Towards data-driven constitutive modelling for granular materials via micromechanics-informed deep learning

Tongming Qu[a], Shaocheng Di[b], Y.T. Feng[a,*], Min Wang[c], Tingting Zhao[d]

[a] *Zienkiewicz Centre for Computational Engineering, College of Engineering, Swansea University, Swansea, Wales, SA1 8EP, UK*

[b] *College of Shipbuilding Engineering, Harbin Engineering University, Harbin, 150001, China*

[c] *Fluid Dynamics and Solid Mechanics Group, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*

[d] *Institute of applied mechanics and biomedical engineering, Taiyuan University of Technology, Taiyuan, Shanxi, 030024, China*

Abstract

The analytical description of path-dependent elastic-plastic responses of a granular system is highly complicated because of continuously evolving microstructures and strain localisation within the system undergoing deformation. This study offers an alternative to the current analytical paradigm by developing micromechanics-informed machine-learning based constitutive modelling approaches for granular materials. A set of critical variables associated with the constitutive behaviour of granular materials are identified through an incremental stress-strain relationship analysis. Depending on the strategy to exploit the priori micromechanical knowledge, three different training strategies are explored. The first model uses only the measurable external variables to make stress predictions; the second model utilises a directed graph to link all the external strain sequences and internal microstructural evolution variables into a single prediction model comprised of a series of sub-mappings, and the third model explicitly integrates the physically important non-temporal properties with external strain paths into training through an enhanced Gated Recurrent Unit (GRU). These three models show satisfactory agreement with unseen test specimens based on multi-directional loading cases. The features and applications of each model are explained. Furthermore, the key factors for constitutive training, potential applications and deficiencies of the current work are also discussed in detail.

Keywords: Deep learning, Data-driven, Elastic-plastic constitutive model, Gated Recurrent Unit (GRU), Granular materials, Micromechanics, Discrete element modelling

*Corresponding author: y.feng@swansea.ac.uk

1. Introduction

Constitutive behaviour of materials is one of the most intensely researched fields in engineering science owing to its complexity and importance in engineering practice. From a macroscopic perspective, the elastic-plastic response of granular materials highly depends on the path of deformation. Its stress-strain behaviour exhibits anisotropy (Nemat-Nasser and Zhang, 2002; Zhu et al., 2006; Yang et al., 2008; Anandarajah, 2008; Chang and Yin, 2010), distortional hardening (Voyiadjis et al., 1995), viscoplasticity (Di Prisco et al., 2002), and strain localisation features (Anand and Gu, 2000; Voyiadjis et al., 2005; Hashiguchi and Tsutsumi, 2007; Qu et al., 2019a). From a microscopic perspective, discrete grains in granular materials transfer forces via inter-particle contacts, and highly inhomogeneous and discontinuous force networks are developed inside the material to balance the external loads. Kuhn and Daouadji (Kuhn and Daouadji, 2018a,b) pointed that many basic principles that we take for granted in conventional elasto-plasticity are not consistent with meticulous particle-scale numerical observations. Although much effort has been made to phenomenological constitutive models of granular materials (Zhu et al., 2010; Lai et al., 2016; Sun et al., 2018; He et al., 2019; Yang et al., 2020; Zhang et al., 2021), it is still a great challenge to develop a unified theoretical constitutive model due to complex microstructural evolution within granular materials (Antony and Kuhn, 2004; Nguyen et al., 2016; Qu et al., 2019b).

The constitutive behaviour is a time sequence problem, in essence. Modelling elastic-plastic constitutive relations via a deep neural network (DNN) suitable for time series prediction is a potential scheme to address the above challenge. As a data-driven method, DNN is a hypothesis function representing the relationship between input and output data by sequentially using a series of linear matrix multiplication and nonlinear mapping with activation functions.

The idea of using artificial neural networks (ANNs) to represent the constitutive behaviour of granular materials (e.g. sand) has a long history (Ellis et al., 1995; Ghaboussi and Sidarta, 1998; Shin and Pande, 2000; Javadi et al., 2003; Hashash et al., 2003, 2004; Banimahd et al., 2005; Jung and Ghaboussi, 2006; Hashash et al., 2006). Owing to the recent development of computational science and a deeper understanding of the data-driven research paradigm, developing a more reliable DNN model with fewer data samples becomes possible. Again, the application of deep learning in characterising material behaviour has been receiving increasing attention (Jenab et al., 2016; Liu and Wu, 2019; Pandya et al., 2020). Specifically, fully connected DNNs have been used to represent temperature- and rate-dependent plasticity models (Li et al., 2019) and von Mises plasticity with isotropic hardening (Zhang and Mohr, 2020).

The recurrent neural networks (RNNs) have been applied to train various constitutive laws (Ali et al., 2019; Settgast et al., 2020; Gorji et al., 2020; Karapiperis et al., 2021), e.g. plasticity of composite materials (Mozaffar et al., 2019; Wu et al., 2020), the stress-strain behaviour of aluminium (Fernández et al., 2020), and polypropylene (Jordan et al., 2020). Abueidda et al. (Abueidda et al., 2021) compared several popular sequence learning methods in application to path-dependent plasticity and thermo-viscoplasticity and found that both GRU and TCN (temporal convolutional network) are able to accurately predict the history-dependent materials but TCN has a greater computational efficiency on GPUs than GRU. In addition, Wang et al. (Wang and Sun, 2019; Wang et al., 2019, 2020) pioneer the application of reinforcement learning and adversarial learning for the traction-separation law of interfaces and constitutive behaviour of granular materials.

Although the data science method has a unique strength in extracting rules from data, as demonstrated for granular matters in (Wang and Sun, 2019), the learning pattern from data tends to lack interpretability and can be spurious. On the other hand, the mechanical theory has a clear interpretable and rigorous logic, but may have to introduce some idealised approximations for complex problems. Thus the resulting constitutive model may suffer from low accurate predictions for real problems.

In this study, we attempt to develop a new research paradigm which takes advantage of the unique capability of data-driven methods in extracting patterns from empirical data, but utilises some prior knowledge acquired from theoretical analysis as guidance to investigate the constitutive behaviour of granular materials. The data representing the constitutive behaviour of granular materials is generated by the discrete element modelling of triaxial testing. An analytical stress-strain equation for granular materials serves as the prior knowledge to determine the key variables involving in data training. The recurrent neural network (RNN) used for time series prediction problems is adopted to model the path/history-dependent elastic-plastic constitutive models of granular matters.

The paper is structured as follows: Section 2 provides an incremental stress-strain relation for granular materials to guide deep learning. Section 3 introduces three different training approaches according to different strategies to exploit the priori micromechanical knowledge. One utilises only the measurable principal strain sequences as inputs (suitable for the experimental condition). The second one predicts stress responses via a directed graph based constitutive model constituted by some sub-networks, which link all the discovered internal and external variables associated with the constitutive behaviour of granular materials. The third approach adopts an enhanced GRU architecture incorporating non-temporal physical variables

for stress predictions. Section 4 introduces the data preparation and implementation details of deep learning models. The prediction results of several constitutive modelling approaches are demonstrated based on conventional and true triaxial loading conditions. Section 5 discusses the critical factors for training a reliable prediction model, potential applications of such a data-driven constitutive model and the deficiencies of the current study. Conclusions are drawn in Section 6. Appendix A offers a brief introduction to the enhanced GRU architecture incorporating physics-invariant quantities. Appendix B gives a detailed account of selecting some key hyperparameters for the several training approaches used.

2. An analytical stress-strain relation for granular materials

In this session, an analytical stress-strain relation for granular materials is used to discover the key factors behind the complex constitutive behaviour and these recognised factors will be incorporated in the training of deep learning models to be described in the next section. In this work, the constitutive analysis is investigated based on a cubic representative volume element (RVE) subjected to strain-dominated triaxial testing conditions shown in Figure 1. Other complex mechanical states will be explored in the future.

Normally, path-dependent elastic-plastic constitutive relations are formulated in an incremental form. A total stress-strain expression needs to be calculated with path integrals of an incremental constitutive relation. According to the principle of solid mechanics, the incremental elastic-plastic relation can be expressed as:

$$\Delta\sigma_{ij} = C_{ijmn}\Delta\varepsilon_{mn} \quad (i,j,m,n = 1,2,3) \tag{1}$$

where $\Delta\sigma_{ij}$ and $\Delta\varepsilon_{mn}$ are stress and strain increments, respectively; C_{ijmn} is a stiffness tensor; and 1, 2 and 3 denote x, y and z coordinates in the global space, respectively (see Figure 1).

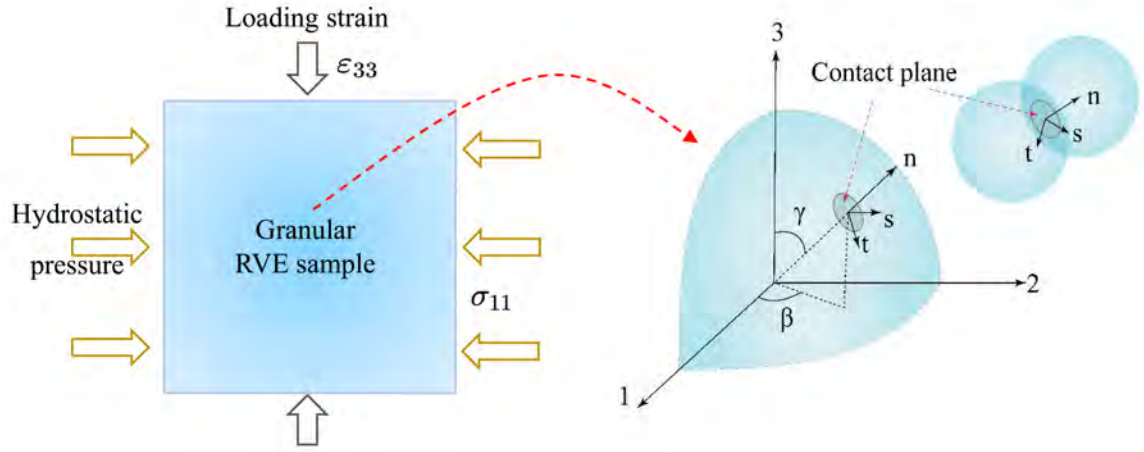


Figure 1: Illustration of an RVE sample and the coordinate systems for interparticle contacts

Assuming that the deformation of the granular assembly is statistically uniform in the space (Voigt's hypothesis), and following the principle of conservation of energy between the granular system and the corresponding continuum, we have derived the formulation of C_{ijmn} in our previous work (Qu et al., 2019b). For a given granular assembly, by assuming that the interactions between particles obey a linear contact model, C_{ijmn} can be expressed as:

$$C_{ijmn} = \frac{(k_n - k_s)}{V} \sum_{k=1}^{N_c} \left(L^k\right)^2 \alpha_i^k \alpha_j^k \alpha_m^k \alpha_n^k + \frac{k_s}{V} \sum_{k=1}^{N_c} \left(L^k\right)^2 \delta_{in} \alpha_j^k \alpha_m^k \tag{2}$$

where k_n and k_s are the particle-scale normal and shear contact stiffnesses, respectively; V is the volume of the granular assembly; N_c is the number of mechanical contacts; L^k is the distance of contact k (i.e. the distance of two contacting spherical centres); α_i^k is the i^{th} component of the direction vector of contact k (the same to α_j^k, α_m^k and α_n^k); δ_{in} is Kronecker's delta.

To characterise the axial stress and strain relationship of a representative volume element, the axial stress increment can be expanded from Eq.(1):

$$\Delta\sigma_{33} = C_{3311}\Delta\varepsilon_{11} + C_{3322}\Delta\varepsilon_{22} + C_{3333}\Delta\varepsilon_{33} \tag{3}$$

where $\Delta\varepsilon_{11}$ and $\Delta\varepsilon_{22}$ are the two lateral strain increments while $\Delta\varepsilon_{33}$ is the loading strain increment; and C_{3311} , C_{3322} and C_{3333} are the components of the equivalent stiffness tensor and can be calculated from Eq. (2) as follows:

$$C_{3311} = \frac{(k_n - k_s)}{V} \sum_{k=1}^{N_c} \left(L^k\right)^2 \alpha_3^k \alpha_3^k \alpha_1^k \alpha_1^k \tag{4}$$

$$C_{3322} = \frac{(k_n - k_s)}{V} \sum_{k=1}^{N_c} \left(L^k\right)^2 \alpha_3^k \alpha_3^k \alpha_2^k \alpha_2^k \tag{5}$$

$$C_{3333} = \frac{(k_n - k_s)}{V} \sum_{k=1}^{N_c} \left(L^k\right)^2 \alpha_3^k \alpha_3^k \alpha_3^k \alpha_3^k + \frac{k_s}{V} \sum_{k=1}^{N_c} \left(L^k\right)^2 \alpha_3^k \alpha_3^k \tag{6}$$

The material properties of particles are non-temporal quantities for a certain granular sample while the microstructural features are temporal quantities, which evolve gradually over the whole range of a deformation process. C_{ijmn} is combined with the non-temporal particle properties (contact stiffnesses) and the temporal microstructural fabric tensor. Therefore the elastic stiffness tensor C_{ijmn} is not constant and will evolve dynamically during external loading.

One microscopic origin responsible for complex constitutive laws of granular media is that the external deformation history or path tends to make internal grains move around each other permanently. This irreversible movement that arises in granular materials affects the subsequent deformation due to the changes in the inherent microstructures or local stiffness of granular assemblies. For general granular materials without considering grain breakage or material degradation, the evolution of elastic-plastic constitutive relations stems from the irreversible evolution of microstructures or fabric features of granular media. The above formulations are derived from the small-strain assumption and thus cannot directly describe the microstructural evolution process, but they are useful to understand the critical variables associated with stress-strain responses of granular materials. For a quasi-static loading condition, the analytical formulation holds true at every single moment while the stiffness tensor C_{ijmn} evolves dynamically during shearing.

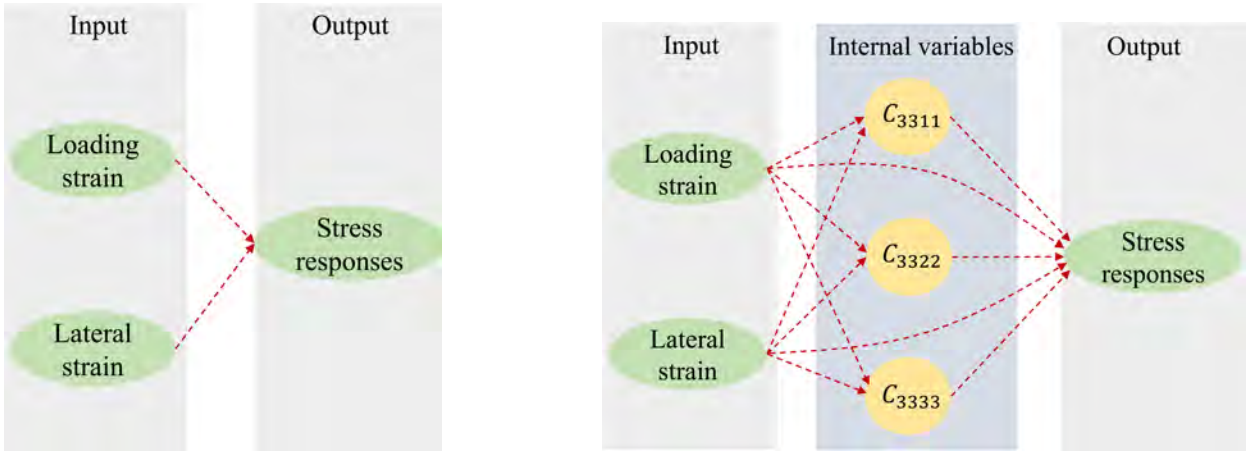
3. Constructing data-driven stress-strain relations with machine learning

3.1. Model A and model B: representing stress-strain relations via a directed graph connected with deep neural networks

The incremental stress-strain relations presented in Section 2 determine the primary variables for capturing the constitutive behaviour of granular materials. Particularly, Equation (3) reveals that the lateral strains $\Delta\varepsilon_{11}$ and $\Delta\varepsilon_{22}$, loading strain $\Delta\varepsilon_{33}$ and several components of the stiffness tensor, i.e. C_{3311}, C_{3322} and C_{3333}, are critical to characterising the stress-strain responses in the loading direction. Among all the related variables, the components of the elastic stiffness tensor C_{3311}, C_{3322} and C_{3333} are internal variables which cannot be observed directly during experimental testing. In contrast, the lateral strains are external variables which can be explicitly measured under experimental conditions.

In the triaxial mechanical condition, we use the deviatoric stress, which is the difference between the major and minor principal stresses, to reflect the evolution of loading stress. Depending on whether the internal variables are incorporated in constitutive modelling, two strategies are available based on the micromechanical formulation given in Section 2. The first strategy is to use all the measurable external variables only, i.e. both the loading strain and lateral strain, but abandon the internal variables to approximate the deviatoric stress. All the training variables are measurable and thus this model can be developed based on the experimental environments. The other strategy is somehow to introduce the micromechanical structural information into training.

A key conceptual step in developing the second training models in this study is utilising a directed graph to represent the complex stress-strain relations. The introduction of a directed graph enables to incorporate all the critical variables (both internal and external) and to construct a complete information flow from strain to stress. In graph theory, a directed graph is a graph that is made up of a set of vertices connected by edges. The vertices represent a series of physical variables while the edges denote certain connections amid these variables. The directed edges are drawn as arrows indicating the direction of information flow from source (or predecessor) nodes to target vertices. Some applications of a directed graph in computational mechanics can be found in (Sun et al., 2013; Sun, 2015; Wang and Sun, 2018). Following the rule of a directed graph, Model A, as shown in Figure 2a, is the training model developed based on the first strategy. Its corresponding input-output pair (NN-A) is: [ε_{11}, ε_{22}, ε_{33} $\rightarrow$ stress responses]. In contrast, Model B, based on the second strategy of training, is more complicated.



(a) Model A: with only external variables

(b) Model B: incorporating both external and internal variables

Figure 2: The directed graph representations of stress-strain relations for granular materials under conventional triaxial testing conditions

As shown in Figure 2b, model B starts with strain variables and ends with stress variables, with the internal variables C_{3311}, C_{3322} and C_{3333} being intermediate vertices. Each edge linking two vertices is represented by deep neural networks, which have proven to be capable of approximating any complex mappings (Cybenko, 1989; Hornik et al., 1989). All the sub-networks constitute a single prediction model linking the strains (inputs) and the stress (output) but involving microstructural variables. The basic idea behind such a directed graph is to construct a constitutive model linking strain to stress directly but also make full use of microstructural information at the same time. Besides, instead of analytically or statistically tracing the yield surface in the phenomenological framework (Shaverdi et al., 2013), we leverage the powerful prediction capability of deep neural networks to describe the complex evolution

of microstructures in granular materials undergoing deformation.

The whole directed graph is unfolded as follows: 1) identify the predecessor nodes of the terminal node (stress responses); 2) recognise all predecessor nodes of intermediate nodes by recursively going towards the source nodes from the target nodes (upstream) in the whole directed graph, until the final predecessor node is a start node (an input variable) only. Under such principles, two information flow paths in Figure 2b can be found: one is [loading and lateral strains $\rightarrow$ stress responses] and the other is [loading and lateral strains $\rightarrow$ C_{3311}, C_{3322} and $C_{3333} \rightarrow$ stress responses]. Both information flows jointly determine the final stress prediction. To form a whole stress-strain pair incorporating internal microstructural evolution in deep learning framework, these two information flows can be implemented by four sub-ANNs (input-output pairs):

(1) NN-B1: [ε_{11}, ε_{22}, $\varepsilon_{33} \rightarrow C_{3311}$]

(2) NN-B2: [ε_{11}, ε_{22}, $\varepsilon_{33} \rightarrow C_{3322}$]

(3) NN-B3: [ε_{11}, ε_{22}, $\varepsilon_{33} \rightarrow C_{3333}$]

(4) NN-B4: [ε_{11}, ε_{22}, ε_{33}, C_{3311}, C_{3322} and $C_{3333} \rightarrow$ stress responses].

Note that the sub-ANN, [ε_{11}, ε_{22}, $\varepsilon_{33} \rightarrow$ stress responses], is not treated as a separate sub-network for training in the current directed graph as it does not incorporate the evolution of microstructures in prediction. This strain-to-stress mapping can be regarded as a special type or part of the fourth sub-ANN above when C_{3311}, C_{3322} and C_{3333} have no contribution to the stress prediction. Instead of artificially determining the role of microstructural variables in reproducing the constitutive behaviour, the weights of C_{3311}, C_{3322} and C_{3333} in the current directed graph are automatically discovered by deep learning.

In the training phase, the microstructural states (C_{3311}, C_{3322} and C_{3333}) are obtained via discrete element modelling (DEM). All these sub-ANNs are trained by supervised learning with ground truth data. In the prediction phase, i.e. after these sub-networks have been well trained, one complete information flow can be constructed by enforcing the prediction outputs of the first three sub-networks to be partial inputs of the fourth sub-network. At that time, the microstructural states (C_{3311}, C_{3322} and C_{3333}) will no longer be required.

In consideration of the path-dependent features of constitutive behaviour, the recurrent neural networks (RNNs), which are special ANN architectures suitable for time-series prediction issues, are ideal candidates to train the sub-networks. Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Cho et al., 2014) are successful RNN architectures in processing long sequences, due to their satisfactory capability of mitigating exploding gradient or vanishing gradient issues. Both of them introduce a gate

mechanism to regulate what information from previous memory needs to be kept around and what previous data can be forgotten. Although existing literature shows that both LSTM and GRU have close prediction performance (Chung et al., 2014), GRU requires fewer trainable parameters and has a relatively higher training efficiency. Thus the GRU architecture is used to train all the sub-networks in this study.

3.2. Model C: integrating physics-invariant properties with external strain paths via an enhanced GRU architecture

Apart from the scheme of exploiting the internal variables through a directed graph, another strategy is to integrate only the physics-invariant properties associated with elastic-plastic behaviour into DNN training. Equation 2 has demonstrated that the elastic stiffness tensor C_{ijmn} is made of non-temporal contact stiffnesses and temporal microstructural tensor in the volume of a specimen. The contact stiffnesses are certainly critical properties governing stress-strain responses of granular materials. For non-cohesive granular materials, the frictional strength between two grains is an important microscopic origin for macroscopic strength. Thus the friction coefficient of particles and the confining pressure are also critical ingredients for the constitutive behaviour. The contact stiffnesses and frictional coefficients of particles usually keep non-temporal when subjected to external loadings, provided that the particle breakage and material degradation do not happen. Under specific loading conditions, such as conventional triaxial testing environment, the confining stress usually remains constant as well. A major challenge for taking ANN models into more realistic problems (e.g. constitutive modelling) is how to incorporate these non-temporal physical properties during training.

In this work, we adopt an enhanced GRU architecture reported in (Mozaffar et al., 2019) to address this issue. A detailed introduction about the architecture and mathematics behind the enhanced GRU can be found in Appendix A. The additional non-temporal features (particle stiffnesses and frictional coefficient) will be used as extra input variables, together with the temporal principal strain sequences to predict the final stress responses via the enhanced GRU architecture. This model is named as "Model C" in the following text.

3.3. Accuracy evaluation of trained prediction models

The accuracy of the prediction models is evaluated by quantifying the overall discrepancy between the actual values and the predicted values. In this work, we adopt two metrics to evaluate the prediction accuracy. One is a score metric, which can give a straightforward but not very rigorous understanding about prediction capability, and the other is the SMAE (scaled

mean absolute error), which is commonly used as the cost function when training a DNN model. Prior to formulating the score metric, the scaled squared error (SSE) for every single point i in the j^{th} stress-strain sample should be calculated:

$$SSE_{ij} = (\bar{y}_{ij}^{True} - \bar{y}_{ij}^{Prediction})^2 \tag{7}$$

where $\bar{y}_{ij}^{True}$ and $\bar{y}_{ij}^{Prediction}$ are the scaled actual and prediction values of the i^{th} point in the j^{th} stress-strain sample, respectively.

After obtaining all the SSE values on the j^{th} stress-strain prediction curve, an empirical cumulative distribution function (eCDF) F_j can be computed as follows:

$$F_j(SSE_{rj}) = \frac{r}{N^j} \quad \left(r = 1, ..., N^j\right) \tag{8}$$

where N^j is the number of data points on the j^{th} stress-strain curve; and all SSE_{ij} are arranged in ascending order. Following the scheme given in Wang and Sun (2019), the following accuracy score is adopted based on the above eCDF:

$$A_{score} = max\left(\frac{log\left[max(\varepsilon_{P\%}, \varepsilon_{crit})\right]}{log(\varepsilon_{crit})}, 0\right) \tag{9}$$

where $\varepsilon_{P\%}$ is the SSE value corresponding to $P\%$ in the eCDF, and it is used as a representative to evaluate the score of predications; ε_{crit} is the critical SSE which can be regarded as "satisfactorily accurate" when $\varepsilon_{P\%} \leq \varepsilon_{crit}$. Normally $\varepsilon_{crit} \ll 1$. In this work we assume $P\%$=90% and ε_{crit}=0.001.

The other metric, i.e. SMAE, is defined as follows:

$$SMAE_j = \frac{1}{N^j}\sum_{i=1}^{N^j}\left|\bar{y}_{ij}^{True} - \bar{y}_{ij}^{Prediction}\right| \tag{10}$$

where $SMAE_j$ is the scaled mean absolute error of the j^{th} stress-strain curve.

Once a DNN model has been trained, one can give predictions over all the training/validation/test data specimens. In this case, a more comprehensive metric is the average SMAE or the average score over the entire dataset. Here both the SMAE metric and the score metric are used to evaluate the prediction performance of the proposed data-driven constitutive modelling strategies.

4. Results and comparison of the three data-driven constitutive training approaches

4.1. Data preparation and the implementation of machine learning

In the current work, all the training data of triaxial testing is provided by discrete element modelling wherein a total of 4037 spherical particles with their radii uniformly distributed

between 2mm and 4mm are used to generate the specimens. These specimens are isotropically consolidated to a confining pressure of 200 kPa. The normal and tangential contact stiffnesses are 10^5 N/m and 5×10^4 N/m, respectively. The interparticle frictional coefficient is 0.5, the particle density is 2600 kg/m^3, the local damping coefficient is 0.5.

The GRU neural networks are built on Keras platform, which is an open-source model-level library allowing convenient construction of machine learning models. The low-level tensor operations behind Keras is performed by Tensorflow, a symbolic tensor manipulation library developed by Google.

Before constructing the deep learning models, the data from DEM requires preprocessing. One reason is that the raw input data with a large difference can increase the learning time and impede the convergence of the networks. Particularly, the input variables in realistic problems tend to be different in terms of units, scales, and distributions. In this study, standardisation is used to reshape the raw input data to the scaled data with a zero mean and a standard deviation of 1. The standardised data effectively reduces the risk of getting stuck in local optima and makes the training process faster. The other motivation for preprocessing data is that the data structure of input sequences must follow the prescribed format in a GRU model. All the input data must be a specific 3D array where the first dimension denotes the samples, the second dimension is the time steps and the last dimension represents the input features.

4.2. Case 1: conventional triaxial compression with multiple loading direction reversals

In the first case, conventional triaxial compression tests are used to generate stress-strain sequence pairs for deep training. We restrict the maximum axial loading strain to 12% with complex loading-unloading paths incorporating monotonic, one, two and three unloading-reloading cycles. These unloading and reloading strain values are mutually different and randomly sampled with a physical restriction that the reloading strain is always lower than its preceding unloading strain. A total of 220 datasets are prepared. After shuffling the database with a certain random seed (all the DNN training follows the same random seed to avoid potential information leaking), 100 groups of specimens are preserved as test specimens (the monotonic cases and the cases with 4, 5 and 6 unloading-reloading cycles, respectively, are artificially selected to test the AI models); while the remaining 120 groups of simulations are used for training and validation data with a partition ratio of 4:1. To provide an overview of the specimen distribution, the training, test and validation data specimens in this work are marked in Figure 3, but it should be noted that the classification here is simply an example and these groups can be shuffled and selected randomly. The other thing to be noted is that some loading

paths incorporate multiple unloading-reloading cycles and thus one loading path may include several unloading-reloading points in Figure 3.

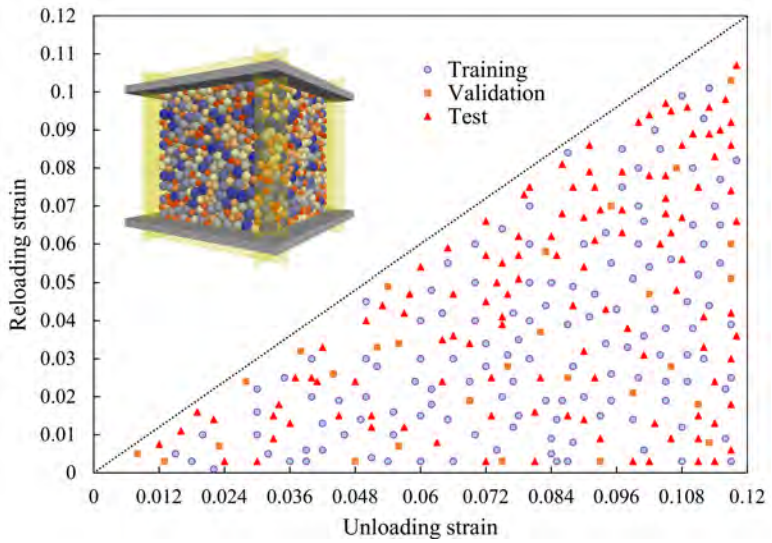


Figure 3: Sampling points for conventional triaxial compressions

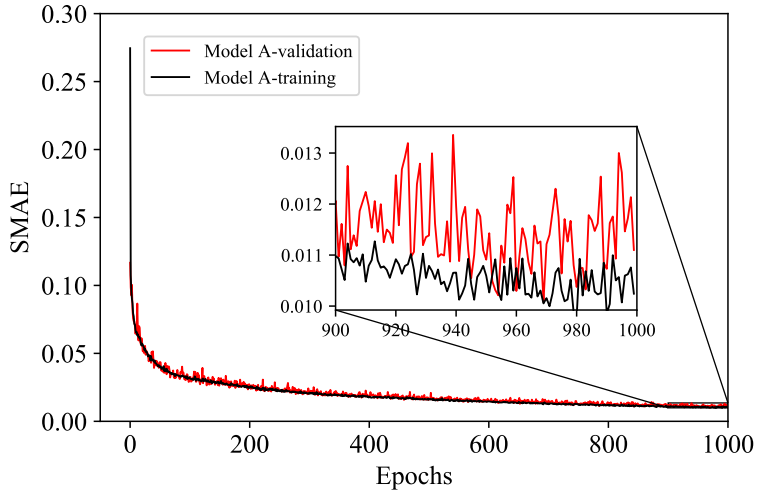
When the data has been preprocessed for training, the next step is to construct, train and validate machine learning models. To discover suitable network architectures and hyperparameter combinations for each model, a series of parametric studies are performed. The detailed process can be found in Appendix B. In this study, the adopted architecture and hyperparameters for Model A, Model B (NN-B1, NN-B2, NN-B3 and NN-B4) and Model C are shown in Table 1. The learning curves for the selected ANNs can be found in Figure 4.

Table 1: Network architectures and some key hyperparameters

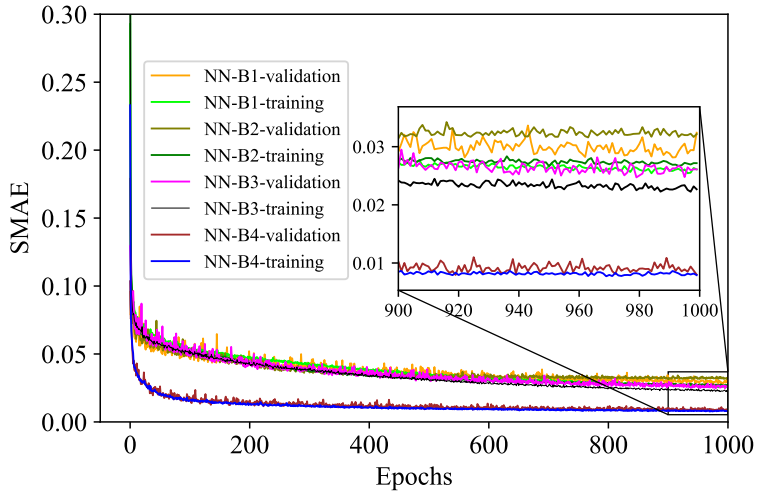
ANNs	Architecture	Timesteps	Batch size	Learning rate
Model A	GRU:100	40	64	0.001
NN-B1	GRU:120-GRU:120	50	128	0.001
NN-B2	GRU:120-GRU:120-Dense:20	30	128	0.01
NN-B3	GRU:120-Dense:100	60	64	0.01
NN-B4	GRU:100-Dense:20	40	128	0.01
Model C	GRU:40-GRU:40	55	128	0.01

For all the networks, the *tanh* activation function is used for the GRU layers and the linear activation function is applied to the output layer. The adaptive moment estimation (Adam) optimizer is used to update the weights iteratively with 1000 epochs. The scaled mean absolute error (SMAE) is used as the loss function. After finishing the training, the performance of the

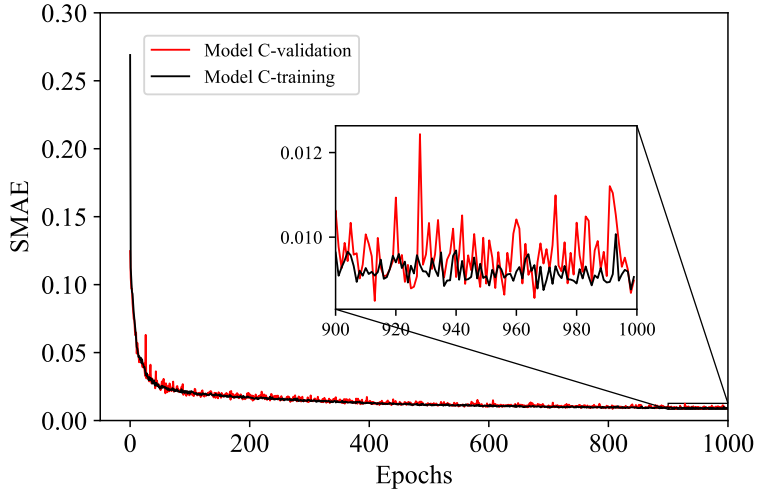
GRU model is evaluated on 100 groups of test data that have not been seen to the model during training.



(a) Model A



(b) Model B



(c) Model C

Figure 4: Learning curves of the selected ANNs in conventional triaxial loading conditions

4.2.1. Prediction results of Model A

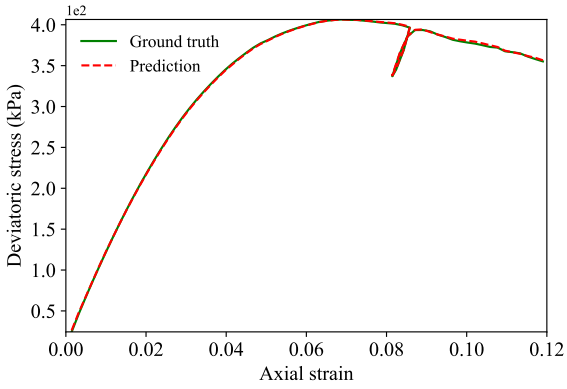
Model A predicts stress responses from only measurable principal strain sequences. The average SMAE and prediction score on the 100 groups of unseen test specimen are 0.0189 and 0.967, respectively. 60% predictions obtain a prediction score of 1.0, which demonstrates that the prediction accuracy of the trained model is acceptable. The best prediction has a SMAE of 0.007 with a score of 1, while the worst prediction has a SMAE of 0.054 with a score of 0.6458. Some of the typical predictions predicted by model A are shown in Figure 5. Although the worst prediction cannot achieve a high score, the overall tendency of stress responses has been captured. The results also indicate that the DNN model is able to predict the complex cases with more than two unloading-reloading cycles, even though these complex cases are never used for training and validating datasets.

Although Model A only considers the external principal strain variables, the values of these strain sequences are highly related to the internal evolution of microstructures and material properties of particles, i.e. the principal strain sequences implicitly encode the microstructural information inside specimens. This is also one of the reasons why Model A have an excellent prediction accuracy.

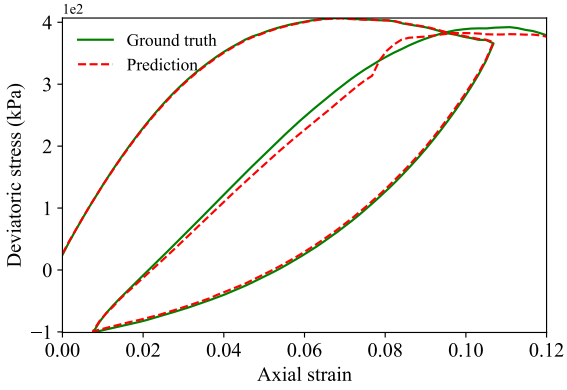
4.2.2. Prediction results of Model B

Model B explicitly incorporates microstructural information into training by introducing a directed graph to connect associated sub-networks. For model B, the average SMAE and prediction score on the 100 groups of test specimens are 0.0197 and 0.979, respectively. 57% predictions obtain a full score of 1.0. The best prediction SMAE is 0.007 with a score of 1 while the worst prediction has a SMAE of 0.044 with a score of 0.678. Some of the typical predictions via model B can be found in Figure 6.

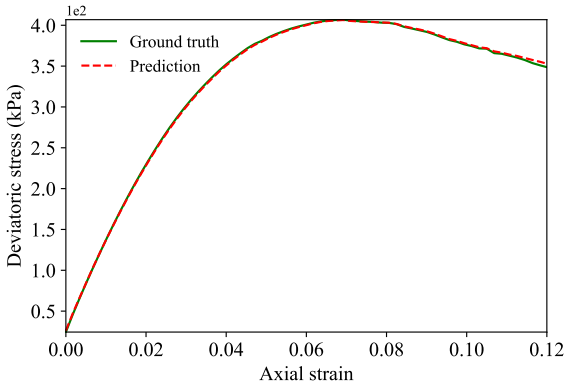
Similar to Model A, Model B can also satisfactorily capture complex unloading-reloading responses of constitutive behaviour. These two models are found to have similar prediction accuracy. However, as Model B is made of 4 different sub-networks, it may take 4 times more computational resources than model A to discover a suitable hyperparameter combination. Thus the results may support that model A is a preferred strategy to train a DNN based stress-strain model. This however arises a question: does the microstructural information not contribute to the stress-strain predictions? According to Figure 4, Table B.1 and Table B.2 in Appendix B, the model which uses microstructural variables directly as inputs (i.e. the sub-network: NN-B4) significantly outperforms Model A in terms of prediction accuracy. Thus it is certain that the microstructural information benefits the stress prediction.



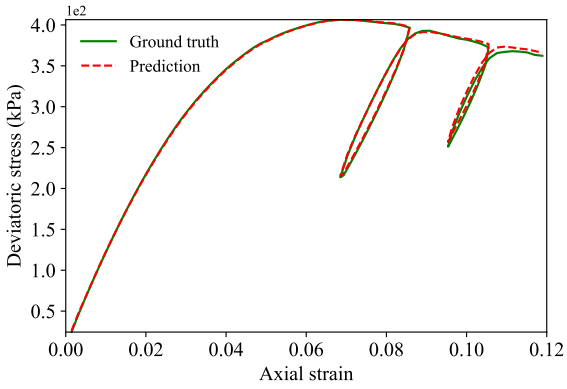
(a) The best prediction, score:1.0



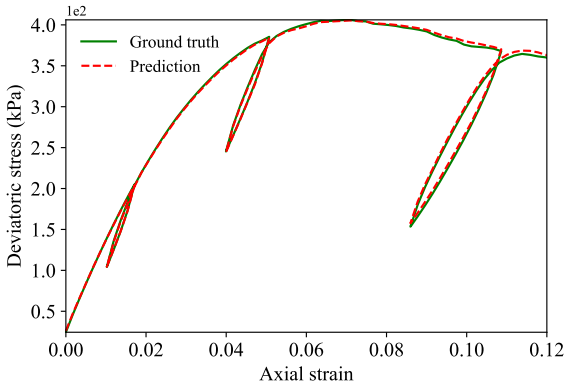
(b) The worst prediction, score:0.646



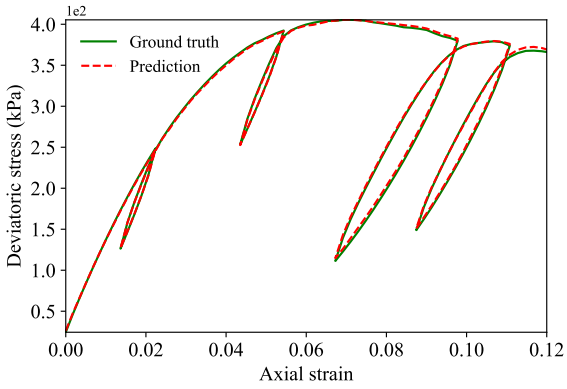
(c) Monotonic loading, score:1.0



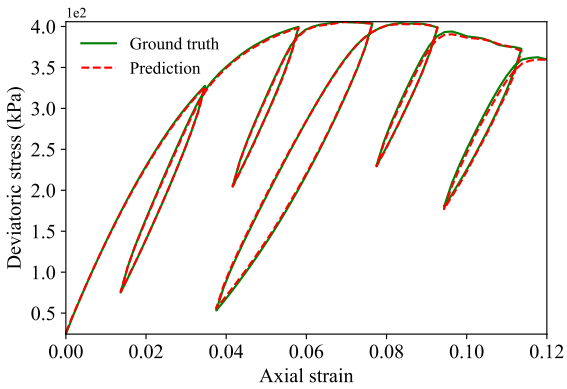
(d) Two unloading-reloading cycles, score:1.0



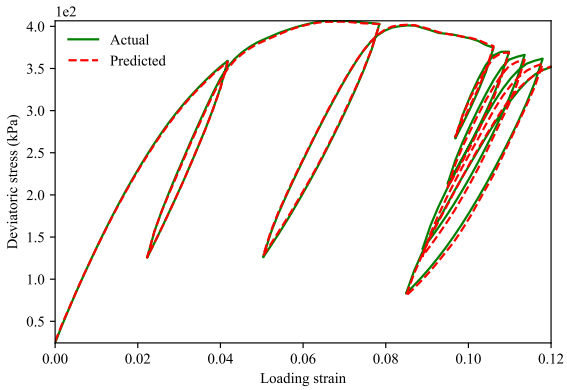
(e) Three unloading-reloading cycles, score:1.0



(f) Four unloading-reloading cycles, score:1.0

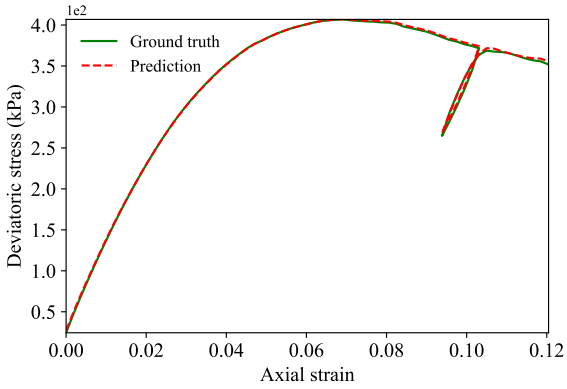


(g) Five unloading-reloading cycles, score:1.0

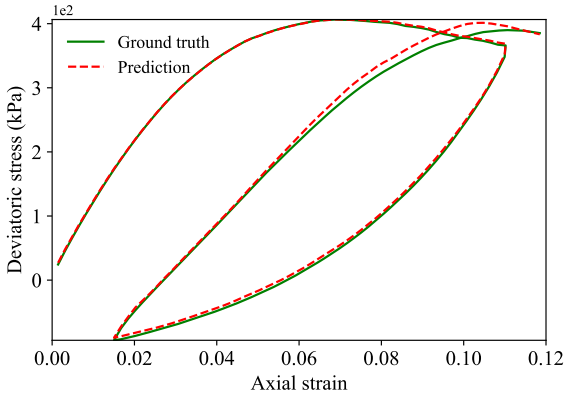


(h) Six unloading-reloading cycles, score:0.830

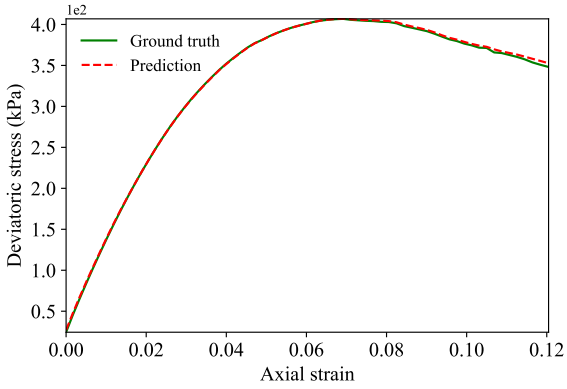
Figure 5: Representative prediction results of Model A



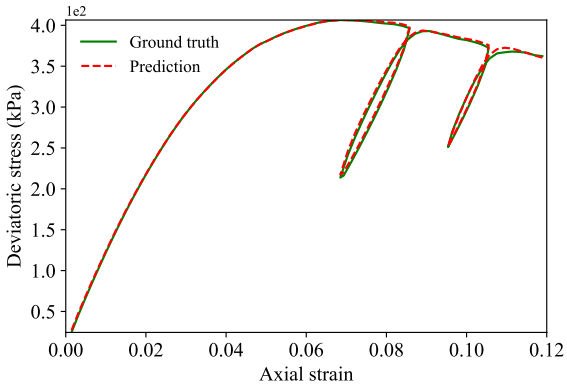
(a) The best prediction, score:1.0



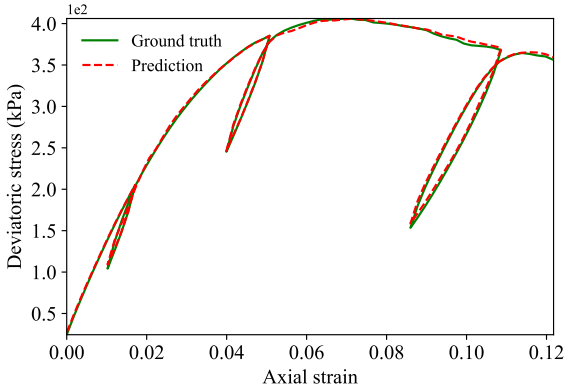
(b) The worst prediction, score:0.678



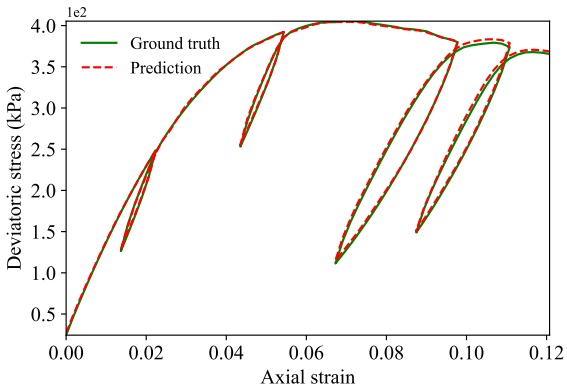
(c) Monotonic loading, score:1.0



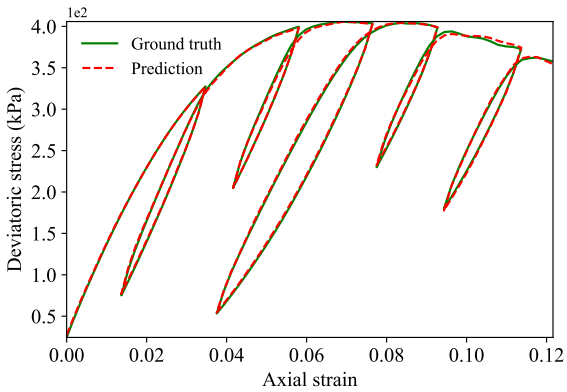
(d) Two unloading-reloading cycles, score:1.0



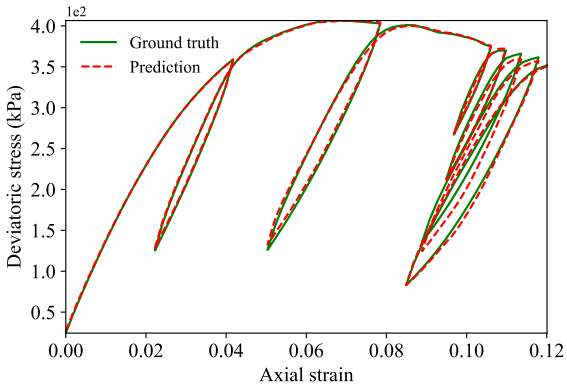
(e) Three unloading-reloading cycles, score:1.0



(f) Four unloading-reloading cycles, score:1.0



(g) Five unloading-reloading cycles, score:1.0



(h) Six unloading-reloading cycles, score:0.830

Figure 6: Representative prediction results of Model B

The problem is that the use of microstructural information as known inputs fundamentally violates the requirement of determining stress responses according to pure strain conditions in a typical constitutive model. To utilise the microstructural information but follow the basic principle of constitutive models, some extra measures like we have done in Model B are necessary.

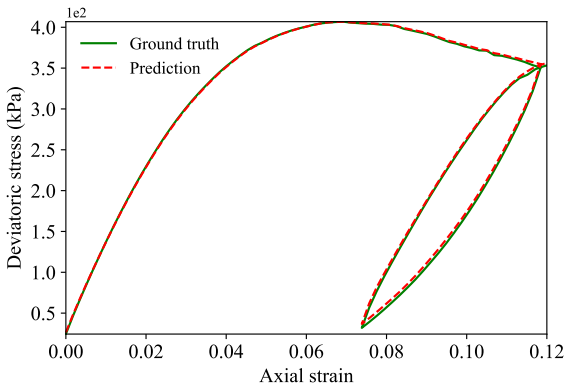
The reasons that microstructural tensors in Model B do not significantly improve prediction are mainly due to a relatively low prediction accuracy given by NN-B1, NN-B2 and NN-B3, as shown in Figure 4 and Table B.2. In model B, the outputs of former networks (NN-B1, NN-B2 and NN-B3) are parts of the inputs for the latter sub-network (NN-B4). If the former sub-networks do not have a satisfactory prediction accuracy, these inaccurate inputs will deteriorate the prediction results of the latter sub-network because artificial neural networks are incapable of recognising "fake" data. In the case that C_{3311}, C_{3322} and C_{3333} are not easy to be predicted by principal strain variables with a high prediction accuracy, it is understandable that the prediction accuracy of Model B (assembled with NN-B1, NN-B2, NN-B3 and NN-B4) is not necessarily high to improve the overall prediction performance compared to Model A.

4.2.3. Prediction results of Model C

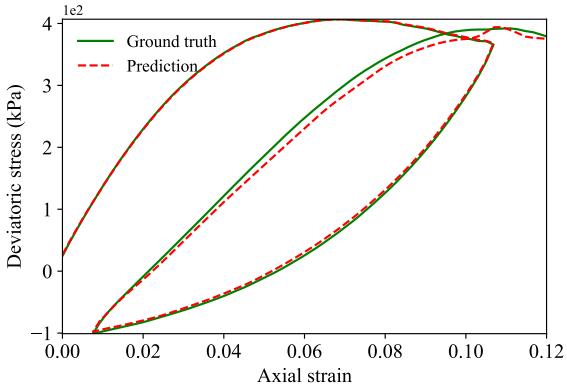
Model C makes full use of non-temporal physical properties in DNN training. The average SMAE and prediction score on the 100 groups of test specimens are 0.0170 and 0.984, respectively. 60% predictions obtain a prediction score of 1.0. The best prediction has a SMAE of 0.007 and a score of 1 while the worst prediction has a SMAE of 0.050 with a score of 0.664. Only two groups of prediction scores are lower than 0.8. Some representative predictions can be found in Figure 7. The overall prediction performance of incorporating non-temporal physical properties slightly outperforms those of models A and B.

4.3. Case 2: true triaxial compression incorporating contant-b and constant-p loading conditions

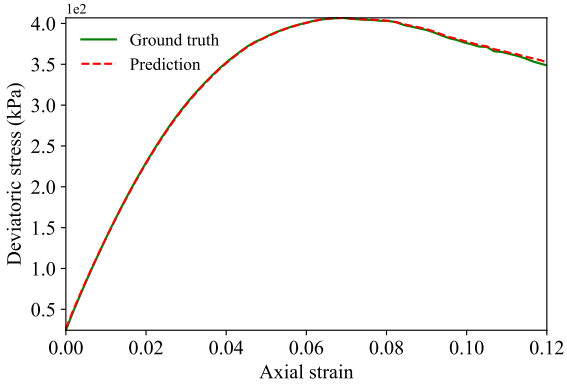
The proposed three DNN training models of granular materials are further examined by considering true triaxial compression loading conditions with unloading-reloading cycles. Two types of typical multi-directional loading cases, 1) the isobaric (constant-p) axisymmetric triaxial loading ($p = -\frac{1}{3}(\sigma_{11} + \sigma_{22} + \sigma_{33})$) and 2) true triaxial compression with constant intermediate principal stress coefficient (constant-b) ($b = \frac{\sigma_{22}-\sigma_{11}}{\sigma_{33}-\sigma_{11}}, \sigma_{11} = \sigma_{22}$), are performed via DEM. The numerical material parameters are the same as the one used in Case 1. For the constant-p condition, the strain paths incorporating both monotonic loading and unloading-reloading loops are prepared. The sampling points for describing loading paths can be found in Figure 8. For the constant-b case, the b value ranges from 0 to 1.0 with an interval of 0.05. In total,



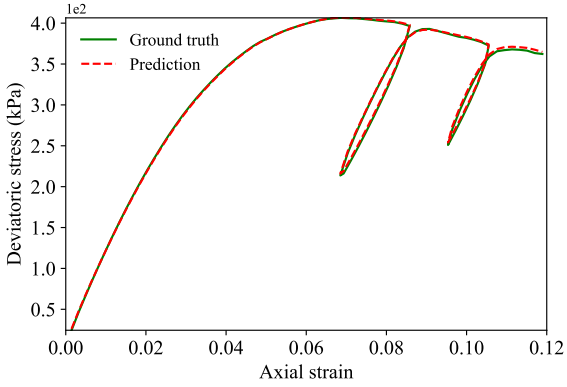
(a) The best prediction, score:1.0



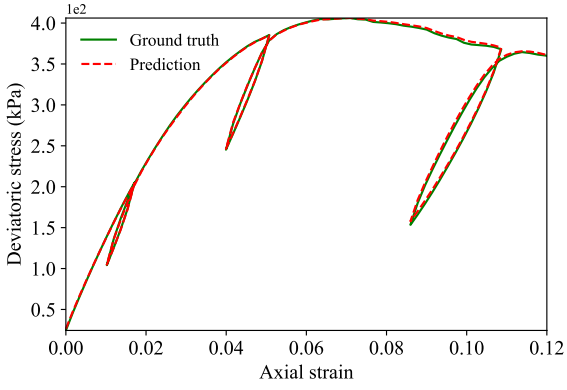
(b) The worst prediction, score:0.664



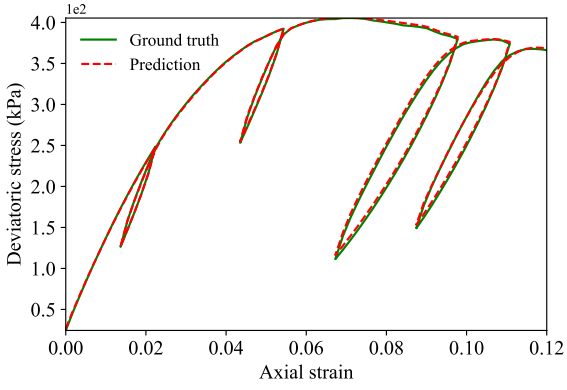
(c) Monotonic loading, score:1.0



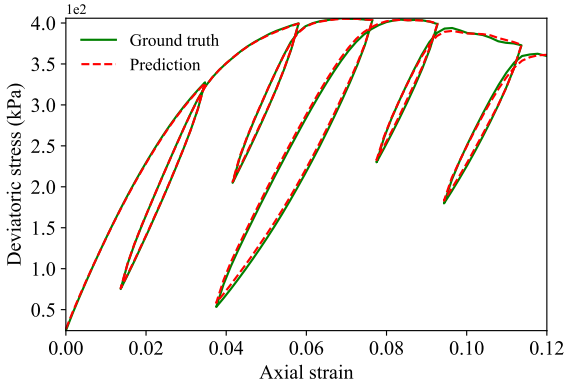
(d) Two unloading-reloading cycles, score:1.0



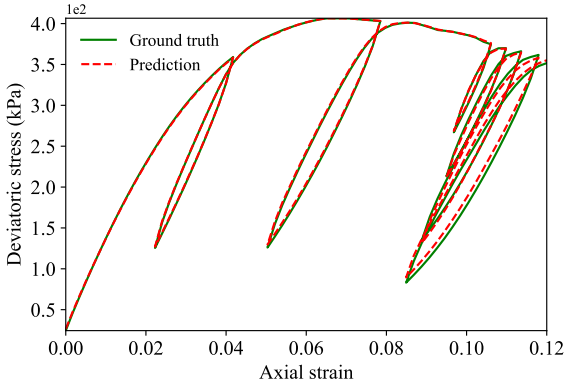
(e) Three unloading-reloading cycles, score:1.0



(f) Four unloading-reloading cycles, score:1.0



(g) Five unloading-reloading cycles, score:1.0



(h) Six unloading-reloading cycles, score:0.899

Figure 7: Representative prediction results of Model C

142 groups of specimens are generated (122 for constant-p and 20 for constant-b cases), with 67 groups for training, 39 groups for validation and 36 groups for testing, respectively.

It is found that the discovered optimal architectures in conventional triaxial compression cases are still able to yield satisfactory predictions in the true triaxial loading conditions. Therefore, most of the architectures and hyperparameters continue to use in Case 2 except for several modifications. Specifically, the batch size and timesteps for sub-network NN-B3 are changed to 128 and 40, respectively; the epoch numbers are reduced to 500 for all the models on the true triaxial loading data, because the loss functions have come to a steady value and more training may cause overfitting. The learning curves for the true triaxial loading cases are shown in Figure 9.

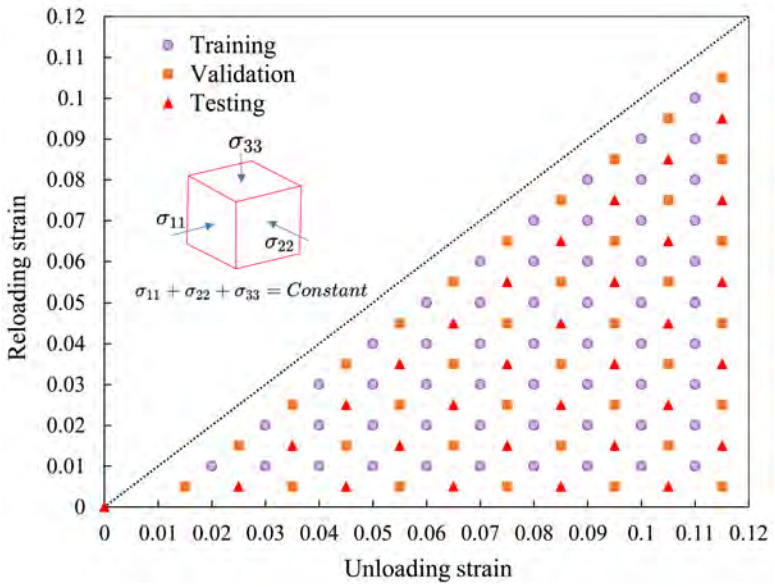
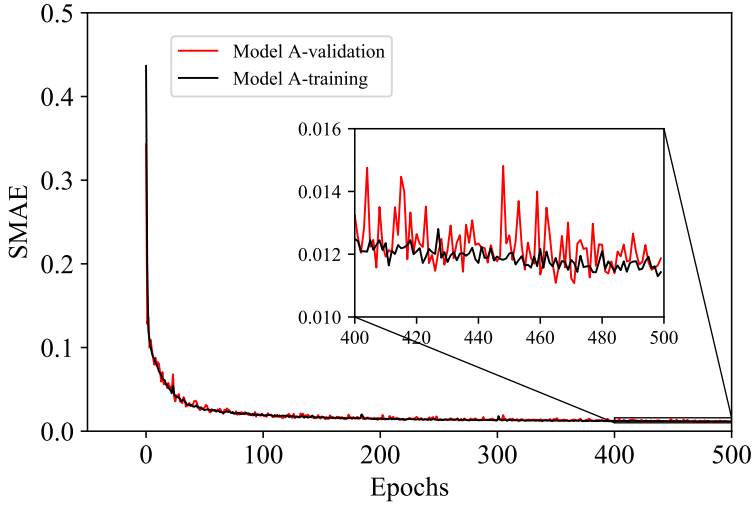


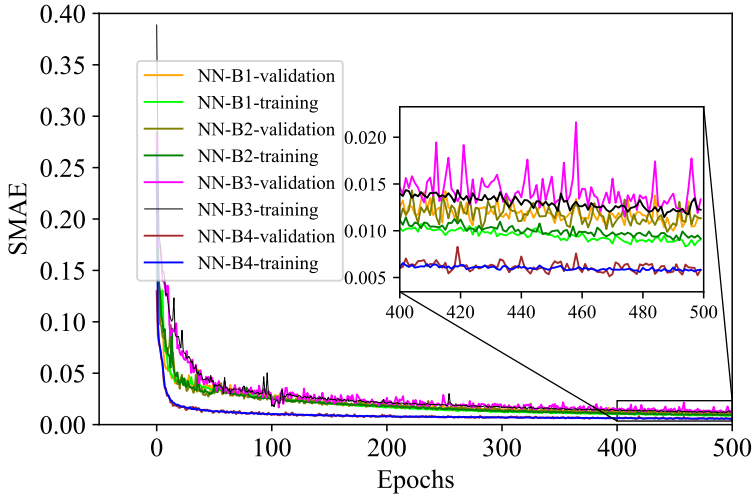
Figure 8: Sampling points for constant-p loading

Figures 10 show some representative results on both the constant-b and the constant-p loading cases. All predictions on the major, intermediate, and minor principal stress components are given. It can be seen that the stress evolutions in the investigated multi-directional conditions are excellently captured by the proposed AI models. In the 36 groups of unseen test specimens, the average SMAE and prediction score are 0.012 and 0.982, respectively; 29 groups of predictions (80%) obtain a score of 1.0. The worst prediction has a SMAE of 0.021 with a score of 0.868. Figures 10(a-c) shows even the worst three predictions have forecasted stress responses satisfactorily.

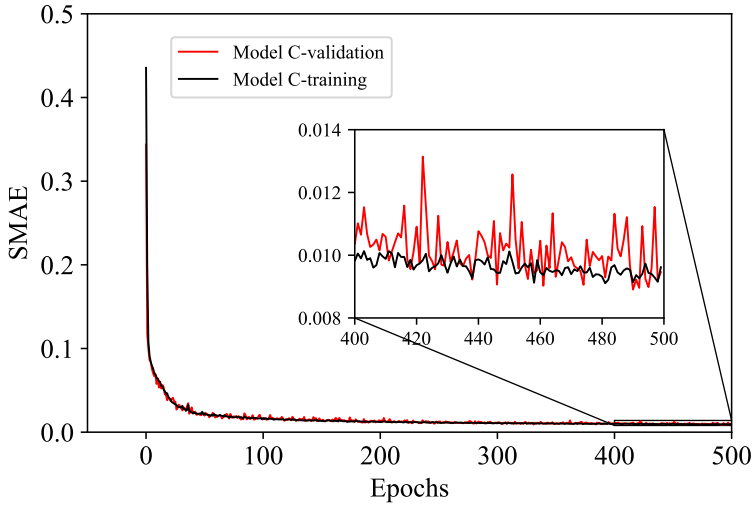
Models B and C also have excellent prediction performance. Model B obtains an average SMAE of 0.013 and an average score of 0.981 on the 36 groups of test specimens. 30 groups of predictions achieve a full credit of 1.0. Model C gives an average SMAE of 0.0099 with



(a) Model A

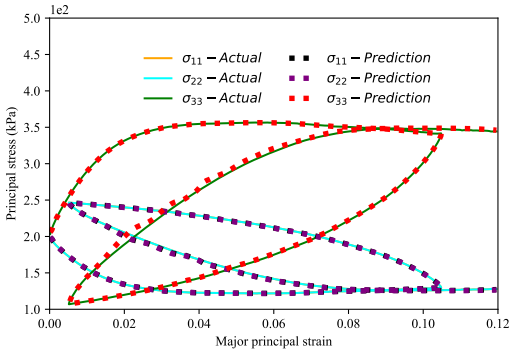


(b) Model B

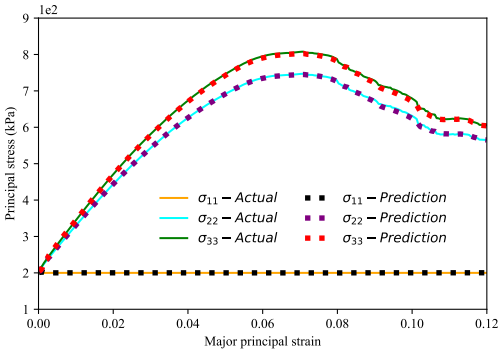


(c) Model C

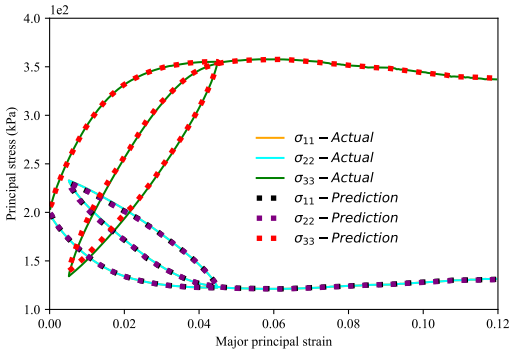
Figure 9: Learning curves of the selected ANNs in true triaxial compression conditions



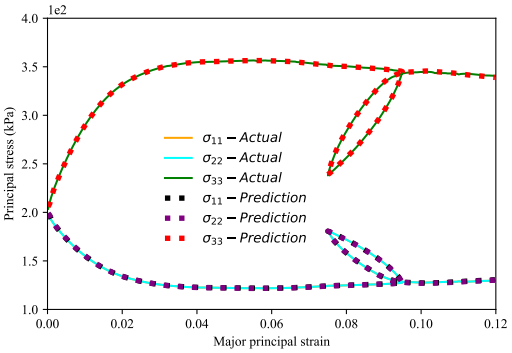
(a) The worst prediction, score:0.868



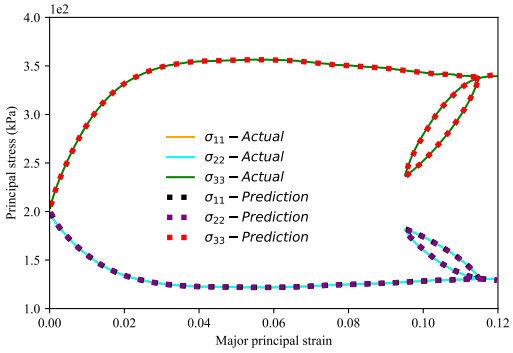
(b) The second worst prediction, score:0.869



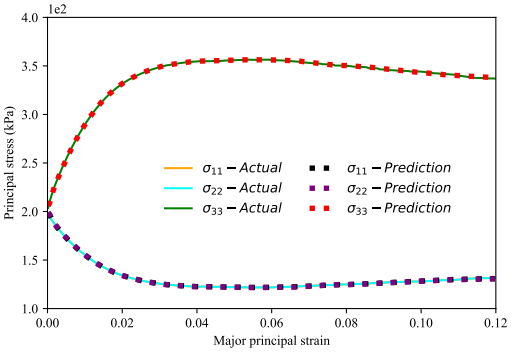
(c) The third worst prediction, score:0.873



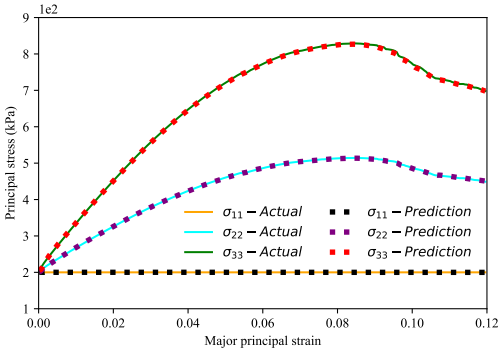
(d) The best prediction, score:1.0



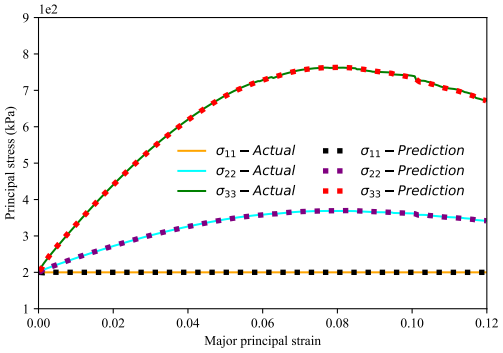
(e) The second best prediction, score:1.0



(f) Monotonic constant-p loading, score:1.0



(g) Constant-b, b=0.5, score:1.0



(h) Constant-b, b=0.3, score:1.0

Figure 10: Representative predictions on true triaxial compression given by Model A

an average score of 0.995. Nearly all the predictions given by Model C obtain a score of 1.0 except for two specimens, whose prediction scores are 0.847 and 0.958, respectively. In the true triaxial loading cases, the results confirm that the previous finding in the conventional triaxial compression cases, i.e., the prediction performance of Model B is similar to that of Model A, while Model C slightly outperforms Models A and B. Considering that the results of Model A in Figure 10 are sufficiently accurate, the representative prediction demonstrations for models B and C are not given here.

4.4. A brief summary on the features and applications of the three DNN-based training approaches

One aim of developing a DEM based data-driven constitutive model is exploring the possibility of extending the data-driven paradigm into experimental environments, with a hope that the deep learning model can replace traditional phenomenological constitutive models. In laboratory experiments, the principle strain information can be directly measured in true triaxial testing apparatus. However, the microstructural evolution information inside a granular specimen and the particle-scale properties are not readily available. Therefore, only Model A can be directly applied to experimental conditions. The prediction results in Figures 5 and 10 demonstrate that the DNN model with only the measurable external variables can reproduce the complex unloading-reloading responses satisfactorily. It is thus possible to develop an experiment-based constitutive prediction model, provided that a sufficient training dataset is available.

Although model B does not significantly outperform model A, the contribution of each microstructural component (i.e. C_{3311}, C_{3322} and C_{3333}) in determining the final stress responses is discovered by the deep learning models. This unique design may provide a possibility of developing an AI-guided scheme to discover some hidden physical rules. Further related research will be explored.

Model C slightly outperforms model A and model B in terms of prediction performance but its downside is that more particle-scale properties are required. Although natural granular materials tend to be heterogeneous and the particle-scale properties are hard to be measured in experiments, these properties are directly available in numerical computations. Thus this training approach can be used as a reinforcement strategy for training a DEM-based constitutive model. The application of such a DEM-based model will be discussed in Section 5.2.

5. Discussion

5.1. Primary factors for training reliable data-driven constitutive models

In this work, our goal is to train deep learning models to reliably predict the constitutive behaviour of granular materials. Three aspects have been explored: (1) the analytical equations of granular materials are utilised to discover key ingredients for describing the constitutive relations of granular materials; (2) the directed-graph based constitutive modelling strategy enables the DNN model to explicitly introduce more microstructural evolution data to assist the training; and (3) the enhanced GRU architecture is introduced to incorporate the non-temporal but physically essential material or environment properties during training.

Among these three factors, the first factor is the key to achieve excellent prediction accuracy. The premise of training a reliable DNN is that a certain pattern or mapping exists among the variables involved in training and the neural network is capable of capturing these inherent mappings. In a conventional triaxial testing environment, the lateral strains are normally ignored but the incremental analytical formulae in Section 2 reveal the critical role of the lateral strains playing in determining the stress responses. Specifically, the mapping between loading stresses and loading strains of granular materials is not an injective or "one-to-one" function. Figure 11 shows the axial stress-strain curves of granular specimens undergoing monotonic shearing. The axial loading rate and path are the same whereas the lateral intermediate principal stress ratio b varies from 0.1 to 0.9. It is evident that the same loading strain causes different stress responses, i.e. the pattern or mapping between the loading stress and strain is not unique. This finding highlights not only the key contribution of micromechanical principles in guiding machine learning but also the importance of introducing comprehensive principal strain information in constitutive modelling.

5.2. Potential applications of DEM based data-driven constitutive modelling

Deep learning normally requires a large amount of data while laboratory tests are expensive and time-consuming. In contrast, DEM provides cheap and flexible modelling data under complex mechanical states. It is feasible to use DEM as virtual "surrogate models" to pave the way for experiments-based constitutive modelling, e.g. to understand how to train a reliable model with the least data specimens. Besides, under the conditions that more advanced particle-scale contact rules are developed to reproduce the realistic granular interactions (Feng et al., 2017; Feng, 2021a,b; Zhao and Feng, 2018) and that the DEM parameters are well-calibrated for representing realistic granular behaviour (Qu et al., 2020a,b), experiments-based granular testing is possible to be replaced with a large number of DEM modelling to reduce costs.

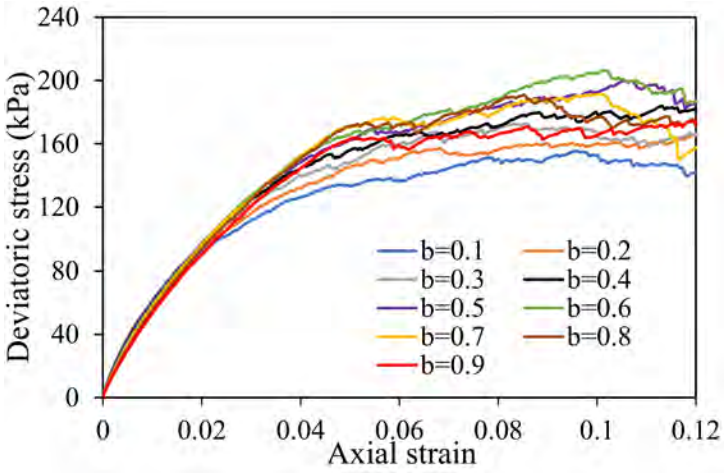


Figure 11: Stress-strain relations of a granular specimen subjected to monotonic shearing under varied intermediate principal stress

The other application is to advance hierarchical multiscale modelling techniques where the constitutive laws are provided by the microscopic modelling (e.g. DEM), instead of assuming a phenomenological constitutive model a priori. The foundation of this method is that the microscopic DEM reasonably reflects the discrete nature of granular media and is capable of capturing the salient macroscopic behaviour of granular materials, although DEM typically simplifies the complexity of real granular media (O'Sullivan, 2011; Gong et al., 2019b,a). Until now, applications of multiscale techniques to the simulation of engineering-scale granular geomaterials are still uncommon due to the required great computational costs. The current FEMxDEM hierarchical multiscale modelling approach interpolates the deformation from the FEM solution to DEM as its boundary conditions. The DEM will deform following this prescribed boundary condition and then return the corresponding stress to FEM (Guo and Zhao, 2014). The whole process is time-consuming as a large number of discrete element models are required. Taking the DEM calculation of triaxial testing in this work as an example, a complete calculation requires around 24 hours (Core(TM) i5-7400, 3GHz) but the DNN model can predict the stress responses in just several seconds. In the case that a reliable DEM-based data-driven model has been trained, the DNN model can replace the original DEM simulations in hierarchical multiscale modelling. Then the efficiency of hierarchical multiscale modelling will be greatly enhanced.

5.3. Sampling intervals in NN-based models

For a given strain path, conventional analytical models give stress predictions that are irrespective of the size of strain increments but NN-based constitutive models are not. As demonstrated by Jung and Ghaboussi (Jung and Ghaboussi, 2006), the NN-based models are not recommended for the problems with the step sizes or sampling intervals which are signifi-

cantly different from those used in the training stage. Otherwise, some forecast deviations will inevitably occur when using these NN-based models for extrapolation. To develop a NN model suitable for applications with different step sizes, the data with different sampling intervals should be used for training.

5.4. The deficiency of current research and future work

The quality of data specimen is the key to obtaining high-accuracy machine learning models. A qualified dataset should be capable of representing all possible situations that the model is intended for. In the current work, the specimens are generated by random or grid sampling in the entire data space. However, it is likely that the data demand can be reduced without deteriorating prediction accuracy. One strategy is to introduce advanced statistical sampling techniques. The other strategy is to develop active machine learning models wherein only the most critical data informed by the model is required to provide. These techniques are expected to use as small a dataset as possible to train a reliable constitutive model.

When data is prepared, the next step is to find suitable network architectures and neuron weights via iterative feedforward and feedback. Many hyperparameters, eg. the number of layers, hidden units, training epochs, learning rate etc. affect the learning process and final results. In this study, we determine these hyperparameters with a large number of parametric studies within the whole search space of hyperparameters. A superior combination of hyperparameters may exist. An adaptive hyperparameter adjusting algorithm can be helpful to further improve the training in this aspect.

Although multiple loading-unloading cycles are considered in this work, there are many other complex mechanical states in realistic engineering problems. In addition, the elastic-plastic responses of granular materials are influenced by the initial fabric, void ratio, size and shape distributions of grains, and mineralogical compositions. Developing a DNN model that is capable of adapting to general strain paths and different granular materials is still highly challenging. In the future, advanced phenomenological models may deserve more considerations in data training to reduce data demands and improve the generalisation capability of the data-driven models by making full advantage of a priori knowledge of available elastic-plastic theory.

6. Conclusions

This study attempts to develop data-driven constitutive modelling strategies for granular materials by integrating micromechanical theory with deep learning models. The derived mechanical formulation identifies critical variables associated with the constitutive behaviour.

Three different training approaches are explored. The first one (Model A) uses only external variables recognised by the analytical formulae (i.e. the principal strain sequences) to approximate stress responses; the second one (Model B) utilises the directed graph to link all the internal and external variables into a single information flow constituted by a series of subnetworks to make predictions, and the third one (Model C) integrates non-temporal physical properties with Model A into training. The proposed three constitutive modelling strategies are found to be capable of predicting stress-strain responses of granular materials with satisfactory agreement on the unseen test specimens. The prediction capability of these three approaches is close to each other with Model C slightly outperforming Model A and Model B. The key findings are as follows:

It is practically feasible to develop a data-driven constitutive model with high accuracy. The direct involvement of comprehensive principal strain information in constitutive training greatly facilitates training reliable machine learning models.

The introduction of microstructural evolution information in a directed graph benefits constitutive modelling but the prerequisite is that the microstructural evolution information is sufficiently accurate. Incorrect microstructural information not only is unhelpful but also impedes reliable predictions.

Provided that the particle-scale properties are known (e.g. DEM conditions) or can be measured, these physically important non-temporal properties can be integrated with temporal strain paths to reinforce DNN-based constitutive prediction.

The combination of a priori knowledge with the data-driven paradigm is promising to solve complex scientific problems. With the aid of micromechanical principles, we avoid searching all possible variables associated with constitutive modelling and prevents the AI-based constitutive modelling to be a combinatorial optimisation problem. Besides, the micromechanical principle also enhances the interpretability of a data-driven constitutive model.

7. Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (NSFC) (Grant Nos. 41606213, 51639004 and 12072217). The authors also would like to thank the five anonymous reviewers for their careful and thoughtful suggestions that have helped improve this paper substantially.

Appendix A. Introduction to the enhanced GRU architecture with incorporating physics-invariant quantities

The original core structures of GRU are the hidden state and the two gates. The hidden state enables to transfer past relevant information along with the sequences. The two gates, *reset* and *update*, are designed to process the data within each GRU cell. The reset gate is used to decide how much past information to forget while the update gate determines what information to throw away and what new information to add.

In the enhanced GRU architecture (see Figure A.1), a secondary hidden state is designed in the formulation to carry non-temporal information through each GRU cell, thus enabling the ANN to develop a desired hypothesis function by fully utilising temporal and non-temporal inputs, and history-dependent hidden states. Although the introduction of these invariant parameters increases the complexity and training costs, it is useful to guarantee the generalisation ability of a trained ANN model. Furthermore, this architecture has better interpretability and thus provides more possibilities for finding the underlying physical laws based on the data-driven paradigm.

The mathematical expressions in the enhanced GRU architecture can be found as follows:

Reset gate (r_t):

$$r_t = \ sigm\left(W_r[h_{t-1}, x_t, h_c] + b_r\right) \tag{A.1}$$

Update gate (z_t):

$$z_t = \ sigm\left(W_z[h_{t-1}, x_t, h_c] + b_z\right) \tag{A.2}$$

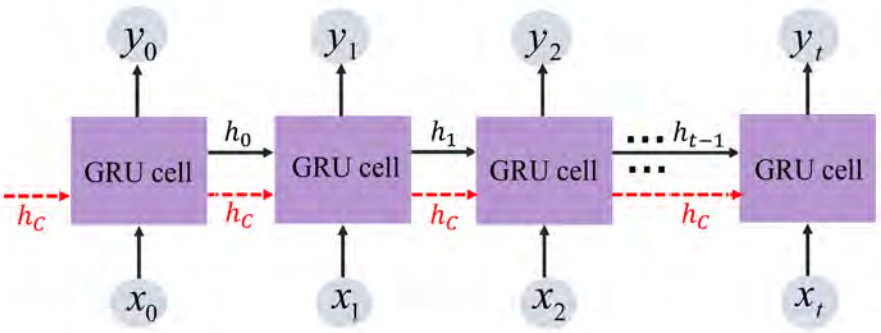
Candidate primary hidden state ($\widetilde{h}_t$):

$$\widetilde{h}_t = \ tanh(W_h[r_t \otimes h_{t-1}, x_t, h_c] + b_h) \tag{A.3}$$

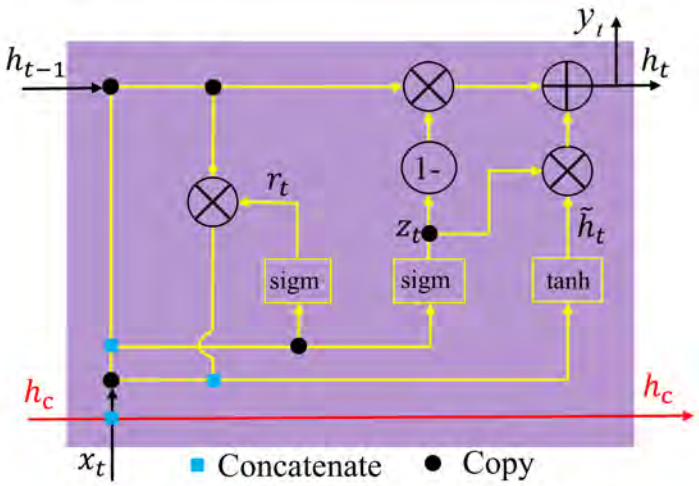
New primary hidden state:

$$h_t = (1 - z_t) \otimes h_{t-1} \oplus z_t \otimes \widetilde{h}_t \tag{A.4}$$

In the above formulations, x_t is the current input at the t^{th} time step; h_{t-1} is the primary hidden state at the $(t-1)^{th}$ time step; h_c is the secondary hidden state used for incorporating physics-invariant variables; $sigm(x)$ is the sigmoid activation function: $sigm(x) = \frac{1}{1+\exp(x)}$; $tanh(x)$ is the hyperbolic tangent function: $tanh(x) = \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}$; W_r, W_z, W_h are weights; b_r, b_z, b_h are biases; the symbols $\otimes$ and $\oplus$ represent element-wise multiplication and addition, respectively.



(a) Unfold sequences for the enhanced RNN architecture



(b) mathematical operations within each enhanced GRU cell

Figure A.1: Schematic diagram for enhanced GRU cells

The actual operations involved in each enhanced GRU cell are shown in Figure A.1b. First, the secondary hidden state is concatenated with the current inputs to form a new vector. Second, this new vector is further concatenated with the previous primary hidden state to form the second new vector. To introduce non-linearity into the model, the sigmoid function acting as a nonlinear active transformation function is applied to obtain the reset gate r_t (Eq. (A.1)). Third, the gate z_t is updated by applying the *sigmoid* function on the second new vector (Eq. (A.2)). Fourth, element-wise multiplication is conducted between the previous primary hidden state and the reset gate r_t, and the multiplication with the first new vector is concatenated to form the third new vector. The candidate primary hidden state $\widetilde{h}_t$ can be calculated by applying the *tanh* function on the third new vector (Eq. (A.3)). Finally, the new primary hidden state h_t can be determined based on Eq.(A.4).

Appendix B. Parametric investigations for determining the hyperparameters of each training model

Appendix B.1. Model A: only loading and lateral strain information are involved

A preliminary network architecture investigation starts from one or two GRU layers, followed by one or zero dense layer, before connecting the output layer. The neuron number in each hidden layer varies from 0 to 120 with a gap of 20. The final architecture will be determined by considering: (1) the amount of SMAE and (2) the complexity of architectures. Specifically, if the difference of two SMAEs is within 10^{-5}, the simplest architecture (the least parameters to be trained) will be selected because a simpler model has less risk of overfitting. The SMAEs of different network architectures can be found in Figure B.1 and the architecture [GRU:100-dense:0] is finally selected. This model requires a total of 31,301 weights and biases to be learned with 31,200 parameters for the GRU layer and 101 parameters for the output layer.

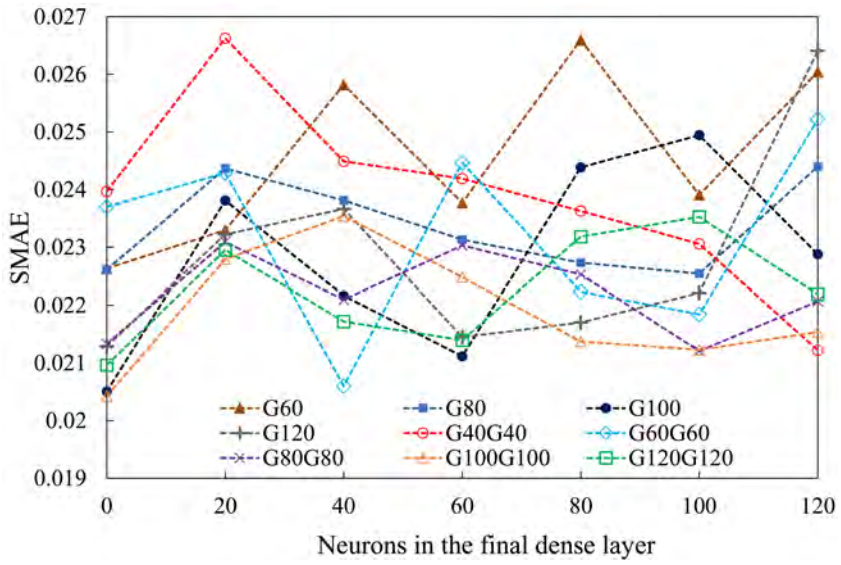


Figure B.1: SMAEs of model A with different network architectures

On the basis of the selected architecture, we further consider the influence of other important hyperparameters, such as timesteps (i.e. the number of lag observations into each GRU unit), batch size and learning rate. The influences of timesteps are shown in Figure B.2. The influences of batch size and learning rate can be found in Table B.1. The final hyperparameters of Model A are: timesteps: 40, batch size: 128, learning rate: 0.01. The inputs are temporal strain variables: ε_{11}, ε_{22}, ε_{33}. The output is the corresponding deviatoric stress sequence. A total of 87 groups of distinct training configurations are considered to determine the architecture and hyperparameters of the final training model.

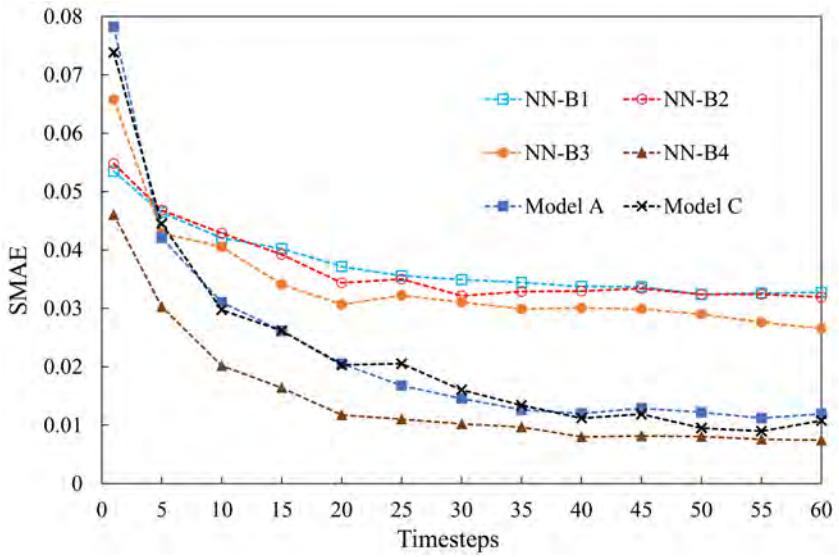


Figure B.2: SMAEs of the selected ANNs (models A, B and C) against various timesteps

Note that the hyperparameter selection for a deep neural network is essentially a very high dimensional combinatorial optimisation problem, it is thus not easy to search all the possible combinations considering available computational resources. Although the parametric study does not cover many other combinations, it gives a relatively reliable architecture and parameters for the model in the searched parameter space.

Appendix B.2. Model B: incorporating microstructural variables

As stated in Section 3.1, model B is made of 4 sub-ANNs by incorporating microstructural information during triaxial testing. The optimal representation of model B requires its constituent sub-networks to reach their optimal prediction accuracy. Therefore, the hyperparameter investigation for these associated sub-ANNs should be carefully conducted as well. Similar to Model A, we explore a rational network architecture first. The performance of each sub-ANN is demonstrated by Figures B.3, B.4, B.5 and B.6.

Table B.1: SMAEs of the selected ANN architectures for model A with different batch sizes and learning rates

Timesteps	Batch size	Learning rate	SMAE
40	16	0.001	0.021023878
40	16	0.01	0.018084617
40	32	0.001	0.019470268
40	32	0.01	0.019951961
40	64	0.001	0.01112244
40	64	0.01	0.01153904
40	64	0.1	0.860573057
40	128	0.001	0.011273821
40	128	0.01	0.011982925
40	128	0.1	0.827519575
40	256	0.001	0.014308579
40	256	0.01	0.013342785
40	256	0.1	0.664513549

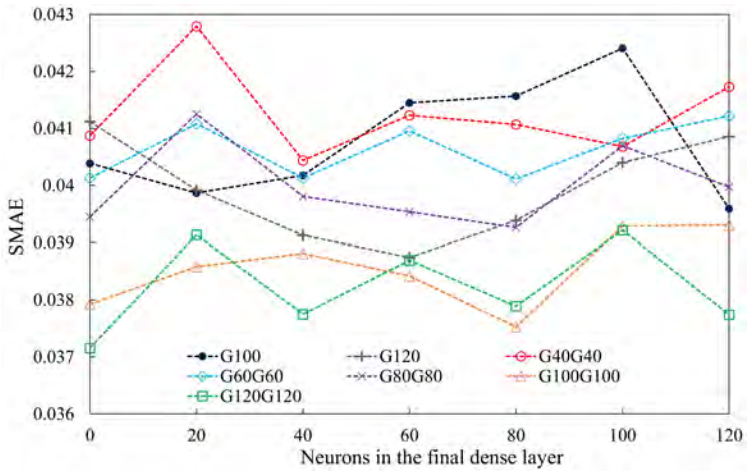


Figure B.3: SMAEs of NN-B1 with different network architectures

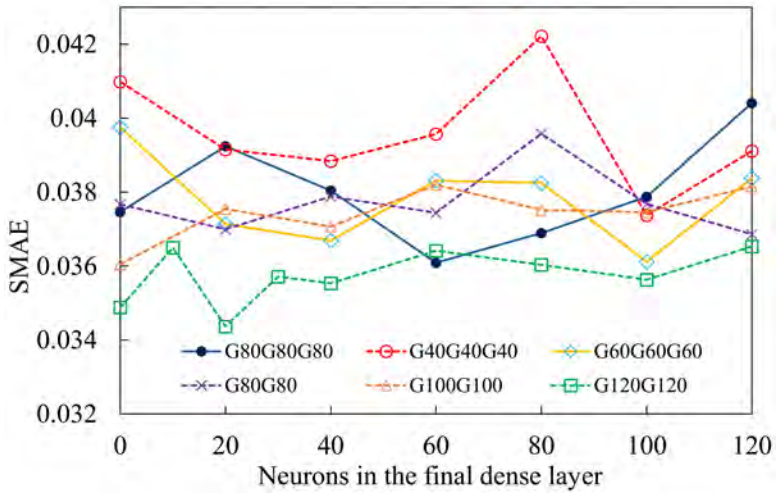


Figure B.4: SMAEs of NN-B2 with different network architectures

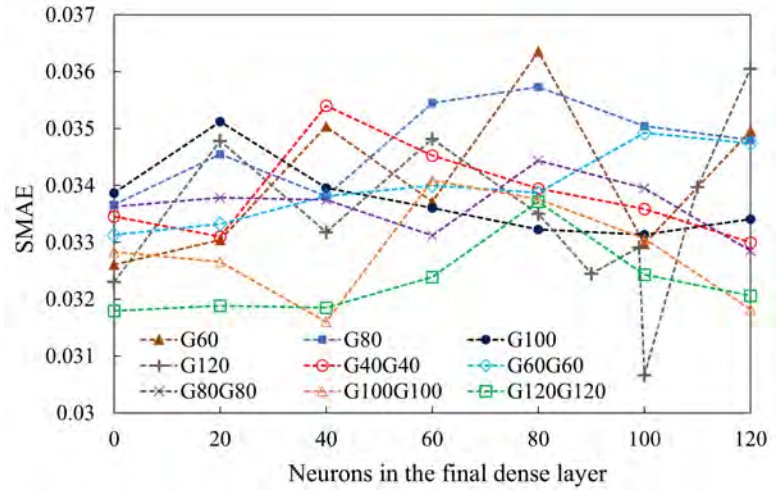


Figure B.5: SMAEs of NN-B3 with different network architectures

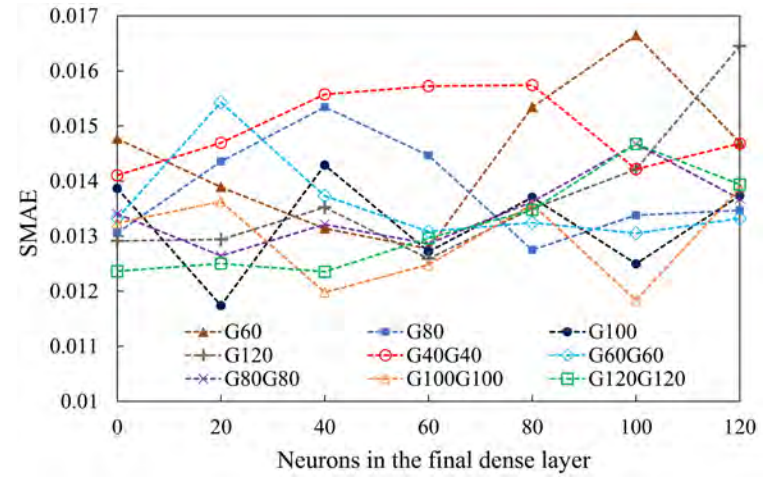


Figure B.6: SMAEs of NN-B4 with different network architectures

Through comparison, the final architectures to be selected and the corresponding trainable parameters for these four sub-ANNs are: (1) NN-B1: [GRU:120-GRU:120-dense:0], a total of

131,521 parameters are trained with 44,640 for the first GRU layer, 86,760 for the second GRU layer, and 121 for the output layer; (2) NN-B2: [GRU:120-GRU:120-dense:20], a total of 133,841 parameters are trained with 44,640 for the first GRU layer, 86,760 for the second GRU layer, 2420 for the dense layer and 21 for the output layer; (3) NN-B3:[GRU:120-dense:100], a total of 56,841 parameters to be learned, with 44,640 for the first GRU layer, 12,100 for the second GRU layer, and 101 for the output layer; (4) NN-B4:[GRU:100-dense:20], a total of 34,141 trainable parameters are required with 32,100 parameters for the GRU layer, 2020 parameters for the dense layer and 21 parameters for the output layer.

The influences of timesteps on the performance of each selected architecture is shown in Figure B.2. Furthermore, the effects of batch size and learning rate are also explored and the results can be found in Table B.2. The discovered optimal network architecture and hyperparameter combinations for each sub-ANN are summarised in Table 1 in Section 4.2. A total of 273 groups of training configurations are considered to determine a suitable hyperparameter combination for Model B.

Appendix B.3. Model C: incorporating non-temporal physical variables

With the incorporation of four physics-invariant variables, the inputs include temporal strain variables: ε_{11}, ε_{22}, ε_{33} and non-temporal physical parameters: the normal and tangential contact stiffnesses of particles, the sliding friction coefficient, and the boundary conditions: confining stress. The output is the corresponding deviatoric stress sequence. Following the same hyperparameter investigation scheme as Model A and Model B, the SMAEs of different network architectures are shown in Figure B.7, wherein the [GRU:40-GRU:40] architecture is selected. This model requires a total of 15,521 weights and biases to be trained, with 5760 parameters for the first GRU, 9720 parameters for the second GRU and 41 parameters for the output layer. Then starting from this architecture, the effect of timesteps on the SMAE of predictions can be found in Figure B.2 and a timestep of 55 is found to achieve the minimum SMAE among the investigated cases. A total of 84 groups of training cases are performed to obtain the final hyperparameter combination for Model C.

Table B.2: SMAEs of the selected ANN architectures with different batch sizes and learning rates

ANNs	Timesteps	Batch size	Learning rate	SMAE
NN-B1	50	16	0.01	0.034787547
	50	32	0.01	0.036005609
	50	64	0.01	0.032427967
	50	128	0.01	0.032401011
	50	16	0.001	0.041222154
	50	32	0.001	0.038220792
	50	64	0.001	0.032896032
	50	128	0.001	0.032357865
NN-B2	30	16	0.01	0.039313793
	30	32	0.01	0.037995158
	30	64	0.01	0.033419531
	30	128	0.01	0.032173722
	30	16	0.001	0.039949795
	30	32	0.001	0.039084379
	30	64	0.001	0.034799603
	30	128	0.001	0.032594303
NN-B3	60	16	0.01	0.037514748
	60	32	0.01	0.041043216
	60	64	0.01	0.026101194
	60	128	0.01	0.026550017
	60	16	0.001	0.042455486
	60	32	0.001	0.033086145
	60	64	0.001	0.026667772
	60	128	0.001	0.027839419
NN-B4	40	16	0.01	0.015798611
	40	32	0.01	0.011401684
	40	64	0.01	0.00879857
	40	128	0.01	0.00801181
	40	16	0.001	0.018063032
	40	32	0.001	0.015245538
	40	64	0.001	0.008079771
	40	128	0.001	0.008377201

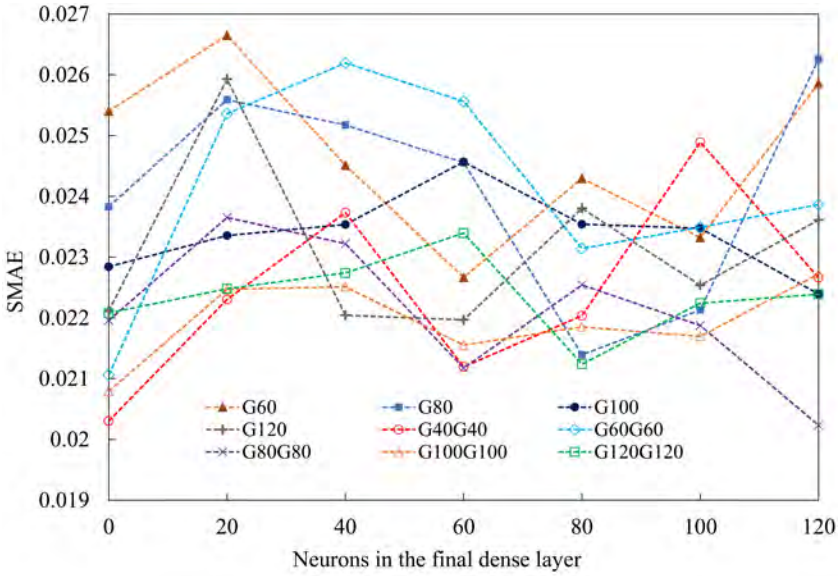


Figure B.7: SMAEs of model C for different network architectures

Table B.3: SMAEs of Model C for varied batch sizes and learning rates

Timesteps	Batch size	Learning rate	SMAE
55	16	0.01	0.022875528
55	32	0.01	0.016673629
55	64	0.01	0.017787122
55	128	0.01	0.008974653
55	256	0.01	0.020686934
55	16	0.001	0.017309264
55	32	0.001	0.018381896
55	64	0.001	0.016616734
55	128	0.001	0.021031122
55	256	0.001	0.021815552

References

Abueidda, D.W., Koric, S., Sobh, N.A., Sehitoglu, H., 2021. Deep learning for plasticity and thermo-viscoplasticity. International Journal of Plasticity 136, 102852.

Ali, U., Muhammad, W., Brahme, A., Skiba, O., Inal, K., 2019. Application of artificial neural networks in micromechanics for polycrystalline metals. International Journal of Plasticity 120, 205–219.

Anand, L., Gu, C., 2000. Granular materials: constitutive equations and strain localization. Journal of the Mechanics and Physics of Solids 48, 1701–1733.

Anandarajah, A., 2008. Multi-mechanism anisotropic model for granular materials. International Journal of Plasticity 24, 804–846.

Antony, S.J., Kuhn, M.R., 2004. Influence of particle shape on granular contact signatures and shear strength: new insights from simulations. International Journal of Solids and Structures 41, 5863–5870.

Banimahd, M., Yasrobi, S., Woodward, P.K., 2005. Artificial neural network for stress–strain behavior of sandy soils: Knowledge based verification. Computers and Geotechnics 32, 377–386.

Chang, C.S., Yin, Z.Y., 2010. Micromechanical modeling for inherent anisotropy in granular materials. Journal of engineering mechanics 136, 830–839.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 .

Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 .

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems 2, 303–314.

Di Prisco, C., Imposimato, S., Aifantis, E., 2002. A visco-plastic constitutive model for granular soils modified according to non-local and gradient approaches. International journal for numerical and analytical methods in geomechanics 26, 121–138.

Ellis, G., Yao, C., Zhao, R., Penumadu, D., 1995. Stress-strain modeling of sands using artificial neural networks. Journal of geotechnical engineering 121, 429–435.

Feng, Y., 2021a. An energy-conserving contact theory for discrete element modelling of arbitrarily shaped particles: Basic framework and general contact model. Computer Methods in Applied Mechanics and Engineering 373, 113454.

Feng, Y., 2021b. An energy-conserving contact theory for discrete element modelling of arbitrarily shaped particles: Contact volume based model and computational issues. Computer Methods in Applied Mechanics and Engineering 373, 113493.

Feng, Y., Han, K., Owen, D., 2017. A generic contact detection framework for cylindrical particles in discrete element modelling. Computer Methods in Applied Mechanics and Engineering 315, 632–651.

Fernández, M., Rezaei, S., Mianroodi, J.R., Fritzen, F., Reese, S., 2020. Application of artificial neural networks for the prediction of interface mechanics: a study on grain boundary constitutive behavior. Advanced Modeling and Simulation in Engineering Sciences 7, 1–27.

Ghaboussi, J., Sidarta, D., 1998. New nested adaptive neural networks (nann) for constitutive modeling. Computers and Geotechnics 22, 29–52.

Gong, J., Liu, J., Cui, L., 2019a. Shear behaviors of granular mixtures of gravel-shaped coarse and spherical fine particles investigated via discrete element method. Powder Technology 353, 178–194.

Gong, J., Nie, Z., Zhu, Y., Liang, Z., Wang, X., 2019b. Exploring the effects of particle shape and content of fines on the shear behavior of sand-fines mixtures via the dem. Computers and Geotechnics 106, 161–176.

Gorji, M.B., Mozaffar, M., Heidenreich, J.N., Cao, J., Mohr, D., 2020. On the potential of recurrent neural networks for modeling path dependent plasticity. Journal of the Mechanics and Physics of Solids , 103972.

Guo, N., Zhao, J., 2014. A coupled fem/dem approach for hierarchical multiscale modelling of granular media. International Journal for Numerical Methods in Engineering 99, 789–818.

Hashash, Y., Jung, S., Ghaboussi, J., 2004. Numerical implementation of a neural network based material model in finite element analysis. International Journal for numerical methods in engineering 59, 989–1005.

Hashash, Y., Marulanda, C., Ghaboussi, J., Jung, S., 2003. Systematic update of a deep excavation model using field performance data. Computers and Geotechnics 30, 477–488.

Hashash, Y.M., Marulanda, C., Ghaboussi, J., Jung, S., 2006. Novel approach to integration of numerical modeling and field observations for deep excavations. Journal of Geotechnical and Geoenvironmental Engineering 132, 1019–1031.

Hashiguchi, K., Tsutsumi, S., 2007. Gradient plasticity with the tangential-subloading surface model and the prediction of shear-band thickness of granular materials. International Journal of Plasticity 23, 767–797.

He, X., Wu, W., Wang, S., 2019. A constitutive model for granular materials with evolving contact structure and contact forces—part i: framework. Granular Matter 21, 16.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

Hornik, K., Stinchcombe, M., White, H., et al., 1989. Multilayer feedforward networks are universal approximators. Neural networks 2, 359–366.

Javadi, A., Tan, T., Zhang, M., 2003. Neural network for constitutive modelling in finite element analysis. Computer Assisted Mechanics and Engineering Sciences 10, 523–530.

Jenab, A., Sarraf, I.S., Green, D.E., Rahmaan, T., Worswick, M.J., 2016. The use of genetic algorithm and neural network to predict rate-dependent tensile flow behaviour of aa5182-o sheets. Materials & Design 94, 262–273.

Jordan, B., Gorji, M.B., Mohr, D., 2020. Neural network model describing the temperature-and rate-dependent stress-strain response of polypropylene. International Journal of Plasticity , 102811.

Jung, S., Ghaboussi, J., 2006. Neural network constitutive model for rate-dependent materials. Computers & Structures 84, 955–963.

Karapiperis, K., Stainier, L., Ortiz, M., Andrade, J., 2021. Data-driven multiscale modeling in mechanics. Journal of the Mechanics and Physics of Solids 147, 104239.

Kuhn, M.R., Daouadji, A., 2018a. Multi-directional behavior of granular materials and its relation to incremental elasto-plasticity. International Journal of Solids and Structures 152, 305–323.

Kuhn, M.R., Daouadji, A., 2018b. Quasi-static incremental behavior of granular materials: Elastic–plastic coupling and micro-scale dissipation. Journal of the Mechanics and Physics of Solids 114, 219–237.

Lai, Y., Liao, M., Hu, K., 2016. A constitutive model of frozen saline sandy soil based on energy dissipation theory. International Journal of Plasticity 78, 84–113.

Li, X., Roth, C.C., Mohr, D., 2019. Machine-learning based temperature-and rate-dependent plasticity model: application to analysis of fracture experiments on dp steel. International Journal of Plasticity 118, 320–344.

Liu, Z., Wu, C., 2019. Exploring the 3d architectures of deep material network in data-driven multiscale mechanics. Journal of the Mechanics and Physics of Solids 127, 20–46.

Mozaffar, M., Bostanabad, R., Chen, W., Ehmann, K., Cao, J., Bessa, M., 2019. Deep learning predicts path-dependent plasticity. Proceedings of the National Academy of Sciences 116, 26414–26420.

Nemat-Nasser, S., Zhang, J., 2002. Constitutive relations for cohesionless frictional granular materials. International Journal of Plasticity 18, 531–547.

Nguyen, G.D., Nguyen, C.T., Nguyen, V.P., Bui, H.H., Shen, L., 2016. A size-dependent constitutive modelling framework for localised failure analysis. Computational Mechanics 58, 257–280.

O'Sullivan, C., 2011. Particulate discrete element modelling: a geomechanics perspective. CRC Press.

Pandya, K.S., Roth, C.C., Mohr, D., 2020. Strain rate and temperature dependent fracture of aluminum alloy 7075: Experiments and neural network modeling. International Journal of Plasticity 135, 102788.

Qu, T., Feng, Y., Wang, M., Jiang, S., 2020a. Calibration of parallel bond parameters in bonded particle models via physics-informed adaptive moment optimisation. Powder Technology .

Qu, T., Feng, Y., Wang, Y., Wang, M., 2019a. Discrete element modelling of flexible membrane boundaries for triaxial tests. Computers and Geotechnics 115, 103154.

Qu, T., Feng, Y., Zhao, T., Wang, M., 2019b. Calibration of linear contact stiffnesses in discrete element models using a hybrid analytical-computational framework. Powder Technology 356, 795–807.

Qu, T., Feng, Y., Zhao, T., Wang, M., 2020b. A hybrid calibration approach to hertz-type contact parameters for discrete element models. International Journal for Numerical and Analytical Methods in Geomechanics 44, 1281–1300.

Settgast, C., Hütter, G., Kuna, M., Abendroth, M., 2020. A hybrid approach to simulate the homogenized irreversible elastic–plastic deformations and damage of foams by neural networks. International Journal of Plasticity 126, 102624.

Shaverdi, H., Taha, M., Kalantary, F., et al., 2013. Micromechanical formulation of the yield surface in the plasticity of granular materials. Journal of Applied Mathematics 2013.

Shin, H., Pande, G., 2000. On self-learning finite element codes based on monitored response of structures. Computers and Geotechnics 27, 161–178.

Sun, W., 2015. A stabilized finite element formulation for monolithic thermo-hydro-mechanical simulations at finite strain. International Journal for Numerical Methods in Engineering 103, 798–839.

Sun, W., Ostien, J.T., Salinger, A.G., 2013. A stabilized assumed deformation gradient finite element formulation for strongly coupled poromechanical simulations at finite strain. International Journal for Numerical and Analytical Methods in Geomechanics 37, 2755–2788.

Sun, Y., Gao, Y., Zhu, Q., 2018. Fractional order plasticity modelling of state-dependent behaviour of granular soils without using plastic potential. International Journal of Plasticity 102, 53–69.

Voyiadjis, G., Thiagarajan, G., Petrakis, E., 1995. Constitutive modelling for granular media using an anisotropic distortional yield model. Acta Mechanica 110, 151–171.

Voyiadjis, G.Z., Alsaleh, M.I., Alshibli, K.A., 2005. Evolving internal length scales in plastic strain localization for granular materials. International journal of plasticity 21, 2000–2024.

Wang, K., Sun, W., 2018. A multiscale multi-permeability poroplasticity model linked by recursive homogenizations and deep learning. Computer Methods in Applied Mechanics and Engineering 334, 337–380.

Wang, K., Sun, W., 2019. Meta-modeling game for deriving theory-consistent, microstructure-based traction–separation laws via deep reinforcement learning. Computer Methods in Applied Mechanics and Engineering 346, 216–241.

Wang, K., Sun, W., Du, Q., 2019. A cooperative game for automated learning of elasto-plasticity knowledge graphs and models with ai-guided experimentation. Computational Mechanics 64, 467–499.

Wang, K., Sun, W., Du, Q., 2020. A non-cooperative meta-modeling game for automated third-party calibrating, validating, and falsifying constitutive laws with parallelized adversarial attacks. arXiv preprint arXiv:2004.09392 .

Wu, L., Kilingar, N.G., Noels, L., et al., 2020. A recurrent neural network-accelerated multi-scale model for elasto-plastic heterogeneous materials subjected to random cyclic and non-proportional loading paths. Computer Methods in Applied Mechanics and Engineering 369, 113234.

Yang, Z., Li, X., Yang, J., 2008. Quantifying and modelling fabric anisotropy of granular soils. Géotechnique 58, 237–248.

Yang, Z., Liao, D., Xu, T., 2020. A hypoplastic model for granular soils incorporating anisotropic critical state theory. International Journal for Numerical and Analytical Methods in Geomechanics 44, 723–748.

Zhang, A., Mohr, D., 2020. Using neural networks to represent von mises plasticity with isotropic hardening. International Journal of Plasticity , 102732.

Zhang, Z., Li, L., Xu, Z., 2021. A thermodynamics-based hyperelastic-plastic coupled model unified for unbonded and bonded soils. International Journal of Plasticity 137, 102902.

Zhao, T., Feng, Y., 2018. Extended greenwood–williamson models for rough spheres. Journal of Applied Mechanics 85.

Zhu, H., Mehrabadi, M.M., Massoudi, M., 2006. Three-dimensional constitutive relations for granular materials based on the dilatant double shearing mechanism and the concept of fabric. International journal of plasticity 22, 826–857.

Zhu, Q., Shao, J.F., Mainguy, M., 2010. A micromechanics-based elastoplastic damage model for granular materials at low confining pressure. International Journal of Plasticity 26, 586–602.