Data-driven Reduced Homogenization for Transient Diffusion Problems with Emergent History Effects

Abdullah Waseem, Thomas Heuzé, Marc G.D. Geers, Varvara G Kouznetsova, Laurent Stainier

▶ **To cite this version:**

Abdullah Waseem, Thomas Heuzé, Marc G.D. Geers, Varvara G Kouznetsova, Laurent Stainier. Data-driven Reduced Homogenization for Transient Diffusion Problems with Emergent History Effects. Computer Methods in Applied Mechanics and Engineering, 2021, 380, pp.113773. 10.1016/j.cma.2021.113773 . hal-03189901

HAL Id: hal-03189901
https://hal.science/hal-03189901

Submitted on 5 Apr 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Data-driven Reduced Homogenization for Transient Diffusion Problems with Emergent History Effects

Abdullah Waseem[1,2], Thomas Heuzé[1], Marc G.D. Geers[2], Varvara G. Kouznetsova[2], and Laurent Stainier*[1]

[1]Institut de Recherche en Génie Civil et Mécanique, GeM, UMR 6183 - CNRS - École Centrale de Nantes-Université de Nantes, 44321 Nantes, France
[2]Department of Mechanical Engineering, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands

Abstract

In this paper, we propose a data-driven reduced homogenization technique to capture diffusional phenomena in heterogeneous materials which reveal, on a macroscopic level, a history-dependent non-Fickian behavior. The adopted enriched-continuum formulation, in which the macroscopic history-dependent transient effects are due to the underlying heterogeneous microstructure is represented by enrichment-variables that are obtained by a model reduction at the micro-scale. The data-driven reduced homogenization minimizes the distance between points lying in a data-set and points associated with the macroscopic state of the material. The enrichment-variables are excellent pointers for the selection of the correct part of the data-set for problems with a time-dependent material state. Proof-of-principle simulations are carried out with a heterogeneous linear material exhibiting a relaxed separation of scales. Information obtained from simulations carried out at the micro-scale on a unit-cell is used to determine approximate values of metric coefficients in the distance function. The proposed data-driven reduced homogenization also performs adequately in the case of noisy data-sets. Finally, the possible extensions to non-linear history-dependent behavior are discussed.

Keywords Data-Driven Mechanics · Computational Homogenization · Model Order Reduction · Non-Fickian Diffusion ·

1 Introduction

Transient mass diffusion phenomena in heterogeneous materials are prevalent in engineering applications, for example, Lithium-ion batteries [1], polycrystalline materials [2], diffusion in porous gels [3], etc. For their analysis, numerical methods like finite elements in conjunction with transient computational homogenization [4, 5] are used. Computational homogenization represents the heterogeneous domain by a homogeneous macro-scale and a heterogeneous micro-scale and solves the transient diffusion phenomena in a coupled two-scale setting. Despite the

*Corresponding author: Laurent.Stainier@ec-nantes.fr

fact that the individual micro-scale constituents might reveal instantaneously linear behavior, the homogenization of transient mass diffusion phenomena in heterogeneous materials provides an emergent non-Fickian diffusion behavior [6, 7]. This lagging and history-dependent diffusion behavior obtained at the macro-scale is due to the transient nature of the mass diffusion occurring at the micro-scale. Diffusion in a heterogeneous material, consisting of inclusions embedded in a matrix material, can be characterized by a characteristic loading time T, a characteristic diffusion time for the inclusions $\tau_i = d^2/\mathcal{D}_i$ and a characteristic diffusion time for the matrix $\tau_m = \ell^2/\mathcal{D}_m$, where d is the inclusion size (e.g. diameter), ℓ is the characteristic size of the representative microscopic domain (e.g. the unit-cell), $\mathcal{D}_i$ and $\mathcal{D}_m$ are the diffusivity constants for the inclusions and the matrix, respectively. In the regime of the relaxed separation of scales ($\tau_m \ll \tau_i \sim T$), the non-Fickian behavior at the macro-scale is due to the slow diffusion inside the inclusions. This gives rise to a lagging behavior at the macro-scale, which is more prominent in the macroscopic storage term than in the diffusion term.

The homogenization in transient regimes is generally computationally very expensive. In previous work [8], a model reduction technique, based on component mode synthesis [9, 10] was developed for transient diffusion phenomena in heterogeneous materials with linear material properties in the relaxed separation of scales regime. The microscopic primary field variable was decomposed into a steady-state and a transient part. Model reduction was achieved by solving an eigenvalue problem and selecting only a few eigenvectors in the reduced bases set. When projected onto the reduced bases subspace, the discretized mass balance equation at the micro-scale provides a set of ordinary differential equations in terms of the activity coefficients of the eigenvectors. At the macro-scale, the macroscopic initial boundary value problem, the ordinary differential equations of the activity coefficients, and the effective, homogenized, constitutive equations entail an enriched-continuum description, where the activity coefficients are the emerging enrichment-variables. These enrichment-variables can be treated as separate fields or as internal-variables, as used in the constitutive theories involving internal variables [11]. This method was later extended to coupled diffusion-mechanics phenomena [12], where diffusion induced stresses were correctly captured with the reduced method. The major limitation of the enriched-continuum formulation is the fact that it relies on the linear material models at the micro-scale to obtain the well-defined eigenvector reduced bases by solving an eigenvalue problem.

In this work, a further extension of the enriched-continuum formulation is proposed, which we call *data-driven reduced homogenization*. It combines the model reduction at the micro-scale, to obtain enrichment-variables playing the role of internal-variables, and the data-driven framework, which was first proposed in [13].

The data-driven computing [13] eliminates the need for a material model in computer simulations and instead directly uses raw data obtained from e.g. experiments or micro-scale simulations. In essence, the data-driven method tries to find a point in the data-set closest to the physical-state of the material obeying compatibility and the balance laws (or vice-versa). It was further extended to noisy data-sets [14], dynamics [15] and inelastic material behavior [16]. Data-driven computational homogenization, was proposed in [17, 18], where the expensive micro-scale calculations were performed first to generate data containing homogenized quantities in an off-line stage, while in an on-line stage, the homogenized macro-scale problem was solved using the data-driven approach. It was also showed that the search through the data-set is much more efficient than solving micro-scale problems in a coupled manner.

The data-driven approach proposed in [13] is fundamentally different from other data-driven methods used in mechanics, where the data is typically used to learn the behavior of the material

in terms of a stress-strain relationship or an energy potential. Classically, this learning process involves linear/non-linear regression through, experimentally collected, data points to build a model. The regression analysis has recently been replaced by techniques such as artificial neural networks, deep learning, etc. borrowed from the field of data science. For example, [19] employed a data-driven method that uses artificial neural networks to obtain a decoupled and efficient computational homogenization for non-linear elastic materials by approximating a density energy function. For a data-set with few points, a data-driven inverse problem was proposed in [20] to recover the entire constitutive manifold. Notably, [21] developed a multi-scale data-driven method using recurrent neural networks, which can capture the history-dependent behavior for plasticity and replaces the micro-scale calculations with a surrogate model. Detailed reviews for modern data-driven model building techniques can be found in [22, 23, 24, 25].

The data-driven reduced homogenization, that will be proposed in this work, entails three stages, i.e. (1) model reduction, (2) data-generation, and (3) data-search.

1. The model reduction at the micro-scale, depending on the material models of the constituents, can be applied to the discrete mass balance equations. It can be categorized as a pre-processing stage. In the context of data-driven reduced homogenization, the central goals of performing model reduction at the micro-scale are to be able to (i) solve a large number of micro-scale problems in a computationally efficient manner during the data-generation stage and (ii) obtain internal-variables to represent the effect of the micro-scale transient behavior at the macro-scale. For the materials with memory, the internal-variables approach provides a computationally efficient way to keep track of the history dependence [16], hence, easing the computational efforts later in the data-search stage.

2. The data-generation stage involves the solution of many micro-scale problems, post-processing, and storage of the results in the form of the macroscopic conjugate quantities. The data-generation stage is typically an off-line stage. To ensure that the data-set contains representative values of conjugate quantities involved in the problem, the micro-scale should be probed under different loading conditions with different frequencies.

3. Finally, the data-search is carried out to find an optimum point that reflects the minimal distance from the current physical-state of the material, satisfying balance equations.

In this work, following [13, 16, 26], a staggered distance-minimizing data-driven solver is adopted. It iteratively minimizes a quadratic distance function, defined on the material phase-space, while looking for a point in the data-set. The compatibility of the macroscopic primary field is enforced directly and the macroscopic balance law is enforced with the help of Lagrange multipliers. To find the physical-state of the material, the stationarity conditions are obtained and then solved by taking all possible variations of the Lagrangian function. Then, the search through the data-set is performed by an array indexing lookup operation. The data-search stage constitutes the on-line stage. The material models (here linear) are known at the micro-scale and the data-driven approach is applied to the macro-scale only. The formulation has to be adapted to the dimension and structure of the phase space. In particular, the history-dependence typically entails very high dimension phase spaces, which can be addressed through various strategies (see [16]). But the structure of the data-driven solver itself remains quite transparent also in case of non-linearity. In this preliminary work, the analysis is limited to the linear material behavior, and the results are compared with the reference

enriched-continuum formulation [8], while the data-driven reduced homogenization approach for non-linear materials will be analyzed in future work. The novel contributions in this paper are:

- introduction of *data-driven reduced homogenization* for macroscopic history-dependent linear diffusion behavior;
- proposing a methodology to evaluate optimal numerical values of the coefficients in the distance function based on the information from micro-scale simulations.

The paper is organized as follows: The diffusion enriched-continuum formulation and model reduction for linear materials are briefly presented in Section 2. The data-driven reduced homogenization is derived in Section 3: first, the data-set and phase-space are defined; next the solution procedure and the algorithm is elaborated for a distance minimizing data-driven solver. All stages of the data-driven reduced homogenization are evaluated with numerical examples in Section 4, where after setting up the micro-scale and macro-scale problems, the data-generation is performed by micro-scale simulations. An important discussion is made on the selection of numerical values of the coefficients in the distance function and the performance and convergence of the proposed method is assessed with a noisy data-set. Future perspectives, along with an outlook to extend the proposed data-driven reduced homogenization method to non-linear history-dependent diffusion materials, are presented in Section 5 and finally the conclusions are given in Section 6.

2 Enriched Continuum for Diffusion Problems

Assuming that the micro-scale material properties and microstructural topology are known, the non-Fickian behavior at the macro-scale can be captured through a multi-scale approach such as transient computational homogenization [4, 5, 7]. For diffusion problems, the macroscopic behavior, in terms of the macroscopic chemical potential $\bar{\mu}$ as the primary unknown field, is obtained by solving a macroscopic transient mass balance equation

$$\begin{aligned} \overline{\boldsymbol{\nabla}} \cdot \bar{\boldsymbol{j}} + \dot{\bar{c}} = 0\,, \qquad & \text{in} \qquad \overline{\Omega}\,, \\ \bar{\mu}(t=0) = \bar{\mu}_0\,, \qquad & \text{in} \qquad \overline{\Omega}\,, \\ \bar{\mu} = \hat{\bar{\mu}}\,, \qquad & \text{on} \qquad \partial\overline{\Omega}_{\bar{\mu}}\,, \\ -\bar{\boldsymbol{j}} \cdot \overline{\boldsymbol{n}} = \hat{\bar{j}}\,, \qquad & \text{on} \qquad \partial\overline{\Omega}_{\bar{j}}\,. \end{aligned} \tag{1}$$

where $\partial\overline{\Omega}_{\bar{\mu}}$ and $\partial\overline{\Omega}_{\bar{j}}$ are the Dirichlet and Neumann sub-parts of the macroscopic boundary $\partial\overline{\Omega}$, respectively, and $\overline{\boldsymbol{n}}$ is the outward unit-normal vector, $\hat{\bar{j}}$ is the prescribed mass influx and $\hat{\bar{\mu}}$ is the prescribed macroscopic chemical potential. The explicit expressions for the macroscopic constitutive equations of the macroscopic mass flux $\bar{\boldsymbol{j}}$ and the rate of change of macroscopic concentration $\dot{\bar{c}}$ are not known and these are to be determined through homogenization, based on the micro-scale material behavior and morphological information.

The micro-scale problem is described by the balance equation

$$\boldsymbol{\nabla} \cdot \boldsymbol{j} + \dot{c} = 0 \tag{2}$$

with the known constitutive equations given by

$$\boldsymbol{j} = -\boldsymbol{M} \cdot \boldsymbol{g}, \qquad \text{where } \boldsymbol{g} = \boldsymbol{\nabla}\mu \text{ and } \mu = \Lambda(c - c_0)\,, \tag{3}$$

with $\boldsymbol{M}$ the mobility tensor, Λ the chemical modulus and c_0 the reference concentration. The material properties are assumed to be known for each micro-structural constituent. Transient computational homogenization involves down-scaling and up-scaling steps: the former consists in imposing the governing macroscopic quantities $(\bar{\mu}, \bar{\boldsymbol{g}})$, with $\bar{\boldsymbol{g}} = \overline{\boldsymbol{\nabla}}\bar{\mu}$, on the micro-scale domain, and the latter involves the computation of the effective conjugate quantities $(\bar{\boldsymbol{j}}, \dot{\bar{c}})$, after solving the transient fully resolved micro-scale problem which can be computationally rather expensive.

In the relaxed separation of scales regime (requiring that the characteristic diffusion time of the matrix τ_m is much smaller than that of the inclusion τ_i which is of the same order of magnitude as the characteristic loading time T i.e. $\tau_m \ll \tau_i \sim T$), for micro-scale constituents with a linear material behavior, a reduced model for transient heat conduction has been proposed in [8]. Whereby the computationally expensive solution of the transient micro-scale problem is replaced by an inexpensive solution of a set of ordinary differential equations at the macro-scale by using computational homogenization along with component mode synthesis. A similar approach to [8] can be adopted for transient mass diffusion problems in heterogeneous materials. First, discretized (e.g using FEM) the microscopic chemical potential field $\underset{\sim}{\mu}$ can be decomposed into its steady-state $\underset{\sim}{\mu}_{ss}$ and transient $\underset{\sim}{\mu}_{tr}$ parts. Next, an eigenvalue problem is solved at the micro-scale to obtain the reduced eigenmodes $\underset{\sim}{\Phi}^{(q)}$, where $q = 1, 2, ..., \mathcal{N}_q$, with $\mathcal{N}_q$ are the reduced number of eigenvectors. Finally, the microscopic discretized problem is projected onto the subspace of the reduced eigenbasis yielding a decoupled system of first-order ordinary differential equations

$$\dot{\underset{\sim}{\eta}} + \underline{\alpha}\,\underset{\sim}{\eta} = -\overset{*}{\underset{\sim}{\bar{d}}}\dot{\bar{\mu}} - \overset{*}{\underset{\sim}{\bar{\boldsymbol{a}}}} \cdot \dot{\bar{\boldsymbol{g}}}\,. \tag{4}$$

where $\underset{\sim}{\eta}$ is the column of the modal amplitudes, having the meaning of activity coefficients or reduced degrees of freedom $\eta^{(q)}$, $\underline{\alpha}$ is the diagonal matrix of eigenvalues $\alpha^{(q)}$, and $\overset{*}{\bar{\boldsymbol{a}}}^{(q)}$ and $\overset{*}{\bar{d}}^{(q)}$ are the coefficients that couple the micro-scale to the macro-scale. Projection onto the reduced degrees of freedom, also provides the expression for the macroscopic constitutive equations of the macroscopic flux

$$\bar{\boldsymbol{j}} = -\underset{\sim}{\bar{\boldsymbol{a}}}^T \dot{\underset{\sim}{\eta}} - \overline{\boldsymbol{B}} \cdot \bar{\boldsymbol{g}} - \bar{\boldsymbol{c}}\dot{\bar{\mu}} - \overline{\boldsymbol{C}} \cdot \dot{\bar{\boldsymbol{g}}}\,, \tag{5}$$

and the rate of change of the macroscopic concentration

$$\dot{\bar{c}} = \underset{\sim}{\bar{d}}^T \dot{\underset{\sim}{\eta}} + \bar{\boldsymbol{e}} \cdot \bar{\boldsymbol{g}} + \bar{f}\dot{\bar{\mu}} + \overline{\boldsymbol{f}} \cdot \dot{\bar{\boldsymbol{g}}}\,. \tag{6}$$

At the macro-scale, equations (1), (4), (5) and (6) present a diffusion enriched-continuum with $\underset{\sim}{\eta}$ as the column of enrichment-variables. The effective coefficients $(\underset{\sim}{\bar{\boldsymbol{a}}}, \overline{\boldsymbol{B}}, \bar{\boldsymbol{c}}, \overline{\boldsymbol{C}})$ and $(\underset{\sim}{\bar{d}}, \bar{\boldsymbol{e}}, \bar{f}, \overline{\boldsymbol{f}})$ are the linear maps between the macroscopic quantities $(\dot{\underset{\sim}{\eta}}, \bar{\boldsymbol{g}}, \dot{\bar{\mu}}, \dot{\bar{\boldsymbol{g}}})$ and $(\bar{\boldsymbol{j}}, \dot{\bar{c}})$, respectively. Their magnitudes and directions depend on the microstructural material properties and microstructural morphology. The reader is referred to [8] for a detailed derivation of the enriched-continuum formulation in the context of transient heat conduction, and to reference [27] for the expressions of the effective coefficients used in equations (4)–(6) and the numerical implementation of the enriched-continuum for mass diffusion problems. The enriched-continuum formulation has also been extended to transient mass diffusion problems coupled to mechanics in [12].

The non-Fickian diffusion at the macro-scale, represented by $\bar{\boldsymbol{j}} = \bar{\boldsymbol{j}}(\dot{\underset{\sim}{\eta}}, \bar{\boldsymbol{g}}, \dot{\bar{\mu}}, \dot{\bar{\boldsymbol{g}}})$ in equation (5) and $\dot{\bar{c}} = \dot{\bar{c}}(\dot{\underset{\sim}{\eta}}, \bar{\boldsymbol{g}}, \dot{\bar{\mu}}, \dot{\bar{\boldsymbol{g}}})$ in equation (6), allows to capture the complex history dependence. The enrichment-variables $\underset{\sim}{\eta}$ play similar role to the internal-variables used in the constitutive theory

of inelastic materials [11]. However, here the macroscopic model is non-classical one since the storage terms $\dot{\bar{c}}$ also depend on the internal-variables, which is usually not the case for inelastic materials. In the data-driven approach, $\underset{\sim}{\eta}$ can serve as an indicator in time for the selection of the conjugate quantities, hence capturing the history-dependent behavior efficiently. On the other hand, if instead of an enriched-continuum, the standard transient computational homogenization scheme [4] would be used for the data-driven approach, there would be no internal-variables $\underset{\sim}{\eta}$ at the macro-scale. Instead, it would require the storage of the complete history of the discrete microscopic fields $\underset{\sim}{\mu}(t)$, for the corresponding macroscopic quantities $(\bar{\mu}, \bar{c}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{j}})$ together with the data. The data-search stage would then consist of searching through the entire history of the discrete microscopic fields $\underset{\sim}{\mu}$ up to a given time t. This would consume an enormous amount of computer resources for data-generation, data-storage, and data-search. Hence, the extraction of an enrichment-variable like quantity through the model reduction at the micro-scale is a crucial step towards an efficient data-driven solver for transient diffusion problems in heterogeneous materials. In the next section, the data-driven homogenization for transient diffusion problems with history effects is formally derived.

3 Data-Driven Reduced Homogenization

In this section, the data-driven reduced homogenization is derived for transient diffusion problems with history effects at the macro-scale. First, the notions of data-set and phase-space are presented. Then, a specific class of data-driven solver, i.e. a distance minimizing data-driven solver, is chosen for the current implementation, which results in a double minimization of the distance function. Next, the solution procedure using a staggered scheme is presented and finally, each step involved of the data-driven simulation algorithm.

For the sake of simplicity, a temporally and spatially discrete macroscopic problem is considered. The time is discretized by backward-Euler scheme, in which a rate term $\dot{\mathcal{F}}$ can be approximated by

$$\dot{\mathcal{F}} \approx \frac{\mathcal{F}^{n+1} - \mathcal{F}^{n}}{\Delta t} \tag{7}$$

where $\Delta t = t^{n+1} - t^{n}$ is the time increment between the current t^{n+1} and previous t^{n} time instance. Spatial discretization of the domain $\overline{\Omega}$ is performed by finite elements containing $m = 1, 2, ..., \mathcal{M}$ material (integration) points and $i = 1, 2, ..., \mathcal{N}$ nodes. The notations and terminologies adopted here, follow from [16, 28].

3.1 Data-Set and Phase-Space

The data-driven reduced homogenization relies on a data-set generated by micro-scale reduced-order simulations. The choice of the macroscopic quantities to be stored in the data-set depends on the expressions of the constitutive equations (4)–(6) and the data extracted from the micro-scale simulations. Here, we choose to store all quantities and their rates (except for the flux rate). The local data-set

$$D_m = \left\{ \left(\bar{\mu}'_m, \dot{\bar{\mu}}'_m, \bar{\boldsymbol{g}}'_m, \dot{\bar{\boldsymbol{g}}}'_m, \underset{\sim}{\eta}'_m, \dot{\underset{\sim}{\eta}}'_m, \bar{\boldsymbol{j}}'_m, \bar{c}'_m, \dot{\bar{c}}'_m \right) \right\}_{I=1}^{n_{dp}}, \tag{8}$$

is available at each material point m of the macroscopic discrete model. In present work, the data-set D_m is constant in time, examples of temporally evolving data-sets i.e. D_m^{n+1} can

be found in [16]. In (8) the prime $\bullet'$ denotes a quantity that belongs to the data-set D_m. Without any prime the quantity belongs to the physical state. I represents a data-point post-processed at each time-step from the micro-scale simulations and n_{dp} is the total number of data-points. It should be noted that after the data-generation stage the homogenization model is disregarded, and thereafter data-driven problems solely rely on the data at hand collected in (8). The local data-sets D_m, in general, can be different for each macroscopic material point m. Collectively, from all the material points, the total number of data-sets available in a discrete system represent a global data-set

$$D = D_1 \times D_2 \times ... \times D_{\mathcal{M}} \,. \tag{9}$$

The physical-state of the material at the homogenized macro-scale can be characterized by a point in the local phase-space Z_m^{n+1}

$$z_m^{n+1} = \left(\bar{\mu}_m, \dot{\bar{\mu}}_m, \bar{\boldsymbol{g}}_m, \dot{\bar{\boldsymbol{g}}}_m, \underset{\sim}{\eta}_m, \dot{\underset{\sim}{\eta}}_m, \bar{\boldsymbol{j}}_m, \bar{c}_m, \dot{\bar{c}}_m\right)^{n+1} \in Z_m^{n+1} \,, \tag{10}$$

evolving in time, whereby the dimensions of the phase-space are $\dim(Z_m^{n+1}) = \mathbb{R}^1 \times \mathbb{R}^1 \times \mathbb{R}^{sd} \times \mathbb{R}^{sd} \times \mathbb{R}^{\mathcal{N}_q} \times \mathbb{R}^{\mathcal{N}_q} \times \mathbb{R}^{sd} \times \mathbb{R}^1 \times \mathbb{R}^1$, in which sd is the spatial dimension of the problem under consideration. Once combined, all the local states z_m^{n+1} make up the global physical-state

$$z^{n+1} = \left\{ \left(\bar{\mu}_m, \dot{\bar{\mu}}_m, \bar{\boldsymbol{g}}_m, \dot{\bar{\boldsymbol{g}}}_m, \underset{\sim}{\eta}_m, \dot{\underset{\sim}{\eta}}_m, \bar{\boldsymbol{j}}_m, \bar{c}_m, \dot{\bar{c}}_m\right)^{n+1} \right\}_{m=1}^{\mathcal{M}} \in Z^{n+1} \,, \tag{11}$$

in the global phase-space $Z^{n+1} = Z_1^{n+1} \times Z_2^{n+1} \times ... \times Z_{\mathcal{M}}^{n+1}$. The physical-state of the material z^{n+1} at the macro-scale should obey the macroscopic compatibility and the discrete macroscopic mass balance laws at each time instance t^{n+1}.

For each material point, in an element of the discretized macroscopic domain, the compatibility is expressed in terms of the discretized macroscopic chemical potential, defined at the nodes $\bar{\mu}_i^{n+1}$ as

$$\bar{\boldsymbol{g}}_m^{n+1} = \sum_{i=1}^{\mathcal{N}} \boldsymbol{\nabla} N_{mi} \bar{\mu}_i^{n+1} \,. \tag{12}$$

The macroscopic mass balance (1), once discretized in space and time and after applying the Dirichlet and the Neumann boundary conditions, reads at each node i and time instance t^{n+1} as

$$-\Delta t \sum_{m=1}^{\mathcal{M}} w_m \boldsymbol{\nabla} N_{mi} \cdot \bar{\boldsymbol{j}}_m^{n+1} + \sum_{m=1}^{\mathcal{M}} w_m N_{mi} (\bar{c}_m^{n+1} - \bar{c}_m^n) = -\Delta t \hat{\bar{j}}_i^{n+1} \qquad \text{where } i = 1, 2, ..., \mathcal{N} \,. \tag{13}$$

In equations (12) and (13), w_m contains information regarding quadrature weights and the volume of the elements, $\boldsymbol{\nabla} N_{mi}$ is the gradient of finite element shape function N_i evaluated at integration point m. Here, for the sake of simplicity of the notation, the finite element shape functions N_i and their gradients $\boldsymbol{\nabla} N_i$, are defined globally on the whole finite element mesh. Here, it should be noted that the macroscopic flux $\bar{\boldsymbol{j}}_m$ as well as the concentration $\bar{c}_m$, in equation (13), are the constitutive quantities which are evaluated at the material points. The primary unknown field is $\bar{\mu}_i$ defined on the nodes. This is different from the typical finite element discretization of the mass balance equation in which usually the concentration field are assumed primary unknown field. Note also that similar expressions could be obtained by alternative

spatial discretization techniques. The compatibility (12) and the terms in the macroscopic mass balance law (13) are coupled through the data-set D_m. Through this coupling the discretized macroscopic chemical potential $\bar{\mu}_i^{n+1}$ is solved at the nodes. The compatibility (12) and the balance law (13) pose restrictions on the state z^{n+1} of the material, hence constraining the phase-space Z^{n+1} as

$$E^{n+1} = \{z^{n+1} \in Z^{n+1} : \text{compatibility (12) and macroscopic mass balance (13)}\}\,. \tag{14}$$

3.2 Distance Minimizing Data-Driven Problem

A distance minimizing data-driven problem, as introduced in [13], seeks a compatible and equilibrated material physical-state $z^{n+1} \in E^{n+1}$ that has a minimum distance to a point in the global data-set D. To work with a distance, first the local phase-space Z_m^{n+1} is equipped with a local norm

$$|z_m^{n+1}| = \Big[\tfrac{1}{2}\,{}^1\bar{\mathcal{C}}_m(\bar{\mu}_m^{n+1})^2 + \tfrac{1}{2}\,{}^2\bar{\mathcal{C}}_m(\dot{\bar{\mu}}_m^{n+1})^2 + \tfrac{1}{2}\,{}^3\bar{\mathcal{C}}_m(\bar{\boldsymbol{g}}_m^{n+1})^2 + \tfrac{1}{2}\,{}^4\bar{\mathcal{C}}_m(\dot{\bar{\boldsymbol{g}}}_m^{n+1})^2 + \tfrac{1}{2}(\underset{\sim}{\eta}_m^{n+1})^T\,{}^5\underline{\mathcal{C}}_m(\underset{\sim}{\eta}_m^{n+1}) + \tfrac{1}{2}(\dot{\underset{\sim}{\eta}}_m^{n+1})^T\,{}^6\underline{\mathcal{C}}_m(\dot{\underset{\sim}{\eta}}_m^{n+1}) + \tfrac{1}{2}\,{}^7\bar{\mathcal{C}}_m(\bar{\boldsymbol{\jmath}}_m^{n+1})^2 + \tfrac{1}{2}\,{}^8\bar{\mathcal{C}}_m(\bar{c}_m^{n+1})^2 + \tfrac{1}{2}\,{}^9\bar{\mathcal{C}}_m(\dot{\bar{c}}_m^{n+1})^2\Big]^{\frac{1}{2}}, \tag{15}$$

where ${}^J\bar{\mathcal{C}}_m$ with $J = 1, 2, ..., 9$ are the coefficients which non-dimensionalize the measure (15) and do not represent any material property. The coefficients ${}^5\underline{\mathcal{C}}_m$ and ${}^6\underline{\mathcal{C}}_m$ are diagonal matrices of size $\mathcal{N}_q \times \mathcal{N}_q$. The numerical values of these coefficients are important for the numerical convergence of the data-driven problem and will be discussed in Section 4.4. Each term in the measure (15) is quadratic, which, under the linear constraints of compatibility and equilibrium, leads to a convex optimization problem. Then, locally, at the material point level, the distance between two points $y_m^{n+1}, z_m^{n+1} \in Z_m^{n+1}$ can be measured as

$$d_m(z_m^{n+1}, y_m^{n+1}) = |z_m^{n+1} - y_m^{n+1}|\,. \tag{16}$$

The global norm can be obtained by taking squares and integrating the local norms over the entire domain

$$|z^{n+1}|_g = \left(\sum_{m=1}^{\mathcal{M}} w_m |z_m^{n+1}|^2\right)^{\frac{1}{2}}, \tag{17}$$

which metrizes the global phase-space Z. Consequently the global distance from a point $y^{n+1} \in Z^{n+1}$ to $z^{n+1} \in Z^{n+1}$ is measured as

$$d(z^{n+1}, y^{n+1}) = |z^{n+1} - y^{n+1}|_g\,. \tag{18}$$

The distance minimizing data-driven problem is then written as a double minimization

$$\min_{y^{n+1}\in D}\ \min_{z^{n+1}\in E^{n+1}}\ d(z^{n+1}, y^{n+1}) = \min_{z^{n+1}\in E^{n+1}}\ \min_{y^{n+1}\in D}\ d(z^{n+1}, y^{n+1})\,. \tag{19}$$

It aims to find a point y^{n+1} in the global data-set D which is closest to a compatible and equilibrated material state E^{n+1}, or equivalently, find a compatible and equilibrated material state E^{n+1} which is closest to a point y^{n+1} in global data-set D while both minimizing the global distance function $d(z^{n+1}, y^{n+1})$.

The double minimization problem (19) is a combination of continuous and discrete optimization problems, the former over the continuous manifold E^{n+1}, the latter in the discrete data-set D. It has a combinatorial complexity, since for each material point m contributing to the global distance-function (18), n_{dp} points can be evaluated and the minimum should be chosen among those. To efficiently solve this computationally intensive combinatorial problem, following [28, 26], a staggered solution scheme is adopted here which freezes the continuous minimization problem while solving the discrete one and vice-versa. It assumes at an iteration k the optimum point in the data-set $\overset{*}{y}{}_k^{n+1} \in D$ to be known and finds a closest state $z_{k+1}^{n+1} \in E^{n+1}$ to that data-set point. This first step represents a projection operation $z_{k+1}^{n+1} = P_{E^{n+1}} \overset{*}{y}{}_k^{n+1}$, where $P_{E^{n+1}}$ denotes the closest point projection from D onto E^{n+1}.

Subsequently, in turn, the point z_{k+1}^{n+1} can be used to find the closest point in the data-set for the next iteration $\overset{*}{y}{}_{k+1}^{n+1} = P_D z_{k+1}^{n+1}$, where P_D denotes the closest point projection from Z^{n+1} onto D. The iterations are continued until there is no other optimum point in the data-set to choose i.e. $P_D z_{k+1}^{n+1} = P_D z_k^{n+1}$.

3.3 Solution Procedure

Assuming a known minimizing point $\overset{*}{y}{}_k^{n+1} \in D$, the projection $z_{k+1}^{n+1} = P_{E^{n+1}} \overset{*}{y}{}_k^{n+1}$ is followed after minimizing the quadratic distance function $d^2(\bullet, \overset{*}{y}{}_k^{n+1})$ subject to the constraints (12) and (13). The compatibility is imposed directly by introducing the chemical potential field as in equation (12) and the discrete mass balance is enforced by using Lagrange multipliers $\bar{\lambda}_i^{n+1}$ at the nodes. The discrete Lagrangian can be written as

$$
\begin{aligned}
\bar{\mathcal{L}}^{n+1} = \sum_{m=1}^{\mathcal{M}} w_m \Bigg[& \tfrac{1}{2}\, {}^1\bar{\mathcal{C}}_m \left(\sum_{i=1}^{\mathcal{N}} N_{mi} \bar{\mu}_i^{n+1} - \overset{*}{\bar{\mu}}{}_m^{n+1} \right)^2 + \tfrac{1}{2}\, {}^2\bar{\mathcal{C}}_m \left(\sum_{i=1}^{\mathcal{N}} N_{mi} \frac{\bar{\mu}_i^{n+1} - \bar{\mu}_i^n}{\Delta t} - \overset{*}{\dot{\bar{\mu}}}{}_m^{n+1} \right)^2 + \\
& \tfrac{1}{2}\, {}^3\bar{\mathcal{C}}_m \left(\sum_{i=1}^{\mathcal{N}} \boldsymbol{\nabla} N_{mi} \bar{\mu}_i^{n+1} - \overset{*}{\bar{\boldsymbol{g}}}{}_m^{n+1} \right)^2 + \tfrac{1}{2}\, {}^4\bar{\mathcal{C}}_m \left(\sum_{i=1}^{\mathcal{N}} \boldsymbol{\nabla} N_{mi} \frac{\bar{\mu}_i^{n+1} - \bar{\mu}_i^n}{\Delta t} - \overset{*}{\dot{\bar{\boldsymbol{g}}}}{}_m^{n+1} \right)^2 + \\
& \tfrac{1}{2}\, {}^5\underset{\sim}{\mathcal{C}}_m \left(\underset{\sim}{\eta}_m^{n+1} - \overset{*}{\underset{\sim}{\eta}}{}_m^{n+1} \right)^2 + \tfrac{1}{2}\, {}^6\underset{\sim}{\mathcal{C}}_m \left(\frac{\underset{\sim}{\eta}_m^{n+1} - \underset{\sim}{\eta}_m^n}{\Delta t} - \overset{*}{\dot{\underset{\sim}{\eta}}}{}_m^{n+1} \right)^2 + \\
& \tfrac{1}{2}\, {}^7\bar{\mathcal{C}}_m \left(\bar{\boldsymbol{j}}_m^{n+1} - \overset{*}{\bar{\boldsymbol{j}}}{}_m^{n+1} \right)^2 + \tfrac{1}{2}\, {}^8\bar{\mathcal{C}}_m \left(\bar{c}_m^{n+1} - \overset{*}{\bar{c}}{}_m^{n+1} \right)^2 + \tfrac{1}{2}\, {}^9\bar{\mathcal{C}}_m \left(\frac{\bar{c}_m^{n+1} - \bar{c}_m^n}{\Delta t} - \overset{*}{\dot{\bar{c}}}{}_m^{n+1} \right)^2 \Bigg] + \\
& \sum_{i=1}^{\mathcal{N}} \left[\left(-\Delta t \sum_{m=1}^{\mathcal{M}} w_m \boldsymbol{\nabla} N_{mi} \cdot \bar{\boldsymbol{j}}_m^{n+1} + \sum_{m=1}^{\mathcal{M}} w_m N_{mi} (\bar{c}_m^{n+1} - \bar{c}_m^n) + \Delta t \hat{\bar{j}}_i^{n+1} \right) \bar{\lambda}_i^{n+1} \right], \qquad (20)
\end{aligned}
$$

where the rate terms in the distance function are approximated using the backward-Euler time discretization introduced in (7).

Next, to find the stationarity conditions for all the variables appearing in the discrete Lagrangian (20), it needs to be perturbed with respect to the admissible fields $\delta\bar{\mu}_i^{n+1}, \delta\underset{\sim}{\eta}_m^{n+1}, \delta\bar{\boldsymbol{j}}_m^{n+1}, \delta\bar{c}_m^{n+1}, \delta\bar{\lambda}_i^{n+1}$. These stationarity conditions are discussed next, one-by-one. The perturbation with respect to

the macroscopic chemical potential field $\bar{\mu}_i^{n+1}$, discretized at the nodes, provides,

$$\delta\bar{\mu}_i^{n+1}:\frac{\partial\overline{\mathcal{L}}^{n+1}}{\partial\bar{\mu}_i^{n+1}}=0\implies$$
$$\sum_{m=1}^{\mathcal{M}}w_m\sum_{j=1}^{\mathcal{N}}\left[N_{mi}\left({}^1\overline{\mathcal{C}}_m+\frac{{}^2\overline{\mathcal{C}}_m}{\Delta t}\right)N_{mj}+\boldsymbol{\nabla}N_{mi}\left({}^3\overline{\mathcal{C}}_m+\frac{{}^4\overline{\mathcal{C}}_m}{\Delta t}\right)\cdot\boldsymbol{\nabla}N_{mj}\right]\bar{\mu}_j^{n+1}=$$
$$\sum_{m=1}^{\mathcal{M}}w_m\sum_{j=1}^{\mathcal{N}}\left[N_{mi}\frac{{}^2\overline{\mathcal{C}}_m}{\Delta t}N_{mj}+\boldsymbol{\nabla}N_{mi}\frac{{}^4\overline{\mathcal{C}}_m}{\Delta t}\cdot\boldsymbol{\nabla}N_{mj}\right]\bar{\mu}_j^{n}+$$
$$\sum_{m=1}^{\mathcal{M}}w_m\left[N_{mi}(\ {}^1\overline{\mathcal{C}}_m\overset{*}{\bar{\mu}}_m^{n+1}+\ {}^2\overline{\mathcal{C}}_m\overset{*}{\dot{\bar{\mu}}}_m^{n+1})+\boldsymbol{\nabla}N_{mi}\cdot(\ {}^3\overline{\mathcal{C}}_m\overset{*}{\bar{\boldsymbol{g}}}_m^{n+1}+\ {}^4\overline{\mathcal{C}}_m\overset{*}{\dot{\bar{\boldsymbol{g}}}}_m^{n+1})\right],\quad(21)$$

which can be written in a matrix-column form as

$$\underline{\underline{\mathbb{K}}}^{\bar{\mu}}\underset{\sim}{\bar{\mu}}^{n+1}=\underline{\underline{\mathbb{M}}}^{\bar{\mu}}\underset{\sim}{\bar{\mu}}^{n}+\underset{\sim}{\mathbb{F}}^{\overset{*}{\bar{\mu}}}.\quad(22)$$

Equation (21) is a transient diffusion equation for $\bar{\mu}$ with corresponding bi-linear forms $(N_{mi}\bullet N_{mj})$ and $(\boldsymbol{\nabla}N_{mi}\bullet\boldsymbol{\nabla}N_{mj})$ as the capacity and diffusivity matrices, respectively. The macroscopic chemical potential field $\underset{\sim}{\bar{\mu}}^{n+1}$ calculated by (22), with given $(\overset{*}{\bar{\mu}}_m^{n+1},\overset{*}{\dot{\bar{\mu}}}_m^{n+1},\overset{*}{\bar{\boldsymbol{g}}}_m^{n+1},\overset{*}{\dot{\bar{\boldsymbol{g}}}}_m^{n+1})$ in the forcing term $\underset{\sim}{\mathbb{F}}^{\overset{*}{\bar{\mu}}}$, is locally compatible with $(\overset{*}{\bar{\mu}}_m^{n+1},\overset{*}{\bar{\boldsymbol{g}}}_m^{n+1})$ in a weak sense and also constrained by the corresponding rate terms $(\overset{*}{\dot{\bar{\mu}}}_m^{n+1},\overset{*}{\dot{\bar{\boldsymbol{g}}}}_m^{n+1})$ in the data-set. The Dirichlet boundary conditions, appearing in equation (1), are enforced on the $\underset{\sim}{\bar{\mu}}$ field, while homogeneous Neumann conditions are considered on the complementary part of the boundary $\partial\overline{\Omega}_{\bar{j}}$.

The perturbation with respect to the enrichment-variables reads

$$\delta\underset{\sim}{\eta}_m^{n+1}:\frac{\partial\overline{\mathcal{L}}^{n+1}}{\partial\underset{\sim}{\eta}_m^{n+1}}=0\implies\underset{\sim}{\eta}_m^{n+1}=\frac{1}{(\Delta t\ {}^5\underline{\mathcal{C}}_m+\ {}^6\underline{\mathcal{C}}_m)}\left[\ {}^6\underline{\mathcal{C}}_m\underset{\sim}{\eta}_m^{n}+\Delta t\ {}^5\underline{\mathcal{C}}_m\overset{*}{\underset{\sim}{\eta}}_m^{n+1}+\Delta t\ {}^6\underline{\mathcal{C}}_m\overset{*}{\dot{\underset{\sim}{\eta}}}_m^{n+1}\right],\quad(23)$$

which means that locally $\underset{\sim}{\eta}_m^{n+1}$ should be consistent with $\overset{*}{\underset{\sim}{\eta}}_m^{n+1}$ and its rate $\overset{*}{\dot{\underset{\sim}{\eta}}}_m^{n+1}$, present in the data-set. Since $\underset{\sim}{\eta}_m$ does not appear in the macroscopic balance equation (13) there are no Lagrange multipliers $\underset{\sim}{\bar{\lambda}}^{n+1}$ in (23); $\underset{\sim}{\eta}_m$ is connected to the other physical-state variables via the data-set only.

The perturbation with respect to the macroscopic mass flux yields

$$\delta\bar{\boldsymbol{j}}_m^{n+1}:\frac{\partial\overline{\mathcal{L}}^{n+1}}{\partial\bar{\boldsymbol{j}}_m^{n+1}}=0\implies\bar{\boldsymbol{j}}_m^{n+1}=\overset{*}{\bar{\boldsymbol{j}}}_m^{n+1}+\Delta t\frac{1}{{}^7\overline{\mathcal{C}}_m}\sum_{i=1}^{\mathcal{N}}\boldsymbol{\nabla}N_{mi}\bar{\lambda}_i^{n+1},\quad(24)$$

which states that the difference between the local macroscopic mass flux $\bar{\boldsymbol{j}}_m^{n+1}$ and its counterpart in the data-set $\overset{*}{\bar{\boldsymbol{j}}}_m^{n+1}$, at iteration k, should be balanced through the Lagrange multipliers field $\bar{\lambda}_i^{n+1}$. The perturbation with respect to the macroscopic concentration can be written as

$$\delta\bar{c}_m^{n+1}:\frac{\partial\overline{\mathcal{L}}^{n+1}}{\partial\bar{c}_m^{n+1}}=0\implies$$
$$\bar{c}_m^{n+1}=\frac{1}{(\Delta t\ {}^8\overline{\mathcal{C}}_m+\ {}^9\overline{\mathcal{C}}_m)}\left[\ {}^9\overline{\mathcal{C}}_m\bar{c}_m^{n}+\Delta t\ {}^8\overline{\mathcal{C}}_m\overset{*}{\bar{c}}_m^{n+1}+\Delta t\ {}^9\overline{\mathcal{C}}_m\overset{*}{\dot{\bar{c}}}_m^{n+1}-\Delta t\sum_{i=1}^{\mathcal{N}}N_{mi}\bar{\lambda}_i^{n+1}\right],\quad(25)$$

which gives a local macroscopic concentration field $\bar{c}_m^{n+1}$ consistent with $(\overset{*}{\bar{c}}_m^{n+1}, \overset{*}{\dot{\bar{c}}}_m^{n+1})$ in the data-set, whereby the difference is rectified by the Lagrange multipliers field $\bar{\lambda}_i^{n+1}$.

Finally, taking the variation with respect to the Lagrange multiplier field $\bar{\lambda}_i^{n+1}$ amounts to

$$\delta\bar{\lambda}_i^{n+1} : \frac{\partial\bar{\mathcal{L}}^{n+1}}{\partial\bar{\lambda}_i} = 0 \implies -\Delta t\sum_{m=1}^{\mathcal{M}} w_m \boldsymbol{\nabla} N_{mi}\cdot\bar{\boldsymbol{j}}_m^{n+1} + \sum_{m=1}^{\mathcal{M}} w_m N_{mi}(\bar{c}_m^{n+1} - \bar{c}_m^{n}) + \Delta t\hat{\bar{j}}_i^{n+1} = 0\,, \quad (26)$$

which is the balance between the internal and the external macroscopic mass fluxes at the nodes. Substituting the expressions of $\bar{\boldsymbol{j}}_m^{n+1}$ and $\bar{c}_m^{n+1}$ from equations (24) (25) into equation (26) and performing some straight forward manipulations provides the system of equations for the Lagrange multiplier field $\bar{\lambda}_i^{n+1}$ as follows

$$\begin{aligned} &-\Delta t^2\sum_{m=1}^{\mathcal{M}} w_m \boldsymbol{\nabla} N_{mi}\cdot\frac{1}{{}^7\bar{\mathcal{C}}_m}\sum_{j=1}^{\mathcal{N}}\boldsymbol{\nabla} N_{mj}\bar{\lambda}_j^{n+1} - \Delta t\sum_{m=1}^{\mathcal{M}} w_m N_{mi}\frac{1}{\Delta t\ {}^8\bar{\mathcal{C}}_m + {}^9\bar{\mathcal{C}}_m}\sum_{j=1}^{\mathcal{N}} N_{mj}\bar{\lambda}_j^{n+1} = \\ &\Delta t\sum_{m=1}^{\mathcal{M}} w_m \boldsymbol{\nabla} N_{mi}\cdot\bar{\boldsymbol{j}}_m^{*} - \Delta t\hat{\bar{j}}_i^{n+1} - \sum_{m=1}^{\mathcal{M}} w_m N_{mi}\Big(\frac{1}{\Delta t\ {}^8\bar{\mathcal{C}}_m + {}^9\bar{\mathcal{C}}_m}\Big[\Delta t\ {}^8\bar{\mathcal{C}}_m\bar{c}_m^{*} + {}^9\bar{\mathcal{C}}_m\bar{c}_m^{n} + \Delta t\ {}^9\bar{\mathcal{C}}_m\dot{\bar{c}}_m^{*}\Big] - \bar{c}_m^{n}\Big)\,, \end{aligned} \quad (27)$$

which in the matrix-column form can be written as

$$\underline{\underline{\mathbb{K}}}^{\bar{\lambda}}\underset{\sim}{\bar{\lambda}}^{n+1} = \underset{\sim}{\mathbb{F}}^{\bar{\lambda}}\,. \quad (28)$$

In equations (24)–(28), the Lagrange multiplier field can be interpreted as an equivalent macroscopic chemical potential which minimizes the difference between the physical-state $(\bar{\boldsymbol{j}}_m^{n+1}, \bar{c}_m^{n+1})$ and the point $(\overset{*}{\bar{\boldsymbol{j}}}_m^{n+1}, \overset{*}{\bar{c}}_m^{n+1}, \overset{*}{\dot{\bar{c}}}_m^{n+1})$ in the data-set, which are present in $\underset{\sim}{\mathbb{F}}^{\bar{\lambda}}$. In equation (28), the Lagrange multiplier field is subject to $\bar{\lambda}_i^{n+1} = 0$ on $\partial\bar{\Omega}_{\bar{\mu}}$ and there is an influx of mass $\hat{\bar{j}}$, which is zero, at the Neumann part of the boundary that naturally appears in the system of equations through the weak form (13). For a variational formulation and a detailed discussion on the boundary conditions on the fields appearing in the data-driven problems, the reader is directed to [26].

In a staggered approach, after solving for $\underset{\sim}{\bar{\mu}}^{n+1}$ and $\underset{\sim}{\bar{\lambda}}^{n+1}$ from equations (22) and (28), the projection $z_{k+1}^{n+1} = P_{E^{n+1}}\overset{*}{y}_k^{n+1}$ can be obtained by evaluating $\underset{\sim}{\eta}_m^{n+1}, \bar{\boldsymbol{j}}_m^{n+1}$ and $\bar{c}_m^{n+1}$ from (23), (24) and (25), respectively. The subsequent projection $\overset{*}{y}_{k+1}^{n+1} = P_D z_{k+1}^{n+1}$ is achieved by a simple search through the data to find a point in the global data-set D which provides the minimum distance to z_{k+1}^{n+1}, as discussed in more detail in the next section.

3.4 Algorithm

The pseudo algorithm for distance minimizing data-driven reduced homogenization is shown in Algorithm 1. The data-driven solver initializes with setting the maximum number of allowed iterations $maxIter$, the data-driven iteration counter k, the allowed tolerance tol and the time stepping variable n. An initial guess for the optimum $\overset{*}{y}_k^{n+1} \in D$ is made. In the preliminary work [13], $\overset{*}{y}_{m,k}^{n+1}$ was initialized by assigning randomly a point in the D_m to each material point m. However, it was observed that an initial guess of $\overset{*}{y}_k^{n+1} = 0$ requires less number of iterations to converge to the desired tolerance, both for steady-state and transient problems.

Algorithm 1: Distance minimizing data-driven reduced homogenization.

- ▶ Initialize: $maxIter = 100$, $k = 0$, $tol = 10^{-12}$, $n = 1$;
- ▶ Assign: $\overset{*}{y}{}_k^{n+1} = 0$;
- ▶ Assemble: $\underset{\sim}{\mathbb{K}}^{\bar{\mu}}, \underset{\sim}{\mathbb{K}}^{\bar{\lambda}}, \underset{\sim}{\mathbb{M}}^{\bar{\mu}}, \underset{\sim}{\mathbb{F}}{}_k^{\overset{*}{\bar{\mu}}}$ and $\underset{\sim}{\mathbb{F}}{}_k^{\overset{*}{\bar{\lambda}}}$;

for $n = 0 \rightarrow T$ **do**

  while $k < maxIter$ **do**

    - ▶ Solve: $\underset{\sim}{\mathbb{K}}^{\bar{\mu}} \underset{\sim}{\bar{\mu}}{}_{k+1}^{n+1} = \underset{\sim}{\mathbb{M}}^{\bar{\mu}} \underset{\sim}{\bar{\mu}}{}_{k+1}^{n} + \underset{\sim}{\mathbb{F}}{}_k^{\overset{*}{\bar{\mu}}}$ (22) and $\underset{\sim}{\mathbb{K}}^{\bar{\lambda}} \underset{\sim}{\bar{\lambda}}{}_{k+1}^{n+1} = \underset{\sim}{\mathbb{F}}{}_k^{\overset{*}{\bar{\lambda}}}$ (28) ;

    for $m = 1 \rightarrow \mathcal{M}$ **do**

      - ▶ Evaluate: $z_{m,k+1}^{n+1} = P_{E^{n+1}} \overset{*}{y}{}_{m,k}^{n+1}$;
      - ▶ Choose: $\overset{*}{y}{}_{m,k+1}^{n+1} = P_D z_{m,k+1}^{n+1}$ such that $d_m(z_{m,k+1}^{n+1}, \overset{*}{y}{}_{m,k+1}^{n+1}) \leq d_m(z_{m,k+1}^{n+1}, D_m)$;
      - ▶ Integrate: $d(z_{k+1}^{n+1}, \overset{*}{y}{}_{k+1}^{n+1})$ from $d_m(z_{m,k+1}^{n+1}, \overset{*}{y}{}_{m,k+1}^{n+1})$;
      - ▶ Assemble: $\underset{\sim}{\mathbb{F}}{}_{k+1}^{\overset{*}{\bar{\mu}}}$ and $\underset{\sim}{\mathbb{F}}{}_{k+1}^{\overset{*}{\bar{\lambda}}}$ using $\overset{*}{y}{}_{k+1}^{n+1}$;

    end

    if $abs\left(d(z_{k+1}^{n+1}, \overset{*}{y}{}_{k+1}^{n+1}) - d(z_k^{n+1}, \overset{*}{y}{}_k^{n+1})\right) \leq tol$ **then**

      - ▶ Terminate ;

    else

      - ▶ k = k + 1 ;

    end

  end

end

Unlike material model based finite element solvers, independently of the existence of a potential non-linearity in the material behavior, the distance minimizing data-driven solver requires the assembly of the matrices $\underset{\sim}{\mathbb{K}}^{\bar{\mu}}, \underset{\sim}{\mathbb{K}}^{\bar{\lambda}}$ and $\underset{\sim}{\mathbb{M}}^{\bar{\mu}}$ only once. For a specific physical phenomenon under consideration, for instance elasticity, diffusion, or history-dependent materials, the same solver can be used for different materials and different data-sets. Then, in the time stepping and data-driven loops, first the pertinent $\underset{\sim}{\bar{\mu}}^{n+1}$ and $\underset{\sim}{\bar{\lambda}}^{n+1}$ problems are solved. In the current formulation these two problems, (22) and (28), are algebraically independent, the only coupling is through the data-set D. However, in some other data-driven problems, as can be seen in the case of dynamics [14] and transient Fickian diffusion [26] (formulated in concentrations), a coupled system of equations emerges after taking variations of the Lagrangian.

Next, at the material point level, the projection $z_{m,k+1}^{n+1} = P_{E^{n+1}} \overset{*}{y}{}_{m,k}^{n+1}$ is performed. It involves evaluating the physical-state of the material $z_{m,k+1}^{n+1} = (\bar{\mu}_m, \dot{\bar{\mu}}_m, \bar{\boldsymbol{g}}_m, \dot{\bar{\boldsymbol{g}}}_m, \underset{\sim}{\eta}_m, \dot{\underset{\sim}{\eta}}_m, \bar{\boldsymbol{j}}_m, \bar{c}_m, \dot{\bar{c}}_m)_{k+1}^{n+1}$, from the previously calculated $\overset{*}{y}{}_k^{n+1}, \underset{\sim}{\bar{\mu}}_{k+1}^{n+1}$ and $\underset{\sim}{\bar{\lambda}}_{k+1}^{n+1}$. The local values of $\bar{\mu}_{m,k+1}^{n+1}$ are evaluated by interpolating $\bar{\mu}_{i,k+1}^{n+1}$ using the finite element shape functions N_{mi}, $\bar{\boldsymbol{g}}_{m,k+1}^{n+1}$ is computed by using compatibility (12), while $\underset{\sim}{\eta}_{m,k+1}^{n+1}, \bar{\boldsymbol{j}}_{m,k+1}^{n+1}$ and $\bar{c}_{m,k+1}^{n+1}$ are calculated by (23), (24) and (25), respectively and their rates $(\dot{\bar{\mu}}_m, \dot{\bar{\boldsymbol{g}}}_m, \dot{\underset{\sim}{\eta}}_m, \dot{\bar{c}}_m)_{k+1}^{n+1}$ using the approximation in (7).

The global distance function (18) is minimized by finding the minima of the local distance function (16), using all the points in the local data-set D_m, at each material point m, and then integrating it using numerical quadrature. The minimum for the local distance function (16) is found through a simple lookup array search algorithm, which amounts to performing

the projection $\overset{*}{y}{}^{n+1}_{m,k+1} = P_D z^{n+1}_{m,k+1}$. Searching through the data-set D_m is the computationally expensive part of data-driven algorithm. If the data-set is large enough, a smart search algorithm, for instance, based on a tree search algorithm [29], should be used to accelerate this step. Then, using the newly found values of $\overset{*}{y}{}^{n+1}_{m,k+1}$, the flux columns $\underset{\sim}{\mathbb{F}}^{\overset{*}{\bar{\mu}}}_{k+1}$ and $\underset{\sim}{\mathbb{F}}^{\overset{*}{\bar{\lambda}}}_{k+1}$ are assembled. Finally, the convergence is checked and the iterations are terminated if there is no change in the optimum data point. When the data-set is perfect, i.e. it contains all the required points to the corresponding states in the phase-space for a given spatial and temporal discretization and boundary conditions, the data-driven solver should be able to find a point in the data-set which matches the point in the phase-space exactly. In that case, the convergence (or stop) criteria can be set as the distance function approaching zero, i.e. $d(z^{n+1}_{k+1}, \overset{*}{y}{}^{n+1}_{k+1}) \to 0$. However, when the data is noisy or incomplete, the distance function might not approach zero and may stagnate, after some iterations, at a certain minimum value. In that case, the convergence criteria based on the stagnation of the global distance function, that is $\mathsf{abs}(d(z^{n+1}_{k+1}, \overset{*}{y}{}^{n+1}_{k+1}) - d(z^{n+1}_{k}, \overset{*}{y}{}^{n+1}_{k})) \leq tol$ as given in Algorithm 1, can be used to terminate the data-driven iterations, where $\mathsf{abs}(\bullet)$ is the absolute value.

4 Numerical Examples

In this section, the proposed framework for data-driven reduced homogenization for transient diffusion problems with history effects is illustrated through numerical examples. First, the problem settings are presented for the micro-scale and the macro-scale. Next, the data-generation step is performed by loading the micro-scale, post-processing, and storing the relevant quantities. After that, the data-driven simulations are carried out, whereby the homogenized chemical potential fields obtained by the data-driven approach are compared with the ones obtained by the regular enriched-continuum formulation. Information from the reduced micro-scale model is used to select the coefficients in the distance function. The micro-scale chemical potential fields are post-processed and also compared. The performance of the data-driven approach using noisy and different data-sets is also analyzed. Finally, a convergence analysis, with respect to the number of points in the data-set, is carried out.

4.1 Problem Settings: Micro-scale and Macro-scale

4.1.1 Micro-scale

The micro-scale consists of a two-dimensional square unit-cell with side length ℓ and a single circular inclusion of diameter d embedded in a matrix. The material properties and linear constitutive material models for the inclusion and the matrix are assumed to be known and complying with the relaxed separation of scales regime ($\tau_m \ll \tau_i \sim T$). Same chemical modulus Λ is assumed for both the inclusion and the matrix. It is not a requirement but rather a convenience for implementing the down-scaling relations $\bar{\mu} = \langle \mu \rangle$ and $\overline{\boldsymbol{\nabla}} \bar{\mu} = \langle \boldsymbol{\nabla} \mu \rangle$ by fixing a point at the micro-scale, for more details about implementation see [8]. Lagrange multipliers, as discussed in [30] for elastodynamics problems, can be used to implement the down-scaling relations with different storage terms (here Λ) for inclusion and matrix. Specific type of boundary conditions are used to fulfill the up-scaling relations, i.e. equivalence of virtual power. The most common ones are (i) zero micro-fluctuation and (ii) periodic micro-fluctuation boundary conditions. In this work, a two-dimensional micro-scale unit-cell attached to a one-dimensional

macro-scale, as shown in Figure 1, represents a slab in an infinite vertical stack of uni-cells and hence periodic micro-fluctuation boundary condition is an obvious choice. The unit-cell is discretized with nearly 4400 linear triangular elements and 2200 nodes. The material properties and other parameters used in the simulations are listed in Table 1. After the assembly of the finite element system and application of the periodic boundary conditions, at the micro-scale, an eigenvalue problem is solved for the smallest 100 eigenvalues $\underline{\alpha}$ and eigenvectors $\underline{\Phi}$. Next, a criterion based on either the energy consistency or coupling terms, as proposed in [8, 12], is used to select a limited number ($\mathcal{N}_q$) eigenmodes that contribute most, in terms of transient effects, to the macro-scale response. For the unit-cell with a single inclusion, as shown in Figure 1, and the material properties given in Table 1, the selection criteria based on the coupling terms provides 6 important eigenvectors as a reduced basis set, see [27] for more details and the contour plots of the selected eigenvectors. Finally, the homogenized coefficients $(\overline{\boldsymbol{a}}, \overset{*}{\underset{\sim}{\boldsymbol{a}}}, \overline{\boldsymbol{B}}, \bar{\boldsymbol{c}}, \overline{\boldsymbol{C}}, \underset{\sim}{\bar{d}}, \overset{*}{\underset{\sim}{\bar{d}}}, \bar{\boldsymbol{e}}, \bar{f}, \overline{\boldsymbol{f}})$ are determined and stored [8]. The values for the first components of these coefficients can be found in Table 2. Note that, even though the macro-scale considered here is a one-dimensional domain, a two-dimensional micro-scale problem has to be solved to obtain the required values of the homogenized coefficients, since in a one-dimensional micro-scale domain the diffusion around the inclusion can not be represented at the macro-scale through homogenization. The reduced order model (4)–(6) is now ready for the data-generation stage.

4.1.2 Macro-scale

The homogenized macroscopic domain $\overline{\Omega}$, both for the enriched-continuum and the data-driven simulations, is a one-dimensional bar of length L with a Dirichlet boundary condition on $\partial\Omega_{\bar{\mu}}$ on the left side of the domain and Neumann no-flux boundary condition on the right side of the domain. It is discretized with 50 linear one-dimensional finite elements, unless stated otherwise, which are integrated using a two-point Gauss quadrature rule. For consistency, in the following the vectorial/tensorial quantities in one-dimension are still shown with a tensorial notation. Total loading time is chosen to be $T = 0.1\tau_i$[s] and both the enriched-continuum and the data-driven problems are discretized in time using the backward-Euler time integration scheme. The reference time step size Δt is taken to be $\Delta t = T \times 10^{-3}$[s]. In a data-driven solver, the time step size can also be obtained based on a term and its rate present in the data-set, for example $\Delta t = \frac{\bar{\mu}_m^{I+1} - \bar{\mu}_m^I}{\dot{\bar{\mu}}_m^{I+1}}$. For the macro-scale simulations, the ramp loading conditions

$$\bar{\mu}_p^{n+1}(t) = \begin{cases} \frac{t}{T_R}\bar{\mu}_{\max}\,, & \text{if } t \leq T_R \text{ (loading)} \\ \bar{\mu}_{\max}\,, & T_R < t \leq T \text{ (relaxation)} \end{cases} \qquad \text{on} \qquad \partial\overline{\Omega}_{\bar{\mu}}\,, \tag{29}$$

are used on the Dirichlet part of the macroscopic boundary $\partial\overline{\Omega}_{\bar{\mu}}$, where $\bar{\mu}_{\max} = \Lambda(c_{\max} - c_0)$ is the maximum attainable chemical potential during mass diffusion and $T_R = T/2$. The first part of the ramp till T_R represents a loading path and the second part from T_R to T represents a relaxation path. The data-driven reduced homogenization solver is initialized with an initial guess $\overset{*}{y}{}_k^{n+1} = 0$ and the stagnation criteria $|d(z_{k+1}^{n+1}, \overset{*}{y}{}_{k+1}^{n+1}) - d(z_{k+1}^{n+1}, \overset{*}{y}{}_k^{n+1})| \leq tol$ is used to terminate the iterations of the staggered scheme. The numerical solution of the enriched-continuum formulation at the macro-scale is used as a reference solution.

Table 1: Default parameters used in the simulations.

Parameter	Symbol	Value	Units
Micro-scale			
Characteristic unit-cell length	ℓ	1×10^{-2}	[m]
Inclusion diameter	d	0.6×10^{-2}	[m]
Mobility in matrix	M_m	1.1×10^{-4}	$[\mathrm{mol^2\,J^{-1}\,m^{-1}\,s^{-1}}]$
Mobility in inclusion	M_i	1.85×10^{-9}	$[\mathrm{mol^2\,J^{-1}\,m^{-1}\,s^{-1}}]$
Reference temperature	θ_0	298	[K]
Boltzman's constant	k_b	8.314	$[\mathrm{J\,K^{-1}\,mol^{-1}}]$
Maximum concentration	$c_{\max}$	24161	$[\mathrm{mol\,m^{-3}}]$
Minimum concentration	c_0	$0.0547 c_{\max}$	$[\mathrm{mol\,m^{-3}}]$
Chemical modulus	$\Lambda = k_b\theta_0/c_0$	1.83	$[\mathrm{J\,m^3\,mol^{-2}}]$
Characteristic diffusion time of inclusion	τ_i	$d^2/M_i\Lambda = 36000$	[s]
Characteristic diffusion time of matrix	τ_m	$\ell^2/M_m\Lambda = 1.69$	[s]
Number of nodes in unit-cell mesh		$\sim 2.2 \times 10^3$	
Macro-scale			
Macroscopic domain length	L	100×10^{-2}	[m]
Total simulation time T	T	$0.1\tau_i$	[s]
Number of elements (reference)		50	
Number of nodes (reference)	$\mathcal{N}$	51	
Time step size	Δt	$T \times 10^{-3}$	[s]
Data-Driven Solver			
Number of data points	n_{dp}	1000	
Maximum number iterations	$maxIter$	100	
Tolerance for the termination criteria	tol	10^{-12}	

Table 2: The values of the coefficients appearing in equations (5) and (6), for the unit-cell shown in Figure 1 with geometrical and material parameters given in Table 1, $\boldsymbol{A}_{(\bullet)}$ indicates the $(\bullet)$ component of the tensor $\boldsymbol{A}$.

Coefficient	$\bar{\boldsymbol{a}}^{(1)}_{(1)}$	$\overline{\boldsymbol{B}}_{(11)}$	$\bar{\boldsymbol{c}}_{(1)}$	$\overline{\boldsymbol{C}}_{(11)}$	in equation (5)
Units	$[\mathrm{mol\,m^{-2}\,s^{-1}}]$	$[\mathrm{J^{-1}\,mol^2\,m^{-1}\,s^{-1}}]$	$[\mathrm{J^{-1}\,mol^2\,m^{-2}}]$	$[\mathrm{J^{-1}\,mol^2\,m^{-1}}]$	
Value	1.3×10^{-7}	0.6×10^{-4}	1.3×10^{-8}	0.6×10^{-5}	
Coefficient	$\bar{d}^{(1)}$	$\bar{\boldsymbol{e}}_{(1)}$	$\bar{f}$	$\bar{\boldsymbol{f}}_{(1)}$	in equation (6)
Units	$[\mathrm{mol\,m^{-3}\,s^{-1}}]$	$[\mathrm{J^{-1}\,mol^2\,m^{-2}\,s^{-1}}]$	$[\mathrm{J^{-1}\,mol^2\,m^{-3}}]$	$[\mathrm{J^{-1}\,mol^2\,m^{-2}}]$	
Value	-33.46	-0.6×10^{-16}	0.534	-0.13×10^{-7}	

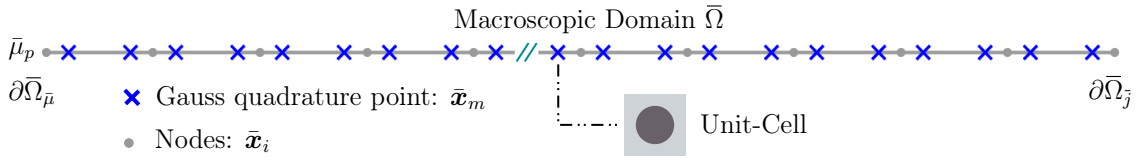


Figure 1: The macroscopic domain $\bar{\Omega}$ with a prescribed macroscopic chemical potential $\bar{\mu}_p$ at the Dirichlet part of the boundary $\partial\bar{\Omega}_{\bar{\mu}}$, and zero-flux $\hat{\bar{j}} = 0$ at the Neumann part of the boundary $\partial\bar{\Omega}_{\bar{j}}$. The finite elements nodes are shown with gray circles. The Gauss quadrature points, where the data-set D_m is available, are shown with blue crosses. The unit-cell from which the reduced order model and the data-set are obtained is also shown: the light-gray part of the unit-cell is the matrix material while the dark gray is the inclusion material.

4.2 Data-Generation from Micro-Scale Simulations

4.2.1 Input Generation $(\bar{\mu}_m^{n+1}, \bar{\boldsymbol{g}}_m^{n+1})$

To obtain the data-set D representative of the problem, the data-generation involves micro-scale simulations, ideally with all possible loading scenarios. In practice, a wide spectrum of loading conditions, i.e. $\bar{\mu}_m^{n+1}$ and $\bar{\boldsymbol{g}}_m^{n+1}$ with varying magnitudes and frequencies, may be needed when a stand-alone micro-scale problem is considered. In the current work, for the validation of the proposed data-driven solver, the loading conditions $(\bar{\mu}_m^{n+1}, \bar{\boldsymbol{g}}_m^{n+1})$ are obtained via a post-processing, at the first Gauss quadrature point of each element and at each time step t^{n+1}, of the solution of the enriched-continuum problem with different loading conditions, given in Table 3. As an example, the outcome for the ramp loading condition (31) is shown in Figure 2. In this

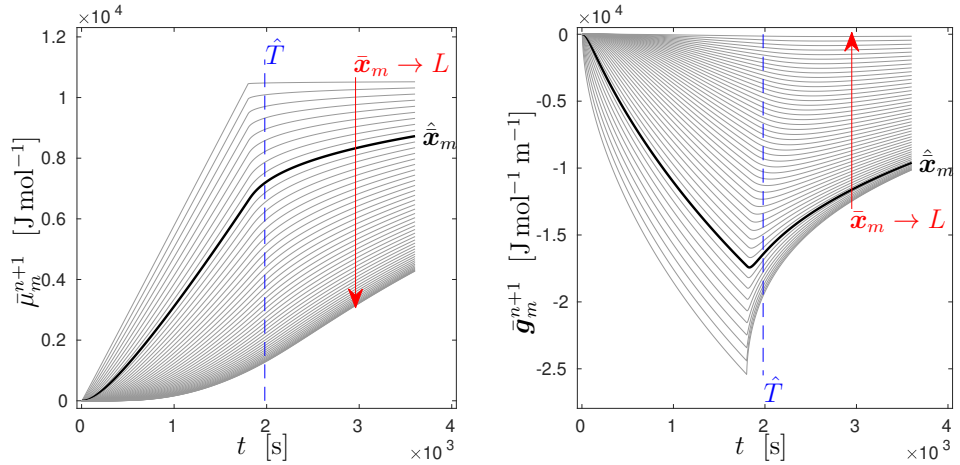


Figure 2: (a) Local macroscopic chemical potential field $\bar{\mu}_m^{n+1}$ and (b) local gradient of macroscopic chemical potential field $\bar{\boldsymbol{g}}_m^{n+1}(t)$ to be used as the input for the micro-scale data-generation step. These are post-processed, from the solution of the enriched-continuum problem, at the first Gauss quadrature point of every element in the macroscopic domain loaded with the ramp condition (31). The dark black line represents $\bar{\mu}_m^{n+1}$ and $\bar{\boldsymbol{g}}_m^{n+1}$ at a reference macroscopic point $\hat{\bar{\boldsymbol{x}}}_m$ which will be used to compare the local elemental quantities. The dashed blue line is at a reference time $\hat{T} = 0.55T$ at which the global quantities, at the nodes, will be compared.

case, $\bar{\mu}_m^{n+1}$ and $\bar{\boldsymbol{g}}_m^{n+1}$ are connected through the macroscopic initial boundary value problem (1)

and each $\bar{\mu}_m^{n+1}$ graph in Figure 2 (a) corresponds to a graph of $\bar{\boldsymbol{g}}_m^{n+1}$ in Figure 2 (b), collectively representing an input in time to the micro-scale reduced problem (4). The ramp effect smooths out as $\bar{\boldsymbol{x}}_m \rightarrow L$, as indicated by the red arrow, which provides different magnitudes and types of loading conditions for the micro-scale problems. In the following, local quantities of interest, such as macroscopic mass-flux $\bar{\boldsymbol{j}}_m^{n+1}$ and the rate of change of macroscopic concentration $\dot{\bar{c}}_m^{n+1}$ will be compared, for the data-driven and the full enriched-continuum solution will be made, at a reference point $\hat{\bar{\boldsymbol{x}}}_m$, located at the first Gauss quadrature point of the tenth element, which is indicated with a dark black line in Figure 2. The global quantities, for example the macroscopic chemical potential field $\bar{\mu}_i^{n+1}$ at the nodes will be compared at a time instance $\hat{T} = 0.55T$ as shown with a dashed blue line in Figure 2.

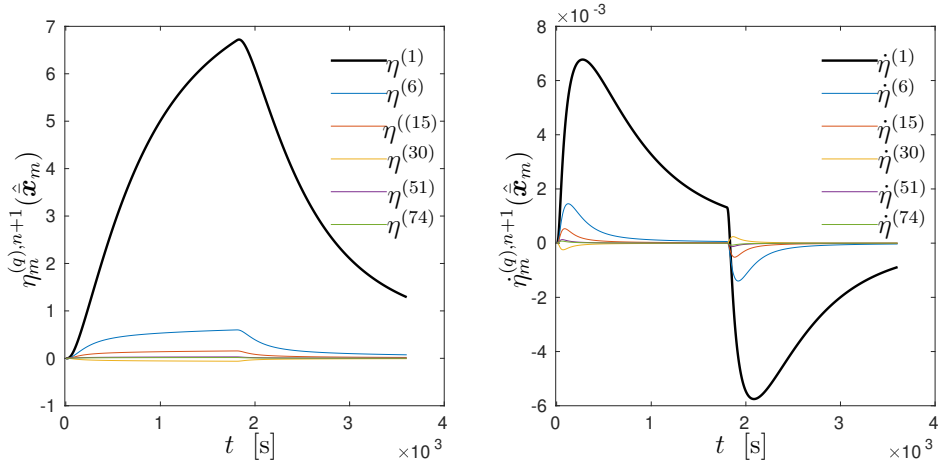
4.2.2 Data-Generation

The data-generation is performed by solving the reduced model (4)–(6), for $\bar{\mu}_m^{n+1}(t)$ and $\bar{\boldsymbol{g}}_m^{n+1}(t)$ computed in the previous section, with the time discretization performed using the approximation in equation (7). The data is stored in a local data-set

$$D_m = \{(\bar{\mu}'_m, \dot{\bar{\mu}}'_m, \bar{\boldsymbol{g}}'_m, \dot{\bar{\boldsymbol{g}}}'_m, \underset{\sim}{\eta}'_m, \dot{\underset{\sim}{\eta}}_m{}', \bar{\boldsymbol{j}}'_m, \bar{c}'_m, \dot{\bar{c}}'_m)\}_{I=1}^{n_{dp}} \tag{30}$$

In current work, for the data-driven simulations the same local data-set D_m is available to all the material points i.e. $D_1 = D_2 = ... = D_{\mathcal{M}}$; in general, a different data-set can be available for each material point.

As stated before, the selection criteria for the dominant eigenvectors, as proposed in [8, 12], provide $\mathcal{N}_q = 6$ for the considered unit-cell. It turns out, however, that out of these six, only one eigenvalue has the largest contribution to the lagging behavior at the macro-scale. This can be verified from the time evolution of $\eta_m^{(q),n+1}$ and $\dot{\eta}_m^{(q),n+1}$ at $\hat{\bar{\boldsymbol{x}}}_m$, as shown in Figure 3. Therefore, it has been chosen to use only $\eta_m^{(1),n+1}$ and $\dot{\eta}_m^{(1),n+1}$ in the data-driven calculations to capture the history-dependent response. The different loading conditions (Table 3) have

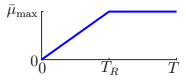
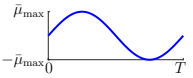
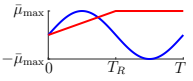
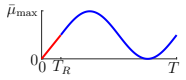


(a) (b)

Figure 3: (a) Time evolution of the enrichment-variables $\eta^{(q),n+1}(t)$ and (b) their rates $\dot{\eta}_m^{n+1}$ at the reference macroscopic point $\hat{\bar{\boldsymbol{x}}}_m$. For the unit-cell shown in Figure 1, the eigenvalue and the corresponding eigenvector corresponding to $\eta^{(1),n+1}$ are dominant one.

provided different data-sets which are indexed as listed in Table 3. The number of data-points

Table 3: Names, expressions, graphs of the macroscopic loading conditions, symbols and the number of data-points n_{dp} for different data-sets used in present study.

Name	Loading condition expression	$\bar{\mu}_p^{n+1}(t)$ Graph	Data-Set	n_{dp}
Ramp	$\bar{\mu}_p^{n+1}(t) = \begin{cases} \frac{tT}{T_R}\bar{\mu}_{\max}\,, & \text{if } t \le T_R \\ \bar{\mu}_{\max}\,, & \text{otherwise} \end{cases}$ (31) where $T_R = T/2$		D_R	1000
Sine	$\bar{\mu}_p^{n+1}(t) = \bar{\mu}_{\max}\sin(\omega t)$ (32) where $\omega = 2\pi/T$		D_S	1000
Ramp + Sine	(31) and (32)		$D_{(R+S)}$	2000
Ramp & Sine	$\bar{\mu}_p^{n+1}(t) = \begin{cases} \frac{tT}{2T_R}\bar{\mu}_{\max}\,, & \text{if } t \le T_R \\ \frac{\bar{\mu}_{\max}}{2}\sin(\omega t) + \frac{\bar{\mu}_{\max}}{2}\,, & \text{otherwise} \end{cases}$ (33) where $T_R = T/7$ and $\omega = 2\pi/(T - T_R)$		$D_{(RS)}$	1000

$n_{dp} = T/\Delta t + 1$ are also given in the table. After the data-generation stage, the information about the micro-scale must be discarded, as the data-driven solver should only rely on the raw data.

Pairs $(\bar{\boldsymbol{g}}_m^{n+1}, \bar{\boldsymbol{j}}_m^{n+1})$, $(\dot{\bar{\mu}}_m^{n+1}, \dot{\bar{c}}_m^{n+1})$ and $(\dot{\bar{\mu}}_m^{n+1}, \dot{\eta}_m^{(1),n+1}, \dot{\bar{c}}_m^{n+1})$ in D_R are visualized in Figure 4 (a), (b) and (c), respectively. The pairs $(\bar{\boldsymbol{g}}_m^{n+1}, \bar{\boldsymbol{j}}_m^{n+1})$ show a negative linear behavior that is independent of the loading path, which indicates that the history effects are not prominent in the diffusion contribution at the macro-scale and can also be sufficiently accurately calculated by the standard volume averaging of the Fickian diffusion behavior at the micro-scale. However, a prominent history dependence and non-Fickian behavior can be observed in the graphs of $(\dot{\bar{\mu}}_m^{n+1}, \dot{\bar{c}}_m^{n+1})$, where there is neither a linear nor logarithmic relation between $\dot{\bar{\mu}}_m^{n+1}$ and $\dot{\bar{c}}_m^{n+1}$ at the macro-scale. This can also be observed in $(\dot{\bar{\mu}}_m^{n+1}, \dot{\eta}_m^{(1),n+1}, \dot{\bar{c}}_m^{n+1})$ graph, which clearly indicates that the history effect emerges from the storage/capacitance term at the macro-scaleand can be tracked by the internal-variable $\dot{\eta}_m^{(1),n+1}$. A correct value of $\dot{\eta}_m^{(1),n+1}$, at a spatial point $\bar{\boldsymbol{x}}_m$ and time t^{n+1}, selected by the projection $\overset{*}{y}{}_{m,k+1}^{n+1} = P_D z_{m,k+1}^{n+1}$, will direct the other quantities, in the data-set D_m, to be either in the loading or the relaxation path, hence keeping track of the history effects. Next, the macroscopic chemical potential field $\bar{\mu}$ and the microscopic chemical potential fields μ obtained from the data-driven reduced homogenization and the enriched-continuum formulation will be compared to provide an indication of the performance of the data-driven approach.

4.3 Homogenized and Microscopic Fields

In this section, the developed data-driven reduced homogenization is verified using the data-set D_R generated from the enriched-continuum problem, with the same loading condition (31) as

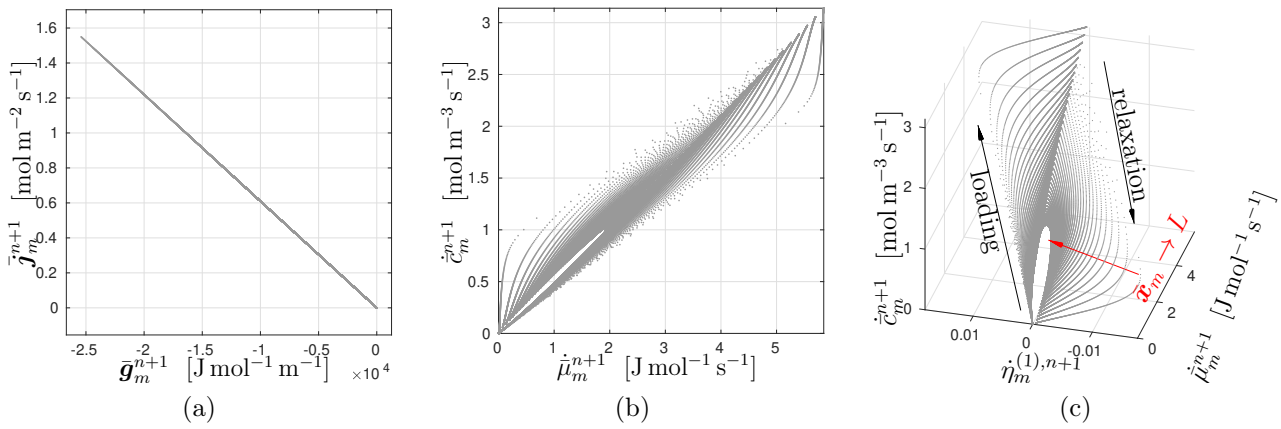


Figure 4: From the data-set D_R, the pairs (a) $(\bar{\boldsymbol{g}}_m^{n+1}, \bar{\boldsymbol{j}}_m^{n+1})$, (b) $(\dot{\bar{\mu}}_m^{n+1}, \dot{\bar{c}}_m^{n+1})$ and (c) $(\dot{\bar{\mu}}_m^{n+1}, \dot{\eta}_m^{(1),n+1}, \dot{\bar{c}}_m^{n+1})$ are visualized.

it is used for the data-driven initial boundary value problem. In this scenario, the data-set can be assumed to be ideal and if the data-driven problem is formulated correctly, both, the enriched-continuum and the data-driven solutions must match very accurately. In Figure 5, the macroscopic chemical potential field at time $\hat{T}$ obtained by the enriched-continuum formulation $\underset{\sim}{\bar{\mu}}_E^{n+1}$ (reference) is shown with the gray line, while the one obtained by the proposed data-driven reduced homogenization method $\underset{\sim}{\bar{\mu}}_D$ using the data-set D_R is shown with the blue line. It can be observed that $\underset{\sim}{\bar{\mu}}_E^{n+1}$ and $\underset{\sim}{\bar{\mu}}_D^{n+1}$ lie on top of each other. The micro-scale chemical potential

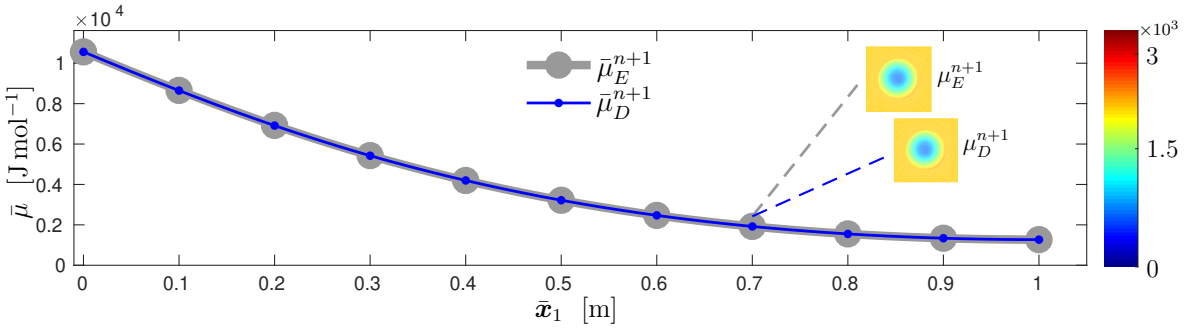


Figure 5: Comparison between the macroscopic chemical potential fields obtained via enriched continuum formulation $\bar{\mu}_E^{n+1}$ (reference), shown with the gray line, and the data-driven reduced homogenization $\bar{\mu}_D^{n+1}$ using D_R, shown with the blue line, at time step $\hat{T}$. The microscopic chemical potential fields μ_E^{n+1} and μ_D^{n+1} are post-processed at $\bar{\boldsymbol{x}}_1 = 0.6842$[m] .The marker is plotted at every tenth node of the finite element mesh.

fields $\underset{\sim}{\mu}_E^{n+1}$ and $\underset{\sim}{\mu}_D^{n+1}$, shown in Figure 6, are post-processed at $\bar{\boldsymbol{x}}_1 = 0.6842$[m] using

$$\underset{\sim}{\mu}_m^{n+1} = \underline{S}\,(\underset{\sim}{I}\bar{\mu}_m^{n+1} + \bar{\boldsymbol{g}}_m^{n+1} \cdot \Delta\underset{\sim}{\boldsymbol{x}}_m) + \underset{\sim}{\Phi}^{(1)}\eta_m^{(1),n+1}\,, \tag{34}$$

where $\underline{S}$ is the Schur-complement of the microscopic finite element matrices, $\underset{\sim}{I}$ is a column of ones and $\Delta\underset{\sim}{\boldsymbol{x}}_m$ is the microscopic position vector connecting the spatial coordinates to the center of the unit-cell. For more details on the post-processing of the microscopic field $\underset{\sim}{\mu}_m$ by using (34) the reader is referred to [8]. The post-processed microscopic fields also reveal an

excellent agreement with the reference simulation where the maximum of the absolute error is of the order of 10^{-11}. To obtain these results, the values of the coefficients ${}^{J}\bar{\mathcal{C}}_m$ were chosen

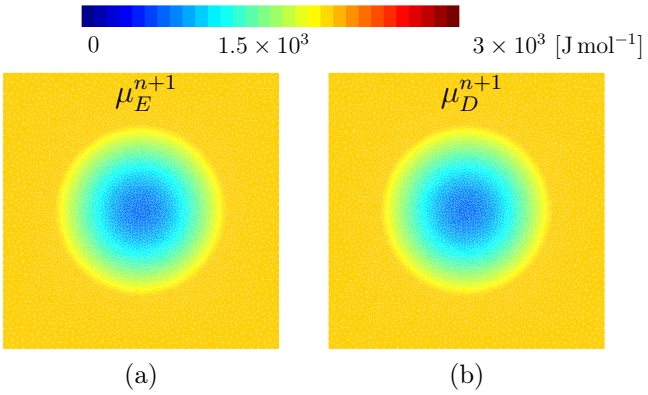


Figure 6: Microscopic chemical potential fields μ_E^{n+1} and μ_D^{n+1} using D_R data-set post-processed at time $\hat{T}$ and $\bar{\boldsymbol{x}}_1 = 0.6842$[m] using equation (34).

based on the information available from the micro-scale, as will be detailed next.

4.4 Numerical Values of the Coefficients ${}^{J}\bar{\mathcal{C}}_m$

In the data-driven simulations, the coefficients ${}^{J}\bar{\mathcal{C}}_m$ in the norm (15) used in the distance function (18) serve two purposes, one is to non-dimensionalize the distance function and second is to give different weights to the parts of the distance function. In a data-set with a large number of data points n_{dp}, the influence of the coefficients ${}^{J}\bar{\mathcal{C}}_m$ is insignificant [31]. However, these coefficients play a crucial role when the data-set has a finite number of data points and if there are inconsistencies in the data-set such as presence of noise or missing points. In a one dimensional problem i.e. $sd = 1$, the solution of the data-driven reduced homogenization exists on a manifold in a $7 + 2\mathcal{N}_q$ dimensional space, which is computationally intractable to fill in completely. Instead, in data-driven simulations, sparse data-sets are used and the coefficients in the distance function should be selected carefully to achieve minimum error with a limited number of iterations.

The total number of coefficients can be reduced by grouping them according to their "classical" thermodynamic conjugacy. In the norm (15) $(\dot{\bar{\mu}}_m, \dot{\bar{c}}_m)$ and $(\bar{\boldsymbol{g}}_m, \bar{\boldsymbol{j}}_m)$ are the conjugate quantities. The coefficient which goes along with one of the conjugate quantities should be equal to the inverse of the other. Some entries of the diagonal matrices ${}^{5}\underline{\mathcal{C}}_m$ and ${}^{6}\underline{\mathcal{C}}_m$ can be neglected if the activity of a particular enrichment-variable $\eta_m^{(q),n+1}$ and its rate $\dot{\eta}_m^{(q),n+1}$ is smaller than that of the other enrichment variables.

The values for these coefficients can be selected by using the information, if available, from the micro-scale calculations. The coefficients which go along with the macroscopic variables appearing in the distance function, are selected to be equal to the corresponding coupling terms for the respective macroscopic variable in the macroscopic constitutive equations (5) and (6). The coefficients whose corresponding macroscopic variables do not appear in equations (5) and (6) and the ones with an insignificant value, as compared to the other coefficients, are chosen to be zero. However, a zero weight in the norm (15) eliminates the influence of the corresponding term on the physics of the problem, care must be taken while setting a coefficient equal to zero. For example, data-driven solution might not be representative of a time-dependent mass

diffusion if the coefficient ${}^{9}\overline{\mathcal{C}}_m$ corresponding to $\dot{\bar{c}}$ is zero. Moreover, the stationarity conditions (21)–(28) also have constraints on which variable can be set to zero. According to (21) the coefficients ${}^{1}\overline{\mathcal{C}}_m$, ${}^{2}\overline{\mathcal{C}}_m$, ${}^{3}\overline{\mathcal{C}}_m$ and ${}^{4}\overline{\mathcal{C}}_m$ can not be set to zero simultaneously. Equation (23) and (25), respectively, restrict (${}^{5}\overline{\mathcal{C}}_m$, ${}^{6}\overline{\mathcal{C}}_m$) and (${}^{8}\overline{\mathcal{C}}_m$, ${}^{9}\overline{\mathcal{C}}_m$) to be zero simultaneously. Also, equation (24) do not allow ${}^{7}\overline{\mathcal{C}}_m$ to be zero. The values for these coefficients used in the simulations are given in Table 4.

Table 4: The values of the coefficients appearing in the norm (15) distance function (18).

Coefficient	Value	Units
${}^{1}\overline{\mathcal{C}}_m$	0	$[\mathrm{J}^{-1}\,\mathrm{mol}\,\mathrm{m}^{-3}\,\mathrm{s}^{-1}]$
${}^{2}\overline{\mathcal{C}}_m$	0.534	$[\mathrm{J}^{-1}\,\mathrm{mol}\,\mathrm{m}^{-3}\,\mathrm{s}]$
${}^{3}\overline{\mathcal{C}}_m$	0.6×10^{-4}	$[\mathrm{J}^{-1}\,\mathrm{mol}\,\mathrm{m}^{-1}\,\mathrm{s}^{-1}]$
${}^{4}\overline{\mathcal{C}}_m$	0	$[\mathrm{J}^{-1}\,\mathrm{mol}\,\mathrm{m}^{-1}\,\mathrm{s}]$
${}^{5}\overline{\mathcal{C}}_m^{(1)}$	0	$[\mathrm{J}\,\mathrm{mol}^{-1}\,\mathrm{m}^{-3}\,\mathrm{s}^{-1}]$
${}^{6}\overline{\mathcal{C}}_m^{(1)}$	33.46	$[\mathrm{J}\,\mathrm{mol}^{-1}\,\mathrm{m}^{-3}\,\mathrm{s}]$
${}^{7}\overline{\mathcal{C}}_m$	1666	$[\mathrm{J}\,\mathrm{mol}^{-3}\,\mathrm{m}\,\mathrm{s}]$
${}^{8}\overline{\mathcal{C}}_m$	0	$[\mathrm{J}\,\mathrm{mol}^{-3}\,\mathrm{m}^{3}\,\mathrm{s}^{-1}]$
${}^{9}\overline{\mathcal{C}}_m$	1.872	$[\mathrm{J}\,\mathrm{mol}^{-3}\,\mathrm{m}^{3}\,\mathrm{s}]$

To test the selected values of the coefficients, the data-driven simulations were conducted with the data-sets $D_{(R+S)}$ and $\widetilde{D}_{(R+S)}$, where $\widetilde{D}_{(R+S)}$ is obtained by adding Gaussian noise to $D_{(R+S)}$ with signal-to-noise ratio of 30. The relative L_2-error norm, between the chemical potential fields $\bar{\mu}_D^{n+1}$ and $\bar{\mu}_E^{n+1}$, is compared. Both, the data-driven problem and the enriched-continuum problem are actuated by the default boundary condition (29). The results are shown in Figures 12, 13, 14 and 15 in Appendices A and B. It can be observed, that the values of the coefficients selected as proposed above (the black lines in the Figures), for most of the cases, yield the smallest values of the relative L_2-error norms. Also, less iterations k are required for the convergence of the staggered scheme. Similar trends have been seen using different data-sets with different number of data points n_{dp} (not shown here for brevity).

4.5 Noisy Data-Set

Uncertainties during the data-generation steps may result in a noisy data-set, which can affect the final result and the convergence of a data-driven solver. To analyze how the proposed data-driven solver behaves in the presence of the noise in the data, a white Gaussian noise, with a signal-to-noise ratio of 30, is added to each element of the original data-sets D_m which results in a data-sets with noise $\widetilde{D}_m = \{(\widetilde{\bar{\mu}}_m{}', \widetilde{\dot{\bar{\mu}}}_m{}', \widetilde{\boldsymbol{g}}_m{}', \widetilde{\dot{\boldsymbol{g}}}_m{}', \widetilde{\underset{\sim}{\eta}}_m{}', \widetilde{\dot{\underset{\sim}{\eta}}}_m{}', \widetilde{\boldsymbol{j}}_m{}', \widetilde{\bar{c}}_m{}', \widetilde{\dot{\bar{c}}}_m{}')\}_{I=1}^{n_{dp}}$. For the pairs $(\widetilde{\boldsymbol{g}}_m^{n+1}, \widetilde{\boldsymbol{j}}_m^{n+1})$ and $(\widetilde{\dot{\bar{\mu}}}_m^{n+1}, \widetilde{\dot{\underset{\sim}{\eta}}}_m^{(1),n+1}, \widetilde{\dot{\bar{c}}}_m^{n+1})$ in the data-set $\widetilde{D}_R$ the noise in the data is shown in Figure 7.

As can be seen in Figure 8, the relative L_2-error between the macroscopic chemical potential fields $\bar{\mu}_E^{n+1}$ and $\bar{\mu}_D^{n+1}$ increases with the addition of noise in the data-set. However, with the amount of added noise, this error is still reasonably small, see Figure 8 (a). The data-driven reduced homogenization also captures the local quantities adequately in the presence of noise as

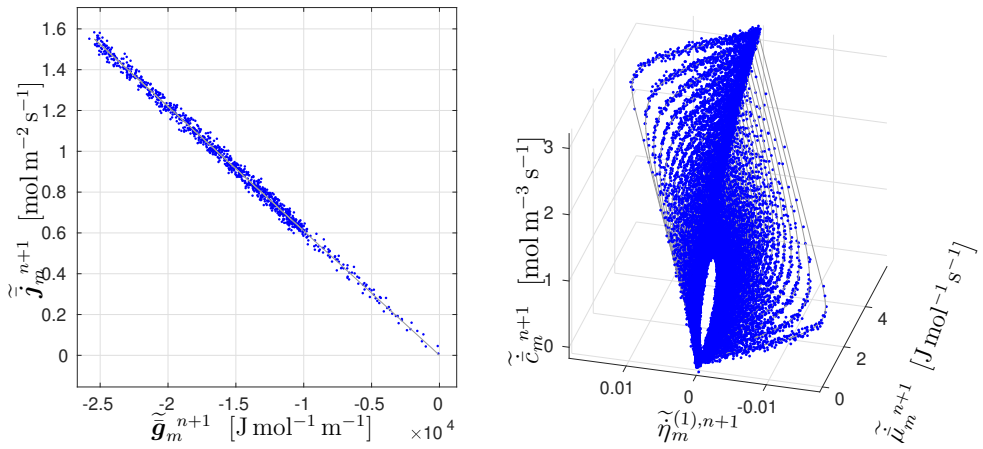


Figure 7: The white Gaussian noise, with the signal-to-noise ratio of 30, was added to the original data-set. From the noisy data-set $\widetilde{D}_R$, (a) the pair $(\widetilde{\bar{\boldsymbol{g}}}_m^{n+1}, \widetilde{\bar{\boldsymbol{j}}}_m^{n+1})$ and (b) the pair $(\widetilde{\dot{\bar{\mu}}}_m^{n+1}, \widetilde{\dot{\bar{\eta}}}_m^{(1),n+1}, \widetilde{\dot{\bar{c}}}_m^{n+1})$ are shown with blue circular markers on top of the corresponding points in the original data-set D_R, shown with light gray lines.

can be seen in Figure 8(b) and (c), where the time evolutions of the macroscopic mass flux $\bar{\boldsymbol{j}}_m^{n+1}$ and the macroscopic concentration $\bar{c}_m^{n+1}$ are evaluated at the macroscopic reference point $\hat{\bar{\boldsymbol{x}}}_m$ computed with the noisy and original data-sets. Different values of the coefficient ${}^J\bar{\mathcal{C}}_m$ in the distance function were also checked with the noisy data-set $\widetilde{D}_{(R+S)}$, see Figure 14 and Figure 15 in Appendix B. In that case, the relative L_2-error increases but there are less differences in the relative L_2-error for different values of the coefficients, which indicates that in the presence of noise the influence of the value of the coefficient is less significant. However, the coefficients ${}^J\bar{\mathcal{C}}_m$ still play an essential role in terms of the convergence towards the expected solution.

To reduce the effect of noise in a data-set and obtain smoother fields, a regression can be performed on neighboring data-points [32]. Noisy data-sets with significant outliers may create a larger problem. In that case, clustering techniques can be used, as proposed in [14]. In the presence of noise, and considering the way in which the data-set is generated, there is a chance that the first and second laws of thermodynamics are not strictly obeyed. To circumvent this problem, [33] formulated the problem in GENERIC framework to guarantee the thermodynamic consistency in data-driven computations.

4.6 Different Data-Sets

For convergence of the data-driven procedure towards a true solution, data-sets used in the simulations should include the states (and their histories) representative for the problem under consideration. To achieve this, a general data-set can be generated by loading the stand-alone micro-scale problem with a complete range of inputs and different loading conditions with different rates. In the following, the performance of the proposed data-driven reduced homogenization is studied on an example where the data-set is obtained under another loading than the final data-driven problem is solved for. To this end the data-set D_S that is obtained by post-processing the enriched-continuum results with a sine loading, as given in Table 3, is used to solve the problem under the ramp loading (29). This provides a challenging test case, because there are (non-physical) negative values of $\bar{\mu}$ present in the data-set, which do not

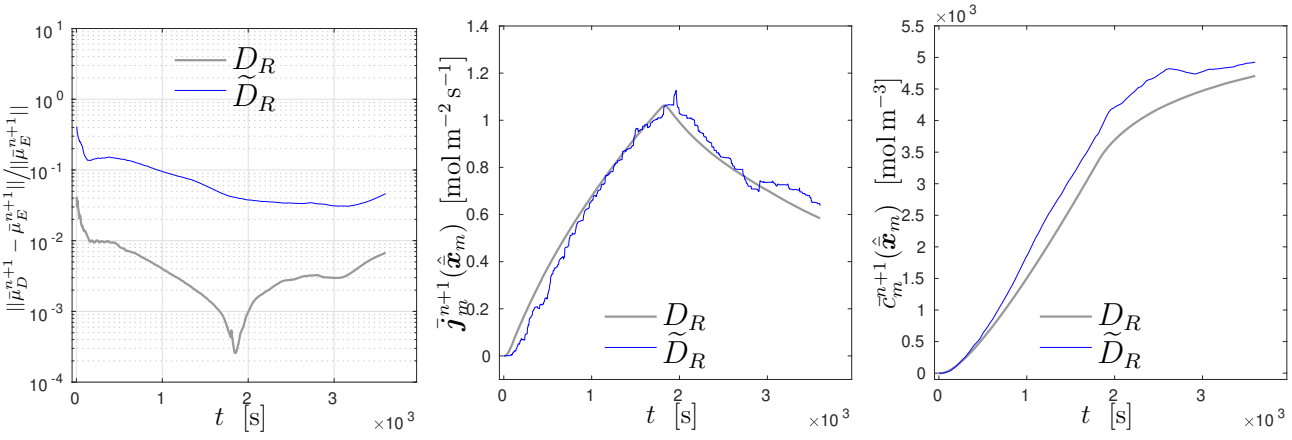


Figure 8: Comparison between the results for the original data-set D_R (gray) and the noisy data-set $\widetilde{D}_R$ (blue) (a) Time evolution of the relative L_2-error for the macroscopic chemical potential $\bar{\mu}^{n+1}$, (b) macroscopic mass flux $\bar{\boldsymbol{j}}_m^{n+1}$ and (c) macroscopic concentration $\bar{c}_m^{n+1}$.

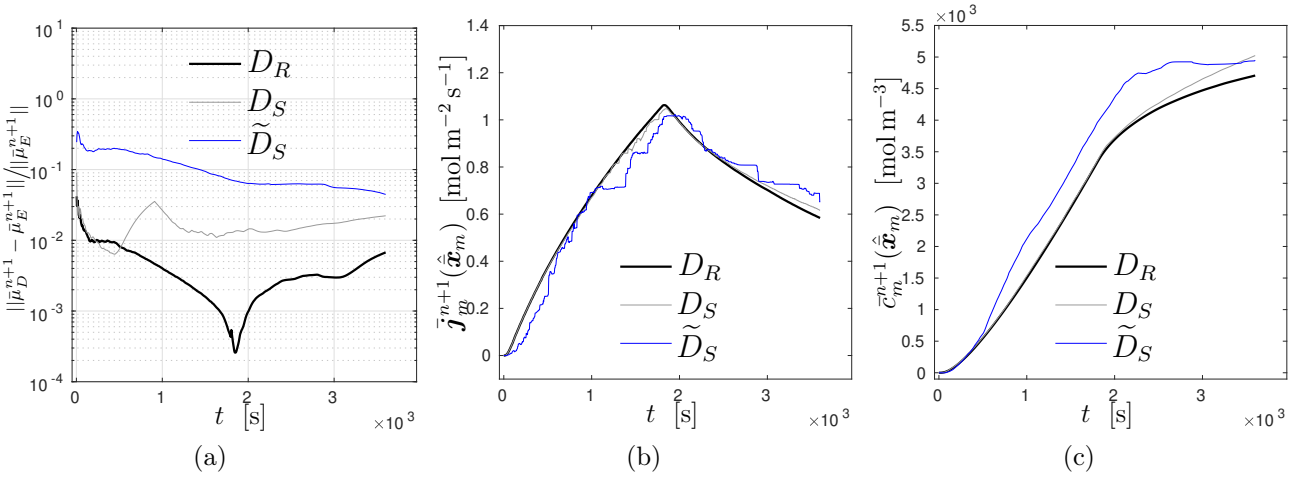


Figure 9: The comparison of the solutions of the data-driven initial boundary value problem with the ramp loading condition (29) using data-sets D_R shown in black, D_S shown in gray and $\widetilde{D}_S$ shown in blue. The time evolution of (a) the relative L_2-error norm, (b) macroscopic mass flux $\bar{\boldsymbol{j}}_m^{n+1}$ at $\hat{\bar{\boldsymbol{x}}}_m$ and (c) the macroscopic concentration $\bar{c}_m^{n+1}$ at $\hat{\bar{\boldsymbol{x}}}_m$.

appear in the solution of the problem with ramp loading. Also the time evolution of the state variables is different from the one present in the data. The results of this analysis are shown in Figure 9.

In this case, the data-driven algorithm is still able to select the representative state z_m^{n+1}, for which the macroscopic mass flux $\bar{\boldsymbol{j}}_m^{n+1}$ and the macroscopic concentration $\bar{c}_m^{n+1}$, evaluated at the reference macroscopic point $\hat{\bar{\boldsymbol{x}}}_m$, are shown in Figure 9 (b) and (c), respectively. There is an increase in the relative L_2-error, as shown in Figure 9(a), when D_S is used instead of D_R and an even larger increase in the case of the noisy data-set $\widetilde{D}_S$. Here, $\widetilde{D}_S$ is obtained by adding white Gaussian noise, with signal-to-noise ratio of 30, to the data-set D_S. Reversely, the data-sets D_R and $\widetilde{D}_R$ cannot be used for the macroscopic initial boundary value problem

under sine loading conditions at all, since the negative values are not present in these data-sets. Therefore this analysis is not presented here.

4.7 Convergence Analysis

The convergence of the proposed data-driven reduced homogenization method with respect to the increase in the number of data-points $n_{dp} = T/\Delta t + 1$ is analyzed here. The data-sets $D_{(RS)}$, generated by the loading condition (33), and $\widetilde{D}_{(RS)}$, with added noise to $D_{(RS)}$, were used in this regard to solve the macroscopic problem with the ramp loading conditions (29). As observed in Figure 10, the increase in the number of data points n_{dp} in the data-set decreases the time averaged relative L_2-error for both the noisy and noiseless data-sets, where, the noisy data-set $\widetilde{D}_{(RS)}$ reveals higher errors than the noiseless data-set $D_{(RS)}$. After a certain data-set coverage, in this case $n_{dp} = 10^3$, the error first reaches a plateau and then slightly increases. This behavior suggests, that, for the problem at hand, the data-set $D_{(RS)}$ has reached its saturation at $n_{dp} = 10^3$ and that it is incomplete by construction, since it does not contain the data-points from all the possible loading conditions with different frequencies. Even with a data-set containing the reference solution, the staggered scheme, adopted for the solution of the double-minimization problem (19), may converge to a local minimum. For an algorithm able to seek the global minimum, see [32]. Next, the computational costs incurred by the

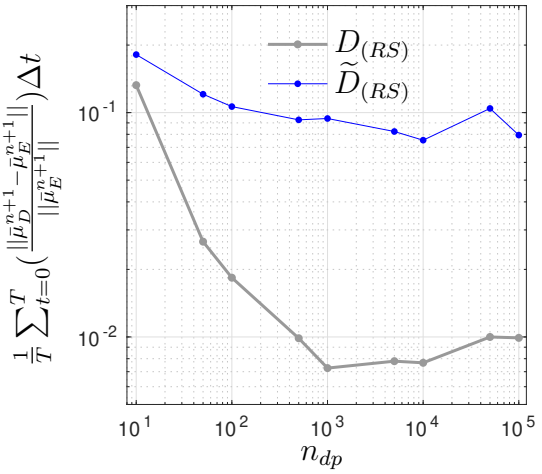


Figure 10: The convergence analysis of the proposed data-driven reduced homogenization upon increasing in the data-set size $D_{(RS)}$ (gray line) and $\widetilde{D}_{(RS)}$ (blue line). $\bar{\mu}_D^{n+1}$ is calculated with data-driven approach and $\bar{\mu}_E^{n+1}$ is calculated with the enriched-continuum approach.

proposed data-driven reduced homogenization are presented with respect to the refined spatial and temporal meshes, larger data-sets and increasing the number of enrichment-variables.

4.8 Computational Cost

Figure 11 presents the computational costs associated with the proposed method using a computer with a Core-i7 4.4 GHz processor and 16Gb memory. Only one variable under consideration is changed at a time and all the other parameters are set to default, as provided in Table 1. The computational time increases with the refined mesh (spatial and temporal), number of data-points n_{dp} and the number of reduced bases $\eta^{(q)}$. As the number of data-points n_{dp} increases, the data-driven problem becomes computationally more expensive because the lookup

search through an array of distance functions in a large data-set is required at each iteration. Similar to the observation in [13], the number of iterations also increases with an increase in the number of data-points. In a large dimensional phase-space z_m^{n+1} such as used here, computationally efficient search schemes, may help to reduce the overall CPU time [34], e.g. a tree search algorithm [29]. For the micro-structure under consideration, as shown in Figure 1, where the total number of reduced bases to adequately represent the micro-scale inertia effects is only six, the computational costs do not increase substantially as the number of reduced bases $\eta^{(q)}$ are increased. By increasing the number of reduced bases, the number of data-points are not increased drastically but number of the quantities in the data-set (8) and the dimension of the phase-space (10) are increased. It should be noted that these simulations were performed for a one-dimensional macroscopic problem, the data-driven formulation is expected to become even more expensive in a two- or a three-dimensional case.

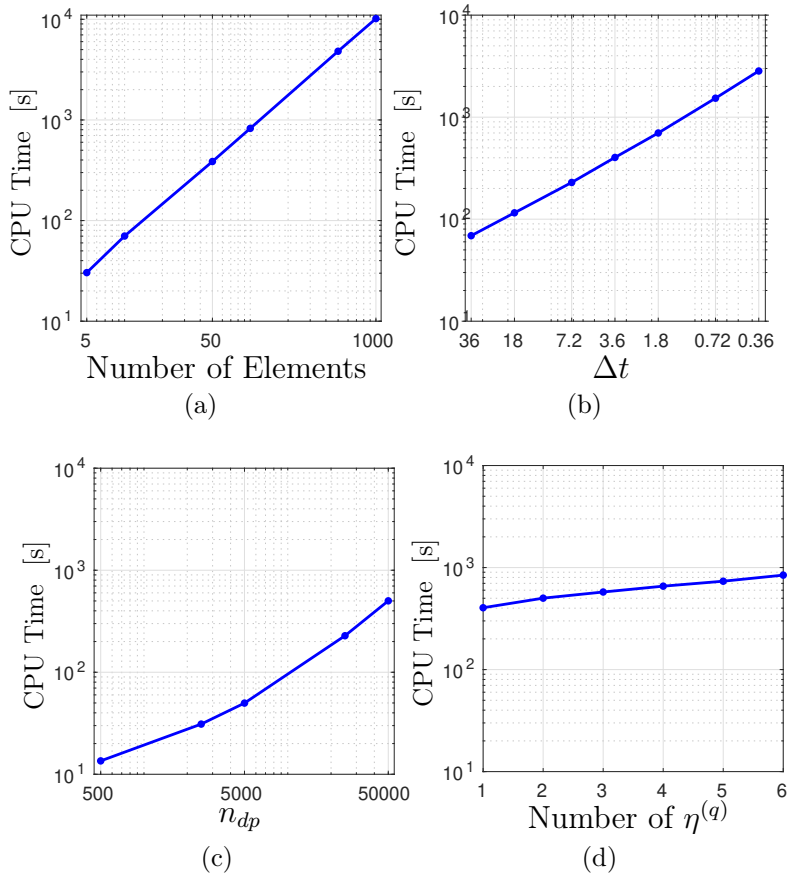


Figure 11: Computational cost of the proposed data-driven reduced homogenization method with (a) number of elements at the macro-scale, (b) the time step size Δt, (c) number of data-points n_{dp} and (d) the number of enrichment-variables $\eta^{(q)}$.

In this paper, the enriched-continuum formulation is only used to generate the data-sets for the data-driven method. The iterative nature of the staggered scheme and the search through the data-set render the data-driven problem more expensive than the enriched-continuum. On the other hand, the enriched-continuum formulation is restricted to linear material models. The data-driven method can prove to be computationally more efficient than the classical homogenization method (without reduction) in the case of a non-linear material behavior where

a closed-form enriched-continuum formulation might not be feasible. This paper provides the basis for that extension relying on the fact that the data-driven solver does not change if the underlying material behavior changes from linear to non-linear. A quantitative comparison can be found in a recently published paper [17], where it is claimed, that for a two-dimensional homogenization problem in elasticity, the computational cost per time step was reduced up to 96.4% with 50^3 data-points and 0.1% accuracy as compared to a classical computational homogenization method.

5 Future Perspectives

The current work establishes a firm foundation, based on data-driven mechanics, for a computationally efficient homogenization methodology for *non-linear history-dependent diffusion* behavior. The model reduction preserves a prominent two-fold advantage of cheap micro-scale calculations and provides the effective history tracking through the enrichment-variables. The challenge for the non-linear case lies in the extraction of a reduced bases set, since an eigenvalue problem is not at hand, complicating the identification of the enrichment-variables. Possible extensions of the proposed data-driven reduced homogenization methodology to the non-linear regime may be inspired by the literature. An example is the nonuniform transformation analysis (NTFA), as proposed in [35], where it is possible to decompose the time-dependent non-linear micro-scale response into a linear and a non-linear part. Then, the micro-scale is divided into several subdomains based on, for example, the material distribution of the constituents. An analytical reduced bases set is found for the non-linear part of the response in each subdomain of the micro-scale unit-cell, while the linear response can be obtained through a simple linear micro-scale calculation. It should be possible to upscale the activity coefficients of the non-linear reduced bases to the macro-scale.

The downside of NTFA is the construction of the analytical reduced bases set. In this context, a more general model reduction method, which relies on reduced bases set by the proper orthogonal decomposition (POD) [36, 37] of the primary micro-scale field, can be used instead. It entails performing micro-scale simulations and collecting the snapshots from time responses of the given micro-structure under various loading scenarios. The responses of these micro-scale simulations, i.e. snapshots of the primary field variable, are collected in a matrix format and the reduced bases set is obtained via a proper orthogonal decomposition. Then, the Galerkin projection onto the reduced bases set can be performed for the micro-scale discrete system of equations providing the evolution equations of the activity coefficients of the reduced bases set, which serve as internal-variables that efficiently capture the history-dependent macroscopic behavior in a data-set. The spatial modes, alone, can not capture the time dependent behavior and their inclusion in the data-set and phase-space should also be avoided, since, it will make the data-driven problem computationally very expensive. A different approach using proper generalized decomposition (PGD) [38] can also be considered. It parameterizes the micro-scale solution in spatial directions, time, constituents and nonlinear behavior. The reduced bases are then constructed iteratively with an alternating direction algorithm.

The structure of the data-driven problems depends on the type of problem under consideration, the terms considered in the distance function, and the form of the evolution equations emerging as a result of the model reduction. The general expressions obtained in (4), (5) and (6) remain valid when the modes are obtained from a different approach, e.g. NTFA, POD, PGD. In non-linear data-driven reduced homogenization, the coefficients ${}^{J}\bar{\mathcal{C}}_m$ in the distance

function might also be approximated by the eigenvalue analysis on the linearized micro-scale material response as presented in section 4.4 or some analytical averaging maybe used technique instead.

The current framework can also be extended to other physical phenomena, where the emergent macro-scale behavior is caused by the underlying microstructure. For example, the homogenized response of locally resonant acoustic metamaterials in the linear regimes was proposed in [10] in which an enrichment-variable emerges at the macro-scale by performing model reduction at the micro-scale. The evolution equation of these enrichment-variables is a second-order ordinary differential equation. To solve this problem by data-driven formulation, a combination of a data-driven approach for dynamics problems, as presented in [15], and history dependent materials using internal-variables, as proposed in [16], can be used. Similar extensions apply to other multi-physics phenomena, as considered in [12], where an enriched-continuum formulation for mass diffusion was coupled to mechanics. There, taking the advantage of the linear material properties and the relaxed separation of scales, a coupled eigenvalue problem was solved to obtain the enrichment-variables representing the history-dependent coupled chemo-mechanical behavior at the macro-scale.

One of the major limitations of any data-driven method, whether a model-free approach [13, 28] or the one which is used to develop e.g. a surrogate micro-scale model [19, 20], is the availability of sufficiently broad and representative data-sets. The data-sets used in the current work were generated by post-processing the material states from the solution of a reduced homogenization scheme. However, such efficient models are approximations, which rely on a number of assumptions, and in most cases are difficult to obtain. In other cases, generation of an adequate data-set, through numerical/physical experiment becomes a tedious task especially if the nonlinearities are involved. One remedy could be the use of on-the-fly micro-scale simulations to fill a gap in the sparse data-set. At the same time, running hundreds and thousands of micro-scale simulations to generate extensive data-sets is expensive and time consuming. This obstacle can be overcome by making the data-sets available in online repositories, that will save a lot of resources and time during the data-generation stage. Hence, the data-generation stage can be replaced by a more efficient data-acquisition stage. In this regard, the Material Genome® [39] project and its application in finding novel materials using data-driven methods [40] may serve as an inspiration.

6 Conclusions

In this work, a *data-driven reduced homogenization* method is proposed for capturing the non-Fickian and history-dependent transient diffusion behavior in heterogeneous materials. It is built on the enriched-continuum formulation, developed earlier in [8] for linear material behavior exhibiting a relaxed separation of scales, and the data-driven mechanics, proposed in [13]. An enriched-continuum is a macroscopically homogenized description of a heterogeneous material in which the transient effects emerging from the micro-scale, through a model reduction, are captured by enrichment-variables at the macro-scale. For linear material properties and the relaxed separation of scales, the model reduction at the micro-scale can be performed by using the eigenvectors, obtained via the solution of an eigenvalue problem at the micro-scale, as the reduced bases. The data-driven method seeks a physical-state of the material closest to a point in the data-set, which can be obtained by experiments (in this work micro-scale simulations). Following [16], instead of using the whole history of the microscopic primary field variables,

the enrichment-variables are used to efficiently keep track of the history-dependent state of the macroscopic behavior. The data-driven reduced homogenization uses a staggered solution scheme [28] to tackle the combinatorial complexity of a mixed, continuous and discrete, double-minimization problem, in which the state and the closest point in the data-set, which minimizes a global distance function, are found iteratively. The macroscopic compatibility is imposed directly and the macroscopic mass balance law is imposed through Lagrange multipliers.

Numerical examples are conducted for a macroscopically isotropic response in a one-dimensional domain, showing an adequate performance and robustness of the proposed methodology. A two-dimensional micro-scale problem, under one-dimensional loading conditions, is considered to obtain the macroscopic quantities and to provide the input for generating the data-sets. The enriched-continuum problem is used as a reference solution, and to generate the data-sets by post-processing the primary field and its gradient, as well. The actual generation of the data-set is done using stand-alone micro-scale simulations with different loading conditions having different frequencies. The obtained point in the data-set can then be added to the already existing data-set. The large number of coefficients in the distance function make the current data-driven problem more prone to numerical errors and instabilities, so a methodology is presented to carefully select the numerical values of these coefficients, based on the information available from the micro-scale simulations. By doing so, a substantial decrease in the number of iterations and numerical error was obtained. Data-driven reduced homogenization captures the homogenized enriched-continuum response very well and also the post-processed micro-scale fields are in close agreement with each other. The proposed data-driven approach performs adequately in the presence of noise in the data-set and also in the case when a different data-set is used. Finally, by increasing the number of points in the data-set the error is reduced substantially, however at the expense of an increased number of iterations and computational effort. Obviously, the present work can be extended to an-isotropic macroscopic and two/three-dimensional behavior.

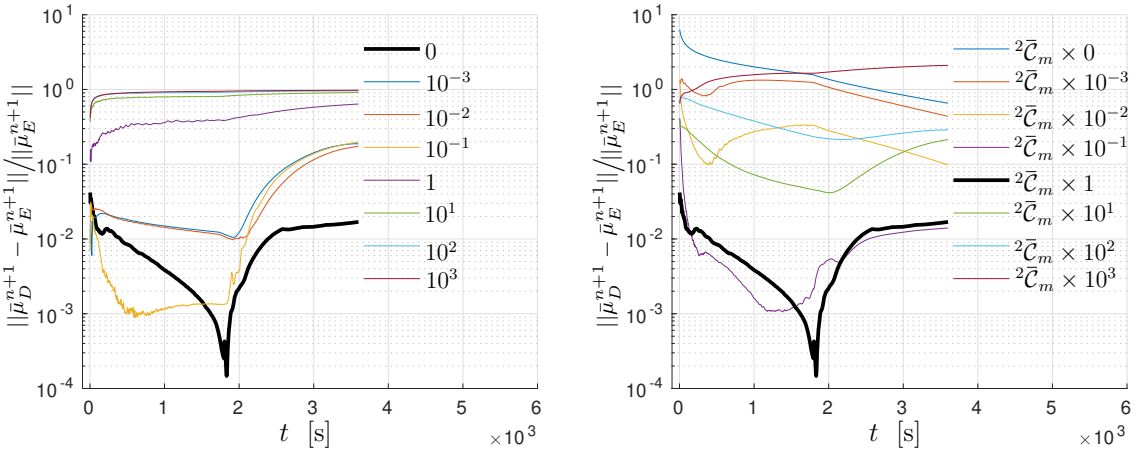
Acknowledgments and Funding: Support for this research was provided by the European Commission through an Erasmus Mundus grant in the framework of the Simulation in Engineering and Entrepreneurship Development (SEED) program. The SEED program is an initiative of 8 universities Partners, managed by EACEA and financed by the European Commission with grant Ref. 2013-0043.

Competing Interests: The authors declare that they have no competing interests.

Author's Contribution: LS and TH initiated the data-driven approach, VK and MG contributed to homogenization and model reduction. AW contributed to everything.

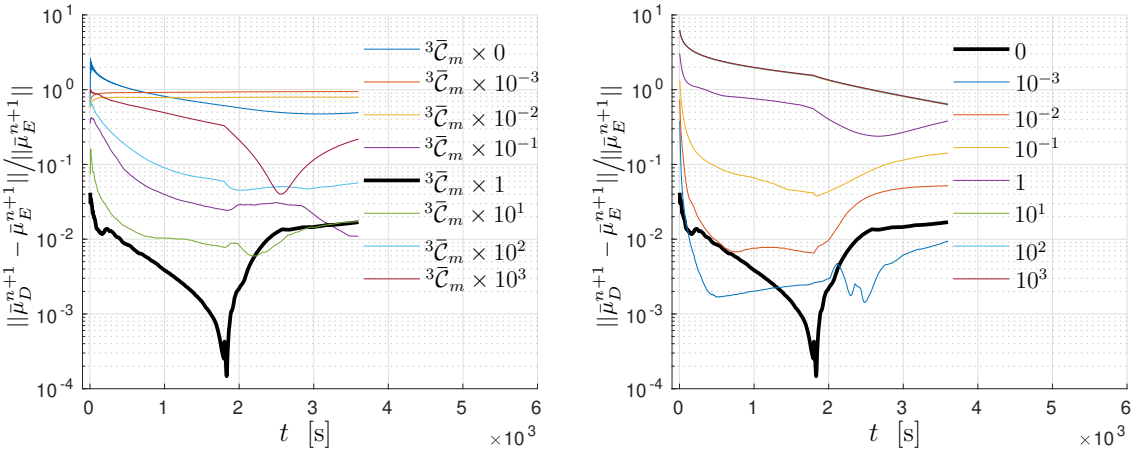
A Relative L_2-Error for Different Coefficient Values Using Data-Set $D_{(R+S)}$

The performance of the data-driven reduced homogenization is checked with different values of coefficients ${}^{J}\overline{\mathcal{C}}_m$, appearing in the definition of the norm (15) for the distance function (18), with the data-set $D_{(R+S)}$. The results are presented in Figure 12 and 13.



(a) Different numerical values for ${}^1\bar{\mathcal{C}}_m$

(b) Different numerical values for ${}^2\bar{\mathcal{C}}_m$



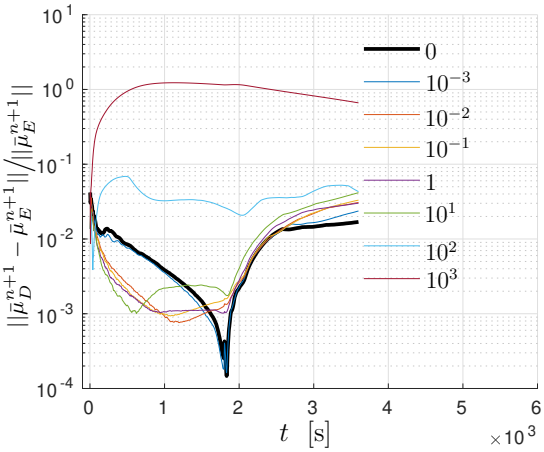
(c) Different numerical values for ${}^3\bar{\mathcal{C}}_m$

(d) Different numerical values for ${}^4\bar{\mathcal{C}}_m$

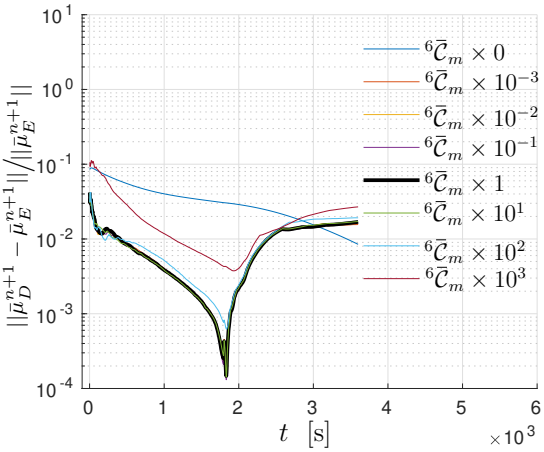
Figure 12: For the coefficients ${}^1\bar{\mathcal{C}}_m$ ${}^2\bar{\mathcal{C}}_m$ ${}^3\bar{\mathcal{C}}_m$and ${}^4\bar{\mathcal{C}}_m$ the time evolution of the relative L_2-error norm, calculated as $||\bar{\mu}_D^{n+1} - \bar{\mu}_E^{n+1}||/||\bar{\mu}_E^{n+1}||$, where $\bar{\mu}_D^{n+1}$ is the chemical potential field obtained by the data-driven reduced homogenization (proposed) using the data-set $D_{(R+S)}^{n+1}$ and $\bar{\mu}_E^{n+1}$ is the chemical potential field obtained by the enriched-continuum formulation (reference) under boundary conditions (29). The default values and the units of the coefficients are given in Table 4. The relative L_2-error computed with the proposed default value for the coefficients ${}^J\bar{\mathcal{C}}_m$ is marked with the black lines.

B Relative L_2-Error for Different Coefficient Values Using Data-Set with Noise $\widetilde{D}_{(R+S)}$

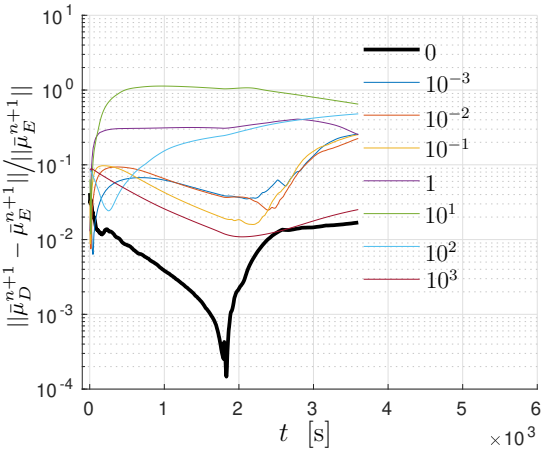
The performance of the data-driven reduced homogenization is checked with different values of coefficients ${}^J\bar{\mathcal{C}}_m$, appearing in the definition of the norm (15) for the distance function (18), with a noisy data-set $\widetilde{D}_{(R+S)}$, the result is shown in Figure 14 and 15. An increase in the relative L_2-error is observed as compared to the data without noise $D_{(R+S)}$, With the introduction of the noise, the error is comparatively less influenced by the numerical values of the coefficients and more by the noisiness of the data.



(a) Different numerical values for ${}^5\bar{\mathcal{C}}_m^{(1)}$



(b) Different numerical values for ${}^6\bar{\mathcal{C}}_m^{(1)}$



(c) Different numerical values for ${}^8\bar{\mathcal{C}}_m$

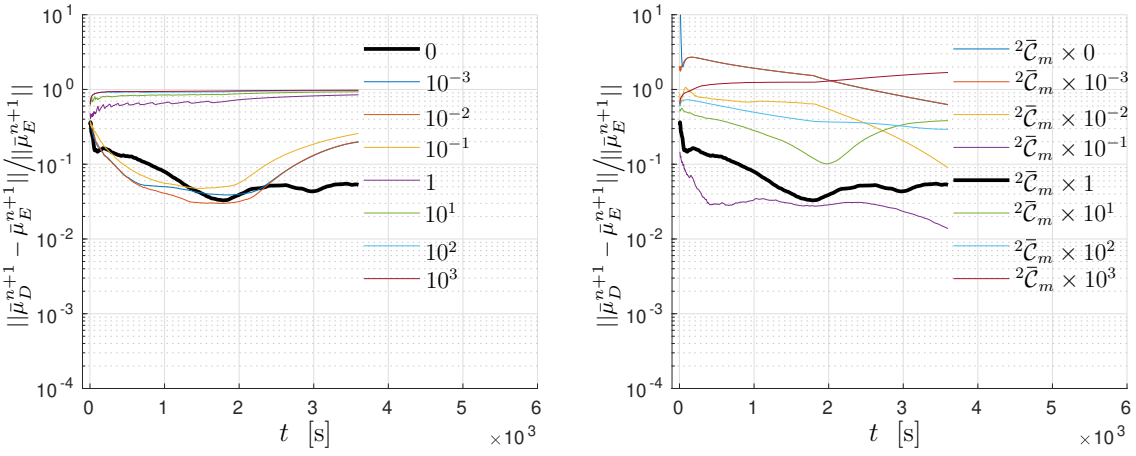
Figure 13: For the coefficients ${}^5\bar{\mathcal{C}}_m^{(1)}$ ${}^6\bar{\mathcal{C}}_m^{(1)}$ and ${}^8\bar{\mathcal{C}}_m$ the time evolution of the relative L_2-error norm, calculated as $||\bar{\mu}_D^{n+1} - \bar{\mu}_E^{n+1}||/||\bar{\mu}_E^{n+1}||$, where $\bar{\mu}_D^{n+1}$ is the chemical potential field obtained by the data-driven reduced homogenization (proposed) using the data-set $D_{(R+S)}^{n+1}$ and $\bar{\mu}_E^{n+1}$ is the chemical potential field obtained by the enriched-continuum formulation (reference) under boundary conditions (29). The default values and the units of the coefficients are given in Table 4. The relative L_2-error computed with the proposed default value for the coefficients ${}^J\bar{\mathcal{C}}_m$ is marked with the black lines.

References

[1] Shin-ichi Nishimura, Genki Kobayashi, Kenji Ohoyama, Ryoji Kanno, Masatomo Yashima, and Atsuo Yamada. Experimental visualization of lithium diffusion in LixFePO4. *Nature materials*, 7(9):707–711, 2008.

[2] Robert W Balluffi, Samuel M Allen, and W Craig Carter. *Kinetics of materials*. John Wiley & Sons, 2005.

[3] Erin M Johnson, David A Berk, Rakesh K Jain, and William M Deen. Hindered diffusion



(a) Different numerical values for ${}^1\bar{\mathcal{C}}_m^{(1)}$

(b) Different numerical values for ${}^2\bar{\mathcal{C}}_m$

(c) Different numerical values for ${}^3\bar{\mathcal{C}}_m$

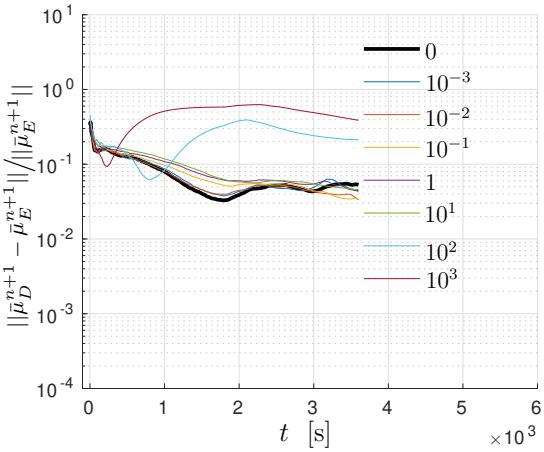
(d) Different numerical values for ${}^4\bar{\mathcal{C}}_m$

Figure 14: For the coefficients ${}^1\bar{\mathcal{C}}_m$ ${}^2\bar{\mathcal{C}}_m$ ${}^3\bar{\mathcal{C}}_m$and ${}^4\bar{\mathcal{C}}_m$ the time evolution of the relative L_2-error norm, calculated as $||\bar{\mu}_D^{n+1} - \bar{\mu}_E^{n+1}||/||\bar{\mu}_E^{n+1}||$, where $\bar{\mu}_D^{n+1}$ is the chemical potential field obtained by the data-driven reduced homogenization (proposed) using the data-set $\widetilde{D}_{(R+S)}^{n+1}$ and $\bar{\mu}_E^{n+1}$ is the chemical potential field obtained by the enriched-continuum formulation (reference) under boundary conditions (29). The default values and the units of the coefficients are given in Table 4. The relative L_2-error computed with the proposed default value for the coefficients ${}^J\bar{\mathcal{C}}_m$ is marked with the black lines.

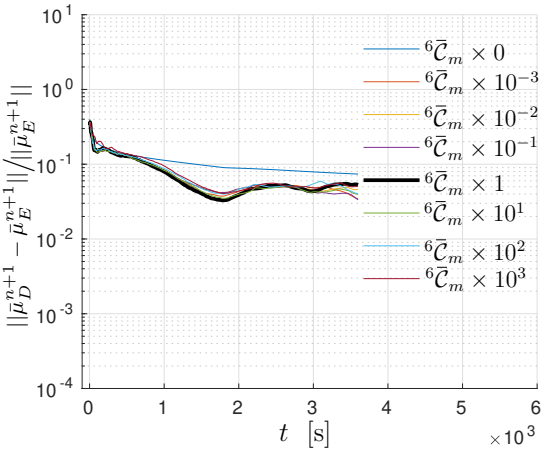
in agarose gels: test of effective medium model. *Biophysical journal*, 70(2):1017–1023, 1996.

[4] F. Larsson, K. Runesson, and F. Su. Variationally consistent computational homogenization of transient heat flow. *International Journal for Numerical Methods in Engineering*, 81(13):1659–1686, 2010.

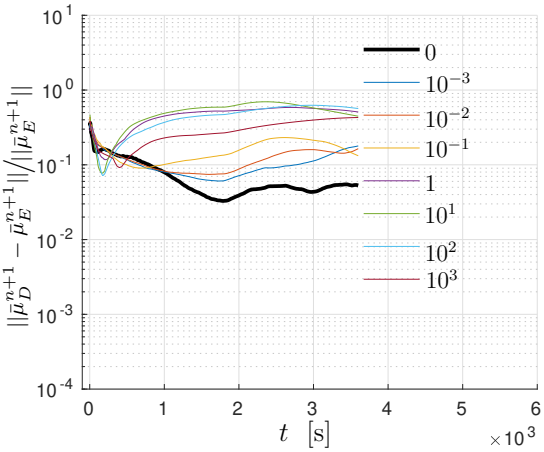
[5] G. R. Ramos, T. dos Santos, and R. Rossi. An extension of the Hill–Mandel principle for transient heat conduction in heterogeneous media with heat generation incorporating finite RVE thermal inertia effects. *International Journal for Numerical Methods in Engineering*, 111(6):553–580, 2017.



(a) Different numerical values for ${}^5\bar{\mathcal{C}}_m^{(1)}$



(b) Different numerical values for ${}^6\bar{\mathcal{C}}_m^{(1)}$



(c) Different numerical values for ${}^8\bar{\mathcal{C}}_m$

Figure 15: For the coefficients ${}^5\bar{\mathcal{C}}_m^{(1)}$ ${}^6\bar{\mathcal{C}}_m^{(1)}$ and ${}^8\bar{\mathcal{C}}_m$ the time evolution of the relative L_2-error norm, calculated as $||\bar{\mu}_D^{n+1} - \bar{\mu}_E^{n+1}||/||\bar{\mu}_E^{n+1}||$, where $\bar{\mu}_D^{n+1}$ is the chemical potential field obtained by the data-driven reduced homogenization (proposed) using the data-set $\widetilde{D}_{(R+S)}^{n+1}$ and $\bar{\mu}_E^{n+1}$ is the chemical potential field obtained by the enriched-continuum formulation (reference) under boundary conditions (29). The default values and the units of the coefficients are given in Table 4. The relative L_2-error computed with the proposed default value for the coefficients ${}^J\bar{\mathcal{C}}_m$ is marked with the black lines.

[6] J. L. Auriault, C. Boutin, and C. Geindreau. *Homogenization of coupled phenomena in heterogenous media*, volume 149. John Wiley & Sons, 2010.

[7] L. Brassart and L. Stainier. Effective transient behaviour of heterogeneous media in diffusion problems with a large contrast in the phase diffusivities. *Journal of the Mechanics and Physics of Solids*, 124:366–391, 2019.

[8] A. Waseem, T. Heuzé, L. Stainier, M. G. D. Geers, and V. G. Kouznetsova. Model reduction in computational homogenization for transient heat conduction. *Computational Mechanics*, 65(1):249–266, 2020.

[9] Roy Craig and Mervyn Bampton. Coupling of substructures for dynamic analyses. *AIAA journal*, 6(7):1313–1319, 1968.

[10] A. Sridhar, V. G. Kouznetsova, and M. G. D. Geers. Homogenization of locally resonant acoustic metamaterials towards an emergent enriched continuum. *Computational mechanics*, 57(3):423–435, 2016.

[11] B. D. Coleman and M. E. Gurtin. Thermodynamics with internal state variables. *The Journal of Chemical Physics*, 47(2):597–613, 1967.

[12] A. Waseem, T. Heuzé, L. Stainier, M. G. D. Geers, and V. G. Kouznetsova. Enriched continuum for multi-scale transient diffusion coupled to mechanics. *Advanced Modeling and Simulation in Engineering Sciences*, 7(1):1–32, 2020.

[13] T. Kirchdoerfer and M. Ortiz. Data-driven computational mechanics. *Computer Methods in Applied Mechanics and Engineering*, 304:81–101, 2016.

[14] T. Kirchdoerfer and M. Ortiz. Data driven computing with noisy material data sets. *Computer Methods in Applied Mechanics and Engineering*, 326:622–641, 2017.

[15] T. Kirchdoerfer and M. Ortiz. Data-driven computing in dynamics. *International Journal for Numerical Methods in Engineering*, 113(11):1697–1710, 2018.

[16] R. Eggersmann, T. Kirchdoerfer, S. Reese, L. Stainier, and M. Ortiz. Model-free data-driven inelasticity. *Computer Methods in Applied Mechanics and Engineering*, 350:81–99, 2019.

[17] R. Xu, J. Yang, W. Yan, Q. Huang, G. Giunta, S. Belouettar, H. Zahrouni, T. B. Zineb, and H. Hu. Data-driven multiscale finite element method: From concurrence to separation. *Computer Methods in Applied Mechanics and Engineering*, 363:112893, 2020.

[18] J. Yang, R. Xu, H. Hu, Q. Huang, and W. Huang. Structural-genome-driven computing for composite structures. *Composite Structures*, 215:446–453, 2019.

[19] J. Yvonnet, D. Gonzalez, and Q. C. He. Numerically explicit potentials for the homogenization of nonlinear elastic heterogeneous materials. *Computer Methods in Applied Mechanics and Engineering*, 198(33-36):2723–2737, 2009.

[20] R. Ibañez, D. Borzacchiello, J. V. Aguado, E. Abisset-Chavanne, E. Cueto, P. Ladevèze, and F. Chinesta. Data-driven non-linear elasticity: constitutive manifold construction and problem discretization. *Computational Mechanics*, 60(5):813–826, 2017.

[21] F. Ghavamian and A. Simone. Accelerating multiscale finite element simulations of history-dependent materials using a recurrent neural network. *Computer Methods in Applied Mechanics and Engineering*, 357:112594, 2019.

[22] F. E. Bock, R. C. Aydin, C. J. Cyron, N. Huber, S. R. Kalidindi, and B. Klusemann. A review of the application of machine learning and data mining approaches in continuum materials mechanics. *Frontiers in Materials*, 6:110, 2019.

[23] F. J. Montáns, F. Chinesta, R. Gómez-Bombarelli, and J. N. Kutz. Data-driven modeling and learning in science and engineering. *Comptes Rendus Mécanique*, 347(11):845–855, 2019.

[24] D. Huang, J. N. Fuhg, C. Weißenfels, and P. Wriggers. A machine learning based plasticity model using proper orthogonal decomposition. *arXiv preprint arXiv:2001.03438*, 2020.

[25] F. Chinesta, E. Cueto, E. Abisset-Chavanne, J. L. Duval, and F. El Khaldi. Virtual, digital and hybrid twins: a new paradigm in data-based engineering and engineered data. *Archives of Computational Methods in Engineering*, pages 1–30, 2018.

[26] L. T.K. Nguyen, M. Rambausek, and M. A. Keip. Variational framework for distance-minimizing method in data-driven computational mechanics. *Computer Methods in Applied Mechanics and Engineering*, 365:112898, 2020.

[27] A. Waseem, T. Heuzé, M. G. D. Geers, and V. G. Kouznetsova. Two-scale analysis of transient diffusion problems through a homogenized enriched continuum. *European Journal of Mechanics-A/Solids*. doi: https://doi.org/10.1016/j.euromechsol.2021.104212.

[28] L. Stainier, A. Leygue, and M. Ortiz. Model-free data-driven methods in mechanics: material data identification and solvers. *Computational Mechanics*, 64(2):381–393, 2019.

[29] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. MIT press, 2009.

[30] D Roca, Oriol Lloberas-Valls, Juan Cante, and Javier Oliver. A computational multiscale homogenization framework accounting for inertial effects: Application to acoustic metamaterials modelling. *Computer methods in applied mechanics and engineering*, 330:415–446, 2018.

[31] S. Conti, S. Müller, and M. Ortiz. Data-driven problems in elasticity. *Archive for Rational Mechanics and Analysis*, 229(1):79–123, 2018.

[32] Y. Kanno. Simple heuristic for data-driven computational elasticity with material data involving noise and outliers: a local robust regression approach. *Japan Journal of Industrial and Applied Mathematics*, 35(3):1085–1101, 2018.

[33] D. González, F. Chinesta, and E. Cueto. Thermodynamically consistent data-driven computational mechanics. *Continuum Mechanics and Thermodynamics*, 31(1):239–253, 2019.

[34] Robert Eggersmann, Laurent Stainier, Michael Ortiz, and Stefanie Reese. Efficient data structures for model-free data-driven computational mechanics. *arXiv preprint arXiv:2012.00357*, 2020.

[35] V. Sepe, S. Marfia, and E. Sacco. A nonuniform TFA homogenization technique based on piecewise interpolation functions of the inelastic field. *International Journal of Solids and Structures*, 50(5):725–742, 2013.

[36] E Monteiro, Julien Yvonnet, and Qi-Chang He. Computational homogenization for nonlinear conduction in heterogeneous materials using model reduction. *Computational Materials Science*, 42(4):704–712, 2008.

[37] F. Fritzen and M. Leuschner. Reduced basis hybrid computational homogenization based on a mixed incremental formulation. *Computer Methods in Applied Mechanics and Engineering*, 260:143–154, 2013.

[38] Hajer Lamari, Amine Ammar, Patrice Cartraud, Grégory Legrain, Francisco Chinesta, and Frédéric Jacquemin. Routes for efficient computational homogenization of nonlinear materials using the proper generalized decompositions. *Archives of Computational methods in Engineering*, 17(4):373–391, 2010.

[39] Yingli Liu, Chen Niu, Zhuo Wang, Yong Gan, Yan Zhu, Shuhong Sun, and Tao Shen. Machine learning in materials genome initiative: A review. *Journal of Materials Science & Technology*, 2020.

[40] Paul Raccuglia, Katherine C Elbert, Philip DF Adler, Casey Falk, Malia B Wenny, Aurelio Mollo, Matthias Zeller, Sorelle A Friedler, Joshua Schrier, and Alexander J Norquist. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533(7601): 73–76, 2016.