MULTISCALE MODEL. SIMUL.
Vol. 21, No. 2, pp. 641–679

LEARNING MARKOVIAN HOMOGENIZED MODELS IN VISCOELASTICITY*

KAUSHIK BHATTACHARYA†, BURIGEDE LIU‡, ANDREW STUART§, AND MARGARET TRAUTNER§

Abstract. Fully resolving dynamics of materials with rapidly varying features involves expensive fine-scale computations which need to be conducted on macroscopic scales. The theory of homogenization provides an approach for deriving effective macroscopic equations which eliminates the small scales by exploiting scale separation. An accurate homogenized model avoids the computationally expensive task of numerically solving the underlying balance laws at a fine scale, thereby rendering a numerical solution of the balance laws more computationally tractable. In complex settings, homogenization only defines the constitutive model implicitly, and machine learning can be used to learn the constitutive model explicitly from localized fine-scale simulations. In the case of one-dimensional viscoelasticity, the linearity of the model allows for a complete analysis. We establish that the homogenized constitutive model may be approximated by a recurrent neural network that captures the memory. The memory is encapsulated in the evolution of an appropriate finite set of hidden variables, which are discovered through the learning process and dependent on the history of the strain. Simulations are presented which validate the theory. Guidance for the learning of more complex models, such as arise in plasticity, using similar techniques, is given.

Key words. homogenization, operator learning, viscoelasticity, surrogate modeling, machine learning

MSC codes. 74D05, 74Q15, 35J47, 65M32, 74S30

DOI. 10.1137/22M1499200

1. Introduction.

Many problems in continuum mechanics lead to constitutive laws which are history dependent. This property may be inherent to physics beneath the continuum scale (for example, in plasticity [35, 36]) or may arise from homogenization of rapidly varying continua [3, 28] (for example, in the Kelvin–Voigt (KV) model of viscoelasticity [11]). When history dependence is present, Markovian models that capture the history dependence are desirable for both interpretability and computability. In some cases theory may be used to justify Markovian models which capture this history dependence, but in many cases data plays a central role in finding such models. The goal of this paper is to study data-driven methods to learn Markovian models for history dependence and to provide theoretical underpinnings for understanding them.

*Received by the editors May 31, 2022; accepted for publication (in revised form) February 2, 2023; published electronically May 19, 2023. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

https://doi.org/10.1137/22M1499200

Funding: The work of the first, second, and third authors was sponsored by the U.S. Army Research Laboratory and was accomplished under Cooperative Agreement W911NF-12-2-0022. The work of the fourth author was funded by the Department of Energy Computational Science Graduate Fellowship under award DE- SC002111.

†Mechanical and Civil Engineering, California Institute of Technology, Pasadena, CA 91125 USA (bhatta@caltech.edu).

‡Department of Engineering, University of Cambridge, Cambridge, UK (bl377@eng.cam.ac.uk).

§Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125 USA (astuart@caltech.edu, mtrautne@caltech.edu).

The paper [20] adopted a data-driven learning approach to uncovering history-dependent homogenized models arising in crystal plasticity. However, the resulting constitutive model is not causal and instead learns causality approximately from computations performed at the level of the cell problem. The paper [21] introduces a different approach, learning *causal* constitutive models of plasticity. In order to give rigorous underpinnings to the empirical results therein, the present work is devoted to studying the methodology from [21] in the setting of linear one-dimensional viscoelasticity. Here we can use theoretical understanding to justify and validate the methodology; we show that machine-learned homogenized models can accurately approximate the dynamics of multiscale models at much cheaper evaluation cost. We obtain insight into desirable choice of training data to learn the homogenized constitutive model, and we study the effect of the multiple local minimizers which appear in the underlying optimization problem. Furthermore, the rigorous underpinnings enable us to gain insight into how to test model hypotheses. We demonstrate that hypothesizing the correct model leads to robustness with respect to changes in time discretization in the causal model: the model can be trained at one time-step and used at others, and the model can be trained with one time-integration method and used with others. In contrast, hypothesizing an incorrect model leads to intolerable sensitivity with respect to the time-step. Thus training at one time-step and testing at other levels of resolution provides a method for testing model form hypotheses. We work primarily with the one-dimensional KV model for viscoelasticity for which the constitutive model depends only on strain and strain rate. We will also briefly touch upon the standard linear solid (SLS) model for which the constitutive relation depends only on the strain and the strain history and perform numerical experiments in a one-dimensional elasto-viscoplastic material; in so doing we show that the ideas presented extend beyond the specifics of the one-dimensional KV setting.

In subsection 1.1 we describe the overarching mathematical framework adopted, and subsection 1.2 contains a detailed literature review. This is followed, in subsection 1.3, by a statement of our contributions and an overview of the paper. In subsection 1.4 we summarize notation used throughout the remainder of the paper.

1.1. Setup. Consider the problem of material response on an arbitrary spatial domain $\mathcal{D} \subset \mathbb{R}^d$ where the material properties vary rapidly within the domain. We denote by $u_\epsilon \in \mathbb{R}^d$ the displacement, where $\epsilon : 0 < \epsilon \ll 1$ denotes the scale of the material fluctuations. Denote by $\mathcal{T} = (0, T)$ the time domain of interest. We consider continuum models which satisfy dynamical equations of the form

$$\rho \partial_t^2 u_\epsilon = \nabla \cdot \sigma_\epsilon + f, \quad (x,t) \in \mathcal{D} \times \mathcal{T}, \tag{1.1a}$$

$$\sigma_\epsilon(x,t) = \Psi_\epsilon^\dagger \left(\nabla u_\epsilon(x,t), \partial_t \nabla u_\epsilon(x,t), \{\nabla u_\epsilon(x,\tau)\}_{\tau \in \overline{\mathcal{T}}}, x, t \right), \quad (x,t) \in \mathcal{D} \times \mathcal{T}, \tag{1.1b}$$

$$u_\epsilon = u^*, \quad \partial_t u_\epsilon = v^*, \quad (x,t) \in \mathcal{D} \times \{0\}, \tag{1.1c}$$

$$u_\epsilon = 0, \quad x \in \partial\mathcal{D}, \quad (x,t) \in \partial\mathcal{D} \times \mathcal{T}. \tag{1.1d}$$

From these equations we seek $u_\epsilon : \mathcal{D} \times \mathcal{T} \mapsto \mathbb{R}^d$. Equation (1.1a) is the balance equation with inertia term $\rho \partial_t^2 u_\epsilon$ for known parameter $\rho \in \mathbb{R}^+$, resultant stress term $\nabla \cdot (\sigma_\epsilon)$ where $\sigma_\epsilon \in \mathbb{R}^{d \times d}$ is the internal stress tensor, and known external forcing $f \in \mathbb{R}^d$; equations (1.1c) and (1.1d) specify the initial and boundary data for the displacement. Equation (1.1b) is the constitutive law relating properties of the strain ∇u_ϵ to the stress σ_ϵ via map $\Psi_\epsilon^\dagger$. In this paper we will consider this model with inertia ($\rho > 0$) and without inertia ($\rho \equiv 0$).

1.1.1. Constitutive model. The constitutive model is defined by $\Psi_\epsilon^\dagger : \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times d} \times C(\overline{\mathcal{T}};\mathbb{R}^{d\times d}) \times \mathcal{D} \times \mathcal{T} \to \mathbb{R}^{d\times d}$. It takes time t as input, which enables the stress at time t to be expressed only in terms of the strain history up to time t, $\{\nabla u_\epsilon(x,\tau)\}_{\tau=0}^t$, and not on the future of the strain for $\tau \in (t,T]$. It takes x as input to allow for material properties which depend on the rapidly varying x/ϵ; it is also possible to allow for material properties which exhibit additional dependence on the slowly varying x, but we exclude this case for simplicity.

Such constitutive models include a variety of plastic, viscoelastic, and viscoplastic materials. In this paper we focus mainly on the one-dimensional KV viscoelastic setting in order to highlight ideas; in this case $\Psi_\epsilon^\dagger$ is independent of the history of the strain. However, we will briefly demonstrate that similar concepts relating to the learning of constitutive models also apply to the case of an SLS, for which $\Psi_\epsilon^\dagger$ is independent of the strain rate but does depend on the history of the strain; the SLS contains the KV and Maxwell models of viscoelasticity as special cases. We also provide a demonstrative application of our methodology to the setting of elasto-viscoplasticity. Furthermore, the paper [21] includes empirical evidence that similar concepts relating to the learning of constitutive models also apply in higher dimensions.

1.1.2. Homogenized constitutive model. The goal of homogenization is to find constitutive models which eliminate small-scale dependence. To this end, we first discuss the form of a general homogenized problem: we seek the equation satisfied by u_0, an appropriate limit of u_ϵ as $\epsilon \to 0$. Then map $\Psi_0^\dagger$ defines the constitutive relationship in a homogenized model for u_0 of the form

$$\rho\partial_t^2 u_0 = \nabla\cdot\sigma_0 + f, \quad (x,t)\in\mathcal{D}\times\mathcal{T}, \tag{1.2a}$$

$$\sigma_0(x,t) = \Psi_0^\dagger\left(\nabla u_0(x,t), \partial_t\nabla u_0(x,t), \{\nabla u_0(x,\tau)\}_{\tau\in\overline{\mathcal{T}}}, t\right), \quad (x,t)\in\mathcal{D}\times\mathcal{T}, \tag{1.2b}$$

$$u_0 = u^*, \quad \partial_t u_0 = v^*, \quad (x,t)\in\mathcal{D}\times\{0\}, \tag{1.2c}$$

$$u_0 = 0, \qquad (x,t)\in\partial\mathcal{D}\times\mathcal{T}. \tag{1.2d}$$

The key property of this homogenized model is that parameter ϵ no longer appears. Furthermore, since we assumed that the multiscale model material properties depend only on the rapidly varying scale x/ϵ and not on x, we have that $\Psi_0^\dagger$ does not depend explicitly on x; it does, however, still have spatial dependence through the local values of strain, strain rate, and strain history: $\Psi_0^\dagger : \mathbb{R}^{d\times d}\times\mathbb{R}^{d\times d}\times C(\overline{\mathcal{T}};\mathbb{R}^{d\times d})\times\mathcal{T}\to\mathbb{R}^{d\times d}$.

If the homogenized model is identified correctly, then dynamics under the multiscale model $\Psi_\epsilon^\dagger$, i.e., u_ϵ, can approximated by dynamics under the homogenized model $\Psi_0^\dagger$, i.e., u_0. This potentially facilitates cheaper computations since length-scales of size ϵ need not be resolved. We observe, however, that for KV viscoelasticity, the homogenized model contains history dependence (memory) even though the multiscale model does not. Markovian history dependence is desirable for two primary reasons: first, Markovian models encode conceptual understanding, representing the history dependence in a compact, interpretable form; second, Markovian expression reduces computational cost from $\mathcal{O}(|\mathcal{T}|^2)$ in the general memory case to $\mathcal{O}(|\mathcal{T}|)$ in the Markovian case. In the general media setting, for a multitude of models in viscoelasticity, viscoplasticity, and plasticity, the homogenized model will depend on the memory in a non-Markovian manner. However, it is interesting to determine situations in which accurate Markovian approximations can be found.

1.1.3. Markovian homogenized constitutive model. We will seek to identify *hidden variables* ξ, closely related to the *internal variables* used in the mechanics

literature [21], and functions $\mathcal{F}, \mathcal{G}$ such that, for $B \in C(\mathcal{T}; \mathbb{R}^{d\times d})$, $\Psi_0^\dagger$ can be approximated by

$$\Psi_0\left(B(t), \partial_t B(t), \{B(\tau)\}_{\tau\in\overline{\mathcal{T}}}, t\right) = \mathcal{F}\left(B(t), \partial_t B(t), \xi(t)\right), \tag{1.3a}$$

$$\partial_t \xi(t) = \mathcal{G}\left(\xi(t), B(t)\right), \tag{1.3b}$$

$$\xi(0) = 0. \tag{1.3c}$$

Note that ξ carries the history dependence on B through its Markovian evolution. We assume that $\xi \in C(\mathcal{T}; \mathbb{R}^r)$ for some integer r and hence that $\mathcal{F} : \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times d} \times \mathbb{R}^r \to \mathbb{R}^{d\times d}$ and that $\mathcal{G} : \mathbb{R}^r \times \mathbb{R}^{d\times d} \to \mathbb{R}^r$. In dimension $d > 1$ there will be further symmetries that should be built into the model, but as the concrete analysis in this paper is in dimension $d = 1$ we will not detail these symmetries here [34].

In general such a Markovian model can only *approximate* the true model, and the nature of the physics leading to a good approximation will depend on the specific continuum mechanics problem. To determine $\mathcal{F}$ and $\mathcal{G}$ in practice we will parameterize them as neural networks, which enables us to use general purpose optimization software to determine suitable values of the parameters. Within computational implementations of the learned homogenized models, the neural networks $\mathcal{F}$ and $\mathcal{G}$ act pointwise in time to generate the stress and time derivatives of the hidden variables at each time-step. In doing so we identify an operator class $\Psi_0(\cdot\,;\theta)$ and parameter space Θ such that, for some judiciously chosen $\theta^* \in \Theta$, $\Psi_0(\cdot\,;\theta^*) \approx \Psi_0^\dagger$.

In this paper we will concentrate on justifying a Markovian homogenized approximation in the context of one-dimensional KV viscoelasticity. Our justification will use theory that is specific to one-dimensional linear viscoelasticity, and we demonstrate that the approach also works for the general SLS, which includes the KV model as a particular limit. Furthermore, the paper [21] contains evidence that the ideas we develop apply beyond the confines of one-dimensional linear viscoelasticity and into nonlinear plasticity in higher spatial dimensions.

The specific property of one-dimensional viscoelasticity that we exploit to underpin our analysis (and which applies to the SLS and therefore also to the KV model) is that, for piecewise-constant media, the homogenized model has a memory term which can be represented in a Markovian way. Therefore, to justify our strategy of approximating by Markovian models we will do the following: first, approximate the rapidly varying medium by a piecewise-constant rapidly varying medium; second, homogenize this model to find a Markovian description; and, finally, demonstrate how the Markovian description can be learned from data at the level of the unit cell problem, using neural networks. For more general problems we anticipate a similar justification holding, but with different specifics leading to the existence of good approximate Markovian homogenized models. The benefit of the one-dimensional viscoelastic setting is that, through theory, we obtain underpinning insight into the conceptual approach more generally, and in particular for plasticity; this theory underpins the numerical experiments which follow.

1.2. Literature review. The continuum assumption for physical materials approximates the inherently particulate nature of matter by a continuous medium and thus allows the use of partial differential equations to describe response dynamics. We refer the reader to [13, 37, 12] for a general background. In continuum mechanics, the governing equations are derived by combining universal balance laws of physics (balance of mass, momenta, and energy) with a constitutive relation that describes the properties of the material being studied. This is typically specified as the relation

between a dynamic quantity like stress or energy and kinematic quantities like strain and its history. The constitutive relation of many materials are history dependent, i.e., the state of stress at an instant depends on the history of deformation. It is common in continuum mechanics to incorporate this history dependence through the introduction of internal variables, which are referred to as hidden variables in computer science and throughout this paper. We refer the reader to [32] for a systematic formulation of internal variable theories.

Of particular interest in this work are viscoelastic materials. We refer the reader to [7, 18] for a general background. In viscoelastic materials, the state of stress at any instant depends on the strain and its history. There are various models where the stress depends only on strain and strain rate (Kelvin–Voigt), internal variables (standard linear solids), convolution kernels, and fractional time derivatives.

While constitutive laws were traditionally determined empirically, more recently there has been a systematic attempt to understand them from more fundamental solids, and this has given rise to a rich activity in multiscale modeling of materials [10, 39, 41]. Materials are heterogeneous on various length (and time) scales, and it is common to use different theories to describe the behavior at different scales [30]. The goal of multiscale modeling of materials is to use this hierarchy of scales to understand the overall constitutive behavior at the scale of applications. The hierarchy of scales includes a number of continuum scales. For example, a composite material is made of various distinct materials arranged at a scale that is small compared to the scale of application but large enough compared to an atomistic/discrete scale, so the behavior is adequately described by continuum mechanics. Or, for example, a polycrystal is made of a large collection of grains (regions of identical anisotropic material but with differing orientation) that are small compared to the scale of application but large enough for a continuum theory. Homogenization theory leverages the assumption of the separation of scales to average out the effects of fine-scale material variations. To estimate macroscopic response of heterogeneous materials, asymptotic expansion of the displacement field yields a set of boundary value problems whose solution produces an approximation that does not depend on the microscale [3, 33]. The fundamentals of asymptotic homogenization theory are well established [28, 8, 1]. Milton [23] provides a comprehensive survey of the effective or homogenized properties.

Homogenization in the context of viscoelasticity was initiated by Sanchez-Palencia [33, Chapter 6], who pointed out that the homogenization of a KV model leads to a model with fading memory. Further discussion of homogenization theory in (thermo-) viscoelasticity can be found in Francfort and Suquet [11], and a detailed discussion of the overall behavior including memory in Brenner and Suquet [5]. A broader discussion of homogenization and memory effects can be found in Tartar [40]. It is now understood that homogenization of various constitutive models gives rise to memory.

As noted above, according to homogenization theory, the macroscopic behavior depends on the solution of a boundary value problem at the microscale. Evaluating the macroscopic behavior by the solution of a boundary value problem computationally leads to what has been called computational micromechanics [43]. These often involve periodic boundary conditions, and fast Fourier transform–based methods have been widely used since Moulinec and Suquet [26] (see [24, 25] for recent summaries). While these enable us to compute the macroscopic response for a particular deformation history, one needs to repeat the calculation for all possible deformation histories.

Therefore, recent work in the mechanics literature addresses the issue of learning homogenized constitutive models from computational data [27, 42, 20, 21] or

experimental data [2]. This learning problem requires determination of maps that take as inputs functions describing microstructural properties and leads into the topic of operator learning.

Neural networks on finite-dimensional spaces have proven effective at solving long-standing computational problems. A natural question which arises from this success is whether neural networks can be used to solve partial differential equations (PDEs). From a theoretical standpoint, the classical notion of neural networks as maps between finite-dimensional spaces is insufficient. Indeed, a differential operator is a map between infinite-dimensional spaces. Similarly to the case of finite-dimensional maps, universal approximation results for nonlinear operator learning have been developed [6]. In one approach to operator learning, model reduction methods are applied to the operator itself. In this setting, a low-dimensional approximation space is assumed, and the operator approximation is constructed via regression using the latent basis [29, 31]. In a second method, classical dimension reduction maps are applied to the input and output spaces, and an approximation map is learned between these finite-dimensional spaces [4]. An important result of this method is its mesh-invariance property. Data for any operator-learning problem must consist of a finite discretization of the true input and output functions. One side effect of this practical fact is that some existing methods for operator learning depend critically on the choice of discretization. Mesh-invariant methods are desirable for both practical and theoretical purposes. Practically, it would be expensive to train a new model to accommodate a finer data resolution. From a theoretical standpoint, since the true map between infinite-dimensional spaces is inherently resolution independent, the operator approximation ought to have this property as well. Mesh invariance allows the operator approximation to be applied in the use of various numerical approximation methods for PDEs, which is of particular importance when the operator is being used as a surrogate model for the overall PDE system [19]. Invariance with respect to time discretization is also leveraged in work on data-driven learning of PDEs in [17, 16].

Surrogate modeling bypasses expensive simulation computations by replacing part of the PDE with a neural network. In several application settings, including fluid flows and solid mechanics computations, surrogate modeling has met empirical success in approximating the true solution [38, 15]. This work continues to use ideas from physics-informed machine learning; in [15] in particular, the differential operator is incorporated into the cost function via automatic differentiation. Other work in surrogate modeling for solid mechanics proposes a hierarchical network architecture that mimics the heterogeneous material structure to yield an approximation to the homogenized solution [22]. The work of [14] also leverages automatic differentiation and approaches the problem of surrogate modeling of constitutive laws via a parameter identification learning perspective. In our work, we consider parameterized models for the purpose of justifying theory, but the learning framework does not rely on parameter identification.

In this paper we use a recurrent neural network (RNN) as a surrogate model for the constitutive relation on the microscale. Our RNN architecture takes the form of two feed-forward neural networks: one which computes the time derivative of hidden variables, and one which outputs the stress pointwise in space and time. In this manner, the history dependence is contained entirely in the hidden variables rather than directly in the neural network. This leads to an interpretable model. The RNN can then be used to evaluate the forward dynamic response on the microscale cells, whose results are combined with traditional numerical approximation methods to yield the macroscale response. Furthermore, the RNN that we train at a particular

time discretization is also accurate when used at other time discretizations if the correct model form is proposed. However, the RNN is trained for a particular choice of material parameters, and the resulting model cannot be used at different material parameter values; simultaneously learning material parameter dependence is left for future work.

1.3. Our contributions and paper overview. Our contributions are as follows:

1. We provide theoretical underpinnings for the discovery of Markovian homogenized models in viscoelasticity and plasticity; the methods use data generated by solving cell problems to learn constitutive models of the desired form.
2. We prove that in the one-dimensional KV setting, any solution of the multiscale problem can be approximated by the solution of a homogenized problem with Markovian structure.
3. We prove that the constitutive model for this Markovian homogenized system can be approximated by an RNN operator learned from data generated by solving the appropriate cell problem.
4. We provide simulations which numerically demonstrate the accuracy of the learned Markovian model.
5. We offer guidance for the application of this methodology, beyond the setting of one-dimensional viscoelasticity, into multidimensional plasticity.

In section 2, we formulate the KV viscoelastic problem and its homogenized solution. In section 3, we present our main theoretical results, addressing contributions 1, 2, and 3; these are in the setting of one-dimensional KV viscoelasticity. We prove that solution of the multiscale problem can be approximated by solution of a homogenized Markovian memory-dependent model that does not depend on small scales, and we prove that an RNN can approximate the constitutive law for this homogenized problem. Section 4 contains numerical experiments which address contributions 4 and 5; the start of that section details the findings.

1.4. Notation. We define notation that will be useful throughout the paper. Recall that $\mathcal{T} = (0, T)$ is the time domain of interest, and in the one-dimensional setting we let $\mathcal{D} = [0, L]$ be the spatial domain. Let $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ denote the standard inner product and induced norm operations on the Hilbert space $L^2(\mathcal{D}; \mathbb{R})$. Additionally, let $\| \cdot \|_\infty$ denote the $L^\infty(\mathcal{D}; \mathbb{R})$ norm.

It will also be convenient to define the ξ-dependent quadratic form

$$q_\xi(u, w) := \int_{\mathcal{D}} \xi(x) \frac{\partial u(x)}{\partial x} \frac{\partial w(x)}{\partial x} \, dx \tag{1.4}$$

for arbitrary $\xi \in L^\infty(\mathcal{D}; (0, \infty))$; furthermore we define

$$\xi^+ := \operatorname*{ess\,sup}_{x \in \mathcal{D}} \xi(x) < \infty \tag{1.5}$$

and

$$\xi^- := \operatorname*{ess\,inf}_{x \in \mathcal{D}} \xi(x) \geq 0. \tag{1.6}$$

In this paper we always work with ξ such that $\xi^- > 0$. Under these assumptions $q_\xi(\cdot, \cdot)$ defines an inner product, and we can define the norm

$$\|u\|^2_{H^1_0, \xi} := q_\xi(u, u)$$

from it; note also that we may define a norm on $H^1_0(\mathcal{D}; \mathbb{R})$ by

$$\|u\|_{H_0^1}^2 := q_{\mathbb{1}}(u,u),$$

where $\mathbb{1}(\cdot)$ is the function in $L^\infty(\mathcal{D};(0,\infty))$ taking value 1 in D a.e. The resulting norms are all equivalent on the space $H_0^1(\mathcal{D};\mathbb{R})$; this is a consequence of the following lemma.

LEMMA 1.1. *For any $\xi_1,\xi_2 \in L^\infty(\mathcal{D};(0,\infty))$ satisfying properties* (1.5) *and* (1.6), *the norms $\|u\|_{H_0^1,\xi_1}$ and $\|u\|_{H_0^1,\xi_2}$ are equivalent in the sense that*

$$\frac{\xi_2^-}{\xi_1^+}\|u\|_{H_0^1,\xi_1}^2 \le \|u\|_{H_0^1,\xi_2}^2 \le \frac{\xi_2^+}{\xi_1^-}\|u\|_{H_0^1,\xi_1}^2.$$

Proof. For $i=1,2$,

$$\xi_i^- \int_0^1 \left|\frac{\partial u}{\partial x}\right|^2 dx \le \int_0^1 \xi_i(x)\left|\frac{\partial u}{\partial x}\right|^2 dx \le \xi_i^+ \int_0^1 \left|\frac{\partial u}{\partial x}\right|^2 dx.$$

The result follows. ∎

We denote by V the Hilbert space $H_0^1(\mathcal{D};\mathbb{R})$ noting that, as a consequence of the preceding lemma, we may use $q_\xi(u,w)$ as the inner product on this space for any ξ satisfying (1.5) and (1.6). We also define

$$\mathcal{Z} = L^\infty(\mathcal{T};L^2(\mathcal{D};\mathbb{R})), \quad \mathcal{Z}_2 = L^2(\mathcal{T};L^2(\mathcal{D};\mathbb{R}))$$

with norms

$$\|r\|_{\mathcal{Z}} = \operatorname*{ess\,sup}_{t\in\mathcal{T}}(\|r(\cdot,t)\|), \quad \|r\|_{\mathcal{Z}_2} = \left(\int_0^{\mathcal{T}} \|r(\cdot,t)\|^2 dt\right)^{\frac{1}{2}}.$$

We note that $\mathcal{Z}$ is continuously embedded into $\mathcal{Z}_2$.

2. One-dimensional Kelvin–Voigt viscoelasticity. This paper is focused on one-dimensional KV viscoelasticity because the model is amenable to rigorous analysis. The resulting analysis sheds light on the learning of constitutive models more generally. In section 2.1 we present the equations for the model in an informal fashion and in a weak form suitable for analysis; in section 2.2 we homogenize the model and define the operator defining the effective constitutive model.

2.1. Governing equations and weak form. The one-dimensional KV model for viscoelasticity postulates that stress is affine in the strain and strain rate, with affine transformation dependent on the (typically spatially varying) material properties. For a multiscale material varying with respect to x/ϵ we thus have the following definition of $\Psi_\epsilon^\dagger$ from (1.1), in the one-dimensional KV model:

$$\sigma_\epsilon = E_\epsilon \partial_x u_\epsilon + \nu_\epsilon \partial_{xt}^2 u_\epsilon,$$

where $E_\epsilon(x) = E(\frac{x}{\epsilon})$ and $\nu_\epsilon(x) = \nu(\frac{x}{\epsilon})$ are rapidly varying material elasticity and viscosity, respectively. Both E and ν are assumed to be 1-periodic. Then equations (1.1) without inertia ($\rho \equiv 0$) become

$$-\partial_x\left(E_\epsilon \partial_x u_\epsilon + \nu_\epsilon \partial_{xt}^2 u_\epsilon\right) = f, \quad (x,t) \in \partial\mathcal{D}\times\mathcal{T}, \tag{2.1a}$$

$$u_\epsilon(0,x) = u^*, \quad \partial_t u_\epsilon(0,x) = v^*, \quad x \in \partial\mathcal{D}, \tag{2.1b}$$

$$u_\epsilon(x,t) = 0, \quad x \in \partial\mathcal{D}. \tag{2.1c}$$

Any classical solution to equations (2.1) will also solve the corresponding *weak form*: find $u_\epsilon \in \mathcal{C}(\mathcal{T}; V)$ such that

$$q_{\nu_\epsilon}(\partial_t u_\epsilon, \varphi) + q_{E_\epsilon}(u_\epsilon, \varphi) = \langle f, \varphi \rangle \tag{2.2}$$

for all test functions $\varphi \in V$.

2.2. Homogenization. In the inertia-free setting $\rho = 0$ we perform homogenization to eliminate the dependence on the small scale ϵ in (2.1). First, we take the Laplace transform of (2.1), which, for Laplace parameter s and with the hat symbol denoting Laplace transform, gives

$$\begin{aligned} -\partial_x\left((E_\epsilon + \nu_\epsilon s)\partial_x \widehat{u}_\epsilon\right) &= \widehat{f}, \quad x \in \mathcal{D}, \\ \widehat{u}_\epsilon(x,s) &= 0, \quad x \in \partial\mathcal{D}. \end{aligned}$$

The initial condition $u_\epsilon(0,x) = u^*$ is applied upon Laplace inversion. Since $\epsilon \ll 1$, we may apply standard techniques from multiscale analysis [3, 28] and seek a solution in the form

$$\widehat{u}_\epsilon = \widehat{u}_0 + \epsilon \widehat{u}_1 + \epsilon^2 \widehat{u}_2 + \cdots.$$

For convenience, define $\widehat{a}(s,y) = E(y) + \nu(y)s$. Note that $\widehat{a}(s,\cdot)$ is 1-periodic. The leading order term in our approximation, $\widehat{u}_0$, solves the following uniformly elliptic PDE with Dirichlet boundary conditions:

$$-\partial_x\left(\widehat{a}_0(s)\partial_x \widehat{u}_0\right) = \widehat{f} \quad \text{for} \quad x \in \mathcal{D}, \tag{2.3a}$$

$$\widehat{u}_0 = 0 \quad \text{for} \quad x \in \partial\mathcal{D}. \tag{2.3b}$$

Here the coefficient $\widehat{a}_0$ is given by

$$\widehat{a}_0(s) = \int_0^1 \left(\widehat{a}(s,y) + \widehat{a}(s,y)\partial_y \chi(y)\right) dy$$

and $\chi(y) : [0,1] \to \mathbb{R}$ satisfies the *cell problem*

$$-\partial_y\left(\widehat{a}(s,y)\partial_y \chi(y)\right) = \partial_y \widehat{a}(s,y), \tag{2.4a}$$

$$\chi \text{ is 1-periodic}, \quad \int_0^1 \chi(y)dy = 0. \tag{2.4b}$$

Using this, the coefficient $\widehat{a}_0$ can be computed explicitly as the harmonic average of the original coefficient $\widehat{a}$ [28]:

$$\widehat{a}_0(s) = \left\langle \widehat{a}(s,y)^{-1} \right\rangle^{-1} = \left(\int_0^1 \frac{dy}{s\nu(y) + E(y)} \right)^{-1}, \tag{2.5}$$

where $\langle \cdot \rangle$ denotes spatial averaging over the unit cell.

Equations (2.3) indicate that the homogenized map $\Psi_0^\dagger$ appearing in (1.2) is, for one-dimensional linear viscoelasticity, defined from

$$\Psi_0^\dagger\left(\partial_x u_0(x,t), \partial_{xt}^2 u_0(x,t), \{\partial_x u_0(x,\tau)\}_{\tau \in \overline{\mathcal{T}}}, t\right) = \mathcal{L}^{-1}\left(\widehat{a}_0(s)\partial_x \widehat{u}_0\right); \tag{2.6}$$

here $\mathcal{L}^{-1}$ denotes the inverse Laplace transform. Note that (2.5) shows that $\widehat{a}_0$ grows linearly in $s \to \infty$, and computing the constant term in a regular power series expansion at $s = \infty$ shows that we may write

$$\widehat{a}_0 = \nu' s + E' + \widehat{\kappa}(s),$$

where $\widehat{\kappa}(s)$ decays to 0 as $s \to \infty$. Here

$$\nu' = \Big(\int_0^1 \frac{1}{\nu(y)} dy\Big)^{-1}, \quad E' = \Big(\int_0^1 \frac{E(y)}{\nu(y)^2} dy\Big) \Big/ \Big(\int_0^1 \frac{1}{\nu(y)} dy\Big)^2.$$

Details are presented in Appendix B.1. Laplace inversion of $\widehat{a}_0(s)\partial_x \widehat{u}_0$ then yields the conclusion that

$$\Psi_0^\dagger(\partial_x u_0(t), \partial_{xt}^2 u_0(t), \{\partial_x u_0(\tau)\}_{\tau \in \overline{\mathcal{T}}.t}; \theta) = E' \partial_x u_0(t) + \nu' \partial_{xt}^2 u_0(t) + \int_0^t \kappa(t-\tau) \partial_x u_0(\tau)\, d\tau. \tag{2.7}$$

Remark 2.1. When $\rho = 0$, the homogenized solution provably approximates u_ϵ in the $\epsilon \to 0$ limit; see Theorem 3.7. However, although we derived it with inertia set to zero, the homogenized solution given by (2.8) is also valid when the inertia term $\rho \partial_t^2 u_\epsilon$ generates contributions which are $\mathcal{O}(1)$ with respect to ϵ.

The homogenized PDE for one-dimensional viscoelasticity follows by combining equations (1.2) with (2.7) to give

$$\rho \partial_t^2 u_0 = \nabla \cdot \sigma_0 + f, \quad (x,t) \in \mathcal{D} \times \mathcal{T}, \tag{2.8a}$$

$$\sigma_0(t) = E' \partial_x u_0(t) + \nu' \partial_{xt}^2 u_0(t) + \int_0^t \kappa(t-\tau) \partial_x u_0(\tau)\, d\tau \quad (x,t) \in \mathcal{D} \times \mathcal{T}, \tag{2.8b}$$

$$u_0 = u^*, \quad \partial_t u_0 = v^*, \quad (x,t) \in \mathcal{D} \times \{0\}, \tag{2.8c}$$

$$u_0 = 0, \qquad (x,t) \in \partial\mathcal{D} \times \mathcal{T}. \tag{2.8d}$$

The price paid for homogenization is dependence on the strain history. We will show in the next section, however, that we can approximate the general homogenized map with one in which the history dependence is expressed in a Markovian manner.

3. Main theorems: Statement and interpretation. In this section we present theoretical results of three types. First, in subsection 3.1, we show that the solution u_ϵ to (2.1) is Lipschitz when viewed as a mapping from the unit cell material properties $E(\cdot), \nu(\cdot)$ in L^∞ into $\mathcal{Z}$; hence, an $\mathcal{O}(\delta)$ approximation of E, ν by piecewise-constant functions leads to an $\mathcal{O}(\delta)$ approximation of u_ϵ. Second, in subsection 3.2, we demonstrate that the homogenized model based on piecewise-constant material properties can be represented in a Markovian fashion by introducing *hidden variables*; hence, combining with the first point, we have a mechanism to approximate u_ϵ by solving a Markovian homogenized model. Third, in subsection 3.3, we show the existence of neural networks which provide arbitrarily good approximation of the constitutive law arising in the Markovian homogenized model; this suggests a model class within which to learn homogenized, Markovian constitutive models from data. Subsection 3.4 establishes our framework for the optimization methods used to learn such constitutive models; this framework is employed in section 4.

Assumptions 3.1. We will make the following assumptions on E, ν, and f throughout:

1. $f \in L^2(\mathcal{D};\mathbb{R})$ for all $t \in \overline{\mathcal{T}}$; thus $\|f\|_{\mathcal{Z}} < \infty$;
2. $E^+, \nu^+ < \infty$, and $E^-, \nu^- > 0$.

Note that $E^+ = E^+_\epsilon$ and $\nu^+ = \nu^+_\epsilon$, so we will drop the ϵ superscript in this notation.

3.1. Approximation by piecewise-constant material. Consider (2.1) with continuous material properties E and ν. We show in Theorem 3.4 that we can approximate the solution u_ϵ to this system by a solution u^{PC}_ϵ which solves (2.1) with suitable piecewise-constant material properties E^{PC} and ν^{PC}, in such a way that u_ϵ and u^{PC}_ϵ are close. To this end we make precise the definition of piecewise-constant material properties.

DEFINITION 3.2 (piecewise constant). *A material is piecewise constant on the unit cell with L pieces if the elasticity function $E(y)$ and the viscosity function $\nu(y)$ both take constant values on L intervals $[0, a_1), [a_1, a_2), \ldots, [a_{L-1}, 1]$. In particular, $E(y)$ and $\nu(y)$ have discontinuities only at the same $L-1$ points in the unit cell. We use the terminology L-piecewise constant to specify the number of pieces.*

Remark 3.3. The situation in which $E(y)$ and $\nu(y)$ have discontinuities at different values of $y \in (0,1)$ can be reduced to the case in Definition 3.2 by increasing the value of L.

THEOREM 3.4 (piecewise-constant approximation). *Let E and ν be piecewise-continuous functions, with a finite number of discontinuities, satisfying Assumptions* 3.1*; let u_ϵ be the corresponding solution to* (2.1)*. Then, for any $\delta > 0$, there exist piecewise-constant E^{PC} and ν^{PC} (in the sense of Definition* 3.2*) such that solution u^{PC}_ϵ of equations* (2.1) *with these material properties satisfies*

$$\|u^{PC}_\epsilon - u_\epsilon\|_{\mathcal{Z}} < \delta.$$

Note that Theorem 3.4 is stated in the setting of no inertia. The proof depends on the following lemma; proof of both the theorem and the lemma may be found in Appendix A.1. We observe that, since the Lipschitz result is in the L^∞ norm with respect to the material properties, it holds with constant C independent of ϵ, in the case of interest where the material properties vary rapidly on scale ϵ.

LEMMA 3.5 (Lipschitz solution). *Let u_i be the solution to*

$$-\partial_x \left(E_i \partial_x u_i + \nu_i \partial^2_{xt} u_i \right) = f, \quad (x,t) \in \partial\mathcal{D} \times \mathcal{T}, \tag{3.1}$$

$$u_i(x,t) = u^*, \quad (x,t) \in \mathcal{D} \times \{0\}, \tag{3.2}$$

$$u_i(x,t) = 0, \quad (x,t) \in \partial\mathcal{D} \times \mathcal{T}, \tag{3.3}$$

associated with material properties E_i, ν_i for $i \in \{1,2\}$, and forcing f, all satisfying Assumptions 3.1*. Then*

$$\|u_1 - u_2\|_{\mathcal{Z}} \le C \left(\|\nu_1 - \nu_2\|_\infty + \|E_1 - E_2\|_\infty \right)$$

for some constant $C \in \mathbb{R}^+$ dependent on $f, E_i^+, E_i^-, \nu_i^+, \nu_i^-$, and L and independent of ϵ.

3.2. Homogenization for piecewise-constant material. We show in Theorem 3.6 that for piecewise-constant material properties $E(\cdot)$ and $\nu(\cdot)$, the homogenized map $\Psi_0^\dagger$ given in (2.7) can be written explicitly with a finite number of parameters, and in particular the memory is expressible in a Markovian form. This Markovian form implicitly defines a finite number of hidden variables.

THEOREM 3.6 (existence of exact parametrization). *Let $\Psi_0^\dagger$ be the map from strain history to stress in the homogenized model, as defined by (2.7), in a piecewise-constant material with $L'+1$ pieces. Define $\Psi_0^{PC}: \mathbb{R}^2 \times C(\overline{\mathcal{T}};\mathbb{R}) \times \mathcal{T} \times \Theta \to \mathbb{R}$ by*

(3.4a)
$$\Psi_0^{PC}(\partial_x u_0(t), \partial_{xt}^2 u_0(t), \{\partial_x u_0(\tau)\}_{\tau\in\overline{\mathcal{T}}}, t; \theta) = E_0 \partial_x u_0(t) + \nu_0 \partial_{xt}^2 u_0(t) - \sum_{\ell=1}^{L_0} \xi_\ell(t),$$

(3.4b)
$$\partial_t \xi_\ell(t) = \beta_\ell \partial_x u_0(t) - \alpha_\ell \xi_\ell(t),\ \xi_\ell(0) = 0, \quad \ell \in \{1,\dots,L_0\},$$

with parameter space

$$\Theta = \left(E_0 \in \mathbb{R}_+,\ \nu_0 \in \mathbb{R}_+,\ L_0 \in \mathbb{Z}_+,\ \alpha_0 \in \mathbb{R}_+^{L_0},\ \beta_0 \in \mathbb{R}^{L_0}\right). \tag{3.5}$$

Then, under Assumptions 3.1, there exists $\theta^* \in \Theta$ with $(E_0, \nu_0, L_0, \alpha_0, \beta_0) = (E', \nu', L', \alpha, \beta)$ such that

$$\Psi_0^\dagger(\partial_x u_0(t), \partial_{xt}^2 u_0(t), \{\partial_x u_0(\tau)\}_{\tau\in\overline{\mathcal{T}}}, t) = \Psi_0^{PC}(\partial_x u_0(t), \partial_{xt}^2 u_0(t), \{\partial_x u_0(\tau)\}_{\tau\in\overline{\mathcal{T}}}, t; \theta^*)$$

for all $u_0 \in \mathcal{C}^2(\overline{\mathcal{D}} \times \overline{\mathcal{T}};\mathbb{R})$ and $t \in \mathcal{T}$.

The proof of the above theorem may be found in Appendix A.2. The parameters E_0, ν_0, α_0, and β_0 are determined via an appropriate decomposition of $\widehat{a}_0$ in (2.5); details are in the proof. In particular, E_0 and ν_0 are homogenized elasticity and viscosity coefficients, respectively, α_0 are decay rates for the hidden variables ξ, and β_0 are coefficients for each decay term. Note that the model in equations (3.4) is Markovian. Furthermore, although the model in (3.4) requires an input of t for evaluation, the spatial variable x only enters implicitly through the local values of $\partial_x u_0$ and $\partial_{xt}^2 u_0$; the model acts pointwise in space. Thus we have not included x explicitly in the theorem statement, for economy of notation. In what follows it is useful to define u_0^{PC} to be the solution to the following system defined with constitutive model Ψ_0^{PC}:

$$\rho \partial_t^2 u_0^{PC} - \partial_x \sigma_0 = f, \quad (x,t) \in \mathcal{D} \times \mathcal{T}, \tag{3.6a}$$
$$\sigma_0(t) = \Psi_0^{PC}\left(\partial_x u_0^{PC}(t), \partial_{xt}^2 u_0^{PC}(t), \left\{\partial_x u_0^{PC}(\tau)\right\}_{\tau\in\overline{\mathcal{T}}}, t\right), \quad (x,t) \in \mathcal{D} \times \mathcal{T}, \tag{3.6b}$$
$$u_0^{PC}|_{t=0} = u^*, \quad \partial_t u_0^{PC}|_{t=0} = v^*, \quad (x,t) \in \mathcal{D} \times \{0\}, \tag{3.6c}$$
$$u_0^{PC} = 0, \quad (x,t) \in \partial\mathcal{D} \times \mathcal{T}. \tag{3.6d}$$

Using a homogenization theorem, together with approximation by piecewise-constant material properties, we now show that u_ϵ can be approximated by u_0^{PC}; this will follow from the inequality

$$\|u_\epsilon - u_0^{PC}\|_{\mathcal{Z}_2} \le \|u_\epsilon - u_\epsilon^{PC}\|_{\mathcal{Z}_2} + \|u_\epsilon^{PC} - u_0^{PC}\|_{\mathcal{Z}_2}.$$

The first term on the right-hand side may be controlled using Theorem 3.4. The fact that dynamics under constitutive law $\Psi_\epsilon^\dagger$ converge to those under $\Psi_0^\dagger$ as $\epsilon \to 0$ may be used to control the second term; this fact is a consequence of the following theorem.

THEOREM 3.7. *Under Assumptions* 3.1*, the solution* u_ϵ *to equations* (2.1) *converges weakly to* u_0*, the solution to equations* (2.8) *with* $\rho = 0$*, in* $W^{1,2}(\mathcal{T}; V)$*. Thus, for any* $\eta > 0$ *there exists* $\epsilon_{\text{crit}} > 0$ *such that for all* $\epsilon \in (0, \epsilon_{\text{crit}})$*,*

$$\|u_\epsilon - u_0\|_{\mathcal{Z}_2} < \eta. \tag{3.7}$$

Proof. Since $f \in \mathcal{Z}$, continuous embedding gives $f \in \mathcal{Z}_2$. Applying Theorem 3.1 of [11] (noting that the work in that paper is set in dimension $d = 3$, but is readily extended to dimension $d = 1$) establishes weak convergence of u_ϵ to u_0 in $W^{1,2}(\mathcal{T}, V)$. Hence strong convergence in $\mathcal{Z}_2$ follows by compact embedding of $W^{1,2}(\mathcal{T}; V)$ into $\mathcal{Z}_2$. ∎

The following corollary is a consequence of Theorem 3.7.

COROLLARY 3.8. *Under Assumptions* 3.1 *and assuming* E, ν *are piecewise constant, the solution* u_ϵ^{PC} *to equations* (2.1) *converges weakly to* u_0^{PC}*, the solution to equations* (3.6) *with* $\rho = 0$*, in* $W^{1,2}(\mathcal{T}; V)$*. Thus, for any* $\eta > 0$ *there exists* $\epsilon_{\text{crit}} > 0$ *such that for all* $\epsilon \in (0, \epsilon_{\text{crit}})$*,*

$$\|u_\epsilon^{PC} - u_0^{PC}\|_{\mathcal{Z}_2} < \eta. \tag{3.8}$$

Combining this result with that of Theorem 3.4, noting continuous embedding of $\mathcal{Z}$ into $\mathcal{Z}_2$, allows us to approximate u_ϵ by u_0^{PC}:

COROLLARY 3.9. *Let* E *and* ν *be piecewise-continuous functions, with a finite number of discontinuities satisfying, along with* f*, Assumptions* 3.1*; let* u_ϵ *be the corresponding solution to* (2.1)*. Then for any* $\eta > 0$*, there exists* L_{crit} *and* ϵ_{crit} *with the property that for all* $L \geq L_{crit}$ *there are* L*-piecewise-constant* E^{PC} *and* ν^{PC} *such that for all* $\epsilon \in (0, \epsilon_{crit})$*, the solution to* u_0^{PC} *to* (3.6) *with* $\rho = 0$ *satisfies*

$$\|u_\epsilon - u_0^{PC}\|_{\mathcal{Z}_2} < \eta. \tag{3.9}$$

3.3. Neural network approximation of constitutive model. For the specific KV model in dimension $d = 1$ we know the postulated form of Ψ_0^{PC} and can in principle use this directly as a constitutive model. However, in more complex problems we do not know the constitutive model analytically, and it is then desirable to learn it from data from within an expressive model class. To this end we demonstrate that Ψ_0^{PC} can be approximated by an operator Ψ_0^{RNN} which has a similar form to that defined by equations (3.4) but in which the right-hand sides of those equations are represented by neural networks, leading to an RNN structure. Note that with this structure, the neural network outputs the stress at a single point in space and time; in practice, repeated evaluation generates output stress trajectories from the spatio-temporal dynamics. As such, the architecture is not the same as standard long short-term memory (LSTM) RNN models. Instead, the feed-forward neural network $\mathcal{G}$ produces time derivatives of the hidden variables ξ, which are used in a forward Euler step to generate the updated hidden variable value. In this manner, the model acts pointwise but incorporates memory through a hidden variable. Since the model acts pointwise, the feed-forward networks $\mathcal{F}$ and $\mathcal{G}$ are the same at every time-step, justifying the "recurrent" terminology.

Recall the definitions of (E', ν', L') and θ^* in Theorem 3.6. We first define the linear functions $\mathcal{F}^{PC} : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{L'} \to \mathbb{R}$ and $\mathcal{G}^{PC} : \mathbb{R}^{L'} \times \mathbb{R} \to \mathbb{R}$ by

$$\mathcal{F}^{PC}(b, c, r) = E'b + \nu' c - \langle \mathbb{1}, r \rangle, \tag{3.10a}$$

$$\mathcal{G}^{PC}(r, b) = -Ar + \beta b, \tag{3.10b}$$

where $A = \operatorname{diag}(\alpha_\ell) \in \mathbb{R}^{L' \times L'}$ and $\beta = \{\beta_1, \dots, \beta_\ell\} \in \mathbb{R}^{L'}$. We then have

$$\Psi_0^{PC}(\partial_x u_0(t), \partial_{xt}^2 u_0(t), \{\partial_x u_0(\tau)\}_{\tau \in \overline{\mathcal{T}}}, t; \theta^*) = \mathcal{F}^{PC}\left(\partial_x u_0(t), \partial_{xt}^2 u_0(t), \xi(t)\right), \tag{3.11a}$$

$$\dot{\xi}(t) = \mathcal{G}^{PC}(\xi(t), \partial_x u_0(t)), \quad \xi(0) = 0 \tag{3.11b}$$

as in Theorem 3.6.

We seek to approximate this map by Ψ_0^{RNN} defined by replacing the linear functions $\mathcal{F}^{PC}$ and $\mathcal{G}^{PC}$ by neural networks $\mathcal{F}^{RNN} : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{L'} \to \mathbb{R}$ and $\mathcal{G}^{RNN} : \mathbb{R}^{L'} \times \mathbb{R} \to \mathbb{R}$ to obtain

$$\Psi_0^{RNN}(\partial_x u_0(t), \partial_{xt}^2 u_0(t), \{\partial_x u_0(\tau)\}_{\tau \in \overline{\mathcal{T}}}, t) = \mathcal{F}^{RNN}\left(\partial_x u_0(t), \partial_{xt}^2 u_0(t), \xi(t)\right), \tag{3.12a}$$

$$\dot{\xi}(t) = \mathcal{G}^{RNN}(\xi(t), \partial_x u_0(t)), \quad \xi(0) = 0. \tag{3.12b}$$

Let $R > 0$ and define the bounded set $\mathsf{Z}_R = \{w : \mathbb{R}^+ \to \mathbb{R} \mid \sup_{t \in \mathcal{T}} |w(t)| \le R\}$.

THEOREM 3.10 (RNN approximation). *Consider Ψ_0^{PC} defined as by* (3.10) *and* (3.11). *Assume that there exist $\rho > 0$ and $0 \le B < \infty$ such that $\rho < \min_\ell |\alpha_\ell|$ and $\max_\ell |\beta_\ell| \le B$. Then, under Assumptions* 3.1*, for every $\eta > 0$ there exists Ψ_0^{RNN} of the form* (3.12) *such that*

$$\sup_{t \in \mathcal{T}, b, c \in \mathsf{Z}_R} \left| \Psi_0^{PC}\left(b(t), c(t), \{b(\tau)\}_{\tau \in \overline{\mathcal{T}}}, t; \theta^*\right) - \Psi_0^{RNN}\left(b(t), c(t), \{b(\tau)\}_{\tau \in \overline{\mathcal{T}}}, t\right) \right| < \eta.$$

The proof of Theorem 3.10 can be found in Appendix A.3.

Note that Ψ_0^{RNN} both avoids dependence on the fine-scale ϵ and is Markovian. The nonhomogenized map $\Psi_\epsilon^\dagger$ is local in time, while the homogenized map Ψ_0^{RNN} is nonlocal in time and depends on the strain history. Let u_0^{RNN} be the solution to the following system with constitutive model Ψ_0^{RNN}:

$$\rho \partial_t^2 u_0^{RNN} - \partial_x \sigma_0 = f, \quad (x, t) \in \mathcal{D} \times \mathcal{T}, \tag{3.13a}$$

$$\sigma_0(t) = \Psi_0^{RNN}\left(\partial_x u_0^{RNN}(t), \partial_{xt}^2 u_0^{RNN}(t), \left\{\partial_x u_0^{RNN}(\tau)\right\}_{\tau \in \overline{\mathcal{T}}}, t\right), \quad (x, t) \in \mathcal{D} \times \mathcal{T}, \tag{3.13b}$$

$$u_0^{RNN}|_{t=0} = u^*, \quad \partial_t u_0^{RNN}|_{t=0} = v^*, \quad (x, t) \in \mathcal{D} \times \{0\}, \tag{3.13c}$$

$$u_0^{RNN} = 0, \quad (x, t) \in \partial\mathcal{D} \times \mathcal{T}. \tag{3.13d}$$

Ideally we would like an approximation result bounding $\|u_\epsilon - u_0^{RNN}\|_{\mathcal{Z}_2}$, the difference between solution of the multiscale problem (2.1) and the Markovian RNN model (3.13), in the case $\rho = 0$. Using Corollary 3.9 shows that this would follow from a bound on $\|u_0^{PC} - u_0^{RNN}\|_{\mathcal{Z}}$, where u_0^{PC} solves (3.6), in the case $\rho = 0$. We note,

however, that although Theorem 3.10 gives us an approximation result between Ψ_0^{PC} and Ψ_0^{RNN}, proving that u_0^{PC} and u_0^{RNN} are close requires developing a new theory for the fully nonlinear PDE for u_0^{RNN}; developing such a theory is beyond the scope of this work and is difficult for two primary reasons: (i) the monotonicity property of Ψ_0^{RNN} with respect to strain rate is hard to establish globally for a trained model; (ii) the functions $\mathcal{F}^{RNN}, \mathcal{G}^{RNN}$ may not be differentiable. As a result, existence and uniqueness of u_0^{RNN} remain unproven; however, numerical experiments in section 4 indicate that in practice, u_0^{RNN} does approximate u_ϵ well.

Remark 3.11.

- Monotonicity of Ψ_0^{RNN} with respect to strain rate is a particular issue when $\rho = 0$ (no inertia), as in this case it is needed to define an (implicit) equation for $\partial_t u_0$ to determine the dynamics. It is for this reason that our experiments will all be conducted with $\rho > 0$, obviating the need for the determination of an (implicit) equation for $\partial_t u_0$. However, this leads to the issue that the homogenized equation is only valid for a subset of initial conditions, in the inertial setting $\rho > 0$; see Remark 2.1.
- In practice we will train Ψ_0^{RNN} using data for the inputs b, c which are obtained from a random function, or set of realizations of random functions. The choice of the probability measure from which this training data is drawn will affect the performance of the learned model when Ψ_0^{RNN} is evaluated at $b = \partial_x u_0^{RNN}$ and $c = \partial_{xt}^2 u_0^{RNN}$, for displacement u generated by the model given by (3.13).

3.4. RNN optimization. In the following section, we present numerical results using a trained RNN as a *surrogate model*: an efficient approximation of the complex microscale dynamics; in this section we discuss the problem of finding such an RNN. To learn the RNN operator approximation, we are given data

$$\left\{ \left(\partial_x u_0\right)_n, \left(\partial_{xt}^2 u_0\right)_n, \left(\sigma_0\right)_n \right\}_{n=1}^N, \tag{3.14}$$

where the suffix n denotes the nth strain, strain rate, and stress trajectories over the entire time interval $\mathcal{T}$. Each strain trajectory $(\partial_x u_0)_n$ is drawn i.i.d. from a measure μ on $\mathcal{C}(\mathcal{T}; \mathbb{R})$. There is no need to generate training data on the same time interval $\mathcal{T}$ as the macroscale model; we do so for simplicity. Note that since the RNN acts pointwise, for every nth set of trajectories, the RNN is evaluated T times, where $T = \frac{\mathcal{T}}{dt}$ for time discretization dt.

The data for the homogenized constitutive model is given by

$$\sigma_0(t) = \Psi_0^\dagger(\partial_x u_0(t), \partial_{xt}^2 u_0(t), \{\partial_x u_0(\tau)\}_{\tau \in \overline{\mathcal{T}}}, t),$$

defined via solution of the cell problem (2.4); but it may also be obtained as the solution to a forced boundary problem on the microscale, as stated in the following lemma.

LEMMA 3.12. *Let u solve the equations*

$$\partial_y \sigma(y,t) = 0, \quad (y,t) \in \mathcal{D} \times \mathcal{T}, \tag{3.15a}$$

$$\sigma(y,t) = E(y)\partial_y u(y,t) + \nu(y)\partial_{ty}^2 u(y,t), \quad (y,t) \in \mathcal{D} \times \mathcal{T}, \tag{3.15b}$$

$$u(0,t) = 0, \quad u(1,t) = b(t), \quad (y,t) \in \partial\mathcal{D} \times \mathcal{T}, \tag{3.15c}$$

$$u(y,0) = 0, \quad y \in \mathcal{D}. \tag{3.15d}$$

Then

$$\{\sigma(t)\}_{t=0}^{T} = \Psi_0^\dagger \left(b(t), \partial_t b(t), \{b(t)\}_{t=0}^{T}, t \right),$$

where $\Psi_0^\dagger$ *is the map defined in* (2.6).

The proof can be found in Appendix B.2 and justifies the application of data resulting from this problem to the homogenized model. In the following, we denote by $(\widehat{\sigma}_0)_n$ and $\widehat{\dot{\xi}}_n$ the output of $\mathcal{F}^{RNN}$ and $\mathcal{G}^{RNN}$ on data point n:

$$(\widehat{\sigma}_0)_n(t) = \mathcal{F}^{RNN}\left((\partial_x u_0)_n(t), (\partial_{xt}^2 u_0)_n(t), \widehat{\xi}_n(t) \right),$$
$$\widehat{\dot{\xi}}_n(t) = \mathcal{G}^{RNN}\left((\partial_x u_0)_n(t), \widehat{\xi}_n(t) \right), \quad \widehat{\xi}_n(0) = 0.$$

To train the RNN, we use the following relative L^2 loss function, which should be viewed as a function of the parameters defining the neural networks $\mathcal{F}^{RNN}, \mathcal{G}^{RNN}$.

Accessible Loss Function:

$$L_1(\{\sigma_0\}_{n=1}^N, \{\widehat{\sigma}_0\}_{n=1}^N) = \frac{1}{N}\sum_{n=1}^{N} \frac{\|(\sigma_0)_n - (\widehat{\sigma}_0)_n\|_{L^2(\mathcal{T};\mathbb{R})}}{\|(\sigma_0)_n\|_{L^2(\mathcal{T};\mathbb{R})}}. \tag{3.16}$$

Remark 3.13. To test robustness of our conclusions, we also employed relative and absolute L^2 squared loss functions. In doing so we did not observe significant differences in the predictive accuracy of the resulting models.

In the case of a material that is 2-piecewise-constant on the microscale, we can explicitly write down the analytic form of the solution, and thus can also know the values of the hidden variable $\{\xi_n\}_{n=1}^N$ and its derivative $\{\dot{\xi}_n\}_{n=1}^N$, for each data trajectory as expressed in equation (3.11). It is intuitive that training an RNN on an extended data set which includes the hidden variable should be easier than using the original data set (3.14). In order to deepen our understanding of the training process we will include training in a 2-piecewise-constant which uses this hidden data, motivating the following loss function. Since, in general, the hidden variable is inaccessible in the data, we refer to the resulting loss as the inaccessible relative loss function.

Inaccessible Loss Function:

$$L_2(\{(\sigma_0)_n\}_{n=1}^N, \{(\widehat{\sigma}_0)_n\}_{n=1}^N, \{\dot{\xi}_n\}_{n=1}^N, \{\widehat{\dot{\xi}}_n\}_{n=1}^N) \tag{3.17a}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \left(\frac{\|(\sigma_0)_n - (\widehat{\sigma}_0)_n\|_{L^2(\mathcal{T};\mathbb{R})}}{\|(\sigma_0)_n\|_{L^2(\mathcal{T};\mathbb{R})}} + \frac{\|\dot{\xi}_n - \widehat{\dot{\xi}}_n\|_{L^2(\mathcal{T};\mathbb{R})}}{\|\dot{\xi}_n\|_{L^2(\mathcal{T};\mathbb{R})}} \right). \tag{3.17b}$$

This inaccessible loss function helps identify the RNN whose existence in proved in previous sections.

4. Numerical results. The numerical results reach the following conclusions, all of which guide the use of machine-learned constitutive models within complex nonlinear problems beyond the one-dimensional linear viscoelastic setting considered in this work:

I. **Machine-learned constitutive models.** We can find RNNs that yield low-error simulations when used as a surrogate model in the macroscopic

system (1.2), (1.3) to approximate the multiscale system (1.1) in the one-dimensional KV setting with inertia. We also discuss how in some material parameter settings, inertial effects lead to higher error in the homogenized approximation.

II. **Choice of training data.** We describe our choice of data sampling distribution μ and show that it exhibits desirable properties.

III. **Effect of nonconvex optimization.** When the inaccessible loss function is used for training, the trained RNN exhibits desirable properties of (approximate) linearity in its arguments in the domain of interest, as is proved for the homogenized constitutive model (3.10) for piecewise-constant materials. When using the accessible loss function, the trained RNN may perform well as a surrogate model without exhibiting linearity in the equation for evolution of the hidden variables. This is attributable to the existence of local minimizers of the loss function and highlights the need for caution in training constitutive models.

IV. **Model choice.** The correct choice of architecture for the RNN leads to discretization-robustness in time: a model learned with one choice of time discretization dt performs well when tested on another dt; this is not true for poor model choices. Discretization-robustness can thus be used as a guide to model choice.

In subsection 4.1 we demonstrate that in appropriate settings the solution u_0^{RNN} obtained under the dynamics of a trained RNN approximates the true solution u_ϵ well when used in the macroscopic setting; furthermore, this RNN is shown to exhibit linearity in its arguments in the domain of interest. We also discuss the error arising from inertial effects. In subsection 4.2 we discuss the performance of an RNN learned using the accessible loss function. In subsection 4.3 we discuss the discretization-robustness property of the RNN and the choice of data sampling distribution μ. In subsection 4.4, we perform experiments on media with more piecewise-constant pieces and analyze the number of hidden variables required to capture the behavior. In subsection 4.5, we extend the methodology to the case of elasto-viscoplasticity.

4.1. RNN as surrogate model. In this subsection we discuss two RNNs: RNN "A," trained using only the inaccessible loss function in (3.4), and RNN "B," trained with the standard loss function in (3.16), but initialized at parameters obtained via training with the inaccessible loss function. Descriptions of these RNNs, and others we introduce in subsequent subsections, may be found in Table 1. A visualization of typical input and output trajectories from the data used for training and testing may be found in Figure 1. For details on RNN training, see Appendix C.1.

In the first surrogate model experiment, we subject the material to sinusoidal boundary forcing of $b(t) = 0.1\sin(2\pi t)$ starting from 0 initial displacement and ve-

TABLE 1
RNN descriptions.

RNN	Description
A	Trained on 2-piecewise-constant media only with inaccessible loss function (3.4)
B	Trained on 2-piecewise-constant media; initialized at solution found with inaccessible loss function then trained with accessible loss function (3.16)
C	Trained on 2-piecewise-constant media only with accessible loss function
D	Trained on continuous media with accessible loss function

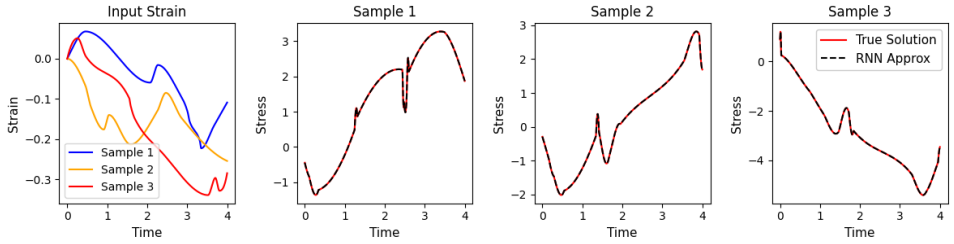
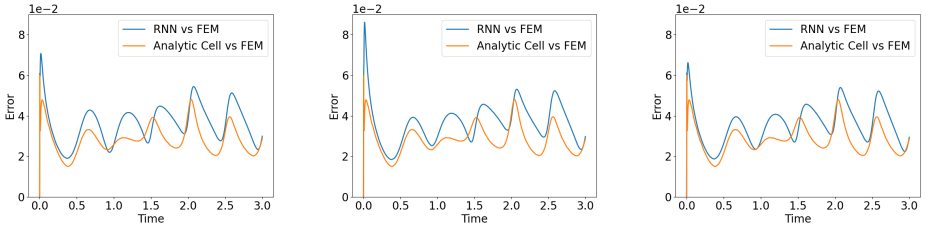


FIG. 1. *Representative data: input strain trajectories and output stress trajectories for three randomly chosen test data samples. The RNN approximation shown was generated with RNN "C."*



(a) RNN "A" (b) RNN "B" (c) RNN "C"

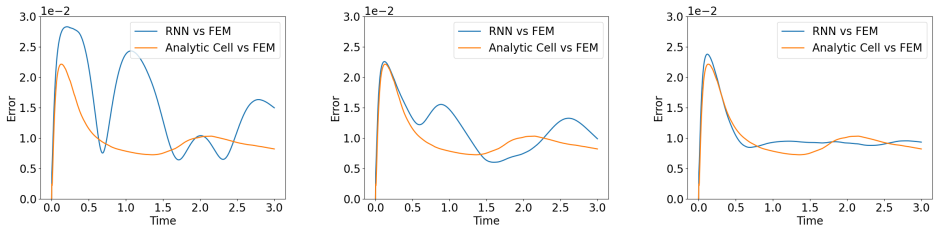
FIG. 2. *Analytic cell and RNN relative error versus FEM solution using sinusoidal forcing; this supports conclusion* I.

locity. As a ground-truth comparison, we use a traditional finite element solver with periodic domain of width 0.04, spatial resolution of $h = 0.005$, and time discretization $dt = 0.1h^2$; we refer to this solution as u_ϵ and name it FEM. In contrast, the RNN-based macroscale computation employs spatial resolution of $h_{cell} = 0.04$, with a time discretization of $dt = 0.4h_{cell}^2$; for economy of notation; this solution is denoted u_0^{RNN} and named RNN. We also compare the results to the displacement obtained using as macroscale constitutive model the analytic solution to the cell problem. To make comparisons we use the relative error given by

$$e(u_0^{RNN}, u_\epsilon)(t) = \frac{\|u_\epsilon(t) - u_0^{RNN}(t)\|_{L^2(\mathcal{D};\mathbb{R})}}{\|u_\epsilon(t)\|_{L^2(\mathcal{D};\mathbb{R})} + 0.01}. \tag{4.1}$$

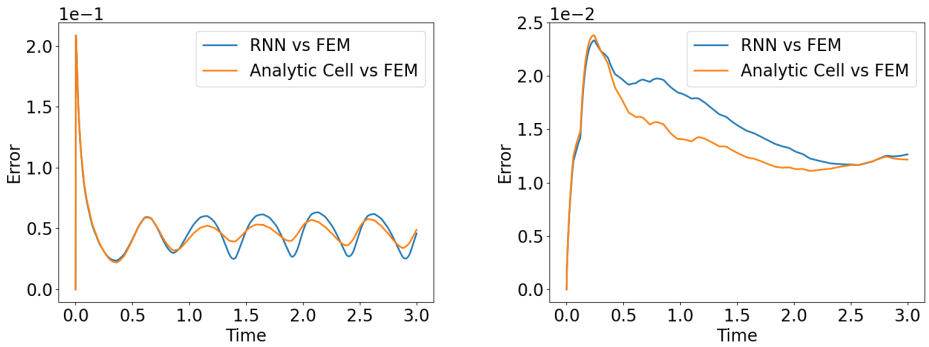
The relative error plots for RNNs "A" and "B" are shown in Figures 2a and 2b. In a second experiment, we subject the material to integrated Brownian motion forcing starting from null initial conditions. The FEM solver uses the same discretizations as in the sinusoidal forcing experiment, and the RNN spatial discretization was $h_{cell} = 0.05$ with time discretization of $dt = 0.4h_{cell}^2$. The results for RNNs "A" and "B" are shown in Figures 3a and 3b.

Both sets of experiments show that the RNN-based macroscopic models accurately reproduce the microscale FEM simulation at far lower computational cost. The RNN-based results have some errors in comparison with the microscale simulation, but the errors are of the same order of magnitude as the errors arising when the exact homogenized constitutive model is used. The initial error between the analytic solution and the FEM solution is due to inertial effects discussed in Remark 2.1. The inertial errors become more significant as the ratio between E and ν varyies more across the interval. Relative error results for an RNN trained on a material with more inertial effects is shown in Figure 4.



(a) RNN "A" (b) RNN "B" (c) RNN "C"

FIG. 3. *Analytic cell and RNN relative error versus FEM solution using integrated Brownian motion forcing; this supports conclusion* I.



(a) Sinusoidal forcing (b) Integrated Brownian motion forcing

FIG. 4. *Relative error of RNN trained on material parameters with higher inertial effects in response to sinusoidal and integrated Brownian motion forcing; this demonstrates conclusion* I.

Both RNNs also exhibit the desirable property of linearity in the inputs in an appropriate domain as discussed in subsection 3.4; this is presented in Figure 5.

4.2. RNN trained with standard loss. We also train a third RNN, denoted RNN "C," using only the accessible loss function. Details of training may be found in Appendix C.1. The performance of this RNN as a surrogate model in the macroscale simulation experiments may be seen in Figures 2c and 3c. While this RNN performs well as a surrogate model, and indeed is comparable in errors to those of RNNs "A" and "B," it does not exhibit a close linear match to the known analytic expression for $\mathcal{G}$, as shown in Figure 5. In the figure, all three RNNs approximate the linear structure of $\mathcal{F}$ well; the difficulty is in obtaining the correct linear dependence in the hidden variable rate, $\dot{\xi}$. Interestingly, by changing the material parameter ν_2 from 0.2 to 2, training via the method of RNN "C" with only the accessible loss function yields an RNN that matches the true linear dependence in $\mathcal{G}$ very well. However, in this parameter regime, inertial effects perturb the simulations on the macroscale to an unacceptable degree, meaning that the homogenization theory that we use as benchmark is not valid, and so we avoid this regime. The inability of RNN "C" to capture the exact linear dependence in $\mathcal{G}$ is unsurprising; indeed, had we guaranteed convergence to the optimal function for any choice of material parameters, we would have entirely sidestepped the problem of high-dimensional optimization inherent to machine learning. In the case of continuous material properties, we do not have a

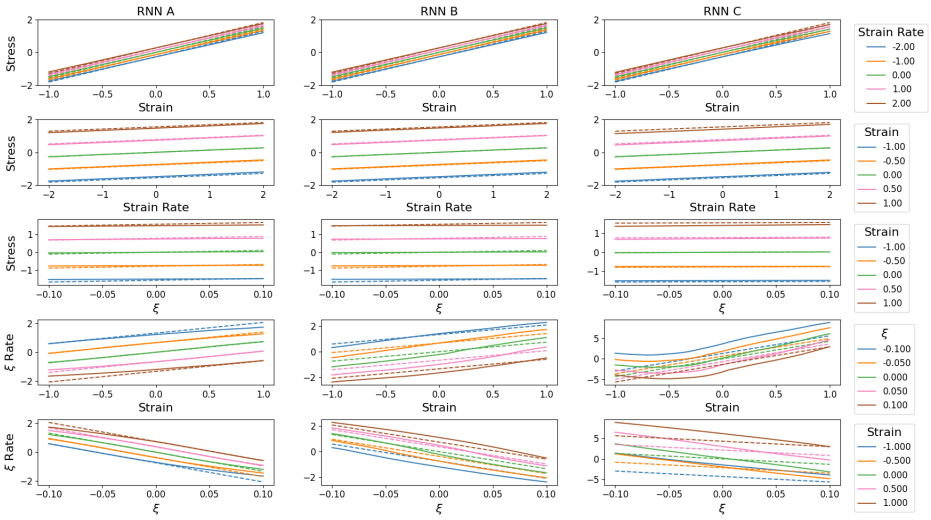
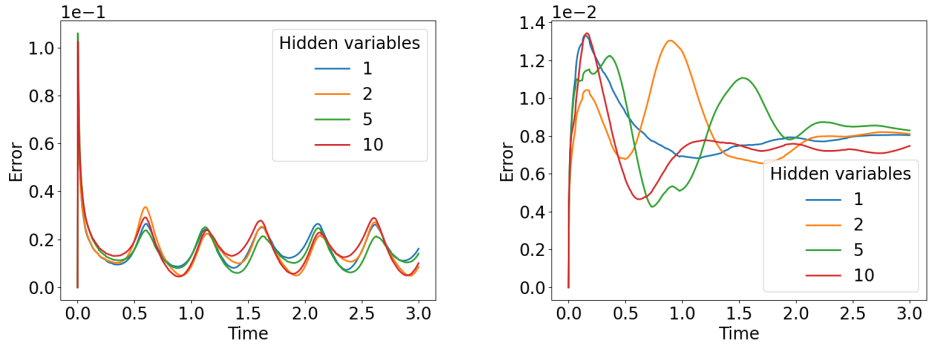


FIG. 5. *RNN outputs versus the truth (dashed) for each of the three candidate RNNs. The columns correspond to RNNS "A," "B," and "C," respectively. The first row shows the strain-stress dependence for five fixed strain rate inputs. The second row shows the strain rate-stress dependence for five fixed strain inputs. The third row shows the ξ, stress relationships for hidden variable ξ for five fixed strain inputs. The fourth row shows the strain, $\dot{\xi}$ relationship for five different fixed values of ξ. Finally, the fifth row shows the $\xi, \dot{\xi}$ relationship for five fixed strain inputs. This supports conclusion* III.

known analytic solution to the microscale problem and thus do not have access to the hidden variable ξ in the train and test data; in this case, we may only use the accessible loss function. We train RNNs, denoted RNN type "D," on continuous media with different numbers of hidden variables ξ_ℓ and use the trained RNNs as surrogate models in the macroscale system subjected to boundary forcing. Training details may be found in Appendix C.2. The relative error of RNN "D" for the sinusoidal and Brownian motion forcing experiments described previously is shown in Figure 6. We note that a similar error is found with all dimensions of the hidden variable, suggesting that 1 hidden variable suffices in this case; the fact that the error does not decrease suggests that the error we see is primarily from homogenization rather than piecewise-constant approximation. In subsection 4.4 we explore the choice of hidden variable count in greater depth.

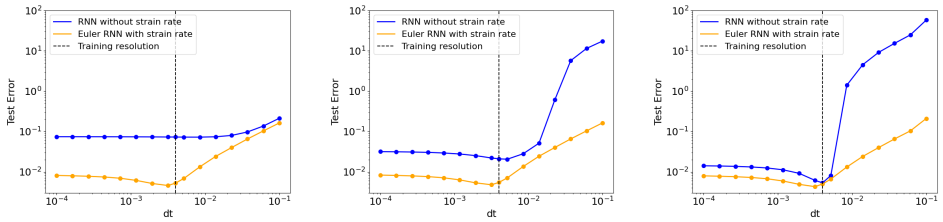
4.3. Time discretization and RNN training. Discretization-robustness is a desirable feature of an RNN surrogate model. To test robustness to changes in time discretization we work in the piecewise-constant media setup leading to RNNs "A," "B," and "C". We evaluate the test error when employing each of the three RNNs using values of time-step dt different from those used in the training. Additionally, to demonstrate the value of postulating the correct model form, we train three additional RNNs via the same methods as described in subsection 4.2 but without strain rate dependence. Figure 7 shows that all three RNNs trained with strain rate as an input parameter were more robust to changes in time discretization than their non-strain-rate counterparts.

To generate the training strain, we sampled trajectories as follows: first, we randomly partitioned the time interval $\mathcal{T}$ into 10 pieces; second, at each point between these time intervals, we generated a value of strain via a balanced random walk from



(a) Sinusoidal forcing (b) Integrated Brownian motion forcing

FIG. 6. *Relative error of continuous-material RNNs "D" with different numbers of hidden variables when used as a surrogate model in a the macroscale system; this supports conclusion* I.



(a) RNN "A" (b) RNN "B" (c) RNN "C"

FIG. 7. *Time discretization error for RNNs "A," "B," and "C." This supports conclusion* IV.

the previous value scaled by the length of the time interval; third, we used a piecewise cubic hermite interpolating polynomial (pchips) function to interpolate between these values of strain. This choice of distribution has the desirable property that it generates data with a variety of strain/strain-rate pairings evenly dispersed throughout the domain of interest rather than introducing large correlations between the two. A scatterplot of the associated values is shown in Figure 8.

4.4. Additional piecewise-constant experiments. We claim that to approximate an N-piecewise-constant material, the RNN ought to have at least $N-1$ hidden variables. Therefore, we train RNNs with different numbers of hidden variables on data from piecewise-constant materials with 3, 5, and 10 pieces. The results are shown in Figure 9. For the 3- and 5-piecewise-constant cases, the error flattens out after 2 and 4 hidden variables, respectively, as expected. For the 10-piecewise-constant case, the error flattens out first at 4 hidden variables and then again at 9 hidden variables, and this can be explained by examining the analytic solution. For the choices of E and ν used, the constitutive law takes the approximate form of

$$\begin{aligned}\sigma_0(t) = E'\partial_x u_0(t) + \nu'\partial_t\partial_x u_0(t) - \int_0^t \partial_x u_0(\tau)\Big(&0.09e^{-1.83(t-\tau)} + 0.16e^{-2.92(t-\tau)} \\ &+ 0.02e^{-4.44(t-\tau)} + 0.20e^{-5.13(t-\tau)} + 0.12e^{-8.18(t-\tau)} + 0.08e^{-9.29(t-\tau)} \\ &+ 0.39e^{-11.3(t-\tau)} + 0.67e^{-15.30(t-\tau)} + 0.06e^{-18.41(t-\tau)}\Big)\, d\tau.\end{aligned}$$

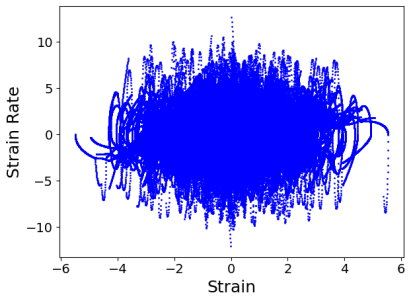


FIG. 8. *Strain and strain rate distributions in the training data generated by μ; this supports the choice of training data discussed in conclusion* II.

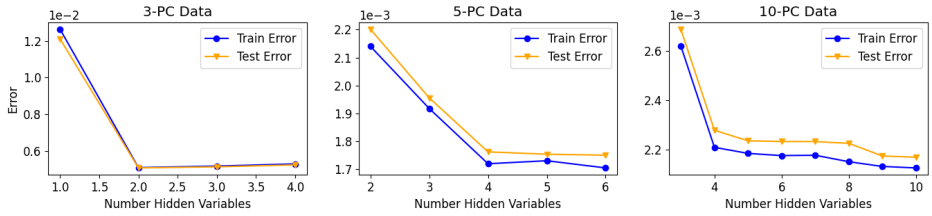


FIG. 9. *Absolute L^2 error of RNNs trained with different numbers of hidden variables on different piecewise-constant materials.*

Note that the exponential decay terms each correspond to one of the nine hidden variables ξ_ℓ, and they are written in order of decreasing exponential term $-\alpha_\ell$. From this we can see that terms with higher magnitudes of $|\alpha_\ell|$ will be negligible compared to the terms with smaller magnitude. The experimental results align with these values; there is a large jump from the fourth exponent (-5.13) to the fifth exponent (-8.18), so the behavior is well captured with only four hidden variables. However, with 9 hidden variables, the model can completely capture the decay terms. This result further justifies the practical use of the piecewise-constant approximation for smooth materials. Analyzing the homogenization decay rates as they relate to the spectrum of the differential operator is an interesting area of future work.

4.5. Elasto-viscoplasticity. The purpose of this example is to demonstrate that the ideas developed in this work have implications beyond linear viscoelasticity. The same RNN architecture is able to learn elastic-viscoplastic dynamics. We present the results of a simple experiment with isotropic rate hardening in one spatial dimension, noting that more detailed studies of the problem, in two and three dimensions, can be found in [21].

Consider the following equations:

$$\partial_y \sigma(y,t) = 0, \quad \sigma(y,t) = E(y)\left(\partial_y u(y,t) - \epsilon_p(y,t)\right), \qquad (y,t) \in \mathcal{D} \times \mathcal{T}, \tag{4.2a}$$

$$\dot{\epsilon}_p(y,t) = \dot{\epsilon}_{p0}\,\mathrm{sign}(\sigma(y,t))\left(\frac{|\sigma(y,t)|^n}{\sigma_h}\right), \qquad (y,t) \in \mathcal{D} \times \mathcal{T}, \tag{4.2b}$$

$$u(0,t) = 0, u(1,t) = b(t), \qquad (y,t) \in \partial\mathcal{D} \times \mathcal{T}, \tag{4.2c}$$

$$u(y,0) = 0, \epsilon_p(y,0) = 0, \qquad y \in \mathcal{D}, \tag{4.2d}$$

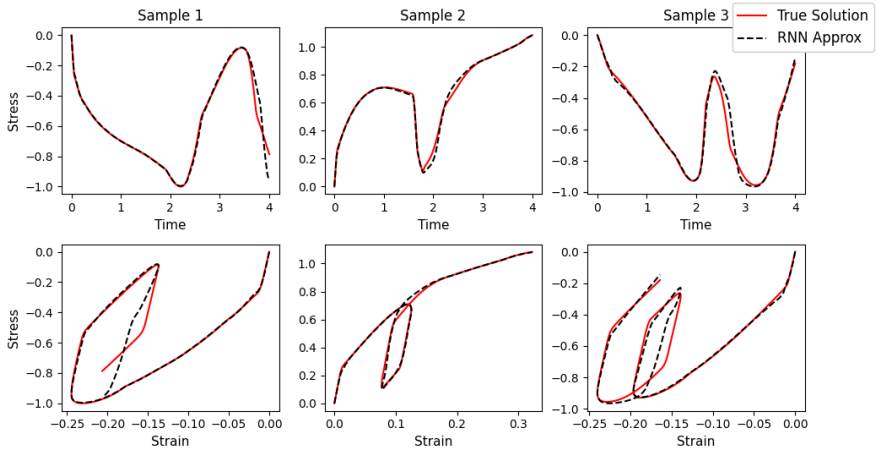


FIG. 10. *RNN trained on elasto-viscoplastic data; a comparison between the true solution and the RNN-predicted solution for three random test samples. Top row: stress trajectories in time. Bottom row: stress-strain trajectories for the same samples.*

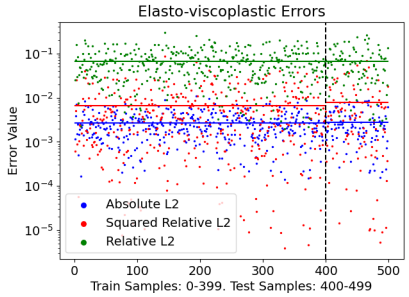


FIG. 11. *Error evaluations of all train and test data points for the elasto-viscoplastic experiments. Solid lines indicate mean error values, which are computed separately for the train and test sets.*

where u is the displacement, ϵ_p is the plastic strain, σ is the stress, and $\dot{\epsilon}_{p0}$, σ_h, and n are constants. We seek to learn the map

$$\{\sigma(t)\}_{t=0}^{T} = \Psi_0^{\dagger}\left(b(t), \{b(t)\}_{t=0}^{T}, t\right).$$

Note that the homogenized constitutive map does not depend on the rate as in the viscoelastic case. Development of the homogenization theory for which these equations arise as cell problems can be found in [21].

We train an RNN with one hidden variable and the same architecture as prescribed for the viscoelastic case but without strain rate dependence, and we use data generated via a numerical solution of equations (4.2). As shown in Figure 10, the RNN is able to learn plastic behavior. Specifically, the strain-stress trajectories exhibit plastic transition. Figure 11 gives a comprehensive picture of the error of this experiment. The mean relative L^2 error is $\approx 7\%$, which is reasonable for plastic experiments. Additionally, the performance on the test set is on par with that of the train set. Further experiments demonstrating discretization-invariance of the correct learning architecture may be found in [21].

5. Conclusions. In this paper, we develop theory underlying the learning of Markovian models for history-dependent constitutive laws. The theory presented applies to the one-dimensional KV case, but the underlying ideas extend to more complex systems, as demonstrated with an experiment with elasto-viscoplasticity. In [21], numerical experiments suggest that the methodology can be useful in higher dimensions as well. Conclusions drawn from our numerical experiments, underpinned by the theory of this paper and enumerated at the start of section 4, provide useful guidance for these more complex nonlinear models in higher spatial dimensions.

Several research directions are suggested by this work. First, when inertial effects are significant, the homogenization theory used in this paper, and underlying the computational work in [21], is not valid; extending the methodology to this setting would be useful. Second, the development of theoretical guidance and methodology for choice of training measure μ will be very important in this field. Third, as is the case with most machine learning applications, convergence of the RNN to the globally optimal solution is not guaranteed; considering learning techniques from reservoir computing could be useful to alleviate this issue as they lead to convex quadratic optimization problems.

Appendix A. Proofs.

A.1. Proof of Theorem 3.4 The proof of Lemma 3.5, which underlies the proof of Theorem 3.4, uses the following two propositions.

PROPOSITION A.1. *Under Assumptions* 3.1*, for all solutions* u *of* (2.2) *the following bounds hold for some constant* C_1*:*

1. $\sup_{t\in\mathcal{T}} \|u\|^2_{H^1_0,\nu} \le \|u|_{t=0}\|^2_{H^1_0,\nu} + (\frac{\nu^+}{E^-})^2 \frac{1}{\nu^-} C_1^2 \|f\|^2_{\mathcal{Z}}$.
2. $\sup_{t\in\mathcal{T}} \|u\|^2_{H^1_0,E} \le \frac{E_+}{\nu_-} \|u|_{t=0}\|^2_{H^1_0,\nu} + (\frac{\nu^+}{E^-})^2 \frac{E^+}{(\nu^-)^2} C_1^2 \|f\|^2_{\mathcal{Z}}$.
3. $\|\partial_t u\|_{H^1_0,\nu} \le \frac{C_1\|f\|_{\mathcal{Z}}}{\nu^-} + \frac{E^+}{\nu^-}\|u\|_{H^1_0,E}$ *for all* $t \in \mathcal{T}$.

Proof. To show the first bound, let $\varphi = u$ in (2.2). We have

$$q_\nu(\partial_t u, u) + q_E(u,u) = \langle f, u\rangle$$

so that

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|u\|^2_{H^1_0,\nu} + \|u\|^2_{H^1_0,E} &\le \|f\|_{H^{-1}}\|u\|_{H^1_0} \\ &\le C_1\|f\|\|u\|_{H^1_0}\end{aligned}$$

for some constant C_1, by compact embedding. Then, using Lemma 1.1,

$$\frac{1}{2}\frac{d}{dt}\|u\|^2_{H^1_0,\nu} + \frac{E^-}{\nu^+}\|u\|^2_{H^1_0,\nu} \le \frac{C_1}{2\delta^2}\|f\|^2 + \frac{\delta^2}{2\nu^-}\|u\|^2_{H^1_0,\nu}$$

for any $\delta > 0$ by Young's inequality. Letting $\delta^2 = \frac{E^-\nu^-}{\nu^+}$, we have

$$\frac{d}{dt}\|u\|^2_{H^1_0,\nu} + \frac{E^-}{\nu^+}\|u\|^2_{H^1_0,\nu} \le \frac{C_1^2\nu^+}{E^-\nu^-}\|f\|^2_{\mathcal{Z}}.$$

Finally, Gronwall's inequality yields

$$\sup_{t\in\mathcal{T}}\|u\|^2_{H^1_0,\nu} \le \|u|_{t=0}\|^2_{H^1_0,\nu} + \left(\frac{\nu^+}{E^-}\right)^2 \frac{C_1^2}{\nu^-}\|f\|^2.$$

The second bound follows from Lemma 1.1. For the third bound, let $\varphi = \partial_t u$ in (2.2). Then

$$q_\nu(\partial_t u, \partial_t u) + q_E(u, \partial_t u) = \langle f, \partial_t u \rangle$$

so that, again using Lemma 1.1, and using the Poincaré inequality,

$$\|\partial_t u\|^2_{H^1_0,\nu} \le \frac{C_1}{\nu^-}\|f\|\|\partial_t u\|_{H^1_0,\nu} + \frac{E^+}{\nu^-}\|u\|_{H^1_0,E}\|\partial_t u\|_{H^1_0,\nu}$$

and

$$\|\partial_t u\|_{H^1_0,\nu} \le \frac{C_1}{\nu^-}\|f\|_{\mathcal{Z}} + \frac{E^+}{\nu^-}\|u\|_{H^1_0,E}. \qquad \square$$

Additionally, we need to bound the difference between two solutions u_1 and u_2 of the PDE in Lemma 3.5 with different material properties. Notice that u_1 and u_2 satisfy

$$\frac{\partial}{\partial x}\left(E_1\left(\frac{\partial}{\partial x}u_1\right) + \nu_1\left(\frac{\partial^2}{\partial t \partial x}u_1\right)\right) = -f,$$
$$\frac{\partial}{\partial x}\left(E_1\left(\frac{\partial}{\partial x}u_2\right) + \nu_1\left(\frac{\partial^2}{\partial t \partial x}u_2\right)\right) = -f + \frac{\partial}{\partial x}\left[(E_1 - E_2)\frac{\partial}{\partial x}u_2 + (\nu_1 - \nu_2)\frac{\partial^2}{\partial t \partial x}u_2\right].$$

Subtracting yields

$$\partial_x\left[E_1\partial_x\gamma + \nu_1\partial^2_{t,x}\gamma\right] = -\partial_x\left[(\Delta E)\partial_x u_2 + (\Delta\nu)\partial^2_{t,x}u\right],$$

where $\gamma = u_1 - u_2$, $\Delta E = E_1 - E_2$, and $\Delta\nu = \nu_1 - \nu_2$. We can rewrite this as an equation for γ in weak form: for all test functions $\varphi \in V$

$$q_{\nu_1}(\partial_t\gamma, \varphi) + q_{E_1}(\gamma, \varphi) = \langle g, \partial_x\varphi\rangle, \quad \gamma|_{t=0} = 0, \tag{A.1}$$

where $g = \Delta E\partial_x u_2 + \Delta\nu\partial^2_{t,x}u_2$. For the following discussion of bounds including both u_1 and u_2, let $E^+ = \max\{E_1^+, E_2^+\}$, $\nu^+ = \max\{\nu_1^+, \nu_2^+\}$, $E^- = \min\{E_1^-, E_2^-\}$, and $\nu^- = \min\{\nu_1^-, \nu_2^-\}$.

PROPOSITION A.2. *Under Assumptions* 3.1*, for all solutions* γ *of* (A.1)*, the following bounds hold:*

1. $\sup_{t\in\mathcal{T}}\|\gamma\|^2_{H^1_0,\nu_1} \le (\frac{\nu^+}{E^-})^2\frac{1}{\nu^-}\|g\|^2_{\mathcal{Z}}$.
2. $\sup_{t\in\mathcal{T}}\|\gamma\|^2_{H^1_0,E_1} \le (\frac{\nu^+}{E^-})^2\frac{E^+}{(\nu^-)^2}\|g\|^2_{\mathcal{Z}}$.
3. $\sup_{t\in\mathcal{T}}\|\partial_t\gamma\|_{H^1_0,\nu_1} \le \frac{\|g\|_{\mathcal{Z}}}{\nu^-} + \frac{E^+}{\nu^-}\|\gamma\|_{H^1_0,E}$.

Proof. To show the first bound, let $\varphi = \gamma$ in (A.1). We have

$$q_\nu(\partial_t\gamma, \gamma) + q_E(\gamma, \gamma) = \langle g, \partial_x\gamma\rangle$$

so that

$$\frac{1}{2}\frac{d}{dt}\|\gamma\|^2_{H^1_0,\nu} + \|u\|^2_{H^1_0,E} \le \|g\|\|\gamma\|_{H^1_0}.$$

Then

$$\frac{1}{2}\frac{d}{dt}\|\gamma\|^2_{H^1_0,\nu} + \frac{E^-}{\nu^+}\|\gamma\|^2_{H^1_0,\nu} \le \frac{1}{2\delta^2}\|g\|^2 + \frac{\delta^2}{2\nu^-}\|\gamma\|^2_{H^1_0,\nu}$$

for any $\delta > 0$ by Young's inequality. Letting $\delta^2 = \frac{E^-\nu^-}{\nu^+}$, we have

$$\frac{d}{dt}\|\gamma\|^2_{H^1_0,\nu} + \frac{E^-}{\nu^+}\|\gamma\|^2_{H^1_0,\nu} \le \frac{\nu^+}{E^-\nu^-}\|g\|^2_{\mathcal{Z}}.$$

Finally, since $\gamma(0) = 0$, Gronwall's inequality yields

$$\sup_{t\in\mathcal{T}}\|\gamma\|^2_{H^1_0,\nu} \le \left(\frac{\nu^+}{E^-}\right)^2 \frac{1}{\nu^-}\|g\|^2.$$

The second bound follows from Lemma 1.1. For the third bound, let $\varphi = \partial_t\gamma$ in (A.1). Then

$$q_\nu(\partial_t\gamma, \partial_t\gamma) + q_E(\gamma, \partial_t\gamma) = \langle g, \partial^2_{xt}\gamma\rangle$$

so that, again using Lemma 1.1,

$$\|\partial_t\gamma\|^2_{H^1_0,\nu} \le \frac{1}{\nu^-}\|g\|\|\partial_t\gamma\|_{H^1_0,\nu} + \frac{E^+}{\nu^-}\|\gamma\|_{H^1_0,E}\|\partial_t\gamma\|_{H^1_0,\nu}$$

and

$$\|\partial_t\gamma\|_{H^1_0,\nu} \le \frac{1}{\nu^-}\|g\|_{\mathcal{Z}} + \frac{E^+}{\nu^-}\|\gamma\|_{H^1_0,E}. \qquad \square$$

To prove the Lipschitz property of the solution in Theorem 3.4, we will need the following lemma.

LEMMA 3.5 (Lipschitz solution). *Let u_i be the solution to*

$$-\partial_x\left(E_i\partial_x u_i + \nu_i\partial^2_{xt}u_i\right) = f, \quad (x,t)\in\partial\mathcal{D}\times\mathcal{T}, \tag{3.1}$$

$$u_i(x,t) = u^*, \quad (x,t)\in\mathcal{D}\times\{0\}, \tag{3.2}$$

$$u_i(x,t) = 0, \quad (x,t)\in\partial\mathcal{D}\times\mathcal{T}, \tag{3.3}$$

associated with material properties E_i, ν_i for $i\in\{1,2\}$, and forcing f, all satisfying Assumptions 3.1. *Then*

$$\|u_1 - u_2\|_{\mathcal{Z}} \le C\left(\|\nu_1 - \nu_2\|_\infty + \|E_1 - E_2\|_\infty\right)$$

for some constant $C\in\mathbb{R}^+$ dependent on $f, E_i^+, E_i^-, \nu_i^+, \nu_i^-$, and L and independent of ϵ.

Proof. Let γ and g be as defined before and after (A.1). Then, by the result of Proposition A.2,

$$\sup_{t\in\mathcal{T}}\|\gamma\|^2_{H^1_0} \le \frac{1}{\nu^-}\sup_{t\in\mathcal{T}}\|\gamma\|^2_{H^1_0,\nu_1} \le \left(\frac{\nu^+}{E^-\nu^-}\right)^2\|g\|^2_{\mathcal{Z}}.$$

To bound the right-hand side,

$$\begin{aligned}
\|g\|_{\mathcal{Z}} &= \|(\Delta E)\partial_x u_2 + (\Delta\nu)\partial^2_{t,x}u_2\|_{\mathcal{Z}}\\
&\le \|(\Delta E)\partial_x u_2\|_{\mathcal{Z}} + \|(\Delta\nu)\partial^2_{t,x}u_2\|_{\mathcal{Z}}\\
&\le \|\Delta E\|_\infty\|\partial_x u_2\|_{\mathcal{Z}} + \|\Delta\nu\|_\infty\|\partial^2_{t,x}u_2\|_{\mathcal{Z}}\\
&\le \sup_{t\in\mathcal{T}}\|u_2\|_{H^1_0}\|\Delta E\|_\infty + \sup_{t\in\mathcal{T}}\|\partial_t u_2\|_{H^1_0}\|\Delta\nu\|_\infty\\
&\le \frac{1}{(\nu^-)^{\frac{1}{2}}}\Big(\sup_{t\in\mathcal{T}}\|u_2\|_{H^1_0,\nu_2}\|\Delta E\|_\infty + \sup_{t\in\mathcal{T}}\|\partial_t u_2\|_{H^1_0,\nu_2}\|\Delta\nu\|_\infty\Big).
\end{aligned}$$

To bound $\|\partial_x u_2\|_{\mathcal{Z}}$ and $\|\partial^2_{t,x}u_2\|$, note that any solution u_2 will satisfy (2.2) for $(u,E,\nu)\mapsto(u_2,E_2,\nu_2)$. The analysis of Proposition A.1 yields

$$\sup_{t\in\mathcal{T}}\|u_2\|_{H_0^1,\nu_2} \le \|u|_{t=0}\|_{H_0^1,\nu} + \left(\frac{\nu^+}{E^-}\right)\frac{C_1}{(\nu^-)^{1/2}}\|f\|_{\mathcal{Z}}$$

and

$$\begin{aligned}\sup_{t\in\mathcal{T}}\|\partial_t u_2\|_{H_0^1,\nu_2} &\le \frac{C_1\|f\|_{\mathcal{Z}}}{\nu^-} + \frac{E^+}{\nu^-}\|u_2\|_{H_0^1,E}\\ &\le \frac{C_1}{\nu^-}\|f\|_{\mathcal{Z}} + \left(\frac{E^+}{\nu^-}\right)^{3/2}\|u|_{t=0}\|_{H_0^1,\nu} + \frac{(E^+)^{3/2}}{(\nu^-)^2}\frac{\nu^+}{E^-}C_1\|f\|_{\mathcal{Z}}.\end{aligned}$$

By the Poincaré inequality, $\|\gamma\|_{\mathcal{Z}} \le C_p \sup_{t\in\mathcal{T}}\|\gamma\|_{H_0^1}$ for some constant C_p and setting

$$\begin{aligned}C = C_p\left(\frac{\nu^+}{E^-(\nu^-)^{\frac{3}{2}}}\right)\max\Bigg\{&\|u|_{t=0}\|_{H_0^1,\nu} + \left(\frac{\nu^+}{E^-}\right)\frac{C_1}{(\nu^-)^{1/2}}\|f\|_{\mathcal{Z}},\\ &\frac{C_1}{\nu^-}\|f\|_{\mathcal{Z}} + \left(\frac{E^+}{\nu^-}\right)^{3/2}\|u|_{t=0}\|_{H_0^1,\nu} + \frac{(E^+)^{3/2}}{(\nu^-)^2}\frac{\nu^+}{E^-}C_1\|f\|_{\mathcal{Z}}\Bigg\}\end{aligned}$$

gives the result. ∎

Now we can prove the piecewise-constant approximation theorem.

THEOREM 3.4 (piecewise-constant approximation). *Let E and ν be piecewise-continuous functions, with a finite number of discontinuities, satisfying Assumptions* 3.1*; let u_ϵ be the corresponding solution to* (2.1)*. Then, for any $\delta > 0$, there exist piecewise-constant E^{PC} and ν^{PC} (in the sense of Definition* 3.2*) such that solution u_ϵ^{PC} of equations* (2.1) *with these material properties satisfies*

$$\|u_\epsilon^{PC} - u_\epsilon\|_{\mathcal{Z}} < \delta.$$

Proof. Let $\mathcal{A}_E$ and $\mathcal{A}_\nu$ be the finite sets of discontinuities of E_ϵ and ν_ϵ, respectively, and let $\mathcal{A} = \mathcal{A}_E \cup \mathcal{A}_\nu$ with elements $a_1, a_2, \dots, a_K$. Partition the interval $\mathcal{D}$ into intervals $D_1 = (a_0, a_1), D_2 = [a_1, a_2), \dots, D_K = [a_{K-1}, a_K)$ such that $\bigcup_{k=1}^K D_k = \mathcal{D}$ and $\bigcap_{k=1}^K D_k = 0$. Let $B_{k,\delta} = \{b_0^k, b_1^k, \dots, b_{N(\delta)}^k\}$ be a uniform partition of D_k such that $b_i^k - b_{i-1}^k = \delta$. Furthermore, define E_ϵ^{PC} and ν_ϵ^{PC} via

$$\begin{aligned}E_\epsilon^{PC}(x) &= \sum_{k=1}^K\sum_{n=1}^N \mathbb{1}_{x\in(b_{n-1}^k, b_n^k]} E\left(\frac{1}{2}b_{n-1}^k + \frac{1}{2}b_n^k\right),\\ \nu_\epsilon^{PC}(x) &= \sum_{k=1}^K\sum_{n=1}^N \mathbb{1}_{x\in(b_{n-1}^k, b_n^k]} \nu\left(\frac{1}{2}b_{n-1}^k + \frac{1}{2}b_n^k\right)\end{aligned}$$

for $x \in \mathcal{D}$, noting that E_ϵ^{PC} and ν_ϵ^{PC} are piecewise constant with $KN(\delta)$ pieces.

E_ϵ and ν_ϵ are continuous on each interval D_k, so for each $\delta' > 0$, there exists a mesh width δ such that with partitions $\{B_{k,\delta}\}_{k=1}^K$

$$\begin{aligned}\sup_{x\in(b_{n-1}^k, b_n^k]}\|E_\epsilon\left(\frac{1}{2}b_{n-1}^k + \frac{1}{2}b_n^k\right) - E_\epsilon(x)\| < \delta',\\ \sup_{x\in(b_{n-1}^k, b_n^k]}\|\nu_\epsilon\left(\frac{1}{2}b_{n-1}^k + \frac{1}{2}b_n^k\right) - \nu_\epsilon(x)\| < \delta'\end{aligned}$$

for all $n \in \{1, \dots, N(\delta)\}$. Thus, $\|E^{PC} - E\|_\infty < \delta'$ and $\|\nu^{PC} - \nu\|_\infty < \delta'$. Since δ' was arbitrary, we can pick $\delta' < \frac{\eta}{C_1}$, where C_1 is as in Lemma 3.5, and the theorem follows by use of the same lemma. ∎

A.2. Proof of Theorem 3.6. We will need the following lemma.

LEMMA A.3 (existence of exact parametrization). *For a piecewise constant material with $L'+1$ pieces and under Assumptions* 3.1, *a_0 in* (2.5) *can be written exactly as*

$$\widehat{a}_0(s) = E' + \nu' s - \sum_{\ell=1}^{L'} \frac{\beta_\ell}{s+\alpha_\ell},$$

where $E', \nu', \beta_\ell \in \mathbb{R}$ and $\alpha_\ell \in \mathbb{R}_+$ for all $\ell \in [L']$.

Proof. Let $E(y)$ and $\nu(y)$ have $L'+1$ constant pieces of lengths $\{d_\ell\}_{\ell=1}^{L'+1}$, each associated to values $\{E_\ell\}_{\ell=1}^{L'+1}$ and $\{\nu_\ell\}_{\ell=1}^{L'+1}$ of E and ν. Then (2.5), rewritten here for convenience,

$$\widehat{a}_0(s) = \left(\int_0^1 \frac{dy}{s\nu(y) + E(y)} \right)^{-1},$$

becomes

$$\widehat{a}_0(s) = \left[\sum_{\ell=1}^{L'+1} \frac{d_\ell}{E_\ell + \nu_\ell s} \right]^{-1} \tag{A.2}$$

$$= \frac{\prod_{\ell=1}^{L'+1}(E_\ell + \nu_\ell s)}{\sum_{\ell=1}^{L'+1} d_\ell \prod_{j\neq\ell}(E_j + \nu_j s)} \tag{A.3}$$

$$= \frac{P(s)}{Q(s)}, \tag{A.4}$$

where $P(s)$ is a polynomial of degree $L'+1$ and $Q(s)$ a polynomial of degree L'. Therefore, there exists a decomposition

$$\frac{P(s)}{Q(s)} = E' + \nu' s - \frac{C(s)}{Q(s)} \tag{A.5}$$

for some constants E' and ν' and polynomial $C(s)$ of degree $L'-1$.

Let $-\alpha_1, \dots, -\alpha_{L'}$ be the roots of $Q(s)$. Then $\frac{C(s)}{Q(s)} = \sum_{\ell=1}^{L'} \frac{\beta_\ell}{s+\alpha_\ell}$ for some constants $\beta_\ell \in \mathbb{C}$ by partial fraction decomposition. We wish to show that $\Re(\alpha_\ell) > 0$ for all roots $-\alpha_\ell$ of $Q(s)$ so that we can take the inverse Laplace transform. Furthermore, we wish to show that, in fact, $-\alpha_\ell \in \mathbb{R}$ for all roots $-\alpha_\ell$ so that $\beta_\ell \in \mathbb{R}$ as well. Since E_j and v_j are positive for all $j \in [L'+1]$, it is clear that if a root $-\alpha_\ell$ is real, then it cannot be positive since $Q(s) = \sum_{\ell=1}^{L'+1} d_\ell \prod_{j\neq\ell}(E_j + \nu_j s)$ has all positive coefficients. We now show that all roots of $Q(s)$ are real. Suppose $a+bi$ is a root of $Q(s)$. Then

$$\begin{aligned} Q(a+bi) &= \sum_{\ell=1}^{L'+1} d_\ell \prod_{j\neq\ell} (E_j + \nu_j(a+bi)) \\ &= \left[\prod_{j=1}^{L'+1} (E_j + \nu_j(a+bi)) \right] \cdot \sum_{\ell=1}^{L'+1} \frac{d_\ell}{E_\ell + \nu_\ell(a+bi)} \\ &= \left[\prod_{j=1}^{L'+1} (E_j + \nu_j(a+bi)) \right] \cdot \sum_{\ell=1}^{L'+1} \left(\frac{d_\ell(E_\ell + \nu_\ell a)}{(E_\ell+\nu_\ell a)^2 + (\nu_\ell b)^2} - \frac{d_\ell(\nu_\ell b)}{(E_\ell+\nu_\ell a)^2 + (\nu_\ell b)^2} i \right). \end{aligned}$$

The term $\prod_{j=1}^{L'+1}(E_j+\nu_j(a+bi))$ is a nonzero constant for $b\neq 0$ since $E_j,\nu_j\in\mathbb{R}_+$. Therefore, for $Q(a+bi)=0$, both the real and imaginary components of the sum on the right must total 0. However, since d_ℓ, ν_ℓ, and the denominator term $(E_\ell+\nu_\ell a)^2+(\nu_\ell b)^2$ are all positive as well, b must equal 0 to make $\mathrm{Im}[Q(a+bi)]=0$. Therefore, all roots of $Q(s)$ are in $\mathbb{R}_-$. Returning to the decomposition, we now have

$$\widehat{a}_0(s)=E'+\nu' s-\sum_{\ell=1}^{L'}\frac{\beta_\ell}{s+\alpha_\ell}, \tag{A.6}$$

where $\beta_\ell\in\mathbb{R}$ and $\alpha_\ell\in\mathbb{R}_+$ for all $\ell\in[L']$. ∎

Now we may prove the theorem.

THEOREM 3.6 (existence of exact parametrization). *Let $\Psi_0^\dagger$ be the map from strain history to stress in the homogenized model, as defined by (2.7), in a piecewise-constant material with $L'+1$ pieces. Define $\Psi_0^{PC}:\mathbb{R}^2\times C(\overline{\mathcal{T}};\mathbb{R})\times\mathcal{T}\times\Theta\to\mathbb{R}$ by*

(3.4a)
$$\Psi_0^{PC}(\partial_x u_0(t),\partial_{xt}^2 u_0(t),\{\partial_x u_0(\tau)\}_{\tau\in\overline{\mathcal{T}}},t;\theta)=E_0\partial_x u_0(t)+\nu_0\partial_{xt}^2 u_0(t)-\sum_{\ell=1}^{L_0}\xi_\ell(t),$$

(3.4b)
$$\partial_t\xi_\ell(t)=\beta_\ell\partial_x u_0(t)-\alpha_\ell\xi_\ell(t),\ \xi_\ell(0)=0,\quad \ell\in\{1,\dots,L_0\},$$

with parameter space

$$\Theta=\left(E_0\in\mathbb{R}_+,\ \nu_0\in\mathbb{R}_+,\ L_0\in\mathbb{Z}_+,\ \alpha_0\in\mathbb{R}_+^{L_0},\ \beta_0\in\mathbb{R}^{L_0}\right). \tag{3.5}$$

Then, under Assumptions 3.1, there exists $\theta^*\in\Theta$ with $(E_0,\nu_0,L_0,\alpha_0,\beta_0)=(E',\nu',L',\alpha,\beta)$ such that

$$\Psi_0^\dagger(\partial_x u_0(t),\partial_{xt}^2 u_0(t),\{\partial_x u_0(\tau)\}_{\tau\in\overline{\mathcal{T}}},t)=\Psi_0^{PC}(\partial_x u_0(t),\partial_{xt}^2 u_0(t),\{\partial_x u_0(\tau)\}_{\tau\in\overline{\mathcal{T}}},t;\theta^*)$$

for all $u_0\in\mathcal{C}^2(\overline{\mathcal{D}}\times\overline{\mathcal{T}};\mathbb{R})$ and $t\in\mathcal{T}$.

Proof. By Lemma A.3, we have that

$$\begin{aligned}\widehat{\sigma_0}&=\widehat{a}_0(s)\partial_x\widehat{u_0}\\&=\left(E'+\nu' s-\sum_{\ell=1}^{L'}\frac{\beta_\ell}{s+\alpha_\ell}\right)\partial_x\widehat{u_0},\end{aligned}$$

where $\beta_\ell\in\mathbb{R}$ and $\alpha_\ell\in\mathbb{R}_+$ for all $\ell\in[L']$. Taking an inverse Laplace transform, we get

$$\sigma_0(t)=E'\partial_x u_0(t)+\nu'\partial_t\partial_x u_0(t)-\sum_{\ell=1}^{L'}\beta_\ell\int_0^t\partial_x u_0(\tau)\exp[-\alpha_\ell(t-\tau)]\,d\tau. \tag{A.7}$$

The above may be reexpressed as equations (3.4) with a choice of parameters $\theta=(E',\nu',L',\alpha,\beta)$ and auxiliary variable ξ. ∎

A.3. Proof of RNN approximation Theorem 3.10. In this subsection we use $|\cdot|, \|\cdot\|$ to denote the modulus and Euclidean norms, respectively, and $\langle\cdot,\cdot\rangle$ to denote the Euclidean inner product. This overlap with the notation from subsection 1.4 should not lead to any confusion as it is confined to this subsection.

To prove Theorem 3.10 we first study the simple case where $\mathcal{F}^{PC}, \mathcal{G}^{PC}$ are uniformly approximated across all inputs; subsequently we will use this to establish Theorem 3.10 as stated.

Assumptions A.4. For any $\delta > 0$, there exist $\mathcal{F}^{RNN}$ and $\mathcal{G}^{RNN}$ such that

$$\sup_{z\in\mathbb{R}^{2+L'}} \left|\mathcal{F}^{PC}(z) - \mathcal{F}^{RNN}(z)\right| \le \delta,$$
$$\sup_{z\in\mathbb{R}^{1+L'}} \left\|\mathcal{G}^{PC}(z) - \mathcal{G}^{RNN}(z)\right\| \le \delta.$$

PROPOSITION A.5. *Under Assumptions* A.4, *if* $\{\alpha_\ell\}$ *in equations* (3.10) *are bounded such that* $0 < \rho < \alpha_\ell$ *for some* ρ *for all* $\ell \in [L']$, *then for any* $\eta > 0$, *there exists a map* Ψ_0^{RNN} *defined as in equations* (3.12) *such that for* Ψ_0^{PC} *defined in equations* (3.11), *for any* $t \in \mathbb{R}^+$ *and functions* $b, c : \mathbb{R}^+ \to \mathbb{R}$,

$$\left|\Psi_0^{PC}(b(t), c(t), \{b(\tau)\}_{\tau\in\overline{\mathcal{T}}}, t; \theta^*) - \Psi_0^{RNN}(b(t), c(t), \{b(\tau)\}_{\tau\in\overline{\mathcal{T}}}, t)\right| \le \eta.$$

Proof. Note that the main difficulty in this proof results from the fact that $\mathcal{F}^{RNN}$ and $\mathcal{F}^{PC}$ act on different hidden variables ξ, which we will denote ξ^{RNN} and ξ^{PC}, and whose first order time derivatives are given by $\mathcal{G}^{RNN}$ and $\mathcal{G}^{PC}$, respectively. We write

$$\begin{aligned}
&\left|\Psi_0^{PC}(b(t), c(t), \{b(\tau)\}_{\tau\in\overline{\mathcal{T}}}, t; \theta^*) - \Psi_0^{RNN}(b(t), c(t), \{b(\tau)\}_{\tau\in\overline{\mathcal{T}}}, t)\right| \\
&= |\mathcal{F}^{PC}(b(t), c(t), \xi^{PC}(t)) - \mathcal{F}^{RNN}(b(t), c(t), \xi^{RNN}(t))| \\
&\le |\mathcal{F}^{PC}(b(t), c(t), \xi^{RNN}(t)) - \mathcal{F}^{RNN}(b(t), c(t), \xi^{RNN}(t))| \\
&\quad + |\mathcal{F}^{PC}(b(t), c(t), \xi^{PC}(t)) - \mathcal{F}^{PC}(b(t), c(t), \xi^{RNN}(t))| \\
&\le \delta + |\mathcal{F}^{PC}(b(t), c(t), \xi^{PC}(t)) - \mathcal{F}^{PC}(b(t), c(t), \xi^{RNN}(t))|
\end{aligned}$$

by Assumptions A.4 since $\mathcal{F}^{PC}$ and $\mathcal{F}^{RNN}$ share arguments in the first term. To bound the second term,

$$\begin{aligned}
|\mathcal{F}^{PC}(b(t), c(t), \xi^{PC}(t)) - \mathcal{F}^{PC}(b(t), c(t), \xi^{RNN}(t))| &= |\langle \mathbb{1}, \xi^{PC}(t) - \xi^{RNN}(t)\rangle| \\
&\le \sqrt{L'}\|\xi^{PC}(t) - \xi^{RNN}(t)\|
\end{aligned}$$

using the known form of $\mathcal{F}^{PC}$ where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^{L'}$.

Let $e_\xi(t) = \xi^{PC}(t) - \xi^{RNN}(t)$. Note that $\xi^{PC}(0) = \xi^{RNN}(0) = 0$, so $e_\xi(0) = 0$. We wish to bound $\|e_\xi(t)\|$. To do so, we first bound $\|\dot{e}_\xi(t)\|$, where $\dot{e}_\xi(t) = \frac{d}{dt}e_\xi(t)$:

$$\begin{aligned}
\dot{e}_\xi(t) &= \dot{\xi}^{PC}(t) - \dot{\xi}^{RNN}(t) \\
&= \mathcal{G}^{PC}(\xi^{PC}(t), b(t)) - \mathcal{G}^{PC}(\xi^{RNN}(t), b(t)) - \mathcal{G}^{RNN}(\xi^{RNN}(t), b(t)) \\
&\quad + \mathcal{G}^{PC}(\xi^{RNN}(t), b(t)) \\
&= \mathcal{G}^{PC}(\xi^{PC}(t), b(t)) - \mathcal{G}^{PC}(\xi^{RNN}(t), b(t)) + q(t),
\end{aligned}$$

where we have defined $q(t) = \mathcal{G}^{PC}(\xi^{RNN}(t), b(t)) - \mathcal{G}^{RNN}(\xi^{RNN}(t), b(t))$ and $\|q(t)\| \le \delta$ by Assumptions A.4. Now note that $\dot{e}_\xi(t) = -Ae_\xi(t) + q(t)$ by the form of $\mathcal{G}^{PC}$, so we can bound

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|e_\xi(t)\|^2 &= \langle e_\xi(t), \dot{e}_\xi(t)\rangle = -\langle e_\xi(t), Ae_\xi(t)\rangle + \langle q(t), e_\xi(t)\rangle \\ &\le -\alpha_{\min}\|e_\xi(t)\|^2 + \left\langle \frac{1}{\alpha_{\min}^{1/2}} q(t), \alpha_{\min}^{1/2} e_\xi(t)\right\rangle \\ &\le -\alpha_{\min}\|e_\xi(t)\|^2 + \frac{1}{2\alpha_{\min}}\|q(t)\|^2 + \frac{\alpha_{\min}}{2}\|e_\xi(t)\|^2 \\ \frac{d}{dt}\|e_\xi(t)\|^2 &\le -\alpha_{\min}\|e_\xi(t)\|^2 + \frac{\delta^2}{\alpha_{\min}}\end{aligned}$$

by Young's inequality. Then by Gronwall's inequality

$$\|e_\xi(t)\|^2 \le \frac{\delta^2}{\alpha_{\min}^2}\left(1 - e^{-\alpha_{\min} t}\right) \tag{A.8}$$

so $\|e_\xi(t)\| < \frac{\delta}{\rho}$ for all time. Returning to the main proof narrative,

$$\begin{aligned}&\left|\Psi_0^{PC}(b(t), c(t), \{b(\tau)\}_{\tau\in\overline{\mathcal{T}}}, t; \theta^*) - \Psi_0^{RNN}(b(t), c(t), \{b(\tau)\}_{\tau\in\overline{\mathcal{T}}}, t)\right| \\ &\le \delta + \sqrt{L'}\|\xi^{PC}(t) - \xi^{RNN}(t)\| \le \delta + \frac{\sqrt{L'}\delta}{\rho}.\end{aligned}$$

Since by Assumptions A.4 δ is arbitrarily small, the theorem result is shown. ∎

Although we did not need to restrict the inputs t, b, and c in Proposition A.5 to compact sets in order to prove it, we will argue that the statement holds under weaker assumptions if the inputs are also bounded. The following weaker assumptions follow from the Universal Approximation Theorem for RNNs [9].

Assumptions A.6. If $D_1 \in \mathbb{R}^{2+L'}$ and $D_2 \in \mathbb{R}^{1+L'}$ are compact sets, then for any $\delta > 0$, there exist $\mathcal{F}^{RNN}$ and $\mathcal{G}^{RNN}$ such that

$$\begin{aligned}\sup_{z\in D_1}\left|\mathcal{F}^{PC}(z) - \mathcal{F}^{RNN}(z)\right| &\le \delta, \\ \sup_{z\in D_2}\left\|\mathcal{G}^{PC}(z) - \mathcal{G}^{RNN}(z)\right\| &\le \delta.\end{aligned}$$

THEOREM 3.10 (RNN approximation). *Consider* Ψ_0^{PC} *defined as by* (3.10), (3.11). *Assume that there exist* $\rho > 0$ *and* $0 \le B < \infty$ *such that* $\rho < \min_\ell |\alpha_\ell|$ *and* $\max_\ell |\beta_\ell| \le B$. *Then, under Assumptions* 3.1, *for every* $\eta > 0$ *there exists* Ψ_0^{RNN} *of the form* (3.12) *such that*

$$\sup_{t\in\mathcal{T}, b, c\in \mathsf{Z}_R}\left|\Psi_0^{PC}\big(b(t), c(t), \{b(\tau)\}_{\tau\in\overline{\mathcal{T}}}, t; \theta^*\big) - \Psi_0^{RNN}\big(b(t), c(t), \{b(\tau)\}_{\tau\in\overline{\mathcal{T}}}, t\big)\right| < \eta.$$

Proof. Notice first that Assumptions A.6 are a weaker version of Assumptions A.4. We will prove the theorem by showing that, for inputs bounded via $t \in \mathcal{T}$ and $b, c \in \mathsf{Z}_R$, we never need the stronger assumption in the proof of Proposition A.5 because the function arguments of $\mathcal{F}^{PC}, \mathcal{F}^{RNN}, \mathcal{G}^{PC}$, and $\mathcal{G}^{RNN}$ never leave a compact set. First we show that $\sup_{t\in\mathcal{T}}\|\xi^{PC}(t)\| \le R_3$ for some $R_3 > 0$. For any $\ell \in \{1, \dots, L'\}$, we have

$$\begin{aligned}\dot{\xi}_\ell^{PC}(t) &= \beta_\ell b(t) - \alpha_\ell \xi_\ell^{PC}(t),\\ |\xi_\ell^{PC}(t)| &\le e^{-\alpha_\ell t}\beta_\ell \left(\sup_{t\in\mathcal{T}}|b(t)|\right)\int_0^t e^{\alpha_\ell t'}\,dt'\\ &\le e^{-\alpha_\ell t}\beta_\ell \left(\sup_{t\in\mathcal{T}}|b(t)|\right)\frac{1}{\alpha_\ell}e^{\alpha_\ell t},\\ \sup_{t\in\mathcal{T}}|\xi_\ell^{PC}(t)| &\le \frac{B}{\rho}R,\end{aligned}$$

so that $\sup_{t\in\mathcal{T}}\|\xi^{PC}(t)\| \le \frac{\sqrt{L'}BR}{\rho}$. Let $R_3 = \frac{\sqrt{L'}BR}{\rho}$. Define $R_4 = \max\{R_3 + \frac{\delta}{\rho}, R\}$ for δ in Assumptions A.6. We will show that $\sup_{t\in\mathcal{T}}\|\xi^{RNN}(t)\| \le R_4$ for ξ^{RNN} defined by ξ^{RNN} in equations (3.12). Then the proof of Proposition A.5 will apply for bounded t, b, and c with the weaker assumptions since all inputs to $\mathcal{F}^{PC}, \mathcal{F}^{RNN}, \mathcal{G}^{PC}$, and $\mathcal{G}^{RNN}$—$b(t)$, $c(t)$, $\xi^{PC}(t)$, and $\xi^{RNN}(t)$—will remain in a compact set for $t\in\mathcal{T}$.

Suppose for the sake of contradiction that there exists a time $t'\in\mathcal{T}$ at which $\|\xi^{RNN}(t')\| > R_4$. Then there exists a time $T' < t' < T$ and $\epsilon > 0$ such that for $t\in[0,T']$, $\|\xi^{RNN}(t)\| \le R_4$ for $t\in(T', T'+\epsilon)$, $\|\xi^{RNN}(t)\| > R_4$, and $\|\xi^{RNN}(T')\| = R_4$ by continuity. In other words, T' is the time at which ξ^{RNN} first crosses the R_4 radius. Then

$$\|e_\xi(T')\| := \|\xi^{RNN}(T') - \xi^{PC}(T')\| \ge R_4 - R_3 \ge \frac{\delta}{\rho}$$

by the triangle inequality. Since $\|\xi^{RNN}(t)\| \le R_4$ for $t\in[0,T']$, the bound on $\|e_\xi(t)\|$ in (A.8) in the proof of Proposition A.5 applies on the interval $t\in[0,T']$ under the weaker Assumptions A.6, and $\|e_\xi(T')\| < \frac{\delta}{\rho}$. This is a contradiction. Therefore, $\sup_{t\in\mathcal{T}}\|\xi^{RNN}(t)\| \le R_4$, and the proof of Proposition A.5 holds with the weaker Assumptions A.6 for bounded inputs $t\in\mathcal{T}$ and $b, c\in\mathsf{Z}_R$. ∎

The bounds on α and β required in Theorem 3.10 are justified because for known material properties E and ν, α and β are determined and finite-dimensional, so they have maximum and minimum values.

Appendix B. Special case solutions.

B.1. Laplace transform limit. Here we derive the form of $\Psi_0^\dagger$ in (2.7) via a power series expansion of the Laplace transform at $s=\infty$. Starting from the definition in (2.6),

$$\Psi_0^\dagger = \mathcal{L}^{-1}\left(\left(\int_0^1 \frac{dy}{s\nu(y)+E(y)}\right)^{-1}\partial_x\widehat{u}_0\right).$$

For $s\gg 1$, we have that

$$\left(\int_0^1 \frac{dy}{s\nu(y)+E(y)}\right)^{-1} \approx \left(\int_0^1 \frac{dy}{s\nu(y)}\right)^{-1} = s\left(\frac{dy}{\nu(y)}\right)^{-1}.$$

Setting $\nu' = \left(\frac{dy}{\nu(y)}\right)^{-1}$, we now subtract out the linear dependence on s and let $z=\frac{1}{s}$. We define

$$F(z) = \hat{a}_0(s) - \nu' s\Big|_{s=z^{-1}}$$

to obtain

$$\begin{aligned}F(z) &= \hat{a}_0\left(z^{-1}\right) - \nu' z^{-1} \\ &= \left(\int_0^1 \frac{z\,dy}{\nu(y)+zE(y)}\right)^{-1} - \left(\int_0^1 \frac{z\,dy}{\nu(y)}\right)^{-1} \\ &= \frac{\int_0^1 \left(\frac{z}{\nu(y)} - \frac{z}{\nu(y)+zE(y)}\right) dy}{\left(\int_0^1 \frac{z\,dy}{\nu(y)+zE(y)}\right)\left(\int_0^1 \frac{z\,dy}{\nu(y)}\right)} \\ &= \frac{z^2 \int_0^1 \frac{E(y)}{\nu(y)(\nu(y)+zE(y))}\,dy}{z^2\left(\int_0^1 \frac{dy}{\nu(y)+zE(y)}\right)\left(\int_0^1 \frac{dy}{\nu(y)}\right)} \\ &= \frac{\int_0^1 \frac{E(y)}{\nu(y)(\nu(y)+zE(y))}\,dy}{\left(\int_0^1 \frac{dy}{\nu(y)+zE(y)}\right)\left(\int_0^1 \frac{dy}{\nu(y)}\right)}.\end{aligned}$$

Since $\inf_{y\in(0,1)} \nu(y) > 0$,

$$\lim_{z\to 0} F(z) = \frac{\int_0^1 \frac{E(y)}{\nu^2(y)}\,dy}{\left(\int_0^1 \frac{dy}{\nu(y)}\right)^2} =: E'.$$

From this same computation, we see that for $\widehat{a}_0(s) = s\nu' + E' + \kappa(s)$, the contribution $\kappa(s)$ consists of lower order terms in s and is such that $\lim_{s\to\infty} \kappa(s) = 0$. Using the fact that the inverse Laplace transform of a product (if it exists) is a convolution, we justify the form of the integral term in (2.7).

Forced boundary problem.

LEMMA 3.12. *Let u solve the equations*

$$\partial_y \sigma(y,t) = 0, \quad (y,t) \in \mathcal{D}\times\mathcal{T}, \tag{3.15a}$$

$$\sigma(y,t) = E(y)\partial_y u(y,t) + \nu(y)\partial_{ty}^2 u(y,t), \quad (y,t) \in \mathcal{D}\times\mathcal{T}, \tag{3.15b}$$

$$u(0,t) = 0, \quad u(1,t) = b(t), \quad (y,t) \in \partial\mathcal{D}\times\mathcal{T}, \tag{3.15c}$$

$$u(y,0) = 0, \quad y \in \mathcal{D}. \tag{3.15d}$$

Then

$$\{\sigma(t)\}_{t=0}^T = \Psi_0^\dagger\left(b(t), \partial_t b(t), \{b(t)\}_{t=0}^T, t\right),$$

where $\Psi_0^\dagger$ is the map defined in (2.6).

Proof. Taking the Laplace transform of (3.15) yields

$$\widehat{\sigma}(s) = (E(y) + \nu(y)s)\partial_y \widehat{u}(y,s).$$

Spatially averaging and noting that $b(t) = \langle \partial_y u(y,t)\rangle$, we have

$$\widehat{b}(s) = \int_0^1 \frac{dy}{(E+s\nu)(y)}\widehat{\sigma}(s). \tag{B.1}$$

Then $\widehat{\sigma}(s) = (\int_0^1 \frac{dy}{(E+s\nu)(y)})^{-1}\widehat{b}(s)$, which is equivalent to $\widehat{\sigma}(s) = \widehat{a_0}(s)\widehat{b}(s)$ using (2.5). The definition of $\Psi_0^\dagger$ in (2.6) completes the proof. ∎

Lemma 3.12 justifies the use of data arising from the system (3.15) to train the map Ψ_0.

Appendix C. Macroscale numeric comparisons.

C.1. RNN training and testing: Piecewise-constant case. We trained three RNNs using the same dataset for the setting of a 2-piecewise-constant material with material parameters $E_1 = 1$, $E_2 = 3$, $\nu_1 = 0.1$, and $\nu_2 = 0.2$. The data was generated using a forward Euler method with time discretization $dt = 0.001$ up to time $T = 4$ on the known analytic solution for the 2-piecewise-constant cell problem. Denote the data by $\{(\partial_x u_0)_n, (\sigma_0)_n\}_{n=1}^N$ as discussed in section 3.4. We repeat the two loss functions here.

Accessible Loss Function:

$$L_1(\{\sigma_0\}_{n=1}^N, \{\widehat{\sigma}_0\}_{n=1}^N) = \frac{1}{N}\sum_{n=1}^N \frac{\|(\sigma_0)_n - (\widehat{\sigma}_0)_n\|}{\|(\sigma_0)_n\|}.$$

Inaccessible Loss Function:

$$L_2(\{\sigma_0\}_{n=1}^N, \{\widehat{\sigma}_0\}_{n=1}^N, \{\xi\}_{n=1}^N, \{\widehat{\xi}_0\}_{n=1}^N) = \frac{1}{N}\sum_{n=1}^N \left(\frac{\|(\sigma_0)_n - (\widehat{\sigma}_0)_n\|}{\|(\sigma_0)_n\|} + \frac{\|(\xi)_n - (\widehat{\xi})_n\|}{\|(\xi)_n\|_{L^2(\mathcal{D},\mathbb{R})}} \right).$$

For each of the following RNNs, the architecture for $\mathcal{F}_{RNN}$ and $\mathcal{G}_{RNN}$ consists of three internal layers of SELU units of 100 nodes separated by linear layers, all followed by a final linear layer. The SELU function is applied elementwise as

$$SELU(x) = s(\max(0, x) + \min(0, \alpha(\exp(x) - 1))),$$

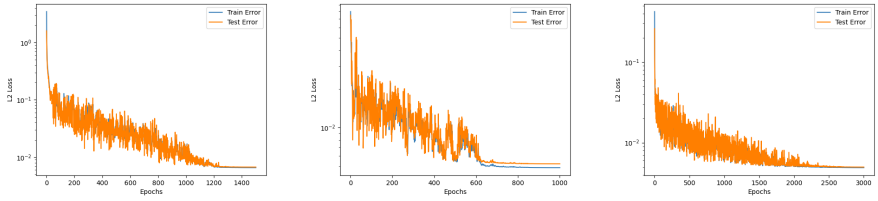
where $\alpha = 1.67326$ and $s = 1.05070$.[1] We trained three different RNNs on the same dataset in the following manner:

- **RNN "A"**: Using only the inaccessible loss function L_2, we trained on $N = 400$ data points with subsampled time discretization of $dt = 0.004$ up to $T = 4$ for 1500 epochs with a batch size of 50.
- **RNN "B"**: First we used the inaccessible loss function L_2 to train on $N = 200$ data points with subsampled time discretization of $dt = 0.004$ up to $T = 2$ for 1500 epochs with a batch size of 40. Then we initialized a new RNN at the parameters of this RNN and trained with the accessible loss function L_1 for 1000 epochs on 200 data with batch size of 40.
- **RNN "C"**: Using only the accessible loss function L_1, we trained on $N = 500$ data points with subsampled time discretization of $dt = 0.004$ up to $T = 4$ for 3000 epochs with a batch size of 50.

The train and test errors are shown for the three RNNs in Figure 12.

C.2. RNN training and testing: Continuous case. We trained several RNNs on data $\{(\partial_x u_0)_n, (\sigma_0)_n\}_{n=1}^N$ for continuous material parameters given by $E(y) = 2 + \tanh(\frac{y-0.5}{0.2})$ and $\nu(y) = 0.5 + 0.1\tanh(\frac{y-0.5}{0.2})$. Each of the four RNNs had 1, 2, 5, and 10 hidden variables (L_0, or the dimension of ξ), respectively. The data was generated by solving the cell problem using a finite difference method with 200 spatial nodes and $dt = h^2$, where h is the spatial discretization. The RNN was trained for 3000 epochs on 500 data. The macroscale simulations were performed with a spatial discretization of $h_{cell} = 0.05$ and a time discretization of $dt = 0.4h_{cell}^2$. They were compared to an FEM solution computed with a spatial discretization of $h = 0.004$ with a material period of 0.04 and time discretization of $dt = 0.1h^2$.

[1] https://pytorch.org/docs/stable/generated/torch.nn.SELU.html



(a) RNN "A" trained using inaccessible loss function (b) RNN "B" initialized at inaccessible loss solution (c) RNN "C" trained using only standard loss function

FIG. 12. *Train and test errors for the three RNNs.*

C.3. RNN training and testing: Additional piecewise-constant experiments. The piecewise constant data was generated by solving the cell problem using a finite difference method with 300 spatial nodes and $dt = 0.005$ over a time length of $T = 10$ for the trajectories. In the training, we sliced the data trajectories by a slice of 2. For the 3-piecewise-constant model, we trained on 500 data points for 3000 epochs with the squared relative loss function. For the 5- and 10-piecewise-constant models, we trained on 600 data points for 4000 epochs with the squared relative loss function. The piecewise constant values were $E_1 = 2$, $E_2 = 8$, $E_3 = 1$, $E_4 = 3$, $E_5 = 7$, $E_6 = 3$, $E_7 = 5$, $E_8 = 6$, $E_9 = 9$, $E_{10} = 4$, $\nu_1 = 0.1$, $\nu_2 = 0.9$, $\nu_3 = 0.7$, $\nu_4 = 0.4$, $\nu_5 = 1.5$, $\nu_6 = 1.2$, $\nu_7 = 0.5$, $\nu_8 = 1.4$, $\nu_9 = 0.5$, and $\nu_{10} = 0.3$. (first 3 for 3-piecewise constant, first 5 for 5-piecewise constant, all 10 for 10-piecewise constant).

C.4. RNN training and testing: Elasto-viscoplasticity experiments. The data for the elasto-viscoplasticity case was generated using a fixed-point iteration scheme with $dt = 0.0001$, 100 spatial elements, and a termination threshold of 0.001. The constant values used were $n = 10$, $E_1 = 5$, $E_2 = 1$, $E_3 = 3$, $E_4 = 2$, $E_5 = 4$, $E_6 = 6$, $E_7 = 1$, $E_8 = 3$, $E_9 = 4$, $\dot{\epsilon}_{p0,1} = 0.05$, $\dot{\epsilon}_{p0,2} = 0.1$, $\dot{\epsilon}_{p0,3} = 0.15$, $\dot{\epsilon}_{p0,4} = 0.07$, $\dot{\epsilon}_{p0,5} = 0.02$, $\dot{\epsilon}_{p0,6} = 0.08$, $\dot{\epsilon}_{p0,7} = 0.04$, $\dot{\epsilon}_{p0,0.12}$, $\dot{\epsilon}_{p0,0.03}$, and each $\sigma_{0,i} = E_i \dot{\epsilon}_{p0,i}$. The RNN was trained without the strain rate variable on 400 data trajectories with a time slice of 8. The model trained for 3000 epochs with absolute error. The samples shown in Figure 10 were chosen using a random number generator.

Appendix D. One-dimensional standard linear solid. In this section we address the model of the one-dimensional Maxwell version of the SLS, whose constitutive law depends only on the strain and strain history. The analysis for the SLS model demonstrates that the ideas presented for the KV model extend beyond that particular setting. In section D.1, we present the governing equations, and in section D.2 we homogenize the system.

D.1. Governing equations. In the setting without inertia, the displacement u_ϵ, strain e_ϵ, and inelastic strain e^p_ϵ are related by

$$-\partial_x \sigma_\epsilon = f, \tag{D.1a}$$

$$e_\epsilon = \partial_x u, \tag{D.1b}$$

$$\sigma_\epsilon = E_{1,\epsilon} e_\epsilon + E_{2,\epsilon}(e_\epsilon - e^p_\epsilon), \tag{D.1c}$$

$$\partial_t e^p_\epsilon = \frac{E_{2,\epsilon}}{\nu_\epsilon}(e_\epsilon - e^p_\epsilon), \tag{D.1d}$$

where $f : \mathcal{D} \times \mathcal{T} \mapsto \mathbb{R}$ is a known forcing, and we impose initial condition $u(x,0) = 0$ and boundary conditions $u(x,t) = 0$ for $x \in \partial\mathcal{D}$. We seek a solution $u_\epsilon : \mathcal{D} \times \mathcal{T} \mapsto \mathbb{R}$. Once more we have small-scale dependence in the material properties through ϵ: we have $E_{i,\epsilon}(x) = E_i(\frac{x}{\epsilon})$ for $i = 1, 2$ and $\nu_\epsilon(x) = \nu(\frac{x}{\epsilon})$ for $0 < \epsilon \ll 1$.

D.2. Homogenization. First, we take the Laplace transform of (D.1) and combine the transformed expressions of (D.1c) and (D.1d) to arrive at

$$\widehat{\sigma}_\epsilon = E_{1,\epsilon}\widehat{e}_\epsilon + E_{2,\epsilon} s \left(s + \frac{E_{2,\epsilon}}{\nu_\epsilon} \right)^{-1} \widehat{e}_\epsilon. \tag{D.2}$$

Letting $\widehat{a}(s) = E_{1,\epsilon} + E_{2,\epsilon} s (s + \frac{E_{2,\epsilon}}{\nu_\epsilon})^{-1}$, the homogenization theory of section 3.2 applies, and we can use the harmonic averaging expression in (2.5) to write

$$\widehat{a}_0(s) = \left\langle (a(s))^{-1} \right\rangle^{-1} = \left(\int_0^1 \frac{s + \frac{E_2}{\nu}}{(E_1 + E_2)s + \frac{E_1 E_2}{\nu}} dy \right)^{-1}, \tag{D.3}$$

where the homogenized solution u_0 solves system 1.2 with $\Psi_0^\dagger$ defined as

$$\Psi_0^\dagger = \mathcal{L}^{-1} \left[\widehat{a}_0(s) \partial_x \widehat{u}_0 \right], \tag{D.4}$$

analogous to the KV case. However, in the case of piecewise-constant E_1, E_2, and ν the inverse Laplace transform yields a different form in the SLS case:

$$\Psi_0^{PC}(\partial_x u_0, t; \theta) = E' \partial_x u_0(t) - \sum_{\ell=1}^{L} \xi_\ell(t), \tag{D.5a}$$

$$\partial_t \xi_\ell(t) = \beta_\ell \partial_x u_0(t) - \alpha_\ell \xi_\ell(t), \quad \ell \in \{1, \dots, L\}, \tag{D.5b}$$

for a material with L pieces. Note that this model does not have dependence on the strain rate, but it has one more hidden variable than the piecewise-constant case for the KV model. The value of E' follows from taking the limit $s \to \infty$ and is given by

$$E' = \left(\int_0^1 \frac{1}{(E_1 + E_2)} dy \right)^{-1}.$$

Full derivation may be found in Appendix D.3.

D.3. SLS derivation. Here we show that the SLS model has one more hidden variable in the piecewise-constant case than the KV model does. This is the analogue of Theorem 3.6 for the SLS model. Starting from (D.3) for $\hat{a}_0(s)$,

$$\begin{aligned} \widehat{a}_0(s) &= \langle \widehat{a}(s)^{-1} \rangle^{-1} \\ &= \left(\int_0^1 \frac{s + \frac{E_2(y)}{\nu}}{(E_1(y) + E_2(y))s + \frac{E_1(y)E_2(y)}{\nu(y)}} \, dy \right)^{-1} \\ &\quad \left(\sum_{i=1}^{L} \frac{(s + \frac{E_{2,i}}{\nu_i}) d_i}{(E_{1,i} + E_{2,i})s + \frac{E_{1,i}E_{2,i}}{\nu_i}} \right)^{-1} \end{aligned}$$

for L-piecewise-constant E_1, E_2, and ν with pieces of length d_i. Let $c_i = \frac{E_{2,i}}{\nu_i}$, $k_i = E_{1,i} + E_{2,i}$, and $p_i = \frac{E_{1,i}E_{2,i}}{\nu_i}$. Continuing,

$$= \left(\sum_{i=1}^{L} \frac{(s+c_i)d_i}{k_i s + p_i} \right)^{-1}$$
$$= \frac{\prod_{i=1}^{L}(k_i s + p_i)}{\sum_{i=1}^{L} d_i (s+c_i) \prod_{j\neq i}(k_j s + p_j)} := \frac{P(s)}{Q(s)}.$$

Note that both $P(s)$ and $Q(s)$ have degree L. There is a unique constant E' such that

$$\frac{P(s)}{Q(s)} = E' + \frac{C(s)}{Q(s)},$$

where $C(s)$ has degree L. Then $\frac{C(s)}{Q(s)}$ decomposes uniquely as $\sum_{\ell=1}^{L} \frac{\beta_\ell}{s+\alpha_\ell}$. Note that this is one more pole than the decomposition for the KV model in Theorem 3.6 has. We will now show that roots of Q are real and negative, which will lead to the expression in (D.5). First notice that if the roots of $Q(s)$ are real, then they must be negative since k_i, c_i, d_i, and p_i are strictly positive for all $i \in [L]$. Suppose for the sake of contradiction that $Q(s)$ has a root with a nonzero imaginary component: $s = a + bi$, where $b \neq 0$. Then

$$\begin{aligned}
Q(a+bi) &= \sum_{\ell=1}^{L} d_i (a+bi+c_\ell) \prod_{j\neq\ell} (k_j(a+bi)+p_j) \\
&= \left(\prod_j (k_j(a+bi)+p_j) \right) \sum_{\ell=1}^{L} \frac{d_\ell(a+bi+c_\ell)}{k_\ell(a+bi)+p_\ell} \\
&= \left(\prod_j (k_j(a+bi)+p_j) \right) \sum_{\ell=1}^{L} \left(\frac{d_\ell a + d_\ell c_\ell + d_\ell bi}{k_\ell a + p_\ell + k_\ell bi} \right) \left(\frac{k_\ell a + p_\ell - k_\ell bi}{k_\ell a + p_\ell - k_\ell bi} \right) \\
&= \left(\prod_j (k_j(a+bi)+p_j) \right) \\
&\quad \times \sum_{\ell=1}^{L} \frac{d_\ell}{(k_\ell a + p_\ell)^2 + (k_\ell b)^2} \left[\left((a+c_\ell)(k_\ell a + p_\ell) + k_\ell b^2 \right) + \left(-k_\ell b c_\ell + b p_\ell \right) i \right].
\end{aligned}$$

If $a+bi$ is a root of Q, then we need $b \sum_{\ell=1}^{L} \frac{d_\ell}{(k_\ell a + p_\ell)^2 + (k_\ell b)^2} (-k_\ell c_\ell + p_\ell) = 0$. Notice that $-k_\ell c_\ell + p_\ell = -\frac{E_{2,\ell}^2}{\nu_\ell}$, which is strictly negative, so for $b \neq 0$, $\Im(Q(a+bi)) < 0$, which is a contradiction. Therefore, $b = 0$, and all the roots of Q are real and negative. Inverting the Laplace transform, we arrive at (D.5).

Acknowledgments. The authors are grateful to Matt Levine for helpful discussions about training RNN models and to Pierre Suquet for helpful discussions about homogenization theory.

REFERENCES

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518, https://doi.org/10.1137/0523084.

[2] F. As'ad, P. Avery, and C. Farhat, *A mechanics-informed artificial neural network approach in data-driven constitutive modeling*, Internat. J. Numer. Methods Engrg., 123 (2022), pp. 2738–2759.
[3] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, American Mathematical Society, 2011.
[4] K. Bhattacharya, B. Hosseini, N. B. Kovachki, and A. M. Stuart, *Model reduction and neural networks for parametric PDEs*, SMAI J. Comp. Math., 7 (2021), pp. 121–157.
[5] R. Brenner and P. Suquet, *Overall response of viscoelastic composites and polycrystals: Exact asymptotic relations and approximate estimates*, Int. J. Solids Struct., 50 (2013), pp. 1824–1838.
[6] T. Chen and H. Chen, *Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems*, IEEE Trans. Neural Netw., 6 (1995), pp. 911–917.
[7] R. C. Christensen, *Theory of Viscoelasticity*, Academic Press, 1982.
[8] D. Cioranescu and P. Donato, *An Introduction to Homogenization*, Oxford University Press, Oxford, 1999.
[9] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Math. Control Signals Syst., 2 (1989), pp. 303–314.
[10] J. Fish, *Multiscale Methods: Bridging the Scales in Science and Engineering*, Oxford University Press on Demand, 2010.
[11] G. A. Francfort and P. M. Suquet, *Homogenization and mechanical dissipation in thermoviscoelasticity*, Arch. Ration. Mech. Anal., 96 (1986), pp. 265–293.
[12] O. Gonzalez and A. M. Stuart, *A First Course in Continuum Mechanics*, Cambridge University Press, 2008.
[13] M. E. Gurtin, E. Fried, and L. Anand, *The Mechanics and Thermodynamics of Continua*, Cambridge University Press, Cambridge, 2010.
[14] E. Haghighat, S. Abouali, and R. Vaziri, *Constitutive Model Characterization and Discovery Using Physics-Informed Deep Learning*, preprint, arXiv:2203.09789, 2022.
[15] E. Haghighat, M. Raissi, A. Moure, H. Gomez, and R. Juanes, *A physics-informed deep learning framework for inversion and surrogate modeling in solid mechanics*, Comput. Method Appl. Mech. Engrg., 379 (2021), 113741.
[16] Y. He, S.-H. Kang, W. Liao, H. Liu, and Y. Liu, *Robust identification of differential equations by numerical techniques from a single set of noisy observation*, SIAM J. Sci. Comput., 44 (2022), pp. A1145–A1175, https://doi.org/10.1137/20M134513X.
[17] S. H. Kang, W. Liao, and Y. Liu, *IDENT: Identifying differential equations with numerical time evolution*, J. Sci. Comput., 87 (2021), pp. 1–27.
[18] R. S. Lakes, *Viscoelastic Solids*, CRC Press, 1998.
[19] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, *Neural Operator: Graph Kernel Network for Partial Differential Equations*, preprint, arXiv:2003.03485, 2020.
[20] B. Liu, N. Kovachki, Z. Li, K. Azizzadenesheli, A. Anandkumar, A. M. Stuart, and K. Bhattacharya, *A learning-based multiscale method and its application to inelastic impact problems*, J. Mech. Phys. Solids, 158 (2022), 104668.
[21] B. Liu, M. Trautner, A. M. Stuart, and K. Bhattacharya, *Learning Macroscopic Internal Variables and History Dependence from Microscopic Models*, preprint, arXiv:2210.17443, 2022.
[22] Z. Liu, C. Wu, and M. Koishi, *A deep material network for multiscale topology learning and accelerated nonlinear modeling of heterogeneous materials*, Comput. Methods Appl. Mech. Engrg., 345 (2019), pp. 1138–1168.
[23] G. W. Milton, *The Theory of Composites*, Cambridge University Press, 2002.
[24] N. Mishra, J. Vondrejc, and J. Zeman, *A comparative study on low-memory iterative solvers for FFT-based homogenization of periodic media*, J. Comput. Phys., 321 (2016), pp. 151–168.
[25] H. Moulinec, P. Suqeut, and G. Milton, *Convergence of iterative methods based on Neumann series for composite materials: Theory and practice*, Internat. J. Numer. Methods Engrg., 114 (2018), pp. 1103–1130.
[26] H. Moulinec and P. Suquet, *A numerical method for computing the overall response of nonlinear composites with complex microstructure*, Comput. Method Methods Appl. Mech. Engrg., 157 (1998), pp. 69–94.
[27] M. Mozaffar, R. Bostanabad, W. Chen, K. Ehmann, J. Cao, and M. Bessa, *Deep learning predicts path-dependent plasticity*, Proc. Natl. Acad. Sci. USA, 116 (2019), pp. 26414–26420.

[28] G. Pavliotis and A. Stuart, *Multiscale Methods: Averaging and Homogenization*, Springer Science & Business Media, 2008.
[29] B. Peherstorfer and K. Willcox, *Data-driven operator inference for nonintrusive projection-based model reduction*, Comput. Methods Appl. Mech. Engrg., 306 (2016), pp. 196–215.
[30] R. Phillips and P. Rob, *Crystals, Defects and Microstructures: Modeling Across Scales*, Cambridge University Press, 2001.
[31] E. Qian, B. Kramer, B. Peherstorfer, and K. Willcox, *Lift & learn: Physics-informed machine learning for large-scale nonlinear dynamical systems*, Phys. D, 406 (2020), 132401.
[32] J. R. Rice, *Inelastic constitutive relations for solids: An internal-variable theory and its application to metal plasticity*, J. Mech. Phys. Solids, 19 (1971), pp. 433–455.
[33] E. Sánchez-Palencia, *Non-homogeneous Media and Vibration Theory*, Lecture Notes in Phys. 127, Springer, 1980.
[34] E. Sanchez-Palencia and A. Zaoui, eds., *Homogenization Techniques for Composite Media*, Lecture Notes in Phys. 272, Springer, 1987.
[35] J. C. Simo, *Numerical analysis and simulation of plasticity*, Handb. Numer. Anal., 6 (1998), pp. 183–499.
[36] J. C. Simo and T. J. Hughes, *Computational Inelasticity*, Interdiscip. Appl. Math. 7, Springer Science & Business Media, 2006.
[37] A. Spencer, *Continuum Mechanics*, Longman Group U.K., Essex, 1980.
[38] L. Sun, H. Gao, S. Pan, and J.-X. Wang, *Surrogate modeling for fluid flows based on physics-constrained deep learning without simulation data*, Comput. Method Appl. Mech. Engrg., 361 (2020), 112732.
[39] E. B. Tadmor, *Modeling Materials: Continuum, Atomistic and Multiscale Techniques*, Cambridge University Press, 2012.
[40] L. Tartar, *Memory effects and homogenization*, Arch. Ration. Mech. Anal., 111 (1990), pp. 121–133.
[41] E. Van Der Giessen, P. A. Schultz, N. Bertin, V. V. Bulatov, W. Cai, G. Csányi, S. M. Foiles, M. G. Geers, C. González, M. Hütter, et al., *Roadmap on multiscale materials modeling*, Model. Simul. Mater. Sci. Eng., 28 (2020), 043001.
[42] L. Wu, N. Kilingar, and L. Noels, *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. Engrg., 369 (2020), 113234.
[43] T. I. Zohdi and P. Wriggers, *Introduction to Computational Micromechanics*, Springer-Verlag, 2005.