frac{G_r - G_0}{G_r + \frac{4}{3}H_0} \tag{2}$$

with $H_0 = \frac{\frac{3}{2}K_r + \frac{4}{3}G_r}{K_r + 2G_r} G_r$, the subscript 0 denoting the (isotropic) matrix phase.

- For an N-phase heterogeneous materials with N isotropic equiaxed inclusions randomly distributed in a representative elementary volume following a polycrystalline-like morphology (i.e., in which no phase clearly functions as a matrix), the Self-Consistent effective bulk K^{SC} and shear G^{SC} moduli are given, respectively, by the implicit relations (Torquato 2002)

$$\sum_{r=1}^{N} f_r \frac{K_r - K^{\text{SC}}}{K_r + \frac{4}{3}G^{\text{SC}}} = 0 \quad \text{and} \quad \sum_{r=1}^{N} f_r \frac{G_r - G^{\text{SC}}}{G_r + H_{SC}} = 0. \tag{3}$$

3.4 Bounds for the elastic properties

From the properties of the constituent phases and their volume fraction, micromechanics offers not only the effective properties but also bounds between which the elastic properties of the heterogeneous material should lie within. It is then possible to cross-check the observed experimental values with the bounds given by the theoretical models based on specific modeling assumptions.

In the present paper, two theoretical bounds defined in terms of the effective bulk K^{eff} and shear G^{eff} moduli are considered (Zaoui 2002):

Voigt-Reuss bounds They are associated with series and parallel models:

$$\left(\sum_{r=1}^{N}\frac{f_r}{K_r}\right)^{-1} \leqslant K^{\text{eff}} \leqslant \sum_{r=1}^{N} f_r K_r; \quad \left(\sum_{r=1}^{N}\frac{f_r}{G_r}\right)^{-1} \leqslant G^{\text{eff}} \leqslant \sum_{r=1}^{N} f_r G_r \tag{4}$$

where the leftmost term is the Reuss estimate and the rightmost term is the Voigt estimate.

Hashin-Shtrikman bounds They are defined for heterogeneous materials with an isotropic distribution of phases for an arbitrary phase geometry based on the variational principle in linear elasticity (Hashin and Shtrikman 1963):

$$\sum_{r=1}^{N}\frac{f_r K_r/[K^- + \alpha^-(K_r - K^-)]}{f_r/[K^- + \alpha^-(K_r - K^-)]} \leqslant K^{\text{eff}} \leqslant \sum_{r=1}^{N}\frac{f_r K_r/[K^+ + \alpha^+(K_r - K^+)]}{f_r/[K^+ + \alpha^-(K_r - K^+)]} \tag{5}$$

$$\sum_{r=1}^{N}\frac{f_r G_r/[G^- + \beta^-(G_r - G^-)]}{f_r/[G^- + \beta^-(G_r - G^-)]} \leqslant G^{\text{eff}} \leqslant \sum_{r=1}^{N}\frac{f_r G_r/[G^+ + \beta^+(G_r - G^+)]}{f_r/[G^+ + \beta^-(G_r - G^+)]} \tag{6}$$

where $G^- = \inf G_r$; $K^- = \inf K_r$; $G^+ = \sup G_r$; $K^+ = \sup K_r$ are the extreme values of the bulk and shear moduli considering all r phases; $\beta^\pm = \frac{6(K^\pm+2G^\pm)}{5(3K^\pm+4G^\pm)}$ and $\alpha^\pm = \frac{3K^\pm}{3K^\pm+4G^\pm}$. HS bounds are narrower than Voigt-Reuss bounds.

The Young's modulus bounds can be directly computed from the lower and upper bounds using (Zimmerman 1992)

$$\frac{9K_{\text{L}}G_{\text{L}}}{3K_{\text{L}} + G_{\text{L}}} \leqslant E^{\text{eff}} \leqslant \frac{9K_{\text{U}}G_{\text{U}}}{3K_{\text{U}} + G_{\text{U}}} \tag{7}$$

where the subscript $_{\text{L}}$ refers to the lower (HS or Reuss) bound; and the subscript $_{\text{U}}$, to the upper (HS or Voigt) bound. For Poisson's ratio, Zimmerman (1992) showed that the correct bounds are

$$\frac{3K_{\text{L}} - 2G_{\text{U}}}{6K_{\text{L}} + 2G_{\text{U}}} \leqslant \nu^{\text{eff}} \leqslant \frac{3K_{\text{U}} - 2G_{\text{L}}}{6K_{\text{U}} + 2G_{\text{L}}} \tag{8}$$

where the largest possible value of ν refers to the largest value of K combined with the smallest value of G, and vice versa. The argument is valid for both Voigt-Reuss and HS bounds.

4 Results and Discussion

MB and ML methods are investigated for predictions and analysis of various properties of cement paste. Then, bounds for elastic properties given by MB methods are compared with experimental observations. Predictions of elastic properties given by MB and ML methods are compared for training and test datasets. Finally, the lack of knowledge on the parametric input is evaluated, and the experimental dataset is enriched with MB observations guided by ML evaluations.

4.1 Micromechanics bounds for dataset curation

Knowing w/c and DOH (or age, from which DOH can be estimated), fractions of phases are evaluated from hydration models, and bounds for E, ν, G, and K are derived from Voigt-Reuss and HS theories (as detailed in Section 3.4). Comparing the experimental elastic properties and the bounds, both lower bounds, being null, are satisfied by all experimental observations. However, some experimental observations of K and ν exceed the upper bounds. Proportions of

values exceeding the theoretical bounds are summarized in Table 3. Since the phase intrinsic properties are associated with a variability/uncertainty on the order of $10-20\,\%$ as reported in Table 2, we also provide bounds estimation accounting for an average $15\,\%$ uncertainty (i.e., the bulk and shear moduli are increased of $15\,\%$ before upper bound calculation). The bounds are computed for Powers and TJ models (with the assumption of gel and capillary water having the same behavior, the bounds computed for Powers and KHP are identical).

Upper bounds	Hydration model	E	ν	K	G
Voigt	Powers	0 % (0 %)	0 % (0.27 %)	0 % (0.82 %)	0 % (0 %)
Voigt	TJ	0 % (0 %)	0 % (0 %)	0 % (0.55 %)	0 % (0 %)
HS	Powers	0 % (0 %)	0 % (0 %)	1.4 % (4.1 %)	0 % (0 %)
HS	TJ	0 % (0 %)	0 % (0.82 %)	1.4 % (3.5 %)	0 % (0 %)

Table 3 Percentage of values exceeding the upper bounds, Voigt and HS, for each elastic constant tested. Bounds computed using Powers or TJ hydration models and considering a $15\,\%$ uncertainty on K and G reported as phase properties in Table 2. Bracketed values refer to bounds computed using average values without the $15\,\%$ uncertainty reported in Table 2.

All values of both shear and Young moduli are below the upper bounds. A few values of the bulk modulus, less than $5\,\%$ for the worst case of the experimental observations, exceed the upper bounds when the uncertainty on the phase elastic moduli is not accounted for. As expected, more points exceed HS than Voigt bound since the HS bounds are tighter. For Poisson's ratio, the proportion of experimental observations exceeding the theoretical bonds is still smaller. It must be noted that a precise experimental evaluation of Poisson's ratio can be a challenge, provided the much smaller range of variation when compared to the elastic moduli. Detailed results on the differences between the experimental values comprised in the training dataset and the theoretical bounds are shown in Appendix D.

By comparing the values according to the hydration model, fewer points are outside the bounds when the TJ model is adopted for K or ν, which provides a more precise description of cement phases than the Powers model. These observations might suggest that the adoption of a precise description of cement paste phase assemblage is critical if theoretical bounds are used to curate databases.

To conclude, experimental Young and shear moduli are in concordance with the bounds. For bulk modulus and Poisson's ratio, only a few points are in contradiction with the theoretical bounds. Depending on the trust given to the model in comparison with the experiments, it could be decided to filter out the database of some experimental observations. However, here, for the proof of concept, all the data is conserved to evaluate the ML performances without arbitration on the experimental results.

4.2 Prediction of elastic properties using ML and Micromechanics

ML and MB methods are evaluated to predict the elastic properties of the samples contained in the training and test datasets.

4.2.1 Reproducing the training dataset observations

ML predictions Knowing the w/c, DOH, and percentage fractions of clinker and gypsum in cement, the four elastic properties are estimated by ML approaches. The validation procedure for one of the validation stages is illustrated in Figure 2.

The accuracy of predictions of elastic constants of cement pastes is compared for the various ML methods tested, see Figure 3. The comparison serves to analyze the consistency and compatibility of the method regarding the database on which they are trained. The qualitative analysis suggests that the prediction of Poisson's ratio is less accurate when compared to predictions of the elastic moduli. Visually, NN, ANN, and GP perform better in predictions.

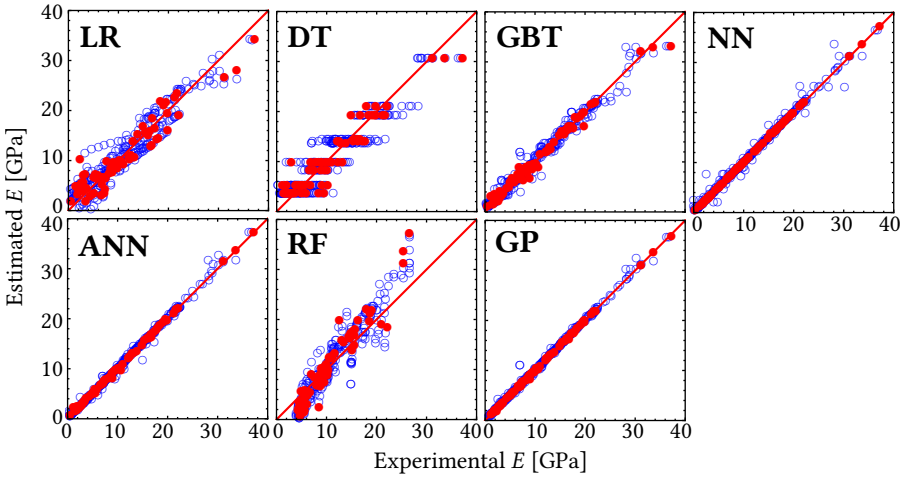


Figure 2 k-fold validation method with $k = 5$. Predicted Young's modulus E plotted against the experimental E at one validation stage out of 5 for the various ML methods tested: 292 values of the 4 training folds are depicted by empty blue dots, full red symbols depict the 73 elements used for validation.

Errors are quantified using the RMSE:

$$\mathrm{RMSE}(x) = \sqrt{\frac{\sum_{i=1}^{n}(x_i^{\mathrm{pred}} - x_i^{\mathrm{exp}})^2}{n}} \tag{9}$$

and MRE:

$$\mathrm{MRE}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_i^{\mathrm{pred}} - x_i^{\mathrm{exp}}|}{x_i^{\mathrm{exp}}} \tag{10}$$

computed as a function of each prediction x_i^{pred} and experimental x_i^{exp} output averaged over the n observations i covering the whole training set obtained from the validation on all k-folds for the elastic constants. Tables 4 and 5 show the RMSE and MRE, respectively, obtained for each ML method prediction. ANN, GP, and NN yield the best accuracy in terms of RMSE and MRE for the elastic constants.

methods	RMSE(E) [GPa]		RMSE(ν) [-]		RMSE(K) [GPa]		RMSE(G) [GPa]	
cross-validation	5-fold	10-fold	5-fold	10-fold	5-fold	10-fold	5-fold	10-fold
LR	2.4	2.4	0.061	0.062	2.2	2.2	1.0	1.0
DT	2.8	2.9	0.042	0.042	2.3	2.3	1.0	1.1
GBT	1.3	1.5	0.037	0.037	1.8	1.8	0.5	0.5
NN	1.1	1.1	0.034	0.032	**1.6**	1.5	0.4	0.4
ANN	**0.6**	**0.6**	0.034	0.033	1.5	**1.4**	**0.3**	**0.2**
RF	3.0	3.0	0.044	0.043	2.6	2.6	1.2	1.2
GP	0.7	0.8	**0.032**	**0.031**	2.0	2.9	**0.3**	**0.2**

Table 4 RMSE of the elastic constants obtained from k-fold cross-validation technique based on 5-fold or 10-fold. Most accurate values marked in bold.

Table 6 shows the (mean) coefficient of determination R^2 of the elastic constants obtained from the k-fold cross-validation technique based on 5-fold or 10-fold. The R^2 is a scale-free quantity, quantifying how a model explains a phenomenon. All models yield high R^2 (closer or higher than 0.9) for the elastic moduli, while the R^2 for ν is overall lower. ANN predictions exhibit higher R^2 in most cases.

MB estimations Knowing the w/c, DOH, and percentage fractions of clinker and gypsum in cement the four elastic characteristics are also predicted by MB methods. Performances are shown in Figure 4. It can be noted that performances vary with the DOH. The homogenization yields predictions of the elastic constants that are, in most cases, better when only the observations

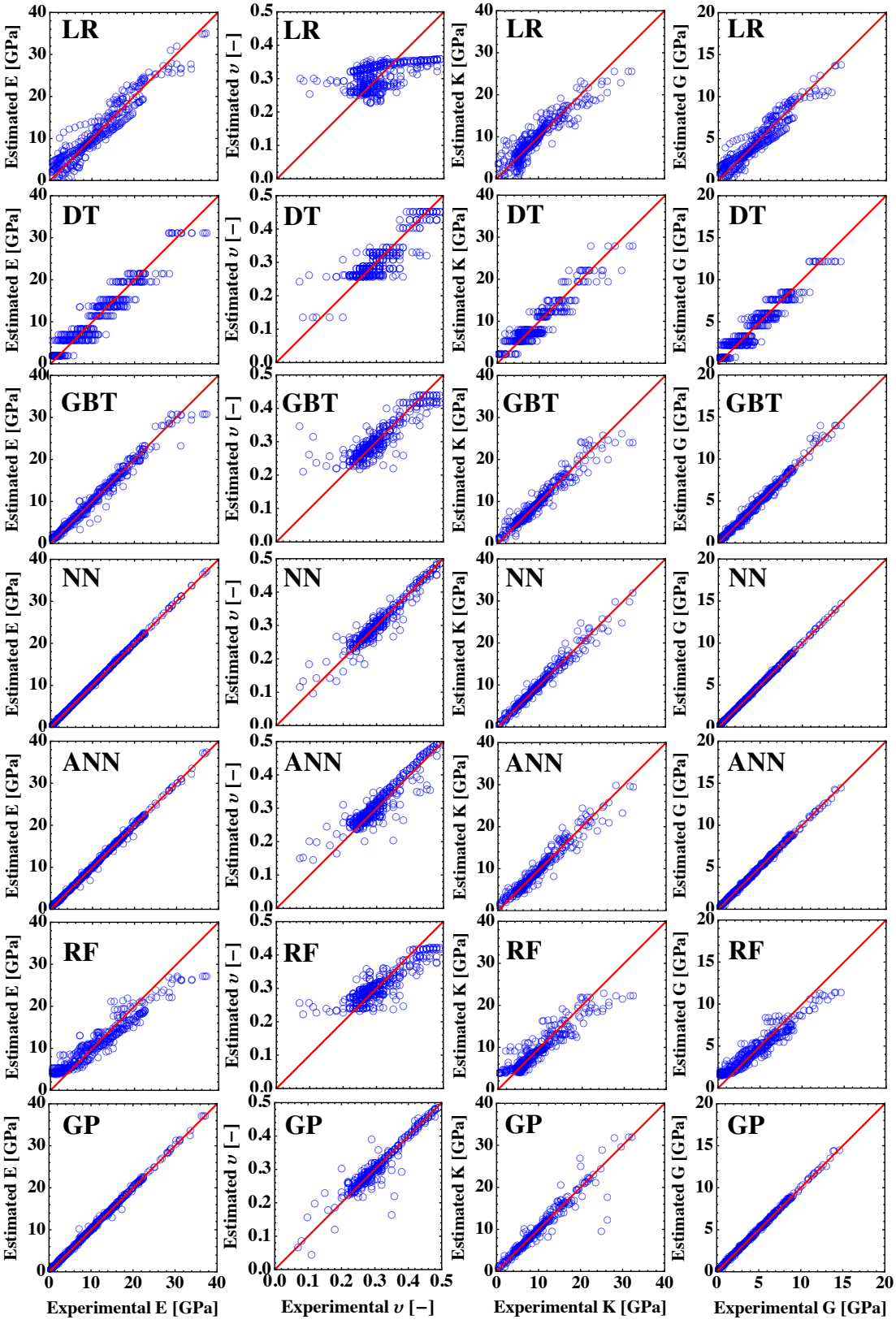


Figure 3 ML performances for elastic prediction of the training set: predicted values based on the various ML methods tested plotted against the experimental elastic constants from the training dataset (Young's modulus E, Poisson's ratio ν, bulk K and shear G moduli).

in the training dataset with DOH $\geqslant 0.7$ (i.e., associated with late ages) are accounted for (this effect can be more pronounced when MT estimates are used). The accuracy of MB estimations is quantified in Tables 7 and 8 using RMSE and MRE, respectively. These parameters were measured for the entire data set and also for the values in the training dataset with DOH $\geqslant 0.7$. The late ages estimates exhibit lower errors overall. Thus a fine description of hydration kinetics, phase

methods	MRE(E) [-]		MRE(ν) [-]		MRE(K) [-]		MRE(G) [-]	
cross-validation	5-fold	10-fold	5-fold	10-fold	5-fold	10-fold	5-fold	10-fold
LR	0.50	0.50	0.18	0.18	0.34	0.34	0.54	0.56
DT	0.53	0.60	0.12	0.11	0.39	0.35	0.55	0.63
GBT	0.19	0.23	0.09	0.09	0.19	0.18	0.19	0.20
NN	0.17	0.16	0.07	0.07	**0.13**	**0.13**	0.18	0.17
ANN	**0.08**	**0.10**	0.08	0.07	0.16	0.15	0.13	0.11
RF	0.69	0.74	0.12	0.11	0.47	0.46	0.80	0.78
GP	0.10	0.09	**0.06**	**0.06**	0.17	0.22	**0.12**	**0.10**

Table 5 MRE of the elastic constants obtained from k-fold cross-validation technique based on 5-fold or 10-fold. Most accurate values marked in bold.

methods	R^2(E) [-]		R^2(ν) [-]		R^2(K) [-]		R^2(G) [-]	
cross-validation	5-fold	10-fold	5-fold	10-fold	5-fold	10-fold	5-fold	10-fold
LR	0.91	0.90	0.25	0.26	0.84	0.86	0.90	0.90
DT	0.84	0.85	0.60	0.68	0.81	0.82	0.85	0.85
GBT	0.97	0.98	0.70	**0.78**	0.89	0.91	0.97	0.98
NN	0.97	0.98	0.74	**0.78**	0.89	0.92	0.97	0.98
ANN	**0.99**	**0.99**	**0.76**	0.76	**0.91**	**0.93**	**0.99**	**0.99**
RF	0.88	0.94	0.65	0.69	0.82	0.83	0.89	0.89
GP	**0.99**	**0.99**	0.71	**0.78**	0.82	0.87	0.98	**0.99**

Table 6 Mean coefficient of determination R^2 of the elastic constants obtained from k-fold cross-validation technique based on 5-fold or 10-fold. Most accurate values marked in bold.

assemblage, and particular effects associated, such as C-S-H gel densification, C-S-H structural and compositional variability, could enhance estimate accuracy. When both error estimates are taken into consideration, the best MB estimates are given by SCPow and SC1s schemes for both cases when only spherical or ellipsoidal inclusions are considered. These four cases are used for comparison with ML methods. The KHP model yields results closer to the ones obtained with Powers model.

Table 9 gathers the coefficients of determination (R^2) of elastic constants. The estimates with ellipsoidal inclusions using the Powers model, MT_{Ellip} Powers and SC_{Ellip} Powers, show a higher R^2 for the elastic moduli. SC Powers (with spherical inclusions) also exhibits one of the highest R^2. The R^2 is overall low, showing the difficulty of the model to properly capture this elastic constant. Since two elastic constants are sufficient to fully determine an isotropic behavior, this last observation suggests that dealing with the elastic moduli (E, K, or G) is a strategy less prone to errors.

As expected, when MB and ML methods predictions are confronted with the training dataset, ML methods display, in general, better accuracy than MB methods, with the less accurate ML methods (RF, DT) yielding predictions with similar accuracy to the best MB estimations. The best RMSE and MRE for Young moduli predictions was with ANN, and the accuracy was roughly 4-fold the accuracy of the best micromechanics estimation. Note however that the comparison according to the training dataset, of course, favors ML methods since these are trained specifically for them, whereas MB methods do not have any *a priori* information on this correlation except the underlying theoretical model assumption.

4.2.2 Reproducing the test dataset observations

Predictions of elastic constants *for a test dataset* by ML methods and homogenization methods are also evaluated. As detailed in Appendix A.1, the test dataset is composed of 58 observations including both static and dynamics measurements of elastic properties from various authors (Constantinides and Ulm 2004; Šavija et al. 2020; Chamrova 2010; Maruyama and Igarashi 2014; Tamtsia et al. 2004; Haecker et al. 2005). The performance of MB and ML methods are analyzed in detail for selected cases in the sequel.

Comparisons with data by Constantinides and Ulm (2004) are given in Figure 5. For that case,

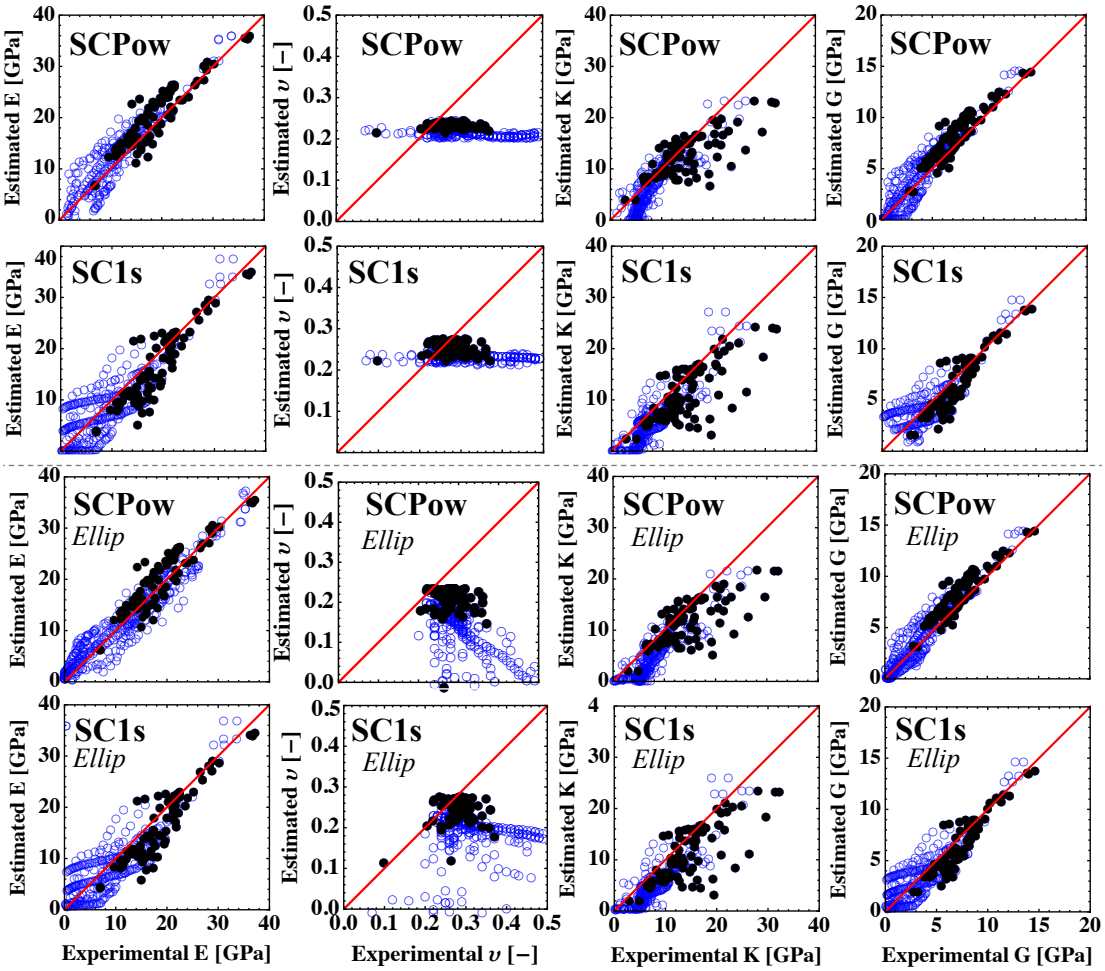
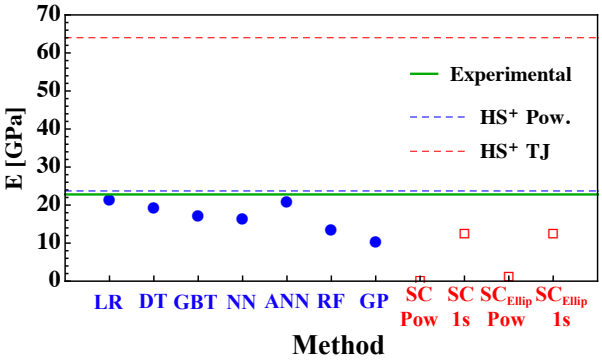


Figure 4 MB performances for elastic prediction of the training set (only the representations with the best performance are shown): experimental elastic constants from the training dataset (Young's modulus E, Poisson's ratio ν, bulk K and shear G moduli) plotted against the estimated values using various homogenization methods. Influence of the DOH on the performances: the empty blue circles correspond to DOH $\leqslant 0.7$ and the solid black circles, to DOH $\geqslant 0.7$, i.e., predictions at late ages.

ML methods (in particular, LR, ANN, and DT) perform better in predicting experimental data than MB techniques. ML and MB yield comparable results when used to predict the Haecker et al. (2005) experimental data, see Figure 6; the exceptions are the cases of homogenization methods using the Powers model. A similar result is obtained by Tamtsia et al. (2004) as visualized in Figure 7. In this case, homogenization with TJ hydration model performs quite well. On this test dataset, we can conclude that ML and MB have similar accuracy, none of them performs significantly better than the other.

Figure 5 Prediction of Young's modulus E to reproduce the experimental observation by Constantinides and Ulm (2004) (solid green line): results from ML (full blue dots) and MB (red empty dots) methods. HS upper bounds using Powers and TJ model (dashed lines) are shown for reference.



The total RMSE and MRE associated with the test dataset for ML methods are shown in Figure 9 (values referring to the original training dataset l_{orig}) and will be discussed in the next

Methods	RMSE(E) [GPa]	RMSE(ν) [-]	RMSE(K) [GPa]	RMSE(G) [GPa]
MT Powers (MTPow)	9.7 (4.0)	0.10 (0.06)	4.5 (**4.1**)	4.1 (1.8)
SC Powers (SCPow)	2.8 (2.8)	0.11 (0.06)	3.4 (4.6)	1.2 (1.2)
MT KHP (MTKHP)	10.2 (7.2)	0.16 (0.06)	5.4 (5.8)	4.1 (2.9)
SC KHP (SCKHP)	8.1 (7.1)	0.10 (0.05)	6.4 (6.5)	3.2 (2.8)
MT 1-scale (MT1s)	6.9 (**2.4**)	0.09 (**0.04**)	3.6 (4.5)	2.8 (**0.8**)
SC 1-scale (SC1s)	3.8 (3.7)	0.10 (0.05)	3.8 (5.4)	1.5 (1.3)
MT 2-scales (MT2s)	6.9 (2.5)	0.09 (**0.04**)	3.6 (4.5)	2.8 (**0.8**)
SC 2-scales (SC2s)	7.6 (9.8)	**0.08** (0.05)	6.7 (8.7)	2.9 (3.8)
MT_{Ellip} Powers (MTPow Ellip)	9.3 (3.9)	0.13 (0.06)	3.7 (4.4)	4.1 (1.8)
SC_{Ellip} Powers (SCPow Ellip)	**2.5** (2.7)	0.30 (0.09)	4.2 (5.2)	**1.1** (1.2)
MT_{Ellip} KHP (MTKHP Ellip)	12.7 (5.18)	**0.08** (**0.04**)	9.0 (6.4)	5.1 (2.2)
SC_{Ellip} KHP (SCKHP Ellip)	3.2 (3.4)	0.30 (**0.04**)	5.0 (5.7)	1.6 (1.5)
MT_{Ellip} 1-scale (MT1s Ellip)	6.6 (2.5)	0.11 (**0.04**)	**3.2** (4.5)	2.8 (**0.8**)
SC_{Ellip} 1-scale (SC1s Ellip)	3.7 (3.6)	0.16 (0.06)	4.1 (5.5)	1.4 (1.3)
MT_{Ellip} 2-scales (MT2s Ellip)	6.6 (2.5)	0.11 (0.05)	**3.2** (4.7)	2.8 (**0.8**)
SC_{Ellip} 2-scales (SC2s Ellip)	7.3 (9.1)	0.72 (0.44)	6.7 (8.2)	2.8 (3.4)

Table 7 RMSE of elastic constants as a measure of the accuracy of the homogenization estimations. Bracketed values correspond to estimations for elements in the dataset with DOH $\geqslant$ 0.7 only. Most accurate values marked in bold.

Methods	MRE(E) [-]	MRE(ν) [-]	MRE(K) [-]	MRE(G) [-]
MT Powers (MTPow)	0.46 (0.23)	0.24 (0.18)	1.1 (0.22)	3.5 (0.27)
SC Powers (SCPow)	0.38 (0.14)	0.27 (0.19)	**0.19** (0.38)	0.41 (0.15)
MT KHP (MTKHP)	3.6 (0.38)	0.52 (0.20)	0.94 (0.28)	2.5 (0.39)
SC KHP (SCKHP)	1.8 (0.38)	0.29 (0.14)	0.73 (0.33)	0.39 (0.77)
MT 1-scale (MT1s)	2.3 (0.11)	0.21 (**0.12**)	0.87 (0.21)	2.6 (0.10)
SC 1-scale (SC1s)	0.68 (0.18)	0.22 (0.13)	0.40 (0.27)	0.76 (0.17)
MT 2-scales (MT2s)	2.2 (**0.10**)	0.20 (**0.12**)	0.85 (**0.20**)	2.6 (**0.09**)
SC 2-scales (SC2s)	0.75 (0.53)	0.20 (0.15)	0.72 (0.56)	0.77 (0.53)
MT_{Ellip} Powers (MTPow Ellip)	0.45 (0.22)	0.32 (0.20)	0.87 (0.22)	3.5 (0.27)
SC_{Ellip} Powers (SCPow Ellip)	**0.37** (0.13)	0.75 (0.27)	0.47 (0.22)	**0.40** (0.16)
MT_{Ellip} KHP (MTKHP Ellip)	0.66 (0.19)	**0.18** (**0.12**)	1.6 (0.30)	4.1 (0.24)
SC_{Ellip} KHP (SCKHP Ellip)	2.0 (0.13)	0.39 (**0.12**)	0.52 (0.26)	0.41 (0.18)
MT_{Ellip} 1-scale (MT1s Ellip)	2.2 (0.11)	0.26 (**0.12**)	0.68 (**0.20**)	2.6 (0.10)
SC_{Ellip} 1-scale (SC1s Ellip)	0.61 (0.17)	0.37 (0.14)	0.42 (0.28)	0.71 (0.16)
MT_{Ellip} 2-scales (MT2s Ellip)	2.2 (0.11)	0.26 (0.13)	0.67 (0.21)	2.5 (**0.09**)
SC_{Ellip} 2-scales (SC2s Ellip)	0.70 (0.49)	0.73 (1.55)	0.73 (0.53)	0.68 (0.48)

Table 8 MRE of elastic constants as a measure of the accuracy of the homogenization estimations. Bracketed values correspond to estimations for elements in the dataset with DOH $\geqslant$ 0.7 only. The most accurate values are marked in bold.

section in comparison with prediction using extended training datasets.

4.3 Missing data and extended database

ML and MB approaches can be used competitively, but the ML approach can also be employed to guide experiments and obtain optimized information over the whole parametric space and MB can be exploited to generate supplementary observations. Thus, direct experimental observations can be evaluated by ML and combined with synthetic observations obtained from MB schemes. Various methods are proposed in the literature to optimize the plan of experiments (Fuhg et al. 2020). Here, we adopt a distance-based approach to identify the zones in which a fewer number

Methods	$R^2(E)$ [-]	$R^2(\nu)$ [-]	$R^2(K)$ [-]	$R^2(G)$ [-]
MT Powers (MTPow)	0.70 (**0.87**)	0.10 (0.00)	0.70 (**0.58**)	0.68 (0.88)
SC Powers (SCPow)	0.90 (0.84)	0.12 (0.00)	0.77 (0.54)	0.90 (0.86)
MT KHP (MTKHP)	0.19 (0.80)	0.00 (0.00)	0.29 (0.51)	0.18 (0.83)
SC KHP (SCKHP)	0.72 (0.82)	0.06 (0.00)	0.72 (0.53)	0.71 (0.84)
MT 1-scale (MT1s)	0.63 (0.85)	0.07 (0.01)	0.67 (0.55)	0.61 (0.87)
SC 1-scale (SC1s)	0.78 (0.83)	0.05 (0.00)	0.74 (0.53)	0.77 (0.86)
MT 2-scales (MT2s)	0.60 (0.85)	0.07 (0.00)	0.65 (0.56)	0.58 (0.87)
SC 2-scales (SC2s)	0.67 (0.84)	0.05 (0.04)	0.63 (0.54)	0.66 (0.86)
MT_{Ellip} Powers (MTPow Ellip)	0.73 (0.86)	**0.19** (0.04)	**0.78** (0.54)	0.68 (**0.89**)
SC_{Ellip} Powers (SCPow Ellip)	**0.91** (0.84)	0.10 (**0.07**)	**0.78** (0.51)	**0.92** (0.86)
MT_{Ellip} KHP (MTKHP Ellip)	0.02 (0.72)	0.05 (0.01)	0.06 (0.46)	0.02 (0.74)
SC_{Ellip} KHP (SCKHP Ellip)	**0.91** (0.85)	0.13 (0.05)	0.77 (0.53)	0.89 (0.87)
MT_{Ellip} 1-scale (MT1s Ellip)	0.64 (0.86)	0.10 (0.00)	0.71 (**0.55**)	0.60 (0.87)
SC_{Ellip} 1-scale (SC1s Ellip)	0.80 (0.84)	0.12 (0.03)	0.75 (0.53)	0.75 (0.53)
MT_{Ellip} 2-scales (MT2s Ellip)	0.63 (0.85)	0.10 (0.00)	0.71 (0.56)	0.59 (0.87)
SC_{Ellip} 2-scales (SC2s Ellip)	0.69 (0.85)	0.03 (0.01)	0.64 (0.55)	0.68 (0.87)

Table 9 Coefficient of determination R^2 of elastic constants as a measure of the accuracy of the homogenization estimations. Bracketed values correspond to estimations for elements in the dataset with DOH $\geqslant$ 0.7 only. Most accurate values marked in bold.

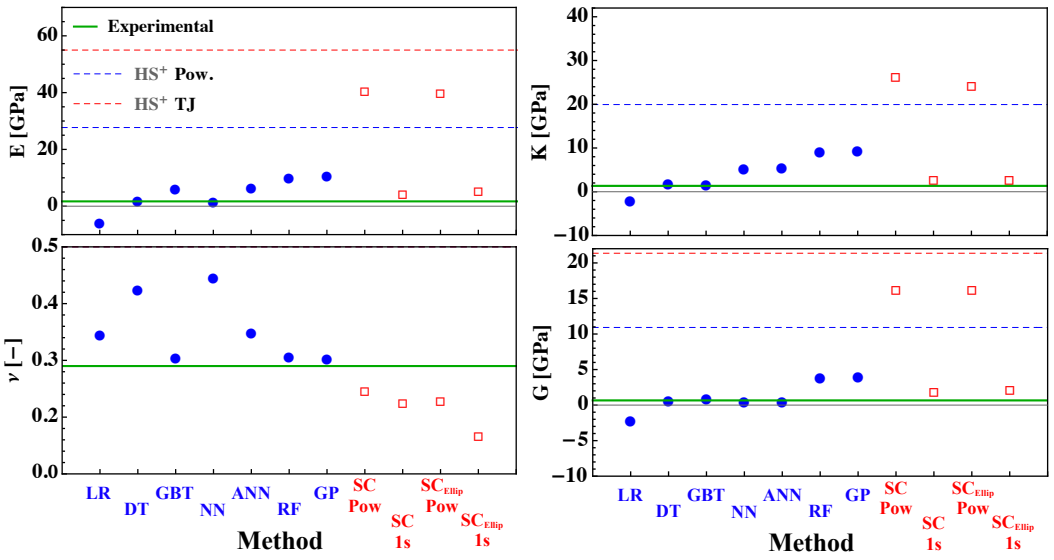


Figure 6 Prediction of the elastic constants (Young's modulus E, Poisson's ratio ν, shear G and bulk K moduli) to reproduce the experimental observations obtained by Haecker et al. (2005) (solid green line): results from ML (full blue dots) and MB (empty red squares) methods. HS upper bounds using Powers and TJ model (dashed lines) are shown for reference.

of experiments have been made. Then, we deploy a k-fold cross validation strategy to exploit domains in which metamodel interpolations are sufficiently accurate and the domain in which information is lacking.

4.3.1 Distance-based approach to identify the domains with missing data

To identify the domains with missing data, we adopt a simple strategy based on the distance of data in a given dimension of input space. In each dimension of the input space, we order the components of the observations in ascending order as shown in Figure 8(top). The normalized difference

$$\Delta O_N = \frac{1}{\sum_i x_i}(x_{i'+1} - x_{i'}) \tag{11}$$

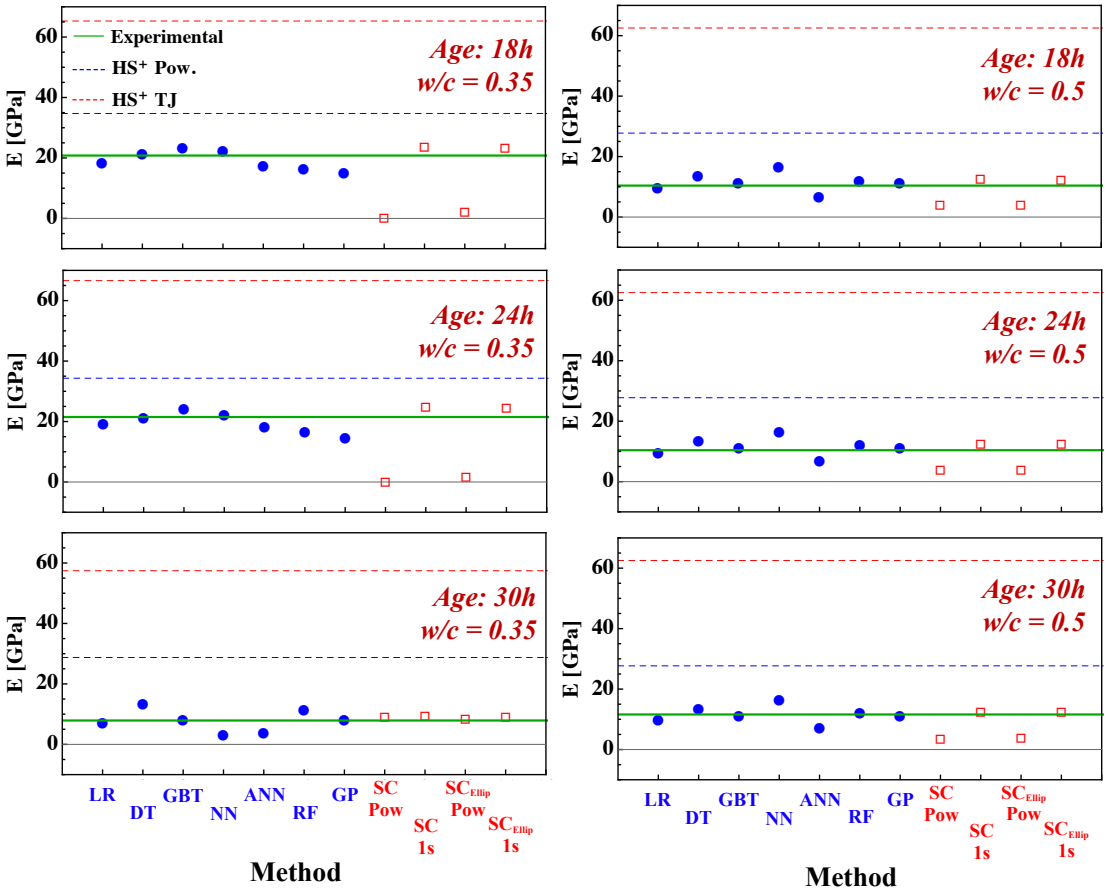


Figure 7 Prediction of Young's modulus E for $w/c = 0.35$ and $w/c = 0.5$ at different ages to reproduce the experimental observations by Tamtsia et al. (2004) (green solid line): results from ML (full blue dots) and MB (empty red squares) methods. HS upper bounds using Powers and TJ model (dashed lines) are shown for reference.

between two components $x_{i'+1}$ and $x_{i'}$ is related to the extent of the domain associated with missing data, where i' denotes the ordered position of an observation. Figure 8 (bottom) shows $\Delta O_N(i')$ for all input components considered here. A large $\Delta O_N(i')$ indicates that two ordered observations $x_{i'+1}$ and $x_{i'}$ are relatively far from each other and that the interval $]x_{i'}, x_{i'+1}[$ is a zone in which data is missing. To identify the most relevant zones in which data is missing according to this approach, we adopt the following criterion: an new observation $x_i^* = \frac{1}{2}(x_{i'} + x_{i'+1})$ is to be generated whenever $\Delta O_N \geqslant c_0$, where c_0 is an arbitrary cut-off. We adopt $c_0 = 0.0001$ for the w/c and DOH, and $c_0 = 0.0005$ for the clinker minerals and gypsum mass fractions, shown as gray dashed lines in Figure 8(bottom). With this approach, the selected x_i^* per input are

- j = DOH [-]: 0.70, 0.95
- $j = w/c$ [-]: 0.28, 0.37, 0.72
- $j = m_{C_3S}$ [%]: 33.05, 52.405, 85.25
- $j = m_{C_2S}$ [%]: 5.40, 11.24, 16.70, 20.70, 31.75, 49.95
- $j = m_{C_3A}$ [%]: 1.10, 9.95
- $j = m_{C_4AF}$ [%]: 4.40, 11.25
- $j = m_{\text{gypsum}}$ [%]: 1.05, 2.00, 2.90, 3.65, 4.69, 6.09.

To generate the new data in these zones, we defined three new datasets:

- The minimum dataset $l_{\min}$ covering all x_i^* for all input vector O_m identified by the strategy above. The minimum number of observations to be generated covering all these values is 6.
- A dataset l_{exist}^{1P} with one new observation by O_{ij}^* per input identified, with each one of the 24 O_{ij}^* values associated with *an already existing* (randomly sampled) set of input.
- A dataset l_{self}^{1P} also with 24 observations with each observation being a random combination of the O_{ij}^* identified by the strategy above.

To generate the new data on elastic constants, we adopt the MB method $\text{SC1s}_{\text{Ellip}}$, which yields

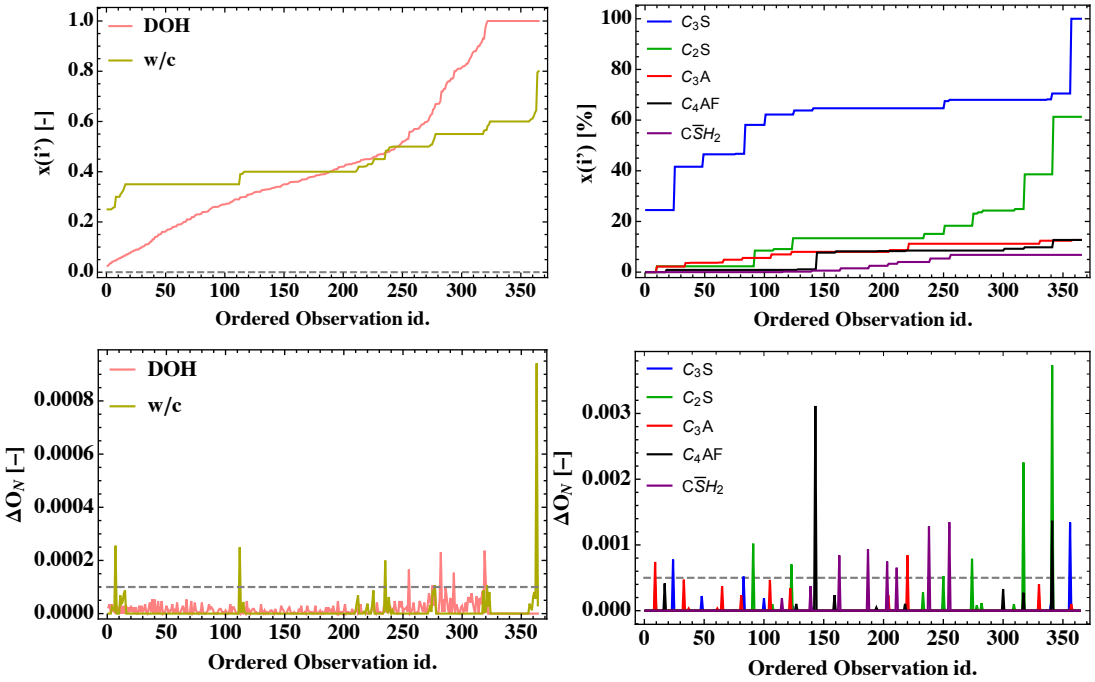


Figure 8 Strategy to identify the zones with missing data. [Top] Ordered input vector $x(i')$ for each input component in $\{\text{age}, w/c, m_{C_3S}, m_{C_2S}, m_{C_3A}, m_{C_4AF}, m_{\text{gypsum}}\}$ as a function of the ordered observation index i'. [Bottom] Normalized difference defined in Equation (11) as a function of index i' indicating ascending order per component. Gray dashed lines depict the limit criterion adopted to identify the most relevant domains with missing data.

one of the best performances of MB methods, as discussed in Section 4.2.1.

In Figure 9, we compare the performance of ML methods trained on the original and extended datasets on the estimation of Young's modulus of the test dataset (RMSE and MRE at the left), and training dataset via a cross-validation approach ($k = 5$ folds) on the training dataset (RMSE and MRE at the right).

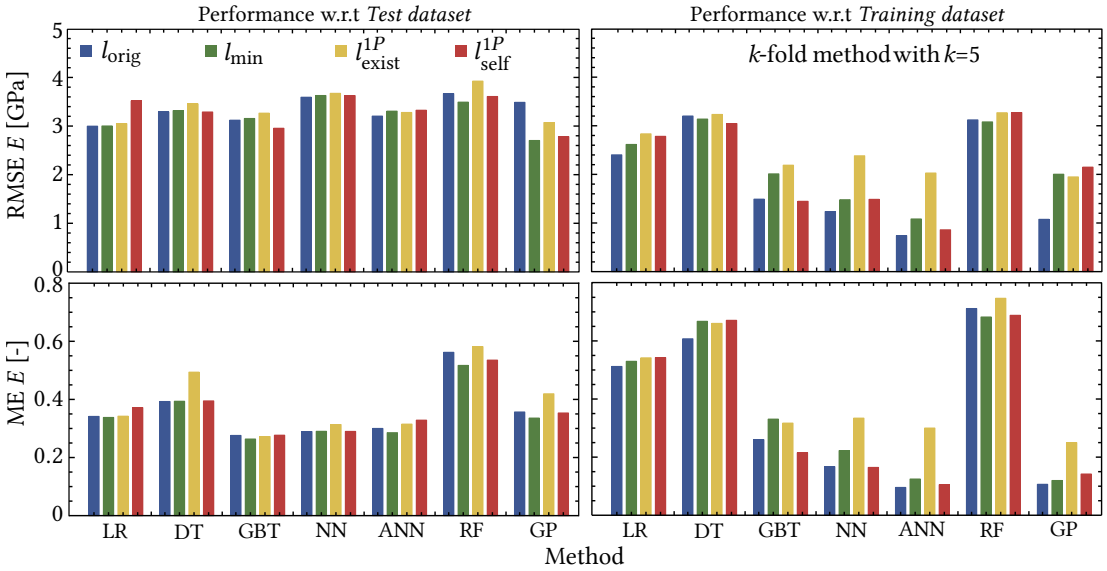


Figure 9 RMSE and MRE computed for the test dataset (left) and using a cross-validation approach ($k = 5$ folds) on the training dataset (right) for various ML methods using only the original training dataset l_{orig}, or extended training datasets l_{min}, l^{1P}_{exist}, or l^{1P}_{self}. New data is generated with $\text{SC1S}_{\text{Ellip}}$.

Regarding the performance when the test dataset is considered, the use of the extended training dataset l^{1P}_{exist} improves the accuracy of ANN predictions, and the use of l^{2P}_{exist} improves the accuracy of GP predictions. Regarding the performance when the training dataset is considered,

the use of extended training datasets generally increases the error as computed by the k-fold methods except for DT, GBT, and RF. When the extended training datasets lead to largest inaccuracies, it must be noted that the increase in RMSE and MRE is not too large (except in some of the l^{1P}_{exist} cases in which the error is duplicated when the training dataset is considered). This observation suggests that using MB methods to generate missing data does not significantly impair the precision of predictions.

These results show that MB methods can be used to generate new data to complete databases for establishing composition-property correlations in cement-based materials leading in some cases (ANN and GP, the best performances in prediction) to an improvement in the prediction accuracy regarding the test dataset.

5 Conclusions

In this article, Machine Learning (ML) and Micromechanics-based (MB) methods were deployed to establish correlations between the composition and the elastic property of OPC pastes. In the exploration of the methods, we identified opportunities for using them as promising allies. ML arises as a proficient tool to exploit a variety of results from different authors with a set of input characteristics, and identify a significant lack of knowledge. Micromechanics is an opportunity to judge experimental results, provides bounds to check ML predictions, and furnishes supplementary/complementary data for ML to be trained on. The main conclusions of this study are as follows:

Methods to link composition and elastic properties The accuracy of ML and MB predictions are comparable for predicting elastic properties of the *test dataset*. When MB and ML are confronted with the *training dataset*, ML gives better accuracy than MB methods, with the less accurate ML methods (RF, DT) yielding predictions with a similar accuracy of the best MB estimations (the comparison according to the training dataset, of course, favors ML methods since these are trained specifically for them, whereas MB methods have not any *a priori* information on this correlation except the underlying theoretical model assumption.). It must be noted that the test dataset used in this work encompasses both dynamic and static measurements, while the ML methods were trained only in dynamic experimental data. Both ML and analytical micromechanics computations performed here are not computer-intensive, especially when compared to often fastidious numerical homogenization approaches. This aspect is a clear advantage of ML and analytical MB methods, notably when a larger exploration of the compositional design space is desired.

Data-driven estimates and importance of reliable databases Even with a relatively small training dataset, ML methods have proven to be reliable and robust in the prediction of elastic properties of cement pastes from their composition for the test dataset. Thus, the effort to build and enlarge the databases on cement composition and properties, for instance, including static measurements in the training data set or even using the same strategy for properties other than elastic properties, may benefit cement and concrete research by providing a reliable tool to tailor the composition of the material for a target property or performance specification.

Providing missing data Analytical micromechanics methods appear proficient in completing the database for input values that have not been explored by experimental campaigns. Indeed, the accuracy of ML and MB being comparable corroborates that MB methods can be used to provide missing data in the databases of cement-based materials despite their well-known variability, and the significant lack of knowledge being robustly identified from the ML approaches. This observation adds to the accumulating evidence showing that MB approaches are a powerful tool to estimate the property from the composition based on a few fundamental component data set and assumptions on cement hydration and microstructure models. Besides providing virtual estimations for concealing missing experimental data, they can also serve to cross-check uncertain or suspicious observations.

The strategy outlined in this study combines MB and ML methods to explore the space of formulation design, it also links the formulation to the effective properties of the materials. It can be extended to other properties in cement and concrete science. This approach arises an

interesting and not costly tool to feed mechanical simulations taking into account the local variability of material properties based on physical and micro-scale properties, even for large simulations. It has been exploited for domains with missing data; it could also be enriched with analysis of subdomains with non-trustable data.

A reviewer of this work indicated another exciting way in which MB and ML can be allies: starting from a mechanistic template—as the one provided by MB methods—and using ML to provide missing parameters or even the constitutive relation. These kinds of hybrid mechanistic–ML models have been proposed in other fields (e.g., Fuhg et al. 2021; Fuhg et al. 2023), but in the case of cement and concrete research, they are yet to be fully worked out. In this direction, one way of using ML to improve MB estimates would be to use data to determine more appropriate localization relations (i.e., the relations determining the contribution of each phase to the effective behavior) instead of relying on the ones provided in the classic homogenization schemes.

A Database collection

Experimental observations, as detailed below, are collected from various papers. The database collected comprises 365 observations with "full" information on which ML methods are trained, and 11 observations with "partial" information used for testing both ML and MB approaches.

A.1 Training dataset

Only data from dynamical measurements of elastic constants are considered in the training dataset.

Helmuth and Turk (1966) The authors do not provide the w/c ratio chosen for experimental formulation, but instead they measure the ratio w_t/c_i between the total water content w_t and the ignited weight c_i at late ages. As proposed by other authors (Achour et al. 2020; Sanahuja et al. 2007), it is possible to estimate the w/c ratio using

$$w_t/c_i = \begin{cases} w/c \dfrac{\kappa_w}{\kappa_h - 1} & \text{if } w/c \leqslant \dfrac{\kappa_h - 1}{\rho_{\text{clinker}}} \\ w/c + \dfrac{1 + \kappa_w - \kappa_h}{\rho_{\text{clinker}}} & \text{otherwise} \end{cases} \tag{A.1}$$

where $\rho_{\text{clinker}} = 3.13$ is the density of clinker. The quantities $\kappa_h = 2.13$ and $\kappa_w = 1.31$ are the volumes of depleted water and formed hydrates, respectively, per unit of clinker consumed by hydration processes.

Boumiz (1996) Only data on cement pastes were published. For some observations, elastic properties were provided without knowledge about age or DOH. In these cases, missing key information has been approximated by local linear least-squares regression. Since the experimental data is pretty smooth, approximating the local behavior (in the range of a few observation points) by a linear fitting is a reasonable choice.

Haecker et al. (2005) The data on cements "H" and "D" were collected as presented by the authors.

Sun et al. (2007) As for (Boumiz 1996), local linear least-squared regression was performed to obtain age and DOH for some experimental observations. Modified Bogue formula was used to compute clinker mineral fractions. Besides elastic constants of cement pastes, the same study reports also results at mortar and concrete scales that can be used in future work.

Wang and Subramaniam (2011) As for (Boumiz 1996), local linear least-squared regression was performed to obtain age and DOH for some experimental observations. The modified Bogue formula was used to compute clinker mineral fractions.

Chamrova (2010) The data was collected as presented by the authors.

Maruyama and Igarashi (2014) As for (Boumiz 1996), the age and DOH were estimated for some observations by local linear least-squares regression.

A.2 Test dataset

Experimental observations from (Tamtsia et al. 2004; Constantinides and Ulm 2004; Lura et al. 2003; Šavija et al. 2020) did not include information regarding either age, DOH, or the pair of

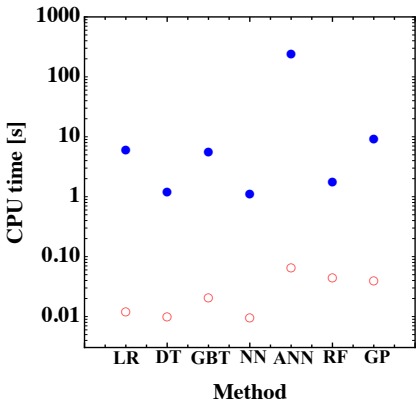
elastic constants necessary to characterize isotropic elastic behavior. Therefore, these samples could not be included in the training dataset, they form the core of the test dataset. The data on cement "L" from (Haecker et al. 2005) is also incorporated in the test dataset.

For the test dataset, both static and dynamical measurements are considered indistinctly.

B Numerical cost of the predictor functions

We compare the CPU time associated with the creation of the predictor functions based on the training dataset, and the realization of one prediction (using the already created predictor functions) in Figure B.1. ANN takes much longer to build the predictor functions than the other methods. Once the predictor function is created, the prediction realization is obtained in a fraction of a second for all the methods.

Figure B.1 CPU time associated with the creation of the predictor functions based on the training dataset (full blue dots), and with one prediction using the already created predictor functions (empty red dots) for the various ML methods.



C Hydration assemblage model

Micromechanics approaches are based on the knowledge of phases intrinsic properties and volume fractions (which evolve with time and DOH). Models used to estimate the phases fraction during the hydration process are briefly exposed in this appendix.

C.1 Powers model

The ratio w/c determines the initial porosity in cement systems and can be used to estimate the porosity as a function of the DOH. In the absence of filler blended in the binder, the Powers model (Powers 1960; Pichler and Hellmich 2011) estimates the volume fractions of the clinker, water (capillary porosity), hydrates and chemical shrinkage (or "air"), respectively as:

$$f_{\text{clinker}} = \frac{1-\text{DOH}}{1+w/c\frac{\rho_{\text{clinker}}}{\rho_{\text{water}}}} = \frac{20(1-\text{DOH})}{20+63w/c} \geqslant 0, \tag{C.1}$$

$$f_{\text{water}} = \frac{\rho_{\text{clinker}}(w/c-0.42\text{DOH})}{\rho_{\text{water}}+w/c\rho_{\text{clinker}}} = \frac{63(w/c-0.42\text{DOH})}{20+63w/c} \geqslant 0 \tag{C.2}$$

$$f_{\text{hydrates}} = \frac{1.42\rho_{\text{clinker}}\text{DOH}}{\rho_{\text{hydrates}}+w/c\rho_{\text{clinker}}/\rho_{\text{water}}} = \frac{43.15\text{DOH}}{20+63w/c} \tag{C.3}$$

$$f_{\text{air}} = 1-f_{\text{clinker}}-f_{\text{water}}-f_{\text{hydrates}} = \frac{3.31\text{DOH}}{20+63w/c} \tag{C.4}$$

with the mass volume of clinker $\rho_{\text{clinker}} = 3.15\,\text{g/cm}^3$, water $\rho_{\text{water}} = 1\,\text{g/cm}^3$ and hydrates $\rho_{\text{hydrates}} = 2.073\,\text{g/cm}^3$ (Pichler and Hellmich 2011).

Following Hansen (1986), the maximum DOH α_{max} is a function of the w/c ratio and depends on curing conditions. For curing without an external water supply, the maximum DOH denoted $\alpha_{\text{max}}^{\text{NW}}$ is reached when water or cement is depleted, thus

$$\alpha_{\text{max}}^{\text{NW}} = \begin{cases} \dfrac{w/c}{\kappa_w/\rho_{\text{clinker}}} & \text{if } w/c \leqslant \kappa_w/\rho_{\text{clinker}}, \\ 1 & \text{otherwise.} \end{cases} \tag{C.5}$$

For curing condition with supplementary external water supply, the maximum DOH denoted $\alpha^{\mathrm{W}}_{\max}$ is reached when cement is depleted or when the entire space available for hydrate growth, i.e., full capillary porosity, is depleted:

$$\alpha^{\mathrm{W}}_{\max} = \begin{cases} \dfrac{w/c\,\rho_{\mathrm{clinker}}}{\kappa_h - 1} & \text{if } w/c \leqslant (\kappa_h - 1)/\rho_{\mathrm{clinker}}, \\ 1 & \text{otherwise.} \end{cases} \tag{C.6}$$

C.2 Königsberger-Hellmich-Pichler model

We adopt the model of the evolution of phase volume fractions proposed by Königsberger et al. (2016), which extends the Powers model. C-S-H densification is accounted for, in agreement with the NMR evidence (Muller et al. 2012). C-S-H Densification is described using three hydration regimes: *regime I* dense C-S-H particles precipitate on cement particle boundaries; *regime II* C-S-H precipitates in a loosely packed configuration where gel porosity appears; and, *regime III* C-S-H precipitation completely fills the capillary porosity. The volume fractions of cement f_{cem} (approximated as clinker), other hydrates f_{CH} (assuming that portlandite CH is the main crystalline hydrate constituting the other hydration products), solid C-S-H f_{sCSH} (considered as a microporous phase with interlayer pores), capillary pore f_{CP}, void volume f_{void} (or chemical shrinkage), and gel pores f_{GP} are, respectively, given by

$$f_{\mathrm{cem}} = \frac{1-\xi}{1+3.185w/c} \geqslant 0 \tag{C.7}$$

$$f_{\mathrm{CH}} = \frac{0.484\xi}{1+3.185w/c} \tag{C.8}$$

$$f_{\mathrm{sCSH}} = \frac{1.105\xi}{1+3.185w/c} \tag{C.9}$$

$$f_{\mathrm{CP}} = \frac{3.185w/c - 0.755\xi}{1+3.185w/c} - f_{\mathrm{GP}} \geqslant 0 \tag{C.10}$$

$$f_{\mathrm{void}} = \frac{0.167\xi}{1+3.185w/c} \tag{C.11}$$

$$f_{\mathrm{GP}} = \begin{cases} 0 & 0 \leqslant \xi \leqslant \xi_{I-II} \\ \dfrac{4.824w/c\xi - 0.799(w/c)^2 - 0.793\xi^2}{(1+3.185w/c)(0.864w/c+1.278\xi)} & \xi_{I-II} < \xi < \xi_{II-III} \\ \dfrac{3.185w/c - 0.755\xi}{1+3.185w/c} & \xi_{II-III} \leqslant \xi \leqslant 1 \end{cases} \tag{C.12}$$

where $\xi_{I-II} = 0.170w/c$ and $\xi_{II-III} = 2.022w/c$ are the transition hydration degrees between hydration regimes.

C.3 Tennis and Jennings model

The phase assemblage in the Tennis and Jennings (2000) model is based on the equations

$$2\mathrm{C_3S} + 10.6\mathrm{H} \longrightarrow \mathrm{C_{3.4} - S_2 - H_8} + 2.6\mathrm{CH} \tag{C.13}$$

$$2\mathrm{C_2S} + 8.6\mathrm{H} \longrightarrow \mathrm{C_{3.4} - S_2 - H_8} + 0.6\mathrm{CH} \tag{C.14}$$

$$\mathrm{C_3A} + 3\mathrm{C\bar{S}H_2} + 26\mathrm{H} \longrightarrow \mathrm{C_6A\bar{S}_3H_{32}} \tag{C.15}$$

$$2\mathrm{C_3A} + \mathrm{C_6A\bar{S}_3H_{32}} + 4\mathrm{H} \longrightarrow 3\mathrm{C_4A\bar{S}H_{12}} \tag{C.16}$$

$$\mathrm{C_3A} + \mathrm{CH} + 12\mathrm{H} \longrightarrow \mathrm{C_4AH_{13}} \tag{C.17}$$

$$\mathrm{C_4AF} + 2\mathrm{CH} + 10\mathrm{H} \longrightarrow 2\mathrm{C_3(A,F)H_6} \tag{C.18}$$

With these stoichiometric relations and the molar volumes of the phases, it is possible to compute the volume fractions of the phases as a function of the DOH. In this approach, the aluminum bearing phases are ettringite ($\mathrm{C_6A\bar{S}_3H_{32}}$ or AFt), monosulfoaluminate ($3\mathrm{C_4A\bar{S}H_{12}}$ or AFm), hydrogarnet ($2\mathrm{C_3(A,F)H_6}$) and $\mathrm{C_4AH_{13}}$. With the progress of hydration, ettringite is

assumed to be completely converted into monosulfoaluminate if water and C_3A are available. No phases bearing carbonates are taken into account.

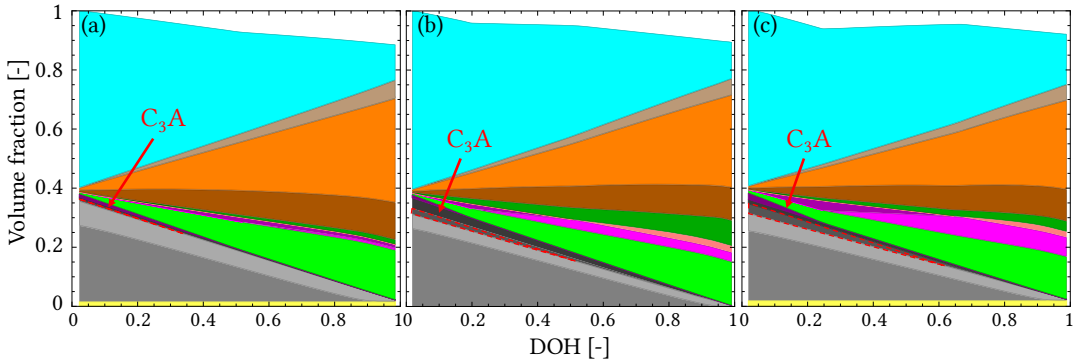
This model distinguishes between LD and HD C-S-H, as well as gel pores. The volumes of C-S-H HD and LD are given, respectively, by

$$V_{\text{HD}} = \frac{M_t - (M_r M_t)}{\rho_{\text{HD}}} \quad \text{and} \quad V_{\text{LD}} = \frac{M_r M_t}{\rho_{\text{LD}}} \tag{C.19}$$

where $\rho_{\text{HD}} = 1750\,\text{kg/m}^3$ and $\rho_{\text{LD}} = 1440\,\text{kg/m}^3$ are the "dried" densities of C-S-H HD and LD, respectively, as reported in (Tennis and Jennings 2000). The LD mass ratio with respect to the total mass of C-S-H denoted M_t and computed from the stoichiometric equations presented above, is denoted $M_r = 3.017(w/c)\text{DOH} - 1.347\text{DOH} + 0.538$. The volume of gel pore reads

$$V_{\text{gel pore}} = V_{\text{LD}} - \frac{M_r M_t}{\rho_{\text{HD}}}. \tag{C.20}$$

The TJ model is used to get the volume fraction of phases as a function of the DOH for three different commercial cements studied in previous works. For a ratio $w/c = 0.5$, the resulting phase assemblages are shown in Figure C.2. The variations on C_3A content lead to significant differences in the emergence of various Al-bearing phases: the fractions of AF-phases and C_4AH_{13} are clearly more significant in systems (b) and (c). The high-C_4AF content of cement (b) leads to a higher fraction of hydrogarnet formed.



Mass fractions of commercial cement [%]			
	(a)	(b)	(c)
C_3S	64.0	60.8	60.3
C_2S	23.1	12.5	17.2
C_3A	1.5	4.3	7.8
C_4AF	2.0	9.9	4.6
$C\bar{S}H_2$	2.0	2.0	4.3
"Inert"	3.5	0.0	4.4

Figure C.2 Volume fraction of phases versus the DOH for three commercial cements studied in references (a) (Honório et al. 2016a; Honório et al. 2018a; Honório et al. 2016b), (b) (Wyrzykowski et al. 2017) and (c) (Termkhajornkit and Barbarulo 2012) with $w/c = 0.5$. The variations on C_3A content lead to significant differences in the emergence of various Al-bearing phases. Colors defined in Figure C.3.

Figure C.3 shows the evolution of the volume fraction of phases with the DOH as estimated using the TJ model for the samples included in the test dataset, see Section 4.2.2. Note that in Figure C.3(c), hydration is stopped at a DOH of approximately 0.9 since there is no water anymore available.

D Database analysis: comparison with theoretical models

The experimental observations are compared with the theoretical bounds of elastic properties provided by analytical estimations for random heterogeneous media. The Voigt-Reuss bounds and the HS bounds, as introduced in Section 3.4, are considered. Figure D.4 shows the differences for each observation included in the training set between the experimental values of the elastic

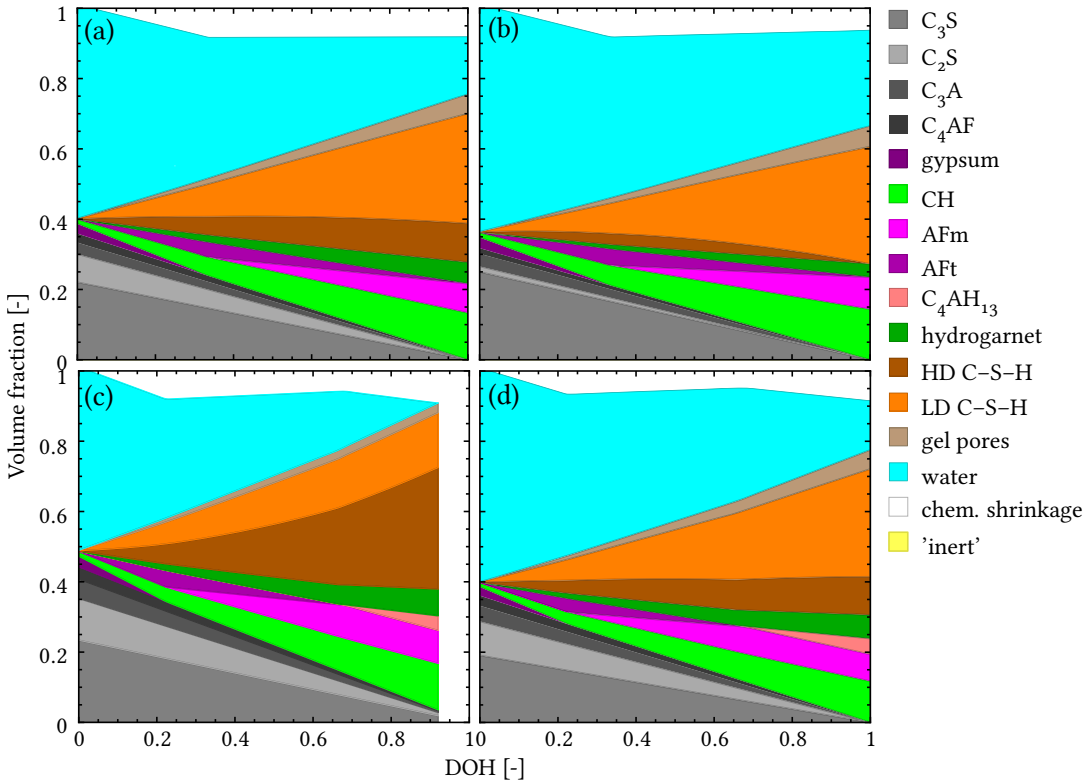


Figure C.3 Evolution of the volume fraction of different phases versus the DOH for three commercial cements studied by (a) (Constantinides and Ulm 2004) for $w/c = 0.5$, (b) (Haecker et al. 2005) for $w/c = 0.6$, (c and d) (Tamtsia et al. 2004) for (c) $w/c = 0.35$ and (d) $w/c = 0.5$.

properties and the upper Voigt and HS bounds obtained from the knowledge of the phase fractions and the homogenization schemes. The positive values correspond to experimental observations exceeding the upper bounds estimated from the theoretical models. All experimental observations for E and G are lower than the upper bounds, whereas a few experimental values for ν and K exceed the upper bounds. Variability, and uncertainties in experimental determination and bound computation may explain this observation. Only some experimental values of K exceed the bounds, their numbers in the training database are identified and depicted in Figure D.4(e) so that the reader can easily extract them from the data collection in case of interest.

Positions of the experimental observations with respect to the lower bounds are not shown since lower bounds being null and the elastic constants non-negative, all experimental observations within the training dataset satisfy the theoretical lower bounds.

E Leave-One-Out Cross-Validation

The Leave-One-Out Cross-Validation (LOOCV) is a technique to exploit the domains associated with larger prediction error or exhibits a marked non-linear behavior (Fuhg et al. 2020). This approach consists in using a k-fold cross-validation with $k = n$, n being the total number of observations. For each observation $i \in [1, n]$, a surrogate model $\mathcal{M}_{-i}$ is trained on $n - 1$ observations, which constitute a subset $\mathcal{M}_{-i}$. This training stage can be computationally expensive. The accuracy is finally computed using (Fuhg et al. 2020)

$$e_{\text{LOOCV}}(\mathbf{x}_i) = |\mathcal{M}(\mathbf{x}_i) - \mathcal{M}_{-i}(\mathbf{x}_i)| \quad \forall i \in [i, n] \tag{E.1}$$

where $\mathcal{M}(\mathbf{x}_i)$ is the metamodel of interest evaluated for the input $\mathbf{x}_i$. A small $e_{\text{LOOCV}}(\mathbf{x}_i)$ means that suppressing the observations i will not significantly affect the metamodel. In other words, the interpolations made around $\mathbf{x}_i$ are sufficiently accurate. Conversely, a large $e_{\text{LOOCV}}(\mathbf{x}_i)$ means that the information around $\mathbf{x}_i$ is lacking.

For ANN and GP, the five observations with larger e_{LOOCV} are $\{5, 6, 11, 16, 17\}$, all from (Helmuth and Turk 1966), and $\{16, 17, 69, 263, 364\}$, respectively. In the case of GP, data regarding

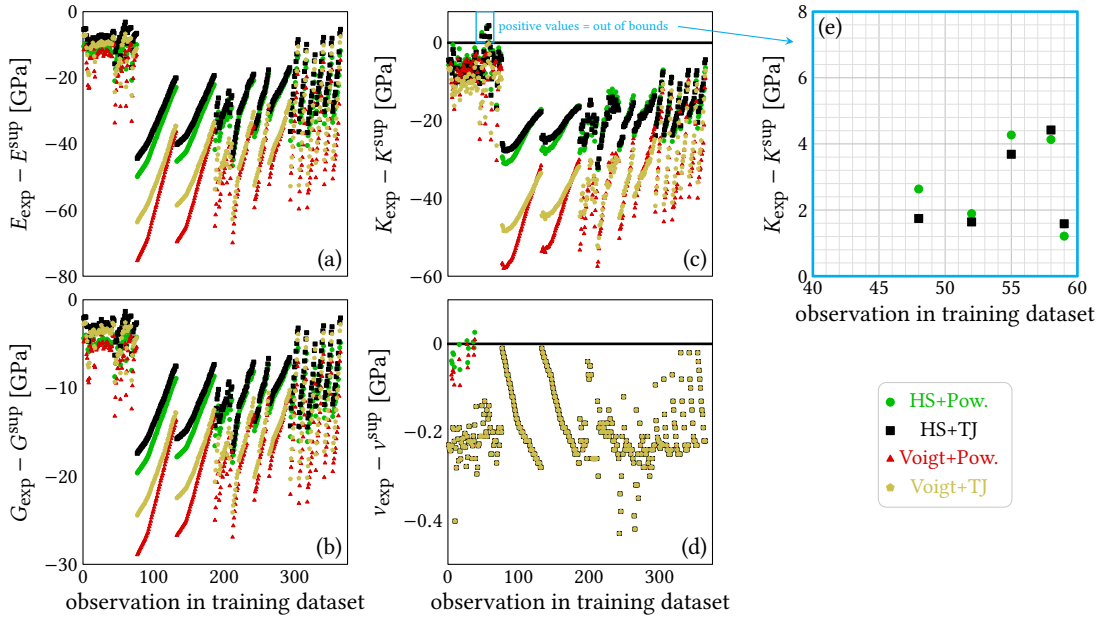


Figure D.4 Difference between the experimental values in the dataset (subscript 'exp') and the Voigt and HS upper bounds (superscript 'sup') of (a) E, (b) G, (c) K and (d) ν. (e) Identification numbers of the few experimental observations of K exceeding the theoretical upper bounds.

larger w/c values is lacking. Such information can be useful to guide future experimental campaigns and optimize experiment design.

References

Achour, M., F. Bignonnet, J.-F. Barthélémy, E. Rozière, and O. Amiri (2020). Multi-scale modeling of the chloride diffusivity and the elasticity of Portland cement paste. *Construction and Building Materials* 234:117124. [DOI], [HAL].

Acker, P. (2001). Micromechanical analysis of creep and shrinkage mechanisms. *Creep, shrinkage and durability mechanics of concrete and other quasi-brittle materials.*

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

Aller, L. H., I. Appenzeller, B. Baschek, K. Butler, C. De Loore, H. W. Duerbeck, M. F. El Eid, H. H. Fink, T. Herczeg, T. Richtler, H. Schneider, M. Scholz, W. Seggewiss, W. C. Seitter, J. Trümper, P. Ulmenschneider, R. Wehrse, and V. Weidemann (1996). *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology - New Series 'Group 6 Astronomy and Astrophysics' Volume 3 'Voigt: Astronomy and Astrophysics'. Extension and Supplement to Volume 2 'Stars and Star Clusters'*. Vol. 456. Springer.

Bary, B. and S. Béjaoui (2006). Assessment of diffusive and mechanical properties of hardened cement pastes using a multi-coated sphere assemblage model. *Cement and Concrete Research* 36(2):245–258. [DOI], [HAL].

Behnood, A., J. Olek, and M. A. Glinicki (2015). Predicting modulus elasticity of recycled aggregate concrete using M5 model tree algorithm. *Construction and Building Materials* 94:137–147. [DOI].

Ben Chaabene, W., M. Flah, and M. L. Nehdi (2020). Machine learning prediction of mechanical properties of concrete: Critical review. *Construction and Building Materials* 260:119889. [DOI].

Bengio, Y. and Y. Grandvalet (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research* 5:1089–1105. [OA].

Boumiz, A (1996). Mechanical properties of cement pastes and mortars at early ages: Evolution with time and degree of hydration. *Advanced Cement Based Materials* 3(3-4):94–106. [DOI].

Breiman, L. (1996). Bagging predictors. *Machine Learning* 24(2):123–140. [DOI], [OA].

Breiman, L., J. Friedman, R. Olshen, and C. Stone (2017). Regression Trees. *Classification and*

regression trees. Taylor & Francis. Chap. 8, pp 216–265. [DOI].

Bullard, J. W., E. J. Garboczi, P. E. Stutzman, P. Feng, A. S. Brand, L. Perry, J. Hagedorn, W. Griffin, and J. E. Terrill (2019). Measurement and modeling needs for microstructure and reactivity of next-generation concrete binders. *Cement and Concrete Composites* 101:24–31. [DOI].

Chamrova, R. (2010). Modelling and measurement of elastic properties of hydrating cement paste. PhD thesis. Switzerland: École Polytechnique Fédérale de Lausanne. [DOI].

Constantinides, G. and F.-J. Ulm (2004). The effect of two types of C-S-H on the elasticity of cement-based materials: Results from nanoindentation and micromechanical modeling. *Cement and Concrete Research* 34(1):67–80. [DOI], [HAL].

Duan, Z., S. Kou, and C. Poon (2013). Prediction of compressive strength of recycled aggregate concrete using artificial neural networks. *Construction and Building Materials* 40:1200–1206. [DOI].

Fuhg, J., C. Böhm, N. Bouklas, A. Fau, P. Wriggers, and M. Marino (2021). Model-data-driven constitutive responses: Application to a multiscale computational framework. *International Journal of Engineering Science* 167:103522. [DOI], [ARXIV].

Fuhg, J., A. Fau, and U. Nackenhorst (2020). State-of-the-art and comparative review of adaptive sampling methods for kriging. *Archives of Computational Methods in Engineering* 28(4):2689–2747. [DOI], [OA].

Fuhg, J. N., A. Fau, N. Bouklas, and M. Marino (2023). Enhancing phenomenological yield functions with data: challenges and opportunities. *European Journal of Mechanics - A/Solids* 99:104925. [DOI], [HAL].

Ghabezloo, S., J. Sulem, and J. Saint-Marc (2009). The effect of undrained heating on a fluid-saturated hardened cement paste. *Cement and Concrete Research* 39(1):54–64. [DOI], [ARXIV].

Golafshani, E. M. and A. Behnood (2018). Automatic regression methods for formulation of elastic modulus of recycled aggregate concrete. *Applied Soft Computing* 64:377–400. [DOI].

Guihard, V., F. Taillade, J.-P. Balayssac, B. Steck, and J. Sanahuja (2019). Permittivity measurement of cementitious materials and constituents with an open-ended coaxial probe: combination of experimental data, numerical modelling and a capacitive model. *RILEM Technical Letters* 4:39–48. [DOI], [OA].

Haecker, C.-J., E. Garboczi, J. Bullard, R. Bohn, Z. Sun, S. Shah, and T. Voigt (2005). Modeling the linear elastic properties of Portland cement paste. *Cement and Concrete Research* 35(10):1948–1960. [DOI], [HAL].

Hansen, T. C. (1986). Physical structure of hardened cement paste. A classical approach. *Materials and Structures* 19(6):423–436. [DOI].

Hashin, Z. and S. Shtrikman (1963). A variational approach to the theory of the elastic behaviour of multiphase materials. *Journal of the Mechanics and Physics of Solids* 11(2):127–140. [DOI].

Helmuth, R. A. and D. H. Turk (1966). *Elastic moduli of hardened Portland cement and tricalcium silicate pastes: effect of porosity*. Tech. rep. 90. Highway Research Board, pp 135–144. [HAL].

Hlobil, M. (2020). Distribution of hydration products in the microstructure of cement pastes. *Acta Polytechnica CTU Proceedings* 27(0):84–89. [DOI], [OA].

Honório, T., B. Bary, and F. Benboudjema (2016a). Multiscale estimation of ageing viscoelastic properties of cement-based materials: A combined analytical and numerical approach to estimate the behaviour at early age. *Cement and Concrete Research* 85:137–155. [DOI], [HAL].

Honório, T., B. Bary, and F. Benboudjema (2018a). Thermal properties of cement-based materials: Multiscale estimations at early-age. *Cement and Concrete Composites* 87:205–219. [DOI].

Honório, T., B. Bary, F. Benboudjema, and S. Poyet (2016b). Modeling hydration kinetics based on boundary nucleation and space-filling growth in a fixed confined zone. *Cement and Concrete Research* 83:31–44. [DOI].

Honório, T., T. Bore, F. Benboudjema, E. Vourc'h, and M. Ferhat (2020a). Dielectric properties of the pore solution in cement-based materials. *Journal of Molecular Liquids* 302:112548. [DOI], [OA].

Honório, T., L. Brochard, and B. Bary (2018b). Statistical variability of mechanical fields in thermo-poro-elasticity: Multiscale analytical estimations applied to cement-based materials at early-age. *Cement and Concrete Research* 110:24–41. [DOI], [HAL].

Honório, T., H. Carasek, and O. Cascudo (2020b). Electrical properties of cement-based materials:

Multiscale modeling and quantification of the variability. *Construction and Building Materials* 245:118461. [DOI], [OA].

Honório, T., P. Guerra, and A. Bourdot (2020c). Molecular simulation of the structure and elastic properties of ettringite and monosulfoaluminate. *Cement and Concrete Research* 135:106126. [DOI], [OA].

Kasperkiewicz, J., J. Racz, and A. Dubrawski (1995). HPC strength prediction using artificial neural network. *Journal of Computing in Civil Engineering* 9(4):279–284. [DOI].

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems (NIPS'17)*. 31st International Conference on Neural Information Processing Systems (Long Beach, USA, Dec. 4, 2017–Dec. 9, 2017), 3149–3157. [DOI], [HAL].

Königsberger, M., C. Hellmich, and B. Pichler (2016). Densification of C-S-H is mainly driven by available precipitation space, as quantified through an analytical cement hydration model based on NMR data. *Cement and Concrete Research* 88:170–183. [DOI].

Königsberger, M., T. Honório, J. Sanahuja, B. Delsaute, and B. Pichler (2021). Homogenization of nonaging basic creep of cementitious materials: A multiscale modeling benchmark. *Construction and Building Materials* 290:123144. [DOI], [HAL].

Lide, D. R., ed. (1997). *Handbook of Chemistry and Physics*. 77th ed. CRC Press. ISBN: 9780849304774.

Lura, P., O. M. Jensen, and K. van Breugel (2003). Autogenous shrinkage in high-performance cement paste: An evaluation of basic mechanisms. *Cement and Concrete Research* 33(2):223–232. [DOI].

Manzano, H. (2009). Atomistic Simulation studies of the Cement Paste Components. PhD thesis. Spain: Universidad del País Vasco. [HDL].

Maruyama, I. and G. Igarashi (2014). Cement reaction and resultant physical properties of cement paste. *Journal of Advanced Concrete Technology* 12(6):200–213. [DOI], [OA].

Monteiro, P. J. and C. Chang (1995). The elastic moduli of calcium hydroxide. *Cement and Concrete Research* 25(8):1605–1609. [DOI].

Muller, A., K. Scrivener, A. Gajewicz, and P. McDonald (2012). Densification of C–S–H measured by 1H NMR relaxometry. *The Journal of Physical Chemistry C* 117(1):403–412. [DOI].

Mura, T. (1987). *Micromechanics of Defects in Solids*. 2nd ed. Mechanics of Elastic and Inelastic Solids. Martinus Nijhoff Publishers. [DOI].

Nematzadeh, Z., R. Ibrahim, and A. Selamat (2015). Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques. 10th Asian Control Conference (ASCC) (Kota Kinabalu, Malaysia, May 31, 2015–June 3, 2015). IEEE. [DOI].

Olson, G. B. (1997). Computational design of hierarchically structured materials. *Science* 277(5330):1237–1242. [DOI], [HAL].

Patel, R., Q. T. Phung, S. Seetharam, J. Perko, D. Jacques, N. Maes, G. De Schutter, G. Ye, and K. Van Breugel (2016). Diffusivity of saturated ordinary Portland cement-based materials: A critical review of experimental and analytical modelling approaches. *Cement and Concrete Research* 90:52–72. [DOI].

Pichler, B. and C. Hellmich (2011). Upscaling quasi-brittle strength of cement paste and mortar: A multi-scale engineering mechanics model. *Cement and Concrete Research* 41(5):467–476. [DOI].

Pichler, B., C. Hellmich, J. Eberhardsteiner, J. Wasserbauer, P. Termkhajornkit, R. Barbarulo, and G. Chanvillard (2013). Effect of gel-space ratio and microstructure on strength of hydrating cementitious materials: An engineering micromechanics approach. *Cement and Concrete Research* 45:55–68. [DOI].

Powers, T. C. (1960). Physical properties of cement paste. *Chemistry of Cement. Proceedings of the Fourth International Symposium* (Washington D.C., USA, Oct. 2, 1960–Oct. 7, 1960), pp 577–613. [OA].

Powers, T. C. and T. L. Brownyard (1946). Studies of the physical properties of hardened Portland cement paste. *ACI Journal Proceedings* 43(9):101–132. [DOI].

Rodriguez, J., A. Perez, and J. Lozano (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(3):569–575. [DOI].

Sanahuja, J., L. Dormieux, and G. Chanvillard (2007). Modelling elasticity of a hydrating cement paste. *Cement and Concrete Research* 37(10):1427–1439. [DOI].

Šavija, B., H. Zhang, and E. Schlangen (2020). Micromechanical testing and modelling of blast furnace slag cement pastes. *Construction and Building Materials* 239:117841. [DOI], [OA].

Speziale, S., F. Jiang, Z. Mao, P. Monteiro, H.-R. Wenk, T. Duffy, and F. Schilling (2008). Single-crystal elastic constants of natural ettringite. *Cement and Concrete Research* 38(7):885–889. [DOI].

Sun, Z., E. Garboczi, and S. Shah (2007). Modeling the elastic properties of concrete composites: Experiment, differential effective medium theory, and numerical simulation. *Cement and Concrete Composites* 29(1):22–38. [DOI].

Tamtsia, B., J. Beaudoin, and J. Marchand (2004). The early age short-term creep of hardening cement paste: load-induced hydration effects. *Cement and Concrete Composites* 26(5):481–489. [DOI], [OA].

Tennis, P. and H. Jennings (2000). A model for two types of calcium silicate hydrate in the microstructure of Portland cement pastes. *Cement and Concrete Research* 30(6):855–863. [DOI], [HAL].

Termkhajornkit, P. and R. Barbarulo (2012). Modeling the coupled effects of temperature and fineness of Portland cement on the hydration kinetics in cement paste. *Cement and Concrete Research* 42(3):526–538. [DOI].

Torquato, S. (2002). *Random Heterogeneous Materials: Microstructure and Macroscopic Properties*. Springer. [DOI].

Ulm, F.-J., G. Constantinides, and F. H. Heukamp (2004). Is concrete a poromechanics materials?–A multiscale investigation of poroelastic properties. *Materials and Structures* 37(1):43–58. [DOI], [HAL].

Velez, K., S. Maximilien, D. Damidot, G. Fantozzi, and F. Sorrentino (2001). Determination by nanoindentation of elastic modulus and hardness of pure constituents of Portland cement clinker. *Cement and Concrete Research* 31(4):555–561. [DOI].

Wang, X. and K. Subramaniam (2011). Ultrasonic monitoring of capillary porosity and elastic properties in hydrating cement paste. *Cement and Concrete Composites* 33(3):389–401. [DOI].

Wangler, T., N. Roussel, F. Bos, T. A. Salet, and R. Flatt (2019). Digital concrete: A review. *Cement and Concrete Research* 123:105780. [DOI], [OA].

Williams, C. and C. E. Rasmussen (1996). Gaussian processes for regression. *Advances in Neural Information Processing Systems (NIPS'95)*. 8th International Conference on Neural Information Processing Systems (Denver, USA, Nov. 27, 1995–Nov. 30, 1995), pp 1–7. [OA], [HAL].

Wolfram (2021). *Mathematica*. Version 13.0.0. [URL].

Wyrzykowski, M., J. Sanahuja, L. Charpin, M. Königsberger, C. Hellmich, B. Pichler, L. Valentini, T. Honório, V. Smilauer, K. Hajkova, G. Ye, P. Gao, C. Dunant, A. Hilaire, S. Bishnoi, and M. Azenha (2017). Numerical benchmark campaign of COST Action TU1404 - microstructural modelling. *RILEM Technical Letters* 2:99–107. [DOI], [OA].

Yan, K. and C. Shi (2010). Prediction of elastic modulus of normal and high strength concrete by support vector machine. *Construction and Building Materials* 24(8):1479–1485. [DOI].

Yeh, I.-C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research* 28(12):1797–1808. [DOI].

Yeh, I.-C. and L.-C. Lien (2009). Knowledge discovery of concrete material using Genetic Operation Trees. *Expert Systems with Applications* 36(3):5807–5812. [DOI].

Young, B., A. Hall, L. Pilon, P. Gupta, and G. Sant (2019). 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* 115:379–388. [DOI].

Zaoui, A. (2002). Continuum micromechanics: Survey. *Journal of Engineering Mechanics* 128(8):808–816. [DOI], [HAL].

Zimmerman, R. W. (1992). Hashin-Shtrikman bounds on the Poisson ratio of a composite material. *Mechanics Research Communications* 19(6):563–569. [DOI].

Authors' contributions TH: Conceptualization, Methodology, Software, Visualization, Data curation, Investigation, Formal analysis, Writing - Original Draft, Writing - Review and Editing, Supervision. SAH: Conceptualization, Investigation, Formal analysis, Writing - Original Draft. AF: Conceptualization, Methodology, Writing - Original Draft, Writing - Review and Editing.

Supplementary Material The database with observations on cement pastes with cement composition and elastic constants is available at the permalink https://github.com/tuliohf/cdi/blob/6e93428oob6b8a7ac898712e34af523aob8ofbc4/Dataset_Elastic_Constants_OPC_Pastes.

Acknowledgements None.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare that they have no competing interests.

Journal's Note JTCAM remains neutral with regard to the content of the publication and institutional affiliations.