xLSTM for constitutive modelling of sand: from data-driven extrapolation to thermodynamic consistency

Toiba Noor [1*], *Gunturi V.* Ramana [1], and *Rajdip* Nayek [2]

[1] Department of Civil and Environmental Engineering, Indian Institute of Technology, Delhi, New Delhi
[2] Department of Applied Mechanics, Indian Institute of Technology, Delhi, New Delhi

Abstract. Constitutive modeling of granular soils is inherently a path-dependent problem in which the stress response at any loading step depends on the entire history of deformation and on the initial state of the specimen. Long Short-Term Memory (LSTM) networks have become the dominant recurrent architecture for data-driven constitutive models of soil, yet their scalar gating mechanism is architecturally mismatched to the covariance-rich loading histories observed in triaxial stress paths. This paper proposes the use of xLSTM [1] (Extended Long Short-Term Memory) for constitutive modeling of sand, and demonstrates its superiority over a standard LSTM baseline on a challenging extrapolation benchmark of drained triaxial compression tests on Karlsruhe sands. The mLSTM block at the core of xLSTM uses a matrix-valued memory cell updated by a covariance rule, enabling it to represent second-order interactions between loading-history features that scalar-gated LSTMs fundamentally cannot. Further a ThermoXLSTM, a thermodynamics-informed extension is introduced, in which effective stresses are derived from a learnable Helmholtz free-energy potential via automatic differentiation, enforcing the first and second laws of thermodynamics by construction. Model testing is done for the smallest and highest initial confining pressures (50 and 400 kPa) for different relative-density groups – conditions absent from training – creating a genuine extrapolation scenario. On this benchmark, xLSTM achieves R^2>0.92 for mean effective stress p', deviatoric stress q, and volumetric strain ε_v, improving NMSE by 34% and 18% over LSTM for p' and q respectively. ThermoXLSTM yields slightly lower accuracy metrics but guarantees thermodynamically admissible stress predictions regardless of the input conditions. Together, these models establish xLSTM as the preferred recurrent backbone for path-dependent geomechanical constitutive modeling and demonstrate that physics constraints can be embedded without sacrificing architectural advantage.

1 Introduction

For cohesionless granular soils such as sand, the stress-strain response is *nonlinear*, *pressure-dependent*, and critically *path-dependent*. Classical constitutive models encode this response through yield surfaces, hardening laws and flow rules [2, 3]. While transparent, these models require domain expertise and extensive calibration of multiple parameters and can still deteriorate when applied to stress levels or density ranges outside of those used for calibration. Machine learning models offer a complementary path, wherein the constitutive laws can be learned directly from experimental data without prescribing to an explicit functional form. Due to the path dependent behavior of sands, recurrent sequence models like Long Short-Term Memory (LSTM) networks have been extensively used for data-driven constitutive modeling [4, 5].

Despite widespread adoption, in a standard LSTM [6], the cell state is updated via element-wise multiplication with gate *vectors*, meaning each dimension of the memory is atténuâtes independently. Sand behaviour under shearing is governed by the *joint* evolution of axial strain, the stress ratio $\eta = q/p'$, and the volumetric strain E_v, quantities whose co-evolution encodes covariance like interactions across timesteps. Element-wise gating cannot explicitly store such cross-feature associative structure: the cell state is a vector, and no direct outer-product or matrix-valued memory update is possible within the standard LSTM formulation. This stands in contrast to the matrix memory cell (mLSTM) of xLSTM [1], which employs an explicit outer-product update to the cell state, enabling the model to store and retrieve associative, covariance-like structure across the loading history. This is structurally analogous to the second-order coupling in stress-strain constitutive relations, suggesting that xLSTM is architecturally better matched to constitutive modeling of path-dependent geomaterials than standard LSTM.

Despite improvement in prediction accuracy, these pure data-driven constitutive models can produce thermodynamically inadmissible predictions, violating the Clausius-Duhem inequality or producing negative dissipation [7]. This is particularly dangerous under extrapolation, where the model operates outside its training distribution and physical constraints can no longer be implicitly enforced through data coverage. The thermodynamics constrained nueral networks derive stress as the automatic-differentiation gradient of

* Corresponding author: cez217542@iitd.ac.in

a learnable free-energy potential, guaranteeing admissibility by construction [8].

This paper makes three contributions. (i) We present the first application of xLSTM to data-driven constitutive modeling of sand, with an architectural justification rooted in the covariance structure of loading history. (ii) Genuine extrapolation benchmark with test specimens at the extremes of the initial confining pressures at different relative density that are absent from training with xLSTM achieving $R^2 > 0.92$ on p', q, and E_v,consistently outperforming LSTM. (iii) We propose ThermoXLSTM, which embeds a learnable Helmholtz free-energy potential and non-negative dissipation into the xLSTM backbone, enforcing thermodynamic admissibility at a modest accuracy cost.

2 Methodology

2.1 Experimental data and problem formulation

The dataset comprises 25 drained triaxial compression tests on Karlsruhe sand [9]. Tests are distributed over five initial confining pressures $p_0 \in \{50, 100, 200, 300, 400\}$ kPa and five relative-density groups covering the range $I_D = 0.15$–0.95, with corresponding initial void ratios $e_0 = 0.697$–0.996. The full list of initial conditions is given in Table 1. Each test sequence records, at successive axial strain increments: axial strain ε_a, radial strain ε_r, volumetric strain ε_v, mean effective stress $p' = (\sigma'_a + 2\sigma'_r)/3$, and deviatoric stress $q = \sigma'_a - \sigma'_r$. Sequence lengths range from approximately 180 to 260 timesteps.

The constitutive modeling task is posed as a sequence-to-sequence prediction problem. At each timestep t, model receives a fixed 8-dimensional input vector:

$$\boldsymbol{x_t} = \begin{bmatrix} e_0^*,\ I_D^*,\ p_0^*,\ \varepsilon_a(t),\ \varepsilon_r(t), p'(t-1), \\ q(t-1), \varepsilon_v(t-1) \end{bmatrix}^{\mathrm{T}} \quad (1)$$

where superscript $*$ denotes initial conditions (initial void ratio, initial relative density, and initial confining pressure respectively) broadcast as constants over the sequence length, and the final three entries are the autoregressive recurrent feedback of the previous stress-strain state. The model predicts:

$$\hat{\mathbf{y}}_t = [p'(t),\ q(t),\ \varepsilon_v(t)]^{\mathrm{T}} \quad (2)$$

At inference, the true values in columns 6-8 of Equation (1) are replaced by the model's own predictions at $t-1$.

Test indices (10 experiments) correspond to the smallest and largest initial confining pressure. The 15 remaining specimens form the training set. This is a deliberate extrapolation split: models train exclusively on intermediate- initial confining pressure specimens (p_0 = 100 - 300 kPa) and must generalise to the extremes (p_0 = 50 and 400 Kpa) that were never seen during training.

Table 1. Initial conditions for all 25 Karlsruhe sand test specimens. Test specimens (extrapolation evaluation) are marked with *.

ID	p_0 (kPa)	e_0	I_D	Set
0*	50	0.996	0.15	Test
1	100	0.975	0.21	Train
2	200	0.975	0.21	Train
3	300	0.970	0.22	Train
4*	400	0.960	0.25	Test
5*	50	0.880	0.46	Test
6	100	0.862	0.51	Train
7	200	0.859	0.52	Train
8	300	0.848	0.55	Train
9*	400	0.847	0.55	Test
10*	50	0.840	0.57	Test
11	100	0.819	0.63	Train
12	200	0.824	0.63	Train
13	300	0.822	0.64	Train
14*	400	0.814	0.68	Test
15*	50	0.743	0.82	Test
16	100	0.758	0.79	Train
17	200	0.748	0.81	Train
18	300	0.734	0.85	Train
19*	400	0.753	0.80	Test
20*	50	0.734	0.85	Test
21	100	0.735	0.85	Train
22	200	0.706	0.92	Train
23	300	0.697	0.95	Train
24*	400	0.718	0.89	Test

2.2 Data preprocessing and scaling

The following normalisation procedure is applied consistently across all three models.

2.2.1 Strain normalisation

All strains are converted from percentage to fraction and divided by a reference strain $\varepsilon_{\mathrm{ref}} = 0.10$, mapping the range [0,10%] to [0,1].

2.2.2 Group-based RMS normalisation for stresses

Stress magnitudes vary by nearly an order of magnitude across confining pressures (e.g. peak $q \approx 120$ kPa at $p_0 = 50$ kPa versus $q \approx 900$ kPa at $p_0 = 400$ kPa). To prevent high-pressure sequences from dominating the loss, p' and q are normalised within each p_0 group using root-mean-square (RMS) scaling:

$$\tilde{p}'(t) = \frac{p'(t)}{\mathrm{rms}_{p_0}(p')}, \qquad \tilde{q}(t) = \frac{q(t)}{\mathrm{rms}_{p_0}(q)} \quad (3)$$

This places scaled targets in $\mathcal{O}(1)$ regardless of confining pressure, yielding balanced gradient magnitudes across all pressure groups.

2.2.3 Volumetric strain

Since ε_v can be negative for dilative dense sand, global signed min-max normalisation is applied:

$$\widetilde{\varepsilon_v} = \frac{(\varepsilon_v - \varepsilon_v^{\min})}{(\varepsilon_v^{\max} - \varepsilon_v^{\min})} \tag{4}$$

2.2.4 Initial conditions

e_0, I_D, and p_0 are each scaled to [0,1] using global min-max values.

2.3 LSTM baseline

The LSTM baseline [6] uses a single-layer LSTM with hidden size $h = 128$, an ELU activation, and a linear output layer:

$$\begin{aligned} \boldsymbol{h}_t,\, c_t &= \mathrm{LSTM}(\boldsymbol{x}_t,\, \boldsymbol{h}_{t-1},\, c_{t-1}), \\ \hat{\boldsymbol{y}}_t &= \mathbf{W}_{\text{out}}\, \mathrm{ELU}(\boldsymbol{h}_t) \end{aligned} \tag{5}$$

where $\boldsymbol{h}_t \in \mathbb{R}^{128}$ and $c_t \in \mathbb{R}^{128}$ are the hidden and cell states, $\mathbf{W}_{\text{out}} \in \mathbb{R}^{3\times 128}$ is the learned output projection matrix, and 3 corresponds to the number of model outputs. Training slides a window of width $W = 25$ over each sequence; within each window, the recurrent feedback columns of Equation (1) are replaced by the model's own predictions, making training fully autoregressive inside the window. Training uses the Adam optimiser at learning rate 10^{-3} for 2000 epochs without a scheduler. The LSTM is implemented in TensorFlow/Keras.

2.4 xLSTM

The xLSTM model [1] replaces the LSTM backbone with a stack of residually connected xLSTM blocks. A linear input projection maps the 8-dimensional input to embedding dimension $d = 356$, which is processed by $L = 5$ xLSTM blocks, followed by a linear output head:

$$\begin{aligned} \boldsymbol{z_t} &= \mathbf{W}_{\text{in}}\boldsymbol{x_t}, \\ \boldsymbol{h}_{1:T} &= \mathrm{xLSTMBlockStack}(\boldsymbol{z_{1:T}}), \\ \hat{\boldsymbol{y}}_{\boldsymbol{t}} &= \mathbf{W}_{\text{out}}\boldsymbol{h_t} \end{aligned} \tag{6}$$

Each xLSTM block applies pre-layer normalisation, a residual skip connection, and an mLSTM or sLSTM sublayer.

2.4.1 mLSTM (The covariance memory update)

The decisive architectural difference from standard LSTM lies in the mLSTM memory cell. In a standard LSTM, the cell-state update is:

$$\boldsymbol{C}_t = \boldsymbol{f}_t \odot \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \odot \widetilde{\boldsymbol{C}}_t, \quad \boldsymbol{f}_t, \boldsymbol{i}_t \in \mathbb{R}^h \tag{7}$$

where the element-wise multiplication by gate vectors $\boldsymbol{f}_t$ and $\boldsymbol{i}_t$ operates independently per feature dimension, limiting the cell state to a vector $\boldsymbol{C}_t \in \mathbb{R}^h$ that cannot directly store cross-feature correlations.

In the mLSTM block, the cell state is instead a *matrix* $\boldsymbol{C}_t \in \mathbb{R}^{d_v \times d_k}$, updated via an outer-product (covariance) rule:

$$\boldsymbol{C}_t = \tilde{f}_t\, \boldsymbol{C}_{t-1} + \tilde{\imath}_t\, \boldsymbol{v}_t \boldsymbol{k}_t^{\mathrm{T}} \tag{8}$$

where $\boldsymbol{v}_t \in \mathbb{R}^{d_v}$ and $\boldsymbol{k}_t \in \mathbb{R}^{d_k}$ are value and key vectors computed from the input, and $\tilde{f}_t, \tilde{\imath}_t \in \mathbb{R}$ are scalar gates computed via *exponential gating*:

$$\begin{aligned} \tilde{f}_t &= \exp(f_t), \\ \tilde{\imath}_t &= \exp(i_t - m_t), \\ m_t &= \max(f_t + m_{t-1},\ i_t) \end{aligned} \tag{9}$$

with the running maximum m_t providing numerical stabilisation [1]. Retrieval uses a query vector $\mathbf{q}_t \in \mathbb{R}^{d_k}$:

$$\begin{aligned} \widetilde{\boldsymbol{h}}_t &= \frac{\boldsymbol{C}_t\, \boldsymbol{q}_t}{\max\left(\left|\boldsymbol{n}_t^{\mathrm{T}} \boldsymbol{q}_t\right|,\ 1\right)}, \\ \boldsymbol{n}_t &= \tilde{f}_t\, \boldsymbol{n}_{t-1} + \tilde{\imath}_t\, \boldsymbol{k}_t \end{aligned} \tag{10}$$

where $\boldsymbol{n}_t \in \mathbb{R}^{d_k}$ tracks the cumulative sum of past input activations and normalises the retrieved vector $\widetilde{\boldsymbol{h}}_t \in \mathbb{R}^{d_v}$.

The outer product $\boldsymbol{v}_t \boldsymbol{k}_t^{\mathrm{T}}$ in Equation (8) is a rank-one covariance update: it writes a direct pairwise association between the key and value vectors at each timestep into the matrix memory. For constitutive modeling of sand, this is architecturally significant. The dilatancy rate $\dot{\varepsilon_v}/\dot{\varepsilon_a}$, for example, depends on both the current stress ratio $\eta = q/p'$ *and* its rate of evolution, a second-order interaction between the stress-ratio history and the strain-rate history. The coupling between the volumetric strain at an earlier step and the mean effective stress at a later step is another such cross-feature interaction. These are exactly the associations that a matrix memory can represent and a scalar gate cannot. The query-key retrieval in Equation (10) then allows the model to attend selectively to which past joint association is most relevant to the current prediction step.

The xLSTM stack also includes sLSTM blocks, which retain scalar memory but add *memory mixing* [1]: the ability to revise stored associations when new conflicting evidence arrives. This is relevant for capturing the transition from contractive to dilative behavior as a dense specimen progresses through its peak deviatoric stress where prior stored associations about the dilatancy direction must be updated.

2.4.2 Training

Full-sequence training with a mixed teacher-forcing and scheduled-sampling strategy [10] is used. The teacher-forcing ratio α decays from 1.0 to 0.1 over the final 75% of training:

$$\alpha(k) = \begin{cases} 1.0\,, & k < 0.25\,K \\ \max\left(0.1, 1 - 0.9\dfrac{0.25k}{0.75k}\right), & k \geq 0.25K \end{cases} \quad (11)$$

where k is the current epoch and $K = 620$ the total. The total training loss is:

$$\mathcal{L}_{\text{xLSTM}} = \mathcal{L}_{\text{TF}} + (1-\alpha)\,\mathcal{L}_{\text{SS}} \quad (12)$$

where $\mathcal{L}_{\text{TF}}$ and $\mathcal{L}_{\text{SS}}$ are the teacher-forcing and scheduled-sampling weighted MSE losses respectively. A linear time-weighting (decaying from 10 to 1 over the sequence) emphasises the early-loading regime where the elastic-to-plastic transition is most demanding to capture.

2.5 ThermoXLSTM

ThermoXLSTM embeds the thermodynamic constraint into the xLSTM backbone [8]. Rather than predicting stresses directly, the model learns a Helmholtz free-energy decomposition from which thermodynamically admissible stresses are derived by automatic differentiation.

Because stress is derived from strain via energy, the input recurrent feedback columns are replaced by the principal effective stresses at the previous step:

$$\begin{aligned} \boldsymbol{x}_t^{\text{thermo}} = [&e_0^*,\ I_D^*,\ p_0^*,\ \varepsilon_a(t),\ \varepsilon_r(t),\\ &\sigma'_a(t-1),\ \sigma'_r(t-1),\ \varepsilon_v(t-1)]^{\mathrm{T}} \end{aligned} \quad (13)$$

Outputs are the principal effective stresses $[\sigma'_a, \sigma'_r]$ (from which p' and q can be recovered analytically) and ε_v. An xLSTM block stack (3 layers, $d = 356$) processes the input sequence to produce hidden representations $\boldsymbol{h}_{1:T}$.

A plastic-strain head decomposes total strains into elastic and plastic parts:

$$\begin{aligned} \boldsymbol{\varepsilon}_t^p &= \tanh\left(\mathbf{W}_p \boldsymbol{h}_t\right) \in [-1,1]^2 \\ \boldsymbol{\varepsilon}_t^e &= \boldsymbol{\varepsilon}_t^{\text{total}} - \boldsymbol{\varepsilon}_t^p \end{aligned} \quad (14)$$

The tanh activation bounds plastic strains to a physically reasonable range.

Two sub-networks, one for elastic stored energy, one for plastic stored energy compute the Helmholtz free energy:

$$\begin{aligned} \Psi_t^e &= \mathcal{N}_e\left(\boldsymbol{\varepsilon}_t^{\text{total}},\ \boldsymbol{\varepsilon}_t^e\right) \in \mathbb{R}^+, \\ \Psi_t^p &= \mathcal{N}_p\left(\boldsymbol{\varepsilon}_t^{\text{total}},\ \boldsymbol{\varepsilon}_t^p\right), \\ \Psi_t &= \Psi_t^e + \Psi_t^p \end{aligned} \quad (15)$$

where $\mathcal{N}_e$ uses Softplus output to enforce $\Psi^e \geq 0$ (non-negative stored elastic energy), and $\mathcal{N}_e, \mathcal{N}_p$ are multi-layer perceptrons with inputs $[\boldsymbol{\varepsilon}^{\text{total}}, \boldsymbol{\varepsilon}^e]$ or $[\boldsymbol{\varepsilon}^{\text{total}}, \boldsymbol{\varepsilon}^p]$ respectively.

Effective stresses are computed as the derivative of the free energy with respect to total strain via PyTorch automatic differentiation:

$$\boldsymbol{\sigma'}_t = \frac{\partial \Psi_t}{\partial \boldsymbol{\varepsilon}_t^{\text{total}}} \quad (16)$$

Equation (16) is an exact analytic relation—not a network output—and satisfies the first law of thermodynamics at every timestep by construction, including for inputs outside the training distribution.

A dissipation network with Softplus output enforces the Clausius-Duhem inequality:

$$\begin{aligned} \mathcal{D}_t &= \mathcal{N}_D\left(\boldsymbol{\varepsilon}_t^{\text{total}},\ \boldsymbol{\varepsilon}_t^p,\ \dot{\boldsymbol{\varepsilon}}_t^p\right) \geq 0, \\ \dot{\boldsymbol{\varepsilon}}_t^p &\approx \boldsymbol{\varepsilon}_t^p - \boldsymbol{\varepsilon}_{t-1}^p \end{aligned} \quad (17)$$

ε_v is not directly constrained by the above stress framework for an open system and is predicted by a separate linear head: $\widehat{\varepsilon_{v_t}} = \boldsymbol{w}_v^T \boldsymbol{h}_t$.

2.5.1 Training loss

$$\begin{aligned} \mathcal{L}_{\text{thermo}} = \lambda_d \underbrace{\frac{1}{|\mathcal{M}|}\sum_{t\in\mathcal{M}} \| \hat{\boldsymbol{y}}_t - \boldsymbol{y}_t \|_{\boldsymbol{w}}^2}_{\text{weighted MSE}} \\ + \lambda_p \underbrace{\left(\mathbb{E}[\text{ReLU}(-\Psi_t)]\right.}_{\Psi \geq 0} \\ + \underbrace{\left.\mathbb{E}[\text{ReLU}(-\mathcal{D}_t)]\right)}_{\mathcal{D} \geq 0} \end{aligned} \quad (18)$$

with $\lambda_d = 1.0$, $\lambda_p = 0.1$, and per-output weights $\boldsymbol{w} = [1.0,\ 1.5,\ 2.0]$. The ReLU terms incur no cost when constraints are already satisfied, activating only to penalise violations. Note that these are *soft* constraints: they encourage admissibility strongly during training but do not provide a mathematical hard guarantee in the strict sense. Training uses Adam at learning rate $= 2 \times 10^{-4}$, OneCycleLR, gradient clipping at 1.0, 420 epochs.

A summary of hyperparameters and training settings for all three models is given in Table 2.

3 Results

3.1 Quantitative comparison on extrapolation test set

Table 3 reports the mean prediction metrics on outputs averaged over all ten test sequences. We report mean squared error (MSE), root MSE (RMSE), mean absolute error (MAE), normalised MSE (NMSE), and the coefficient of determination R^2.

Table 2. Model hyperparameters and training configuration.

Parameter	LSTM	xLSTM	ThermoX.
Framework	TF/Keras	PyTorch	PyTorch
d_{model} / h	128	356	356
Layers	1	5	3
Params (approx.)	70K	635K	480K
Optimiser	Adam	Adam	Adam
LR	10^{-3}	2×10^{-4}	2×10^{-4}
LR schedule	—	OneCycleLR	OneCycleLR
Epochs	2000	620	420
Batch size	Full	8	8
Output weights	$[1, 1.5, 2]$	$[1, 1, 1]$	$[1, 1.5, 2]$

Mean effective stress $\tilde{p}'$: xLSTM achieves $R^2 = 0.928$, improving over LSTM ($R^2 = 0.890$) by $\Delta R^2 = +0.038$, corresponding to a 34% reduction in NMSE (0.0725 vs 0.1095).

Deviatoric stress $\tilde{q}$: xLSTM achieves $R^2 = 0.920$ vs 0.902 for LSTM ($\Delta R^2 = +0.018$, 18% NMSE reduction).

Volumetric strain $\tilde{\varepsilon_v}$: $R^2 \approx 0.929$ for both xLSTM and LSTM.

ThermoXLSTM : ThermoXLSTM achieves $R^2 = 0.917$ on $\tilde{p}'$ and $R^2 = 0.898$ on q, slightly below the pure xLSTM results. This is expected because the physics constraint restricts stress predictions to the thermodynamically admissible subset, reducing the effective degrees of freedom of the model. For $\tilde{\varepsilon_v}$, ThermoXLSTM achieves slightly lower performance than xLSTM, consistent with the pattern on stress outputs.

Table 3. Prediction metrics on outputs, averaged across 10 extrapolation test sequences. ↓ lower is better; ↑ higher is better. Best result per row in bold.

	MSE↓	RMSE↓	MAE↓	NMSE↓	R^2 ↑
LSTM($\tilde{p}'$)	0.00105	0.02921	0.02649	0.1095	0.890
xLSTM($\tilde{p}'$)	**0.00070**	**0.02371**	**0.02147**	**0.0725**	**0.928**
Thermo($\tilde{p}'$)	0.00350	0.05804	0.03913	0.0832	0.917
LSTM($\tilde{q}$)	0.00382	0.05589	0.04985	0.0983	0.902
xLSTM($\tilde{q}$)	**0.00292**	**0.04947**	**0.04484**	**0.0802**	**0.920**
Thermo($\tilde{q}$)	0.00141	0.03715	0.02114	0.1018	0.898
LSTM($\tilde{\varepsilon_v}$)	0.000279	0.01522	**0.01176**	**0.0709**	0.929
xLSTM($\tilde{\varepsilon_v}$)	**0.000270**	**0.01496**	0.01338	0.0713	**0.929**
Thermo($\tilde{\varepsilon_v}$)	0.000276	0.01510	0.01360	0.0720	0.928

3.2 Qualitative analysis: stress paths and volumetric behaviour

3.2.1 Loose specimen ($I_D \approx 0.15$, $p_0 = 50$ kPa)

This specimen represents a contractive sand at low confining pressure—the loosest and lowest-pressure case in the dataset. Both LSTM and xLSTM reproduce the monotonically increasing q-ε_a response with broadly similar accuracy; the specimen shows no softening and moderate volumetric contraction. xLSTM provides closer tracking of the p' path.

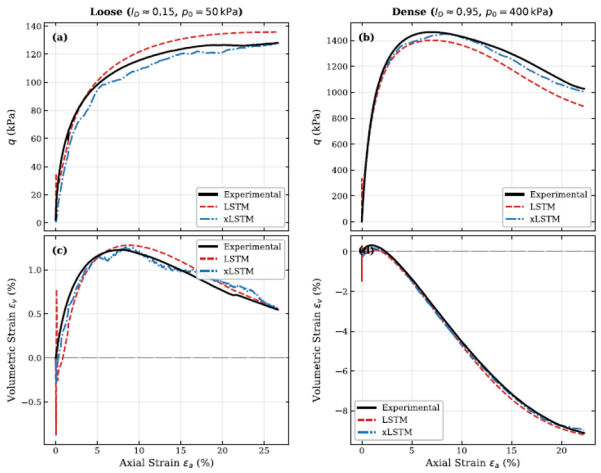

Fig. 1: Predicted versus experimental responses for the two representative test cases. Top row: deviatoric stress q versus axial strain ε_a. Bottom row: volumetric strain ε_v versus ε_a. Left column (a, c): loose specimen ($I_D \approx 0.15$, $p_0 = 50$ kPa). Right column (b, d): dense specimen ($I_D \approx 0.95$, $p_0 = 400$ kPa). Solid black: experimental; red dashed: LSTM; blue dash-dot: xLSTM.

3.2.2 Dense specimen ($I_D \approx 0.95$, $p_0 = 400$ kPa):

This is the most demanding extrapolation case: dense sand at high confining pressure exhibits a pronounced deviatoric stress peak followed by strain softening, and characteristic dilation ($\varepsilon_v < 0$) beginning before the peak. The LSTM baseline underestimates the peak deviatoric stress. xLSTM captures both the peak-and-softening pattern in q-ε_a and the sign reversal of ε_v substantially better. This is consistent with the larger NMSE improvement on q than on ε_v (Table 3): the peak and softening are controlled by the history-dependent stress-ratio and dilatancy coupling that matrix memory can represent more faithfully.

3.3 Thermodynamic admissibility of thermoXLSTM

A defining property of ThermoXLSTM is that stress is always the gradient of a finite-valued free energy (Equation (16)) and dissipation is constrained non-negative (Equation (17)). To quantify the consequence of this, we tracked the free energy Ψ_t and incremental dissipation $\mathcal{D}_t$ along all ten test sequences for both xLSTM and ThermoXLSTM.

Pure xLSTM, predicting stresses directly, occasionally produces small negative dissipation

increments, particularly in the early elastic regime of dense specimens at low confining pressure, exactly the extrapolation regime where the model has least data support. ThermoXLSTM produces no such violations: $\Psi_t \geq 0$ and $\mathcal{D}_t \geq 0$ are maintained across all timesteps of all ten test sequences.

This admissibility guarantee comes at a modest accuracy cost. We argue this is an acceptable trade-off for engineering deployment: a constitutive model that is slightly less accurate in R^2 but guaranteed thermodynamically consistent is preferable to one with marginally better metrics that can produce physically inadmissible stress paths when embedded in finite element analysis far from the training distribution.

4 Conclusion

We have presented the first systematic application of xLSTM to data-driven constitutive modeling of sand, together with ThermoXLSTM, a thermodynamics-informed extension that enforces physical admissibility by construction.

Three findings stand out:

1. xLSTM consistently and meaningfully outperforms LSTM on a genuine extrapolation benchmark: $R^2 = 0.928$ vs 0.890 on mean effective stress and $R^2 = 0.920$ vs 0.902 on deviatoric stress, with NMSE reductions of 34% and 18% respectively. The improvement is architecturally motivated: the matrix-memory mLSTM cell stores and retrieves second-order interactions between loading-history features, the co-evolution of axial strain, stress ratio, and volumetric strain that the scalar-gated LSTMs cannot represent.

2. The extrapolation test design—targeting specimens at the extremes of the initial confining pressure is crucial. The performance gap between xLSTM and LSTM would be substantially smaller on an interpolation test set. xLSTM's advantage is largest precisely where constitutive models are most uncertain in practice: dense dilative sand exhibiting peak and post-peak softening.

3. ThermoXLSTM extends the xLSTM backbone with a learnable Helmholtz free-energy potential and non-negative dissipation, guaranteeing thermodynamically admissible predictions by construction at a modest accuracy cost. This provides a physically trustworthy extension suitable for finite element deployment under loading conditions not covered by training data.

Together, these results establish xLSTM as the preferred recurrent backbone for data-driven constitutive modelling of path-dependent geomaterials, and demonstrate that physics-informed inductive bias can be embedded in a state-of-the-art sequence architecture without sacrificing its fundamental advantage. The field should look beyond LSTM as a default choice. It is also worth noting that the method is applicable to both drained and undrained loading conditions provided the data is available.

This work was supported by Prime Minister's Research Fellowship.

References

1. M. Beck, K. Pöppel, M. Spanring, A. Auer, O. Prudnikova, M. Kopp, G. Klambauer, J. Brandstetter, S. Hochreiter, xLSTM: extended long short-term memory, in Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Vancouver, Canada, December 10–15 (2024). https://arxiv.org/abs/2405.04517
2. K.H. Roscoe, J.B. Burland, On the generalised stress–strain behaviour of 'wet' clay, in Engineering Plasticity, ed. by J. Heyman, F.A. Leckie (Cambridge University Press, Cambridge, 1968), pp. 535–609
3. M. Taiebat, Y.F. Dafalias, SANISAND: simple anisotropic sand plasticity model. Int. J. Numer. Anal. Methods Geomech. **32**, 915–948 (2008). https://doi.org/10.1002/nag.651
4. N. Zhang, S.-L. Shen, A. Zhou, Y.-F. Jin, Application of LSTM approach for modelling stress–strain behaviour of soil. Appl. Soft Comput. **100**, 106959 (2021). https://doi.org/10.1016/j.asoc.2020.106959
5. L. Hadam, M. Amrane, N. Handel, R. Demagh, S. Messast, A. Gheris, Deep sequence learning architectures for predicting cyclic soil stress-strain behavior: comparative evaluation of LSTM, GRU, TCN, and hybrid models. Transp. Geotech., 101822 (2025). https://doi.org/10.1016/j.trgeo.2025.101822
6. S. Hochreiter, J. Schmidhuber, Long short-term memory. Neural Comput. **9**, 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
7. F. Masi, I. Stefanou, P. Vannucci, V. Maffi-Berthier, Material modeling via thermodynamics-based artificial neural networks, in Geometric Structures of Statistical Physics, Information Geometry, and Learning (SPIGL'20), Springer Proceedings in Mathematics & Statistics, vol. 361, ed. by F. Barbaresco, F. Nielsen (Springer, Cham, 2021), pp. 308–329. https://doi.org/10.1007/978-3-030-77957-3_16
8. M.M. Su, Y. Yu, T.H. Chen, N. Guo, Z.X. Yang, A thermodynamics-informed neural network for elastoplastic constitutive modeling of granular materials. Comput. Methods Appl. Mech. Eng. **430**, 117246 (2024). https://doi.org/10.1016/j.cma.2024.117246
9. T. Wichtmann, T. Triantafyllidis, An experimental database for the development, calibration and verification of constitutive models for sand with focus to cyclic loading: part I—tests with monotonic loading and stress cycles. Acta Geotech. **11**, 739–761 (2016). https://doi.org/10.1007/s11440-015-0402-z
10. S. Bengio, O. Vinyals, N. Jaitly, N. Shazeer, Scheduled sampling for sequence prediction with recurrent neural networks, in Advances in Neural Information Processing Systems **28** (NIPS 2015), Montréal, Canada, December 7–12 (2015). https://arxiv.org/abs/1506.03099