A physics-informed neural network approach to modelling elastoplastic soils and the implicit finite-element coupling

Mingpeng Liu [a,*], Qinghua Zhang [b], Raul Fuentes [a]

[a] Institute of Geomechanics and Underground Technology, RWTH Aachen University, Germany.

[b] School of Qilu Transportation, Shandong University, China.

* Corresponding author. Email address: mingpeng.liu@rwth-aachen.de

Abstract

This study presents a physics-informed neural network (PINN) that captures the elasto-plastic behaviour of soils under complex strain and stress paths. The PINN uses the void ratio and plastic strain as the outputs of the neural networks and then includes them in a general elasto-plastic stress-strain relationship as an additional loss function. The PINN provides a more stable prediction and a superior performance than deep neural networks in forecasting the stress, strain, plastic strain, and void ratio. The PINN was then incorporated into a FEM framework as a replacement for the constitutive model to solve boundary value problems (BVPs). Three cases, biaxial test, layered-soil compression, and cavity expansion, showed that the FEM-PINN framework results in excellent agreements of deviatoric stress, plastic shear strain, and void ratio compared to state-of-the-art numerical methods as a benchmark. The approach captures the stress-history memory of soils as well as the volumetric and shear response of soil, including contractive and dilative behaviour. The stress concentration and strain localization characteristics were also reproduced by the FEM-PINN framework in the BVPs. Critically this approach also achieves outstanding performance in plasticity without the need to define plastic yield functions or hardening rules. Added to the comparable accuracy, the PINN is significantly better from a computational effort point of view.

Keywords: Physics-informed neural network; Elastoplastic soil; Constitutive modelling; Finite element-neural network coupling; Boundary value problems

1. Introduction

Soil presents complex mechanical behaviours, such as path dependency, stress dilatancy, state variables-dependency, and critical state characteristics, to name but a few [1]. In an attempt to predict soil behaviour, constitutive models are usually used, e.g., the Modified Cam-clay (MCC) model [2] and the Clay And Sand Model (CASM) [3]. Among those developed models in the past, elastoplastic models coupled with the critical state concept [2] achieved great popularity [4-6] due to their comprehensive description of soil. Additionally, they follow a well-understood philosophy and methodology that can be explained in relatively simple terms mathematically and this enables seamless implementation in numerical frameworks. Among the many numerical frameworks, the finite element method (FEM) combined with constitutive models remains the most widely spread tool to predict soil behaviour and therefore, for the design of engineering problems [7-9].

Over the past few years, the successful application of neural networks, also commonly referred to Deep Learning (DL), has been an alternative to phenomenological models [10, 11]. Based on that, some DL models have been recently developed to represent the mechanical behaviours of soil by directly learning the stress-strain relationships from simulations or measurements [12, 13]. These DL models eliminate in principle the requirement for any idealizations of material mechanics or calibration of constitutive parameters [14]. The neural network can replace conventional constitutive modelling, including the potential to learn complex path-dependent behaviour if appropriate state variables are used [13]. For all of the above reasons, DL models have been recently proposed as surrogate models and used in FEM. Zhang et al. [15-17] investigated the application of the long short-term memory (LSTM) neural network in modelling complex history-dependent soil constitutive behaviours as well as incorporating it into finite elements for the boundary value problems (BVPs) of biaxial tests and foundations with internal erosion. Guan et al. [18, 19] successfully applied several neural networks to learn the stress-strain relationship and to model the BVPs of cavity expansion and strip footing. However, such pure DL models also come with some drawbacks. They usually require a large amount of training data and have poor extrapolation ability [12]. This may be rooted in the

lack of proper physical constraining of the output [10, 12]. In addition, such DL models usually take only the stress and strain variables as an output of neural networks so as to improve prediction accuracy, but discard some important information in soil, e.g., plastic strain or void ratio [20]. The lack of physical information in DL models has hindered their application.

These drawbacks of DL models have led some scholars to develop physical-based neural networks that enforce the outputs to comply with a framework of physical equations [16, 21], Physical-Informed Neural Network (PINN) can theoretically be applied to any given physical process described by partial differential equations (PDEs) [22]. The PINN endows the outputs of neural networks with a clear physical meaning and constrains the outputs into a framework of physical relationships, thereby improving the prediction accuracy of models [23, 24]. Recently, some scholars extended the restriction of PDEs in PINN to more general respecting physical-based laws, including physic-based postulates in constitutive modelling that are not necessarily expressed in terms of PDEs [10, 12, 24]. In fact, PINN is applied to solid mechanics [21], fluid mechanics [25], thermodynamics [26], and elastoplastic mechanics [10], etc. Vlassis and Sun [27, 28] deployed the derivative of the network outputs as the loss function to incorporate a yield surface and flow rule for modelling conventional elasto-plastic materials. Xu et al. [29] proposed using the tangent elasto-plastic matrix to constrain the outputs of the neural network into classical elasto-plastic constitutive relationships. With the re-constructed tangent matrix, the neural network was successfully and effectively used to solve the boundary value problem of thin plate tests. Su et al. [30] used the plastic strain as the output and constrained it into elastoplastic constitutive modelling with a thermodynamics-informed neural network.

However, many of these PINN models ignore internal state variables [10], such as void ratio and plastic strain, even though they are key in the calculation of stress and constitutive relationships. This results in inaccurate prediction of these variables and overall performance in BVPs. Additionally, neural networks usually perform not so well in forecasting soil internal state variables due to the lack of constraining these soil properties into a physical relationship. Besides this, when coupling with

FEM, these soil internal variables should participate in the formation of a stiffness matrix. Finally, other elements such as stress history memory, stress reversals, and dilative, compressive behaviour during and before shear band formation are not considered jointly.

To fill the above gaps, this work focuses on developing a Physical-Informed Neural Network to model soil elastoplastic behaviour, in which a general stress-strain relationship is used to constrain the outputs of the neural network approximation. Critically, we consider internal state variables (void ratio and plastic strain) as part of the process to prove whether a more universal solution to soil behaviour prediction can be achieved. The hypothesis is that this will improve the accuracy of the networks whilst directly recovering the soil's physical information. Section 2 introduces the PINN-based constitutive modelling framework. Section 3 illustrates how to develop the PINN, including the data generation, network architecture, and loss function. Section 4 shows the validation of the developed model and Section 5 incorporates the PINN into the FEM framework to enhance its practical application whilst comparing it to boundary value problems.

2. Methodology

2.1 Strain-controlled neural network-based constitutive modelling

For a strain-controlled constitutive model, the current stress state is dependent on the previous stress and strain history, and soil internal state variables (e.g., void ratio) [12, 15]. Following this logic, a neural network-based modelling strategy shown in Fig. 1 can be described as:

$$\left(\boldsymbol{\sigma}^{i+1}, J^{i+1}\right) = \mathcal{NN}\left(\boldsymbol{\sigma}^{i}, \boldsymbol{\varepsilon}^{i}, \Delta\boldsymbol{\varepsilon}^{i}, J^{i}\right) \tag{1}$$

where $\boldsymbol{\sigma}^{i}$, $\boldsymbol{\varepsilon}^{i}$, $\Delta\boldsymbol{\varepsilon}^{i}$, and $\boldsymbol{J}^{i}$ represent the stress tensor, strain tensor, strain increment tensor, and soil internal state variables at step i, respectively; $\boldsymbol{\sigma}^{i+1}$ and $\boldsymbol{J}^{i+1}$ are the stress tensor and internal state variables at step i+1. $\mathcal{NN}$ represents the neural network. Many deep neural networks possessing a highly nonlinear mapping ability between inputs and outputs have been developed to capture constitutive relationships, such as the multi-layer perceptron (MLP) [10, 12], the long short-term memory (LSTM) neural network [15, 31], or some hybrid networks [11].

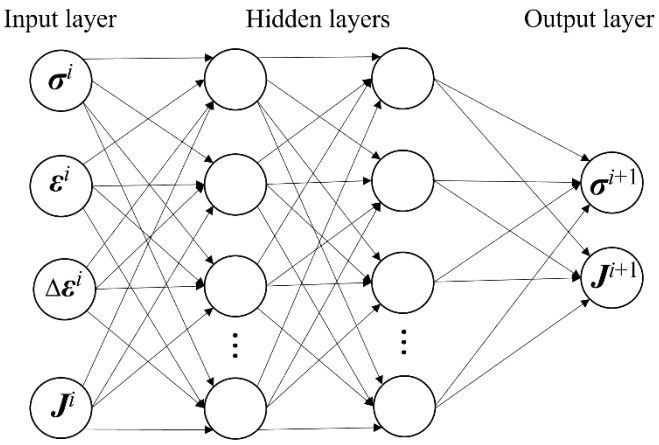


Fig. 1. Schema of neural network-based constitutive modelling.

2.2 Classical elasto-plastic model

The elasto-plastic theory is written in the incremental form in this paper as it matches the finite element method. Traditional elastoplastic models decompose the total strain into elastic and plastic components and can be expressed in Eq. (2):

$$\Delta\boldsymbol{\varepsilon} = \Delta\boldsymbol{\varepsilon}_e + \Delta\boldsymbol{\varepsilon}_p \tag{2}$$

where $\Delta\boldsymbol{\varepsilon}$, $\Delta\boldsymbol{\varepsilon}_e$, and $\Delta\boldsymbol{\varepsilon}_p$ denote the total, elastic, and plastic strain increment tensor, respectively. The elastic strain can be solved using the simple general elasticity theory. However, the specification of plastic strains requires further complex calculations. Eq. (3) describes the conventional elasto-plastic constitutive model [12]:

$$\Delta\boldsymbol{\sigma} = \boldsymbol{D}_e : \left(\Delta\boldsymbol{\varepsilon} - \Delta\boldsymbol{\varepsilon}_p\right) \tag{3-1}$$

$$F(\boldsymbol{\sigma}, \mathcal{H}) \leq 0 \tag{3-2}$$

$$\Delta\boldsymbol{\varepsilon}_p = \mathrm{d}\lambda \frac{\partial g(\boldsymbol{\sigma}, \mathcal{H})}{\partial \boldsymbol{\sigma}} \tag{3-3}$$

$$\mathrm{d}\mathcal{H} = H(\mathcal{H}, \boldsymbol{\sigma}, \boldsymbol{\varepsilon}, \Delta\boldsymbol{\sigma}, \Delta\boldsymbol{\varepsilon}, \boldsymbol{J}) \tag{3-4}$$

Eq. (3-1) describes the general stress-strain relationship, in which $\boldsymbol{\sigma}$ represents the stress tensor; $\boldsymbol{D}_e$ is the elastic stiffness matrix and $F(\boldsymbol{\sigma}, \mathcal{H})$ is the yield function or loading surface in which $\mathcal{H}$ are the hardening variables. Eq. (3-3) is the flow rule, where g denotes the plastic potential function and $\mathrm{d}\lambda$ is the plastic multiplier. Eq. (3-4) is the hardening law.

The elastic stiffness matrix $\boldsymbol{D}_e$ can be resolved by the general elasticity theory, as shown in Eq. (4):

$$\boldsymbol{D}_e = \left(K - \frac{2}{3}G\right)\mathbf{I}\otimes\mathbf{I} + 2G\boldsymbol{\Pi} \tag{4}$$

where K and G are the elastic bulk and shear moduli, respectively; $\mathbf{I}$ and $\boldsymbol{\Pi}$ are second and fourth-order identity tensors; $\otimes$ denotes the dyadic product. K and G can be obtained from soil properties [3]:

$$K = \frac{1+e}{\kappa}p \tag{5-1}$$

$$G = \frac{3(1-2\mu)}{2(1+\mu)}K \tag{5-2}$$

where μ is the Poisson's ratio; κ is the slope of the swelling line; e is the void ratio; p is the mean effective stress defined as $p = \mathrm{tr}(\boldsymbol{\sigma})/3$.

2.3 Physical-informed neural network for elasto-plastic soil

In this paper, we aim to develop a physics-informed model to describe the elasto-plastic information of soil. For this purpose, we designed the framework shown in Fig. 2. Inspired by Eghbalian et al [12], we use the void ratio e and plastic strain tensor $\boldsymbol{\varepsilon}_p$ as the internal state variables, together with the stress as the output. These parameters with the knowledge of other variables such as μ and κ allow calculation of all variables to fully characterize soil behaviour. The advantages of using e and $\boldsymbol{\varepsilon}_p$ as the output of neural networks are 1) no need to define analytical expressions of yield surface or plastic potential functions; 2) the output of networks can directly form the material stiffness matrix; 3) e and $\boldsymbol{\varepsilon}_p$ are necessary for the expression of the stress-strain relationship, which is further used as a loss function.

In Fig. 2, the main body of the neural network is chosen as MLP as recommended by Haghighat et al. [10] and Eghbalian et al. [12]. Details of the MLP approach are shown in Appendix A. Accordingly, with the current step i of stress, strain, strain increment, and internal state variables ($\boldsymbol{\sigma}^i$, $\boldsymbol{\varepsilon}^i$, $\Delta\boldsymbol{\varepsilon}^i$, $\boldsymbol{J}^i$), the MLP is employed to predict the next step i+1 of stress and internal state variables ($\boldsymbol{\sigma}^{i+1}$, J^{i+1}). Strain increment information should be given specifically to determine the loading path.

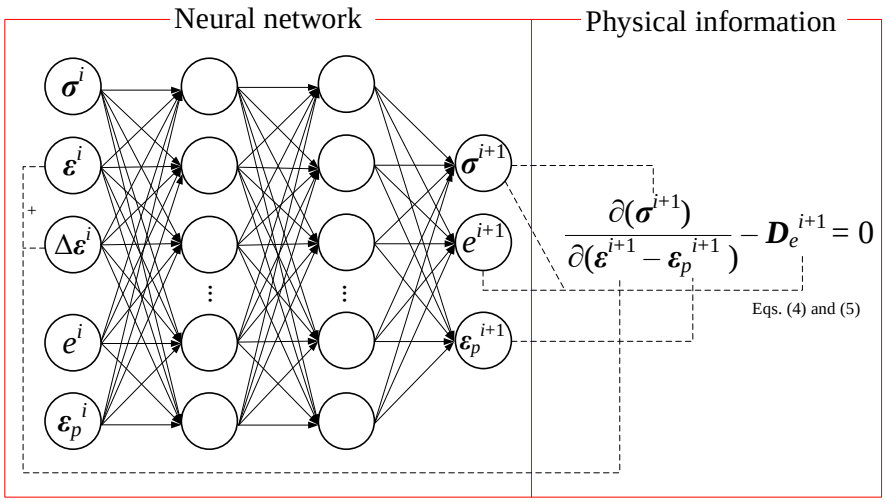


Fig. 2. The PINN for elasto-plastic soil.

The problem's physics is introduced in the network by using a PDE as part of the error function, defined from Eq. (3-1):

$$\frac{\partial(\boldsymbol{\sigma}^{i+1})}{\partial(\boldsymbol{\varepsilon}^{i+1}-\boldsymbol{\varepsilon}_p^{i+1})} - \boldsymbol{D}_e^{i+1} = 0 \tag{6}$$

where $\boldsymbol{D}_e{}^{i+1}$ is the elastic stiffness matrix at the i+1 step. The stress and plastic strain tensor in Eq. (6) are the predicted output of the neural network, while $\boldsymbol{D}_e{}^{i+1}$ can be further calculated using Eqs. (4) and (5). By Eq. (6), the predicted void ratio, stress, and plastic strain are constrained into an elasto-plastic stress-strain relationship and thereby improve the prediction precision for physical information.

Hence, the loss function $\mathcal{L}$ has two parts – see Eq. (7).

$$\mathcal{L} = \mathcal{L}_1\left[\left(\boldsymbol{\sigma}, e, \boldsymbol{\varepsilon}_p\right)_{\text{pred}}, \left(\boldsymbol{\sigma}, e, \boldsymbol{\varepsilon}_p\right)_{\text{real}}\right] + \mathcal{L}_2\left[\frac{\partial(\boldsymbol{\sigma}^{i+1})}{\partial(\boldsymbol{\varepsilon}^i-\boldsymbol{\varepsilon}_p^i)} - \boldsymbol{D}_e^{i+1}, 0\right] \tag{7-1}$$

$$\mathcal{L}_1\left[y_{\text{pred}}, y_{\text{real}}\right] = \text{MSE} = \frac{1}{N_T}\sum_{i=1}^{N_T}\left(y_{\text{pred}} - y_{\text{real}}\right)^2 \tag{7-2}$$

$$\mathcal{L}_2[\circ, 0] = \|\circ\|_2 \tag{7-3}$$

where $\mathcal{L}_1$ represents the error between the predicted output compared to the true values, and $\mathcal{L}_2$ denotes the error of the PDE. Subscripts "pred" and "real" refer to the predicted values of the neural network and real values. The $\mathcal{L}_1$ loss function uses the mean squared error (MSE) and N_T is the number of sample values. The $\mathcal{L}_2$ loss function tries to reduce the L_2 norm ($\|\circ\|_2$) to zero as an indication of the error of the PDE - see Eq. (7-3).

3. Development of PINN

3.1 Data generation and processing

3.1.1 Bounding surface model

To test the ability of our approach, we selected a constitutive model, the bounding surface model (BSM), capable of capturing the very complex behaviour of soils. The model uses seven material parameters: λ, κ, μ, M, N, Γ, and n. The parameter definition and detailed introduction of BSM is given in Appendix B. Some of the complex behaviours, that this model can capture, is that plasticity can occur before reaching the yield surface, in which case plastic strain would emerge in every loading step [32]. Additionally, BSM can also extend to general soils including clay and sand, which also makes it attractive for its generalization ability.

The adopted constitutive parameters of the bounding surface model are summarized in Table 1. The initialization parameters used are pre-consolidation pressure p_c, initial void ratio $e_{0,}$ and initial stress state p_0. We set the value of p_c to 100 kPa and e_0 to 0.85, respectively. The p_0 is set to a confining pressure of 100 kPa to model a normally consolidated soil.

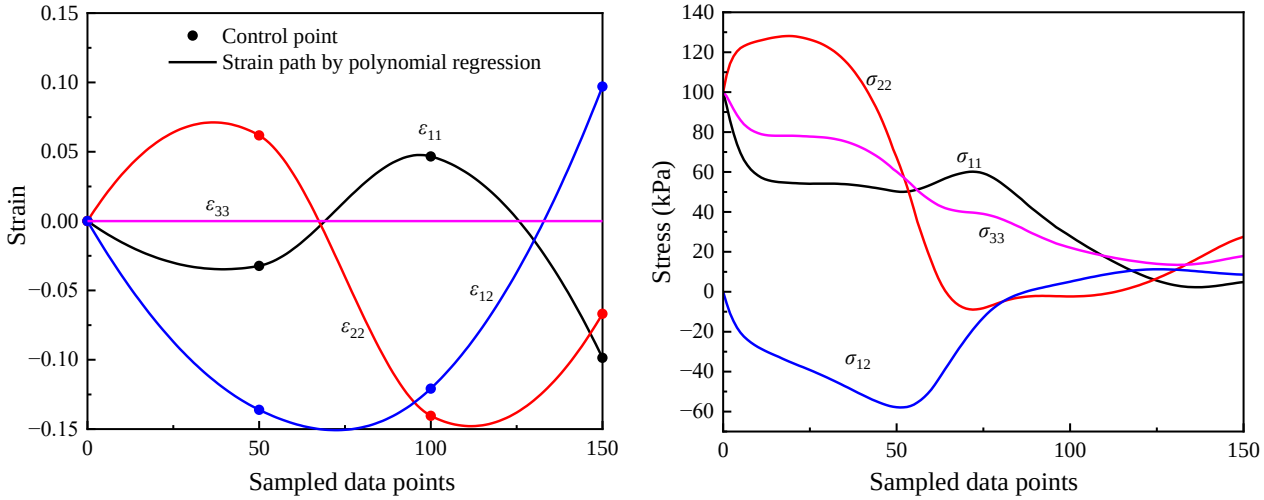
Table 1. Parameters of the bounding surface model.

λ	κ	μ	M	N	Γ	n
0.15	0.05	0.3	1.2	2.69	2.62	1.6

3.1.2 Random loading path

To consider complex strain-loading paths that the soil may experience, we use the representative volume elements (RVEs) to generate strain paths since RVEs can map the properties of soils to the FEM domain. To demonstrate the superiority of the PINN in processing complex stress-strain data and capturing the inherent elastoplastic relationship, the random strain-loading path proposed by Mozaffar et al. [13] is employed to generate training data in this study, which is also used by Zhang et al. [15] and He et al. [33]. For convenience and efficiency, we degrade the real three-dimensional problems in RVEs to plane strain conditions, in which case the strain and stress components are (ε_{11}, ε_{22}, ε_{33}, ε_{12}) and (σ_{11}, σ_{22}, σ_{33}, σ_{12}) with ε_{33} equals to zero. Next, we display how to generate the

complex strain paths. Considering in all cases the soil starts from the initial state before loading, the strain path should start from $\varepsilon_{11} = \varepsilon_{22} = \varepsilon_{12} = 0$, as shown in Fig. 3(a). Then, the strain path will evolve in the next 150 loading steps. For one strain path, three control points are set at 50, 100, and 150 loading steps and their strain values are randomly taken from the range of [-0.15, 0.15]. The strain-loading path is subsequently constructed by interpolating between control points using polynomial regression, and the same method to strain components ε_{11}, ε_{22}, and ε_{12}, as shown in Fig. 3(a). Correspondingly, the BSM can calculate the stress path (σ_{11}, σ_{22}, σ_{33}, σ_{12}), as displayed in Fig. 3(b). The stress-strain path presents a complex case, including even strain direction reversals, which helps to test the robustness and generalization of PINN.



(a) (b)

Fig. 3. Data sampling strategy for plane strain condition (ε_{33}=0): (a) random strain-loading path generated (dot denote control points); (b) corresponding stress path.

3.1.3 Data processing

With the BSM, 20 strain-loading paths are generated for training and testing the neural network. The generated samples are displayed in Fig. 4, where the stress and strain are shown by the spherical and deviatoric tensors. As can be observed in Fig. 4, adopting a random strain path guarantees that a wide range of volumetric and shear strains, as well as their increments, are included. Most of the generated volumetric strain as well as its increment are mainly distributed around zero, as shown in Figs. 4(b), (c), (d), and (e). Similarly, the generated mean effective stress p is mainly distributed around the initial confining pressure. The shear strain, its increment, and its plastic components have a wide range

because of the continued change of strain paths.

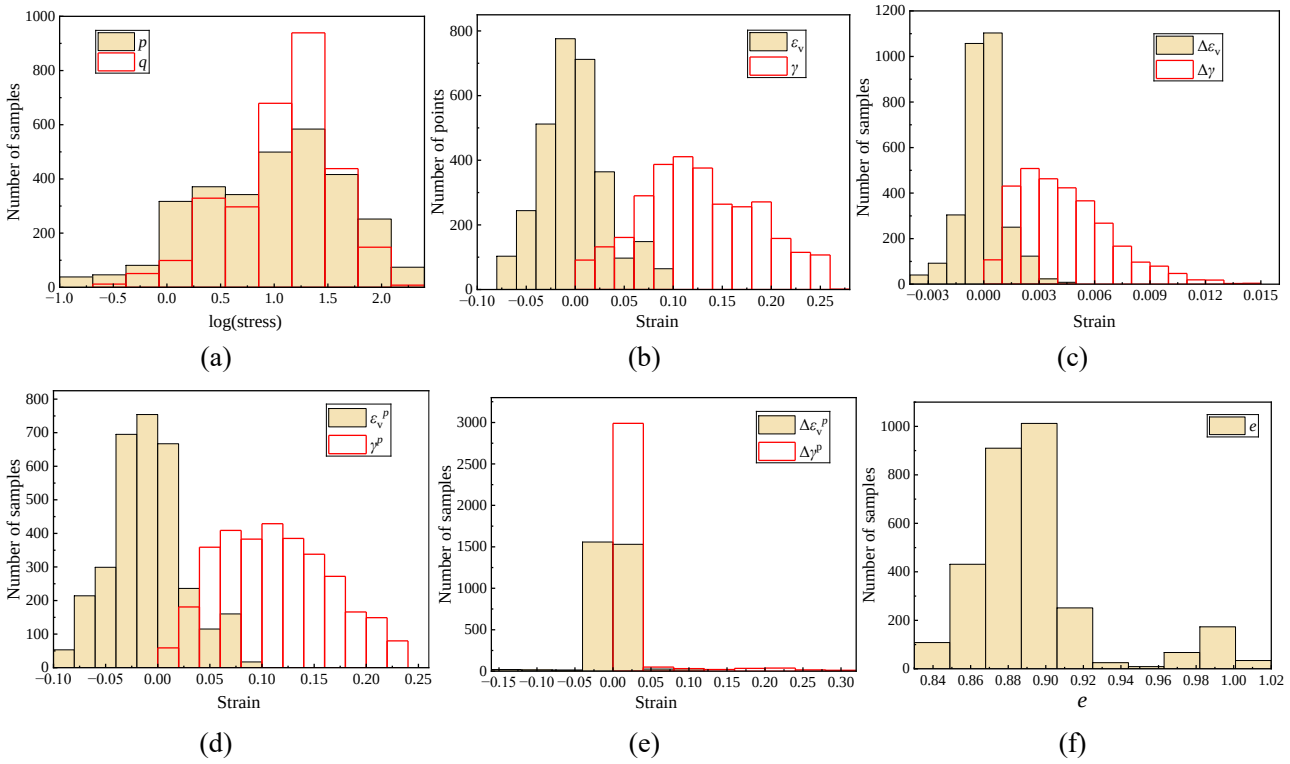


Fig. 4. Distribution of generated samples. (a) mean effective stress p and deviatoric stress q; (b) volumetric strain ε_v and shear strain γ; (c) volumetric strain increment $\Delta\varepsilon_v$ and shear strain increment $\Delta\gamma$; (d) plastic volumetric strain ε_v^p and plastic shear strain γ^p; (e) plastic volumetric strain increment $\Delta\varepsilon_v^p$ and plastic shear strain increment $\Delta\gamma^p$; (f) void ratio e.

Normalization is preferably adopted to rescale variables with different scales to lower their influence on the model performance and reduce computational costs. The following equation was used to normalize the input data to the common range of 0 ~ 1:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{9}$$

where x is the raw input variables before normalization, x_{norm} is the input variables after normalization, x_{min} and x_{max} are the minimum and maximum values of the input variables, respectively.

3.2 Configurations and hyperparameters of PINN

The configurations and hyperparameters of neural networks should be determined, including the number of layers, the number of neurons in each layer, and the learning rate. For comparing the

performance of PINNs with MLP, we varied the number of neurons in three-hidden-layer neural networks, as shown in Table 2, while other hyperparameters are kept the same. The Relu function is adopted as the activation function for hidden layers as it can well address the vanishing and exploding gradient problem [34]. A widely used learning rate is set to be 0.001 [12, 15]. The adaptive moment estimation (Adam) optimizer is utilized since it makes the neural networks converge faster and better handle sparse gradients [34, 35]. The number of epochs is set to 1000 to ensure sufficient training. The input layer has 15 neurons corresponding to 1 for void ratio, 4 for stress components, 3 for strain components, 3 for strain increment components, and 4 for plastic strain components. The output layer has 9 neurons, of which 1 for void ratio, 4 for stress components, and 4 for plastic strain components. The loss function of MLP is the MSE in Eq. (7-2), while the PINN uses the custom loss function Eq. (7-1).

All produced data are firstly divided into training and testing datasets using the split ratio of 80% : 20%. Besides, the training data is shuffled in each epoch during training to enhance the robustness of neural networks. Additionally, two indexes: the coefficient of determination (R^2) and mean absolute percentage error (MAPE) are applied to evaluate the performance of the developed models:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{\text{real}} - y_{\text{pred}}\right)^2}{\sum_{i=1}^{n}\left(y_{\text{real}} - \overline{y}_{\text{real}}\right)^2} \tag{10}$$

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_{\text{pred}} - y_{\text{real}}}{y_{\text{real}}}\right| \times 100\% \tag{11}$$

where $\overline{y}_{\text{real}}$ is the mean real value and other parameters are the same as in Eq. (7-2).

Table 2. Architecture and hyperparameters of neural networks

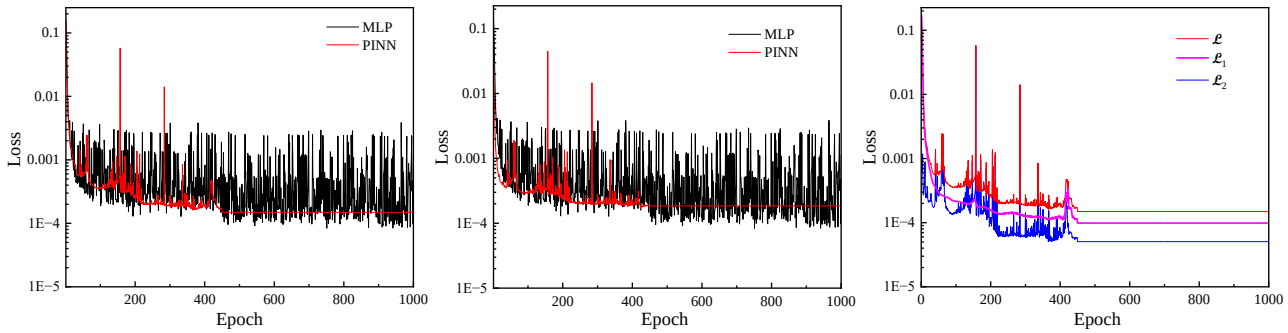
Model	Architecture	f	Learning rate	Loss function	Others
PINN	15-40-40-40-9 15-60-60-60-9 15-80-80-80-9 15-100-100-100-9	Relu	0.001	Eqs. (7-1) to (7-3)	Optimizer = Adam Batch size = 128 Epoch = 1000
MLP	15-40-40-40-9 15-60-60-60-9 15-80-80-80-9 15-100-100-100-9	Relu	0.001	Eq. (7-2)	

Note: 15-40-40-40-9, for example, denotes the neural network including an input layer with 15

variables; three hidden layers with 40 neurons in each hidden layer; and an output layer with 15 variables. *f* denotes the activation function.

3.3 Training of PINN

This section presents the training results of neural networks. Figs. 5(a) and (b) depict the evolution of the loss function for MLP and PINN in training and testing sets, respectively, taking the network architecture "15-80-80-80-9" in Table 2 as an example. The training curve of traditional MLP presents obvious fluctuation, while the loss function of PINN converges after 430 epochs and achieves a lower average loss value. More distinctly, the loss function in the testing set generated by the PINN is also lower than that by the MLP, thus revealing the improved generalization ability of PINN. Fig. 5(c) displays the evolution of the custom loss function Eq. (7) as well as its decomposition. The fluctuation of the total loss function is mainly due to $\mathcal{L}_2$ (the error of the PDE), while the $\mathcal{L}_1$ part (the error between real values and predictions) achieves great improvement in convergence rate and stability. Since $\mathcal{L}_1$ is the same loss function as used in MLP, the $\mathcal{L}_1$ part loss reduces from 2.66e-4 for MLP to 9.86e-5 for PINN, corresponding to a 62.9% improvement. Accordingly, we can conclude that the loss function is reduced by applying the physical constraint to the output of the neural network.



(a) (b) (c)

Fig. 5. Comparison of the loss function during the training process in (a) training set, (b) testing set, and (c) decomposition of Eq. (7).

4. Validation of the developed model

4.1 Comparison between different networks

For the sake of investigating the optimal network architecture, Fig. 6 displays the performance of

MLP and PINN for different hidden-layer neurons in terms of R^2 and MAPE. The R^2 for the PINN in both training and testing sets are all higher than those from MLP. Additionally, the hidden layers with 80 neurons reach the highest R^2 for the testing set. Generally, more neurons in layers could improve the nonlinear prediction ability but have a higher risk of overfitting. In this study, 100 neurons reduced the R^2 value compared to 80 neurons, illustrating a possible overfitting phenomenon. Similarly, the MAPE for the PINN in training and testing datasets are all lower than the MLP. Besides, the MAPE with 80 neurons achieves the lowest value in the testing set for both MLP and PINN, showing a better generalization ability. The network architecture is determined to be the hidden layers with 80 neurons as it provides the best accuracy with the lowest overfitting potential.

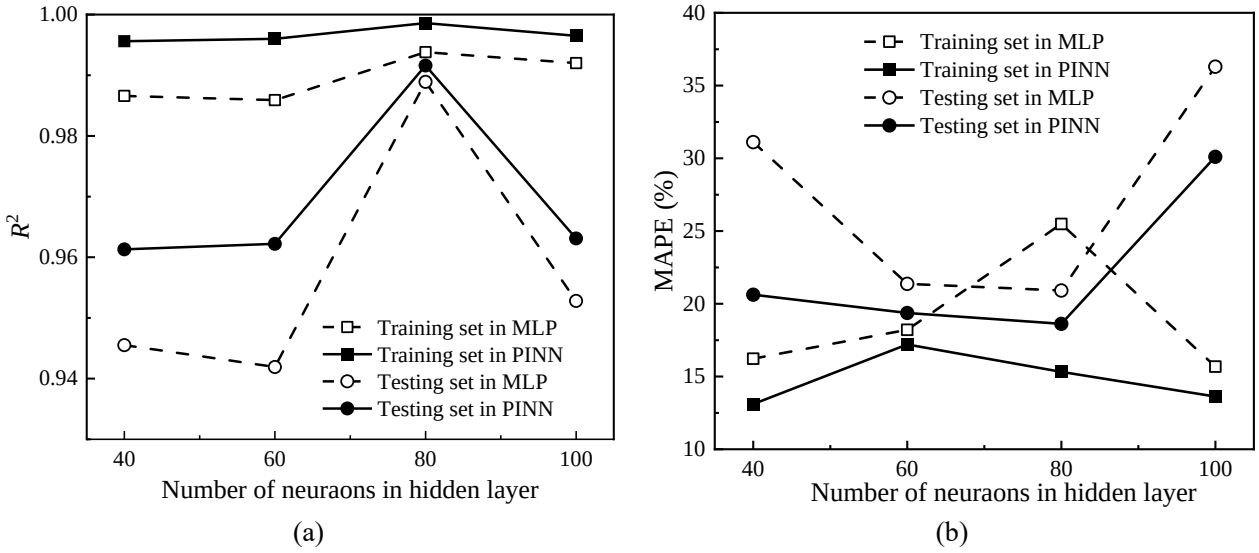


Fig. 6. Comparison of (a) R^2 and (b) MAPE for different network architectures in training and testing sets.

After choosing the optimal network, the R^2 and MAPE for output variables are shown in Fig. 7(a) with respect to void ratio, stress tensor, and plastic strain tensor. The PINN for each group variable still performs better than the MLP with higher R^2 and lower MAPE in both training and testing sets. The stress performs the worst, followed by plastic strain, especially in terms of MAPE compared to the void ratio. The reason is thought to be that both tensors have four components while the void ratio has only one, resulting in the neural network being harder to capture all components. To delve deeper into this, the R^2 of each stress and plastic strain component in the PINN result is shown in Fig. 7(b). All stress and plastic strain components present a similar and very high R^2 value. The σ_{33} and $(\varepsilon_p)_{33}$ perform a little bit worse than other components but very marginally.

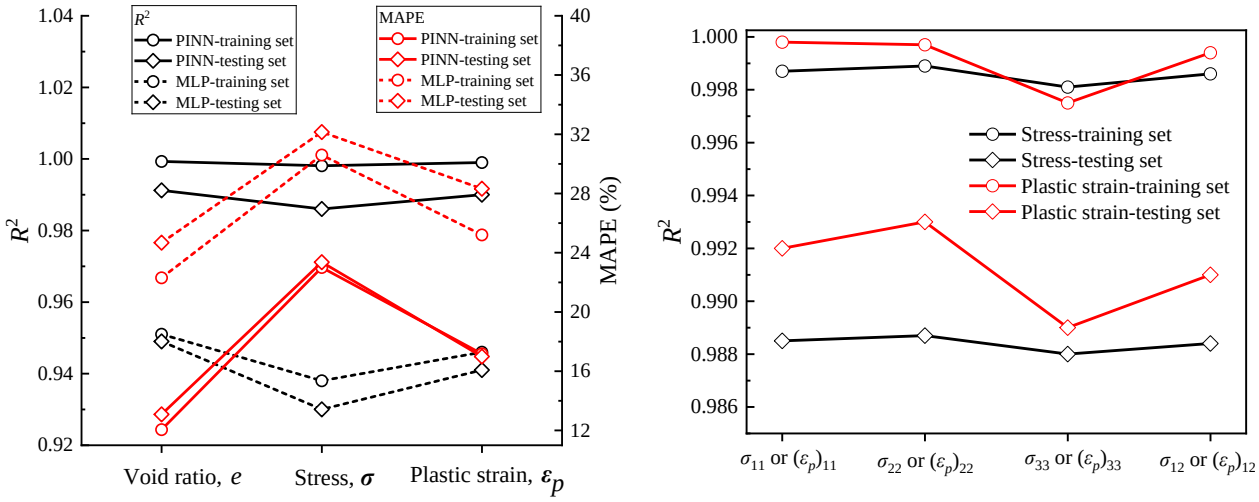


Fig. 7. Comparison of R^2 and MAPE for (a) different group variables, and (b) different components of stress and plastic strain in the PINN result.

The performance of the selected network architecture for a variation in the size of the training dataset is presented in Fig. 8(a) for R^2 and Fig. 8(b) for MAPE, determined for the testing set. The R^2 basically increases with the used data samples while the MAPE decreases, signifying a higher fitting degree and less error with more data used in the training process. Meanwhile, for all the sizes of training samples, the R^2 for PINN is always higher than that for MLP, while the MAPE for PINN is lower than MLP. However, as expected the difference tends to become smaller as the size of the dataset increases. This confirms what we know: that PINNs assist in making better predictions in the absence of too much data. Interestingly, the performance of PINN degrades slightly when the training samples reach around 2100 (about the 14th loading path). This phenomenon was also discovered by Zhang et al. [15] and may be because this loading path has less points to participate in the training since 20% of points are always used for testing, causing a reduction in R^2 regarding this path. Besides, when the number of training samples reaches about 3000 (this paper used), the R^2 does not increase and MAPE does not decrease obviously. Therefore, the optimal sample size can be regarded as 3000 as no more improvements are observed.

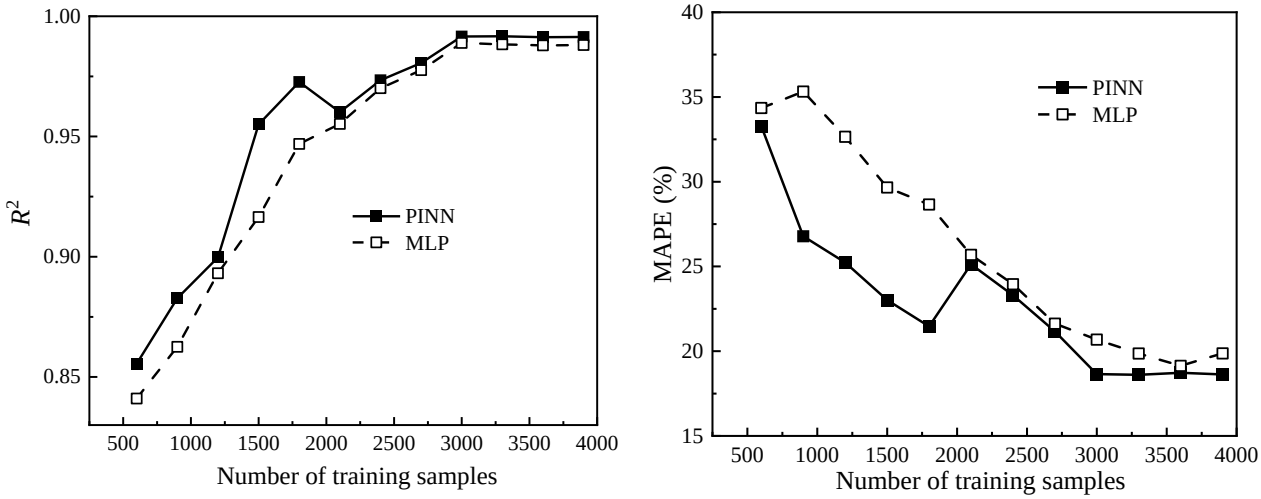


Fig. 8. Performance of PINN and MLP with 80-neurons hidden layers as a function of the size of training dataset: (a) R^2, and (b) MAPE for the testing set.

4.2 Prediction on complete loading paths

Next, the recall mode is employed to predict a complete loading path, as illustrated in Algorithm 1. Herein, the trained network is called multiple times. The outputs of the current step are used as the inputs of the next step, in such a way to update the constitutive relationships, including stress, strain, plastic strain, and void ratio. As the algorithm suggests, using the neural network to predict the constitutive response is significantly easier to implement than the implicit integration required in classical elasto-plastic theory [36].

Algorithm 1 Recall mode to predict a strain-loading path

Require: Pre-trained network, number of loading steps (nstep)

Initialization: $\boldsymbol{\sigma}^0$, $\boldsymbol{\varepsilon}^0$, e^0, $\boldsymbol{\varepsilon}_p{}^0$

for i = 0, nstep **do**

Prescribe the strain increment tensor $\Delta\boldsymbol{\varepsilon}^i$

Calculate the output of the neural network $\left(\boldsymbol{\sigma}^{i+1}, e^{i+1}, \boldsymbol{\varepsilon}_p^{i+1}\right) = \mathcal{NN}\left(\boldsymbol{\sigma}^i, \boldsymbol{\varepsilon}^i, \Delta\boldsymbol{\varepsilon}^i, e^i, \boldsymbol{\varepsilon}_p^i\right)$

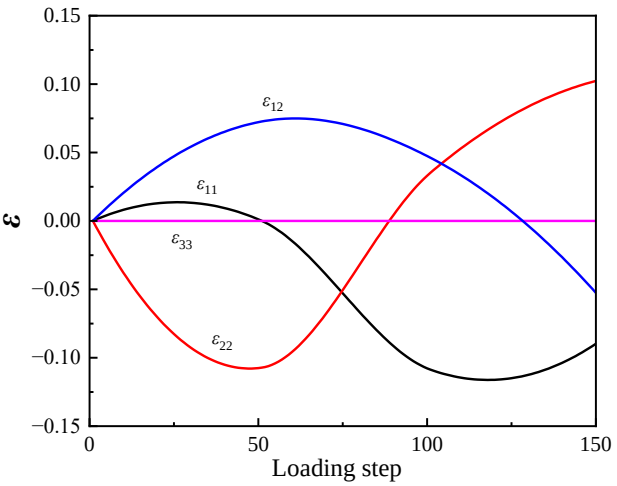
Update the state variables $\boldsymbol{\sigma}^i = \boldsymbol{\sigma}^{i+1}$, $\boldsymbol{\varepsilon}^i = \boldsymbol{\varepsilon}^i + \Delta\boldsymbol{\varepsilon}^i$, $e^i = e^{i+1}$, $\boldsymbol{\varepsilon}_p{}^i = \boldsymbol{\varepsilon}_p{}^{i+1}$

end for

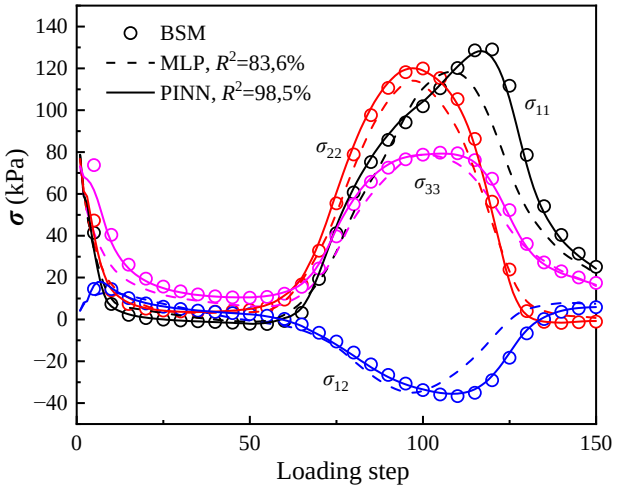
By way of example, Fig. 9 displays a better prediction of constitutive response by PINN for a complete strain-loading path. Fig. 9(b) presents the evolution of stress, in which PINN shows

excellent agreement with the generated data using the BSM and outperforms the results by MLP, which underestimates the stress components. In addition, Fig, 9(c) shows that PINN also achieves great prediction of the plastic strain and its evolution. The predicted void ratio in Fig. 9(d) shows the PINN also achieves a better agreement with the BSM, while the MLP greatly overestimates the value of the void ratio in the evolution process. In conclusion, the elastoplastic physical constraint for the output of the neural network improves the prediction accuracy significantly.

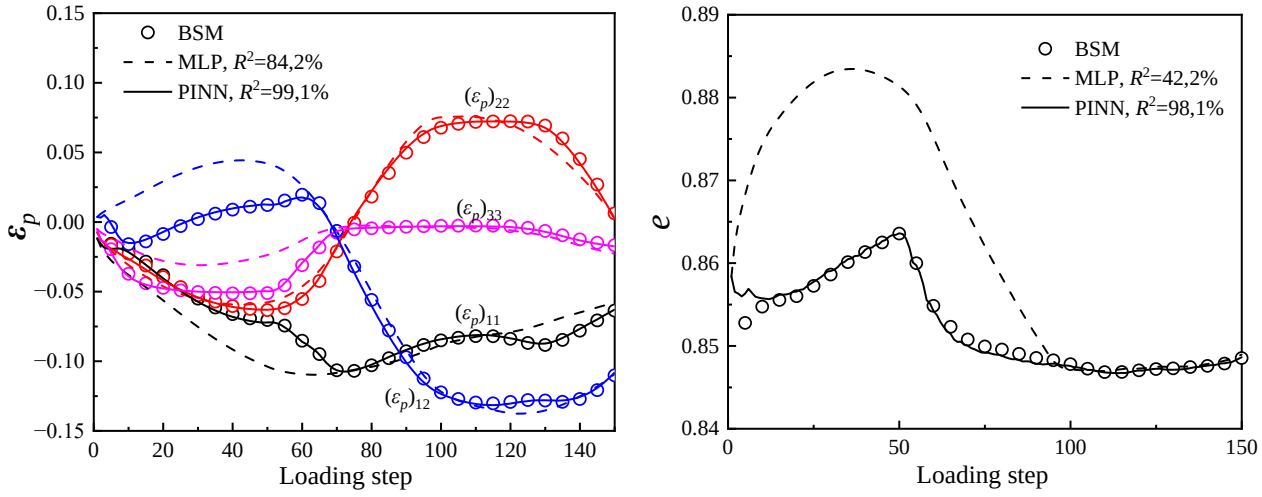
Fig. 10 provides a case of PINN predictions where both the stress and strain components deviate slightly from the bounding surface model. Whilst not perfect, the performance of PINN is still better than MLP, and the PINN can also successfully capture the evolution tendency of stress and strain. The reason for the worse prediction case is found to be that we randomly take 80% of total data generated from all loading paths to train the network, so inevitably some loading paths have less points participating in the training, corresponding to a worse prediction on the complete stress and strain paths. Besides, although the R^2 in training and testing sets have reached a very high value, there are still some discrepancies in predicting the whole loading path. This is because the recall mode uses the output in the last step instead of the real value as the input in the current step, thereby causing the error accumulation.



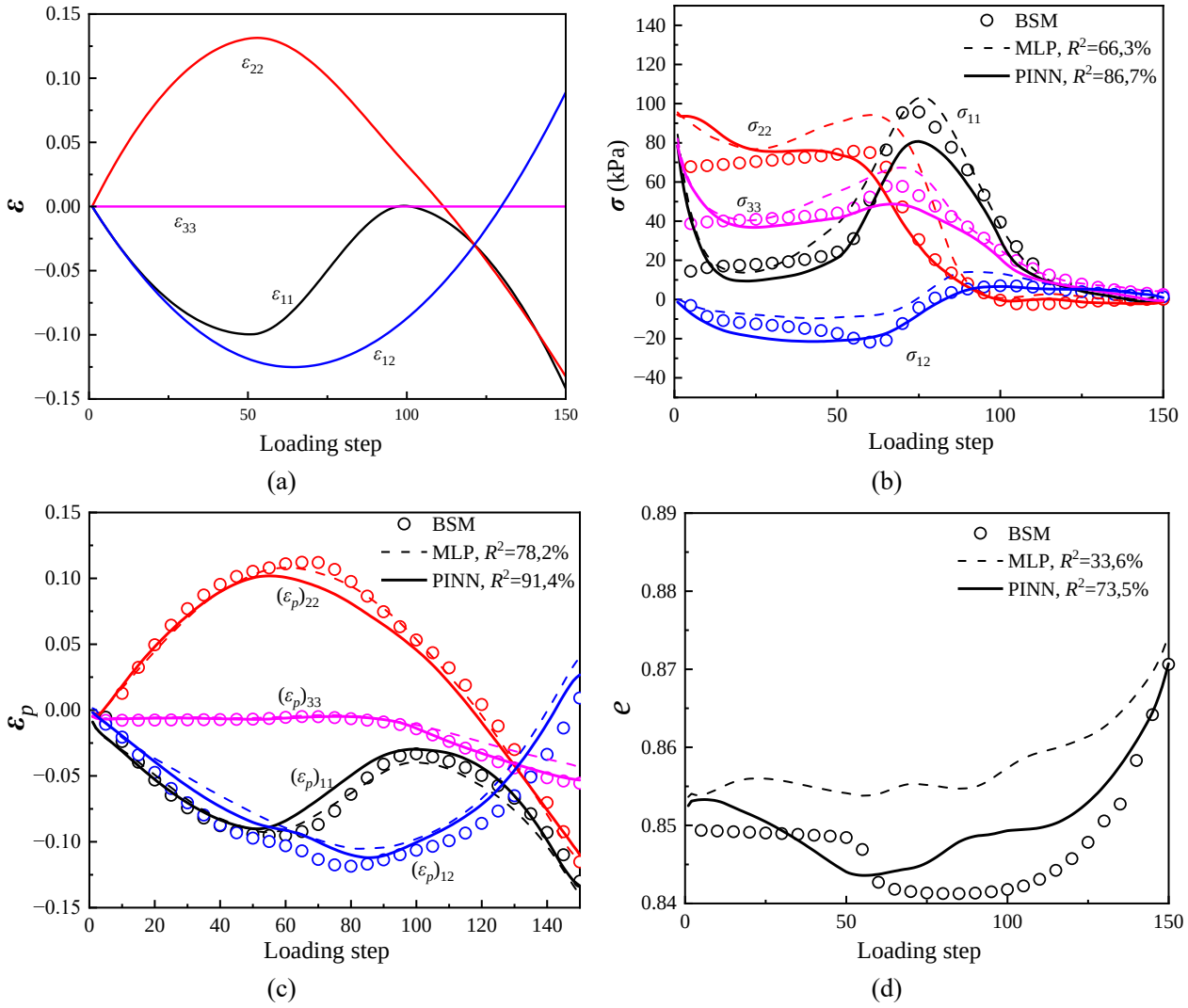
(a)



(b)



340 Fig. 9. Better prediction of constitutive relationship by PINN and comparison with bounding surface model: (a)
341 strain path; (b) predicted and real stress path; (c) predicted and real plastic strain; and (d) predicted and real void
342 ratio.



343 Fig. 10. Worse prediction of constitutive relationship by PINN and comparison with bounding surface model: (a)

strain path; (b) predicted and real stress path; (c) predicted and real plastic strain; and (d) predicted and real void ratio.

5. Incorporation into FEM

5.1 Implicit algorithm for FEM-PINN coupling

To enhance the application of the model to real cases, we incorporate the PINN into FEM for solving some boundary value problems (BVPs). Compared with elastoplastic constitutive models, the PINN does not need elastoplastic iteration and thereby requires fewer computational resources [12]. Fig. 11 and Algorithm 2 show the schema of the implementation of PINN into the FEM framework using an implicit algorithm. The Gauss points in the finite element domain give the displacement tensor $\boldsymbol{u}$. Accordingly, the strain can be calculated via the shape function and its gradient. The strain in the current step is then transmitted to the neural network to predict the stress and internal state variables, with the recorded stress and state variables. In the FEM-PINN framework, the predicted stress and material stiffness matrix are required to return to Gauss points to assemble the global stiffness matrix. The tangential stiffness matrix is found a higher convergence rate and stable calculation results, which makes sense since the incremental quantities are calculated based on previous timesteps [18, 33]. Referring to Guan et al. [18], Eqs. (4) and (5) are employed to assemble the tangential stiffness matrix, in which the required stress tensor and void ratio are directly predicted by the neural network. Finally, the node displacement and material stress can be obtained by several Newton-Raphson iterations, thereby finishing the FEM-PINN calculation.

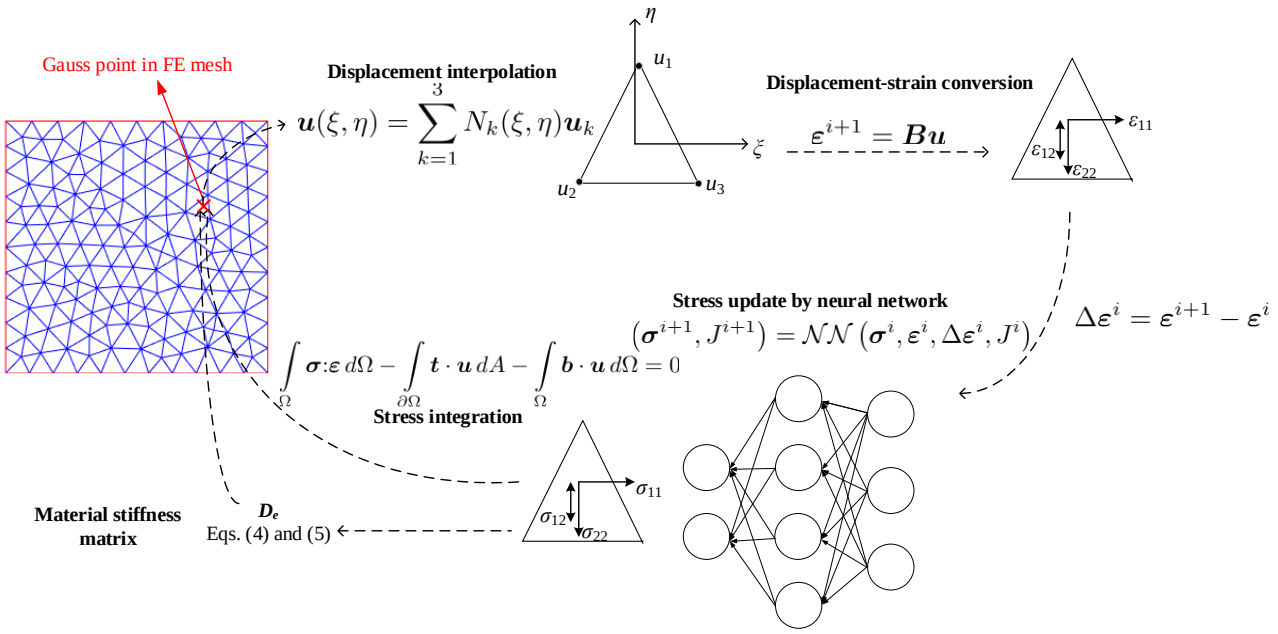


Fig. 11. Schema of the incorporation of neural network into finite element framework. $\boldsymbol{u}$ denotes the displacement vector; N denotes the shape function; $\boldsymbol{B}$ denotes the operation matrix (gradient of shape function); Ω denotes the geometric domain; $\partial\Omega$ denotes the boundary; $\boldsymbol{t}$ denotes the surface traction force; $\boldsymbol{b}$ denotes the body force.

Algorithm 2 Implicit algorithm for FEM-PINN coupling

Require: Geometry, meshing, loading, boundary conditions, and pre-trained neural network

Initialization: $\boldsymbol{\sigma}^0$, $\boldsymbol{\varepsilon}^0=0$, e^0, $\boldsymbol{\varepsilon}_p{}^0=0$, total load step N_s, and maximal New-Raphson iterations N_{max}

for $i = 1, N_s$ **do**

Apply global loading step i

for $j = 1, N_{max}$ **do**

Calculate the strain increment tensor $\Delta\boldsymbol{\varepsilon}^{i-1,j}$ via shape function and node displacement $\boldsymbol{U}$

Calculate stress tensor and internal state variables via the neural network

$\left(\boldsymbol{\sigma}^{i},e^{i},\boldsymbol{\varepsilon}_p^{i}\right)=\mathcal{NN}\left(\boldsymbol{\sigma}^{i-1},\boldsymbol{\varepsilon}^{i-1},\Delta\boldsymbol{\varepsilon}^{i-1,j},e^{i-1},\boldsymbol{\varepsilon}_p^{i-1}\right)$

Assemble element stiffness matrix via Eqs. (4) and (5)

Assemble global stiffness matrix $\boldsymbol{K}$ with the integration of element stiffness matrix

Calculate the internal force $\boldsymbol{F}_{int}=\boldsymbol{K}\boldsymbol{U}$ and Residual force $\boldsymbol{R}=\boldsymbol{F}_{int}-\boldsymbol{F}_{ext}-\boldsymbol{F}_b$

if $\boldsymbol{R}$ < tolerance, **break** current loop

elseif $\boldsymbol{R}$ > tolerance

Calculate the displacement increment $\Delta\boldsymbol{U}=\boldsymbol{K}^{-1}\boldsymbol{R}$

Update node displacement vector $\boldsymbol{U}=\boldsymbol{U}+\Delta\boldsymbol{U}$

end if

end for

Update the state variables at Gauss points $\boldsymbol{\sigma}^{i-1}=\boldsymbol{\sigma}^{i}$, $\boldsymbol{\varepsilon}^{i-1}=\boldsymbol{\varepsilon}^{i}$, $e^{i-1}=e^{i}$, $\boldsymbol{\varepsilon}_p{}^{i-1}=\boldsymbol{\varepsilon}_p{}^{i}$

end for

Note: $\boldsymbol{F}_{ext}$ and $\boldsymbol{F}_{b}$ denote the external force and body force vectors, respectively.

In the next sections, we select the BVPs of the biaxial test (with three sub-cases), layered-soil compression, and cavity expansion to examine the validity of the FEM-PINN framework. As the Gauss point in the finite element domain would experience various strain paths that are quite different from the loading paths in the training process, we first use the BSM to run cases as the benchmark and generate a new database. Then we deploy the transfer-learning to re-train these BVPs-based data, during which the network architecture is kept constant but the weight matrix between layers will be re-trained using the new dataset. After that, the re-trained neural network will play the role of calculating stress and internal state variables in the FEM-PINN framework. The FEM, constitutive model, and neural network are all built up in a custom Matlab program.

5.2 Case 1: Biaxial test

5.2.1 Proportional strain loading

The biaxial test with proportional strain-control loading is used. Inspired by Zhang et al. [15], we designed an elementary test with two three-node triangular elements as shown in Fig. 12(a). The size of the square is 100×100 mm. The left and bottom boundaries are fixed in horizontal and vertical directions, respectively. We impose a proportional strain by controlling the ratio of displacements of the top (u_{22}) and right boundaries (u_{11}):

$$u_{22} = \chi u_{11} \tag{12}$$

where χ controls the strain-loading proportion. $\chi = 1$ denotes a constant volume response. $\chi > 1$ and $\chi < 1$ correspond to the dilation and contraction in volume, respectively. We set u_{22} to 15 mm in all cases, corresponding to a 15% strain. Besides, the BSM is used as the benchmark for comparison with the PINN. The constitutive parameters and initial state are the same as that in section 3.1.1. Five χ values are investigated: 0.6, 0.8, 1.0, 1.2, and 1.5, with corresponding strain paths shown in Fig. 12(b). The results in the top-right element are extracted to be analysed.

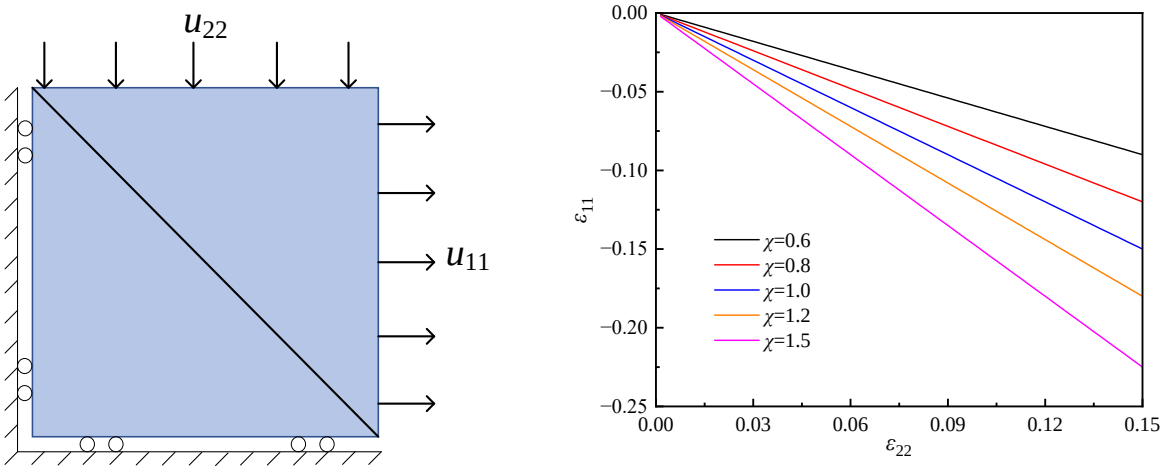
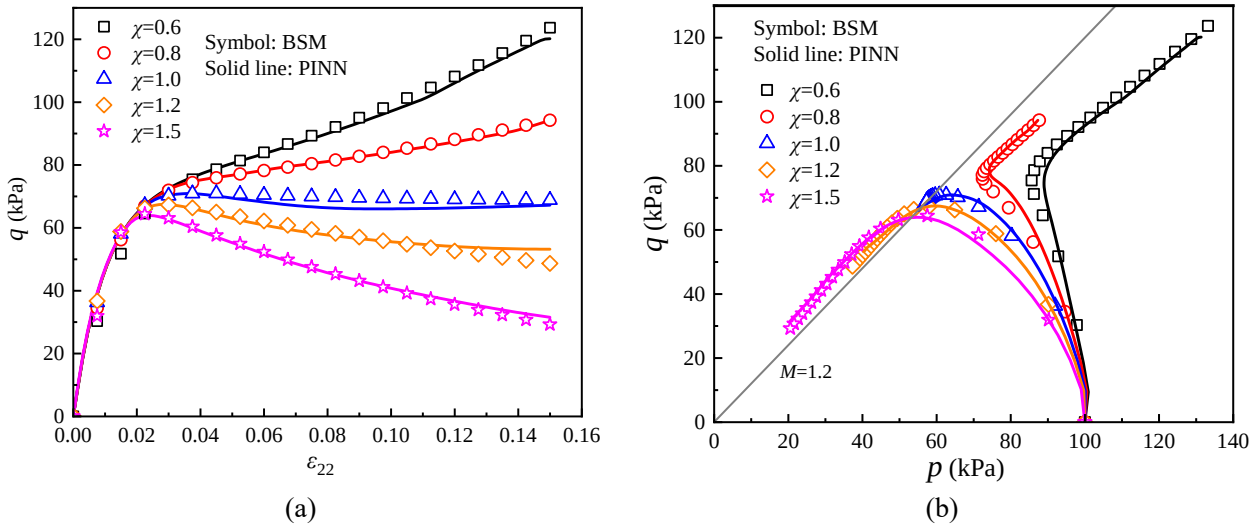


Fig. 12. Biaxial test in a finite element: (a) triangular elements, and (b) proportional strain-loading paths with different χ values.

Fig. 13 displays the predicted evolution of deviatoric stress, stress paths, plastic shear strain, and void ratio using the BSM and PINN. Excellent agreements with the exact BSM-based results are achieved by the FEM-PINN framework for both stress and strain responses. The decrease of stress and transition of strain-hardening to strain-softening behaviours with increasing χ values are captured by the FEM-PINN predictions. In the *p*-*q* plane, the FEM-PINN also successfully predicts the stress paths for all χ values, in which the stress always increases with strain for $\chi < 1$. For the constant-volume response (χ=1), the final computed *M* value coincides with the prescribed value at the critical state. Besides, critically, the FEM-PINN accurately predicts the plastic strain for all χ values as shown in Fig. 13(c). Likewise, the evolution of void ratio in the loading process is also well captured.



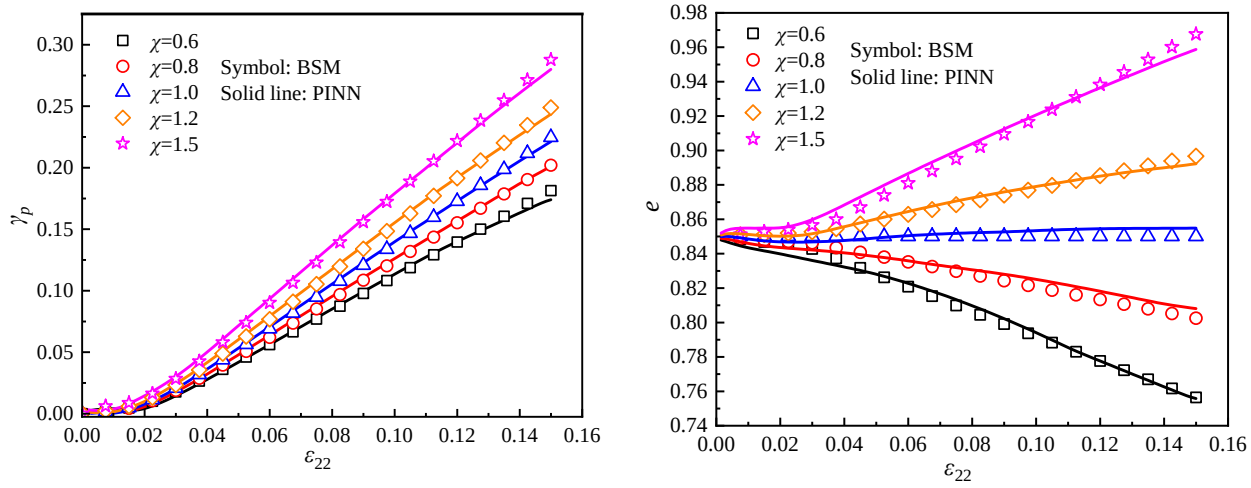


Fig. 13. Results of biaxial test with different χ values in (a) deviatoric stress q; (b) stress paths of effective stress p versus deviatoric stress q; (c) plastic shear strain γ^p; and (d) void ratio e. The same colour represents the same case.

5.2.2 Loading-unloading test

To examine the ability of PINN to maintain the well-known memory of soils regarding their stress-strain history, the loading-unloading-reloading path is designed using the FE domain in Fig. 12(a) with the same boundary condition. In this strain path, we keep $\varepsilon_{11} = \varepsilon_{22}$ and $\varepsilon_{12} = 0$, as shown in Fig. 14(a). The axial strain will first increase to 2%, which is achieved by applying the displacement of the top and right boundary to 2 mm. Then unload the strain to 1% by inversely applying the displacement of boundaries by 1 mm. Finally, reload the strain to 4% by again applying the displacement of the top and right boundary until 4 mm. In this loading-unloading cycle, the constitutive parameters and initial state of soil elements keep the same as in Section 5.2.1.

Fig. 14(b) shows the evolution of deviatoric stress with axial strain using BSM and PINN. It can be observed that PINN achieves great agreement with the BSM in the loading-unloading-reloading cycle. In the strain range of 1% - 2%, the unloading line did not coincide with the reloading line, which is a feature of the bounding surface model that the plastic strain can occur even within the yield surface. This performance is also captured by the FEM-PINN.

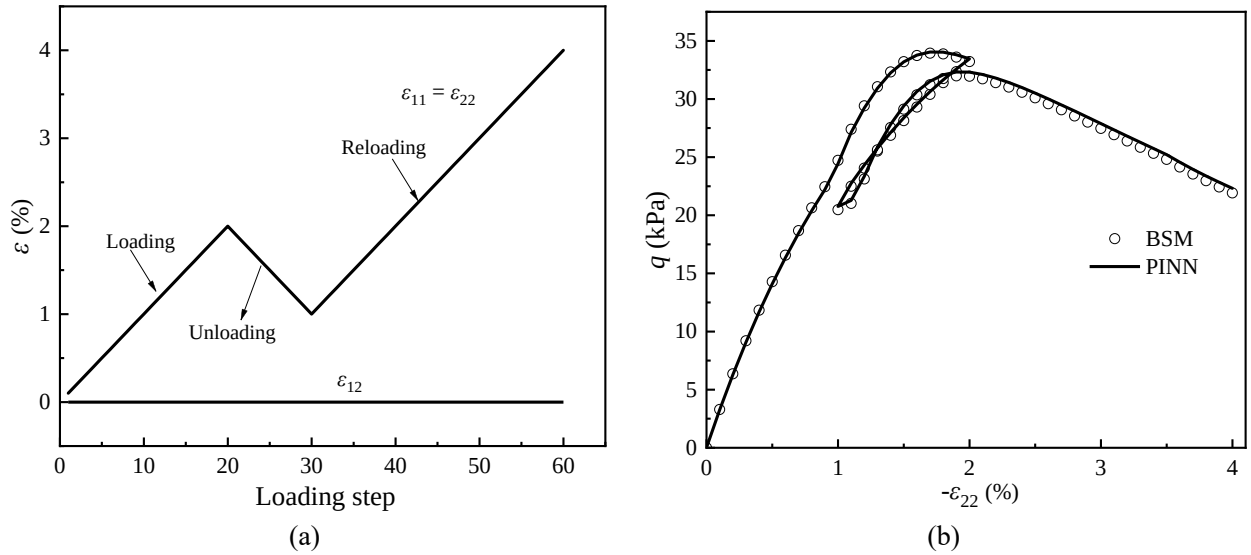


Fig. 14. Illustration of biaxial loading-unloading test: (a) loading path; and (b) results of deviatoric stress q.

5.2.3 Biaxial shear

This example aims to verify the ability of PINN in shear localization problems. The biaxial shear BVP is shown in Fig. 15 with a homogeneous soil column of size 0.08×0.02 m. The constant confining pressure of p_0 = 100kPa is applied. A top-boundary displacement of u_{22} = 0.006 m is applied, corresponding to 7.5% axial strain, while the bottom boundary is fixed in the vertical direction. A small hole is used to induce in the left centre as an imperfection to trigger shear localization [33], as shown in Fig. 15. The constitutive parameters and initial state of soil elements also keep the same in Section 5.2.1. Two Gauss points are selected to display the process of loading in detail.

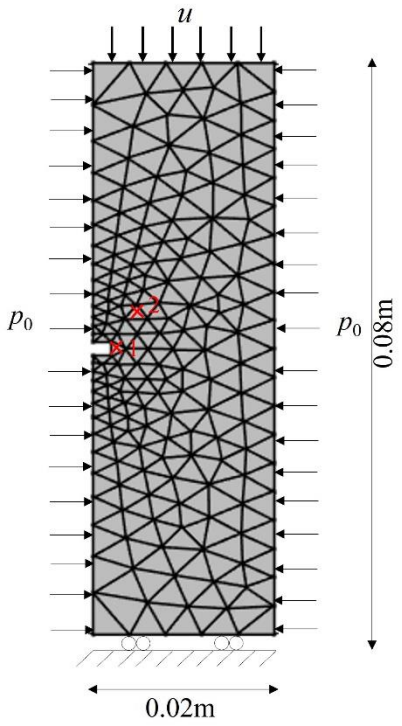


Fig. 15. FE domain of the biaxial shear test, as well as two selected Gauss points.

The contours plot shear strain at the end of loading are displayed in Fig. 16. It can be observed that the shear strain field of the PINN matches the BSM well, as shown in Figs. 16(a) and (b). Besides, both methods successfully described the expected development of a shear band around the imperfection. Fig. 16(c) presents the relative error (defined as $|\frac{y_B - y_P}{y_B}|$, where y_B and y_P denote the results of BSM and PINN, respectively) with respect to the shear strain. As supposed, the relative error at all Gauss points is extremely low and less than 7%, indicating an excellent agreement between BSM and FEM-PINN calculations.

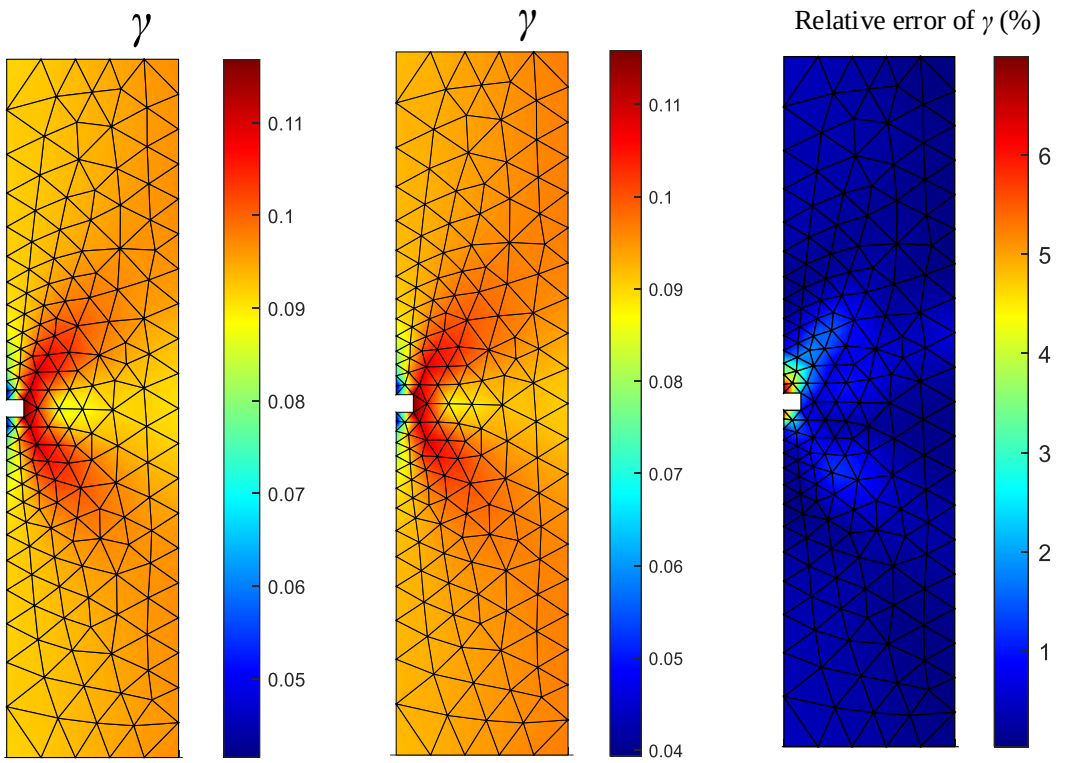


Fig. 16. Contours of FEM results of shear strain γ for biaxial shear test using (a) BSM and (b) PINN, as well as (c) their relative error.

The selected two Gauss points in Fig. 15 are further plotted in Fig. 17 regarding the evolution of deviatoric stress and void ratio with plastic shear strain. The deviatoric stress in Fig. 17(a) presents a good agreement between the BSM and PINN for two Gauss points. Point 1 presents more stress evolution as it is closer to the imperfection hole. The void ratio in point 1 shows some discrepancies between the BSM and PINN, as displayed in Fig. 17(b). For Gauss point 2, its void ratio almost remains constant after the shear strain of 5%, which indicates point 2 has approached the critical state since it is located at the shear band. As expected, the FEM-PINN also reproduced this critical state feature.

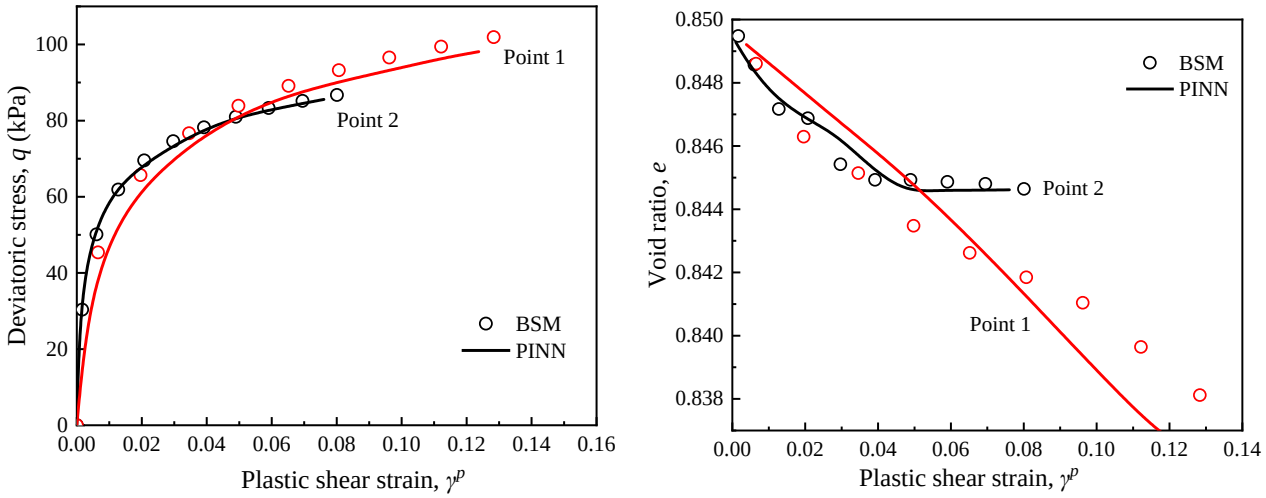


Fig. 17. Results of biaxial shear test for two points in the FE domain. The variation of (a) deviatoric stress q and (b) void ratio e with γ^p using the BSM and PINN.

5.3 Case 2: Layered-soil compression

The second case is the layered-soil compression, whose FE domain is shown in Fig. 18. A soil column with a height of 0.2 m and width of 0.1 m is set up, with the lateral side fixed in the horizontal direction and the bottom fixed in the vertical direction. The soil in this case has the same constitutive parameters as that in Section 3.1.1 and a pre-consolidated pressure of 100 kPa. From the top to bottom, the soil column is divided into three layers, with the initial stress of p_0 = 100 kPa, 50 kPa, and 100 kPa (corresponding to over consolidation ratio (OCR) of 1, 2, and 1) as well the initial void ratio of e_0 = 0.85, 0.88, and 0.82, respectively. Soil thickness for each layer is depicted in Fig. 18. The top boundary will be subject to a total vertical displacement of u_{22} = 0.006 m. We calculated all Gauss points but selected three of them from each layer to detail the results of finite element simulation, as shown in Fig. 18. Soil gravity is neglected.

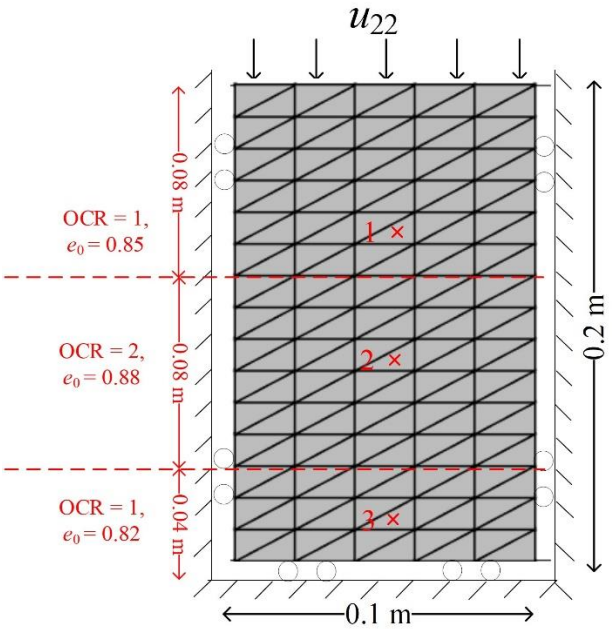
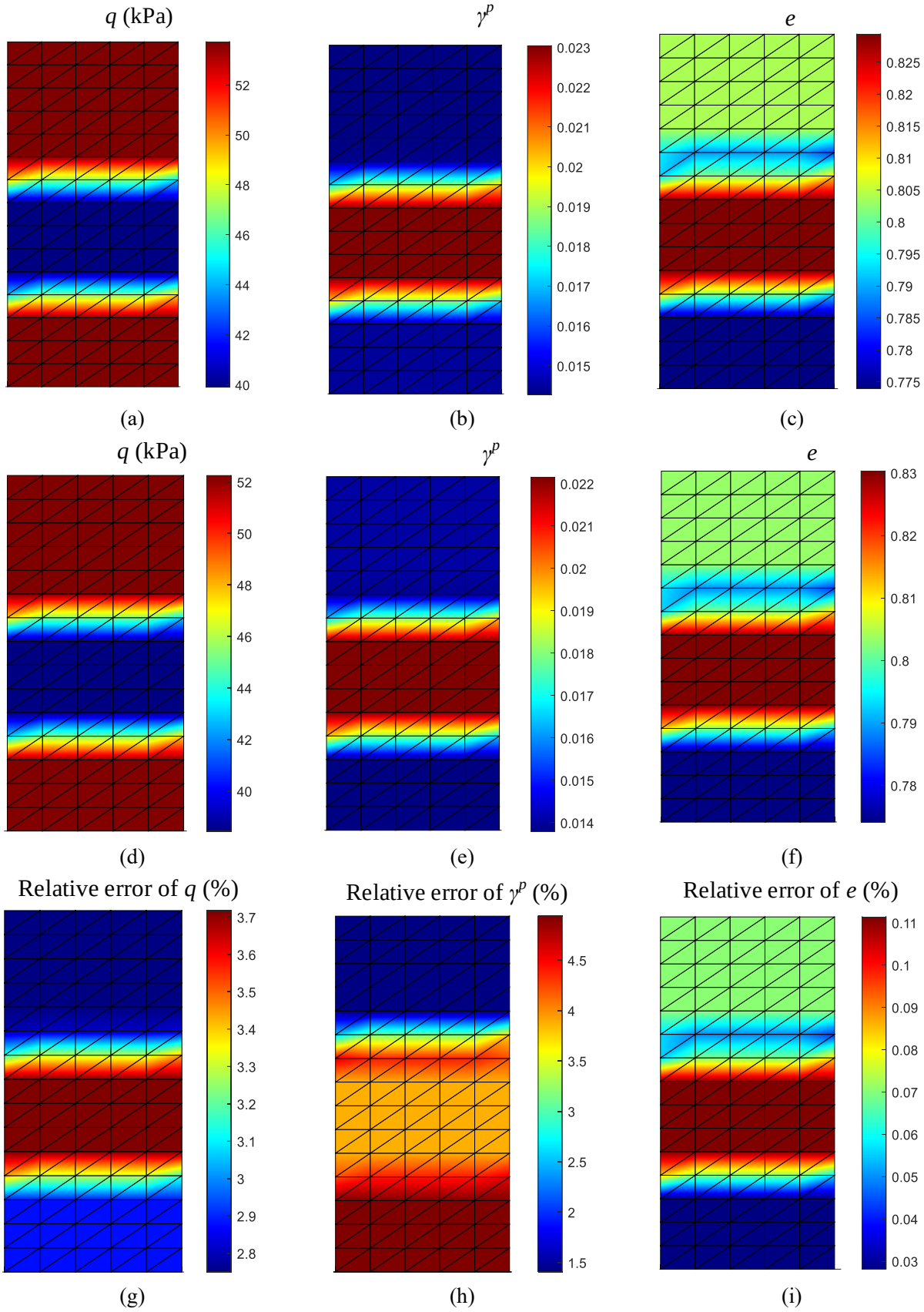


Fig. 18. FE domain of the layered-soil compression, as well as three selected Gauss points.

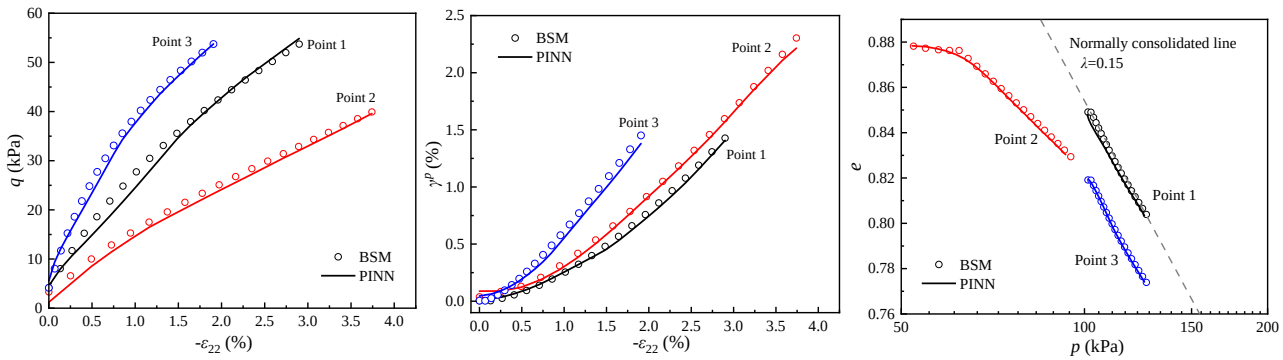
The comparison of the deviatoric stress, plastic shear strain, and void ratio contours using the BSM and PINN at the end of the simulation are present in Fig. 19. It can be observed that the FEM-PINN can reproduce satisfactory results that are very similar to those calculated using the benchmark BSM. The soil of OCR=2 presents the lowest deviatoric stress and highest plastic shear strain due to the small initial stress state. The predicted void ratio contour by FEM-PINN coincides well with the BSM in all three soil layers. To quantify the difference between the BSM and FEM-PINN, Figs. 19(g), (h), and (i) show the relative error with respect to the deviatoric stress, plastic shear strain, and void ratio. The relative error at all Gauss points is extremely low and less than 5%, indicating an excellent agreement between BSM and FEM-PINN calculations.



476 Fig. 19. Contours of FEM results for layered-soil compression using the bounding surface model in (a) deviatoric
477 stress q; (b) plastic shear strain γ^p; and (c) void ratio e, as well as using the PINN in (d) deviatoric stress q; (e) plastic

shear strain γ^p; and (f) void ratio e.; and their relative error in (g) deviatoric stress q; (h) plastic shear strain γ^p; and (i) void ratio e.

Fig. 20 shows the detailed deviatoric stress, plastic shear strain, and void ratio for three selected Gauss points. It can be observed that the PINN exhibits great agreement with the BSM-based results both in stress and deformation responses. The soil of OCR=2 and e_0=0.88 experiences the longest vertical strain, during which the evolution of deviatoric stress and plastic shear strain is accurately captured by the FEM-PINN framework. Additionally, the void ratio in three soil layers decreased with the increasing mean effective stress as the applied compression loading. The calculated void ratios of soil with OCR=1 by the BSM and PINN are precisely located at the normally consolidated line (NCL) with the pre-set λ = 0.15. The soil of OCR=2 and e_0=0.88 starts from the left side to gradually approach the NCL. As expected, this feature is also reproduced by PINN.



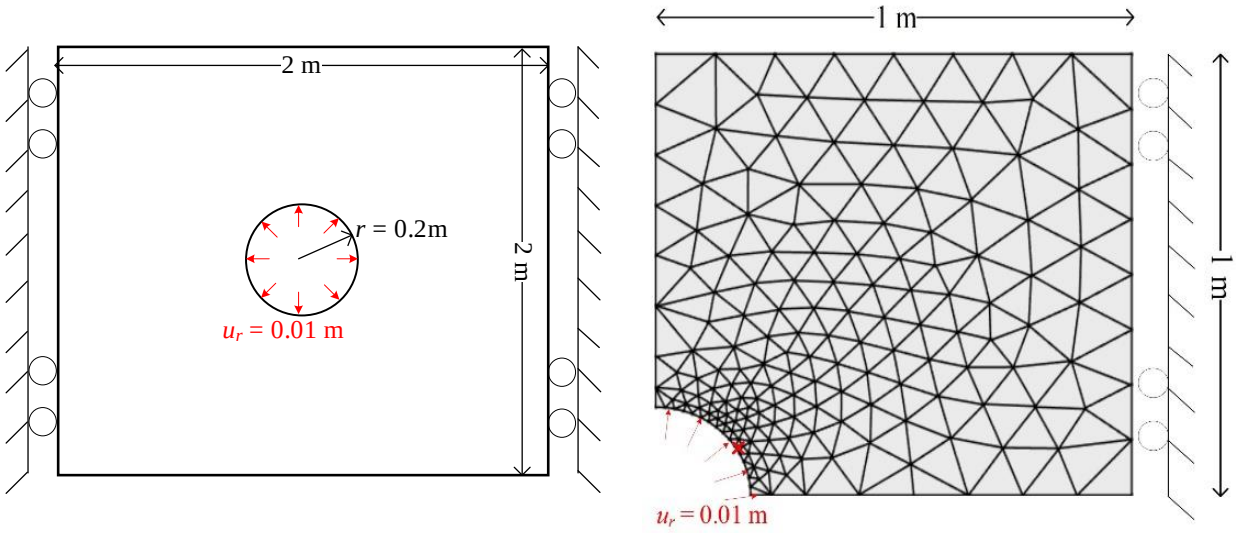
(a) (b) (c)

Fig. 20. Results of layered-soil compression for three points in FE domain in (a) deviatoric stress q; (b) plastic shear strain γ^p; and (c) void ratio e with mean effective stress p, using bounding surface model and PINN.

5.4 Case 3: Cavity expansion

The third case investigated a typical cavity expansion problem to demonstrate the capability of the proposed FEM-PINN framework to capture complex constitutive responses. Fig. 21(a) shows the geometry and boundary conditions of the cavity expansion problem. A cavity inside the centre of a 2×2 m square soil domain has a radius of 0.2 m and would expand along the radial direction by 0.01 m. The left and right boundaries are fixed to horizontal displacements, but the top and bottom boundaries are free. Only a quarter of the domain was used, as displayed in Fig. 21(b). One Gauss

point on the edge of the cavity is selected to analyse as a representative. The constitutive parameters, initial stress, and void ratio are the same as that in Section 3.1.1.



(a) (b)

Fig. 21. FE domain of the cavity expansion: (a) geometry and boundary conditions; (b) meshing and selected Gauss point.

Fig. 22 compares the result contours of the bounding surface model and FEM-PINN in domain displacement, deviatoric stress, and plastic shear strain, where a very good match is observed. The field displacement contours in Figs. 22(a) and (b) are almost the same, which demonstrates the FEM-PINN framework achieved good success and can reproduce soil deformation in the expansion process. In addition, the stress field also exhibits excellent agreement, as shown in Figs. 22(d) and (e). Stress concentration emerges around the cavity edge, which has been reported by Guan et al. [18], and a larger stress region is present near the right side of the cavity because of the constraint of boundary displacement. The FEM-PINN also well reproduces the strain localization of plastic shear strain near the cavity edge area in Figs. 22(g) and (h). Figs. 22(c), (f), and (i) show the relative error for each variable between the BSM and PINN is very small with values less than 8.5%.

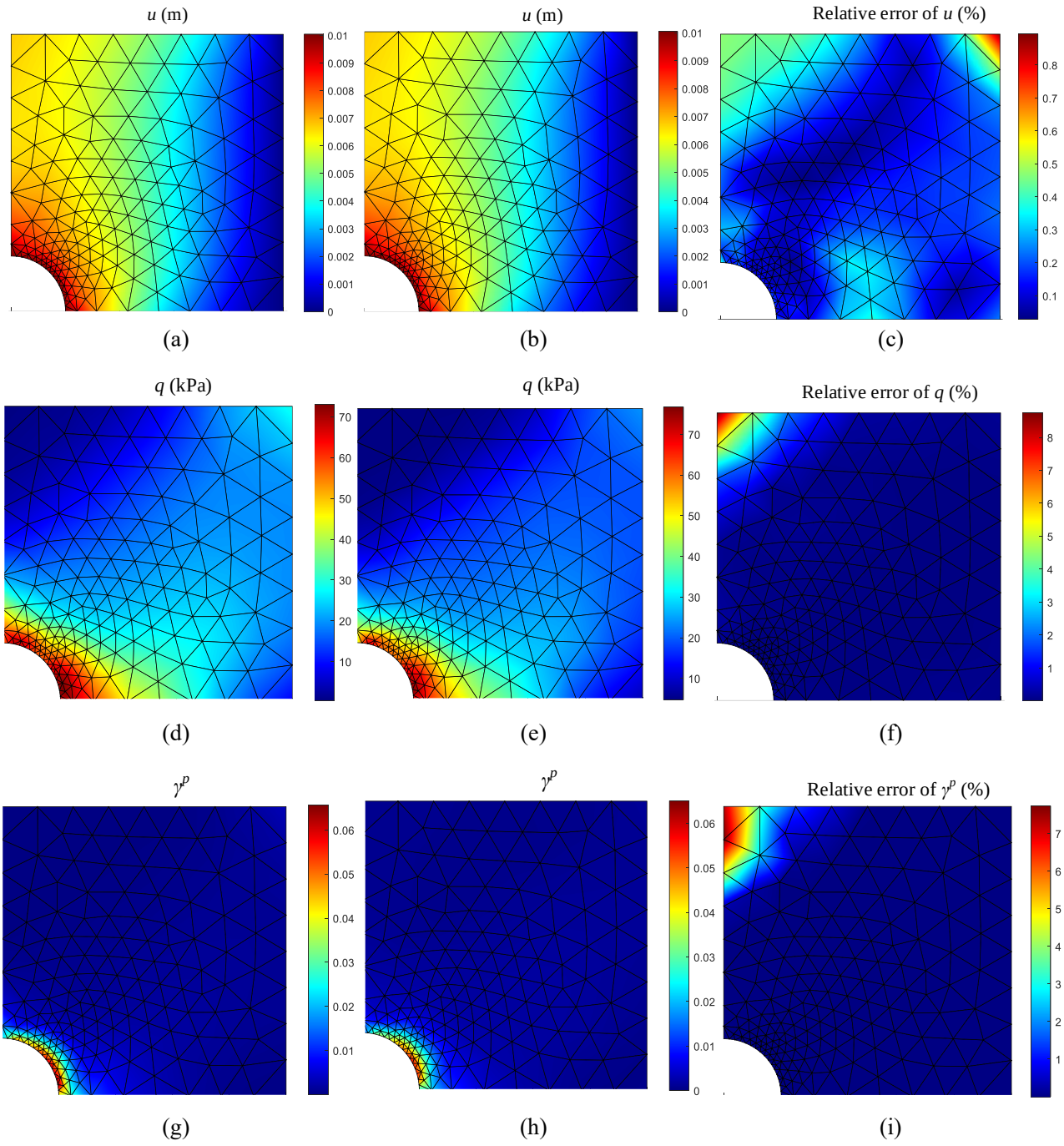


Fig. 22. Contours of FEM results for cavity expansion using bounding surface model in (a) total displacement u; (d) deviatoric stress q; and (g) plastic shear strain γ^p, as well as using the PINN in (b) total displacement u; (e) deviatoric stress q; and (h) plastic shear strain γ^p, and their relative error in (c) total displacement u, (f) deviatoric stress q; and (i) plastic shear strain γ^p.

Fig. 23 shows the evolution of stress and plastic strain components with the cavity radial displacement for the selected Gauss point. It can be observed that the stress and plastic strain during cavity

expansion would undergo different loading paths for every component. Considering this fact, the FEM-PINN framework still provides well-acceptable results and successfully captures the evolution tendency for each stress and strain component. In summary, the PINN for modelling elasto-plastic soil in this study can surrogate the stress, strain, and internal variables with practical incorporation with the FEM framework.

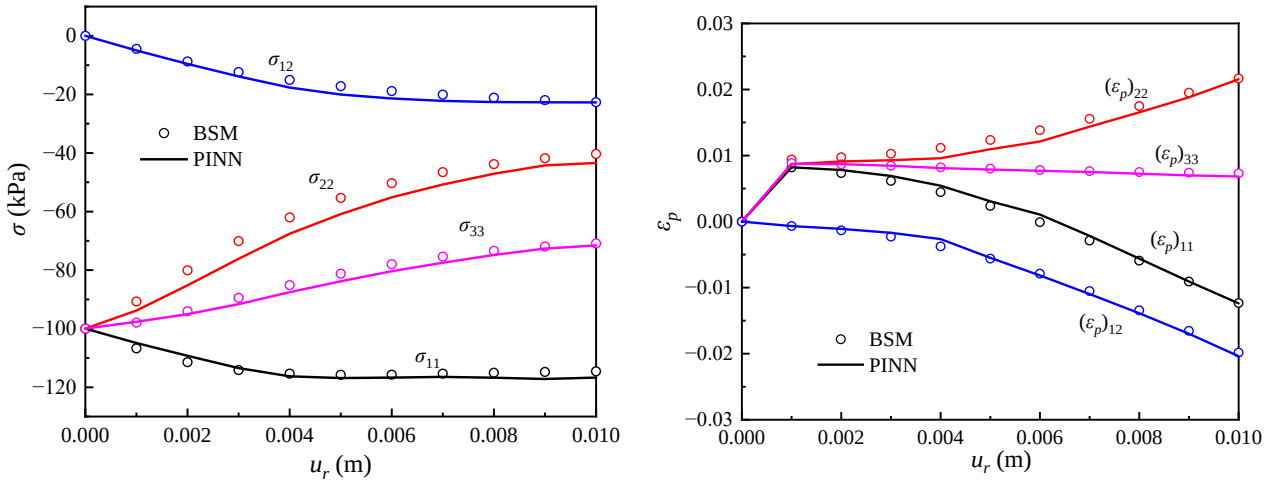


Fig. 23. Results of cavity expansion for selected Gauss point in evolution of (a) stress components; and (b) plastic shear strain components, using bounding surface model and PINN.

6. Conclusion

This study proposed a physical-informed neural network approach to capture the elasto-plastic behaviour of soils under different strain and stress paths. The general elasto-plastic stress-strain relationship with strain decomposition is used as an additional loss function to constrain the outputs to a framework of elasto-plasticity. The results demonstrate that the loss function which includes the restriction of physical information, can achieve a lower error than traditional deep networks based on MLP. The PINN provides a more stable prediction and a superior performance than the deep network in forecasting stress and strain in complex strain-loading paths. The proposed PINN can also accurately capture the evolution of plastic strain components and void ratio. Critically, it achieves this without the specification of any yield function or hardening laws.

The PINN was then incorporated into the FEM framework as a replacement for the constitutive model

to test its performance in boundary value problems. Three cases were examined: biaxial test, layered-soil compression, and cavity expansion. The results show that the FEM-PINN calculation exhibited excellent performance of deviatoric stress, plastic shear strain, and void ratio in comparison to the benchmark bounding surface model. In addition, the stress concentration and strain localization characteristics were also reproduced by the FEM-PINN framework. When coupled with FEM, the neural network can provide reliable and effective means of modelling soil behaviours under complex loading paths.

The PINN model overcomes non-physical constraints of outputs of neural networks and therefore improves the prediction precision. The PINN for elasto-plastic soil in this study can be a practical approach to model complex soil behaviours in engineering applications with less computational cost. This study provides an alternative strategy and potential for developing an interesting interpretable deep learning for elasto-plastic soil as well as its practical use. Future work will involve validation of the approach for multiple soil parameters and conditions which will allow comparison to real soil behaviour. We will also experiment with cyclic and rapid loading conditions.

Acknowledgement

The authors would like to acknowledge the funding support by the China Scholarship Council (CSC).

Appendix A: Multi-layer perceptron (MLP) neural network

Fig. A1 shows a typical MLP, which contains an input layer, multiple hidden layers, and an output layer [18]. A set of neurons are arranged in each layer and connected through weights and bias. The input data are firstly presented through the input layer and then passed through the hidden layer(s) being processed by the neural network to eventually predict values for the output layer. This process can be expressed mathematically by considering a feedforward propagation process that the output of the *i*-th neuron in the *k*th layer (y_i^k) can be given by assuming n neurons in the (*k*-1)-th layer [37]:

$$y_i^k = f\left(\sum_{j=1}^{n} \omega_{ij}^k \, x_j^{k-1} + b_i^k\right) \tag{A1}$$

where x_j^{k-1} is the output of the j-th neuron in the (k-1)-th layer; ω_{ij}^k is the weight between the two neurons of the x_j^{k-1} and y_i^k; b_i^k is the bias of the i-th neuron in the kth layer. f signifies the activation function, and typically used activation functions are sigmoid, hyperbolic tangent (Tanh), rectified linear unit (Relu), and leaky Relu activation functions [5].

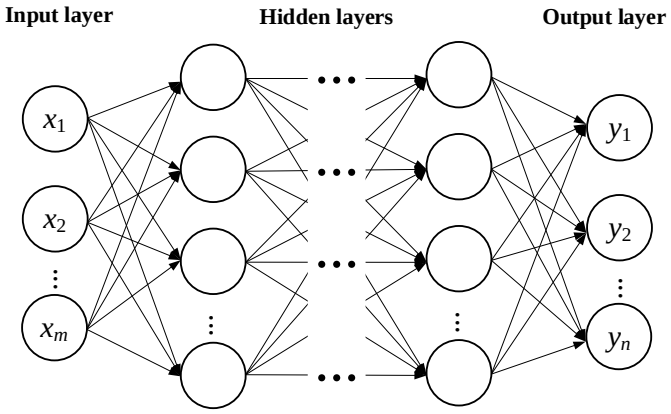


Fig. A1. Typical structure of an MLP. x_1, x_2, …, x_m represent the elements in input vector **x**; y_1, y_2, …, y_n represent the elements in output vector **y.**

The training implementation of MLP can be considered into two phases: forward calculation and backward propagation. During the forward process, the value of each hidden neuron is calculated by summing up the values of input neurons multiplied by corresponding connection weights. The error between the output and the real values can be calculated and then minimized by the backward algorithm that updates the connection weights.

Appendix B: Bounding surface model

The bounding surface model adopted in this research is based on the CASM by Yu [3] and the model by Zhou et al. [38]. The state bounding surface is described by a general stress-state relation (see Eq. (B-1)) with the definition of state parameter ξ (see Eq. (B-2) and Fig. B1, defined by Been and Jefferies [39]):

$$\left(\frac{\eta}{M}\right)^n = 1 - \frac{\xi}{\xi_R} \tag{B-1}$$

$$\xi = v + \lambda \ln p' - \Gamma \tag{B-2}$$

$$\xi_R = N - \Gamma = (\lambda - \kappa)\ln r \tag{B-3}$$

where p' is the mean effective stress; M is the slope of the critical state line (CSL) in the p' - q space where q denotes the deviatoric stress; $\eta = q / p'$ is the stress ratio; $v = (1 + e)$ is known as specific volume and e is void ratio; λ is the slope of the CSL in the v - $\ln p'$ space; κ represents the slope of the swelling line; ξ_R is the positive reference state parameter which denotes the vertical distance between CSL and a normal consolidation line (NCL); n denotes a material constant controlling the curvature of the yield surface; Γ and N denote the v value of CSL and NCL at p' = 1 kPa, respectively; $r = p'_0 / p'_x$ is known as the spacing ratio, in which p'_0 is the consolidation pressure and p'_x is the mean stress at the intersection between the CSL and the swelling line.

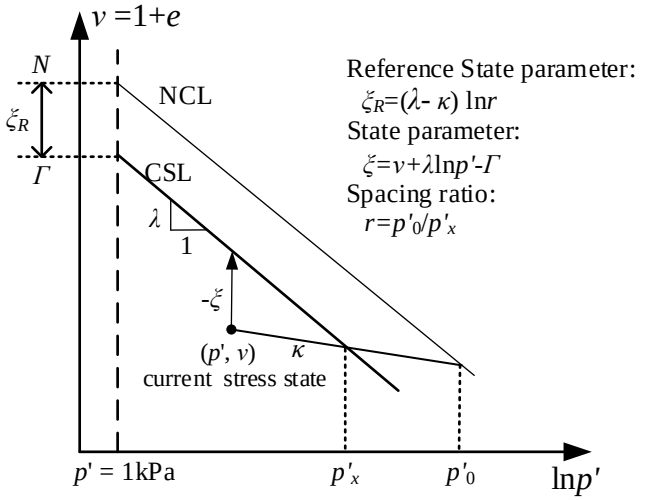


Fig. B1. Definitions of state parameter and critical state parameter.

The elastic volumetric ($d\varepsilon_V^e$) and deviatoric strain ($d\varepsilon_q^e$) increments are calculated using the bulk modulus K and the shear modulus G:

$$d\varepsilon_V^e = \frac{dp'}{K} = \frac{\kappa dp'}{(1+e)p'} \tag{B-4}$$

$$d\varepsilon_q^e = \frac{dq}{3G} = \frac{2(1+\mu)dq}{9(1-2\mu)K} \tag{B-5}$$

where μ denotes Poisson's ratio.

The normal consolidation line (NSL) of soil is expressed as:

$$v = N - \lambda \ln p' \tag{B-6}$$

The critical state line (CSL) is shown as follows:

$$q = Mp' \tag{B-7}$$

$$v = \Gamma - \lambda \ln p' \tag{B-8}$$

Fig. B2 depicts the bounding surface in p' – q space. The formula of the bounding surface can be deduced by integrating Eqs. (B-1) to (B-3) in terms of stress:

$$F = \left(\frac{q}{Mp'}\right)^n + \frac{\ln(p'/p_0')}{\ln r} \tag{B-9}$$

Parameter r can be expressed as a function of Γ, N, λ, and κ with respect to the critical state of soil:

$$r = \exp\left(\frac{N-\Gamma}{\lambda-\kappa}\right) \tag{B-10}$$

The radius mapping rule is employed, in which the mapping centre point is the original point in stress space. The mapping rule is expressed as:

$$\frac{p'}{p_m'} = \frac{q}{q_m} = \frac{\rho}{\rho_m} \tag{B-11}$$

where p' and q also signify the stress components in the real point; p'_m and q_m are the stress components in the mapping point; ρ and ρ_m are the distance to the mapping centre from the real point and mapping point, respectively.

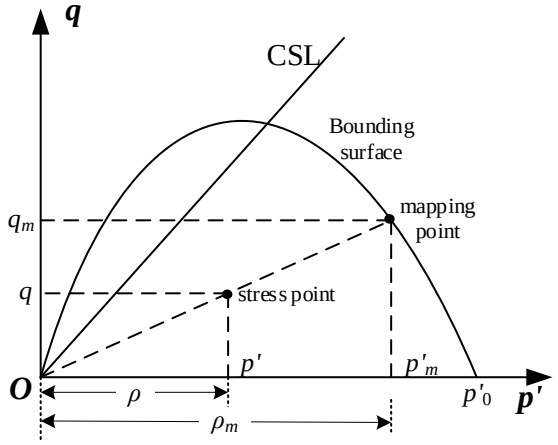


Fig. B2. Schema of bounding surface model.

The bounding surface model adopts the non-associated flow rule, in which the stress-dilation function D is expressed as follows:

$$D = \frac{\mathrm{d}\varepsilon_V^p}{\mathrm{d}\varepsilon_q^p} = \frac{M^2(\rho/\rho_m) - \eta^2}{2\eta} \tag{B-12}$$

where $\mathrm{d}\varepsilon_V^p$ and $\mathrm{d}\varepsilon_q^p$ are elastic volumetric and deviatoric strain increments, respectively. The

following isotropic hardening law is adopted to relate the variation of the bounding surface $\mathrm{d}p'_0$ to the internal variable $\mathrm{d}\varepsilon_V^p$ as follows:

$$\mathrm{d}p'_0 = \frac{\nu}{\lambda-\kappa} p'_0 \, \mathrm{d}\varepsilon_V^p \tag{B-13}$$

The plastic volumetric strain increment should keep the criterion that the stress state should remain on the bounding surface during plastic evolution, as follows:

$$\frac{\partial f}{\partial p'} \mathrm{d}p' + \frac{\partial f}{\partial q} \mathrm{d}q + \frac{\partial f}{\partial p'_0} \frac{\partial p'_0}{\partial \varepsilon_V^p} \mathrm{d}\varepsilon_V^p = 0 \tag{B-14}$$

Substituting $\mathrm{d}\varepsilon_V^p$ by the $D\mathrm{d}\varepsilon_q^p$ can deduce to:

$$\mathrm{d}\varepsilon_q^p = \frac{1}{K_p} \left\{ \frac{\partial F}{\partial \boldsymbol{\sigma}} \right\}^{\mathrm{T}} \{\mathrm{d}\boldsymbol{\sigma}\} \tag{B-15}$$

where K_p is the plastic modulus and expressed as:

$$K_p = -\frac{\partial F}{\partial p_0} \frac{\partial p_0}{\partial \varepsilon_V^p} D = \frac{\nu}{(\lambda-\kappa)\ln r} \frac{M^2-\eta^2}{2\eta} \tag{B-16}$$

The above formula is only suitable to the point on the bounding surface. For the points within the bounding surface, Zhou et al. [38] used a relatively simple form:

$$K_p = \frac{\nu}{(\lambda-\kappa)\ln r} \frac{M^2(\rho/\rho_m)-\eta^2}{2\eta} \tag{B-17}$$

in which the introduction of ρ/ρ_m reflects that the state points closer to the bounding surface will have a smaller plastic modulus.

The Eqs. (B-1) to (B-17) formulate the theoretical basis of the bounding surface model. As mentioned above, there are seven parameters that need to be calibrated: λ, κ, μ, M, N, Γ, and n, by which other parameters and corresponding constitutive law can be derived from these seven parameters.

Reference

[1] Wu, M., Z. Xia and J. Wang, Constitutive modelling of idealised granular materials using machine learning method. Journal of Rock Mechanics and Geotechnical Engineering, 2023, 15(4): 1038-1051.

[2] Roscoe, K.H. and J.B. Burland. On the generalised stress-strain behaviour of 'wet' clay1968. Engineering Plasticity: Cambridge University Press, Cambridge, pp 535–609.

[3] Yu, H.S., CASM: a unified state parameter model for clay and sand. International Journal for

654 Numerical and Analytical Methods in Geomechanics, 1998, 22(8): 621-653.
655 [4] Zhang, N., A. Zhou, Y.-F. Jin, et al., An enhanced deep learning method for accurate and robust
656 modelling of soil stress–strain response. Acta Geotechnica, 2023.
657 [5] Zhang, P., Z.-Y. Yin and Y.-F. Jin, State-of-the-Art Review of Machine Learning Applications in
658 Constitutive Modeling of Soils. Archives of Computational Methods in Engineering, 2021, 28(5):
659 3661-3686.
660 [6] Zhang, Q., B. Mortazavi, X. Zhuang, et al., Exploring the mechanical properties of two-
661 dimensional carbon-nitride polymer nanocomposites by molecular dynamics simulations. Composite
662 Structures, 2022, 281: 115004.
663 [7] Kouretzis, G.P., D. Sheng and D. Wang, Numerical simulation of cone penetration testing using a
664 new critical state constitutive model for sand. Computers and Geotechnics, 2014, 56: 50-60.
665 [8] Staubach, P., J. Machaček, B. Bienen, et al., Long-Term Response of Piles to Cyclic Lateral
666 Loading Following Vibratory and Impact Driving in Water-Saturated Sand. Journal of Geotechnical
667 and Geoenvironmental Engineering, 2022, 148(11): 04022097.
668 [9] Zhang, Q., S. Ritzert, J. Zhang, et al., A multi-field decomposed model order reduction approach
669 for thermo-mechanically coupled gradient-extended damage simulations. arXiv preprint
670 arXiv:2407.02435, 2024.
671 [10] Haghighat, E., S. Abouali and R. Vaziri, Constitutive model characterization and discovery using
672 physics-informed deep learning. Engineering Applications of Artificial Intelligence, 2023, 120:
673 105828.
674 [11] Guan, Q.Z. and Z.X. Yang, Hybrid deep learning model for prediction of monotonic and cyclic
675 responses of sand. Acta Geotechnica, 2023, 18: 1447–1461.
676 [12] Eghbalian, M., M. Pouragha and R. Wan, A physics-informed deep neural network for surrogate
677 modeling in classical elasto-plasticity. Computers and Geotechnics, 2023, 159: 105472.
678 [13] Mozaffar, M., R. Bostanabad, W. Chen, et al., Deep learning predicts path-dependent plasticity.
679 Proceedings of the National Academy of Sciences, 2019, 116(52): 26414-26420.
680 [14] Motevali Haghighi, E. and S. Na, Revisiting Preparation of Phase Space for Learning Path-
681 Dependent Behavior via Deep Neural Networks. Journal of Engineering Mechanics, 2022, 148(12):

682 06022003.
683 [15] Zhang, P., Z.-Y. Yin, Y.-F. Jin, et al., Physics-constrained hierarchical data-driven modelling
684 framework for complex path-dependent behaviour of soils. International Journal for Numerical and
685 Analytical Methods in Geomechanics, 2022, 46(10): 1831-1850.
686 [16] Zhang, P., Z.-Y. Yin, Y.-F. Jin, et al., Physics-Informed Multifidelity Residual Neural Networks
687 for Hydromechanical Modeling of Granular Soils and Foundation Considering Internal Erosion.
688 Journal of Engineering Mechanics, 2022, 148(4): 04022015.
689 [17] Zhang, P., Z.Y. Yin, Y.F. Jin, et al., An AI-based model for describing cyclic characteristics of
690 granular materials. International Journal for Numerical and Analytical Methods in Geomechanics,
691 2020, 44(9): 1315-1335.
692 [18] Guan, Q.Z., Z.X. Yang, N. Guo, et al., Finite element geotechnical analysis incorporating deep
693 learning-based soil model. Computers and Geotechnics, 2023, 154: 105120.
694 [19] Guan, Q., Z. Yang, N. Guo, et al., Deep learning-accelerated multiscale approach for granular
695 material modeling. International Journal for Numerical and Analytical Methods in Geomechanics,
696 2024, 48: 1372–1389.
697 [20] Zhang, A. and D. Mohr, Using neural networks to represent von Mises plasticity with isotropic
698 hardening. International Journal of Plasticity, 2020, 132: 102732.
699 [21] Haghighat, E., M. Raissi, A. Moure, et al., A physics-informed deep learning framework for
700 inversion and surrogate modeling in solid mechanics. Computer Methods in Applied Mechanics and
701 Engineering, 2021, 379: 113741.
702 [22] Raissi, M., P. Perdikaris and G.E. Karniadakis, Physics-informed neural networks: A deep
703 learning framework for solving forward and inverse problems involving nonlinear partial differential
704 equations. Journal of Computational Physics, 2019, 378: 686-707.
705 [23] Meng, X. and G.E. Karniadakis, A composite neural network that learns from multi-fidelity data:
706 Application to function approximation and inverse PDE problems. Journal of Computational Physics,
707 2020, 401: 109020.
708 [24] Vahab, M., E. Haghighat, M. Khaleghi, et al., A Physics-Informed Neural Network Approach to
709 Solution and Identification of Biharmonic Equations of Elasticity. Journal of Engineering Mechanics,

2022, 148(2): 04021154.
[25] Cai, S., Z. Mao, Z. Wang, et al., Physics-informed neural networks (PINNs) for fluid mechanics: a review. Acta Mechanica Sinica, 2021, 37(12): 1727-1738.
[26] Masi, F. and I. Stefanou, Multiscale modeling of inelastic materials with Thermodynamics-based Artificial Neural Networks (TANN). Computer Methods in Applied Mechanics and Engineering, 2022, 398: 115190.
[27] Vlassis, N.N. and W. Sun, Component-Based Machine Learning Paradigm for Discovering Rate-Dependent and Pressure-Sensitive Level-Set Plasticity Models. Journal of Applied Mechanics, 2021, 89(2).
[28] Vlassis, N.N. and W. Sun, Sobolev training of thermodynamic-informed neural networks for interpretable elasto-plasticity models with level set hardening. Computer Methods in Applied Mechanics and Engineering, 2021, 377: 113695.
[29] Xu, K., D.Z. Huang and E. Darve, Learning constitutive relations using symmetric positive definite neural networks. Journal of Computational Physics, 2021, 428: 110072.
[30] Su, M.M., Y. Yu, T.H. Chen, et al., A thermodynamics-informed neural network for elastoplastic constitutive modeling of granular materials. Computer Methods in Applied Mechanics and Engineering, 2024, 430: 117246.
[31] He, G.-F., P. Zhang, Z.-Y. Yin, et al., Multi-fidelity data-driven modelling of rate-dependent behaviour of soft clays. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 2023, 17(1): 64-76.
[32] Dafalias, Y.F. and M.T. Manzari, Simple Plasticity Sand Model Accounting for Fabric Change Effects. Journal of Engineering Mechanics, 2004, 130(6): 622-634.
[33] He, Y. and S.J. Semnani, Machine learning based modeling of path-dependent materials for finite element analysis. Computers and Geotechnics, 2023, 156: 105254.
[34] Pan, Y. and L. Zhang, Mitigating tunnel-induced damages using deep neural networks. Automation in Construction, 2022, 138: 104219.
[35] Ruder, S., An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747, 2016.

738 [36] Potts, D.M. and A. Gens, A critical assessment of methods of correcting for drift from the yield
739 surface in elasto-plastic finite element analysis. International Journal for Numerical and Analytical
740 Methods in Geomechanics, 1985, 9(2): 149-159.
741 [37] Su, M., N. Guo and Z. Yang, A multifidelity neural network (MFNN) for constitutive modeling
742 of complex soil behaviors. International Journal for Numerical and Analytical Methods in
743 Geomechanics, 2023, 47: 3269–3289.
744 [38] Zhou, C., K.Y. Fong and C.W.W. Ng, A new bounding surface model for thermal cyclic behaviour.
745 International Journal for Numerical and Analytical Methods in Geomechanics, 2017, 41(16): 1656-
746 1666.
747 [39] Been, K. and M.G. Jefferies, A state parameter for sands. Géotechnique, 1985, 35(2): 99-112.

748