# Adaptive Goal-oriented Data Sampling in Data-Driven Computational Mechanics

Anna Gorgogianni, Konstantinos Karapiperis, Laurent Stainier, Michael Ortiz, José E Andrade

# Adaptive Goal-oriented Data Sampling in Data-Driven Computational Mechanics

**Anna Gorgogianni[1], Konstantinos Karapiperis[2], Laurent Stainier[3], Michael Ortiz[4], José E. Andrade[5]**

**Abstract:** Data-Driven (DD) computing is an emerging field of Computational Mechanics, motivated by recent technological advances in experimental measurements, the development of highly predictive computational models, advances in data storage and data processing, which enable the transition from a material data-scarce to a material data-rich era. The predictive capability of DD simulations is contingent on the quality of the material data set, i.e. its ability to closely sample all the strain-stress states in the phase space of a given mechanical problem. In this study, we develop a methodology for increasing the quality of an existing material data set through iterative expansions. Leveraging the formulation of the problems treated with the DD paradigm as distance minimization problems, we identify regions in phase space with poor data coverage, and target them with additional experiments or lower-scale simulations. The DD solution informs the additional experiments so that they can provide better coverage of the phase space of a given application.

We first illustrate the convergence properties of the approach through a DD finite element simulation of a linear elastic cylinder under triaxial compression. The same numerical experiment is then performed on a specimen of Hostun sand, a material with complex history-dependent behavior. Data sampling is performed with Level-Set Discrete Element Method (LS-DEM) calculations of unit cells representative of this granular material, subjected to loading paths determined by the proposed method. It is shown that this adaptive expansion of the data set, tailored for a particular application, leads to convergent and accurate DD predictions, without the computational cost of using large databases with potentially redundant or low-quality data.

**Keywords:** data-driven computing, multiscale modeling, data acquisition strategies

## 1 Introduction

In computations of the mechanical behavior of material bodies, constitutive modeling has been an essential task; together with a set of constraints and conservation laws, such as compatibility and equilibrium, the constitutive equations provide the closure relations to the associated boundary value problems. The predictive capability of this conventional computing paradigm is largely governed by the accuracy of the employed constitutive model in predicting material behavior. However, such predictions have been traditionally challenged by the scarcity of experimental material data, leading to material models that had to extrapolate far beyond the range of data used for their calibration. Other challenges and sources of uncertainty in this process include the non-unique choice of the functional forms of material laws. Moreover, the calibration procedure can be far from trivial, particularly for models aiming to capture complex material behavior and introducing a significant number of parameters, some of which may lack physical interpretation.

Motivated by progress in Data Science, and aiming to reduce the aforementioned sources of bias and empiricism in the classical material modeling step, various alternatives based on machine learning and data-driven techniques have been proposed in the literature. One could view such attempts as having largely followed two different directions. The first direction introduces surrogates for constitutive

[1]Division of Engineering and Applied Science, California Institute of Technology, Pasadena, CA 91125, USA
[2]Mechanics & Materials Laboratory, Department of Mechanical and Process Engineering, ETH Zürich, 8092 Zürich, Switzerland
[3]Institut de Recherche en Génie Civil et Mécanique (GeM), Ecole Centrale de Nantes, 44321 Nantes cedex 3, France
[4]Graduate Aerospace Laboratories, California Institute of Technology, Pasadena, CA 91125, USA
[5]Division of Engineering and Applied Science, California Institute of Technology, Pasadena, CA 91125, USA. Corresponding author

laws learnt directly from the available material data, this way eliminating the bias introduced by the modeler when proposing a specific stress-strain relationship. Early efforts in this first class of models include the training of neural networks (NN) with experimental data for predictions of mechanical behavior [9]. In [10], the performance of NN-based constitutive models of elastoplastic behavior in finite element analysis was investigated, and improved convergence compared to the use of classical elastoplastic models was reported. Other examples include successful applications of deep learning [20, 21], with resulting material models that can be predictive beyond the training data set, as well as computationally efficient to incorporate in the analysis.

The second direction, initiated by [16], does not resort to any explicit material model. With the mathematical formulation of the Data-Driven distance-minimizing method [16] as well as its extensions [2, 5, 6, 17, 18], predictions of mechanical behavior can be made by directly incorporating the material data in the analysis, in lieu of material models. Material data are generated through physical experiments as well as physics-based lower-scale computational models, this way avoiding the phenomenology of classical constitutive models. No extrapolation or prediction beyond a given data set is attempted. The boundary value problems of mechanics are solved by reformulating them as distance minimization problems between two sets; the constraint set, consisting of the local strain-stress states satisfying essential constraints, conservation laws and boundary conditions, and the material data set, consisting of a finite collection of strain-stress states obtained experimentally. The data-driven solution consists of pairs of mechanically admissible strain-stress states and their corresponding nearest-neighboring states from the material data set. The distance between the constraint set and the material data set quantifies the error made when seeking to satisfy as best as possible the conservation laws and the captured material behavior.

A major source of this error can be the issue of data scarcity; the experiments providing material data may be insufficient to cover all mechanical strain-stress states of a particular application. Another possibility is a low-quality material database, created from experiments under loading paths not pertinent to the considered mechanical problem. The process of generating material data, also referred to as data or phase space sampling, should thus aim to minimize the outlined above sources of error in data-driven computations, and is the goal of this study.

Various approaches to phase space sampling have been proposed in the literature [7, 13, 16, 19], however, only few of them have been implemented. Phase space sampling strategies can be broadly categorized into two types, depending on the existence or not of a priori information for the considered mechanical problem. These two approaches are termed as offline and on-the-fly sampling respectively [13]. In the case of offline sampling, the preexisting information could consist of experimental measurements of the local strains or strain histories at discrete points of the deforming specimen. The material data set can then be completed by lower-scale computations, e.g. molecular dynamics, discrete element method, lower-scale finite element method, that will provide the remaining components of the state, i.e. local stress response. Such prior information though is not required to guide the process of phase space sampling. It is indeed a salient feature of the Data-Driven distance-minimizing method [16], which is adopted in this study, that it can provide real-time information on the errors, expressed in terms of the distance between the mechanically admissible and the material strain-stress states. Detection of large error values, at a given material point and a given time step of the data-driven simulation, can be used as a criterion for an on-the-fly expansion of the data set [13], until the errors become sufficiently small.

Following the concept of the on-the-fly data sampling strategy, the present study introduces an efficient way to populate an existing database focused on the demands of a particular application, i.e. goal-oriented. The way to populate the database does not involve any data-fitting or other empirical process, but is informed by the data-driven solution of the mechanical states, which are those respecting the physical laws. An initial DD simulation of a considered problem is performed with an existing material database. The material points whose strain-stress states are not well-covered by the database are then identified by analyzing the error distribution in the data-driven simulation.

These material points are targeted for additional lower-scale experiments, providing higher-quality material data to expand the database with. The process continues with more such iterations of decoupled DD simulations and lower-scale experiments at the material point level, until the error of the DD simulations in satisfying conservation laws and material behavior is minimized.

The paper is organized as follows: Section 2 provides a brief review of the data-driven distance-minimizing method for the case of history-dependent material behavior, which is the main focus of the present study. Section 3 analyzes the proposed adaptive goal-oriented data sampling strategy. The remaining of the paper is devoted to numerical applications of the method, initially for the case of linear elastic material behavior and then for the case of complex nonlinear material behavior. We conclude with a discussion on the performance of the proposed framework in aiding convergence and increasing the accuracy of data-driven predictions, the computational benefits of the data-driven method, as well as the greater paradigm shift it has introduced.

# 2 Data-Driven distance-minimizing method

## 2.1 Reformulation of boundary value problems

We present here the Data-Driven formulation of a purely mechanical problem for the general case of a system discretized in both space and time, [5, 13, 16]. The Data-Driven framework retains all formerly developed numerical schemes, e.g. finite elements, time integration techniques. Consider a body discretized in $N$ nodes and $M$ material points. The body is subjected to applied forces, assembled in the global nodal force vector, $\boldsymbol{f} = \{\boldsymbol{f}_i\}_{i=1}^{N}$, and undergoes nodal displacements, $\boldsymbol{u} = \{\boldsymbol{u}_i\}_{i=1}^{N}$ (Fig. 1a)), which are the primary unknowns of the classical finite element method. The Data-Driven analysis takes place in the global phase space $Z$, consisting of the collection of the local strain-stress pairs at all $M$ material points of the discretized system, i.e. $\boldsymbol{z} = \{(\boldsymbol{\epsilon}_e, \boldsymbol{\sigma}_e)\}_{e=1}^{M} \in Z$. The essential constraints and conservation laws pertinent to the case of mechanical problems considered herein, are compatibility and equilibrium, expressed with the following equations:

$$\boldsymbol{\epsilon}_{e,k} = \boldsymbol{B}_e \boldsymbol{u}_k, \quad \forall e = 1, \dots, M \tag{1}$$

$$\sum_{e=1}^{M} w_e \boldsymbol{B}_e^T \boldsymbol{\sigma}_{e,k} = \boldsymbol{f}_k \tag{2}$$

where $\boldsymbol{u}_k, \boldsymbol{f}_k, \boldsymbol{\epsilon}_k, \boldsymbol{\sigma}_k$ denote the displacements, forces, strains and stresses at time $t_k$ respectively, $\boldsymbol{B}_e$ is a discrete strain operator for material point $e$, and $\{w_e\}_{e=1}^{M}$ are elements of volume associated with each material point. Eqs. (1) and (2) impose constraints on the internal strain-stress states that can be deemed admissible. The set of compatible and equilibrated internal states forms the constraint set $C$, Fig. 1b), which is a subset of the global phase space $Z$, on which the internal state of the system should lie:

$$C_k = \{\boldsymbol{z} \in Z | \, (1) \text{ and } (2)\} \tag{3}$$

where the subscript $k$ in Eq. (3) implies that the constraint set could be time-dependent, for instance due to time-dependent loading.

The contribution of the Data-Driven paradigm is that it eliminates the uncertainty introduced by empirical constitutive modeling, since it abandons any material law in functional form, but instead it directly incorporates material data in the analysis. Material data can be obtained by on site measurements, from laboratory experiments, or in silico, i.e. via computational models such as molecular dynamics, lower-scale finite element method, discrete element method, Fig. 1a). We denote as $\boldsymbol{y} = \{(\boldsymbol{\epsilon}'_e, \boldsymbol{\sigma}'_e)\}_{e=1}^{M}$ a point in the global data set $D$, consisting of a finite collection of strain-stress data points describing local material behavior of the corresponding points, $D = D_1 \times \cdots \times D_M$. Each material point could in general be represented by a different data set, $D_e$, for instance due to heterogeneity of material properties. Besides, in the case of inelastic material behavior, which is the focus of

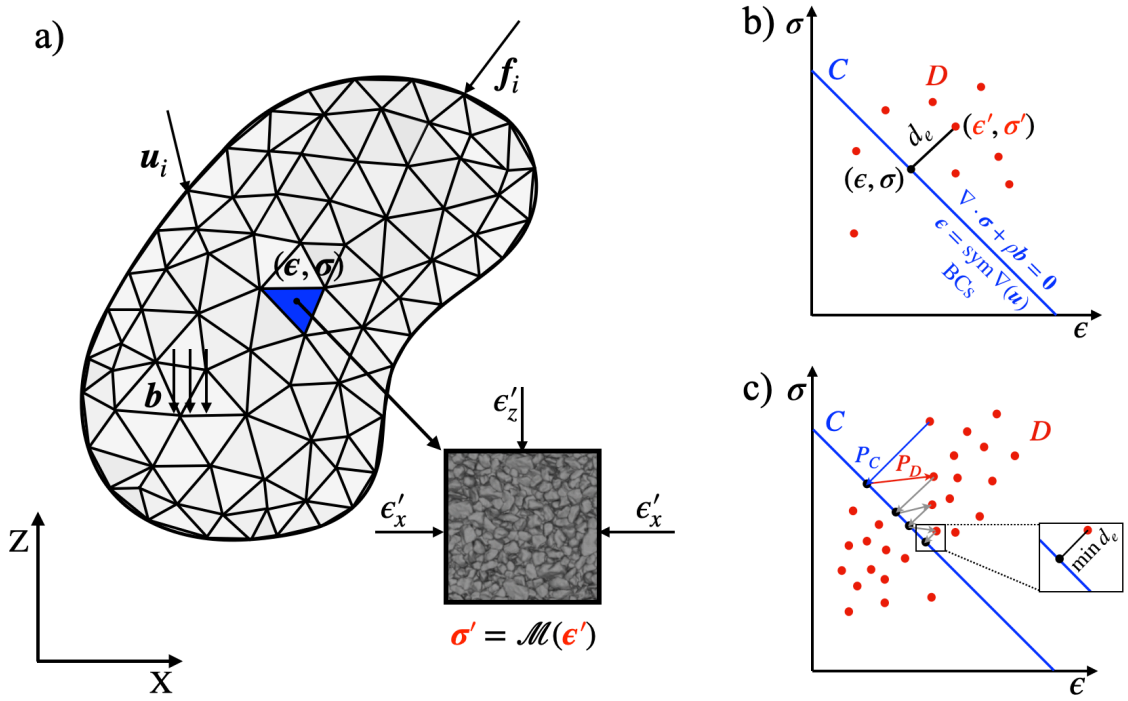


Figure 1: Overview of the Data-Driven Method: a) Finite element model of a boundary value problem and micromechanical model used to generate material data, b) Illustration of a mechanical strain-stress state in the constraint set $C$, a material strain-stress state in the data set $D$, and their distance $d_e$, c) Sequence of projection operations aiming to minimize the distance between sets $C$ and $D$, by minimizing the local distances $d_e$, $\forall\, e = 1, \ldots, M$. The black point in the inset denotes the data-driven mechanical solution, whereas the red point denotes the data-driven material solution.

this study, the data set is time-dependent; it consists of all material states that can be accessed from the current state given the past local history of deformation:

$$D_k = \left\{ \left( \boldsymbol{\epsilon}'_{e,k}, \boldsymbol{\sigma}'_{e,k} \right)_{e=1}^{M} | (\text{local history of each material point}) \right\} \tag{4}$$

The solution to the considered problem should satisfy as closely as possible the pertinent constraints, Eqs. (1) and (2), and material behavior. The latter is captured through the point data set $D_k$, created by experimental measurements. Given the finiteness of the data set, such a solution can be found provided the data-driven problem is formulated as a distance minimization problem between the constraint set $C_k$ and the data set $D_k$:

$$\min_{\boldsymbol{z} \in C_k} \min_{\boldsymbol{y} \in D_k} d^2 \left( \boldsymbol{z}_k, \boldsymbol{y}_k \right) = \min_{\boldsymbol{y} \in D_k} \min_{\boldsymbol{z} \in C_k} d^2 \left( \boldsymbol{z}_k, \boldsymbol{y}_k \right) \tag{5}$$

$$d^2 \left( \boldsymbol{z}_k, \boldsymbol{y}_k \right) = ||\boldsymbol{z}_k - \boldsymbol{y}_k||^2 \tag{6}$$

This requires a definition of the distance $d$ in phase space:

$$d^2 \left( \boldsymbol{z}_k, \boldsymbol{y}_k \right) = \sum_{e=1}^{M} w_e d_e^2 \left( \boldsymbol{z}_{e,k}, \boldsymbol{y}_{e,k} \right) \tag{7}$$

where $d_e$ is the corresponding local distance given by

$$d_e^2 \left( \boldsymbol{z}_{e,k}, \boldsymbol{y}_{e,k} \right) = \frac{1}{2} \{ \mathbb{C}_e \left( \boldsymbol{\epsilon}_{e,k} - \boldsymbol{\epsilon}'_{e,k} \right) \cdot \left( \boldsymbol{\epsilon}_{e,k} - \boldsymbol{\epsilon}'_{e,k} \right) + \mathbb{C}_e^{-1} \left( \boldsymbol{\sigma}_{e,k} - \boldsymbol{\sigma}'_{e,k} \right) \cdot \left( \boldsymbol{\sigma}_{e,k} - \boldsymbol{\sigma}'_{e,k} \right) \} \tag{8}$$

with $\boldsymbol{z}_{e,k} = (\boldsymbol{\epsilon}_{e,k}, \boldsymbol{\sigma}_{e,k})$, $\boldsymbol{y}_{e,k} = (\boldsymbol{\epsilon}'_{e,k}, \boldsymbol{\sigma}'_{e,k})$ being two given instantaneous local strain-stress states in the constraint set and the local material data set respectively. $\mathbb{C}_e$ is a symmetric positive definite matrix of purely numerical nature. It serves as a weight which makes the two terms of Eq. (8), expressing the distance deviations in strain and stress respectively, having the same units and similar magnitude. In the present analysis, the matrix $\mathbb{C}_e$ is chosen to have the same form as the elasticity matrix, since such a selection results in good convergence behavior [6]. Another option would be to solve for the optimal $\mathbb{C}_e$ as described in [12], this way avoiding any bias introduced by the choice of the metric tensor and rendering the scheme completely parameter-free.

The solution to the data-driven problem, Eq. (5), requires identifying the optimal assignment of strain-stress states from the material data set to all points of the discretized system, that results in the minimum global distance. This is a problem of combinatorial complexity, since the number of possible assignments of local material states grows exponentially with the number of material points [6,22]. To circumvent this complexity, the solution of the global distance minimization problem can be obtained heuristically through a staggered scheme, i.e. by solving a series of two minimization sub-problems. In the first sub-problem of a given iteration $j$, a data point of the global data set, $\boldsymbol{y}_k^{(j)} = \{(\boldsymbol{\epsilon}^*_{e,k}, \boldsymbol{\sigma}^*_{e,k})\} \in D_k$, is projected to the closest point in the constraint set, $\boldsymbol{z}_k^{(j+1)} = \{(\boldsymbol{\epsilon}_{e,k}, \boldsymbol{\sigma}_{e,k})\} \in C_k$. The global data point $\boldsymbol{y}_k^{(j)}$ represents a possible assignment of strain-stress states from the data set to all material points of the discretized system. The projection onto the constraint set $P_C$, defined by Eq. (9),

$$\boldsymbol{z}_k^{(j+1)} = P_C(\boldsymbol{y}_k^{(j)}) = \min_{\boldsymbol{z}_k \in C_k} \|\boldsymbol{z}_k - \boldsymbol{y}_k^{(j)}\|^2 \tag{9}$$

is performed by solving the following system of equations, obtained from employing the Lagrange multiplier method to solve the constrained distance minimization problem

$$\left(\sum_{e=1}^{M} w_e \boldsymbol{B}_e^T \mathbb{C}_e \boldsymbol{B}_e\right) \boldsymbol{u}_k = \sum_{e=1}^{M} w_e \boldsymbol{B}_e^T \mathbb{C}_e \boldsymbol{\epsilon}^*_{e,k} \tag{10}$$

$$\left(\sum_{e=1}^{M} w_e \boldsymbol{B}_e^T \mathbb{C}_e \boldsymbol{B}_e\right) \boldsymbol{\eta}_k = \boldsymbol{f}_{k+1} - \sum_{e=1}^{M} w_e \boldsymbol{B}_e^T \boldsymbol{\sigma}^*_{e,k} \tag{11}$$

where $\boldsymbol{\eta}_k$ denote the Lagrange multipliers. The closest compatible and equilibrated global strain-stress point $\boldsymbol{z}_k^{(j+1)} = \{(\boldsymbol{\epsilon}_{e,k}, \boldsymbol{\sigma}_{e,k})\}$ can then be calculated by

$$\boldsymbol{\epsilon}_{e,k} = \boldsymbol{B}_e \boldsymbol{u}_k, \qquad \forall e = 1, \ldots, M \tag{12}$$

$$\boldsymbol{\sigma}_{e,k} = \boldsymbol{\sigma}^*_{e,k} + \mathbb{C}_e \boldsymbol{B}_e \boldsymbol{\eta}_{e,k}, \qquad \forall e = 1, \ldots, M \tag{13}$$

In the second sub-problem of a single iteration, a nearest-neighbor search is performed for every material point, in order to find the material states $\boldsymbol{y}_k^{(j+1)}$ which are the closest to the corresponding states in $\boldsymbol{z}_k^{(j+1)}$, regarding the local distance (Eq. (8)). This projection onto the data set, denoted as $P_D$, is defined below:

$$\boldsymbol{y}_k^{(j+1)} = P_D(\boldsymbol{z}_k^{(j+1)}) = \min_{\boldsymbol{y}_{e,k} \in D_{e,k}} \|\boldsymbol{y}_{e,k} - \boldsymbol{z}_{e,k}^{(j+1)}\|^2 \quad \forall e = 1, \ldots, M \tag{14}$$

For every time step $t_k$, a series of iterations of these projection operations is performed as illustrated in Fig. 1c), until the assignment of the material strain-stress states to all points of the discretized body does not change any more. At the first time step of the simulation, the initial assignment of strain-stress states from the data set would be random, whereas at the next time steps, such assignment can be informed by the previously found optimal local states. One should also note that this heuristic solver does not necessarily yield the global minimizer, but the solution is usually a good approximation of it [6].

### 2.2 Universal energy-based parametrization of material history

The Data-Driven framework does not rely on any assumption of material behavior, it can thus be applied to any type of materials, e.g. elastic or inelastic [5, 16]. The present study focuses on the case of inelastic material behavior, which is characterized by irreversibility of deformation and history dependence. In terms of the data-driven constrained distance minimization problem outlined in the preceding section, the above implies that optimality of a material state is now understood not only in the sense of that state minimizing the local distance, Eq. (8), but also of that state not resulting in violation of the thermodynamics laws, which induce additional constraints, this time on the data set. The first and second laws write as follows:

$$\dot{\mathcal{D}} = \boldsymbol{\sigma} : \dot{\boldsymbol{\epsilon}} - \dot{\mathcal{A}} \geq 0 \tag{15}$$

or, when approximating the rates of Eq. (15) within a time-discrete setting, and for a given material point $e$

$$\mathcal{D}_{e,k+1} - \mathcal{D}_{e,k} = \frac{\boldsymbol{\sigma}_{e,k} + \boldsymbol{\sigma}_{e,k+1}}{2} : (\boldsymbol{\epsilon}_{e,k+1} - \boldsymbol{\epsilon}_{e,k}) - (\mathcal{A}_{e,k+1} - \mathcal{A}_{e,k}) \geq 0 \tag{16}$$

where $\mathcal{D}_{e,k}$, $\mathcal{A}_{e,k}$ denote the dissipation and free energy density at material point $e$ and time step $t_k$. The time-dependent data set can then be expressed by Eq. (17), and illustrated by Fig. 2.

$$D_{e,k+1} = \{(\boldsymbol{\epsilon}_{e,k+1}, \boldsymbol{\sigma}_{e,k+1}) \,|\, (\boldsymbol{\epsilon}_{e,k}, \boldsymbol{\sigma}_{e,k}), (16)\} \tag{17}$$

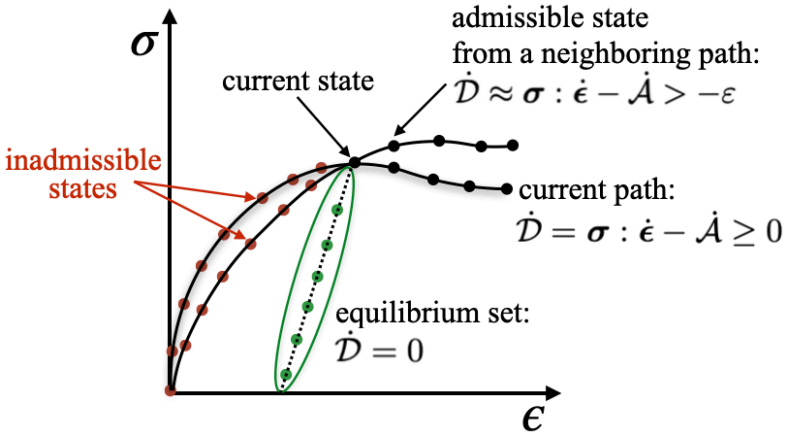


Figure 2: Energy-based parametrization of material history in the data set; The material strain-stress states which can be accessed from the current state are those that comply with the laws of thermodynamics (or within some tolerance, e.g. $\varepsilon << 1$)

This parametrization of material history can readily be obtained as long as the lower-scale computational model used for predictions of material behavior can provide the free energy and dissipation in addition to the strain-stress response (Section 4.2.2). Each strain-stress point in the phase space is then augmented by the corresponding values of free energy and dissipation. The data-driven nearest neighbor search at every time step can then take place only within the subset of thermodynamically admissible states described by Eq. (17), which is considered to contain those material states attainable from the current state given the local history of deformation. Though this universal, energy-based parametrization of material history, the data-driven solver can navigate through inelastic material data sets, and effectively differentiate between loading and unloading, as shown in [13]. It should also be pointed out that the thermodynamics-based history parametrization allows the data-driven solver to additionally consider neighboring material states from different material response paths as potential solutions to the distance minimization problem, provided these states have undergone equivalent history, as defined by an appropriate relaxation of the constraints in Eq. (16) and illustrated in Fig.2.

## 3 Adaptive Goal-oriented Data Sampling

The present study introduces an efficient way to sample the phase space of a considered mechanical problem analyzed within the data-driven framework. The underlying concept is to iteratively expand an initial data set until all material points of the discretized system are well-covered by their data set, i.e. the local distances between the mechanical and the assigned nearest-neighboring material states become sufficiently small for all material points at all time steps of the data-driven simulation. The suggested approach is characterized as adaptive, since it can iteratively adapt to meet the demands of any mechanical problem, and goal-oriented, since its results depend on the particular application.

The proposed data sampling strategy, also referred to as phase space sampling, can be implemented as described below, and illustrated with Fig. 3. A data-driven finite element simulation of the considered mechanical problem is performed with an initial material database. As the most general case, the initial database is considered to be scarce and generated in the absence of any prior information for the mechanical states of the problem. At the end of the simulation, when a converged data-driven solution is obtained, the distribution of local distances $d_e$ between the mechanical and their assigned closest strain-stress states from the material data set is analyzed. High values of the distances indicate that the mechanical states of the corresponding material points have not been well-covered by the data set. Provided all other potential sources of large distances, such as errors in the physical or virtual experiments or inaccurate material representation, are eliminated, the reason for the distance deviations can be safely attributed to insufficient representation of the corresponding mechanical states by the strain-stress states present in the current data set.

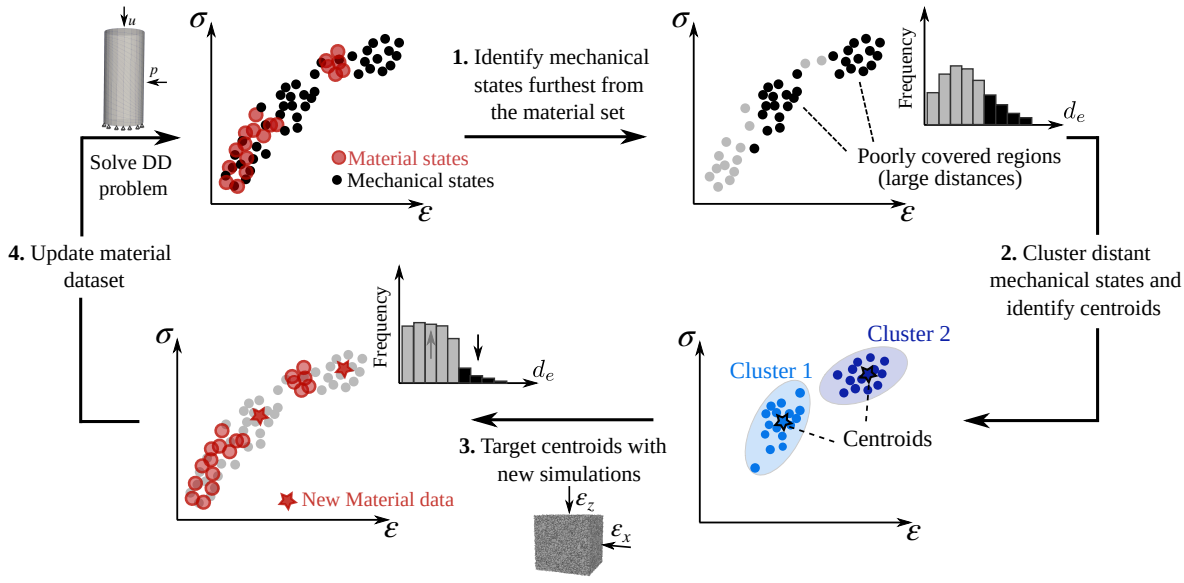


Figure 3: Illustration of a typical iteration of adaptive goal-oriented data sampling

In the next step, the mechanical states of the "problematic" material points characterized by large distance deviations, are used to inform the loading to impose on additional lower-scale simulations of representative unit cells, associated with the corresponding points. In order to avoid unnecessary lower-scale simulations, since they can be computationally expensive, a clustering analysis is first performed in order to group mechanical states which are similar. The clustering analysis allows us then to consider only the truly distinct mechanical states in the unit cell computations, this way avoiding adding redundant data in the database.

The strain-stress data points obtained from the unit cell simulations are then added to the material

data set. A new data-driven simulation is performed, which uses the expanded data set. Given that the loading applied to the unit cells was informed by the mechanical states of the previous data-driven simulation, each data set expansion will result in decreasing distances $d_e$ between the mechanical and their corresponding material strain-stress states. After every new expansion, the distribution of local distances will change, indicating the different material points in need of better-quality data. The mechanical states of these points will also change, indicating the regions of the phase space that have not yet been well-covered by the expanding data set. The process continues with more iterations of the decoupled data-driven finite element simulations of the global problem and the lower-scale unit cell simulations at the material point level, until all material points are well-covered by their data set and the data-driven prediction of mechanical behavior is accurate.

To account for history-dependent material behavior, the outlined data acquisition strategy requires only minor adjustments, which are the following; The identification of poorly covered regions of the phase space is based on the distances between the mechanical strain-stress paths traversed in the data-driven simulation, and the strain-stress paths present in the database. We define as path distance of a given material point $e$ the sum in time of the instantaneous distances between the mechanical and material states of the point, i.e. $\sum_k w_e d_e(\boldsymbol{z}_{e,k}, \boldsymbol{y}_{e,k})$, which we denote as $\sum_k w_e d_{e,k}$ for brevity. The mechanical paths which are similar, i.e. their distance $\sum_k w_e d_{e,k}$ is sufficiently small, are then grouped into clusters. The centroids of the detected clusters in this case represent mechanical strain-stress histories instead of strain-stress states. These histories are imposed as the loading paths to additional lower-scale simulations, triggering new material response strain-stress paths that will augment the database, upon an efficient parametrization of material history, e.g. as in Section 2.2.

# 4 Numerical Studies

## 4.1 Triaxial compression test of linear elastic cylinder

To evaluate the performance of the proposed data sampling strategy in aiding convergence of the Data-Driven solution, we first apply it in the case of history-independent material behavior, by simulating a triaxial compression test on a linear elastic homogeneous cylinder, under quasi-static conditions. The same experiment will subsequently be simulated on a specimen of the same geometry, but made of a heterogeneous granular material with complex history-dependent behavior. The cylindrical specimen has a diameter of 1 cm and height of 2.3 cm. The experiment starts with an initial stage of isotropic compression to 10 kPa. After the isotropic compression stage, the radial cell pressure remains constant at $\sigma_r = 10\,\text{kPa}$, while an increasing vertical displacement is applied at the top surface of the cylinder, up to an ultimate compressive axial strain of magnitude $\epsilon_a = 15\%$. The specimen is discretized with a regular mesh of linear hexahedral elements, with eight Gauss integration points. The average dimensions of the hexahedral elements are $h_x = h_y = 0.15\,\text{cm}$, and $h_z = 0.14\,\text{cm}$ (Fig. 4).

The initial material database consists of 100 strain-stress data points following linear elasticity, with values of the elastic constants $E = 100\,\text{kPa}$ and Poisson's ratio $\nu = 0.3$. The material data is plotted in Fig. 5 in the $p-q$ and $\epsilon_a-\epsilon_V$ spaces, where $p = -(1/3)\,\text{tr}\,\boldsymbol{\sigma}$ = pressure, $q = \sqrt{3/2}\,||\text{dev}\,\boldsymbol{\sigma}||$ = deviatoric stress, $\epsilon_a = -\epsilon_{zz}$ = axial strain, and $\epsilon_V = \text{tr}\,\boldsymbol{\epsilon}$ = volumetric strain. The data is plotted for the triaxial compression stage, which is why the volumetric strain starts from a non-zero value, representing the uniform contraction of the specimen at the end of the isotropic compression. As can be seen in Fig. 5, the initial database does not provide a continuous coverage of the phase space. It is the goal of the adaptive sampling process to identify the missing information in the database and iteratively expand it so as to better sample the mechanical states developed during this triaxial compression test. As described in Section 3, the augmentation of the existing database starts by analyzing the spatial distribution of the distances between the mechanical and material strain-stress states of the DD solution, in order to identify the material points in need of better coverage. We indeed chose to calculate the local distances for the total number of finite elements instead of the total number

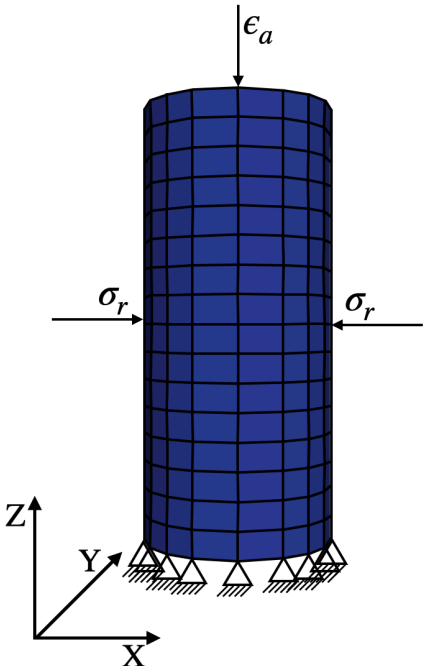


Figure 4: Finite element model of the linear elastic triaxial compression cylinder, simulated with the Data-Driven method

of material, i.e. Gauss integration points. Such a selection reduces memory requirements without compromising accuracy for the linear elements used in the simulation. The result of this analysis is shown in the histogram of iteration 1 of Fig. 7a), where one can observe the non-uniform coverage of the corresponding elements by the initial, not sufficiently dense database of Fig. 5. The mechanical strain-stress states of the elements with distances larger than a selected distance threshold are subjected to clustering using the density-based spatial clustering algorithm DB-SCAN [8], a process that will be described in detail in Section 4.2.4. The distance threshold in this study was taken as the median value of the local distances of the current DD simulation, and will decrease upon successive iterations of the data-sampling process as more material points are better covered by the augmented database. Given that in this example material behavior is known, the truly distinct strain states identified by the clustering are directly used to calculate the corresponding stress data from linear elasticity. In the case of a material with similarly history-independent behavior but with a given microstucture, the corresponding stress data would be obtained by a unit cell calculation. The resulting strain-stress data points are then added to the existing database. A new DD simulation is then performed, which uses the enriched database. The mechanically admissible strain-stress states developed in the new DD simulation will be used to inform the next database expansion, and the process continues with more such iterations until the phase space of strain-stress states is well-covered, as illustrated in Fig. 6, and convergence is reached, as shown in Fig. 7.

## 4.2 Triaxial compression test of cylinder of Hostun sand

### 4.2.1 Description of analysis

We now proceed by considering the triaxial compression test of a cylinder of Hostun sand, in order to investigate the performance of the suggested data acquisition strategy in aiding convergence and accuracy of the data-driven computations, for the case of complex history-dependent material behavior. Both physical and in silico experiments are available to validate the DD simulation. In the physical experiment [1], the cylindrical specimen contains 53,939 angular grains, and its geometry is similar

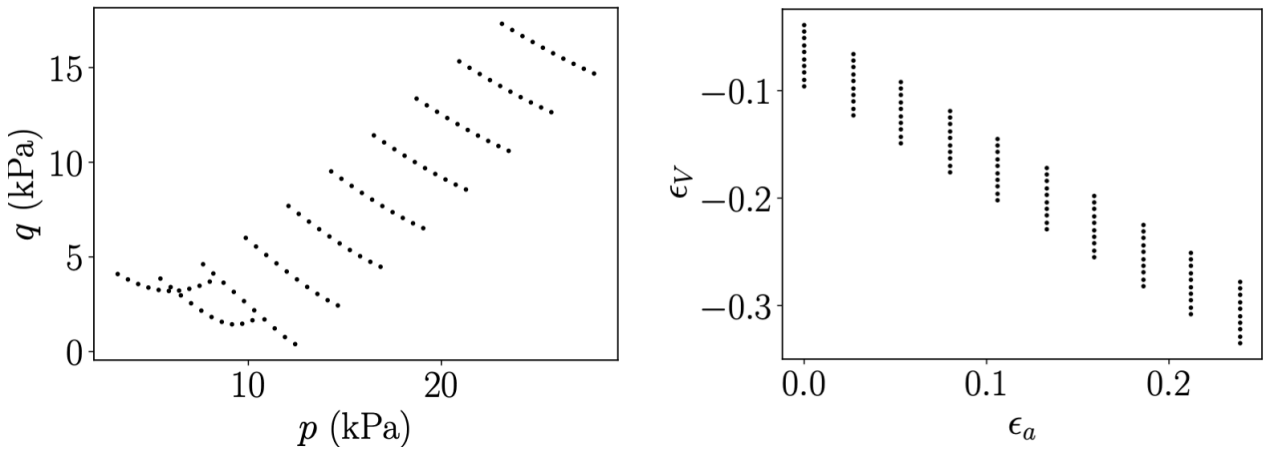


Figure 5: Representation of the strain-stress data points present in the initial database of the linear elastic material in $p - q$ and $\epsilon_a - \epsilon_V$ spaces

to that of the elastic cylinder of Fig. 4. The experiment starts with an initial stage of isotropic compression to 100 kPa. The application of the radial cell pressure $\sigma_r$ requires encasing the specimen with a flexible membrane, whereas the vertical compression is enforced through a freely rotating platen placed at the top surface of the cylinder, which applies a vertical displacement under quasi-static conditions. After the isotropic compression stage, the radial cell pressure remains constant whereas the vertical compression increases, inducing failure through shear-band formation.

A virtual experiment of this triaxial compression test was performed with LS-DEM, a variant of the discrete element method to be described in Section 4.2.2. For every grain of the actual sample, which is captured by XRCT scans of the entire cylinder [1], LS-DEM can reconstruct a virtual grain through a level-set imaging algorithm [15]. The virtual grains are equivalent to the physical ones in that they have identical shape and initial configuration. The virtual sample is subjected to exactly the same loading conditions as the physical sample, by simulating both the flexible membrane as well as the top platen. The LS-DEM simulation has been shown to be highly predictive, able to capture both the macroscopic stress-strain response, as well as the onset and spatio-temporal evolution of shear band in the specimen [15], Fig. 8a).

Given the proven predictive capability of LS-DEM, we will use this tool to perform unit cell simulations for sampling the phase space of local strain-stress states in our DD continuum finite element simulation. The finite element model of the cylinder, of same geometry as the virtual sample, is discretized with an unstructured mesh of linear hexahedral elements, with eight Gauss integration points. The orientation of the finite elements in the region where the shear band will be formed is the same as the orientation of the shear band in the experiment, Fig. 8b), corresponding to an angle of 48° with the horizontal. The meshing in the surrounding region gradually transitions towards regularity. The average dimensions of the hexahedral elements are $h_x = h_y = 0.15$ cm, and $h_z = 0.14$ cm. Starting from the isotropic compression stage at a 100 kPa, the specimen is subjected to triaxial compression up to an ultimate axial strain of approximately 15%, similar to the experiment.

### 4.2.2 Level-Set Discrete Element Method (LS-DEM) for phase space sampling

The Level-Set Discrete Element Method (LS-DEM) [14] is a variant of the Discrete Element Method (DEM) [4], which has the ability to accurately capture any morphology of the grains. Similarly to DEM, the material microstructure is explicitly modeled as a collection of particles, each representing a physical grain. Rigid body kinematics is assumed for the particles, while some small overlap is allowed

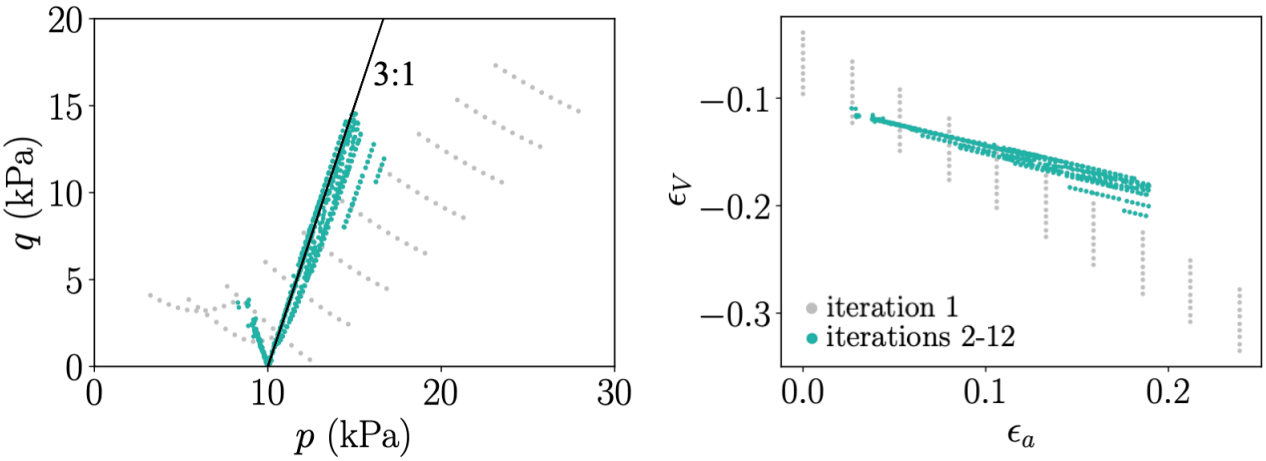


Figure 6: Representation of the strain-stress data points present in the initial (iteration 1) and the expanded with adaptive sampling (iterations 2-12) databases of the linear elastic material, in $p-q$ and $\epsilon_a - \epsilon_V$ spaces

to occur locally at the particle contacts.

The total force acting at contact $c$ of a given particle can be resolved to a normal and tangential component (Fig. 9b)), as follows:

$$\mathbf{f}^c = \mathbf{f}_n^c + \mathbf{f}_t^c \tag{18}$$

It has been shown [15] that the constitutive behavior of sand can adequately be described by a linear elastic (Hookean) contact law capped by Coulomb friction, which will thus be adopted in the present study. Neglecting any thermal effects, the contact law is described by:

$$\mathbf{f}_n^c = k_n \delta_n \mathbf{n} \tag{19}$$

$$\mathbf{f}_t^c \leftarrow \boldsymbol{R}\,\mathbf{f}_t^c - k_t \|\Delta \mathbf{s}\|, \quad \|\mathbf{f}_t^c\| \not> \mu \,\|\mathbf{f}_n^c\| \tag{20}$$

where $k_n$, $k_t$ denote the normal and tangential elastic stiffness respectively, $\delta_n$ is the particles' inter-penetration, $\mathbf{n}$ is the contact normal, $\Delta\mathbf{s}$ is the time increment of tangential contact displacement, $\boldsymbol{R}$ is a matrix that rotates the contact normal from the current to the previous time step, and $\mu$ is the friction coefficient. The total force acting at a given particle is $\mathbf{f} = \sum_{c \in C^p} \mathbf{f}^c$, where the summation is performed over all contacts $C^p$ of the particle. Once the total moment $\mathbf{m}$ acting on the particle is also calculated, time integration of Newton's equations of motion is performed to update the particle's kinematics.

The generation of granular material data sets for this triaxial compression application is performed through LS-DEM simulations of unit cells representative of material behavior of the corresponding material points of the specimen, Fig. 9a). The universal, thermodynamics-based parametrization of material history described in Section 2.2, will be used herein. LS-DEM can provide all required for the energy-based parametrization, Eq. (16), state variables, i.e. stress, strain, free energy density, and dissipation density. Assuming quasi-static conditions, the average stress tensor of the granular assembly is given by [3]:

$$\overline{\boldsymbol{\sigma}} = \frac{1}{V} \sum_c \mathbf{f}^c \otimes \mathbf{l}^c \tag{21}$$

where the summation is performed over all contacts $c$ of all particles in the assembly (unit cell), and $\mathbf{l}^c$ are the branch vectors connecting the centroids of contacting particles. The average strain $\bar{\boldsymbol{\epsilon}}$ is

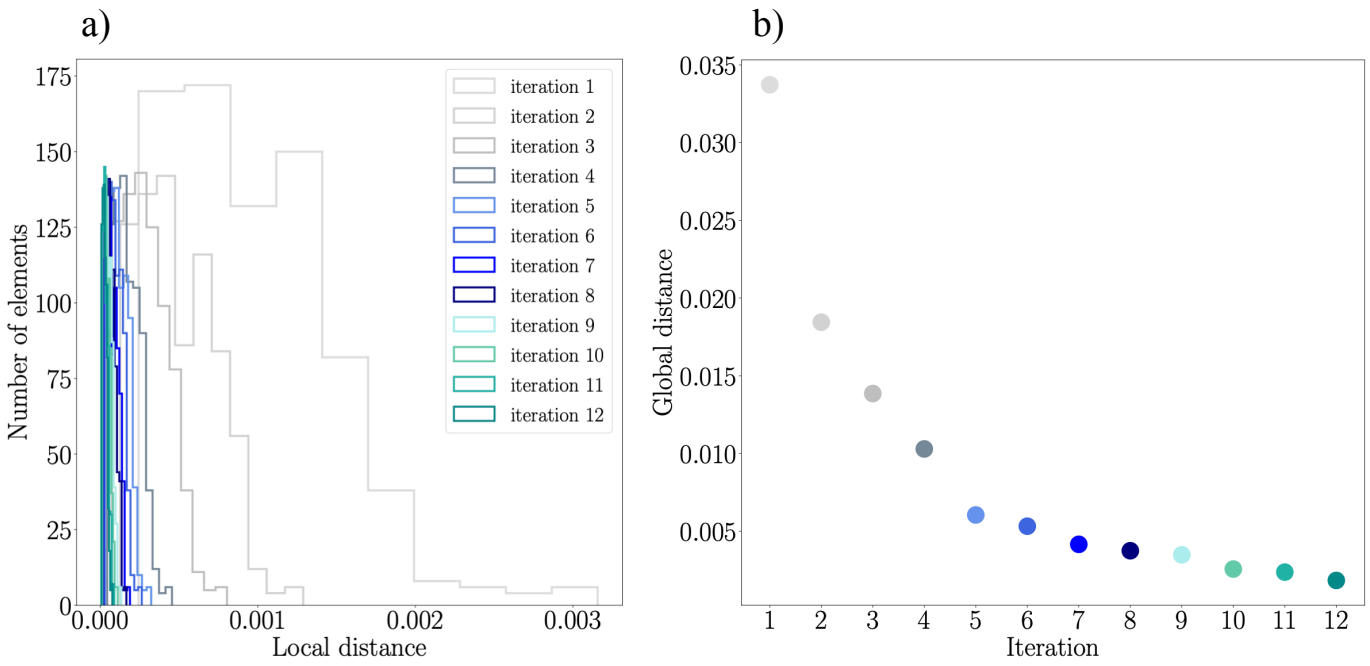


Figure 7: Performance of adaptive sampling for the Data-Driven simulations of linear elastic material behavior; a) Distribution of values of local distances between mechanical paths and material response strain-stress paths with iteration and, b) Global distance at the final time step of each DD simulation

obtained directly from the boundary deformation of the unit cell, and the free energy density due to interpenetration at the particle contacts is given by:

$$\mathcal{A} = \sum_c \mathcal{A}^c = \frac{1}{2V} \sum_c \left( \frac{\|\mathbf{f}_n^c\|^2}{k_n} + \frac{\|\mathbf{f}_t^c\|^2}{k_t} \right) \tag{22}$$

The LS-DEM unit cell simulations consider periodic boundary conditions, for which case one can calculate the increment of dissipation by the Hill-Mandel macrohomogeneity condition [11], reproduced below

$$d\mathcal{D} = \overline{\boldsymbol{\sigma}} : \overline{d\boldsymbol{\epsilon}} - d\mathcal{A} \tag{23}$$

In the present simulations dissipation arises from the frictional slip at the contacts,

$$d\mathcal{D} = \sum_c d\mathcal{D}^c = \frac{1}{V} \sum_c \mathbf{f}_t^c \cdot d\mathbf{u}^{c,\,\text{slip}} \tag{24}$$

where $d\mathbf{u}^{c,\,\text{slip}} = (\mathbf{f}_t^{c,t} - \mathbf{f}_t^{c,t+dt})/k_t$. The average stress $\bar{\boldsymbol{\sigma}}$ and strain $\bar{\boldsymbol{\epsilon}}$ quantities, augmented by the free energy and dissipation density, Eq. (22) and Eq. (23), can then be used as the input material data of the corresponding material points of the data-driven finite element model.

### 4.2.3 Creation of initial database

For illustrative purposes, we present in detail in Sections 4.2.3-4.2.5 the steps involved in a typical iteration of the adaptive data sampling, along with the specifics of its application in this case of a granular material with history-dependent behavior. The DD simulation of the first iteration of the

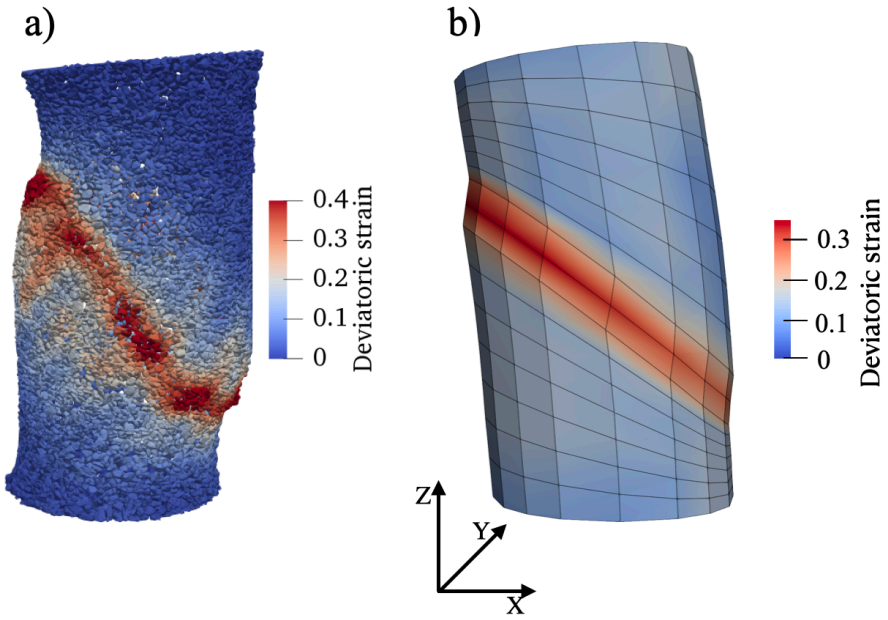


Figure 8: a) Shear band formation in the inelastic triaxial compression cylinder, as simulated with LS-DEM and, b) Finite element discretization of the cylinder and shear band formation in the red corresponding Data-Driven simulation

adaptive sampling process is based on a scarce material database. Two unit cells, extracted from the LS-DEM (virtual) sample, were used to provide material response data to form the initial database; one was extracted from a location outside the shear band, and the second one was extracted from within the region where the shear band would be formed. These two unit cells are considered to represent entire spatial regions; the region outside and the region within the shear band, respectively. However, due to material heterogeneity, unit cells from more spatial locations can be extracted in subsequent iterations of the data sampling. In the DD simulation, every material point is associated to the corresponding unit cell depending on its location and it only accesses the corresponding generated material data, a restriction which will later be lifted.

The extracted unit cells have approximately $5,000$ grains each (Fig. 9a)). The void ratios are 0.43 and 0.41 for the cell outside and the cell inside the shear band respectively. Each cell was subjected to two loading histories, obtained by homogenization of the response of the triaxial compression LS-DEM cylinder, over the corresponding spatial regions represented by the two cells. The suggested data acquisition strategy though does not require any prior information of the mechanical states of the problem. Convergence of the stress-strain response was confirmed by considering cells of different sizes, i.e. number of grains. The material response histories are used to create the initial database, which contains 260 strain-stress data points, augmented by the corresponding values of free energy and dissipation. Given the scarce initial database, the distance between the mechanical and material strain-stress paths traversed in the DD simulation is non-negligible, as shown in iteration 1 of Fig. 13a), and the DD simulation is not able to capture the experimental response, as shown in iteration 1 of Figs. 11. It is the goal of the adaptive sampling process to expand the database in the direction of decreasing distance.

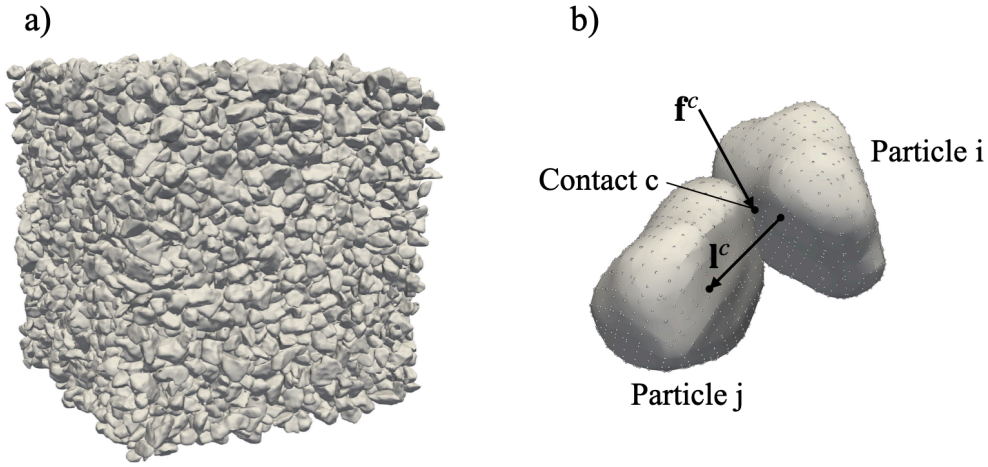


Figure 9: a) LS-DEM model of a unit cell representative of the considered granular material (Hostun sand), and b) Contact force and branch vector

#### 4.2.4 Analysis of mechanical strain-stress paths of the DD simulation

The next step in the adaptive data sampling process is to identify the material elements which have not been well-covered by the current database and analyze their mechanical paths to inform additional unit cell LS-DEM simulations. A clustering analysis of the mechanical paths of these elements is performed using DB-SCAN, [8]. Every cluster identified by the algorithm consists of mechanical paths which are sufficiently close to one another with respect to the custom distance metric $d_p$:

$$d_p^2 = \sum_k d^2 \left( \boldsymbol{z}_{e1,k}, \boldsymbol{z}_{e2,k} \right) \tag{25}$$

where index $k$ runs over all time steps for which the DD mechanical states were stored, the distance terms $d^2$ are calculated following the local distance definition of Eq. (8), and $\boldsymbol{z}_{e1,k}$, $\boldsymbol{z}_{e2,k}$ now denote the strain-stress states in the mechanical paths of any two elements $e1$ and $e2$, at a given time step $t_k$. The algorithm requires two input parameters; the radius $r$, which is the maximum distance between the path represented by the centroid of a cluster and a mechanical path at the boundaries of the cluster, and the minimum number of mechanical paths $s_{\min}$ needed to form a cluster. In this study, selecting $r = 0.1\overline{d}_p$, where $\overline{d}_p$ is the average of the path distances of the $n_p$ in number, poorly-covered elements, and $s_{\min} = \lceil n_p/10 \rceil$ where $\lceil x \rceil$ = least integer that is greater than or equal to $x$, was found to aid accuracy of the data-driven simulations, as will be shown by the results that follow (Section 4.2.6).

Fig. 10 shows two different clusters, indicated with different color, formed from the mechanical paths of the poorly-covered finite elements in the initial Data-Driven simulation. The paths are plotted in $p-q-\epsilon_V$ and $p-q-\epsilon_S$ spaces, where $p$ = pressure, $q$ = deviatoric stress, $\epsilon_V$ = volumetric strain, and $\epsilon_S = \sqrt{2/3}\,||\text{dev}\,\boldsymbol{\epsilon}||$ = deviatoric strain. One of the clusters is characterized by considerably higher volumetric and deviatoric strains; this cluster was formed by the mechanical paths of the poorly-covered elements located within the developed shear band. One can also notice that the maximum volumetric strain reached by the clustered mechanical paths is slightly larger than 0.06. Meanwhile, from Fig. 11b), it is seen that the target maximum volumetric strain, which is the one provided by the experiments, is around $\epsilon_{V,\max} = 0.065$. If the volumetric strain response of the unit cell within the shear band to the clustered mechanical path of $\epsilon_{V,\max} \approx 0.06$ closely follows it, then this will help the DD solution of the next iteration move closer to the experimental one, since it is the behavior of the elements within the shear band that governs the global response after localization.

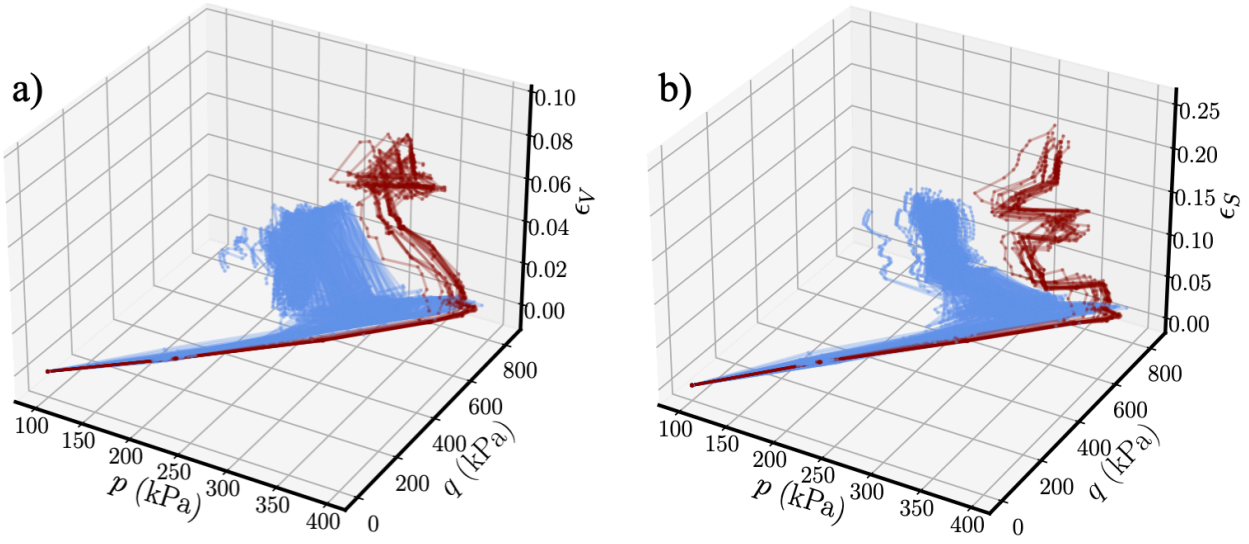


Figure 10: Clustering of mechanical paths of the poorly-covered elements, traversed in the initial DD simulation of the inelastic cylinder in; a) $p - q - \epsilon_V$ space and, b) $p - q - \epsilon_S$ space

#### 4.2.5 Database expansion with informed lower-scale unit cell simulations

For every cluster identified in the analysis of the mechanical strain-stress paths, we sample one mechanical path, selected as the path closest to the average path of the cluster, which is obtained by time-averaging of the corresponding strain and stress points from all paths within the cluster. The sampled mechanical paths are then imposed as the loading paths to the corresponding unit cells, depending on the spatial region these paths represent; mechanical paths sampled from locations within the shear band will be subjected to the unit cell extracted from the shear band, and accordingly for the paths representing the surrounding region.

The LS-DEM unit cell simulations consider mixed strain-stress control conditions, such that the dilatancy of granular material behavior is not inhibited. Generally, the choice of the control type could vary with the different cells, and with every iteration of the data sampling process. After bifurcation, the elements outside the shear band are expected to unload elastically along the triaxial compression path, whereas the elements within the shear band will be approximately under simple shear in the $x - z$ plane, Fig. 8b). Thus, it is deemed reasonable to control the $\epsilon_{zz} = \epsilon_a$ strain component, and all remaining stress components, for the unit cell extracted from outside the shear band. For the cell within the shear band, it is the $\epsilon_{xz}$ strain component that can be controlled, along with all remaining stress components. Another way of selecting the type of mixed-control conditions could be by observing the Data-Driven prediction of the macroscopic response, $\sigma_1/\sigma_3 - \epsilon_a$ and $\epsilon_V - \epsilon_a$ plots, in the previous iteration, and aiming to find a control type for which the response of the unit cells will be such that the next Data-Driven simulation, with the expanded database, approaches the experiment. One such case is analyzed in Section 4.2.6.

It is particularly important in the adaptive sampling process, to aim at achieving convergence and accuracy almost simultaneously. The goal of the adaptive sampling is to expand a given database so that convergence of the Data-Driven solution is achieved, $d^2 \to 0$ (Eq. (7)). However, whether the converged solution will be accurate depends strongly on the type of mixed-control conditions, as well as the heterogeneity of the granular sample, which results in variability of the response of the extracted from it unit cells. The risk of not focusing on achieving convergence and accuracy simultaneously can

be illustrated by considering the following; If, at a given iteration of the adaptive sampling process, the Data-Driven solution starts deviating from the experiment, then the sampled loading paths to be subjected to the unit cells will be such that they minimize the distance with the inaccurate Data-Driven solution. Thus, convergence will be approached, but accuracy will be compromised. It is also important to note that the thermodynamics-based parametrization of material history in the data sets does not result in unique selections of optimal material states from the Data-Driven solver. It is therefore crucial to scrutinize every database expansion, and aim in finding the optimal parameters of the clustering algorithm, since these affect what loading paths will be imposed to the unit cells, as well the optimal type of mixed-control conditions, which affects the material response, where optimality in both cases is understood in the sense of the newly generated data, through which the DD solver will have to navigate, aiding accuracy. Once the material data from such informed experiments is generated, then, in the next iteration, the initial assignment of local strain-stress states to all material points of the finite element model can be informed by the newly added data, so as to improve convergence. Given the sensitivity of the data-driven inelastic simulations to the initial assignment of local material states, it is indeed worthwhile making an informed assignment, whenever possible.

The adaptive data sampling process continues with a new Data-Driven simulation, which uses the expanded database. The resulting Data-Driven solution will inform a new set of experiments to further improve the quality of the current database. The series of iterations, illustrated with Fig. 3, is repeated until sufficient accuracy is achieved, as shown in Figs. 11-12.

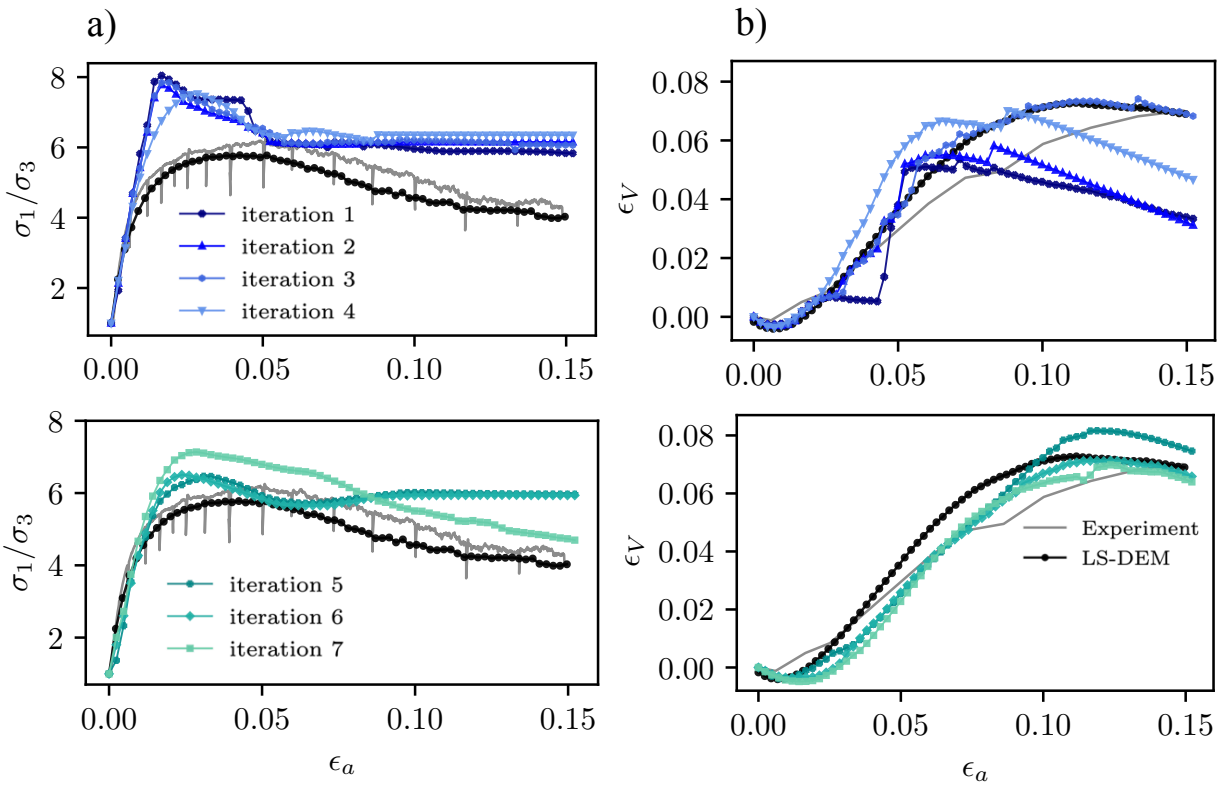


Figure 11: Data-driven mechanical solutions of the macroscopic behavior of the inelastic cylinder, upon successive iterations; a) Stress ratio vs axial strain and, b) Volumetric strain vs axial strain

#### 4.2.6 Results and discussion

In the preceding section, we illustrated in detail a single iteration of the adaptive sampling process. Here we present the results for all iterations attempted. As seen in Figs. 11-13, a total of seven

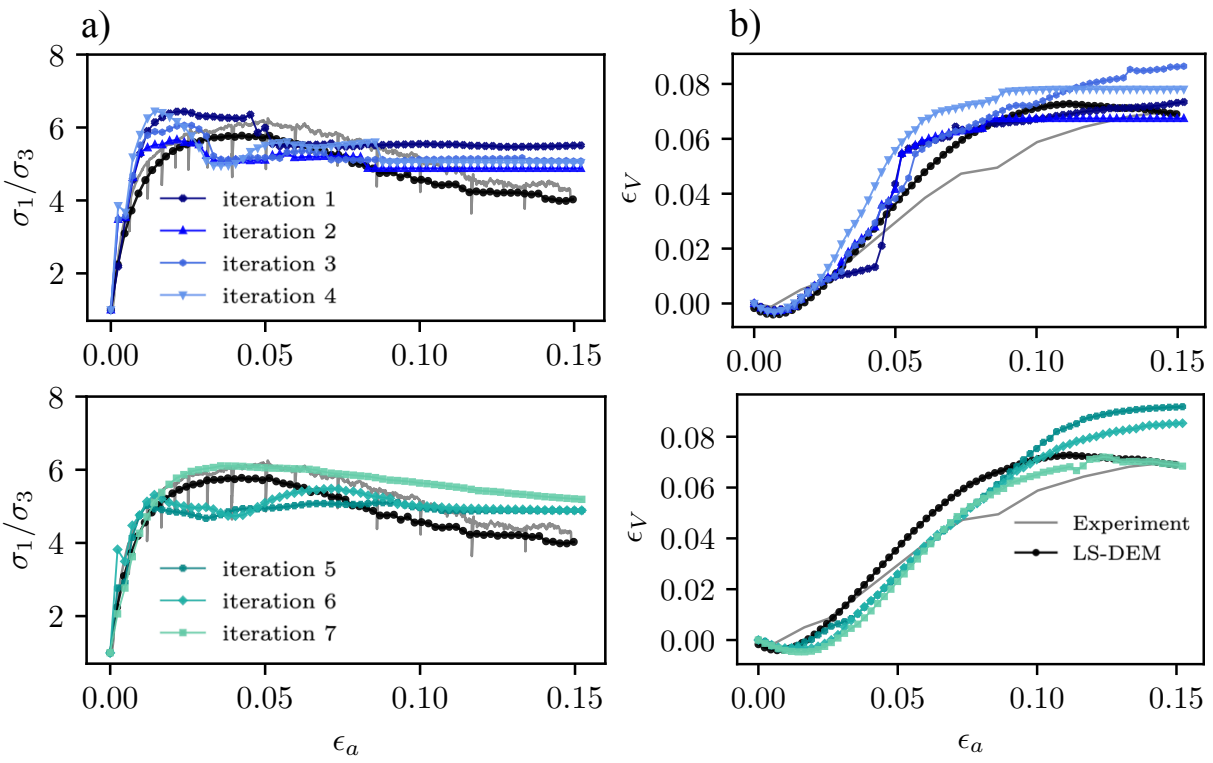


Figure 12: Data-driven material solutions of the macroscopic behavior of the inelastic cylinder, upon successive iterations; a) Stress ratio vs axial strain and, b) Volumetric strain vs axial strain

iterations was sufficient to considerably improve the accuracy and convergence of this data-driven simulation of triaxial compression test.

Figs 11 and 12 present the data-driven predictions of macroscopic behavior in terms of stress ratio versus axial strain, $\sigma_1/\sigma_3 - \epsilon_a$, and volumetric strain versus axial strain, $\epsilon_V - \epsilon_a$. The stress ratio is computed as the ratio $\frac{\sigma^{\mathrm{t}}_{zz}}{(\sigma_{xx}+\sigma_{yy})/2}$, where $\sigma^{\mathrm{t}}_{zz}$ is calculated by averaging the nodal stresses among the nodes located at the upper 15% of the specimen height, in order to be consistent with the computation of the stress ratio in the corresponding LSDEM simulation used for validation. The other stress components, $\sigma_{xx}$ and $\sigma_{yy}$, are computed by averaging the nodal stresses in the entire finite element model of the cylinder. The volumetric strain $\epsilon_V$ is computed by averaging the volumetric strains at all material points of the discretized specimen. We remind that in Data-Driven we have two solutions; the mechanical solution, which respects the physical laws and essential constraints, and the material solution, which respects material behavior. At convergence, both solutions should be sufficiently close to each other. In terms of the mechanical solution with respect to iteration, we see that, in iterations 1-3, the mechanical prediction of $\sigma_1/\sigma_3 - \epsilon_a$ behavior remains almost constant, whereas the $\epsilon_V - \epsilon_a$ prediction considerably changes, until it almost overlaps with the LS-DEM prediction at iteration 3. The change in the corresponding material solutions of iterations 1-3 is much less pronounced. Thus, the distance reduction between the mechanical and material solutions is mostly due to the change in the mechanical volumetric strain response, which drastically approaches the corresponding material solution. The change in the global distance with respect to iteration can be seen in Fig. 13b), where one can more clearly observe the drastic distance reduction from iteration 1 to iteration 3. The corresponding distribution of the distances between the mechanical and material response strain-stress paths, as shown in Fig. 13a), also changes; As more elements become better-covered by the expanding database, the distribution becomes narrower and transitions towards the lower values of path distances.

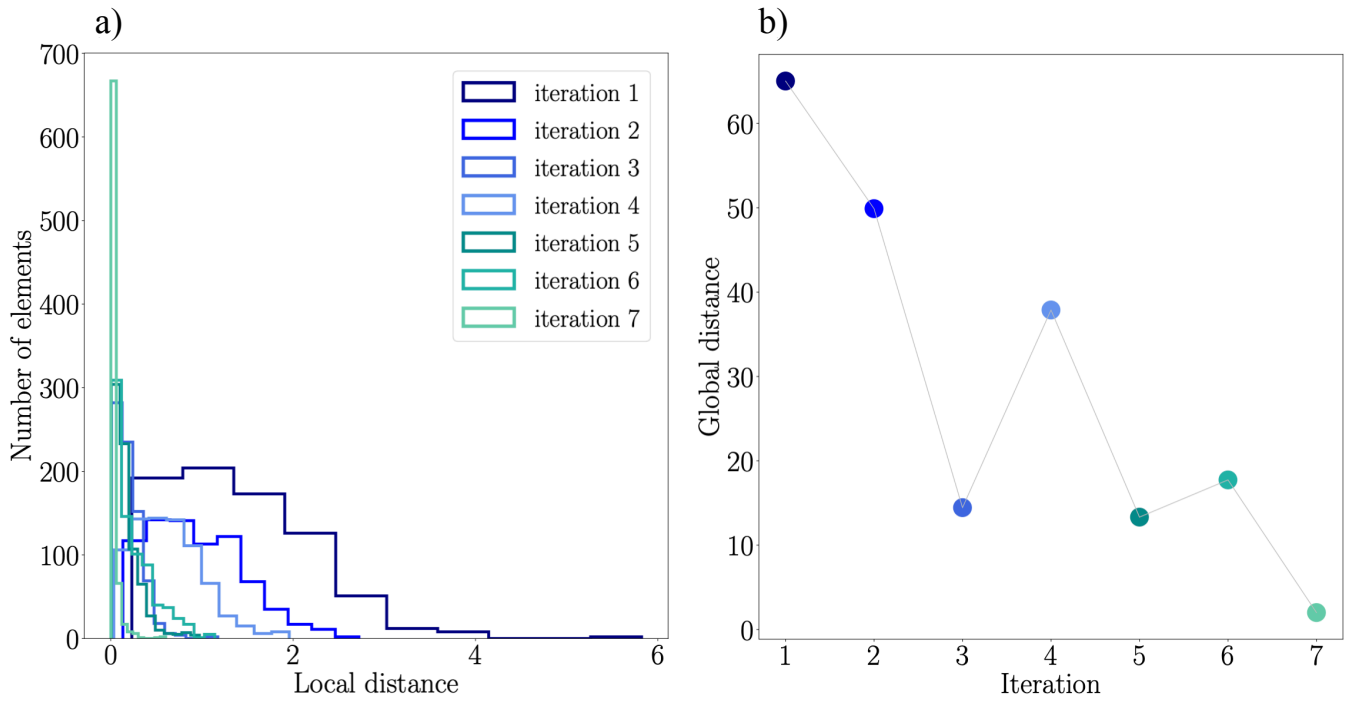


Figure 13: Performance of adaptive sampling for the Data-Driven simulations of inelastic material behavior; a) Distribution of values of local distances between mechanical and material response strain-stress paths with iteration and, b) Global distance at the final time step of each DD simulation

As already commented, it is crucial in the data sampling process to aim at achieving convergence and accuracy almost simultaneously. Even though iterations 1-3 contribute significantly to convergence and the mechanical solution of volumetric strain response is as accurate as it can be, since it coincides with the LS-DEM prediction, the stress ratio vs axial strain response has hardly changed. Thus, one should now focus on improving the accuracy of the mechanical stress prediction. To achieve that, when performing the unit cell simulations to create the database of iteration 4, we changed the control type of the unit cell outside the shear band, such that the peak developed $\sigma_{zz} = \sigma_1$ stress of the unit cell is closer to the peak stress in the experiment. This helped the global stress response start approaching the experiment, as seen in iteration 4 of Fig. 11a). However, the volumetric strain response now largely deviates from the experiment, Fig. 11b), and from the corresponding material solution. Given that the global distance is a summation of the distance deviations in strain and stress space, the global distance momentarily increased from iteration 3 to iteration 4, Fig. 13b). We accepted though, for this instantaneous increase, in our attempt to reach a solution with errors that are balanced in both spaces.

Besides, one should also keep in mind, that, in the case of inelastic material behavior, the data-driven nearest-neighbor search takes place only within the subset of thermodynamically admissible states. This can lead the nearest-neighbor search assign material states that are sub-optimal in terms of minimizing the distance to the corresponding mechanical states, if the truly optimal states violate thermodynamics. Thus, illustration of convergence with respect to database expansion may not be as clear in this application on inelastic material behavior as it has been in the elastic case.

Lastly, we have mentioned that in order to capture localization in this triaxial compression test, which arises mostly from material heterogeneity and not from the boundary conditions, we considered two distinct databases; one representing the material outside the shear band, and a second one,

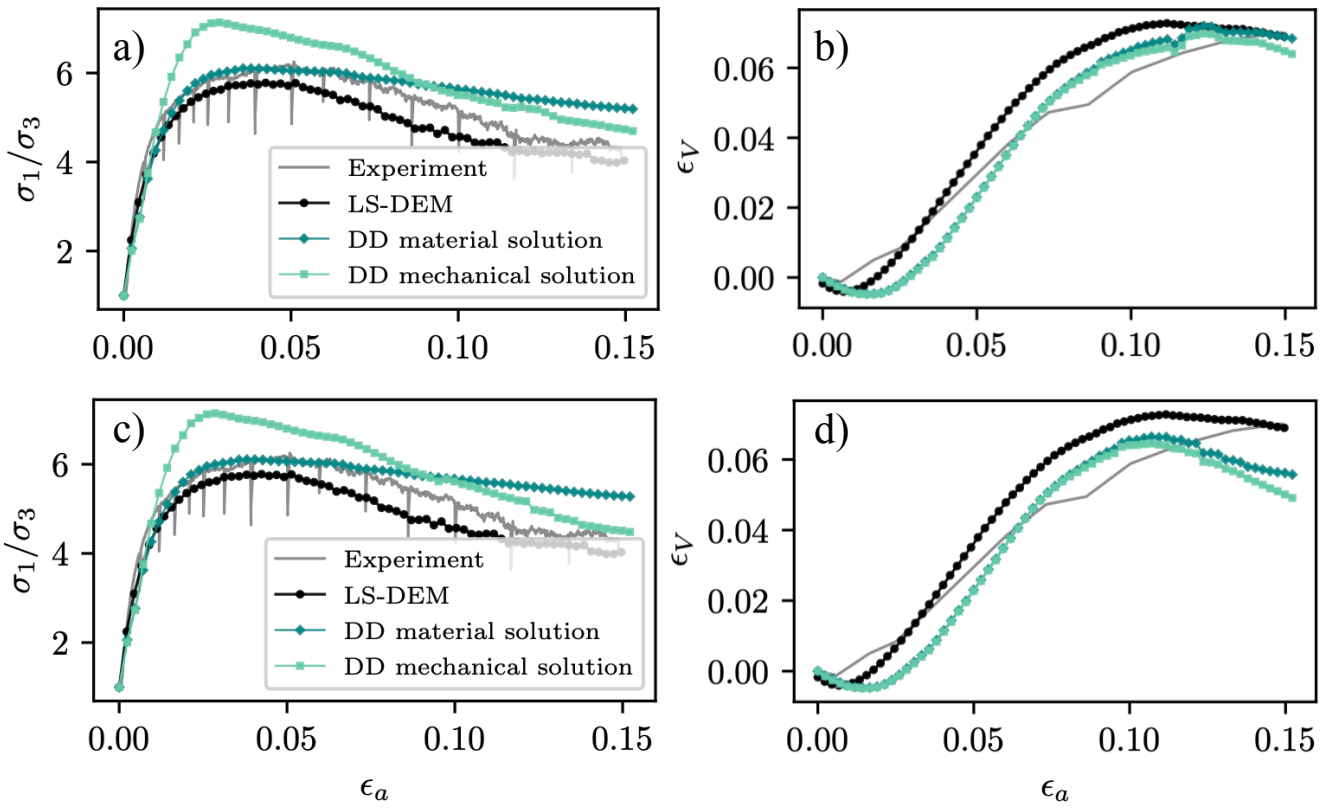


Figure 14: Data-driven simulation results of iteration 7 for the case of inelastic material behavior: a), b) by using different databases for the elements within and outside the shear band and, c), d) by allowing all elements to access data from both databases

representing the material within the shear band. In iterations 1-5, each database could only be accessed by the corresponding material points, depending on their location. In the last iterations however, we lift this restriction and allow all material points of the finite element mesh to access all available data, while informing only the initial assignment of material states so as to capture material heterogeneity. This serves only as the starting point for the data-driven iterative solver. Upon such initialization, data from both databases are available to all material points. As seen in Fig. 14, the results remain fairly insensitive, illustrating the efficiency of the thermodynamically-motivated parametrization of material history in navigating through the unified database and correctly capturing the global response. The failure mechanism of the specimen, as predicted by the last Data-Driven simulation, is also sufficiently well captured, as shown in Fig. 8b).

## 5 Conclusions

The Data-Driven distance-minimizing method [16] reformulates the solution of boundary value problems of mechanics directly in stress and strain data, this way bypassing the use of empirical constitutive models. The method was motivated by the development of highly predictive computational models, able to capture material behavior across scales, as well as advances in Data Science, the combination of which can foster the creation and efficient organization of databases of mechanical data, obtained from physical as well as high-fidelity virtual experiments. The method is also rather appealing from a computational point of view; few of its important features include the ability to standardize finite ele-

ment (FE) solvers given that they are no longer tied to constitutive models, the linearity of the system of equations defining the projection of a strain-stress data point onto the constraint set, Eqs. (10)-(13), the decoupling of the search for the optimal data points from this equilibrium projection, as well as the real-time information on the solution errors that the data-driven solver provides. Recent research efforts have implemented the data-driven solver in commercial FE software [22], and aim to increase the computational efficiency of the data-driven method so that it is comparable, and possibly higher, than that of the conventional, constitutive model-tied FE method.

Given that the predictability of data-driven computations is governed by the quality of material data, the process of generating material data sets pertinent to a considered application will remain a topic worth exploring. To the best of the authors' knowledge, this study provides the first demonstration of a data acquisition strategy, applicable to any mechanical problem and type of material behavior, that results in convergent data-driven predictions without any a priori knowledge of the strain-stress states to be encountered. The underlying concept is to exploit the reformulation of boundary value problems as distance minimization problems that the data-driven method requires. The suggested approach uses the distance deviations between the constraint set and the material data set to inform new experiments, which are now targeted to better sample the phase space of the considered problem. The process involves a set of iterations of decoupled data-driven simulations followed by lower-scale experiments at the material point level that provide the higher-quality data to expand an existing database with. The lower-scale experiments are solely informed by the compatible and equilibrated local strain-stress states provided by the data-driven solution, thus no empiricism is involved during the data population.

Application of the method in numerical simulations of triaxial compression on linear elastic material, and subsequently on a granular material with complex history-dependent behavior, clearly shows its ability to identify and complete any missing information of strain-stress states in an existing database. The results in both cases are rather promising; starting from scarce material databases, the method is able to effectively populate them within only a minimum number of iterations involved, i.e. database expansions. The latter has important implications on the computational cost of the data-driven simulations. It is known that the most time-consuming step of the data-driven solution process is the nearest-neighbor (NN) search within the material data set. The search time increases with an increasing database size, thus, the creation of databases of size that is just sufficient to cover the phase space of the considered application, as suggested by the proposed method, avoids any unnecessary overload in the NN search. One could also consider removing any poor-quality material data points upon iterations of the adaptive data sampling, to accelerate the NN search.

It should be noted that the purpose of this study is to illustrate the concept underlying our proposed data sampling method. This very first implementation of the adaptive sampling algorithm lacks features which can greatly improve its efficiency. Such features have already been implemented in other studies, and can be incorporated in the current algorithm as well. A few suggestions are: 1) Implementation of a data structure for acceleration of the NN search in the augmented database [6]. In the case of history-dependent material behavior, one could expect that the most efficient data structure could depend on the representational paradigm considered, e.g. internal-variable or energy-based history parametrization. 2) Optimization of the metric tensor to aid convergence, as described in [12]. In general, the implemented decoupled scheme implies increased simulation time compared to the alternative of on-the-fly database expansion, [13]. However, an on-the-fly sampling scheme would involve higher memory requirements for the case of history-dependent material behavior, since one would need to store the prior states of the granular unit cells, before applying incremental loading to trigger new material responses on-the-fly.

Lastly, one can comment on the greater paradigm shift that the data-driven approach has initiated; while in the conventional approach, modelers would have been mostly pleased with the experimental validation of their computations of mechanical behavior, in the data-driven era, modelers are given a new direction to look into; this is the one of relying on a material model-free computational framework

to perform better-informed experiments and fill in any crucial missing information of material behavior.

### Acknowledgements

L. Stainier and M. Ortiz gratefully acknowledge the financial support of the Deutsche Forschungsgemeinschaft (DFG) and French Agence Nationale de la Recherche (ANR) through the project "Direct Data-Driven Computational Mechanics for Anelastic Material Behaviours" (ANR-19-CE46-0012-01, RE 1057/47-1, project number 431386925) within the French-German Collaboration for Joint Projects in Natural, Life and Engineering (NLE) Sciences. L. Stainier acknowledges the support from the "NExT program of Nantes Université, through the iDDrEAM IRP project". M. Ortiz is grateful for support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via project 211504053 - SFB 1060; project 441211072 - SPP 2256; and project 390685813 - GZ 2047/1 - HCM. A. Gorgogianni acknowledges the financial support provided by the Drinkward fellowship of the California Institute of Technology.

## References

[1] E. Andò, S. A. Hall, G. C. Viggiani, J. M. Desrues, and P. Bésuelle. Grain scale experimental investigation of localised deformation in sand: a discrete particle tracking approach. *Acta Geotech.*, 7(1):1–13, 2012.

[2] P. Carrara, L. De Lorenzis, L. Stainier, and M. Ortiz. Data-driven fracture mechanics. *Comput. Methods Appl. Mech. Eng.*, 372, 12 2020.

[3] J. Christoffersen, M. M. Mehrabadi, and S. Nemat-Nasser. A Micromechanical Description of Granular Material Behavior. *J. Appl. Mech.*, 48(2):339–344, 06 1981.

[4] P. A. Cundall and O. D. L. Strack. A discrete numerical model for granular assemblies. *Géotechnique*, 29(1):47–65, 1979.

[5] R. Eggersmann, T. Kirchdoerfer, S. Reese, L. Stainier, and M. Ortiz. Model-free data-driven inelasticity. *Comput. Methods Appl. Mech. Eng.*, 350:81–99, 2019.

[6] R. Eggersmann, L. Stainier, M. Ortiz, and S. Reese. Efficient data structures for model-free data-driven computational mechanics. *Comput. Methods Appl. Mech. Eng.*, 382, 8 2021.

[7] R. Eggersmann, L. Stainier, M. Ortiz, and S. Reese. Model-free data-driven computational mechanics enhanced by tensor voting. *Comput. Methods Appl. Mech. Eng.*, 373:113499, 1 2021.

[8] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining KDD96*, pages 226–231. AAAI Press, 1996.

[9] J. Ghaboussi, J. H. Garrett, and X. Wu. Knowledge-based modeling of material behavior with neural networks. *J. Eng. Mech.*, 117(1):132–153, 1991.

[10] Y.M.A. Hashash, S. Jung, and J. Ghaboussi. Numerical implementation of a neural network based material model in finite element analysis. *Int. J. Numer. Methods Eng.*, 59:989–1005, 2 2004.

[11] R. Hill. The essential structure of constitutive laws for metal composites and polycrystals. *J. Mech. Phys. Solids*, 15(2):79–95, 1967.

[12] K. Karapiperis, M. Ortiz, and J. E. Andrade. Data-driven nonlocal mechanics: Discovering the internal length scales of materials. *Comput. Methods Appl. Mech. Eng.*, 386:114039, 2021.

[13] K. Karapiperis, L. Stainier, M. Ortiz, and J. E. Andrade. Data-driven multiscale modeling in mechanics. *J. Mech. Phys. Solids*, 147:104239, 2021.

[14] R. Kawamoto, E. Andò, G. Viggiani, and J. E. Andrade. Level set discrete element method for three-dimensional computations with triaxial case study. *J. Mech. Phys. Solids*, 91:1–13, 2016.

[15] R. Kawamoto, E. Andò, G. Viggiani, and J. E. Andrade. All you need is shape: Predicting shear banding in sand with LS-DEM. *J. Mech. Phys. Solids*, 111:375–392, 2018.

[16] T. Kirchdoerfer and M. Ortiz. Data-driven computational mechanics. *Comput. Methods Appl. Mech. Eng.*, 304:81–101, 6 2016.

[17] T. Kirchdoerfer and M. Ortiz. Data driven computing with noisy material data sets. *Comput. Methods Appl. Mech. Eng.*, 326:622–641, 11 2017.

[18] T. Kirchdoerfer and M. Ortiz. Data-driven computing in dynamics. *Int. J. Numer. Methods Eng.*, 113(11):1697–1710, 2018.

[19] A. Leygue, M. Coret, J. Réthoré, L. Stainier, and E. Verron. Data-based derivation of material response. *Comput. Methods Appl. Mech. Eng.*, 331:184–196, 2018.

[20] K. Linka, M. Hillgärtner, K.P. Abdolazizi, R. C. Aydin, M. Itskov, and C.J. Cyron. Constitutive artificial neural networks: A fast and general approach to predictive data-driven constitutive modeling by deep learning. *J. Comput. Phys.*, 429, 3 2021.

[21] M. Mozaffar, R. Bostanabad, W. Chen, K. Ehmann, J. Cao, , and M. A. Bessa. Deep learning predicts path-dependent plasticity. *Proc. Natl. Acad. Sci.*, 116, 2019.

[22] E. Prume, L. Stainier, M. Ortiz, and S. Reese. A data-driven solver scheme for inelastic problems. In *Proceedings in Applied Mathematics and Mechanics*, 2022.