Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5dd00383k

GEOMIND: a hybrid generative artificial intelligence model for geopolymer design and optimization

Sébastien Rousseau, [ab] Assil Bouzid, [a] Sylvie Rossignol [a] and Ameni Gharzouni *[a]

Geopolymers are an emerging class of eco-friendly materials with a wide range of applications. Nevertheless, achieving compounds for a specific application requires extensive experimental efforts in finding the accurate formulation of precursors. Using artificial intelligence, we tackle the challenging task of formulating accurate geopolymer mixtures that meets a predefined set of properties that the final materials should feature. This task goes beyond the prediction of materials' properties and focuses on the actual material design. To this end, we build a high-quality in-house experimental database of geopolymer formulations and their physical properties. We develop a customized trained machine learning framework based on two variational autoencoder modules. The first predicts the formulations that correspond to an array of target properties and the second corrects the requested properties to better match the predicted formulation. Furthermore, our model embeds a geopolymer feasibility bloc that ensures that the predicted materials can be synthesized. Overall, this framework is able to predict formulations and their corresponding properties with less than 10% error bar on a set of key properties of the final material encompassing the viscosity, the density and the compressive strength. The suggested methodology paves the way for the systematic application of AI-based materials design in the development of eco-friendly novel materials for different applications.

Received 25th August 2025
Accepted 3rd June 2026

DOI: 10.1039/d5dd00383k

rsc.li/digitaldiscovery

1. Introduction

Geopolymers are innovative and sustainable alternative mineral binders. They are synthesized through the alkaline activation of aluminosilicate precursors at room temperature. The resulting three-dimensional network exhibits high mechanical strength and good resistance to high temperatures and aggressive environments.[1,2] These properties make them suitable for a wide range of applications.[3] Indeed, they have demonstrated potential in large-scale applications such as the construction sector as an alternative to Portland cement due to their comparable mechanical performances and lower environmental impact.[4] Additionally, their thermal stability makes them promising candidates for fire-resistant materials[5] and their dielectric properties enable their integration in advanced technologies such as electromagnetic and microwave applications.[6] However, the performances of geopolymers are highly dependent on the choice of precursors (*i.e.*, aluminosilicate and alkaline sources). Among the aluminosilicate sources, metakaolin is commonly used. Yet, its reactivity can be affected by several parameters including the starting kaolinite, the presence of impurities (Si/Al molar ratio) and the calcination temperature, time and process.[7] The alkali activator also has a crucial impact on the geopolymer properties. For instance, it was demonstrated that potassium based geopolymers exhibit lower viscosity and higher mechanical strength compared to sodium based ones.[8] Given the large number of precursors, identifying optimal geopolymer formulations remains a complex and challenging task that requires extensive experimental work. In this context, the use of artificial intelligence (AI) algorithms can be useful in supporting the experimental work and efficiently guiding the design of tailored geopolymer formulations for specific applications.

In the last few years, the use of machine learning models and methods has emerged widely in materials science in general and in novel materials design and process optimizations specifically.[9] This rise is owing to the capabilities offered by machine learning methods to solve complex problems and establish correlations in high-dimensional spaces. Nevertheless, this capacity is strongly linked to the quality of the data used for training the models and their representativity of the problem to be solved.

The application of AI methods in the field of geopolymers has gained growing attention. Recently, several studies have focused on predicting material properties from theoretical chemical formulations. Specifically, within this scope most of the available literature focuses on the prediction of the

[a]*Institut de Recherche sur les Céramiques, UMR 7315 CNRS - Université de Limoges, Centre Européen de la Céramique, 12 rue Atlantis, 87068, Limoges Cedex, France. E-mail: ameni.gharzouni@unilim.fr*

[b]*XLIM, UMR CNRS 7252, 16 Rue Jules Vallès, 19100 Brive La Gaillarde, France*

compressive strength of geopolymers using various machine learning techniques. Chen *et al.*[10] have demonstrated that an Artificial Neural Networks (ANNs) model outperforms the Gene Expression Programming (GEP) model in predicting the compressive strength of fly ash and slag-based geopolymers with a mean absolute error (MAE) of 5.85 MPa. Other studies[11–20] compare regressor machine learning models such as the K-Nearest Neighbor (KNN), Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB) and eXtreme Gradient Boosting (XGB), and Artificial and Deep Neural Network (A/DNN) for predicting the compressive strength material properties. Ensemble learning is often used with these machine learning models to achieve an improved global accuracy of predictions. The obtained results reveal that the XGB and the deep learning methods are generally the most accurate. However, no existing generative AI model in the materials design literature currently addresses the specific task of geopolymer precursor design, particularly targeting different properties at the same time.

The most closely related tasks reported in the literature deal with precursor selection for synthesis of thermodynamically stable complex solid-state oxide materials,[21] and precursor recommendation model for different inorganic materials.[22] Other authors[23] have also succeeded in generating novel molecules using an autoencoder model trained on hundreds of thousands of existing chemical structures. Xie *et al.*[24] have also developed a graph-based deep learning model that learns from raw crystal structures and accurately predicts multiple material properties, enabling data-driven materials design.

The main difficulty in the development of an AI supported materials design algorithm is the development of a high-quality experimental database that remains a highly costly task. The widely adopted strategies are either to collect data from the existent literature to build the training database[25] or to use hypothetical or empirical data-augmentation techniques.[26–29] While these strategies are tempting, they raise critical issues. In particular, when data are collected from the literature, one cannot guarantee the consistency of the database in the sense that collected samples might be generated under different conditions and using various processing roots. This can lead to model-fitting problems and hazardous prediction capabilities. As for the data augmentation, it usually relies on two sub-models, a generator to predict new samples and a selector to validate the samples adding them to the database. Several studies highlighted that the accuracy of property predictions is poorly improved as the added data are strongly correlated with the known data.[16,25] Consequently, the model tends to specialize on the training data instead of generalizing.

In this work, we tackle the materials design problem and present a machine learning algorithm that (i) is capable of suggesting formulations of geopolymer precursors and solutions that meet a given target material's properties, (ii) takes into account simultaneously three key properties of geopolymer materials: density, viscosity and compressive strength, (iii) is capable of correcting the user request of three unrealistic properties by suggesting the closest possible properties/formulation couple that is realistic, (iv) takes into account geopolymerization rules and expert-knowledge to avoid unrealistic mixtures. The proposed scheme is deeply grounded on a high-quality homemade material database containing more than 100 samples. Our proposed algorithm outperforms state of the art methods at various levels, providing thereby a tool that will guide the synthesis of geopolymer materials.

2. Experimental and AI methodology

2.1. Synthesis and characterization

2.1.1 Precursors and sample preparation. Geopolymer samples were synthesized using five metakaolins (named Mx with x varying from 1 to 5) and three commercial potassium silicate solutions denoted as S1, S3 and S3′ and a sodium based one named SNa with Si/M (with M = K or Na) as summarized in Table 1. In order to modify the Si/M molar ratio, potassium or sodium hydroxide pellets were dissolved in the starting silicate solutions. Then, metakaolins were added (see Fig. 1). The obtained mixtures were placed in a closed sealable polystyrene mold at room temperature (20 °C).

2.1.2 Characterization techniques. The initial viscosity η_0 (directly after mixing the metakaolin and the alkaline solution) was measured with a rotational Brookfield Viscometer DV2T

Table 1 Nomenclature, supplier and chemical composition of the different precursors

			Chemical composition (wt%)				
Precursors	Name	Supplier	SiO_2	Al_2O_3	H_2O	M_2O (M = K or Na)	Purity (%)
Metakaolins	M1	Imerys	55.0	40.0			
	M2		54.0	39.0			
	M3		54.0	46.0			
	M4		52.4	45.3			
	M5	Argeco	59.9	35.3			
Alkaline silicate solutions	S1	Woellner	14.3		79.3	6.4	
	S3′		23.4		54.9	21.7	
	S3		18.7		59.4	21.9	
	SNa		27.5		64.2	8.3	
Alkali hydroxides	KOH	Sigma-Aldrich					85.2
	NaOH						97.0

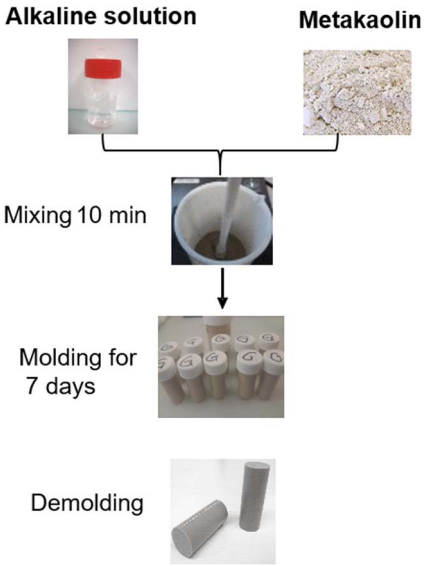

Fig. 1 Synthesis protocol of geopolymer materials.

coupled with a low shear, low viscosity cylindrical spindle LV-04 (64), and the rotational speed was varied, starting at 100 rpm (for viscosity up to 6 Pa s) and ending at 1 rpm (up to 1300 Pa s). Cylindrical polystyrene vessels ($Ø = 28$ mm) were used.

Uniaxial compression tests were undertaken using an Instron 5969 instrument equipped with a 50 kN load cell, and Bluehill3 software. The tests were performed on ten cylindrical samples with an aspect ratio of 2 ($Ø = 15$ mm, $h = 30$ mm) after 7 days of endogenous consolidation at room temperature. The crosshead speed was about 0.5 mm min^{-1}.

2.2. Machine learning framework

In this work, we resort to machine learning techniques with a particular focus on deep learning approaches and specifically artificial neural network.[30] MultiLayer Perceptrons (MLPs) are commonly used for regression and classification tasks.[31] These feedforward networks with nonlinear activations and hidden layers provide robust predictive performance across diverse datasets. For generative tasks, Variational Autoencoders (VAEs) are adopted,[32] where the encoder maps input data to a latent space with a probabilistic prior (typically Gaussian) and the decoder reconstructs samples from latent variables, enabling efficient exploration of the underlying data distribution.

Bayesian optimization (BO) is another machine learning method that is widely used to efficiently tune machine learning models' hyperparameters, the prediction of material properties[13] and material selection.[15] It optimizes a given objective function over continuous domains with typically less than 20 optimization dimensions.

Additionally, eXtreme Gradient Boosting Regressors (XGBRs) are employed as ensemble tree-based models, iteratively correcting residual errors to produce accurate and robust predictions.[33] As illustrated in Fig. 2, the model is initialized with a first decision tree, then additional trees are sequentially added to the model during the supervised fitting phase. Each new tree

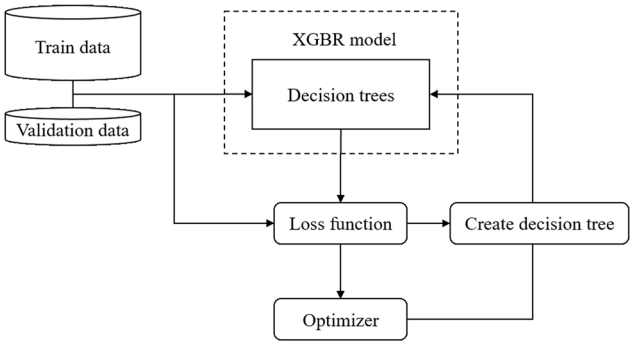

Fig. 2 XGBR data-driven building and updating diagram.

is trained to correct the residual errors of the ensemble model so far and adjusts its prediction. The final output of the model is obtained as the sum of the predictions of all the trees.

3. Results and discussion

3.1. Database construction and data correlations

Experimental data are collected in an in-house database containing 112 data samples of laboratory-manufactured geopolymer materials. Importantly, all the samples were produced under the same experimental conditions as detailed above. The database contains samples obtained by various combinations of 11 precursors, their mixtures' molar ratios and properties in the fresh (initial viscosity, density) and consolidated states (density, compressive strength). The selected metakaolins cover a wide range of Si/Al molar ratios and reactivity as they have different purities, calcination methods and suppliers (Table 1). Similarly, the used silicate solutions cover distinct silica and alkali contents and were modified by the addition of alkali hydroxide and/or mixing the commercial solutions. Actually, we aimed to explore the full range of molar ratios compatible with geopolymer formation. Therefore, we explore a wide-enough range of compositions that samples the whole geopolymer domain in the Al–Si–M/O ternary diagram (with M being an alkali cation) as shown in Fig. 3A.[34,35] In order to represent the chemical infeasibility of some particular mixtures of precursors, 14 non-feasible samples that do not induce the formation of a geopolymer network are added to the database.

For practical considerations, classes of geopolymer precursors are defined. Specifically, S1, S3, S3′ and SNa refer to silicate solutions. KOH and NaOH refer to solutions that were modified by adding potassium hydroxide and sodium hydroxide, respectively. M1, M2, M3, M4, M5 classes refer to the various metakaolins used in this work. Following these definitions, a sample is a data point that is described by 15 features: 11 precursor mass ratios (1 value per precursor class) and 4 material properties. A precursor mass ratio value lies between 0 and 1 giving the proportion of the precursor in the mixture. Consequently, precursor data format is a vector whose sum is equal to 1. The considered properties are the initial viscosity, fresh mixture density, final density, and compressive strength.

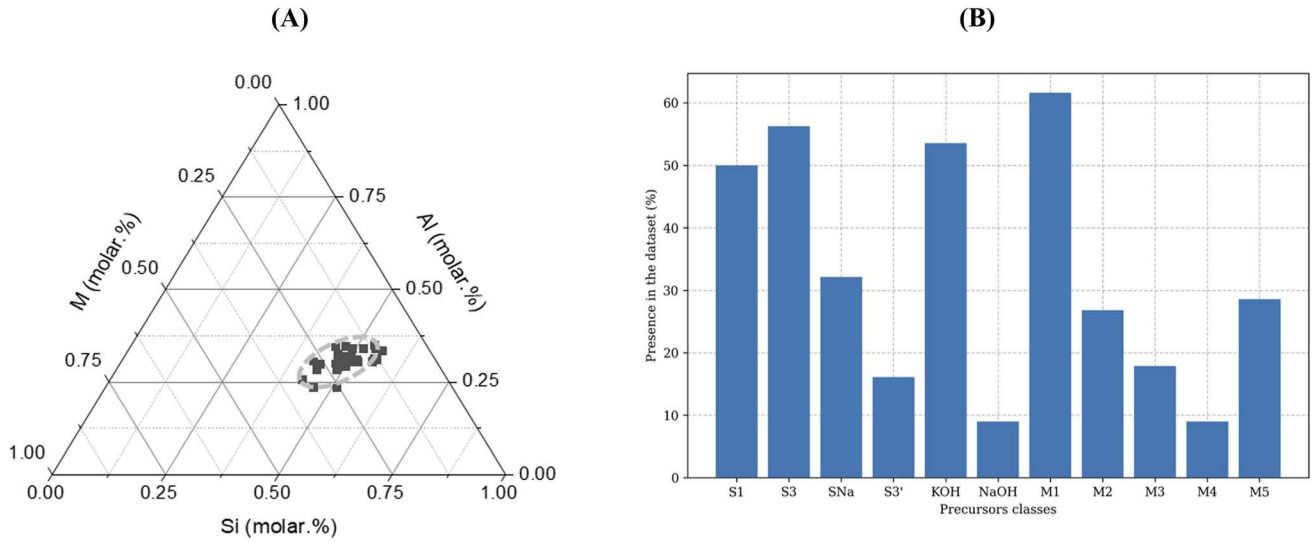

Fig. 3 (A) Positioning of the database samples in the geopolymer feasibility domain within the Al–Si–M/O ternary diagram (with M being an alkali cation) and (B) distribution of the different precursors in the dataset.

For the 14 unfeasible samples, a numerical value of −1 in the property fields indicates an impossible property measurement.

Fig. 3B presents the distribution of samples that contain each precursor in the database. We observe that all the classes are well represented with a fraction of samples containing a given precursor varying between 9 and 61%, indicating well-balanced data. This data-balanced representation of the samples is of paramount importance for developing machine learning models that can span a wide chemical space in an unbiased manner.

Fig. 4 shows the distribution of the measured experimental properties and their correlations over the 98 feasible mixtures. The color map indicates the concentration of samples and is constructed through a linear interpolation and a Gaussian filter for the sake of visibility.

Generally, dense and viscous geopolymers exhibit the highest compressive strength. Fig. 4 hints that the compressive strength of the consolidated material tends to increase along with the density and viscosity of the mixture. Nevertheless, this dependency looks nonlinear and feature complex correlations. The correlation of density and viscosity shows no clear low and shows, similar to other correlations maps, that for achieving a given target viscosity, one can have an array of samples with different compressive strengths. Such a correlation is actually expected and strongly depends on the nature of the precursor used to formulate the materials. As small changes in the precursor composition can affect the final material properties, the materials design task should account for the data complexity and take advantage of all the existent data in formulating the geopolymer materials. As such, a multi-

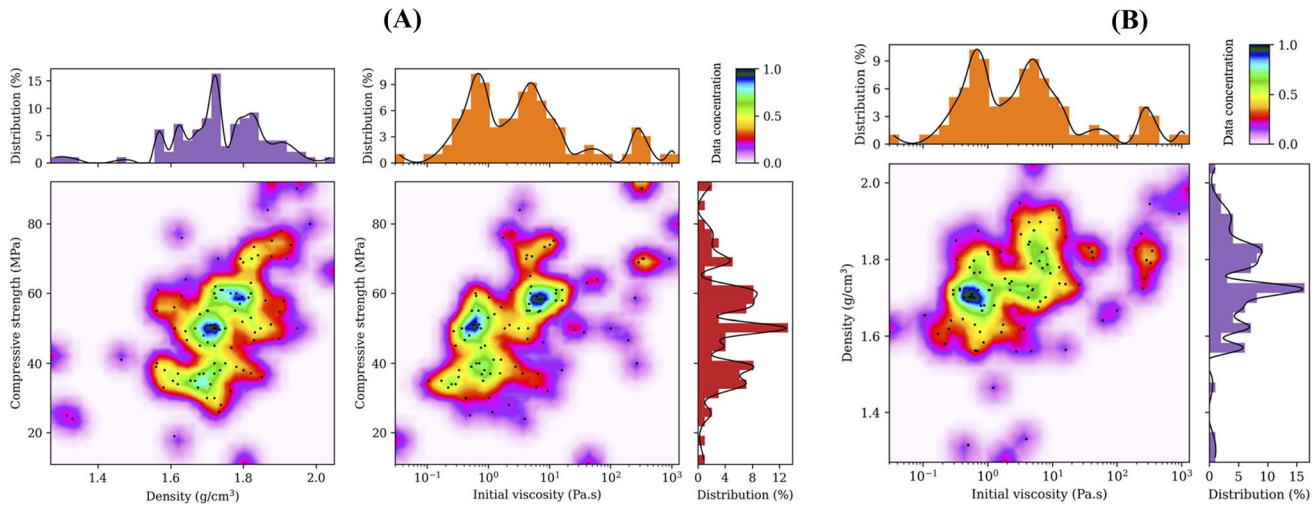

Fig. 4 Inter-correlated two-dimensional histograms showing (A) compressive strength as a function of material density and initial viscosity and (B) material density as a function of initial viscosity across the 98 feasible samples in the database. The color scale represents the normalized data concentration (red/purple: low, green/yellow: intermediate and blue: high data-point concentration).

property prediction approach is required and seems an adequate strategy to achieve realistic precursor/property formulations.

3.2. Machine learning modeling

3.2.1 Data preprocessing. Geopolymer formulations can be varied according to the target properties. However, their number must be determined before the models are trained, as neural networks expect descriptors represented by tensors with fixed dimensions as inputs and outputs. As described in Section 3.1, we consider a vector of 4 properties: the initial viscosity and the density of the fresh mixtures and the consolidated material density and compressive strength. The second entry is a vector of 11 precursors with mass ratios between 0 and 1. All the properties in datasets are normalized with a mean of 0 and a standard deviation of 1, except for the viscosity as it shows an extremely wide range between 0.2 Pa s and 1314 Pa s. Therefore, this property is categorized into 3 classes: low, medium and high viscosity, and formalized in a vector of length 3 where all values are located according to their class and normalized between 0 and 1, the value −1 is imputed to the other locations. Low viscosity is below 2 Pa s, which means the material could be projectable, while high viscosity is above 100 Pa s. By adopting this trick of categorizing viscosity into three classes, the model will predict first the adequate range of viscosity, then its value. Without this categorization, we encountered limited model performances.

We note that among the 112 samples of the database, a small fraction of less than 5% had one missing experimental dataset (20). Therefore, we applied and tested a data-augmentation strategy and assessed its impact on the performance and accuracy as it will be explained below.

3.2.2 Model training and evaluation. Considering the state-of-art and objectives, we consider VAE models featuring a feed-forward neural network architecture with densely connected layers through numerous perceptron neurons. The model hyperparameters are obtained through extensive tests where they were varied. The final optimized hyperparameters of each VAE model are provided in the SI (Tables S1 and S2). The model training phase is configured with an RMSprop optimizer to perform stochastic gradient descent updating the average of squared gradients which directly involve training rate variations over time. As the in-house database contains 112 samples, we resort to stratified k-fold cross-validation specifically using 28 folds in order to maximize the size of training sets while maintaining representative distributions of the different meta-kaolin types. This approach was intended to provide a robust evaluation of model performance within the explored formulation domain, though this may limit generalization to novel inputs. In this way, the accuracy levels of the models are clearly defined and all predicted property results can be observed and compared with experimental properties when samples come from a validation set.

3.2.2.1 The simulator model. Before trying to develop a model that formulates geopolymers for a specific application, we first develop a model that predicts the properties of the material from its chemical formulation, as commonly done in the recent literature. This model is referred to hereafter as the Simulator model and is illustrated in Fig. 5. In order to predict properties for only feasible geopolymers, we embed expert knowledge into the Simulator model as a final bloc. Specifically, this bloc calculates molar ratios between silica, alumina, alkali cation and water from precursor mass ratios and verifies the feasibility of the mixture according to geopolymerization laws based on established existence domains of geopolymer materials. If the mixture is chemically unfeasible or does not correspond to known geopolymer domains, the model returns negative values of properties (−1). Preliminary experiments have shown that without this block, the model suggested formulations that were not practically feasible: either

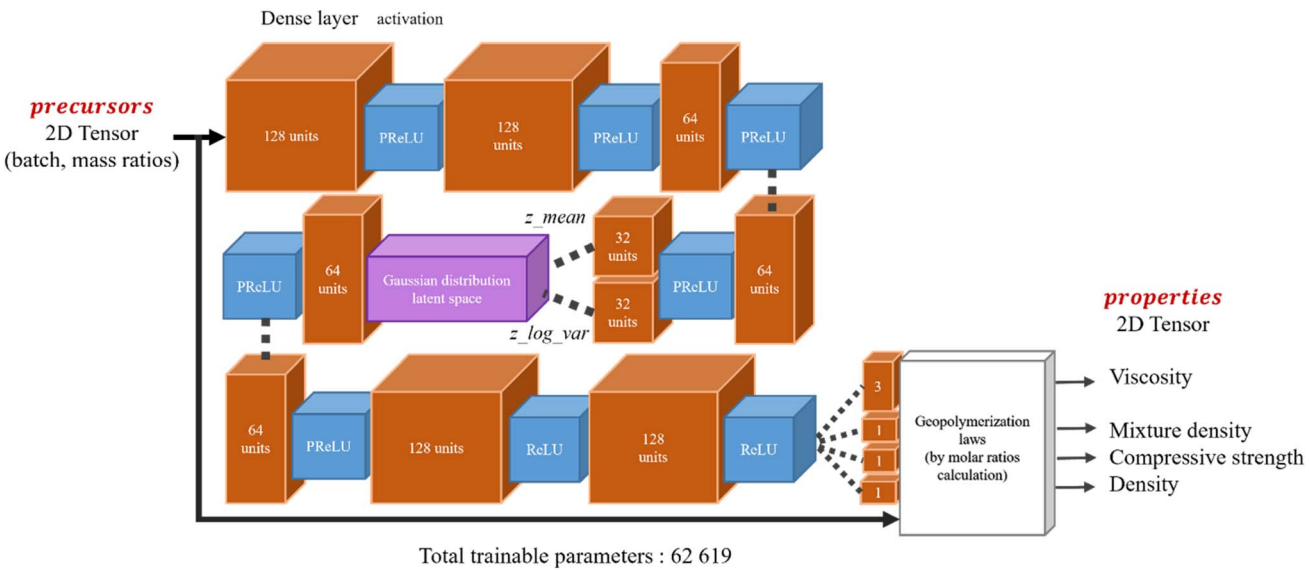

Fig. 5 The architecture of the Simulator model.

excessively viscous mixtures that could not be mixed due to a large amount of metakaolin, or overly liquid formulations that led to material sedimentation and produced friable samples that could be broken manually.

Fig. 6A shows that the Simulator model fits properly on training and validation sets and achieves low MAE on the properties. This result demonstrates that precursors and properties are directly linked and that the VAE can capture this nonlinear relationship. Furthermore, the adopted data preprocessing strategy is convenient for achieving stable and efficient neural networks. In order to quantitatively assess the role of the expert-knowledge block, we conducted a comparison of the Simulator model's cross-validation, trained either with or without molar ratio calculations (Fig. 6B). One finds that the hybrid model achieves a better accuracy as demonstrated by the evolution of the loss curves shown in Fig. 6B. Indeed, with the expert knowledge, a faster and more stable decrease in the MAE is observed, indicating that incorporating the expert knowledge significantly improves the model's learning efficiency. While additional metrics, such as feasibility rates or constraint violation statistics, could offer further insights, the current analysis, combining experimental observations and measurable performance improvements, supports the positive added value of the expert-knowledge block.

In addition to that, we exploit the Simulator model predictive capacity to impute the few sparse missing entries in the

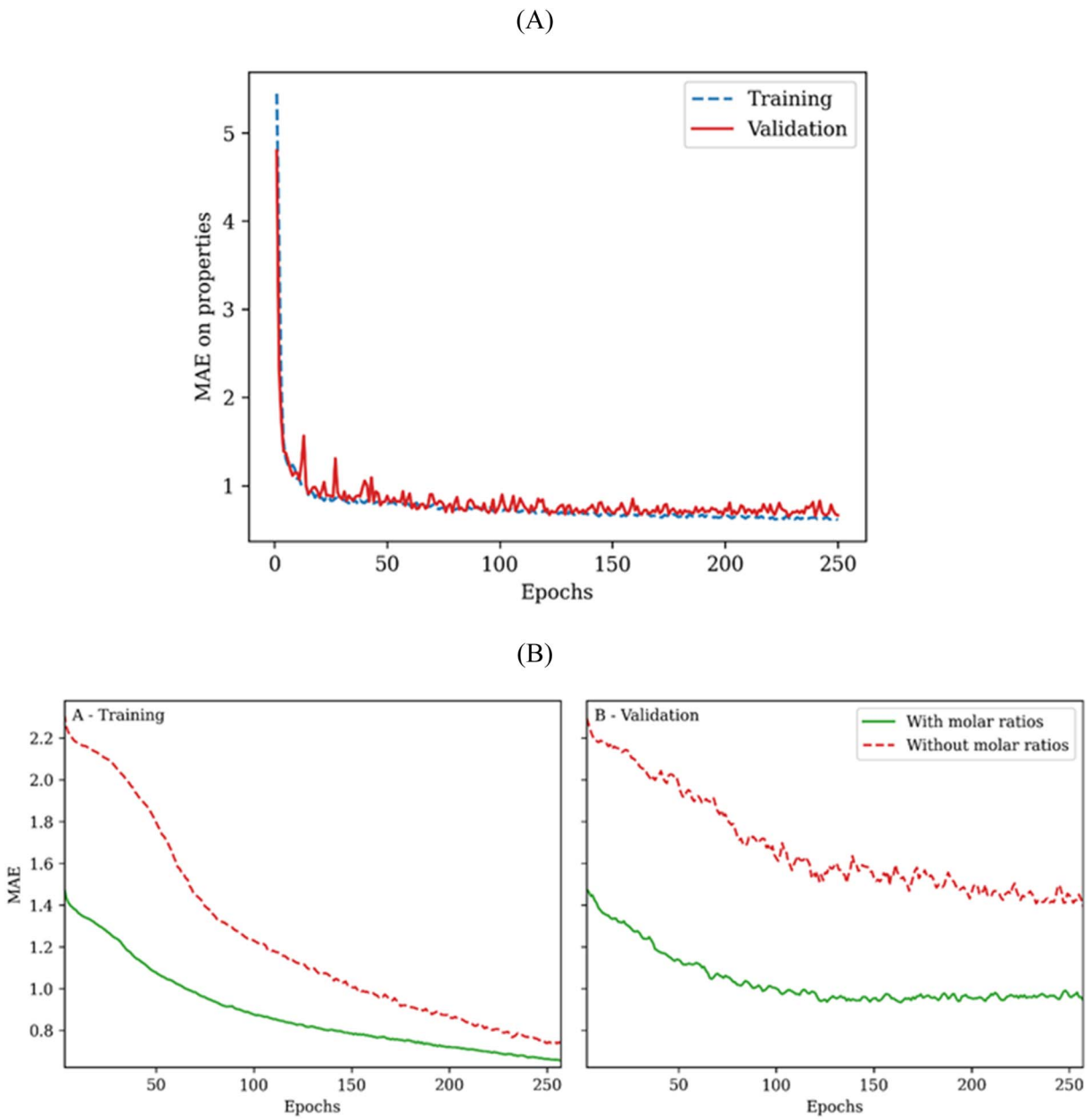

Fig. 6 (A) The Simulator model's loss curves and (B) the Simulator model's cross-validation mean loss curves (moving average), trained either with or without molar ratio calculations.

database. Actually, missing data imputation is regularly addressed in the deep learning literature with various methods depending on the application and the considered data type.[36] Missing data are either interpolated, randomly imputed, fixed at a normalized value, discriminated or generated. Generally, generative methods outperform the other techniques, as they follow an explicit modeling of the data distribution. In our specific case, samples with a missing data point are by large a minority (less than 5%), which makes it possible to generate them with the Simulator model. In practice, these samples are excluded from the training process, and the specific missing properties are then predicted from the converged model. Subsequently, we compare the models trained on the database with the few missing datapoints that retrained on the slightly data augmented database and consider the same test set. Interestingly, we find that models fitted on the data-augmented database are slightly more precise than models trained on the data that include few missing values, as illustrated in Table 2. Hence, we stick to the data-augmented database for the remaining part of the work.

While useful, using the Simulator model for designing materials is a non-trivial task as a given final set of properties might be the result of various precursor mixtures. Therefore, as the simulator is non-reversible, it is highly desirable to develop a model that does the inverse task compared to the simulator.

3.2.2.2 The formulator model. The Formulator model, illustrated in Fig. 7, attempts to predict precursor mass ratios from material properties. The Formulator model, illustrated in Fig. 7, attempts to predict precursor mass ratios from material properties. In contrast to the Simulator model, it cannot benefit from a molar ratio calculator (the expert knowledge) as these ratios are computed from precursors that are here the output of the model and an infeasible mixture should nevertheless end up with its precursor mass ratios, particularly with the database which also includes infeasible samples presenting valid precursor mass ratios with negative property values. Yet, these property values have a large number of possible precursors (at least 14 cases represented in the database). To ensure the predicted precursor fractions sum to 1.0, a post-hoc L1 normalization step is applied to the Formulator output. This step has a negligible impact on property predictions.

Fig. 8 shows the training and validation loss curves of the Formulator model. In strong contrast to the efficient contrast of the Simulator model, the Formulator model struggles to find balance in the validation set. The instability of the Formulator is due to the presence of samples in the dataset with radically different precursors that can produce similar material properties. This creates large training errors, as the model may predict certain precursors based on properties, then finds out that those predictions are incorrect for specific samples during loss

Table 2 Mean absolute errors between predicted and experimental properties with and without data augmentation

Property	Viscosity (Pa s)	Mixture density (g cm^{-3})	Compressive strength (MPa)	Material density (g cm^{-3})
Missing data proportion	5.4%	2.7%	3.6%	6.2%
MAE without data augmentation	39.2	0.097	7.9	0.142
MAE with data augmentation	24.5	0.093	7.6	0.121

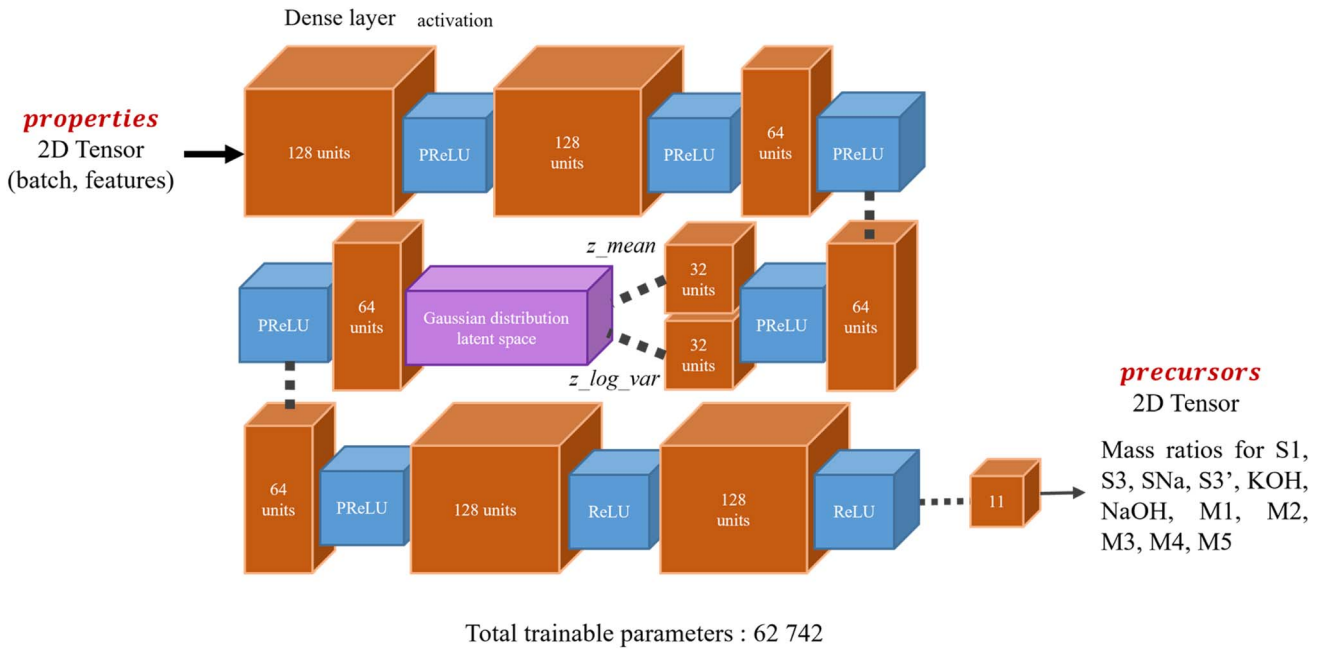

Fig. 7 The Formulator model architecture.

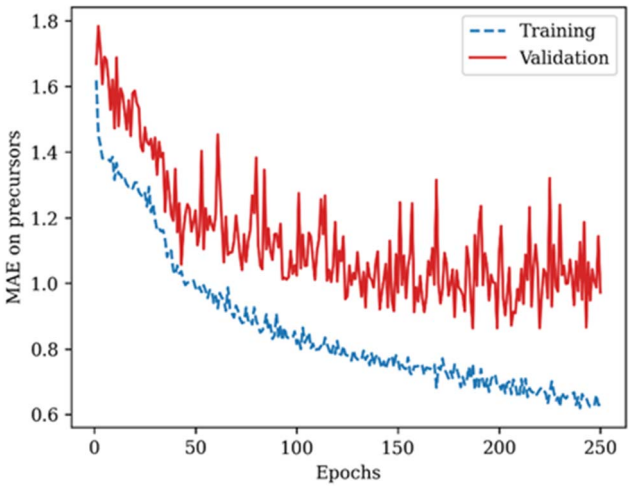

Fig. 8 The Formulator model's loss curves.

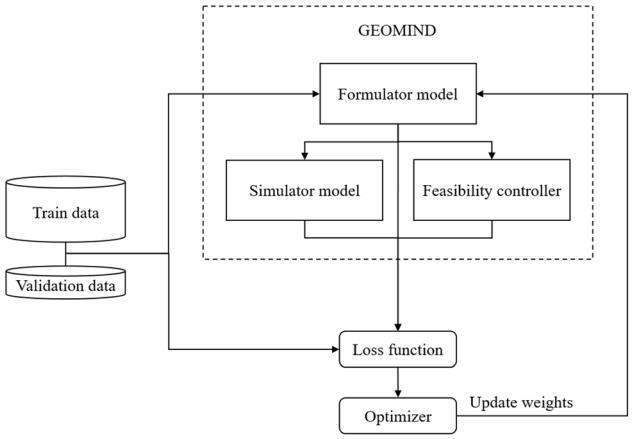

Fig. 9 GEOMIND framework.

calculation. The learning phase is unstable leading to unreliable and unfeasible results when synthesizing samples based on precursor mixtures predicted by this model result. When the samples were feasible, the error bar on the initial target properties passed as input to the model and their experimental counterpart was dramatically large. Such discrepancies are inherently related to the vast chemical space that the model tries to capture.

These results suggest that a novel and customized learning strategy should be developed in order to map the large precursor space and the 4-fold property space in an efficient manner that minimizes the outlies.

3.2.2.3 GEOMIND framework. We here take advantage of both the Simulator and the Formulator models and introduce the hybrid generative artificial intelligence "Geopolymer Engineering Optimization using Machine Intelligence for Novel Designs" termed hereafter GEOMIND, which aims to achieve the material design task. Specifically, within GEOMIND the Formulator's learning process is guided through the supervision of the Simulator model and the expert knowledge. The Formulator model doesn't fit itself, it is trained by GEOMIND in an analogous way, that is purely architectural, not mathematical or algorithmic, to how a generator model is trained by an antagonist model through the supervision of a discriminator model in a Generative Adversarial Network. However, unlike GANs, in our scheme there is no adversarial training, no minimax optimization, and no alternating updates between the two models. As such, GEOMIND takes advantage from the high accuracy of the Simulator model in predicting materials' properties and uses this feature to construct a customized training loss function that will be minimized during the training of the Formulator model. Here, the trainable weights of the Simulator are fixed while those of the Formulator model are optimized during the learning process. Fig. 9 illustrates the framework of GEOMIND training. GEOMIND is differentiated from prior inverse design schemes like conditional VAEs or bi-directional VAE[37,38] in several aspects. Indeed, we do not use any labels to steer generation from the latent space. Instead, we control precise numerical properties within a fine-grained representation space directly from the model input, rather than from the latent space. In addition, here we uniquely target the inverse design problem with the focus on a multi-property request, while most of the existing schemes focused on a single property.

The advantage of this particular architecture is that the Simulator ensures that the target properties are reachable in the latent space, it gives its own predicted properties from theoretical precursors formulated, themselves predicted from target properties. Both the Simulator and the Formulator share the same training dataset but in the case of an outsider data sample, each of these models will produce a different response. Through GEOMIND, convergence of target and predicted properties means that a presumed material is realistic. The Simulator has to be trained before the Formulator because its trainable weights have to be fixed within the framework shown in Fig. 9. The customized loss function L considers various MAE between properties, precursors and molar ratios to calculate the training cost as it is defined in the following equation:

$$L = w_1\mathrm{MAE}_{\mathrm{precursors}} + w_2\mathrm{MAE}_{\mathrm{sums}} + w_3\mathrm{MAE}_{\frac{\mathrm{Si}}{\mathrm{M}}\mathrm{sol}} + w_4\mathrm{MAE}_{\frac{\mathrm{Si}}{\mathrm{Al}}} + w_5\mathrm{MAE}_{\frac{\mathrm{Solid}}{\mathrm{Liquid}}} + w_6\mathrm{MAE}_{\mathrm{properties}} \quad (1)$$

where w is configurable loss weights for the optimizer and $\mathrm{MAE}_{\mathrm{sums}}$ refers to the absolute differences between precursor fractions and 1. This ensures that the Formulator model will learn to predict 100% of the mixture composition. These hyperparameters ($w1$–$w6$) were adjusted based on the relative importance of each term in the loss function through empirical testing (trial and error) rather than through an optimization procedure. The evolving nature of our database makes any fixed optimization less meaningful, as the optimal weights would likely shift with the addition of new data.

Molar ratios are computed by the Feasibility Controller module that checks the feasibility of a predicted mixture based on its chemical composition molar ratios, to guarantee that the precursors are chemically realistic and manufacturable in relation with geopolymerization. The main aim of embedding this expert knowledge within the learning scheme is to increase

the Formulator's performance and reliability of the formulation predictions.

Fig. 10 shows the training and the validation cross-validation mean loss curves of the Formulator model when it is trained to minimize the precursor's MAE within GEOMIND or by itself. We find that the formulator performance and stability within GEOMIND's framework are considerably increased, thereby highlighting the benefits of the suggested methodology for the design of new geopolymer materials. The final loss values of the Formulator model with both training methods are provided in Table S3 in the supplementary file. Within the GEOMIND scheme, the Simulator helps mitigate this issue by recognizing reliable (compatible) precursor choices for the target properties, even if they don't perfectly match the training sample, thereby smoothing out the high training cost and refining the predictions. This approach improves precursor discrimination and property targeting, as clearly demonstrated by the stable performance.

After achieving a well converged model, we now focus on its performance on predicting the properties and the precursors. For the sake of comparison, we vary either the formulator architecture (VAE *vs.* MLP) or the latent space form (Gaussian *vs.* Student's t-distribution). Fig. 11 shows the precursor and property validation loss curves with GEOMIND obtained as the mean of their respective k-fold cross-validation curves, and Table S4 (SI) summarizes the final loss values of the different Formulator models. Property loss represents the loss calculated between target properties and the Simulator's predicted properties within the GEOMIND framework. Overall, the three models show similar performance with VAE generative models being slightly more robust. This trend is expected to be more pronounced in the case of larger datasets.

3.3. Model performance and material design

Once trained and validated, we now focus on the model performances against experimental data to ensure that the development of new materials moves beyond the theoretical phase, paving the way for practical applications in industry and research.

3.3.1 The performance of the simulator model. The Simulator model is evaluated by comparing its predicted properties to experimental data. Furthermore, we also confront the

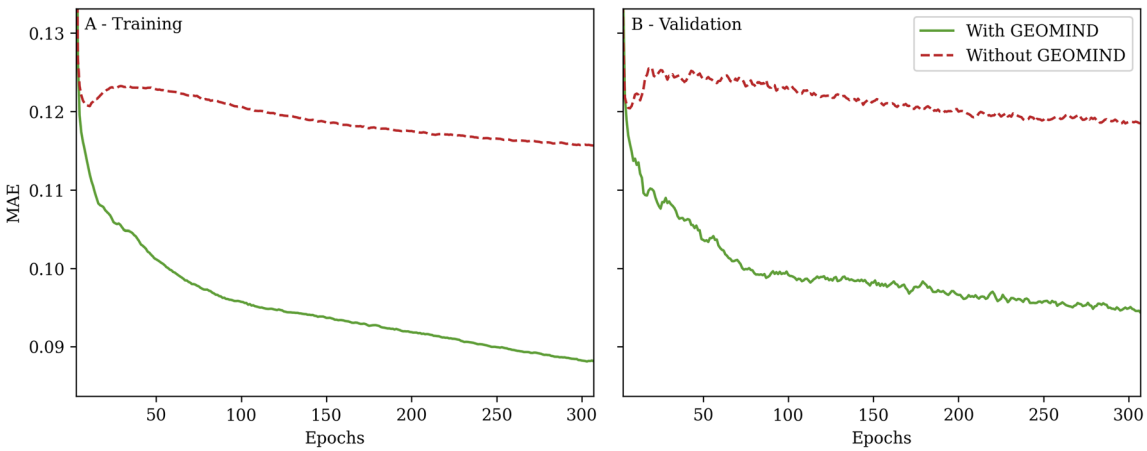

Fig. 10 Loss curves (moving average) of the Formulator model trained either with or without GEOMIND.

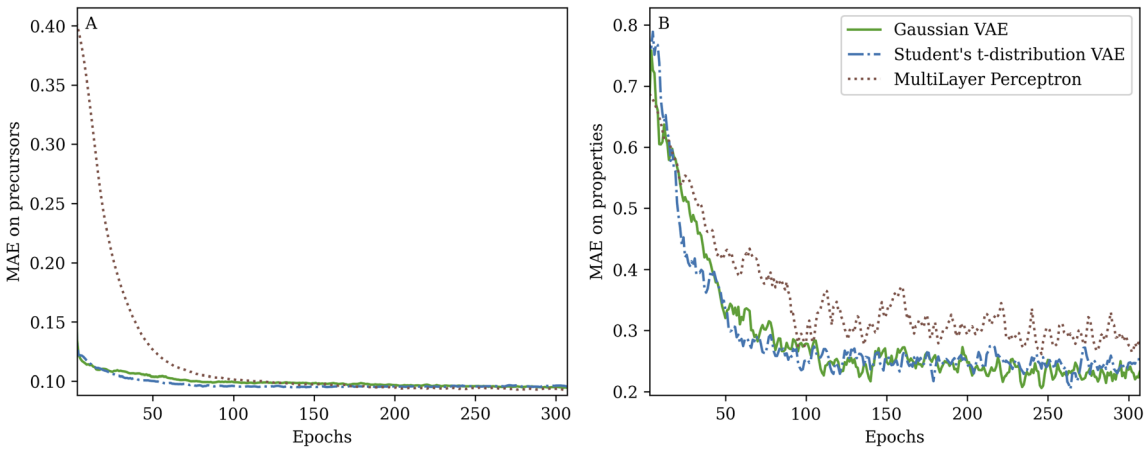

Fig. 11 Loss curves of different models of the Formulator in the GEOMIND framework.

Simulator model to the widely used XGBR as implemented in the xgboost package and trained on the same database. Table 3 shows the MAE and normalized MAE (nMAE) between predicted and experimental properties for both models. The normalized MAE is obtained by dividing the MAE by the difference between the maximum and minimum values of the property in the dataset. This normalization allows comparison across properties with different ranges.[39,40] We find that the Simulator model, which we recall is based on a VAE, features errors almost twice as low as the XGBR model indicating its higher accuracy. Nevertheless, one can notice that the normalized MAEs are relatively high, with some errors exceeding 10%. This result is not surprising as it strongly depends on the considered property. For instance, when examining the viscosity, the normalized MAE is very small, less than 2%, however it translates into a high MAE of around 24.5 Pa s. When looking at the other properties, their MAEs are in the order of the experimental error bar which consolidates, once more, the accuracy of the Simulator model in predicting the materials' properties from a set of precursor mixture.

Fig. S1 and S2 in the supplementary file show a direct comparison between the Simulator's predictions on all feasible samples in the database and the reference experimental data. As explained in Section 3.2.2., all values presented come from a validation set, which means that none of these values were learned during the Simulator training phase. As expected, we find a linear correlation between the various predicted and experimental predictions. The data dispersion around the diagonal line reflects the errors reported in Table 3. We note that although accurate, few outliers are present and lead to a large deviation of the predicted properties from the experimental results which affects the MAE. Overall, the Simulator model, designed as a VAE and embedding expert knowledge of the geopolymerization domain provides an accurate framework to predict materials' properties from input chemical mixtures while relying on a moderate size database. The accuracy of the model is naturally expected to increase by increasing the size of the database.

3.3.2 Material design with GEOMIND and experimental validation. In practice, GEOMIND is used as follows. The user formulates a set of target properties that the final material should have. The formulator model suggests the set of precursors that need to be mixed to achieve the target properties. At this stage, the user might ask for a set of properties that are divergent. Therefore, the predicted formulation is entered to the simulator that will predict the actual properties that correspond to the predicted material. If the user request is reasonable and corresponds to a manufacturable geopolymer, the properties predicted by the simulator will be very close to the target user request. Otherwise, it will suggest the actual correct properties that the user should expect. In this way, GEOMIND achieves a striking balance between predicting feasible formulations that correspond to real properties and that are as close as possible to the user request (target properties). The final predicted properties are the ones that should be compared directly to experimental measurements on the sample made following the predicted formulation.

To evaluate the capacities of GEOMIND, we use it to predict fifteen new compositions that were then synthesized experimentally and their properties measured and compared to the predictions (Table S5 of the SI and Table 4). To further quantify the uncertainty of GEOMIND predictions, we used nonparametric bootstrap resampling with 5000 iterations to determine the confidence intervals (CIs) for each property. In each iteration, a new dataset of the same size as the original was created by randomly sampling GEOMIND-prediction/experimental–observation pairs from the original data with replacement, so that some pairs could be selected more than once while others might not be selected in a given iteration. The MAE was recalculated for each resampled dataset, and the resulting bootstrap distribution was used to derive a 95% confidence interval from the 2.5th and 97.5th percentiles. The results are summarized in Table 4, and the bootstrap distributions are shown in Fig. S2 of the SI. It can be seen that the overall normalized MAE of the final predicted properties is less than 10% compared to the experimental measurements. For compressive strength, the MAE of 7.6 MPa is flanked by a 95% CI of [4.73, 10.8] MPa, indicating a consistent and reliable predictive performance despite the inherent variability in experimental measurements. The initial viscosity exhibits a wider 95% CI of [1.08, 22.2] Pa s around its MAE of 8.44 Pa s, reflecting greater uncertainty due to the broader range and higher variability of viscosity values in the dataset. However, this uncertainty remains within an acceptable range for practical applications. In contrast, mixture density and material density show exceptionally tight 95% CIs ([0.04, 0.06] g cm^{-3} and [0.01, 0.03] g cm^{-3}, respectively) around their low MAE values (0.05 g cm^{-3} and 0.02 g cm^{-3}). This highlights the high precision of GEOMIND in predicting these properties, as evidenced by the minimal spread in the bootstrap distributions (see Fig. S2). Taken together, these results demonstrate that while achieving low predictive errors, GEOMIND provides statistically robust estimates of uncertainty,

Table 3 MAE and normalized MAE between predicted and experimental properties for the database (112 samples)

Model	Error	Initial viscosity (Pa s)	Mixture density (g cm^{-3})	Compressive strength (MPa)	Material density (g cm^{-3})
Simulator	MAE	24.5	0.09	7.6	0.12
	nMAE (%)	1.9	14	9.4	15.5
XGBR	MAE	44.3	0.17	13.3	0.28
	nMAE (%)	3.4	36	16.5	37

Table 4 MAE and normalized MAE between experimental, predicted and target properties on fifteen test samples. 95% confidence interval (CI) of the predicted properties' MAE is also provided

	Experimental *vs.* target properties		Experimental *vs.* predicted properties	
Property	MAE	nMAE (%)	MAE [95% CI]	nMAE (%)
Initial viscosity (Pa s)	14.09	1.07	8.44 [1.08, 22.2]	0.64
Mixture density (g cm^{-3})	0.17	26.87	0.05 [0.04, 0.06]	8.37
Compressive strength (MPa)	11.40	14.07	7.60 [4.73, 10.8]	9.39
Material density (g cm^{-3})	0.17	21.94	0.02 [0.01, 0.03]	3.40

further validating its reliability for guiding geopolymer formulation design.

Among the tested materials, we consider three various use-cases, encompassing a material with conflicting target properties, a material with a random set of properties and a material with properties outside the knowledge domain. We note that none of these materials' target properties are present in the training database. The results of these tests are presented in Fig. 12 and compared to experiments. The colormap on this figure was obtained by a triangulation of the experimentally available values and the application of a Gaussian smoothing filter. It allows visualizing on the same graph the three main properties of the consolidated geopolymer material.

For sample 1, the user request is to obtain a final material with low density (1.6 g cm^{-3}) and viscosity (1.62 Pa s) and high compressive strength (82 MPa). We find that while it is located in the bottom left part of the map, it's dark color that denotes a high compressive strength does not fit with the color of the surroundings. This reflects a non-consistent user request with conflicting target properties. After prediction of the corresponding formulation, it is passed to the simulator to cross-check its actual predicted properties. Interestingly, GEOMIND predicts properties that essentially correct the compressive strength to be lower than the user request (51 MPa) while keeping the density and the viscosity within the same range. The experimental realization of these samples demonstrates that GEOMIND's predictions are accurate.

As for sample 2, we impose a user request of a material featuring properties outside the knowledge domain of the model. In this case, GEOMIND keeps both the compressive strength and the density within the same range while drastically reducing the material's viscosity to achieve manufacturable geopolymer compounds. Finally, the prediction of a sample with random properties shows that even in this case GEOMIND is able to predict formulation/properties that turn out to be very accurate compared to the experimental realization. The trade-off between the three user requested properties can be explained by the custom loss function that penalizes predictions leading to non-manufacturable materials. As a result, if a set of target properties is unrealistic, GEOMIND automatically adjusts its prediction to the closest achievable alternative within the geopolymer feasible space. For example, when a user requests samples with ultra-low viscosity and ultra-high strength, GEOMIND does not arbitrarily prioritize one property over another, but simply identifies the Pareto-optimal solution which represents the best possible compromise where no property can be improved without worsening another property. This is achieved through the loss function

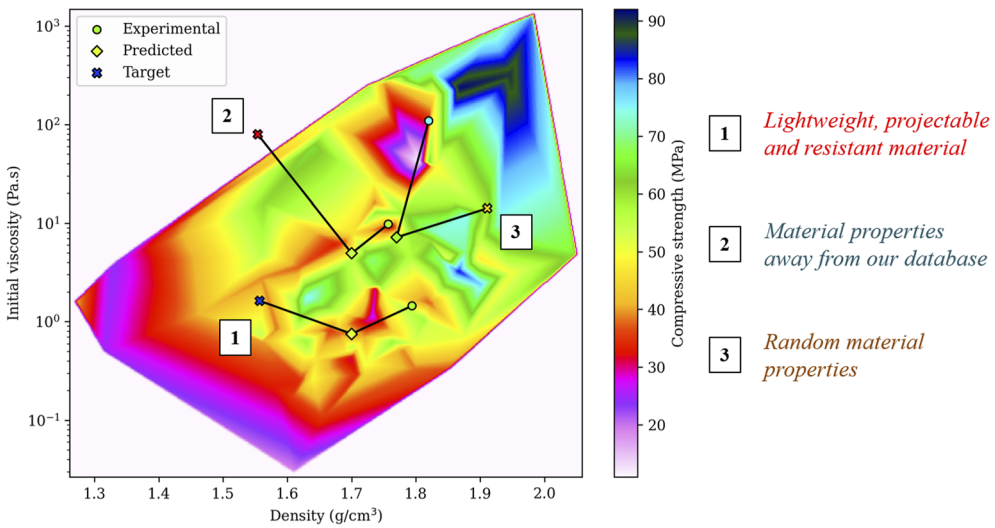

Fig. 12 (×) Target, (◆) predicted and (●) experimental properties of three example samples represented on database property heatmap showing the evolution of the initial viscosity and the compressive strength as a function of the material density.

minimization. While the exact mathematical formulation of this correction isn't explicitly derived, the logic is embedded in how the model optimizes its loss function.

In order to assess the added value of the expert knowledge, a set of nine samples, formulated using the GEOMIND framework without integrating the feasibility controller block (Fig. 9), was synthesized. The results have shown that only five out of the nine samples were feasible corresponding to a success rate of 55.6%. This issue was not encountered previously when GEOMIND integrated the feasibility controller block, which confirms the limitation of GEOMIND when operating without the expert knowledge. This outcome highlights that the absence of geopolymer feasibility domain guidance and chemical constraints affect the model's ability to consistently suggest feasible geopolymer formulations. Furthermore, the properties of the five feasible samples were measured and compared to the target and predicted properties (Table S6 of the SI) and their MAE and nMAE were determined (Table 5). Although the nMAE on the viscosity value remains low, we observed an increase of the nMAE of the compressive strength and the material density up to 15% which further underscores the importance of the expert knowledge in the architecture of GEOMIND to achieve lower nMAE (see Table 4).

For the sake of comparison, we also use Bayesian optimization to identify compositions that match the target properties on the Simulator model and confront it to GEOMIND. The methodology and the obtained results are further detailed in the SI. The final predicted properties compared to those measured on the sample corresponding to the predicted mixture show good accuracy with GEOMIND outperforming the BO method. It should be noted that the comparison with Bayesian optimization in this study is purely qualitative, focusing solely on the accuracy of predicted material properties relative to experimental and target values. While this approach effectively assesses the predictive fidelity of the model, it does not address quantitative metrics such as sample efficiency, runtime, or design success rate under matched constraints.

Overall, these tests demonstrate the problem-solving capabilities of the GEOMIND methodology in finding a compromise between a realistic and manufacturable geopolymer and an idealistic request of target properties.

Table 5 MAE and normalized MAE between experimental, predicted and target properties of the five feasible test samples proposed by GEOMIND without the feasibility controller block

	Experimental *vs.* target properties		Experimental *vs.* predicted properties	
Property	MAE	nMAE (%)	MAE	nMAE (%)
Initial viscosity (Pa s)	17.32	1.32	9.43	0.72
Compressive strength (MPa)	24.20	29.88	12.00	14.81
Material density (g cm^{-3})	0.21	26.41	0.12	15.64

4. Conclusion

This study demonstrates a powerful use case of artificial intelligence for the efficient design of geopolymer materials with tailored properties. Unlike the majority of existent literature on geopolymer design, we here build an in-house high-quality database of 112 samples. This database includes balanced data representation across precursor compositions and resulting material properties. This key development ensures the consistency of the developed machine learning approach and provides a solid base for further improvements. The developed machine learning framework is based on variational autoencoder modules that were thoroughly trained. The final model performances were validated against experimental data. The key findings of this work are:

- The possibility of accurately and simultaneously predicting, not one, but four key properties of geopolymers in both the fresh (viscosity and density of the mixture) and consolidated states (density and compressive strength), aligning well with experimental results with an overall normalized MAE of less than 10%.
- The development of GEOMIND, a new hybrid framework machine learning architecture based on two VAE models and one module encoding expert-knowledge. Specifically, the first model generates the formulations from target properties, and the second model predicts their corresponding properties in order to better align with experimental reality. The model was guided by an expert knowledge module that includes chemical constraints (silica/alumina, silica/alkali, solid/liquid) to ensure that all generated formulations are feasible, leading to more reliable geopolymer design. The training of the model was performed using a customized loss function that considers precursor composition errors and property deviations demonstrating high stability and good convergence.

GEOMIND was used to predict formulations based on an array of target properties input by the user. The results demonstrate an outstanding performance where the model is able to properly guide the design by striking a balance between the various requested properties achieving an MAE of 12.2 Pa s, 0.063 g cm^{-3}, 7.2 MPa and 0.024 g cm^{-3} for the initial viscosity, mixture density, compressive strength and final density, respectively. In addition, the developed model is able to extrapolate to unseen formulations and target properties, as confirmed by experimental validation on fifteen diverse samples. It is important to note that this validation set is not exhaustive. Future studies should include a larger and more diverse validation set to fully assess the generalizability of the model across the broader geopolymer design space.

Consequently, the developed model is a key accelerator for an efficient design of geopolymer materials which can boost the industrial adoption of this new class of materials. Undoubtably, the accuracy of the model can be further enhanced by increasing the size of the database. This can be achieved by implementing an active learning loop where the model is used to predict new compounds that will be manufactured, while refining the model on the fly to broaden its scope and enhance

its fidelity. Subsequently, the hyperparameter values could be optimized using a suitable optimization method such as the Bayesian optimization. Future work will also explore uncertainty quantification techniques, such as Monte Carlo dropout, ensemble methods, or Bayesian neural networks, to provide confidence intervals for predictions. Although a quantitative comparison with the Bayesian optimization under matched constraints was beyond the scope of this study, future work should evaluate comparative sample efficiency, runtime, and design rate to provide a better assessment of relative performance. Ultimately, GEOMIND can also be applied to other classes of materials including alkali-activated materials and cements, and be generalized to the design of other precursor-based materials with the necessity to adapt the raw-material classes and molar-ratio constraints to reflect the different chemistry of these materials. Furthermore, the flexibility of GEOMIND allows for integration of new classes by updating the training dataset, without requiring architectural changes. However, performance on entirely unseen precursors would depend on their similarity to the existing classes and the quality of the new training data.

Author contributions

A. Gharzouni and A. Bouzid conceived the research idea and supervised the project. A. Gharzouni and S. Rossignol performed the data collection, database construction, and experimental validation. S. Rousseau developed the VAE frameworks with supervision by A. Bouzid. S. Rousseau, A. Bouzid and A. Gharzouni contributed to the discussions and jointly wrote and reviewed the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

Data availability

The full experimental dataset (112 samples) is confidential due to patented compositions but is available upon reasonable request to the corresponding author (ameni.gharzouni@unilim.fr). A representative database subset (10 samples) is available on GitHub (https://github.com/Geopolymer-AI/GEOMIND) as well as codes used to prepare, analyze and plot the data, train, test and use the GEOMIND model. An archived version of the associated repository is available *via* Zenodo at https://doi.org/10.5281/zenodo.20286234.

Supplementary information is available. See DOI: https://doi.org/10.1039/d5dd00383k.

Acknowledgements

This work was totally funded by the LabEx Sigma-lim ANR-10-LBX-0074, laboratory of excellence launched by the French Ministry of Higher Education and Research (https://www.unilim.fr/labex-sigma-lim) between IRCER (https://www.ircer.fr) and XLIM (https://www.xlim.fr) Research Institutes. The authors gratefully acknowledge Romain Négrier for insightful discussions.

References

1 P. Duxson, A. Fernández-Jiménez, J. L. Provis, G. C. Lukey, A. Palomo and J. S. van Deventer, *J. Mater. Sci.*, 2007, **42**, 2917–2933.
2 Q. Cligny, E. Hyvernaud, A. Gharzouni, D. Brandt and S. Rossignol, *Eng. Rep.*, 2024, **6**, 12.
3 J. L. Provis and J. S. J. van Deventer, *Geopolymers: Structure, processing, properties and industrial applications*, 2009.
4 C. Shi, A. F. Jiménez and A. Palomo, *Cement Concr. Res.*, 2011, **41**, 750–763.
5 L. Ouamara, A. Gharzouni, B. Naït-Ali, J. Jouin, G. Babule, P. Duport, C. Chinaya, E. Guillaume and S. Rossignol, *Open Ceram.*, 2023, **16**, 100462.
6 I. N. Vlasceanu, A. Gharzouni, O. Tantot, M. Lalande, C. Elissalde and S. Rossignol, *Open Ceram.*, 2020, **2**, 100013.
7 A. Gharzouni, E. Joussein, B. Samet, S. Baklouti and S. Rossignol, *J. Non-Cryst. Solids*, 2015, **410**, 127–134.
8 D. Khale and R. Chaudhary, *J. Mater. Sci.*, 2007, **42**, 729–746.
9 C. H. Chan, M. Sun and B. Huang, *EcoMat*, 2022, **4**, e12194.
10 B. Chen and W. Liu, *Comput. Model. Eng. Sci.*, 2024, **141**, 57705.
11 W. Huo, Z. Zhu, H. Sun, B. Ma and L. Yang, *J. Clean. Prod.*, 2022, **380**, 135159.
12 J. Shen, *et al.*, *Constr. Build. Mater.*, 2022, **360**, 129600.
13 P. Gupta, N. Gupta and K. K. Saxena, Innov. Emerg, *Technol.*, 2023, **10**, 2350003.
14 M. Rathnayaka, D. Karunasinghe, C. Gunasekara, K. Wijesundara, W. Lokuge and D. W. Law, *Constr. Build. Mater.*, 2024, **419**, 135519.
15 M. I. Khan and Y. M. Abbas, *Mater. Today Commun.*, 2023, **35**, 105793.
16 M. Verma, *Asian J. Civ. Eng.*, 2023, **24**, 2659–2668.
17 H. Tanyildizi, *Environ. Sci. Pollut. Res.*, 2024, **31**, 41246–41266.
18 H. A. Al-Jamimi, W. A. Al-Kutti, S. Alwahaishi and K. S. Alotaibi, *Case Stud. Constr. Mater.*, 2022, **17**, e01238.
19 A. Ahmad, W. Ahmad, F. Aslam and P. Joyklad, *Case Stud. Constr. Mater.*, 2022, **16**, e00840.
20 S. Chen, H. Cao, Q. Ouyang, X. Wu and Q. Qian, *Mater. Des.*, 2022, **223**, 111092.
21 N. J. Szymanski, P. Nevatia, C. J. Bartel, Y. Zeng and G. Ceder, *Nat. Commun.*, 2023, **14**, 6956.
22 T. He, H. Huo, C. J. Bartel, Z. Wang, K. Cruse and G. Ceder, *Sci. Adv.*, 2023, **9**, eadg8180.
23 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2018, **4**, 268–276.
24 T. Xie and J. C. Grossman, *Phys. Rev. Lett.*, 2018, **120**, 145301.
25 W. Zheng, Z. Shui, Z. Xu, X. Gao and S. Zhang, *J. Build. Eng.*, 2023, **76**, 107396.

26 H. A. T. Nguyen, D. H. Pham and Y. Ahn, *Appl. Sci.*, 2024, **14**, 3601.
27 K. M. Jablonka, G. M. Jothiappan, S. Wang, B. Smit and B. Yoo, *Nat. Commun.*, 2021, **12**, 2312.
28 R. Xin, *et al.*, *J. Phys. Chem. C*, 2021, **125**, 16118–16128.
29 J. Moon, *et al.*, *Nat. Mater.*, 2024, **23**, 108–115.
30 F. Rosenblatt, *Proc. IRE*, 1960, **48**, 301–309.
31 K. Hornik, M. Stinchcombe and H. White, *Neural Netw.*, 1989, **2**, 359–366.
32 L. Pinheiro Cinelli, M. A. Marins, E. A. Barros Da Silva and S. Lima, *Netto*, 2021, 111–149.
33 J. H. Friedman, *Ann. Stat.*, 2001, **29**, 1189–1232.
34 X. X. Gao, A. Autef, E. Prud'homme, P. Michaud, S. Basma, E. Joussein and S. Rossignol, *J. Sol-Gel Sci. Technol.*, 2013, **65**, 220–221.
35 A. Gharzouni, I. Sobrados, E. Joussein, S. Baklouti and S. Rossignol, *Colloids Surf. A Physicochem. Eng. Asp.*, 2016, **511**, 212–221.
36 N. B. Ipsen, P.-A. Mattei and J. Frellsen, How to deal with missing data in supervised deep learning?, *In Proceedings of 2022 International Conference on Learning Representations*, 2022.
37 Q. Liu, P. Cai, D. Abueidda, S. Vyas, S. Koric, R. Gomez-Bombarelli and P. Geubelle, *Comput. Methods Appl. Mech. Eng.*, 2025, **438**, 117848.
38 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2018, **4**, 268–276.
39 A. Fentis, M. Rafik, L. Bahatti, O. Bouattane and M. Mestari, *Energy Rep.*, 2022, **8**, 3221–3233.
40 G. Müller-Plath and H.-J. Lüdecke, *Res. Stat.*, 2024, **2**, 2317172.