J. Mech. Phys. Solids 217 (2026) 106807

Contents lists available at ScienceDirect

Journal of the Mechanics and Physics of Solids

journal homepage: www.elsevier.com/locate/jmps

A neural-network framework to learn history-dependent constitutive laws and identifiability of internal variables

Mayank Raj *, Lianghao Cao , Andrew Stuart , Kaushik Bhattacharya

Division of Engineering and Applied Sciences, California Institute of Technology, Pasadena, 91125, CA, USA

ARTICLE INFO

Keywords:
Polyconvex neural network
Crystal plasticity
Internal variables
Numerical homogenisation

ABSTRACT

The identification of constitutive laws is ubiquitous in engineering: in modeling of materials, where experimental data are fitted to mathematical models or learning surrogate models to beat the FE^2 computational cost of multiscale numerical simulations. However, these models of constitutive laws, unless equipped with a energetic formulation, are not necessarily consistent with (a) the second law of thermodynamics; (b) stability of the material under extreme applied strain; and (c) the mathematical theory underpinning the existence of a solution of the non-linear momentum balance equation. In this work, we present a causal and energetic formulation, consistent with the aforementioned properties, of learning a history-dependent constitutive law. Moreover, we empirically discover, and, under controllability assumptions, theoretically show that the internal variables that are inferred from a learned constitutive model are identifiable up to a linear transform. This characterisation of the class of internal variables sheds light on an equivalence class of models that lead to the same stress-strain map. The framework is deployed to learn the Taylor-averaged response of a polycrystalline magnesium unit cell. We achieve a 2% relative error in the prediction of the Taylor-averaged response.

1. Introduction

Modeling the mechanical response of a material at the length scale relevant to its engineering application, known as the macro-scale, is often complicated. This is because the macro-scale response of a material is the cumulative result of physical processes occurring at a length scale finer than the scale of engineering applications. Moreover, the physics of the mechanical response of a material is often simpler and, hence, much better understood at a fine length scale. In this context, homogenisation refers to obtaining the effective macroscopic response of a material based on knowledge of its microscopic behavior. Although the analytic derivation of equations governing the macroscopic response from known microscopic physics may be possible in a simplistic setting (Bensoussan et al., 2011; Pavliotis and Stuart, 2008), doing so for many classes of materials is often intractable. Therefore, a computational approach to homogenisation, where data from a fine-scale process may be used to model the constitutive law at a coarse scale, is indispensable. In recent years, data-driven methods have been explored to beat the FE^2 computational cost of two-scale homogenisation. The use of a trained neural network (NN) to capture the effective properties of a hierarchical superconductor, and its application in the design of composites, is presented in Lefik et al. (2009). In Mozaffar et al. (2019), recurrent neural networks (RNNs) are trained in order to capture the effective response of a fixed representative volume element (RVE) when subjected to significant distortional

* Corresponding author.
E-mail addresses: mraj@caltech.edu (M. Raj), lianghao@caltech.edu (L. Cao), astuart@caltech.edu (A. Stuart), bhatta@caltech.edu (K. Bhattacharya).

https://doi.org/10.1016/j.jmps.2026.106807
Received 13 May 2026; Received in revised form 8 July 2026; Accepted 6 August 2026
Available online 8 August 2026

hardening; the framework is also extended to learn response of the RVE to varying microstructure. In Ghavamian and Simone (2019) academic examples are used to demomnstrate that RNNs can be used as surrogate models for the micro-level problem, thereby reducing the computation cost of FE^2 in a two-scale material simulation. Further literature on the use of RNNs to accelerate the computation of two-scale material simulation includes Logarzo et al. (2021), Wu et al. (2020). In Wu et al. (2020) the effective response of a fiber reinforced matrix is learned; the fiber is assumed to be hyperelastic, and the matrix to obey J2 plasticity. In Im et al. (2021) a framework is developed which uses the long short-term memory (LSTM) architecture to learn surrogate models for elasto-plastic materials. The paper Bonatti and Mohr (2022) develops an architecture that is consistent at different time discretizations. More recent work on the use of NN can be seen in the As'ad and Farhat (2026) for a woven fabric of viscoelastic material, and Liu et al. (2023) for plastic material with history-dependent stress-strain response. In the data-driven approach to homogenisation, the response of the unit cell is fit to a neural network model that is used as the constitutive law in a macroscale FE solver.

Apart from the homogenisation setting, neural networks are also being used to fit a constitutive model to experimental data. Historically, many finite-deformation constitutive laws have been developed to model the macroscopic non-linear response of materials. However, these classical non-linear constitutive laws are designed for very specific materials and usually do not generalize well to similar, but different, materials. Modeling finite-deformation constitutive laws becomes even more complicated for material behavior that depends on the history of the applied load. History-dependent behavior may arise in cases such as overall viscoelastic response of a rapidly varying continua (Francfort and Suquet, 1986; Brenner and Suquet, 2013) and plasticity (Simo and Hughes, 2006). Hence, there has been an attempt to leverage the generalisation ability of neural networks to learn the constitutive response of a wide class of materials.

In this work, the constitutive law of the homogenized response of a polycrystalline unit cell, subject to finite deformation, is learnt using NN models. The mechanical response of metals under finite deformation is inelastic and hence is modeled by introducing a set of internal variables. Internal variables are used in the thermodynamics formalism of mechanics for modelling inelastic materials. The idea is to expand the set of thermodynamic state variables by a set of so-called internal variables $(\xi_1, \dots, \xi_n)$ to model the transition between two thermodynamic equilibrium states. The first application of internal variables in mechanics can be traced back to Coleman and Gurtin (1967), where thermodynamics restrictions on the internal variables are derived, and their connection with dynamical stability is analyzed. Application of this theory, in particular, in metal plasticity is done in Rice (1971), where internal variables are used to represent the average behavior of local microstructural rearrangements due to dislocations. A historical review of internal state variable theory can be found in Horstemeyer and Bammann (2010). In the classical approach, internal variables and their evolution equations are postulated *a priori* as a part of the constitutive law. Their evolution equation is calibrated using experimental observations such as yield in metal plasticity. At the same time, internal variables could also arise from an approximation theoretic approach towards the approximation of integrals that capture history-dependence. The physical meaning of internal variables depends on their application. In metal plasticity, internal variables are postulated to represent plastic strain that, in turn is believed to capture average deformation due to dislocation rearrangements. However, in the case of viscoelasticity, internal variables arise out of the approximation of integral Bhattacharya et al. (2023) and hence, do not represent any physical or experimentally measurable quantity. In the more recent data-driven approach towards constitutive modeling, internal variables and their evolution law are discovered from data.

Recurrent neural network-based constitutive laws that directly map strain to stress or displacement to force do not impose *a priori* by architecture (a) the second law of thermodynamics, (b) material stability, and (c) mathematical conditions that guarantee the existence of solutions of the governing equations. These mathematical conditions are described for hyperelastic material in Ball (1977, 1976) and extended to history-dependent material in Mielke (2003), Ortiz and Stainier (1999). The existence of a solution for history-dependent governing equations relies on the polyconvexity of the energy potential (Ball, 1976) and the coercivity of both the energy and dissipation potentials (Mielke, 2004; Francfort and Mielke, 2006; Mainik and Mielke, 2005; Mielke et al., 2018). The research reported in this paper builds on the existing literature on modeling the homogenized macroscopic mechanical response of a polycrystal. The response is assumed to be viscoplastic, and a mapping from strain history to stress is sought. To this end, a novel NN framework, based on internal variable theory, is introduced. The main contributions of the work are now summarised.

(C1) A NN framework is introduced, consistent with the second law of thermodynamics, material stability, objectivity, and the theory underpinning the existence of a solution of the non-linear momentum balance equation, to learn a constitutive law map of strain history to stress.

There has been recent work where some of these constraints are imposed: polyconvexity, material stability, and objectivity for hyperelastic material in As'ad et al. (2022), Klein et al. (2022), the second law of thermodynamics in the case of one-dimensional elasto-plastic material in Flaschel et al. (2025) and two-dimensional viscoelastic material in As'ad and Farhat (2023). However, this framework has not been tested on three-dimensional metal visco-plasticity in the two-scale homogenisation setting. Latest approach towards visoplastic homogenisation is introduced in Liu et al. (2023), Karimi and Bhattacharya (2024) where NN framework invariant in time-discretisation is to learn history-dependent constitutive laws by learning evolution of internal variables. However, the architecture there is not consistent with the foregoing mathematical properties.

(C2) The polyconvex model is trained on the space of symmetric deformation gradients and is extended to the space of non-symmetric deformation gradients in a manner consistent with objectivity.

The current literature shows there are no abstract forms of constitutive laws that are both polyconvex and objective. Both are simultaneously achieved by invariant-based models where more assumptions on material symmetry are made such as in Schröder et al. (2010). A data-driven approach to imposing both polyconvexity and objectivity uses an objectivity-based additional term

in the loss function as in Klein et al. (2022). The data-driven approach performs poorly in terms of stress-prediction (Klein et al., 2022), hence we present an approach where objectivity, along with polyconvexity, is achieved by definition of energy function.

(C3) The framework is demonstrated by fitting the Taylor-averaged response of a polycrystalline magnesium microstructure, achieving a relative error of 2%, and resulting in a substantial speed-up over the original model without any compromise on accurate prediction.

(C4) Empirically, it is shown that the internal variables are identifiable up to a linear transform; theoretical results supporting this observation are provided.

In the case of inelastic material, it has been reported that internal variables learned from data are not unique (Flaschel et al., 2025); however, there is no theoretical characterisation of the set of these different but equivalent sets of internal variables.

The paper is organized as follows. Section 2 provides the mathematical background concerning modeling inelastic history-dependent mechanical response, which includes the second law of thermodynamics, polyconvexity, growth properties of potential pertinent to material stability, and objectivity. We also present, in Section 2, the application of the Legendre transform in the formulation of explicit kinetics. Section 3 presents the NN framework, (C1), that covers a description of the NN architecture used to parameterize various potentials, methods to enforce objectivity and polyconvexity of the energy potential, the method to enforce convexity of the dissipation potential, and the functions used to enforce the growth of potentials; in the same section we also provide details of data generation and training (C2). Section 4 presents (C3) and (C4), demonstrating the advantage of the proposed energy decomposition and efficacy of the resulting framework in terms of achieved prediction accuracy. The empirical verification of the identifiability of internal variables up to linear transform and supporting theoretical results (C5) are also contained in Section 4. Finally, a concluding discussion on limitations and further scope of this work is presented in Section 5.

2. Mathematical framework

Let $\mathbf{F}(t) \in \mathbb{R}^{3\times3}$ and $\mathbf{P}^\dagger(t) \in \mathbb{R}^{3\times3}$ denote, respectively, the deformation gradient and first Piola-Kirchhoff stress. The behavior of a viscoplastic material is a non-Markovian map of the form:

$$\Psi^\dagger : \{\mathbf{F}(\tau) : \tau \in (0,t)\} \mapsto \mathbf{P}^\dagger(t), \tag{1}$$

The superscript † is used throughout this article to distinguish the true map, from its approximation. In Section 2.1, we lay out a Markovian formulation to model the history-dependent constitutive law, and state the physical properties desired from such a constitutive law. An explicit formulation of the dynamics, capturing the history-dependence and employing internal variables, is presented in Section 2.2.

2.1. Markovian constitutive laws

The map (1) is non-Markovian when used as a constitutive law in the balance of linear momentum equation. However, for interpretability and computational efficiency, it is desirable to deploy Markovian models whenever possible. Of course, the choice of such a Markovian approximation must allow for an adequate representation of physics. Markovian approximation of non-Markovian models is often done by introducing a set of internal variables $\xi(t) \in \mathbb{R}^n$.

In an energetic formulation, the stress and evolution of internal variables are determined from derivatives of the energy density potential $\mathcal{W} : \mathbb{R}^{3\times3} \times \mathbb{R}^n \to \mathbb{R}$ and the dissipation potential $\widehat{D} : \mathbb{R}^n \to \mathbb{R}$. Specifically

$$\mathbf{P}(t) = \frac{\partial \mathcal{W}(\mathbf{F}(t), \xi(t))}{\partial \mathbf{F}}, \tag{2a}$$

$$\frac{\partial \widehat{D}(\dot{\xi}(t))}{\partial \dot{\xi}} = -\frac{\partial \mathcal{W}(\mathbf{F}(t), \xi(t))}{\partial \xi}, \quad \xi(0) = 0. \tag{2b}$$

We use $\dot{()}$ to denote the time derivative, as in $\dot{\xi}(t)$. It should be noted that $\mathbf{F}, \xi, \dot{\xi}$ are used as dummy variables for arguments of $\mathcal{W}$ and $\widehat{D}$; whereas $\mathbf{F}(t), \xi(t)$ and $\dot{\xi}(t)$ represents these variables at time t. We follow this convention throughout this paper. This energetic Markovian formulation leads to several desirable properties that we now list.

The Second Law of Thermodynamics. The second law of thermodynamics, in the form of the Clausius–Duhem inequality, states that the dynamics of the process must be dissipative. Assuming the process to be isothermal, the Clausius–Duhem inequality can be written as

$$\frac{\partial \widehat{D}(\dot{\xi}(t))}{\partial \dot{\xi}} \cdot \dot{\xi}(t) \geq 0. \tag{3}$$

A sufficient condition that ensures that (3) is satisfied for all $t \in \mathbb{R}^+$ is that $\widehat{D}$ be a convex function with a minimum at zero. The method to impose convexity is presented in Section 3.2.

Objectivity. The mechanical response of a material must be indifferent to the reference frame of an observer. As a consequence, it must satisfy $\mathcal{W}(\mathbf{QF}, \cdot) = \mathcal{W}(\mathbf{F}, \cdot)$ for all $\mathbf{Q} \in \mathrm{SO}(3)$. Therefore, $\mathcal{W}$ must admit a representation such that $\mathcal{W}(\mathbf{F}, \xi) = \widehat{\mathcal{W}}(\mathbf{C}, \xi)$ where $\mathbf{C} := \mathbf{F}^\top\mathbf{F}$, is called the Cauchy strain tensor. Eqs. (2a) and (2b) can be written in terms of the Cauchy strain tensor to satisfy objectivity. To this end, we have

$$\mathbf{S}(t) = \frac{\partial \mathcal{W}(\mathbf{C}(t), \xi)}{\partial \mathbf{C}} \tag{4a}$$

$$\frac{\partial \widehat{D}(\dot{\xi}(t))}{\partial \dot{\xi}} = -\frac{\partial \mathcal{W}(\mathbf{C}(t), \xi(t))}{\partial \xi}, \quad \xi(0) = 0, \tag{4b}$$

where the symmetric tensor $\mathbf{S} := \mathbf{F}^{-1}\mathbf{P}$ in (4a) is called the second Piola-Kirchhoff stress. The benefit of formulating the problem in terms of $\mathbf{C}$ is that it reduces the dimensionality of the computation, as both $\mathbf{S}$ and $\mathbf{C}$ are symmetric (in contrast to their counterparts $\mathbf{P}$ and $\mathbf{F}$). At the same time, a drawback of this formulation is that, without making any further simplifying assumptions, it is hard to restrict $\mathcal{W}$ to be a polyconvex function of $\mathbf{F}$. In this work, we learn the energy $\mathcal{W}$ on the space of symmetric deformation gradients and use polar decomposition to extend the function to non-symmetric deformation gradients in a manner that is consistent with objectivity. This extension is presented in Section 4.2.4 and further discussion is done in Section B.3.

Polyconvexity. The notion of polyconvexity was introduced in the seminal work on existence theorems for nonlinear elasticity momentum balance PDEs (Ball, 1976, 1977). Energy density $\mathcal{W}(\mathbf{F}, \xi)$ is polyconvex in $\mathbf{F}$ iff it admits a representation $\mathcal{W}(\mathbf{F}, \xi) = g(\mathbf{F}, \operatorname{adj}\mathbf{F}, \det\mathbf{F}, \xi)$ where $g(\cdot, \cdot, \cdot, \xi)$ is a convex function for all $\xi \in \mathbb{R}^n$. The polyconvexity of the energy density potential allows for characterisation of a wide range of physical assumptions under which the non-linear momentum balance equation has a solution. Polyconvexity is imposed by imposing convexity, which is discussed in Section 3.2, and more discussion on the polyconvex model is done in Section 3.3. Apart from polyconvexity, the existence of a solution also relies on certain growth properties of the potential, which are discussed next.

Material Stability. Apart from being a necessity for the existence theorems as mentioned in the previous paragraph, other growth conditions may also be desired to prevent unrealistic mechanical behavior when a material is subjected to extreme strain; hence the name 'material stability'. Here, we discuss two different growth conditions that have physical interpretations and significance. It is expected that

$$\frac{\mathcal{W}(\mathbf{F}, \xi)}{|\mathbf{F}|^3} \to \infty \quad \text{as} \quad |\mathbf{F}| \to \infty; \tag{5}$$

where $|\cdot|$ represents the Frobenius norm. The limit in (5) implies that a line segment of positive length cannot be produced from an infinitesimal cube using a finite amount of energy. Another growth condition, which is not necessary for the existence theorems, but is physically desirable, is as follows

$$\mathcal{W}(\mathbf{F}, \xi) \to \infty \quad \text{as} \quad \det\mathbf{F} \to 0^+. \tag{6}$$

The limit in (6) ensures that a piece of material cannot be compressed to a point with a finite amount of energy or compressive stress. These limiting behaviors are imposed using certain functions that are presented in Section 3.1.

We constrain the neural network surrogates of $\mathcal{W}$ and $\widehat{D}$ to be consistent with the second law of thermodynamics and material stability. Before we describe the NN parameterisation of $\mathcal{W}$ and $\widehat{D}$, and discuss the training procedure, it is worthwhile to note that both (2b) and (4b) are implicit ODEs. However, owing the convexity of $\widehat{D}$ as a consequence of the second law of thermodynamics, it is possible to derive an equivalent explicit form of (2b) and (4b). This is the subject of discussion in the next Section 2.2.

2.2. Legendre transform and explicit formulation of dynamics

Since (2b) is an implicit dynamical system, we use the convexity of $\widehat{D}$ to invert (2b). To this end, we work with the Legendre transform (Simo and Hughes, 1998; Rockafellar, 1970) of $\widehat{D}$: we define $\mathcal{D} : \mathbb{R}^n \to \mathbb{R}$ by

$$\mathcal{D}(d) = \sup_{\dot{\xi} \in \mathbb{R}^n} \left(\dot{\xi} \cdot d - \widehat{D}(\dot{\xi}) \right). \tag{7}$$

If $\widehat{D}$ is continuously differentiable,

$$d = \frac{\partial \widehat{D}(\dot{\xi})}{\partial \dot{\xi}}; \quad \dot{\xi} = \frac{\partial \mathcal{D}(d)}{\partial d}. \tag{8}$$

Using (8), (2) can be written as

$$\mathbf{P}(t) = \frac{\partial \mathcal{W}(\mathbf{F}(t), \xi(t))}{\partial \mathbf{F}} \tag{9a}$$

$$d(t) = -\frac{\partial \mathcal{W}(\mathbf{F}(t), \xi(t))}{\partial \xi} \tag{9b}$$

$$\dot{\xi}(t) = \frac{\partial \mathcal{D}(d(t))}{\partial d}, \quad \xi(0) = 0. \tag{9c}$$

Hence, when working with (9), learning $\mathcal{D}$ is equivalent to learning $\widehat{D}$. Once $\mathcal{W}$ and $\mathcal{D}$ in (9) are learned from data using samples of $(\mathbf{F}, \mathbf{P}^\dagger)$, we obtain the trajectories of internal variables $\xi(t)$ as a by-product of the prediction of stress from strain. These internal variables are not uniquely identifiable given the dataset. In Section 4.3, we present an analysis that asserts that the internal variables learned by (9) from data are unique up to a linear transform.

3. Methodology

In Section 3.1, we present a method to ensure material stability by incorporating growth in the learned constitutive law. Section 3.2 describes different architectures of neural networks that are used as building blocks in parameterizing $\mathcal{W}$ and $\mathcal{D}$. Section 3.3 is the

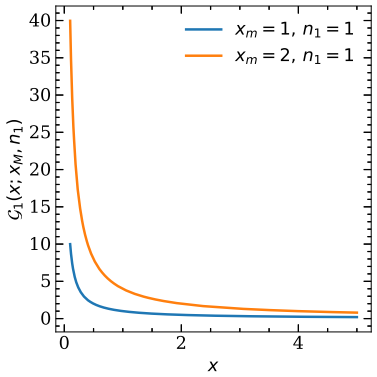

Fig. 1. Visualisation of $\mathcal{G}_1$ for different parametric values of x_m.

decomposition of the energy function into isochoric and volumetric components. This decomposition turns out to be beneficial for accurate prediction. Details about the dataset and the NN training are presented, respectively, in Sections 3.4 and 3.5.

3.1. Functions to impose material stability

A NN can be chosen to approximate any continuous function arbitrarily well in the support of the dataset (Pinkus, 1999). However, an optimized NN does not necessarily have the desired growth outside of the support of the dataset; this issue is addressed for dissipative fluid mechanics problems in Li et al. (2022). Therefore, the growth of the energy potential, as discussed in Section 2.1, cannot be learned by a NN from the data and must be incorporated separately. The two methods proposed in Li et al. (2022) to control the growth of an NN model are either (a) to augment the dataset with artificial data generated from an artificial model with desired properties (growth in our case), or (b) by adding functions to the NN model which enforce the desired properties without interfering with the learning process. We adopt the latter approach. To this end, for $x \in \mathbb{R}$, the functions $\mathcal{G}_i, i = 1, 2$ are defined. These functions $\mathcal{G}_1 : \mathbb{R}^+ \to \mathbb{R}$ and $\mathcal{G}_2 : \mathbb{R} \to \mathbb{R}$ have the property that they are almost constant in the support of the dataset, but have polynomial growth outside the support. We set

$$\mathcal{G}_1(x; x_m, n_1) = \frac{1}{n_1} \frac{x_m^{n_1+1}}{x^{n_1}}; \quad x > 0,\ x_m > 0, \tag{10}$$

where

$$x_m = \min_{x \in \text{Dataset}} x. \tag{11}$$

We note that, for x_m, the minimum must be taken over a positive-valued quantity, denoted by the dummy variable x in (11), such as the Jacobian of the deformation gradient, derived from the samples in the dataset. Real non-negative number $n_1 \in \mathbb{R}^+$ controls the singular behaviour at the origin. We also define $\mathcal{G}_2 : \mathbb{R} \to \mathbb{R}$ as

$$\mathcal{G}_2(x; x_M, \alpha, n_2) = \frac{x_M}{\alpha n_2} \log(1 + \exp(\alpha(\left|x/x_M\right|^{n_2} - 1))), \tag{12}$$

where

$$x_M = \max_{x \in \text{Dataset}} |x|, \tag{13}$$

and $n_2 \in \mathbb{R}^+$ is the desired order of growth. x_M is a measure of the size of the data support, and the parameter $\alpha \in \mathbb{R}^+$, which controls the transition to growth on the boundary of the support, needs to be appropriately chosen. We visualize the two functions $\mathcal{G}_i; i = 1, 2$. Fig. 1 shows the parametric dependence $\mathcal{G}_1(x; x_m, n_1)$ on x_m, for fixed n_1. Fig. 2 shows parametric dependence of $\mathcal{G}_2(x; x_M, \alpha, n_2)$ on x_M, α and n_2. The parameter x_M controls the size of the flat part of this function, n_2 is the order of singularity, and α controls how rapidly the function transitions from the flat part to its polynomial behavior. Recall that the parameters x_m and x_M are chosen in such a way that the dataset lies in the relatively flatter part of $\mathcal{G}_1$ and $\mathcal{G}_2$. This ensures that $\mathcal{G}_1$ and $\mathcal{G}_2$ enforce the desired growth of potentials without interfering with the learning, that happens in the support of the dataset, of NNs $\mathcal{N}_H$ and $\mathcal{N}_V$. We show the use of these functions leads to physically desirable predictions that concern the staibility of a material under extreme values of applied strain $\mathbf{F}$ in Section 4.2.3.

Instead of adding the $\mathcal{G}_i$, $i = 1, 2$, to the NNs, it is computationally beneficial and more accurate to add their derivatives to the appropriate stresses derived from the NNs; this enables us to bypass the need for automatic differentiation. To this end, the derivatives of $\mathcal{G}_i$, $i = 1, 2$, denoted respectively by $\mathcal{H}_1 : \mathbb{R} \to \mathbb{R}$ and $\mathcal{H}_2 : \mathbb{R} \to \mathbb{R}$ are given as

$$\mathcal{H}_1(x; x_m, n_1) := \frac{\partial \mathcal{G}_1(x; x_m, n_1)}{\partial x} = -\left(\frac{x_m}{x}\right)^{n_1+1}; \tag{14a}$$

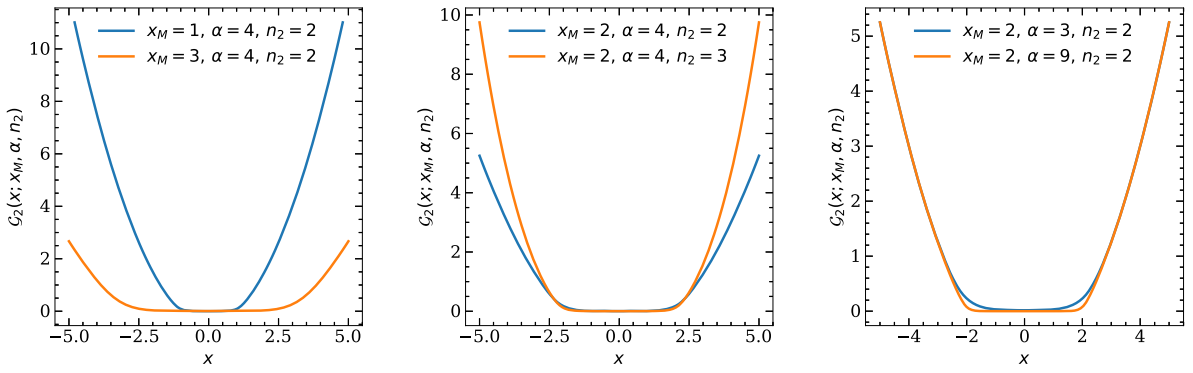

Fig. 2. Effect of different parameters used in $\mathcal{G}_2$ on its gorwth. $\mathcal{G}_2$ is used to enforce the growth of the potential outside the support of the data.

$$\mathcal{H}_2(x; x_M, \alpha, n_2) := \frac{\partial \mathcal{G}_2(x; x_M, \alpha, n_2)}{\partial x} = \frac{|x|^{n_2-2}x}{x_M^{n_2-1}(1 + \exp(-\alpha(|x/x_M|_2^{n_2} - 1)))}. \tag{14b}$$

We will use the notational shorthands κ and Ξ to denote (x_m, n_1) and (x_M, α, n_2), respectively. The functions described here are used , along with NNs, to parameterize energy and dissipation potential $\mathcal{W}$ and $\mathcal{D}$. Before we discuss the exact parameterisation in Section 3.3, we present the NN architectures at an abstract level in Section 3.2, that are used in the parameterisation of potentials.

3.2. Neural network architecture

In this Subsection, we discuss the NN architecture used in this work: ($\mathcal{A}_1$) a NN that is convex in terms of its arguments.

$\mathcal{A}$. **Input Convex Neural Network.** We use a shallow (meaning single hidden layer) architecture of the input convex NN described in Amos et al. (2017). The architecture can be written as

$$h_1 = \text{ReLUS}\big(W_1 h_0 + b_1\big); \tag{15a}$$

$$h_2 = W_2 h_1 + b_2. \tag{15b}$$

The elements of weight and bias matrices, W_1 and b_1, are real-valued, while those of W_2 are positive real numbers. The activation function ReLUS$(\cdot)$ is the square of the ReLU activation function, i.e., for $x \in \mathbb{R}$, ReLUS$(x) = x^2 \mathbb{1}_{\{x>0\}}$. The positivity of the element of W_2 is ensured by writing it as the element-wise square of another matrix, i.e, $W_2 = \widetilde{W}_2 \odot \widetilde{W}_2$, where $\widetilde{W}_2$ takes values in a real space. It is to be noted that the positivity of W_2 is intended to enforce convexity of the architecture. We also highlight that the choice of activation function, ReLUS compared to the popularly used ReLU is because the architecture, in this work, is used to parameterize potentials; samples of which are not available. Hence, these NNs are trained on samples of their derivatives, which are stresses. Therefore, this choice of activation function ensures that the gradients of the loss function, which will be in terms of the double derivatives of the activation function, do not vanish. Moreover, noting that the ReLUS$(\cdot)$ has quadratic growth, the activation function limits the architecture to have just one hidden layer; as the growth of the NN would be 2^{L-1} when L hidden layers are used, consequently, making the training of such architectures next to impossible due to blow-ups during their optimisation. More disucussion on this is given in Appendix B.1. Architectures described here will be used to parameterize functions $\mathcal{W}$, and $\mathcal{D}$, described in Section 3.3.

3.3. Polyconvex energy function

We parameterize the potential $\mathcal{W}$, in the following manner to enforce polyconvexity. We adopt a hydrostatic-deviatoric decomposition of $\mathcal{W}$. To this end, we have

$$\mathcal{W}(\mathbf{F}, \xi) = \mathcal{W}_H\big(I_F\big) + \mathcal{W}_V(\mathbf{K}, \xi) \tag{16}$$

where $I_F = \det \mathbf{F}$ and $\mathbf{K} := \mathbf{F}/I_F^{1/3}$; $\mathcal{W}_H : \mathbb{R} \to \mathbb{R}$, $\mathcal{W}_V : \mathbb{R}^{d\times d} \times \mathbb{R}^n \to \mathbb{R}$. We define $\widetilde{\mathbf{P}} := \mathbf{PF}$, and then, straightforward calculation leads to

$$\text{hyd}\,\widetilde{\mathbf{P}} = I_F \frac{\partial \mathcal{W}_H(I_F)}{\partial I_F}\mathbf{I}_3, \tag{17a}$$

$$\text{dev}\,\widetilde{\mathbf{P}} = \frac{1}{I_F^{1/3}}\left(\frac{\partial \mathcal{W}_V(\mathbf{K}, \xi)}{\partial \mathbf{K}}\mathbf{F} - \left(\frac{\partial \mathcal{W}_V(\mathbf{K}, \xi)}{\partial \mathbf{K}} \cdot \mathbf{F}\right)\mathbf{I}_3\right). \tag{17b}$$

To incorporate limiting behaviour, we further decompose $\mathcal{W}_H$ and $\mathcal{W}_V$ as

$$\mathcal{W}_H(I_F) := \mathcal{N}_H(I_F) + \mathcal{G}_1(I_F; \kappa_2) + \mathcal{G}_2(I_F, \Xi_3), \tag{18a}$$

$$\mathcal{W}_V(\mathbf{K},\xi) := \mathcal{N}_V(\mathbf{F},\operatorname{adj}\mathbf{F},I_F,\xi)+\mathcal{G}_2(|n_{\mathbf{F}}|;\Xi_4); \tag{18b}$$

where $n_{\mathbf{F}} := (\mathbf{F},\operatorname{adj}\mathbf{F},I_F)$ represents vectorized concatenation; $\mathcal{N}_H : \mathbb{R}\to\mathbb{R}$ and $\mathcal{N}_V : \mathbb{R}^{d\times d}\times\mathbb{R}^{d\times d}\times\mathbb{R}^+\times\mathbb{R}^n\to\mathbb{R}$. $\mathcal{W}_H$ is trained on the samples of

Polyconvexity of $\mathcal{W}(\mathbf{F},\xi)$ in terms of $\mathbf{F}$ is ascertained by using convex NN architecture, $\mathcal{A}$, for both $\mathcal{N}_H$ and $\mathcal{N}_V$. $\mathcal{N}_H$ is trained on samples of $(I_F,\operatorname{hyd}\widetilde{\mathbf{P}}^\dagger)$, while $\widetilde{\mathcal{N}_V}$ is trained on samples of $(\mathbf{F},\operatorname{dev}\widetilde{\mathbf{P}}^\dagger)$. The dissipation potential is parameterized as follows

$$\mathcal{D}(\dot\xi) = \mathcal{N}_D(\dot\xi)+\mathcal{G}_2(\dot\xi;\Xi_5), \tag{19}$$

where $\mathcal{N}_D$ is parameterized using architecture $\mathcal{A}_2$ and $\mathcal{G}_2$, with appropriate parameters Ξ_5, is used for coercivity of $\mathcal{D}$. A benefit of our data-driven approach is that it does not require us to make any *a priori* distinction between the elastic and plastic responses of the material. The clear delineation between these two components of the model, which is commonplace in traditional models of elasto-plastic materials, simplifies conceptual modeling, and has been beneficial for this reason, but is arguably not always justified from first principles. In our approach the stored energy can depend explicitly on the state variables, so plasticity can influence elastic response; as a consequence, we cannot train the elastic and plastic components of the model separately. Moreover, our neural-network-based approach, while parameter-intensive, does not require any assumption on the material model class and hence can be deployed to model a wide range of materials.

3.4. Data

The method used to generate data for this study is ad-hoc but based on physical intuition. It is to be noted that internal variables ξ are physically unobserved, and hence the dataset comprises only deformation gradient and stress trajectories. We denote the dataset as

$$\mathrm{D} = \left\{(\mathbf{F}^{(i)},\mathbf{P}^{\dagger(i)})\right\}_{i=1}^{N}$$

where, for a natural number M, $\mathbf{F}^{(i)} : \{0,\dots,M\}\to\mathbb{R}^d$ and $\mathbf{P}^{\dagger(i)} : \{0,\dots,M\}\to\mathbb{R}^d$ are discretized trajectories of deformation gradient and stress. It is to be noted that the time step associated with the dataset is chosen to be $\Delta t = 1/M$. The data comes from solving a finer-scale crystal plasticity model of a representative volume element (unit cell). Each unit cell is a three-dimensional polycrystalline cube with 128 randomly oriented grains. Each grain has three basal, three prismatic, and six pyramidal slip systems, in addition to six tensile twin systems (treated as pseudoslips). The material parameters were chosen to represent magnesium. The homogenized stress is obtained by using Taylor's averaging, where the solution of the equilibrium equation is replaced by the assumption that the deformation gradient is uniform across the unit cell (Kocks et al., 2000; Kovachki et al., 2022). Details of the data generation can be found in (Liu et al., 2022). This dataset constitutes an interesting numerical homogenisation problem for two primary reasons. Firstly, the existence of any true potentials $\mathcal{W}^\dagger$ and $\mathcal{D}^\dagger$ for the homogenized response is not established; so finding them computationally provides insight. Secondly, the dimensionality, n, and samples of internal variables are unknown, and again computation provides insight. The dataset is decomposed into mutually disjoint train and test sets.

3.5. Training the neural networks

Here we present details pertaining to training the NNs. The function $\mathcal{W}$ and $\mathcal{D}$ describing the mapping from $\mathbf{F}$ to $\mathbf{P}$ in (4), are to be learned through the NNs $\mathcal{N}_H,\mathcal{N}_V$ and $\mathcal{D}$ described in Section 3.3. The NN $\mathcal{N}_H$, concerning only the hydrostatic stress that is independent of the strain-history and hence of the internal variable ξ, is trained on the input-output pairs $(I_3,\operatorname{hyd}\widetilde{\mathbf{P}}^\dagger)$, the mapping between them is given in (17a). To this end, using (17a), the loss is defined as

$$\mathcal{L}_1(\Theta_1) = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=0}^{M}\left(\operatorname{hyd}\widetilde{\mathbf{P}}_m^{\dagger(n)}-\operatorname{hyd}\widetilde{\mathbf{P}}_m^{(n)}\right)^2 \tag{20a}$$

where n and m are respectively indexing over trajectories and time steps. The optimum value, Θ_1^*, of the parameter is found by minimizing $\mathcal{L}_1$ i.e.,

$$\Theta_1^* = \underset{\Theta_1}{\operatorname{argmin}}\,\mathcal{L}_1(\Theta_1). \tag{21}$$

Notably, the contribution of growth functions $\mathcal{H}_1$ and $\mathcal{H}_2$ in stress is included in the loss $\mathcal{L}_1$. Now, we describe learning the NN $\mathcal{N}_V$, which, in contrast with $\mathcal{N}_H$, is coupled with the NN $\mathcal{D}$ through the ODE describing the dynamics of the internal variables. Hence, training the NNs $\mathcal{N}_V$ and $\mathcal{D}$ requires, firstly, solving a dynamical system and then optimizing their parameters. Using Eqs. (9) and (18b), the evolution in time of internal variable ξ and the deviatoric stress $\operatorname{dev}\widetilde{\mathbf{P}}$ is given as

$$\operatorname{dev}\widetilde{\mathbf{P}} = \frac{\partial\mathcal{N}_V(\mathbf{F},\operatorname{adj}\mathbf{F},I_F,\xi)}{\partial\mathbf{F}}+\frac{\partial\mathcal{G}_2(|n_F|;\Xi_4)}{\partial\mathbf{F}} \tag{22a}$$

$$\dot\xi(t) = \frac{\partial\mathcal{D}(d;\Theta_3)}{\partial d},\quad \text{where } d = -\frac{\partial\mathcal{N}_V(\mathbf{F},\operatorname{adj}\mathbf{F},I_F,\xi;\Theta_2)}{\partial\xi};\quad \xi(0)=0. \tag{22b}$$

The loss corresponding to $\operatorname{dev}\widetilde{\mathbf{P}}$ is computed as

$$\mathcal{L}_2(\Theta_2,\Theta_3) = \frac{1}{MN}\sum_{n=1}^{N}\sum_{t=0}^{M}\frac{\left\|\operatorname{dev}\widetilde{\mathbf{P}}_m^{\dagger(n)}-\operatorname{dev}\widetilde{\mathbf{P}}_m^{(n)}\right\|_2^2}{\left\|\operatorname{dev}\widetilde{\mathbf{P}}_m^{\dagger(n)}\right\|_2^2}. \tag{23}$$

We note that the loss $\mathcal{L}_2$ is a function of both Θ_2 and Θ_3, as dev $\mathbf{P}$ depends on both the NNs $\mathcal{N}_V$ and D. It is thermodynamically desired for D to have minimum at zero. Therefore, we define a loss penalizing gradient at zero as follows

$$\mathcal{L}_3(\Theta_3) = \left\| \frac{\partial D(\mathbf{0}, \Theta_3)}{\partial d} \right\|_2^2. \tag{24}$$

However, we show in Section B.4 that this can be learned from data, without having $\mathcal{L}_3$ as a part of the objective to be minimized. NNs $\mathcal{N}_V$ and D are optimized as

$$\Theta_2^*, \Theta_3^* = \underset{\Theta_2, \Theta_3}{\operatorname{argmin}} (\mathcal{L}_2(\Theta_2, \Theta_3) + \lambda \mathcal{L}_3(\Theta_3)), \tag{25}$$

where $\lambda > 0$ is a weighting term. It is to be noted that minimisation of $\mathcal{L}_1$ is done independently of that of $\mathcal{L}_2$ and $\mathcal{L}_3$. Moreover, in Section B.4 and Fig. B.2, we show that minimisation of $\mathcal{L}_2$ is insensitive to the parameter λ, that dovetails with the idea that D can be learned to have a minimum at zero from data. In both the optimisation tasks, (21) and (25), the parameters $\Theta_{1,2,3}$ are randomly initialized, the backpropagation algorithm is used to compute gradients of the loss with respect to the parameters, and the Adam (Kingma and Ba, 2017) scheme is used to update the parameters. The training is implemented in PyTorch (Paszke et al., 2019). In the next Section, we present performance and evaluation of this polyconvex model, empirical evidence of the uniqueness of internal variables, and supporting theory.

4. Results and discussion

The surrogate NN constitutive laws developed in this work are deployed to predict the Taylor-averaged response of a three-dimensional magnesium polycrystal. Below, we list key results of this work and describe the order in which they are presented in this Section. The key results are as follows.

(R1) Polyconvex neural network achieve relative error of 2% in the prediction of Taylor-averaged response of a magnesium polycrystal, while being consistent with various desirable physical properties listed before in Section 2. The learned NN models also show the growth behavior achieved by using the functions $\mathcal{G}_{1,2}$.
(R2) The polyconvex model is learned on the dataset comprising only symmetric deformation gradients, and the model is extended to non-symmetric deformation gradients in a manner consistent with objectivity.
(R3) The internal variables are empirically observed to be unique up to a linear transformation, and supporting theory is provided.

Section 4.1 presents a qualitative analysis and visualisation of the dataset. We then present the results corresponding to the polyconvex model in Section 3.3. There, we show the predictive accuracy (R1) of the polyconvex model and visual comparision between true and predicted stress. In Section 4.2.4, we present (R3) objective extension of the polyconvex neural network. It is followed in Section 4.3 by empirical evidence of that indentifiability of internal variables up to a linear transform, supporting theory, and the behaviour of the model under extreme stress. Section 4.4 contains details on the computational cost associated with training of and prediction by NN models.

4.1. Data visualisation

Fig. 3 presents a sample from the dataset. Each sample constitutes components of the deformation gradient and stress tensor over a unit time interval. Though the deformation gradient tensor, $\mathbf{F}$, is not necessarily symmetric, the dataset is generated by sampling $\mathbf{F}$ to be symmetric. The samples, each of $\mathbf{F}$ and $\mathbf{P}$ are shown in Fig. 3.
We also visualise distributions of different quantities using box plots. It is observed from Fig. 4 that the spread of different components of $\mathbf{F}$, about their means, is comparable, as trajectory components are sampled in a similar manner. On the other hand, the diagonal components of stress $\mathbf{P}$ have a much larger spread compared to the off-diagonal components. If $\mathbf{P}$ were to be learned directly by parameterising the potential $\mathcal{W}$ as a NN, without the decomposition (16), this disparity in the spread of stress components would make the pertinent training of the NN hard. Hence, $\mathcal{W}$ is approximated by learning the NNs $\mathcal{N}_H$ and $\mathcal{N}_V$; where $\mathcal{N}_V$ is trained on samples of dev $\widetilde{\mathbf{P}}$. The right-most plot in Fig. 4 shows the spread of different components of dev $\widetilde{\mathbf{P}}$, which are more comparable among themselves, as compared to that in $\mathbf{P}$. This leads to enhanced predictive accuracy of this framework, which is presented later in Section 4.2.2, and Fig. 6.

4.2. Polyconvex model

In this section, we present results of the NN model described before in Section 3.3 that, by design, is polyconvex. Section 4.2.1 shows that six internal variables are sufficient to capture the history dependency of stress on strain. Section 4.2.2 shows high predictive ability of the polyconvex model. We also show that the deviatoric component of the stress contributes significantly more to the overall error as compared to the hydrostatic component. Section 4.3 shows that different sets of internal variables, obtained from two NN models differing in the initialisation of their parameters at the start of training, are related through a linear transform. Section 4.2.3 shows that, as a result of using growth functions, the surrogate model exhibits physically desirable behaviour when subjected to extreme strain.

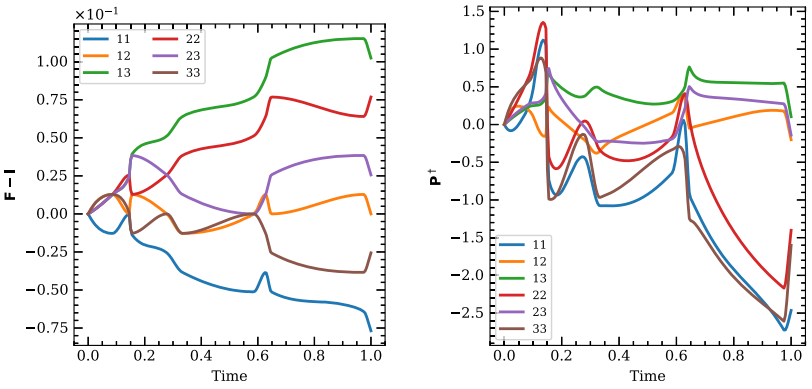

Fig. 3. A sample from the dataset. There are six components of the deformation gradient tensor $\mathbf{F}$ (left) as it is sampled to be symmetric and stress $\mathbf{P}^{\dagger}$ (right).

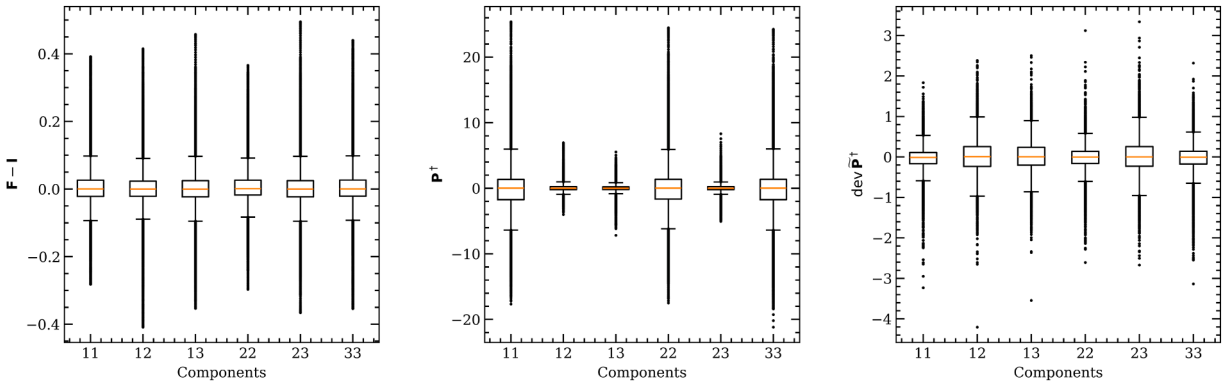

Fig. 4. Distribution of different components of $\mathbf{F}$, $\mathbf{P}^{\dagger}$ and $\mathrm{dev}\,\widetilde{\mathbf{P}}^{\dagger}$. There is more disparity in distributions of components of $\mathbf{P}^{\dagger}$ as compared to those of $\mathrm{dev}\,\widetilde{\mathbf{P}}^{\dagger}$.

4.2.1. *Identification of number of internal variables*

Internal variables $\xi(t) \in \mathbb{R}^n$ are not physically observed quantities. They are introduced to approximate the non-Markovian dependence of $\mathbf{P}$ on $\mathbf{F}$ in a Markovian fashion. Therefore, the dimensionality, n, of internal variables is not known *a priori*; rather it is to be identified by learning NNs with different values of n and choosen based on a balance between a desirable level of accuracy of the predicted $\mathbf{P}$ with respect to $\mathbf{P}^{\dagger}$ and dimension of the model. Fig. 5 presents the relative error in predicted stress $\mathbf{P}$ with respect to the dimensionality of the internal variables. All these NNs, corresponding to different values of n, were trained with their architecture and dataset size kept constant. $\mathcal{N}_V$ and D have 15000 nodes each in their only hidden layers. $\mathcal{N}_H$ has two hidden layers with 300 nodes each. The numbers of input nodes in $\mathcal{N}_V$ and D depend on n and, hence, is equal to $6+n$ and n respectively. It may be observed from Fig. 5 that $n=6$ is the optimal dimensionality of the internal variables for this numerical homogenisation problem, a further increase in n does not offer significant improvement in the prediction of stress, and the dimension $n=6$ does not impose an unwanted computational burden.

4.2.2. *Prediction of stress and error analysis*

After identifying the optimal dimensionality of internal variables and size of the NNs and dataset, we visualize the predictive accuracy of the objective framework and empirically analyze the error. The results correspond to the largest architecture that we trained. The number of nodes used in $\mathcal{N}_V$ and D is 15000 each and that in $\mathcal{N}_H$ is 300. Three thousand samples are used for training and testing the NN model. The dimensionality of the internal variables used is six. Fig. 6 shows comparative performances of frameworks without and with deviatoric-hydrostatic decomposition of energy. We observe in Fig. 6 that the relative error in stress, in the case without decomposition, is 15%, and that in the case with decomposition is 2%. Moreover, when the energy is decomposed, we observe that the error in $\mathrm{hyd}\,\widetilde{\mathbf{P}}$ is much lower than that in $\mathrm{dev}\,\widetilde{\mathbf{P}}$. The error in $\mathbf{P}$ depends on the errors both in $\mathrm{hyd}\,\widetilde{\mathbf{P}}$ and $\mathrm{dev}\,\widetilde{\mathbf{P}}$. Though the error in $\mathrm{dev}\,\widetilde{\mathbf{P}}$ is higher than that in $\mathrm{hyd}\,\widetilde{\mathbf{P}}$, its contribution to the error is $\mathbf{P}$ is not as much due to the lower magnitude of $\mathrm{dev}\,\widetilde{\mathbf{P}}$ compared to that of $\mathrm{hyd}\,\widetilde{\mathbf{P}}$. It is also observed that the errors on the training and test sets are close, reflecting the good generalisation of the model. Fig. 7 presents the visual comparison for a sample from the test dataset. The six plots compare the true stress components with their predicted values. Fig. 8 shows the visual comparison between true and predicted stress corresponding to the maximum error. We observe that the diagonal components of stress, where there is a systematic bias, contribute the most to the relative error.

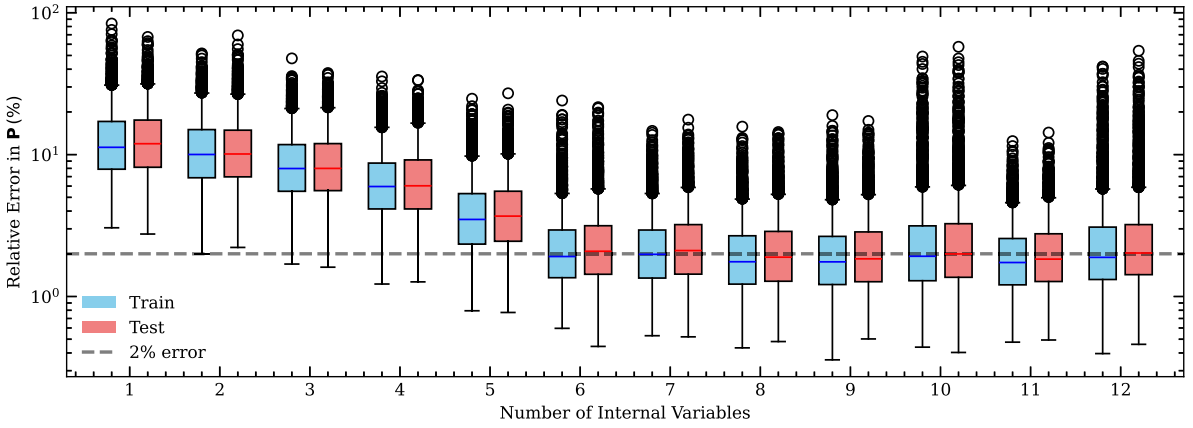

Fig. 5. The use of six internal variables is the optimal choice for maximum prediction accuracy.

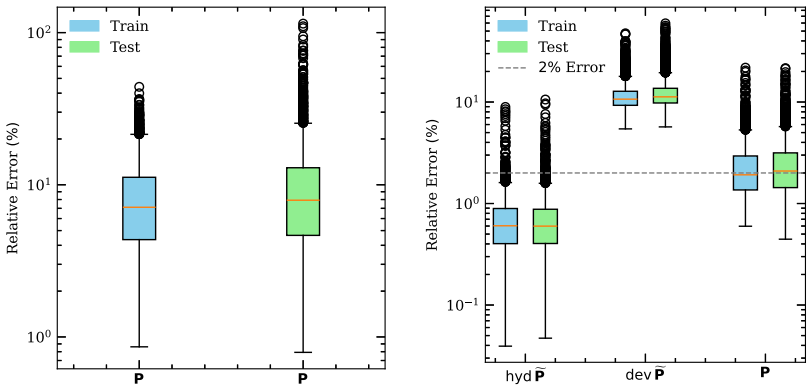

Fig. 6. Relative error in stress. (Left) Training without decomposition of $\mathcal{W}$ into its deviatoric-hydrostatic part. (Right) Training with decomposition of energy into deviatoric-hydrostatic parts as discussed in Section 3.3. It is observed that the decomposition of energy results in a significant improvement in the accuracy of the framework.

We also observe that the stress in the Fig. 8, which is the failure mode, is two orders of magnitude smaller than that in Fig. 7, which corresponds to median error. As the bias is in the diagonal components, it could be attributed to learning of $\mathcal{N}_H$, where the loss $\mathcal{L}_1$, compared to $\mathcal{L}_2$ is not relative. Hence, the small magnitude of stress in the failure mode is not well captured by the $\mathcal{L}_2$ loss. Use of relative loss for $\mathcal{L}_2$ might improve the accuracy of the failure modes and reduce the length of tails of box plots in Fig. 5.

4.2.3. *Visualisation of material stability*

In this Section, we present the effects of using growth functions $\mathcal{G}_i; i = 1, 2$, described in Section 3.1, in the prediction of stress for large strain values. Recall that $\mathcal{G}_i; i = 1, 2$ are used to enforce the physically reasonable property of any constitutive law that a piece of material cannot be compressed to a point with finite stress. Fig. 9 shows the effects of using these functions with $\mathcal{N}_H$ (see (18a)), the hydrostatic part of the model. It can be observed from Fig. 9 that, if $\mathcal{G}_1$ is not used while training, i.e., growth is not imposed, the predicted stress is finite at $\det \mathbf{F} = 0$. This is physically unreasonable, as it implies that a material can be compressed to a point with finite stress, hence would be called an unstable material. Therefore, this problem is rectified by using $\mathcal{G}_1$. From Fig. 9, it is observed that the hydrostatic stress $\mathrm{hyd}\,\widetilde{\mathbf{P}}$ blows up to negative infinity as $\det \mathbf{F}$ approaches zero from the right side.

Finally, we present the effect of using $\mathcal{G}_2$ with $\mathcal{N}_V$ (see (18b)). To understand this, we examine the predicted stress for large shear strain. To this end, we fix the internal variable to be zero and predict stress along $\mathbf{F}$ corresponding to pure shear and simple shear directions. Therefore, along pure shear and simple shear directions, we have

$$\text{(Pure shear)} \quad \mathbf{F}_\lambda = \begin{bmatrix} 1 & \lambda & 0 \\ \lambda & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}; \quad \text{(Simple shear)} \quad \mathbf{F}_\lambda = \begin{bmatrix} 1 & \lambda & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{26}$$

From Fig. 10, we observe the desired growth in stress components for pure shear in the $1-2$ plane. Fig. 11 shows the growth of stress in a simple shear direction. From both Figs. 10 and 11, we observe that normal components of stress along all three directions

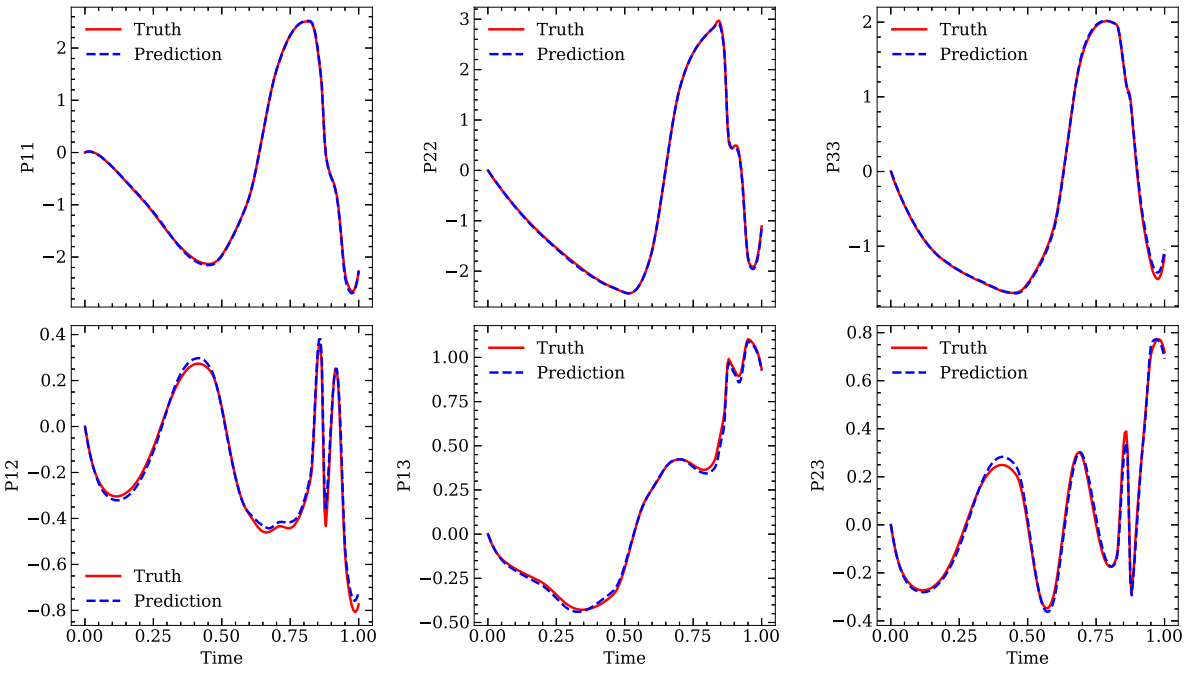

Fig. 7. Visual comparison of different components of true and predicted stress. The sample corresponds to the median error of 2%. The plots in the top row show diagonal components of the stress tensor. The bottom row shows the off-diagonal components. The comparison corresponding to the sample with maximum error is shown in Fig. 8.

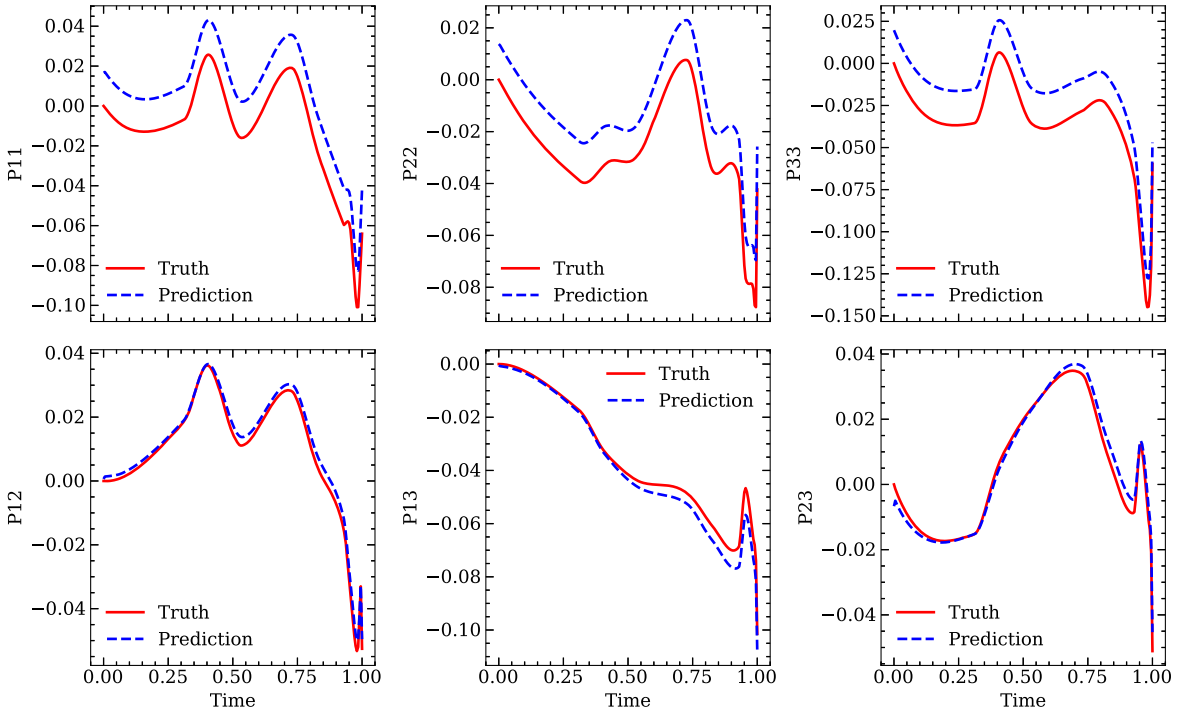

Fig. 8. Comparison of true and predicted stress corresponding to the sample with maximum error.

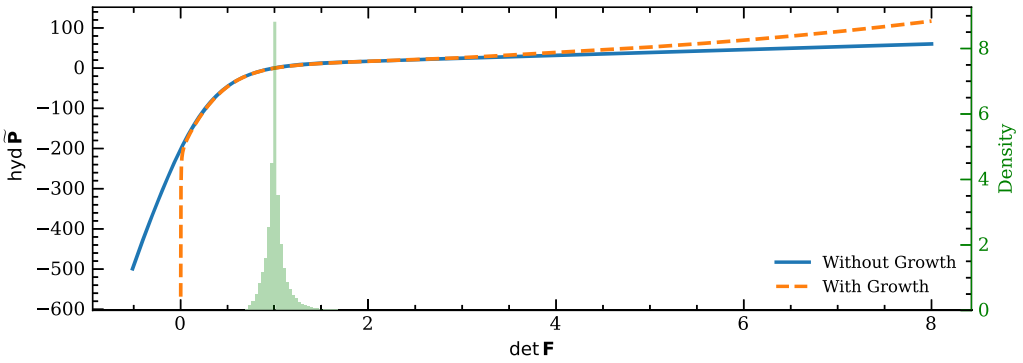

Fig. 9. The hydrostatic stress with and without the growth imposed on the NN model for a fixed value of internal variables.

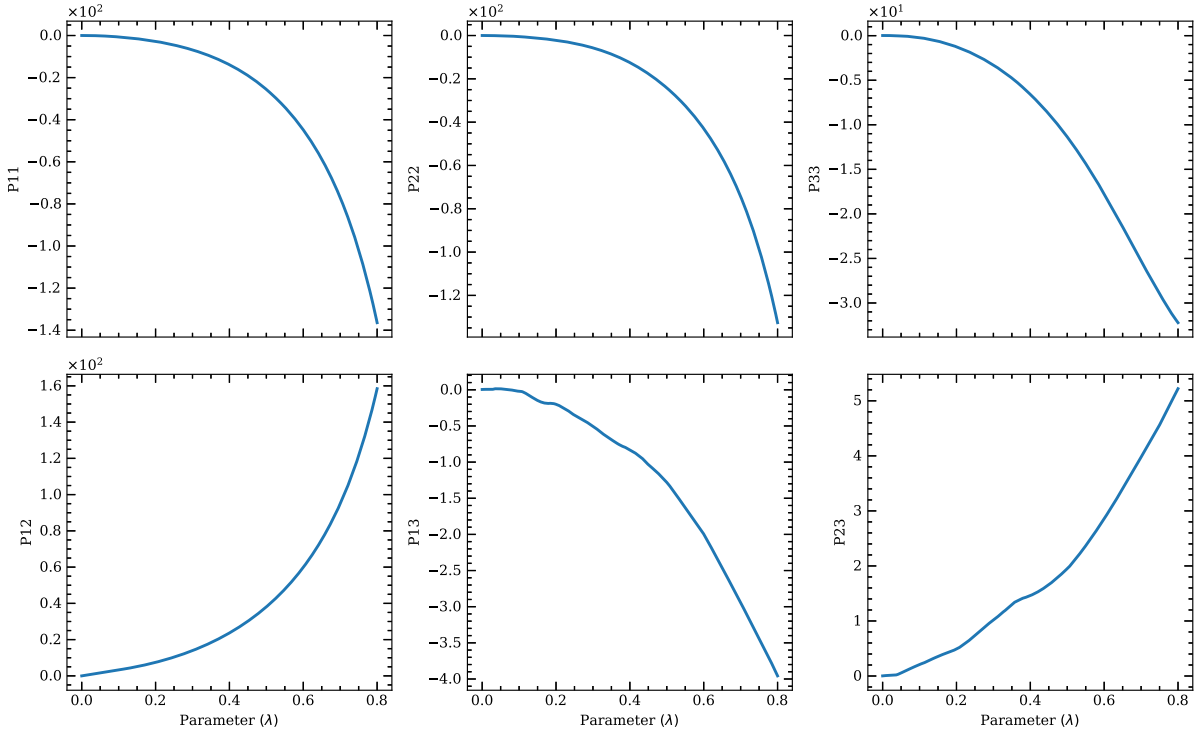

Fig. 10. Increase in stress as shear strain is applied in the 1–2 plane, along the pure shear direction and beyond the support of the dataset (see (26)).

and the shear components in the $1-2$ plane grow rapidly with growing amount of shear, λ. The shear components $P13$ and $P23$, that do not lie in the $1-2$ plane, are almost small compared to the components in the $1-2$ plane.

4.2.4. Objectivity of the polyconvex model

We recall that the polyconvex model is trained on data comprising only symmetric deformation gradients. Hence, it is not objective by design of its architecture. The current literature shows that there are limited constitutive models of general forms that are both polyconvex and objective. While both can be achieved for isotropic materials where the model is expressed in terms of the invariants of the deformation gradient, anisotropy requires further assumptions on material symmetry, such as in Schröder et al. (2010). An invariant-based neural-network model that satisfies both polyconvexity and objectivity is found not to have performed well in terms of predictive accuracy Klein et al. (2022). Hence, in Klein et al. (2022), objectivity has been incorporated as a soft constraint by adding an additional term to the loss function. The term appended to the neural network loss function for incorporating objectivity is as follows

$$\|\mathbf{P}(\mathbf{QF}) - \mathbf{QP}(\mathbf{F})\|^2, \quad \forall\, \mathbf{Q}, \mathbf{F}. \tag{27}$$

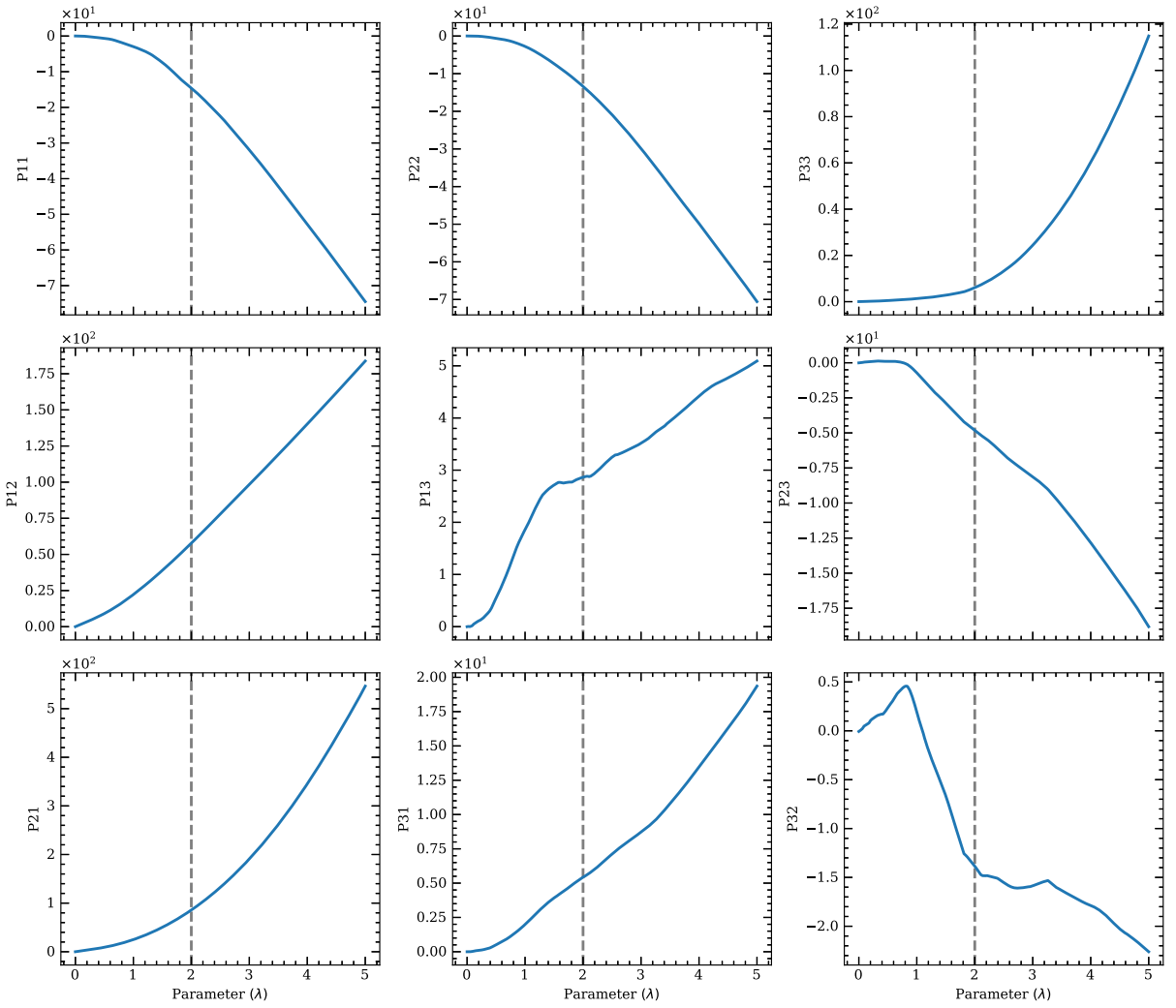

Fig. 11. Increase in stress as shear strain is applied in the 1–2 plane, along the simple shear direction and beyond the support of the dataset (see (26)). The vertical line marks the maximum stretch in the dataset.

where $\mathbf{Q}$ is an orthogonal matrix. In this work, we propose another method that is equivalent to appending (27) to the loss term.

We recall that, in this work, the polyconvex neural network model is trained on a dataset comprising only of symmetric deformation gradient $\mathbf{F}$. Hence, the first Piola-Kirchoff stress $\mathbf{P}_D$, where the subscript D is used to represent map that is learned from data, is accurate over space of symmetric tensor i.e., $\mathbf{P}_D : \mathbb{R}^{3\times3}_{sym,+} \to \mathbb{R}^{3\times3}_{sym}$; where $\mathbb{R}^{3\times3}_{sym}$ is the space of symmetric matrices, and $\mathbb{R}^{3\times3}_{sym,+}$ is the space of space symmetric matrices with positive determinant. However, we extend this learned model to define stress for a non-symmetric deformation gradient as follows

$$\mathbf{P}(\mathbf{F}) := \begin{cases} \mathbf{P}_D(\mathbf{F}), & \text{if } \mathbf{F} \in \mathbb{R}^{3\times3}_{sym,+}; \\ \widetilde{\mathbf{Q}}\mathbf{P}_D(\widetilde{\mathbf{F}}), & \text{if } \mathbf{F} \neq \mathbb{R}^{3\times3}_{sym,+},\ \mathbf{F} = \widetilde{\mathbf{Q}}\widetilde{\mathbf{F}}, \widetilde{\mathbf{Q}} \in \mathrm{SO}(3),\ \widetilde{\mathbf{F}} \in \mathbb{R}^{3\times3}_{sym,+}. \end{cases} \tag{28}$$

$\mathbf{F} = \widetilde{\mathbf{Q}}\widetilde{\mathbf{F}}$ is the polar decomposition of $\mathbf{F}$. Hence, with (28) extension of Piola-Kirchoff stress to non-symmetric deformation gradients, (27) evaluates to zero as follows,

$$\mathbf{P}(\mathbf{QF}) - \mathbf{QP}(\mathbf{F}) = \mathbf{P}(\mathbf{Q}\widetilde{\mathbf{Q}}\widetilde{\mathbf{F}}) - \mathbf{QP}(\widetilde{\mathbf{Q}}\widetilde{\mathbf{F}}) = \mathbf{Q}\widetilde{\mathbf{Q}}\mathbf{P}_D(\widetilde{\mathbf{F}}) - \mathbf{Q}\widetilde{\mathbf{Q}}\mathbf{P}_D(\widetilde{\mathbf{F}}) = \mathbf{0}. \tag{29}$$

4.3. Identifiability of internal variables

An important aspect of data-driven constitutive identification is that the internal variables are not postulated *a priori*, but learned from stress-strain data. Therefore, it happens that the same stress-strain data, but different instances of learning can lead to different

internal variables. In this section, we present the main theoretical results that internal variables are identifiable up to a linear transform. Throughout this subsection, we work with the constitutive law in (4), which expresses the internal variable evolution implicitly. Theorem 4.6 states that given a constitutive model, which is defined by a fixed pair $(\mathcal{W}, \widehat{D})$, as in (4b), and the strain-stress response generated by that model; a linear transformation of the internal variables does not change the strain-stress response. Theorem (4.11) states the converse result, i.e., given two pairs of functions $(\mathcal{W}, \widehat{D})$ and $(\widetilde{\mathcal{W}}, \widetilde{D})$ that lead to the same stress-strain response, and that differ only in their sets of internal variables, then the sets of internal variables must be related by a linear transform. The empirical verification of this identifiability of internal variables up to a linear transform is done later in this section.

Definition 4.1. Let m, n be positive integers, $\mathbb{R}^m$ be the m-dimensional space of real numbers, and $\mathrm{GL}(n, \mathbb{R})$ be the general linear group of order n. For $T > 0$, let $C^k = C^k([0,T]; \mathbb{R}^m)$ denote the class of k-times continuously differentiable functions, and $L^\infty([0,T]; \mathbb{R}^m)$ the essentially bounded functions, from $[0,T]$ into $\mathbb{R}^m$.

Assumption 4.2. Let $\mathcal{W} \in C^1(\mathbb{R}^m \times \mathbb{R}^n, \mathbb{R}^+)$. Let $\widehat{D} \in C^1(\mathbb{R}^n, \mathbb{R}^+)$ be a strictly convex function with minimum value of zero attained at the point zero.

Remark 4.3. In this section, $\mathcal{W}$ is defined to take the vectorized deformation gradient as input. Hence, $m = d(d+1)/2$, which is 6 for $d = 3$. We highlight that the $\mathcal{W}$ is viewed as a function from $\mathbb{R}^m \times \mathbb{R}^n$, where $\mathbb{R}^n$ is the space in which the internal variable lives. We ignore the fact that the deformation gradient must have a positive determinant. Moreover, we note that the rate-independent theories of plasticity are not strictly convex.

Definition 4.4. Let $\mathcal{S} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ and $\mathcal{F} : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ be such that for $f \in \mathbb{R}^m$ and $p, q \in \mathbb{R}^n$

$$\mathcal{S}(f,p) := \frac{\partial \mathcal{W}(f,p)}{\partial f}, \tag{30a}$$

$$\mathcal{F}(f,p,q) := \frac{\partial \widehat{D}(q)}{\partial q} + \frac{\partial \mathcal{W}(f,p)}{\partial p}. \tag{30b}$$

Definition 4.5. Let $F \in L^\infty([0,T]; \mathbb{R}^m)$ and $\xi \in C^1([0,T]; \mathbb{R}^n)$. For time $t \in [0,T]$, using (30a), and (30b), we define a first-order ordinary differential equation and an observable S as follows

$$S(t) = \mathcal{S}(F(t), \xi(t)), \tag{31a}$$

$$\mathcal{F}(F(t), \xi(t), \dot{\xi}(t)) = 0, \quad \xi(0) = 0. \tag{31b}$$

Theorem 4.6. *Given a pair of functions $(\mathcal{W}, \widehat{D})$ satisfying Assumption 4.2. For any $F \in L^\infty$, consider $S(t)$ and $\xi(t)$ given by (31). Then, for every $A \in \mathrm{GL}(n, \mathbb{R})$, there exist functions*

$$\widetilde{\mathcal{W}} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^+, \quad (f, \widetilde{p}) \mapsto \widetilde{\mathcal{W}}(f, \widetilde{p}); \tag{32a}$$

$$\widetilde{D} : \mathbb{R}^n \to \mathbb{R}^+, \qquad \widetilde{q} \mapsto \widetilde{D}(\widetilde{q}) \tag{32b}$$

with $\widetilde{\mathcal{S}} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ and $\widetilde{\mathcal{F}} : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$, defined by,

$$\widetilde{\mathcal{S}}(f, \widetilde{p}) := \frac{\partial \widetilde{\mathcal{W}}(f, \widetilde{p})}{\partial f} \tag{33a}$$

$$\widetilde{\mathcal{F}}(f, \widetilde{p}, \widetilde{q}) := \frac{\partial \widetilde{D}(\widetilde{q})}{\partial \widetilde{q}} + \frac{\partial \widetilde{\mathcal{W}}(f, \widetilde{p})}{\partial \widetilde{p}} \tag{33b}$$

such that, for $\eta(t) := A\xi(t)$ for $t \in [0,T]$,

$$S(t) = \widetilde{\mathcal{S}}(F(t), \eta(t)); \tag{34a}$$

$$\widetilde{\mathcal{F}}(F(t), \eta(t), \dot{\eta}(t)) = 0, \quad \eta(0) = 0. \tag{34b}$$

Moreover, the convexity of $\widehat{D}$ is preserved in $\widetilde{D}$.

Proof. Proof given in the Appendix. □

Remark 4.7. Theorem 4.6 holds for any positive integers m and n, i.e., any dimensionality of $F(t)$ and $\xi(t)$.

Now, we state Theorem 4.11, which is the converse of Theorem 4.6.

Definition 4.8. For a given pair of functions $(\mathcal{W}, \widehat{D})$, $t \in [0,T]$, $C \in L^\infty$, and the corresponding differential Eq. (31b), we define

$$U := \{F(t) : F \in L^\infty, t \in [0,T]\} \subset \mathbb{R}^m, \tag{35a}$$

$$\Upsilon := \{\xi : \mathcal{F}(F(t), \xi(t), \dot{\xi}(t)) = 0, \ \forall F \in L^\infty\} \subset C^1([0,T]; \mathbb{R}^n), \tag{35b}$$

$$P := \{\xi(t) : \xi \in \Upsilon, t \in [0,T]\} \subset \mathbb{R}^n, \tag{35c}$$

$$Q := \{\dot{\xi}(t) : \xi \in \Upsilon, t \in [0,T]\} \subset \mathbb{R}^n. \tag{35d}$$

We also observe that sets P and Q are directly related to set Υ. We desire P and Q to be open subsets of $\mathbb{R}^n$. To this end, we state Assumption 4.9 needed for Theorem 4.11. We later provide examples with particular forms of functions of $(\mathcal{W}, \widehat{D})$ that are consistent with Assumption 4.9.

Assumption 4.9. For a given pair of functions $(\mathcal{W}, \widehat{D})$, consider the ODE defined in (31b). Let there be open sets $P, Q \subset \mathbb{R}^n$ such that for any $(p, q) \in P \times Q$, there exists $t_* \in [0, T]$, $F_* \in L^\infty([0,T]; \mathbb{R}^m)$ and $\xi_* \in C^1([0,1]; \mathbb{R}^n)$ such that

$$\mathcal{F}(F_*(t), \xi_*(t), \dot{\xi}_*(t)) = 0, \quad \xi_*(0) = 0, \quad \forall t \in [0,T]; \tag{36a}$$

$$\xi_*(t_*) = p; \quad \dot{\xi}_*(t_*) = q. \tag{36b}$$

Definition 4.10. $\mathcal{T} \in C^2(\mathbb{R}^n; \mathbb{R}^n)$ is a twice-differentiable diffeomorphism i.e., $\mathcal{T}$ is bijective, and $\mathcal{T}^{-1}$ is twice differentiable.

Theorem 4.11. *For given two pairs of functions $(\mathcal{W}, \widehat{D})$ and $(\widetilde{\mathcal{W}}, \widetilde{D})$; consider $(\mathcal{S}, \mathcal{F})$, and $(\widetilde{\mathcal{S}}, \widetilde{\mathcal{F}})$ defined as per (30), and (33) respectively. Let for all $C \in L^\infty$ and $t \in [0,T]$,*

$$\mathcal{S}(F(t), \xi(t)) = \widetilde{\mathcal{S}}(F(t), \eta(t)), \tag{37a}$$

$$\mathcal{F}(F(t), \xi(t), \dot{\xi}(t)) = 0, \quad \xi(0) = 0, \tag{37b}$$

$$\widetilde{\mathcal{F}}(F(t), \eta(t), \dot{\eta}(t)) = 0, \quad \eta(0) = 0. \tag{37c}$$

Moreover, let $\mathcal{F}$ satisfy Assumption 4.9 and $P \subset \mathbb{R}^n$ be the set as defined in Assumption 4.9. Let $\mathcal{T} : P \to \mathbb{R}^n$ be a smooth diffeomorphism such that $\eta(t) = \mathcal{T}(\xi(t))$. Then, $\mathcal{T}$ must be linear i.e. $\mathcal{T}(\xi(t)) = A\xi(t)$ for all $\xi(t) \in P$, where $A \in \mathrm{GL}(n, \mathbb{R})$.

Proof. The proof of Theorem 4.11 is given in the Appendix. □

We now present a few examples to highlight the importance of the Assumption 4.9 in Theorem 4.11. We present Example 4.12 that is consistent with Assumption 4.9. Example 4.12 is a case where the dynamics of internal variables are governed by a non-linear ODE. It is to be noted that the source of non-linearity is the choice of non-quadratic $\mathcal{D}$; while $\mathcal{W}$ is quadratic.

Example 4.12. Let $m = n$ and define $\mathcal{W}$ and $\widehat{D}$ as follows

$$\mathcal{W}(f, p) = \frac{1}{2}\begin{bmatrix} f \\ p \end{bmatrix}^\top \begin{bmatrix} A_1 & A_2 \\ A_2^\top & A_3 \end{bmatrix} \begin{bmatrix} f \\ p \end{bmatrix}, \tag{38a}$$

$$\widehat{D}(q) = \frac{1}{2(r+1)} |q|^{2(r+1)}; \quad \text{where } r > -1/2, \tag{38b}$$

where $|\cdot|$ is the 2-norm. Hence,

$$\mathcal{F}(F(t), \xi(t), \dot{\xi}(t)) = \left|\dot{\xi}(t)\right|^{2r} \dot{\xi}(t) + A_3 \xi(t) + A_2 F(t). \tag{39}$$

Hence, for any given t_* and $p, q \in \mathbb{R}^n$, we can construct ξ_* and C_* as

$$F_*(t) = -A_2^{-1}\left(\left|\dot{\xi}_*(t)\right|^{2r} \dot{\xi}_*(t) + A_3 \xi_*(t)\right) \tag{40}$$

such that $\mathcal{F}(F_*(t), \xi_*(t), \dot{\xi}_*(t)) = 0$, $\xi_*(t_*) = p$ and $\dot{\xi}_*(t_*) = q$. Hence, $P = Q = \mathbb{R}^n$, and this example is consistent with Assumption 4.9.

Finally, we present a case to highlight the importance of dimensionality. In the following example, the dimensionality of the deformation gradient is less than that of the strain. This makes the example inconsistent with Assumption 4.9. Hence, the technique employed to prove Theorem 4.11 does not apply to Example 4.13.

Example 4.13. Let $m < n$. Consider $\mathcal{W}$ and $\widehat{D}$ as follows.

$$\mathcal{W}(f, p) = \frac{1}{2}\begin{bmatrix} f \\ p \end{bmatrix}^\top \begin{bmatrix} A_1 & A_2 \\ A_2^\top & A_3 \end{bmatrix} \begin{bmatrix} f \\ p \end{bmatrix}, \tag{41a}$$

$$\widehat{D}(q) = \frac{1}{2} q^\top D q, \tag{41b}$$

where $D \in \mathbb{R}^{n \times n}$ is symmetric, $A_3 \in \mathbb{R}^{n \times n}$ is symmetric and $A_2 \in \mathbb{R}^{n \times m}$. Hence, we have

$$D\dot{\xi}(t) + A_3 \xi(t) + A_2 F(t) = 0, \tag{42}$$

For a fixed $t = t_*$, fix $\xi(t_*) = p_*$. Hence,

$$\dot{\xi}(t_*) = -D^{-1}(A_3 p_* + A_2 F(t_*)). \tag{43}$$

Define the set

$$Q := \{\dot{\xi}(t_*) : F(t_*) = u, \forall u \in \mathbb{R}^m\}. \tag{44}$$

Since A_2 has rank of m, the set $Q \subset \mathbb{R}^n$ is m-dimensional. Hence, $\dot{\xi}(t_*)$ cannot take any arbitrary value in $\mathbb{R}^n$.

Remark 4.14. The results about the identifiability of internal variables, established in this section, are agnostic of any structure, such as polyconvexity, which is discussed in Section 3.3, on $\mathcal{W}$, except for the ones that might arise to satisfy Assumption 4.9. Moreover, even though Theorem 4.11 stands on Assumption 4.2, that is violated when $n > m$ as in Example 4.13, Fig. 13 shows the claim of Theorem 4.11 empirically holds even when $n > m$. Hence, advanced techniques that do not rely on the strong Assumption 4.9 are needed to prove this result for a broader class of constitutive laws, including when $n > m$.

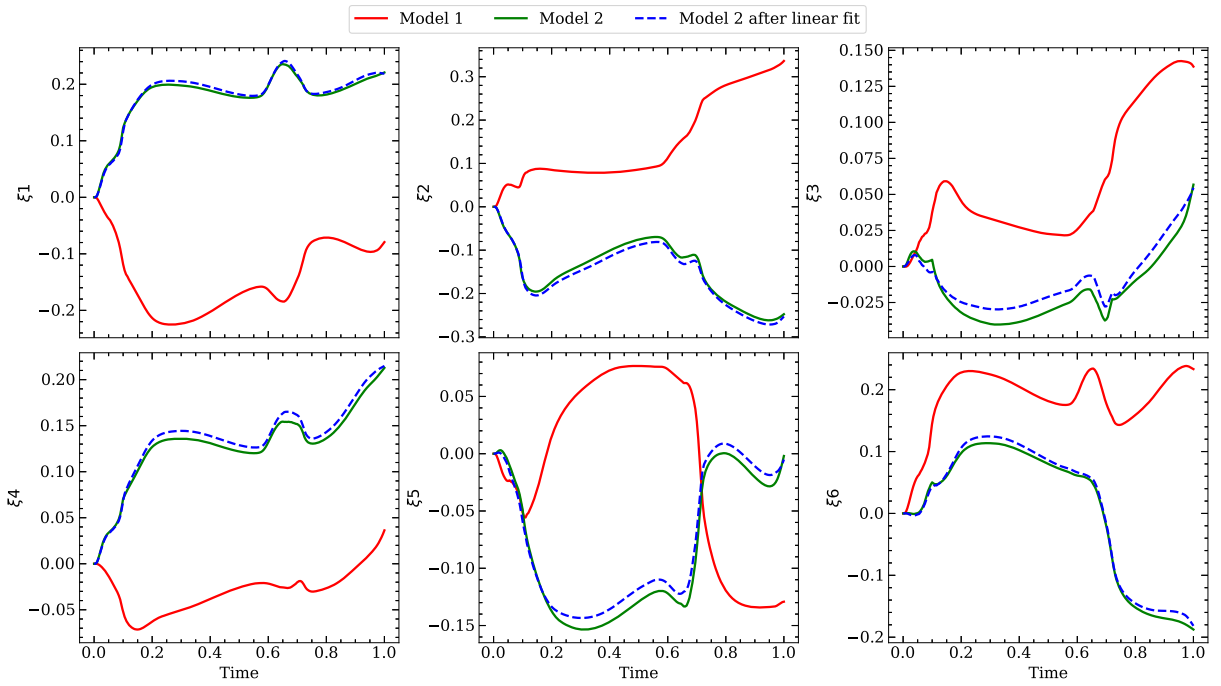

Fig. 12. Visualisation of the linear fit between two sets of internal variables.

As internal variables are physically unobservable, there is nothing to compare the predicted internal variables against. However, as discussed in the theoretical results, the inferred trajectory of internal variables, for a given trajectory of $\mathbf{F}$, should be unique up to scaling. Therefore, we train two different NN models on the same dataset, with the same sets of mutually disjoint training and testing datasets, and compare the internal variables predicted by the two models. Fig. 12 shows the visual comparison between the sets of internal variables predicted using two independently trained NN models, dissimilar in their initialisation of parameters. It is observed from Fig. 13 that, although the two sets of internal variables (labelled as Model 1 and Model 2) do not match, they indeed are linear transformations of one another. Moreover, Fig. 12 shows the quality of this linear fit using the relative error between two sets of internal variables. Note that the linear regression model is fitted to the internal variables corresponding to the training set, and then the same linear regression model is used to check if this scaling holds for the test set. The median error in the linear fit is approximately 5% on both the train and test sets.

The theoretical result provided in this study holds when the number of strain components is more than the number of internal variables, and empirical evidence and examples are provided for the case when the two are the same. It would be interesting to study the identifiability of internal variables in models where the dimensionality of internal variables differs from that of strain and, particularly, the cases of isotropic, kinematic and mixed hardening rules. As our NN-based approach, in its current form, does not separate elastic and plastic energy, it cannot make a distinction between isotropic and kinematic hardening either. Still, an interesting question is whether the matrix A relating the change of internal variables is block diagonal. If so, we can partition the state variables into groups, for example, $X_1 = \{\xi_1, \dots, \xi_m\}$ and $X_2 = \{\xi_{m+1}, \dots, \xi_n\}$. The dimensionality of these groups may provide insight into their physical meaning. This would be consistent with the phenomenological theory of plasticity where isotropic hardening has dimension one and kinematic hardening dimension five. Moreover, it would also be interesting to inspect the case with an isochoric constraint on the plastic strain tensor. The constitutive law formulation used in this work is associative because dissipation depends only on the rate of change of internal variables. Hence, it would be interesting to verify whether the same identifiability result holds for non-associative constitutive laws where the dissipation depends on both the internal variables and their rate. Furthermore, it would be interesting to quantify the rate of decay in the linear fit of two internal variable sets with respect to the decay in the stress prediction error by two independent models.

4.4. Computational cost

We discuss the computational cost of training our NN model. All computations reported are performed on a single NVIDIA Tesla P100-PCIE-16GB GPU. The number of parameters in the neural networks $\mathcal{N}_H$, $\mathcal{N}_V$, and $\mathcal{N}_D$ is 4500, 315000, and 120000, respectively, for six internal variables. The number of parameters in $\mathcal{N}_V$ and $\mathcal{N}_D$ increases by 15000 for every increase in the number of internal variables. The training of $\mathcal{N}_V$ and $\mathcal{N}_D$ is coupled and takes 24 hours. Training $\mathcal{N}_H$ is much faster and takes just an hour. This is so because $\mathcal{N}_H$ does not depend on internal variables and hence corresponds only to the hyperelastic part of the constitutive law.

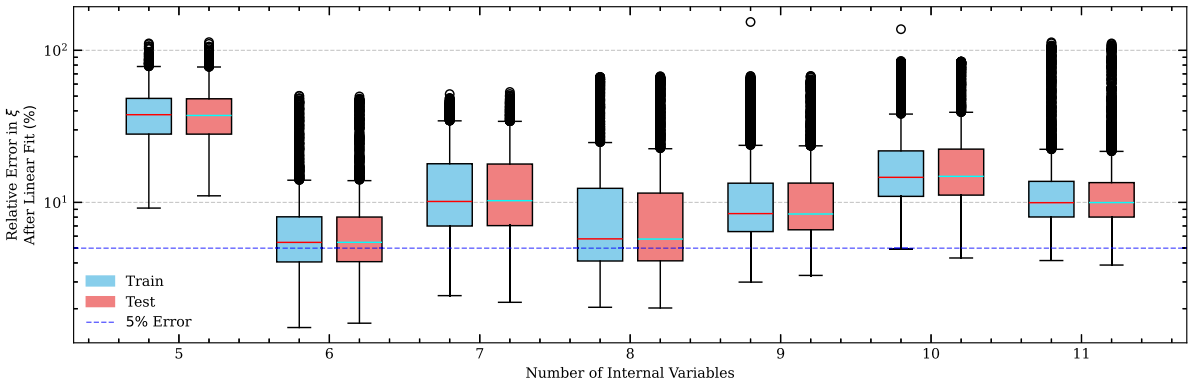

Fig. 13. Error in the linear fit between two sets of internal variables obtained by independently training the same architecture with different initialisations.

We assume that the off-line training cost is negligible; this assumption requires, of course, that the trained model is used repeatedly, thereby amortizing the cost of training. Once the neural networks are trained, the prediction of stress trajectories given strain trajectories takes less than a second per sample that evaluates to 5×10^{-3} seconds per time-step. Predictions for multiple strain samples can be parallelized and take 8 seconds for 4000 samples, with 200 time-steps per sample. The cost of generating the dataset, which is given in Liu et al. (2022), is roughly 0.24 second per time-step. Moreover, the cost comparison of solving a boundary value problem using the two methods: the two-scale FE^2 computation and just the macro problem with a NN model as constitutive law is done in Liu et al. (2022, Table 2); where it is shown that the latter is significantly faster than the former. While the NN model used in that comparison is non-energetic, we do not expect the cost of using an energetic model to be very different as both lead to a non-linear boundary value problem.

5. Conclusion

In this work, we proposed an energy-based NN framework to model a history-dependent constitutive law. The framework, which relies on the internal variable theory, was designed to be causal, interpretable, and incorporate knowledge from physics, namely the second law of thermodynamics and material stability, and consistent with the existence theory of the non-linear equilibrium problem. We outlined methods for the NN-based constitutive law to ensure consistency with material stability and the second law of thermodynamics. We demonstrated the efficacy of this framework for homogenizing a polycrystalline material with a viscoplastic response. An accuracy of 2% in terms of relative error was achieved in the prediction of stress trajectories. Moreover, we theoretically showed that the internal variables obtained as by-products of stress prediction are unique up to a scaling factor. This claim was empirically verified by fitting a linear regression model. While the method is deployed to model viscoplastic response, it can also capture any history-dependent response. These machine learning models can be directly used as surrogates for the homogenized constitutive law of a unit cell in a multiscale FEM-based equilibrium equation solver such as Abaqus. There are many many directions for further research and investigations on learning energy-based issues. Technical challenges pertaining to learning potentials include (a) the choice of right activation for fast and stable training process and (b) the effect of data normalization. These have been discussed in detail in Appendices B.1 and B.2. The identifiability of internal variables up to scaling can be investigated in different plasticity models and the convergence of error in the linear fit between the two sets of internal variables, with respect to error in stress can be studied numerically. This has been discussed at the end of Section 4.3. Moreover, the framework of learning potentials can be extended to (a) learning heterogeneous i.e., microstructure-dependent potentials; and (b) learning a spatially non-local constitutive laws.

CRediT authorship contribution statement

Mayank Raj: Writing – original draft, Writing – review & editing, Validation, Visualization, Formal analysis; **Lianghao Cao:** Writing – review & editing, Software, Methodology, Investigation, Formal analysis; **Andrew Stuart:** Writing – review & editing, Methodology, Formal analysis, Funding acquisition; **Kaushik Bhattacharya:** Writing – review & editing, Supervision, Formal analysis, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We gratefully acknowledge the financial support of the US Office of Naval Research through MURI grant N00014-23-1-2654. The computations presented here were conducted in the Resnick High Performance Computing Center, a facility supported by the Resnick Sustainability Institute at the California Institute of Technology.

Code and data availability

The code is available at https://github.com/m-raj/Learning-Stable-Material-Model.git. Data will be made available upon request.

Appendix A. Proofs

Theorem 4.6. *Given a pair of functions $(\mathcal{W}, \widehat{D})$ satisfying Assumption 4.2. For any $F \in L^\infty$, consider $S(t)$ and $\xi(t)$ given by (31). Then, for every $A \in \mathrm{GL}(n, \mathbb{R})$, there exist functions*

$$\widetilde{\mathcal{W}} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^+, \quad (f, \widetilde{p}) \mapsto \widetilde{\mathcal{W}}(f, \widetilde{p}); \tag{A.45a}$$

$$\widetilde{D} : \mathbb{R}^n \to \mathbb{R}^+, \qquad \widetilde{q} \mapsto \widetilde{D}(\widetilde{q}) \tag{A.45b}$$

with $\widetilde{S} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ and $\widetilde{\mathcal{F}} : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$, defined by,

$$\widetilde{S}(f, \widetilde{p}) := \frac{\partial \widetilde{\mathcal{W}}(f, \widetilde{p})}{\partial f} \tag{A.46a}$$

$$\widetilde{\mathcal{F}}(f, \widetilde{p}, \widetilde{q}) := \frac{\partial \widetilde{D}(\widetilde{q})}{\partial \widetilde{q}} + \frac{\partial \widetilde{\mathcal{W}}(f, \widetilde{p})}{\partial \widetilde{p}} \tag{A.46b}$$

such that, for $\eta(t) := A\xi(t)$ for $t \in [0, T]$,

$$S(t) = \widetilde{S}(F(t), \eta(t)); \tag{A.47a}$$

$$\widetilde{\mathcal{F}}(F(t), \eta(t), \dot{\eta}(t)) = 0, \quad \eta(0) = 0. \tag{A.47b}$$

Moreover, $\widetilde{D}$ is convex.

Proof. We define

$$\widetilde{\mathcal{W}}(f, \widetilde{p}) := \mathcal{W}(f, p) \tag{A.48a}$$

$$\widetilde{D}(\widetilde{q}) := \widehat{D}(q) \tag{A.48b}$$

such that

$$\widetilde{p} = Ap, \quad \widetilde{q} = Aq. \tag{A.49}$$

Hence,

$$\widetilde{S}(f, \widetilde{p}) := \frac{\partial \widetilde{\mathcal{W}}(f, \widetilde{p})}{\partial f} = \frac{\partial \mathcal{W}(f, p)}{\partial f} = S(f, p). \tag{A.50}$$

and

$$A^{-\top} \frac{\partial \mathcal{W}(f, p)}{\partial p} + A^{-\top} \frac{\partial \widehat{D}(q)}{\partial q} = A^{-\top} \mathcal{F}(f, p, q) \tag{A.51}$$

Substituting $f = F(t)$, $p = \xi(t)$ and $q = \dot{\xi}(t)$, we get

$$S(t) = S(F(t), \xi(t)) = \widetilde{S}(F(t), A\xi(t)) \tag{A.52a}$$

$$\widetilde{\mathcal{F}}(F(t), A\xi(t), A\dot{\xi}(t)) = A^{-\top} \mathcal{F}(F(t), \xi(t), \dot{\xi}(t)). \tag{A.52b}$$

We define $\eta(t) = A\xi(t)$, and use (31b) to conclude,

$$S(t) = \frac{\widetilde{\mathcal{W}}(F(t), \eta(t))}{\partial p} \tag{A.53a}$$

$$\widetilde{\mathcal{F}}(F(t), \eta(t), \dot{\eta}(t)) = \frac{\widetilde{\mathcal{W}}(F(t), \eta(t))}{\partial \widetilde{p}} + \frac{\partial \widetilde{D}(\dot{\eta}(t))}{\partial \widetilde{q}} = 0, \quad \eta(0) = 0. \tag{A.53b}$$

□

Lemma A.1. *Let $\mathcal{Q}(p)$ represent an orthogonal matrix, parameterized by p; and let $\mathcal{T}(p)_{i,j}$ represent $\partial\mathcal{T}(p)_i/\partial p_j$. If $\mathcal{T}(p)_{i,j} = \mathcal{Q}(p)_{ij}$, then $\mathcal{T}(p)_{ij} = [\mathcal{Q}_0]_{ij}$ where $\mathcal{Q}_0$ is an orthogonal matrix independent of p.*

Proof. We have given that $\mathcal{T}(p)_{i,j} = \mathcal{Q}(p)_{i,j}$. Differentiating both sides with p_k, we have

$$\mathcal{T}(p)_{i,jk} = \mathcal{Q}(p)_{ij,k}. \tag{A.54}$$

By orthogonality of $\mathcal{Q}(p)$, we have

$$\mathcal{Q}(p)_{ij}\mathcal{Q}(p)_{jm} = \delta_{im}; \tag{A.55}$$

$$\implies \mathcal{Q}(p)_{ij,k} + \mathcal{Q}(p)_{jm,k} = 0 \quad \forall i, m; \quad \text{(Differentiating both sides with } p_k\text{)}, \tag{A.56}$$

$$\implies \mathcal{Q}(p)_{ij,k} = -\mathcal{Q}(p)_{ji,k}; \quad (m = i). \tag{A.57}$$

$$\implies \mathcal{T}(p)_{i,jk} = -\mathcal{T}(p)_{j,ik} \quad \text{(using (A.54))} \tag{A.58}$$

As a mixed partial derivative of $\mathcal{T}$ is continuous, we have

$$\mathcal{T}(p)_{i,jk} = \mathcal{T}(p)_{i,kj}. \tag{A.59}$$

Now, we used (A.58) and (A.59),

$$\mathcal{T}(p)_{i,jk} = \mathcal{T}(p)_{i,kj} \tag{A.60a}$$

$$= -\mathcal{T}(p)_{k,ij} \tag{A.60b}$$

$$= -\mathcal{T}(p)_{k,ji} \tag{A.60c}$$

$$= \mathcal{T}(p)_{j,ki} \tag{A.60d}$$

$$= \mathcal{T}(p)_{j,ik} \tag{A.60e}$$

$$= -\mathcal{T}(p)_{i,jk}. \tag{A.60f}$$

Hence, $\mathcal{T}(p)_{i,jk} = -\mathcal{T}(p)_{i,jk}$. Hence, $\mathcal{T}(p)_{i,jk} \equiv 0$. Hence, $\mathcal{T}(p)_{i,j} = [Q_0]_{ij}$ where Q_0 is independent of p. □

Proposition A.2. *Let P, Q be open sets in $\mathbb{R}^n$. Let D and $\widetilde{D}$ be two strongly convex functions with minima at zero, and their minimum values be zero. Let $\mathcal{T} : P \to \mathbb{R}^n$ be a diffeomorphism. If*

$$D(q) = \widetilde{D}(A(p)q), \quad \forall p, q \in P \times Q, \tag{A.61}$$

where $A(p) := [\partial\mathcal{T}(p)_i/\partial p_j]$; then $\mathcal{T}(p) = A_0 p$ is linear for $A_0 \in \mathrm{GL}(n, \mathbb{R})$.

Proof. As (A.61) is true for all $p, q \in P \times Q$, we pick p_1 and p_2 such that $A(p_1) \neq A(p_2)$. Hence, we have

$$D(q) = \widetilde{D}(A(p_1)q), \quad \forall q \in Q; \tag{A.62a}$$

$$D(q) = \widetilde{D}(A(p_2)q), \quad \forall q \in Q. \tag{A.62b}$$

Using (A.62a) and (A.62b), we have

$$\widetilde{D}(A(p_1)q) = \widetilde{D}(A(p_2)q), \quad \forall q \in Q. \tag{A.63}$$

As $A(p_1)$ is invertible, substitute $q = A(p_1)^{-1}\widetilde{q}$ to obtain

$$\widetilde{D}(\widetilde{q}) = \widetilde{D}(A(p_2)A(p_1)^{-1}\widetilde{q}). \tag{A.64}$$

Now, for $\alpha > 0$, we define the level set $\widetilde{Y}_\alpha$, as

$$\widetilde{Y}_\alpha := \{\widetilde{q} : \widetilde{D}(\widetilde{q}) = \alpha\}. \tag{A.65}$$

Hence, (A.64) implies that

$$\widetilde{q} \in \widetilde{Y}_\alpha \iff A(p_2)A(p_1)^{-1}\widetilde{q} \in \widetilde{Y}_\alpha \tag{A.66}$$

Hence, we have

$$A(p_2)A(p_1)^{-1} : \widetilde{Y}_\alpha \to \widetilde{Y}_\alpha, \tag{A.67}$$

is a bijection. Observing that $\widetilde{Y}_\alpha$ is a bounded set, it follows that

$$A(p_2)A(p_1)^{-1} = \mathcal{Q}_\alpha(p) \quad \forall p_1, p_2 \in P, \quad \alpha \in \mathbb{R}^+. \tag{A.68}$$

Hence,

$$A(p) = \mathcal{Q}_\alpha(p)A_0; \quad A_0 \in \mathrm{GL}(n, \mathbb{R}). \tag{A.69}$$

We apply Lemma A.1 on the diffeomorphism $\mathcal{T} \circ \mathcal{T}_0$, where $\mathcal{T}_0(p) := A_0 p$, to deduce that $\mathcal{Q}_\alpha(p)$ must be independent of p. Also, we denote $\widetilde{A}(\alpha) := \mathcal{Q}_\alpha A_0$, to get

$$A(p) = \widetilde{A}(\alpha), \quad \forall p \in P, \forall \alpha \in \mathbb{R}^+. \tag{A.70}$$

Hence, it must be that

$$A(p) \equiv \widetilde{A}(\alpha) \equiv A_0. \tag{A.71}$$

where $A_0 \in \mathrm{GL}(n, \mathbb{R})$. Hence $\mathcal{T}$ is affine such that $\mathcal{T}(p) = A_0 p + p_0$. But as $\mathcal{D}$ and $\widetilde{\mathcal{D}}$ have minima at zero, $p_0 = 0$. Hence, $\mathcal{T}(p) = A_0 p$, is linear. □

From the Example 4.12, it is observed that Assumption 4.9 is justified for quadratic $\mathcal{W}$ and convex $\mathcal{D}$, when $m \geq n$. On the other hand, from Example 4.13, the assumption is not justified when $n > m$, and $\mathcal{W}$ is a quadratic function. In the Example 4.12, the non-linearity comes from $\mathcal{D}$.

Lemma A.3. $U = \mathbb{R}^m$.

Proof. For any $c_* \in \mathbb{R}^m$ and fixed $t = t_*$, we can construct $C_* \in C^1([0, T]; \mathbb{R}^m)$, given by $C_*(t) = f_* t / t_*$, such that $C_*(t_*) = f_*$, $C_*(0) = 0$. □

Theorem 4.11. *For given two pairs of functions $(\mathcal{W}, \widehat{\mathcal{D}})$ and $(\widetilde{\mathcal{W}}, \widetilde{\mathcal{D}})$; consider $(\mathcal{S}, \mathcal{F})$, and $(\widetilde{\mathcal{S}}, \widetilde{\mathcal{F}})$ defined as per (30), and (33) respectively. Let for all $C \in L^\infty$ and $t \in [0, T]$,*

$$\mathcal{S}(F(t), \xi(t)) = \widetilde{\mathcal{S}}(F(t), \eta(t)), \tag{A.72a}$$

$$\mathcal{F}(F(t), \xi(t), \dot{\xi}(t)) = 0, \quad \xi(0) = 0, \tag{A.72b}$$

$$\widetilde{\mathcal{F}}(F(t), \eta(t), \dot{\eta}(t)) = 0, \quad \eta(0) = 0. \tag{A.72c}$$

Moreover, let $\mathcal{F}$ satisfy Assumption 4.9 and $P \subset \mathbb{R}^n$ be the set as defined in Assumption 4.9. Let $\mathcal{T} : P \to \mathbb{R}^n$ be a smooth diffeomorphism such that $\eta(t) = \mathcal{T}(\xi(t))$. Then, $\mathcal{T}$ must be linear i.e. $\mathcal{T}(\xi(t)) = A\xi(t)$ for all $\xi(t) \in P$, where $A \in \mathrm{GL}(n, \mathbb{R})$.

Proof. Using Assumption 4.9, we restrict the functions $\mathcal{W}$ and $\mathcal{D}$ as follows

$$\mathcal{W}|_{U \times P} : U \times P \to \mathbb{R}^+ \tag{A.73a}$$

$$\widehat{\mathcal{D}}|_Q : Q \to \mathbb{R}^+. \tag{A.73b}$$

We work with the restricted functions for the rest of the proof. For notational convenience, in this proof, we continue to use $\mathcal{W}$ and $\mathcal{D}$ instead of $\mathcal{W}|_{U \times P}$ and $\mathcal{D}|_Q$, respectively.

$$\mathcal{S}(F(t), \xi(t)) = \widetilde{\mathcal{S}}(F(t), \eta(t)) \tag{A.74}$$

$$\implies \frac{\partial \mathcal{W}(F(t), \xi(t))}{\partial p} = \frac{\partial \widetilde{\mathcal{W}}(F(t), \eta(t))}{\partial p} \tag{A.75}$$

Integrating on the set U, (A.76)

$$\implies \mathcal{W}(F(t), \xi(t)) = \widetilde{\mathcal{W}}(F(t), \mathcal{T}(\xi(t))) + h(\xi(t)) \tag{A.77}$$

where $h : \mathbb{R}^m \to \mathbb{R}$ is constant of integration w.r.t, f. Further, differentiating both sides,

$$\frac{\partial \mathcal{W}(F(t), \xi(t))}{\partial p} = \nabla \mathcal{T}(\xi(t))^\top \frac{\partial \widetilde{\mathcal{W}}(F(t), \mathcal{T}(\xi(t)))}{\partial \widetilde{p}} + \nabla h(\xi(t)) \tag{A.78}$$

Using (A.72b), (A.72c) and Assumption 4.9

$$\implies \frac{\partial \widehat{\mathcal{D}}(q)}{\partial q} = \nabla k(p)^\top \frac{\partial \widetilde{\mathcal{D}}(\nabla k(p) q)}{\partial \widetilde{q}} + \nabla h(p), \quad \forall p, q \in P \times Q \subset \mathbb{R}^n \times \mathbb{R}^n. \tag{A.79}$$

Upon substituting $q = 0$ and using the fact that $\widehat{\mathcal{D}}$ and $\widetilde{\mathcal{D}}$ have minima at zero, we conclude $\nabla h(p) = 0, \forall p$. Hence, $h(p) = h_0$ is a constant function. Hence, we have

$$\frac{\partial \widehat{\mathcal{D}}(q)}{\partial q} = \nabla k(p)^\top \frac{\partial \widetilde{\mathcal{D}}(\nabla k(p) q)}{\partial \widetilde{q}}, \tag{A.80}$$

Integrating on the set Q

$$\implies \widehat{\mathcal{D}}(q) = \widetilde{\mathcal{D}}(\nabla k(p) q) + \widetilde{h}(p). \tag{A.81}$$

where $\widetilde{h}(p)$ is constant of integration with respect to p. Using that minima of $\widehat{\mathcal{D}}$ and $\widetilde{\mathcal{D}}$ are zero, upon substituting $q = 0$, we conclude $\widetilde{h}(p) \equiv 0$. Hence,

$$\widehat{\mathcal{D}}(q) = \widetilde{\mathcal{D}}(\nabla k(p) q), \quad \forall p, q \in P \times Q. \tag{A.82}$$

Using Proposition A.2, we conclude $\nabla \mathcal{T}(p) \equiv A_0$. □

Appendix B. Further discussion on learning potentials

B.1. NN architecture and activation function

While learning energy and dissipation functions offer definitive computational advantages in terms of well-posedness and stability of the macroscopic model defined by the constitutive law, the framework does have limitations that need to be addressed. It is observed that optimisation of neural networks is done using samples of their derivatives in the form of stress measurements. Hence, gradient computation for neural network parameters entails double differentiation of the parameterised functions. This poses a significant constraint on the activation function that can be used for fast training. In the broader machine learning community, the successful training of neural networks with multiple hidden layers has been possible due to the use of activation functions with linear growth, the most popular being ReLU. The linear growth is important for two reasons: (i) use of linear growth compared to superlinear growth results in neural network with multiple hidden layers to be *globally* Lipschitz, that leads to stable training; and (ii) sublinear growth such as hyperbolic tangent and sigmoid, would still result in a globally Lipschitz neural network, they are known to suffer from vanishing gradients and slower training. Moreover, the support of ReLU with its first derivative of order one is unbounded. Hence, when learning functions using samples of their derivatives, such as in this work, it is natural to use ReLUS$(\cdot)$, though at the cost of the neural network losing its globally Lipschitz property, leading to training being prone to blow-ups. In this work, it was observed that the use of more than one hidden layer in neural networks with ReLUS as the activation function makes them extremely difficult to train. Surprisingly, neural networks with one hidden layer and ReLUS activation, while not globally Lipschitz, could still be trained without any blow-ups during the training process. Neural networks of different hidden layer sizes were tried to find the optimal architecture. Fig. B.1 shows convergence of error as the size of neural networks is increased. It also shows the convergence of error with increasing size of the train dataset.

B.2. Data normalisation

Another challenge with learning the potential is related to the normalisation of the dataset. Independent normalisations of input and output datasets to have zero means and unit standard deviations generally lead to faster training of neural networks. However, in the case of learning energies, independent normalisation of input and output datasets interferes with the potential. Mathematically, considering only elastic response,

$$P = \frac{\partial \mathcal{W}(F)}{\partial F} \tag{B.83}$$

where P and F are un-normalized data. Let $\widehat{P} = AP + B$ and $\widehat{F} = CF + D$, where the constant matrices A, B, C and D are chosen such that $\widehat{P}$ and $\widehat{F}$ have zero means and identity standard deviation. Hence, we have

$$A^{-1}(\widehat{P} - B) = \frac{\partial \mathcal{W}(C^{-1}(\widehat{F} - D))}{\partial F}. \tag{B.84}$$

Let $\widehat{\mathcal{W}}$ be such that $\widehat{\mathcal{W}}(\widehat{F}) := \mathcal{W}(C^{-1}(\widehat{F} - D)) + \left(\left(AC^{\top}\right)^{-1} B\right)^{\top} \widehat{F}$. Hence,

$$\widehat{P} = AC^{\top} \frac{\partial \widehat{\mathcal{W}}(\widehat{F})}{\partial \widehat{F}}. \tag{B.85}$$

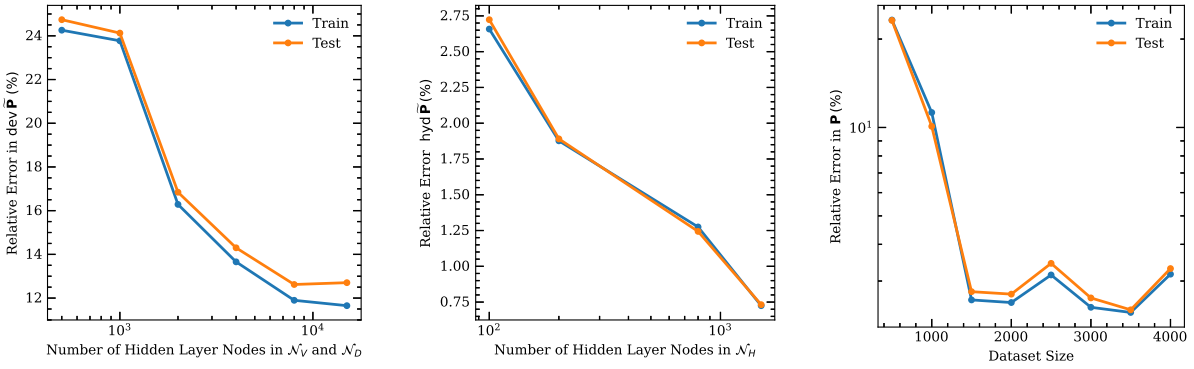

Fig. B.1. (Left) Relative error in deviatoric stress for different sizes of the only hidden layers in neural networks $\mathcal{N}_V$ and $\mathcal{N}_D$. (Middle) Relative Error in hydrostatic stress for different sizes of neural network $\mathcal{N}_H$. (Right) Convergence of error in stress with respect to the size of the training dataset.

It is clear that, unless $AC^\top$ turns out to be identity, normalised stress and normalised strain are not energy conjugates in the classical sense of differentiation. It would be insightful to compare the performances of (B.85) against (B.83), where the energies are to be parameterised as neural networks. However, this comparison is not undertaken in the paper.

B.3. Objectivity

Objectivity, or frame-indifference, is a natural property of constitutive laws that stems from the idea that the response of a material should be independent of an observer's frame of reference. Let $\mathcal{W} : \mathbb{R}^{3\times3}_{+} \to \mathbb{R}$ where $\mathbb{R}^{3\times3}_{+}$ is space of matrices with positive determinant. The generally accepted mathematical definition of objectivity is that $\mathcal{W}$ should be such that, for any deformation gradient $\mathbf{F}$ and any $\mathbf{Q} \in \mathrm{SO}(3)$,

$$\mathcal{W}(\mathbf{F}) = \mathcal{W}(\mathbf{QF}); \quad \mathbf{Q} \in \mathrm{SO}(3). \tag{B.86}$$

We show that, contingent on the definition (B.86) of objectivity, defining $\mathcal{W}$ only on the symmetric deformation gradient tensor and using polar decomposition to extend the function over non-symmetric deformation gradients, is both necessary and sufficient to define an objective $\mathcal{W}$ over all tensors. To this end, let $\mathcal{W}_D : \mathbb{R}^{3\times3}_{sym,+} \to \mathbb{R}$, where $\mathbb{R}^{3\times3}_{sym,+}$ is the space of three-by-three symmetric matrices with positive determinant. We define $\mathcal{W} : \mathbb{R}^{3\times3}_{sym,+} \to \mathbb{R}$ using $\mathcal{W}_D$ as follows.

$$\mathcal{W}(\mathbf{F}) = \begin{cases} \mathcal{W}_D(\mathbf{F}), & \mathbf{F} \in \mathbb{R}^{3\times3}_{sym,+}; \\ \mathcal{W}_D(\widetilde{\mathbf{F}}), & \mathbf{F} = \widetilde{\mathbf{Q}}\widetilde{\mathbf{F}}, \quad \widetilde{\mathbf{Q}} \in \mathrm{SO}(3),\ \widetilde{\mathbf{F}} \in \mathbb{R}^{3\times3}_{sym,+} \end{cases} \tag{B.87}$$

We first show that (B.86) implies (B.87), that is, existence of a $\mathcal{W}_D$. It is straightforward as, using (B.86) twice, we have

$$\mathcal{W}(\mathbf{F}) = \mathcal{W}(\mathbf{Q}\widetilde{\mathbf{Q}}\widetilde{\mathbf{F}}) = \mathcal{W}(\widetilde{\mathbf{F}}), \tag{B.88}$$

where $\mathbf{F} = \widetilde{\mathbf{Q}}\widetilde{\mathbf{F}}$ is the polar decomposition. Hence, for any non-symmetric $\mathbf{F}$, $\mathcal{W}(\mathbf{F})$ can be reduced to be equal to $\mathcal{W}(\widetilde{\mathbf{F}})$ for a symmetric $\widetilde{\mathbf{F}}$. Conversely, we can also show that (B.87) implies (B.86), as using (B.87), we have

$$\mathcal{W}(\mathbf{F}) - \mathcal{W}(\mathbf{QF}) = \mathcal{W}(\widetilde{\mathbf{QF}}) - \mathcal{W}(\mathbf{Q}\widetilde{\mathbf{QF}}) = \mathcal{W}_D(\widetilde{\mathbf{F}}) - \mathcal{W}_D(\widetilde{\mathbf{F}}) = 0. \tag{B.89}$$

B.4. Minimum of D at zero

The constraint in (24), imposing the minimum of the dissipation function at zero, may be enforced after training and it not necessary to have it as a part of the optimisation objective in (25). This can be observed from the loss curves in Fig. B.2, and is theoretically shown as follows. Let $\mathcal{W}$ and D be energy density and dissipation such that D is convex and has a minimum at zero. Hence, the dynamics $\partial\mathcal{W}(F,\xi)/\partial\xi + \partial D(\dot{\xi})/\partial\dot{\xi} = 0$ is thermodynamically consistent. Consider $\widetilde{\mathcal{W}}$ and $\widetilde{D}$ defined as follows, for any $a \in \mathbb{R}^n$, a vector of dimension the same as that of $\dot{\xi}$:

$$\widetilde{\mathcal{W}}(F,\xi) = \mathcal{W}(F,\xi) + a^\top\xi \tag{B.90a}$$

$$\widetilde{D}(\dot{\xi}) = D(\dot{\xi}) - a^\top\dot{\xi}. \tag{B.90b}$$

Clearly, $\widetilde{D}$, which is still a convex function, does not have a minimum at zero. However, the dynamics of ξ defined using $\widetilde{D}$ and $\widetilde{\mathcal{W}}$ is still thermodynamically consistent as

$$\frac{\partial\widetilde{\mathcal{W}}(F,\xi)}{\partial\xi} + \frac{\partial\widetilde{D}(\dot{\xi})}{\partial\dot{\xi}} = 0 \tag{B.91a}$$

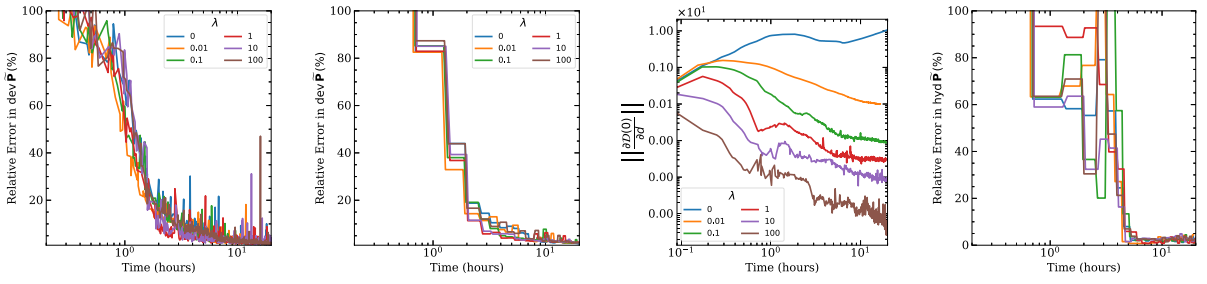

Fig. B.2. Decay in various error terms as the neural networks are trained. (Left-most) Train error and (Middle-left) Test error as $\mathcal{N}_V$ and $\mathcal{N}_D$, neural networks corresponding to deviatoric stress, are trained. Decay in gradient of D at zero for different weights λ. (Right-most) Test error as $\mathcal{N}_H$, the neural network corresponding to hydrostatic stress, is trained. The steps in the test error curves are due to their periodic but less frequent evaluation compared to the evaluation of the training error.

$$\implies \frac{\partial \mathcal{W}(F,\xi)}{\partial \xi} + \frac{\partial \mathcal{D}(\dot{\xi})}{\partial \dot{\xi}} = 0. \tag{B.91b}$$

The dynamics described by $(\widetilde{\mathcal{W}}, \widetilde{\mathcal{D}})$ is exactly the same as that described by $(\mathcal{W}, \mathcal{D})$, hence thermodynamically consistent. Our NN training may learn $(\widetilde{\mathcal{W}}, \widetilde{\mathcal{D}})$ instead of $(\mathcal{W}, \mathcal{D})$, making the constraint in (24) unnecessary. An equivalent interpretation is that once $\widetilde{\mathcal{W}}$ and $\widetilde{\mathcal{D}}$ are learned from data, they may be corrected as per (B.90) to obtain $\mathcal{W}$ and $\mathcal{D}$, with $\mathcal{D}$ having its minimum at zero. In Fig. B.2, we observe that loss curves of relative error in $\operatorname{dev}\widetilde{\mathbf{P}}$ are insensitive to the parameter λ.

References

Amos, B., Xu, L., Kolter, J.Z., 2017. Input convex neural networks. arXiv:1609.07152.

As'ad, F., Avery, P., Farhat, C., 2022. A mechanics-informed artificial neural network approach in data-driven constitutive modeling. Int. J. Numer. Method. Eng. 123 (12), 2738–2759. https://doi.org/10.1002/nme.6957

As'ad, F., Farhat, C., 2023. A mechanics-informed deep learning framework for data-driven nonlinear viscoelasticity. Comput. Method. Appl. Mech. Eng. 417, 116463. https://doi.org/10.1016/j.cma.2023.116463

As'ad, F., Farhat, C., 2026. A staggered training framework for mechanics-informed neural networks in tractable multiscale homogenization with application to woven fabrics. Comput. Method. Appl. Mech. Eng. 452, 118666. https://doi.org/10.1016/j.cma.2025.118666

Ball, J.M., 1976. Convexity conditions and existence theorems in nonlinear elasticity. Arch. Ration. Mech. Anal. 63 (4), 337–403. https://doi.org/10.1007/BF00279992

Ball, J.M., 1977. Constitutive inequalities and existence theorems in nonlinear elastostatics. Nonlinear Analysis and Mechanics, Heriot-Watt Symposium, Vol. 1, https://people.maths.ox.ac.uk/~ball/Articles%20in%20Conference%20Proceedings%20and%20Books/Ball%201977%20Heriot-Watt%20Symposium.pdf.

Bensoussan, A., Lions, J.-L., Papanicolaou, G., 2011. Asymptotic Analysis for Periodic Structures. Vol. 374. American Mathematical Soc.

Bhattacharya, K., Liu, B., Stuart, A., Trautner, M., 2023. Learning Markovian homogenized models in viscoelasticity. Multiscale Model. Simul. 21 (2), 641–679. https://doi.org/10.1137/22M1499200

Bonatti, C., Mohr, D., 2022. On the importance of self-consistency in recurrent neural network models representing elasto-plastic solids. J. Mech. Phys. Solid. 158, 104697. https://doi.org/10.1016/j.jmps.2021.104697

Brenner, R., Suquet, P., 2013. Overall response of viscoelastic composites and polycrystals: exact asymptotic relations and approximate estimates. Int. J. Solid. Struct. 50 (10), 1824–1838. https://doi.org/10.1016/j.ijsolstr.2013.02.011

Coleman, B.D., Gurtin, M.E., 1967. Thermodynamics with internal state variables. J. Chem. Phys. 47 (2), 597–613. https://doi.org/10.1063/1.1711937

Flaschel, M., Steinmann, P., De Lorenzis, L., Kuhl, E., 2025. Convex neural networks learn generalized standard material models. J. Mech. Phys. Solid. 200, 106103. https://doi.org/10.1016/j.jmps.2025.106103

Francfort, G., Mielke, A., 2006. Existence results for a class of rate-independent material models with nonconvex elastic energies. J. für die reine und angewandte Mathematik 2006 (595), 55–91. https://doi.org/10.1515/CRELLE.2006.044

Francfort, G.A., Suquet, P.M., 1986. Homogenization and mechanical dissipation in thermoviscoelasticity. Arch. Ration. Mech. Anal. 96 (3), 265–293. https://doi.org/10.1007/BF00251909

Ghavamian, F., Simone, A., 2019. Accelerating multiscale finite element simulations of history-dependent materials using a recurrent neural network. Comput. Method. Appl. Mech. Eng. 357, 112594. https://doi.org/10.1016/j.cma.2019.112594

Horstemeyer, M.F., Bammann, D.J., 2010. Historical review of internal state variable theory for inelasticity. Int. J. Plast. 26 (9), 1310–1334. Special Issue In Honor of David L. McDowell. https://doi.org/10.1016/j.ijplas.2010.06.005

Im, S., Lee, J., Cho, M., 2021. Surrogate modeling of elasto-plastic problems via long short-term memory neural networks and proper orthogonal decomposition. Comput. Method. Appl. Mech. Eng. 385, 114030. https://doi.org/10.1016/j.cma.2021.114030

Karimi, M., Bhattacharya, K., 2024. A learning-based multiscale model for reactive flow in porous media. Water Resour. Res. 60 (9), e2023WR036303. e2023WR036303 2023WR036303. https://doi.org/10.1029/2023WR036303

Kingma, D.P., Ba, J., 2017. Adam: a method for stochastic optimization. arXiv:1412.6980.

Klein, D.K., Fernández, M., Martin, R.J., Neff, P., Weeger, O., 2022. Polyconvex anisotropic hyperelasticity with neural networks. J. Mech. Phys. Solid. 159, 104703. https://doi.org/10.1016/j.jmps.2021.104703

Kocks, U.F., Tomé, C.N., Wenk, H.-R., 2000. Texture and Anisotropy: Preferred Orientations in Polycrystals and their Effect on Materials Properties. Cambridge university press.

Kovachki, N., Liu, B., Sun, X., Zhou, H., Bhattacharya, K., Ortiz, M., Stuart, A., 2022. Multiscale modeling of materials: computing, data science, uncertainty and goal-oriented optimization. Mech. Mater. 165, 104156. https://doi.org/10.1016/j.mechmat.2021.104156

Lefik, M., Boso, D.P., Schrefler, B.A., 2009. Artificial neural networks in numerical modelling of composites. Comput. Method. Appl. Mech. Eng. 198 (21), 1785–1804. Advances in Simulation-Based Engineering Sciences - Honoring J. Tinsley Oden. https://doi.org/10.1016/j.cma.2008.12.036

Li, Z., Liu-Schiaffini, M., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., Anandkumar, A., 2022. Learning chaotic dynamics in dissipative systems. Adv. Neural Inf. Process. Syst. 35, 16768–16781.

Liu, B., Kovachki, N., Li, Z., Azizzadenesheli, K., Anandkumar, A., Stuart, A.M., Bhattacharya, K., 2022. A learning-based multiscale method and its application to inelastic impact problems. J. Mech. Phys. Solid. 158, 104668. https://doi.org/10.1016/j.jmps.2021.104668

Liu, B., Ocegueda, E., Trautner, M., Stuart, A.M., Bhattacharya, K., 2023. Learning macroscopic internal variables and history dependence from microscopic models. J. Mech. Phys. Solid. 178, 105329. https://doi.org/10.1016/j.jmps.2023.105329

Logarzo, H.J., Capuano, G., Rimoli, J.J., 2021. Smart constitutive laws: inelastic homogenization through machine learning. Comput. Method. Appl. Mech. Eng. 373, 113482.

Mainik, A., Mielke, A., 2005. Existence results for energetic models for rate-independent systems. Calc. Var. Part. Differ. Equ. 22 (1), 73–99. https://doi.org/10.1007/s00526-004-0267-8

Mielke, A., 2003. Energetic formulation of multiplicative elasto-plasticity using dissipation distances. Contin. Mech. Thermodyn. 15 (4), 351–382. https://doi.org/10.1007/s00161-003-0120-x

Mielke, A., 2004. Existence of minimizers in incremental elasto-plasticity with finite strains. SIAM J. Math. Anal. 36 (2), 384–404. https://doi.org/10.1137/S0036141003429906

Mielke, A., Rossi, R., Savaré, G., 2018. Global existence results for viscoplasticity at finite strain. Arch. Ration. Mech. Anal. 227 (1), 423–475. https://doi.org/10.1007/s00205-017-1164-6

Mozaffar, M., Bostanabad, R., Chen, W., Ehmann, K., Cao, J., Bessa, M.A., 2019. Deep learning predicts path-dependent plasticity. Proc. Natl. Acad. Sci. 116 (52), 26414–26420. https://doi.org/10.1073/pnas.1911815116

Ortiz, M., Stainier, L., 1999. The variational formulation of viscoplastic constitutive updates. Comput. Method. Appl. Mech. Eng. 171 (3), 419–444. https://doi.org/10.1016/S0045-7825(98)00219-9

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: an imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F.d., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Pavliotis, G.A., Stuart, A., 2008. Multiscale Methods: Averaging and Homogenization. Vol. 53. Springer Science & Business Media.

Pinkus, A., 1999. Approximation theory of the MLP model in neural networks. Acta Numer. 8, 143–195.

Rice, J.R., 1971. Inelastic constitutive relations for solids: an internal-variable theory and its application to metal plasticity. J. Mech. Phys. Solid. 19 (6), 433–455. https://doi.org/10.1016/0022-5096(71)90010-X
Rockafellar, R.T., 1970. Convex analysis. Princeton Mathematical Series, Princeton University Press, Princeton, N. J.
Schröder, J., Neff, P., Ebbing, V., 2010. Polyconvex energies for trigonal, tetragonal and cubic symmetry groups. In: Hackl, K. (Ed.), IUTAM Symposium on Variational Concepts with Applications to the Mechanics of Materials. Springer Netherlands, Dordrecht, pp. 221–232.
Simo, J.C., Hughes, T. J.R., 1998. Computational Inelasticity. Springer.
Simo, J.C., Hughes, T. J.R., 2006. Computational Inelasticity. Springer. https://doi.org/10.1007/b98904
Wu, L., Nguyen, V.D., Kilingar, N.G., Noels, L., 2020. A recurrent neural network-accelerated multi-scale model for elasto-plastic heterogeneous materials subjected to random cyclic and non-proportional loading paths. Comput. Method. Appl. Mech. Eng. 369, 113234. https://doi.org/10.1016/j.cma.2020.113234