RELIABILITY-AWARE CONSTITUTIVE LEARNING FOR MATERIALS DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate characterization of material behavior at elevated temperatures is essential for engineering design but remains costly and time-consuming due to the extensive experimental testing required across multiple thermo-mechanical conditions. Recent advances in high-throughput mechanical testing workflows coupled with machine learning have enabled rapid prediction of stress-strain behavior from limited experimental measurements. However, existing approaches often lack interpretability and reliability assessment, provide limited support for out-of-distribution generalization to unseen alloy families, and offer no guarantees of physically meaningful stress-strain trajectories. We present a physics-aware constitutive learning framework that combines trajectory-aware training, a shape-aware loss, and a per-temperature mean stress loss to improve trajectory fidelity while preserving temperature-dependent material response. Beyond prediction accuracy, we introduce a reliability framework that quantifies neighborhood support within a low-dimensional constitutive materials space, providing an interpretable measure of extrapolation risk for previously unseen materials. We further propose a coverage-disagreement framework that identifies under-sampled thermo-mechanical regions and provides quantitative guidance for targeted experimental data acquisition. Across leave-one-material-out evaluations and previously unseen alloy families, the proposed framework produces physically consistent stress-strain curves while achieving average prediction errors below 5% on previously unseen materials. Together, these contributions transform constitutive machine learning from a prediction-focused task into a reliability-aware decision-support framework capable of assessing extrapolation risk and guiding future experimentation.

1 INTRODUCTION

Accurate constitutive descriptions of material behavior are fundamental to engineering design, process optimization, and simulation-driven decision making. For metallic alloys, constitutive behavior often varies significantly with temperature, requiring extensive experimental characterization across multiple thermo-mechanical conditions. These challenges have motivated growing interest in machine learning approaches that can accelerate constitutive characterization and reduce experimental burden.

Recent advances in machine learning have demonstrated promising capabilities for constitutive modeling and material characterization. However, predictive accuracy alone is often insufficient for practical deployment. Models are routinely applied to materials that differ from those observed during training, raising a fundamental question: can the resulting prediction be trusted? Furthermore, existing approaches provide limited guidance regarding which materials or thermo-mechanical conditions should be characterized next to improve model performance.

To address these challenges, this work introduces a **reliability-aware constitutive learning framework** that combines physics-informed learning, materials-space reliability assessment, and coverage-disagreement analysis. The overall workflow is illustrated in Figure 1. Starting from measurements on a new custom sample geometry (named MAPS), the framework evaluates materials-space support, predicts constitutive response using physics-informed objectives, and identifies thermo-mechanical regions that would benefit most from additional characterization. The frame-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

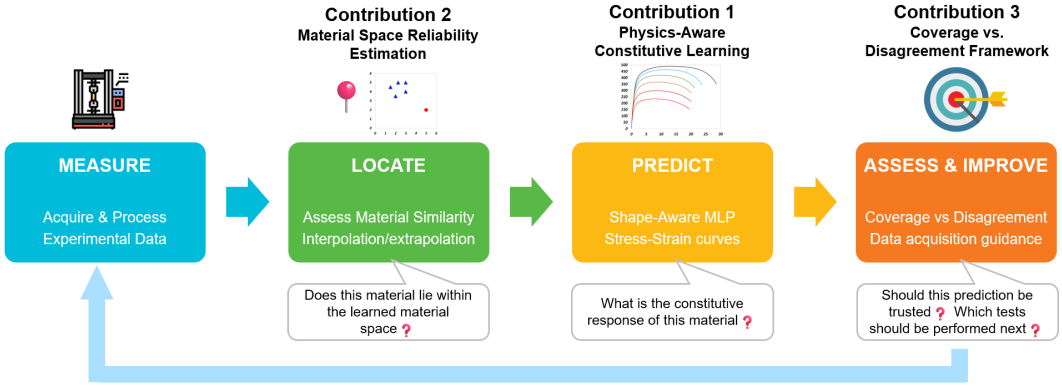

Figure 1: Overview of the proposed reliability-aware constitutive learning framework.

work is validated across 18 materials spanning four alloy families, including leave-one-material-out studies and previously unseen alloy families.

The results demonstrate that constitutive machine learning can provide value beyond predictive accuracy alone by combining prediction, reliability assessment, and experimental planning within a unified decision-support framework.

2 RELATED WORK

Machine learning has been increasingly applied to constitutive modeling and material characterization. Prior work has explored machine-learning surrogates trained on stress-strain data as alternatives to traditional constitutive equations (Hussain et al., 2026; Li et al., 2019; Pandya et al., 2020; Murugesan et al., 2023), as well as data-driven characterization workflows that combine experiments, simulations, and machine learning to recover constitutive behavior from indirect measurements (Weng et al., 2020; van Rooyen & Becker, 2018; Agha et al., 2025). The present work belongs to the latter category. Unlike existing approaches that primarily focus on prediction accuracy, this work additionally addresses prediction reliability for previously unseen materials and guidance for future data acquisition.

Physics-informed machine learning has emerged as an effective approach for improving accuracy, generalization, and physical consistency in scientific applications (Raissi et al., 2019; Michaloglou et al., 2026). Within materials science, physics-informed methods have been applied to constitutive modeling, fatigue and creep prediction, materials design, and multiscale property estimation (Zhang et al., 2021; Jeong et al., 2024; Moon et al., 2025). While existing approaches use physically motivated constraints to improve predictive accuracy and reduce non-physical behavior, comparatively little attention has been devoted to assessing whether a prediction is supported by similar materials within the training dataset or whether it relies on extrapolation beyond previously observed constitutive behavior.

Assessing prediction reliability and guiding future data acquisition remain fundamental challenges in scientific machine learning. Existing approaches include uncertainty quantification, out-of-distribution detection, representation-space analysis, active learning, and uncertainty-guided sampling strategies (Lakshminarayanan et al., 2017; Gal & Ghahramani, 2016; Liu et al., 2020; Ramalho & Miranda, 2020; Chen et al., 2022; Wakabayashi et al., 2022; Cao et al., 2026). However, these methods generally do not distinguish between insufficient representation of material classes and insufficient sampling of thermo-mechanical conditions. The proposed framework addresses this gap through two complementary approaches: a materials-space reliability framework that quantifies support from previously observed materials and a coverage-disagreement framework that identifies under-sampled thermo-mechanical regions.

3 METHODS

3.1 DATASET

The dataset is based on a previously published MAPS specimen benchmark (Agha et al., 2025), a high-throughput thermo-mechanical characterization approach in which a spatial temperature gradient is maintained across the specimen during deformation. Each test records force measurements together with strain and temperature data from 12 regions of interest (ROIs), enabling simultaneous characterization of constitutive response across a range of temperatures within a single experiment.

The dataset consists of virtual MAPS tensile test data generated using finite-element analysis (FEA) spanning 18 engineering materials across four alloy families: aluminum, steel, nickel, and titanium. Approximately 16–18 tests were simulated per material, resulting in a total of 306 tests spanning heater temperatures from 25°C to 600°C. Each test contains approximately 200–300 displacement increments, with each increment represented by a unique combination of force, ROI strain, and ROI temperature measurements. Collectively, the benchmark contains approximately 50,000 observations.

Model development was performed using 6 aluminum and 6 steel alloys from the published benchmark dataset using a leave-one-material-out (LOMO) protocol, in which an entire material was excluded from training and reserved exclusively for testing. To evaluate extrapolation beyond the training distribution, six additional materials spanning nickel, steel, and titanium-based alloy systems were incorporated for external validation. This combination of LOMO evaluation and validation on previously unseen alloy families provides a stringent assessment of both constitutive prediction accuracy and reliability under materials-space extrapolation conditions.

3.2 NEURAL NETWORK ARCHITECTURE

To enable direct comparison with the published benchmark, the neural network architecture was adopted without modification from the baseline model described previously (Agha et al., 2025). The objective of this work was not to improve performance through architectural changes, but rather to investigate the impact of physics-informed learning objectives and reliability-aware assessment. Each MAPS observation consists of force, temperature, and strain measurements collected from 12 predefined regions of interest (ROI). In addition to the raw measurements, centered-difference features were computed for force and ROI strain signals to capture local deformation trends during loading. The final model utilized a 38-dimensional input vector consisting of force, 12 ROI temperatures, 12 ROI strains, and 13 centered-difference features. The network output consisted of 12 stress values corresponding to the constitutive response at each ROI.

Additional architecture, optimization, and training details are provided in Appendix A.

3.3 PHYSICS-INFORMED OBJECTIVES

While the baseline multilayer perceptron provides accurate stress predictions, conventional pointwise objectives do not explicitly encourage physically consistent constitutive behavior. The proposed framework augments the baseline prediction objective with trajectory-aware batching and physics-informed loss functions designed to preserve constitutive shape characteristics and temperature-dependent stress trends.

To preserve spatial relationships between neighboring ROIs, complete time sequence of MAPS tests were maintained during batching. This enables the use of physics-informed objectives that operate on relationships between neighboring ROI predictions.

The baseline prediction objective minimizes the mean squared error (MSE),

$$L_{\mathrm{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\sigma_i^{\mathrm{pred}} - \sigma_i^{\mathrm{actual}}\right)^2 \tag{1}$$

To improve constitutive trajectory fidelity, a shape-aware loss penalizes disagreement between predicted and observed stress differences across neighboring ROIs,

$$L_{\text{shape}} = \frac{1}{N}\sum_{i=1}^{N}\left(\Delta\sigma_i^{\text{pred}} - \Delta\sigma_i^{\text{actual}}\right)^2 \tag{2}$$

A per-temperature mean stress loss was additionally introduced to preserve temperature-dependent stress levels,

$$L_{\text{temp}} = \frac{1}{T}\sum_{t=1}^{T}\left(\bar{\sigma}_t^{\text{pred}} - \bar{\sigma}_t^{\text{actual}}\right)^2 \tag{3}$$

The final training objective combines the three loss terms,

$$L_{\text{total}} = L_{\text{MSE}} + \lambda_{\text{shape}}L_{\text{shape}} + \lambda_{\text{temp}}L_{\text{temp}} \tag{4}$$

where λ_{shape} and λ_{temp} control the contribution of the physics-informed objectives. The final model used $\lambda_{\text{shape}} = 800$ and $\lambda_{\text{temp}} = 1.5$. A schematic illustration of the physics-informed objectives, additional implementation details and hyperparameter studies are provided in Appendix A.

3.4 Materials-Space Reliability Framework

Accurate constitutive predictions alone do not indicate whether a material is well represented by the available training data. To quantify the degree of extrapolation required for an unseen material, a materials-space reliability framework was developed. Each MAPS test is represented by a 63-dimensional material descriptor and projected into a five-dimensional material space using Principal Component Analysis (PCA). Reliability is subsequently evaluated using neighborhood-based metrics that quantify materials-space support and extrapolation risk.

3.4.1 Neighborhood Distance

Each material is represented by the centroid of its tests in the five-dimensional PCA space. The neighborhood distance between two materials is defined as

$$D_{ij} = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2 \tag{5}$$

and normalized according to

$$D_{\text{norm}} = \frac{D}{\sqrt{N}} \tag{6}$$

where N denotes the number of neighboring materials.

3.4.2 Local PC1 Variance

To assess local neighborhood structure, PCA is recomputed using neighboring materials and the variance explained by the first principal component is calculated as

$$V_{\text{local}} = \frac{\lambda_1}{\sum_{k=1}^{5}\lambda_k} \tag{7}$$

where λ_k denotes the eigenvalue associated with the kth local principal component.

3.4.3 Direction Similarity

To evaluate whether neighboring materials exhibit similar temperature-dependent behavior, a direction vector is constructed from the PCA-space trajectory of each material. Direction similarity is then calculated using cosine similarity,

$$S = \frac{\mathbf{V}_i \cdot \mathbf{V}_j}{\|\mathbf{V}_i\|_2\,\|\mathbf{V}_j\|_2} \tag{8}$$

where larger values indicate stronger agreement in constitutive evolution with temperature.

3.4.4 RELIABILITY ASSESSMENT

The three reliability metrics are combined to determine the level of neighborhood support available for a given material. A material is classified as RELIABLE when

$$\begin{aligned} D_{\text{norm}} &< 4.5, \\ V_{\text{local}} &> 0.95, \\ S &> 0.95. \end{aligned} \tag{9}$$

Materials failing any of these criteria are classified as CAUTION. A CAUTION classification does not imply prediction failure; rather, it indicates increased reliance on extrapolation beyond the material relationships represented within the training dataset. These thresholds are established using the aforementioned training data and would need to be adapted for a different dataset.

3.5 COVERAGE-DISAGREEMENT FRAMEWORK

The materials-space reliability framework described in Section 3.4 quantifies whether an entire material is well represented within the available training dataset. However, it does not provide insight into whether specific thermo-mechanical regions within a material have been sampled sufficiently. To address this limitation, a complementary coverage-disagreement framework was developed. As illustrated in Figure 2, the framework evaluates experimental support and response consistency within the thermo-mechanical space. Coverage quantifies the amount of experimental support available within a temperature regime and is defined as

$$\text{Coverage}(T) = N_T \tag{10}$$

where N_T denotes the number of independent observations within temperature bucket T.

To quantify local variability, a disagreement metric was introduced,

$$\text{Disagreement} = \frac{|\sigma_{\max} - \sigma_{\min}|}{|\sigma_{\text{mean}}|} \times 100 \tag{11}$$

where $\sigma_{\max}$, $\sigma_{\min}$, and σ_{mean} denote the maximum, minimum, and mean stress values within a thermo-mechanical cell.

Together, coverage and disagreement partition the thermo-mechanical space into four characteristic regimes: well-constrained regions (high coverage, low disagreement), sparse but consistent regions (low coverage, low disagreement), disagreement despite coverage (high coverage, high disagreement), and under-sampled high-variability regions (low coverage, high disagreement). The final category represents the highest priority for future experimental characterization because it combines limited observational support with elevated uncertainty.

The materials-space reliability framework addresses the question, "Which materials should be added to the training dataset next?" while the coverage-disagreement framework addresses "Which thermo-mechanical conditions should be tested next?" Together, these complementary frameworks provide actionable guidance for material selection and experimental planning.

Additional evaluation details are provided in Appendix B.

4 RESULTS

4.1 DO PHYSICS-INFORMED OBJECTIVES IMPROVE CONSTITUTIVE LEARNING?

Relative to the MSE-only baseline, physics-informed objectives improved constitutive fidelity, physical consistency, and training stability while maintaining competitive prediction accuracy. The final model reduced Shape RMSE (Root Mean Square Error) from 1.36 to 0.97, reduced monotonicity violations from 0.20 to 0.12, and lowered seed-to-seed Mean Absolute Percentage Error (MAPE) variability from 1.18% to 0.86%. Detailed ablation results are provided in Appendix E.

Figure 3 provides a qualitative comparison for AA3003-H14, one of the most challenging leave-one-material-out cases. The MSE-only baseline frequently exhibits non-physical softening behavior

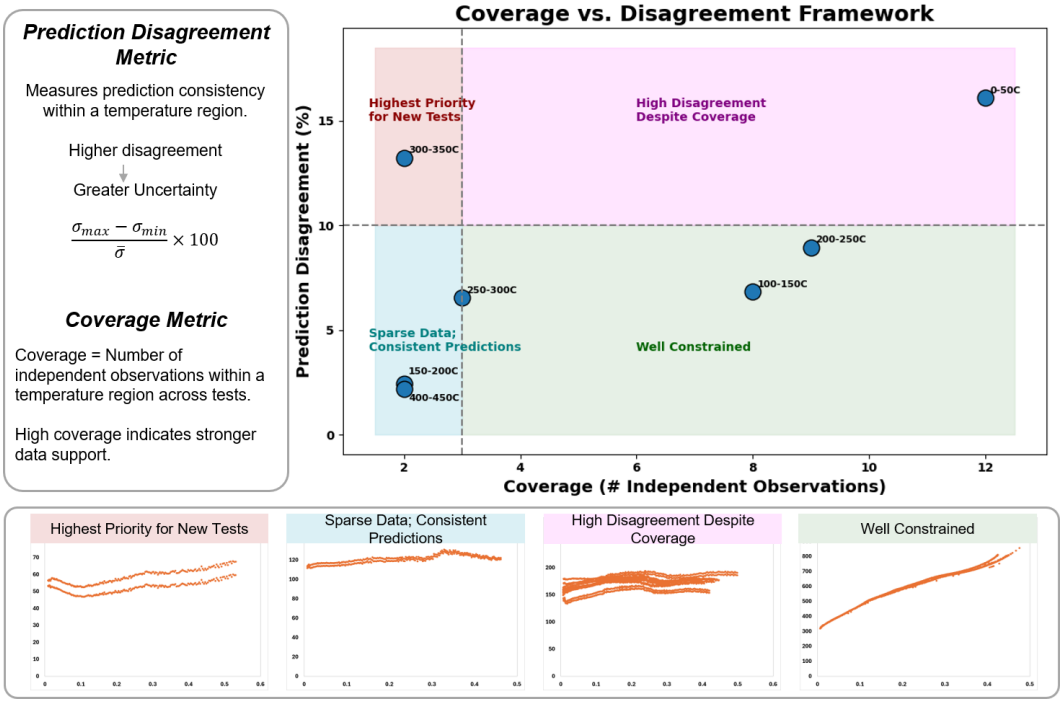

Figure 2: Coverage-disagreement framework.

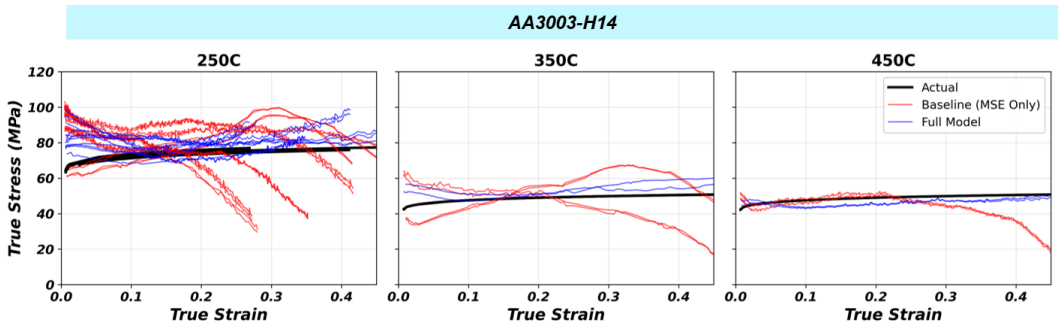

Figure 3: Comparison of baseline and physics-informed constitutive predictions for AA3003-H14, one of the most challenging LOMO evaluation cases.

at elevated temperatures, whereas the full model more closely follows the experimental stress-strain trajectories and preserves constitutive shape.

These results demonstrate that prediction error alone is insufficient for evaluating constitutive learning models. By incorporating shape-based and temperature-dependent constraints, the proposed framework produces more physically realistic constitutive trajectories while improving training stability and reducing non-physical behavior.

4.2 CAN MATERIALS-SPACE RELIABILITY PREDICT GENERALIZATION DIFFICULTY?

To evaluate whether the proposed materials-space reliability framework provides meaningful guidance regarding model generalization, the reliability metrics were computed for each of the 12 LOMO experiments and compared against the resulting prediction performance. Table 1 summarizes the neighborhood distance, local PC1 variance, directional similarity, reliability classification, and prediction statistics for each material. The majority of materials were classified as RELIABLE and

Table 1: Validation of the materials-space reliability framework using leave-one-material-out evaluation.

Material	D_{norm}	Local PC1 Var	Dir. Sim.	Reliability	Mean MAPE (%)	Std MAPE (%)
SS430-O	1.69	1.00	0.99	RELIABLE	3.7	0.87
AA6061-T6	2.06	0.96	0.98	RELIABLE	6.5	0.90
SS316	1.47	1.00	0.97	RELIABLE	4.3	0.44
C40	1.75	0.98	0.99	RELIABLE	3.4	0.45
SS304	1.90	1.00	0.99	RELIABLE	3.7	0.56
C10	2.93	0.99	1.00	RELIABLE	3.5	0.36
AA6063-T6	3.05	0.97	0.95	RELIABLE	8.7	0.90
17-4PH-H900	4.20	1.00	0.99	RELIABLE	6.0	1.18
AA7075-T6	2.33	**0.94**	0.99	CAUTION	8.9	**3.06**
AA2024	2.14	0.97	**0.95**	CAUTION	5.6	1.01
AA5052	3.71	**0.90**	**0.91**	CAUTION	**9.6**	**2.41**
AA3003-H14	**4.98**	0.97	**0.92**	CAUTION	**13.3**	**2.01**

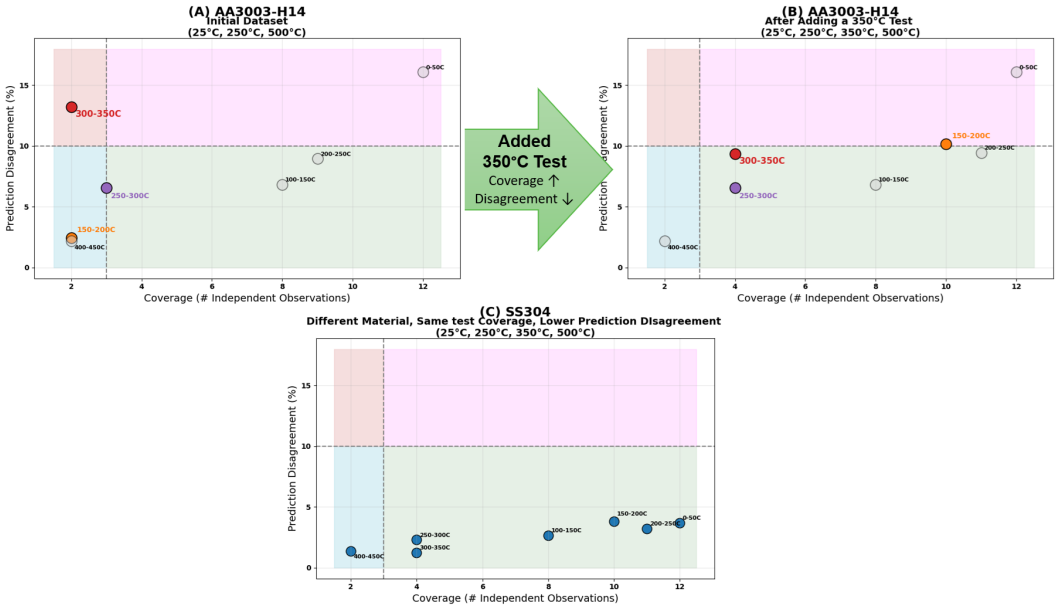

Figure 4: Coverage-disagreement analysis for AA3003-H14 and SS304.

achieved low prediction error and low seed-to-seed variability, indicating that materials located within well-supported regions of material space can be generalized reliably.

In contrast, AA7075-T6, AA2024, AA5052, and AA3003-H14 were classified as CAUTION. These materials exhibited either increased neighborhood distance, reduced directional similarity, or both, and were among the most challenging LOMO cases. AA3003-H14 exhibited the largest normalized distance ($D_{\text{norm}} = 4.98$) and the highest mean prediction error (13.3%), while AA5052 combined reduced directional similarity ($S = 0.91$) with elevated prediction variability (2.41% standard deviation). Overall, materials classified as CAUTION generally exhibited higher prediction error, larger run-to-run variability, or both, demonstrating that the proposed reliability framework provides meaningful information regarding expected generalization difficulty and model robustness.

4.3 CAN COVERAGE AND PREDICTION DISAGREEMENT GUIDE EXPERIMENTAL PLANNING?

The proposed coverage-disagreement framework was evaluated using AA3003-H14, one of the most challenging materials in the dataset. Figure 4 illustrates how the framework identifies thermo-mechanical regions that are likely to benefit from additional experimental characterization.

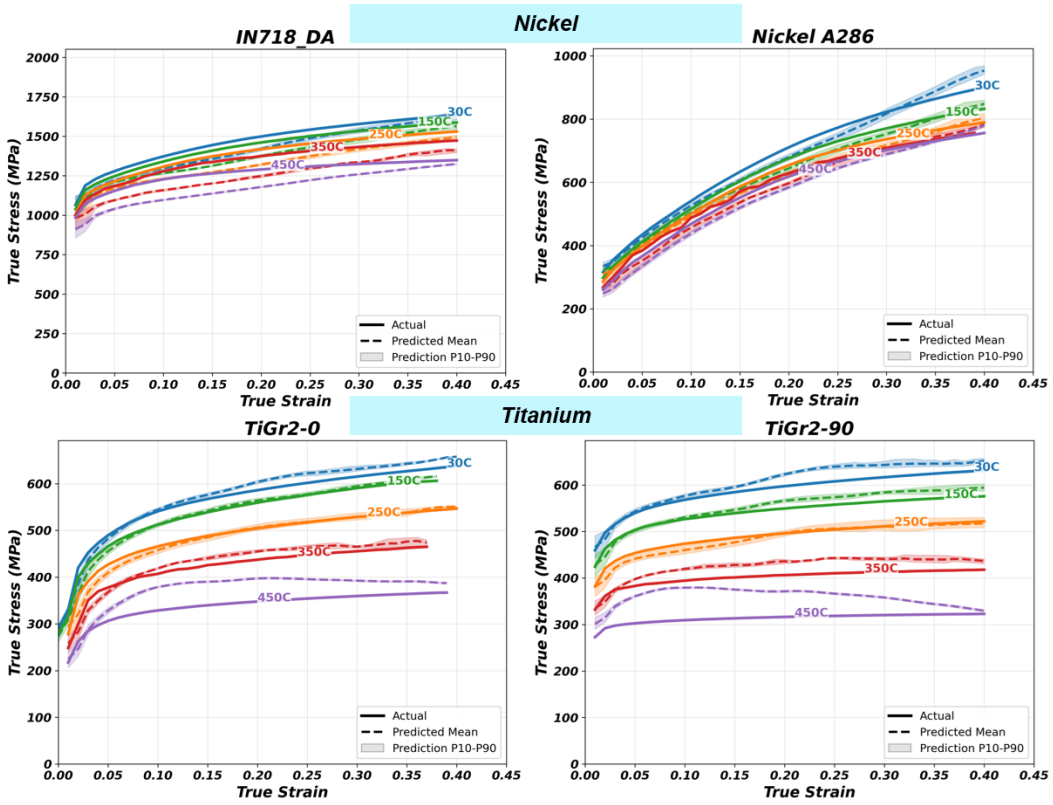

Figure 5: Generalization to previously unseen alloy families.

Figure 4A shows the initial AA3003-H14 dataset, which contains measurements at 25°C, 250°C, and 500°C. The temperature range between approximately 300°C and 350°C exhibits elevated prediction disagreement together with limited local coverage, identifying this region as a candidate for additional experimentation. A new constitutive test was subsequently conducted at 350°C and incorporated into the dataset.

As shown in Figure 4B, the additional measurement increases coverage within the previously undersampled region and reduces the corresponding disagreement. This result demonstrates that the framework can guide targeted data acquisition and dataset expansion. Figure 4C highlights a second important observation. Despite a comparable number of temperature measurements, SS304 exhibits substantially lower disagreement across the temperature range than AA3003-H14. Prediction reliability therefore depends not only on data density but also on the underlying constitutive complexity of the material.

Overall, the coverage-disagreement framework identifies thermo-mechanical regions that are both under-sampled and uncertain, providing actionable guidance for experimental planning and constitutive data acquisition.

4.4 DOES THE FRAMEWORK GENERALIZE TO PREVIOUSLY UNSEEN ALLOY FAMILIES?

To evaluate extrapolation beyond the materials used during model development, the framework was assessed on six previously unseen materials spanning nickel-, steel-, and titanium-based alloy systems. Figure 5 and Table 2 summarize the resulting prediction accuracy and reliability classifications.

The framework generalized successfully to the previously unseen nickel and steel alloys, which were classified as RELIABLE and achieved mean MAPE values between 2.6% and 4.7%.

Table 2: Generalization performance on previously unseen alloy families.

Alloy Family	Material	D_{norm}	PC1 Var	Dir. Sim.	Reliability	MAPE (%)	Std. (%)
Nickel	IN718-DA	3.12	1.00	0.99	RELIABLE	4.7	0.34
Nickel	Nickel A286	1.69	1.00	0.99	RELIABLE	3.1	0.33
Steel	SS15-5PH	2.09	1.00	0.99	RELIABLE	3.3	0.48
Steel	SS301-H12	2.86	1.00	0.96	RELIABLE	2.6	0.10
Titanium	TiGr2-0	5.98	0.99	0.59	CAUTION	3.7	0.51
Titanium	TiGr2-90	3.18	0.99	0.80	CAUTION	3.6	0.54

In contrast, the two titanium alloys were classified as CAUTION. TiGr2-0 exhibited the largest normalized neighborhood distance ($D_{\text{norm}} = 5.98$) and the lowest direction similarity ($S = 0.59$), indicating limited materials-space support and substantially different temperature-dependent behavior relative to the training materials. TiGr2-90 remained below the reliability threshold due to reduced direction similarity ($S = 0.80$). Importantly, reliability should not be interpreted as a direct estimate of prediction error. Despite being classified as CAUTION, both titanium alloys achieved relatively low prediction errors (3.6–3.7% MAPE). Reliability therefore reflects materials-space support and extrapolation risk rather than prediction accuracy alone. These results demonstrate that the framework can generalize beyond the alloy families used during training while simultaneously providing an interpretable assessment of prediction support.

5 LIMITATIONS

The current study evaluates 18 materials across four alloy families and focuses on uniaxial tensile constitutive behavior. Future work should investigate larger material libraries, additional alloy systems, and more complex loading conditions. Furthermore, reliability classifications are defined relative to the material space represented by the current training dataset and may evolve as larger constitutive databases become available.

6 CONCLUSION

This work introduces a reliability-aware scientific machine learning framework for constitutive response prediction, materials-space reliability assessment, and data-driven experimental planning. The proposed approach addresses two fundamental challenges in data-driven constitutive modeling: generating physically realistic stress-strain predictions and quantifying the reliability of those predictions for new materials and thermo-mechanical conditions.

The results demonstrate that shape-based and per-temperature mean stress objectives improve constitutive trajectory fidelity, reduce non-physical behavior, and increase training stability while maintaining competitive prediction accuracy. Beyond predictive performance, the materials-space reliability framework successfully identified materials that were difficult to generalize to under leave-one-material-out evaluation, while the coverage-disagreement framework identified under-sampled thermo-mechanical regions that would benefit from additional characterization. The framework was further validated on previously unseen nickel-, steel-, and titanium-based alloys. Importantly, reliability should be interpreted as a measure of extrapolation risk rather than prediction error alone. Materials classified as CAUTION may still be predicted accurately, but those predictions rely on greater extrapolation beyond the material relationships represented within the training dataset.

More broadly, the materials-space reliability framework addresses the question, “Which materials should be added to the training dataset next?” while the coverage-disagreement framework addresses “Which thermo-mechanical conditions should be tested next?” Together, these complementary capabilities transform the model from a predictive tool into a decision-support framework for material selection, experimental planning, and future dataset expansion.

AI USE STATEMENT

Generative AI tools were used during the preparation of this work for refining manuscript organization, improving readability and clarity of the text, and assisting with software code implementation and debugging. AI tools were also used as a discussion aid for exploring alternative experimental and modeling approaches and to refine the presentation and communication of research ideas.

Generative AI tools were NOT used to generate datasets, conduct experiments, produce experimental results, formulate the scientific contributions of this work, generate figures or tables, perform quantitative analyses, interpret results, or make final research decisions. All model development, experimental design, data generation, analyses, figures, tables, evaluations, interpretations, and conclusions were developed, verified, and approved by the authors.

All AI-assisted outputs, including text suggestions and code fragments, were reviewed, validated, and modified as necessary by the authors prior to inclusion in the project. The authors take full responsibility for the content of this manuscript and all associated research artifacts.

REPRODUCIBILITY STATEMENT

Dataset details, model architecture, training procedures, evaluation metrics, and supplementary ablation studies are provided in Appendices A-E. All experiments were evaluated using a 10-seed leave-one-material-out protocol together with additional validation on previously unseen alloy families.

REFERENCES

Akshat Agha, Janet B Davis, and Sergio L dos Santos e Lucato. Novel approach to rapid and accurate temperature-dependent mechanical testing using machine learning. *Materials & Design*, pp. 115307, 2025.

Bin Cao, Jie Xiong, Jiaxuan Ma, Yuan Tian, Yirui Hu, Mengwei He, Longhan Zhang, Jiayu Wang, Jian Hui, Li Liu, et al. Bgolearn: a unified bayesian optimization framework for accelerating materials discovery. *npj Computational Materials*, 2026.

Shuizhou Chen, He Cao, Qiubao Ouyang, Xing Wu, and Quan Qian. Alds: An active learning method for multi-source materials data screening and materials design. *Materials & Design*, 223: 111092, 2022.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Arif Hussain, Amir Hosein Sakhaei, and Mahmood Shafiee. Machine learning-based constitutive modelling for material non-linearity: A review. *Mechanics of Advanced Materials and Structures*, 33(1):2439557, 2026.

Iksu Jeong, Maenghyo Cho, Hayoung Chung, and Do-Nyun Kim. Data-driven nonparametric identification of material behavior based on physics-informed neural network with full-field data. *Computer Methods in applied mechanics and Engineering*, 418:116569, 2024.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Xueyang Li, Christian C Roth, and Dirk Mohr. Machine-learning based temperature-and rate-dependent plasticity model: Application to analysis of fracture experiments on dp steel. *International Journal of Plasticity*, 118:320–344, 2019.

Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in neural information processing systems*, 33:7498–7512, 2020.

Alkmini Michaloglou, Ioannis Papadimitriou, Ilias Gialampoukidis, Stefanos Vrochidis, and Ioannis Kompatsiaris. Physics-informed neural networks in materials modeling and design: A review. *Archives of Computational Methods in Engineering*, 33(4):5223–5260, 2026.

Hyeonbin Moon, Donggeun Park, Hanbin Cho, Hong-Kyun Noh, Jae Hyuk Lim, and Seunghwa Ryu. Physics-informed neural network-based discovery of hyperelastic constitutive models from extremely scarce data. *Computer Methods in Applied Mechanics and Engineering*, 446:118258, 2025.

Mohanraj Murugesan, Jae-Hyeong Yu, Kyu-Seok Jung, Sung-Min Cho, Krishna Singh Bhandari, Wanjin Chung, and Chang-Whan Lee. Supervised machine learning approach for modeling hot deformation behavior of medium carbon steel. *steel research international*, 94(2):2200188, 2023.

Kedar S Pandya, Christian C Roth, and Dirk Mohr. Strain rate and temperature dependent fracture of aluminum alloy 7075: Experiments and neural network modeling. *International Journal of Plasticity*, 135:102788, 2020.

M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. ISSN 0021-9991.

Tiago Ramalho and Miguel Miranda. Density estimation in representation space to predict model uncertainty. In *International workshop on engineering dependable and secure machine learning systems*, pp. 84–96. Springer, 2020.

Melody van Rooyen and Thorsten Hermann Becker. High-temperature tensile property measurements using digital image correlation over a non-uniform temperature field. *The Journal of Strain Analysis for Engineering Design*, 53(3):117–129, 2018.

Yuki K Wakabayashi, Takuma Otsuka, Yoshiharu Krockenberger, Hiroshi Sawada, Yoshitaka Taniyasu, and Hideki Yamamoto. Bayesian optimization with experimental failure for high-throughput materials growth. *npj Computational Materials*, 8(1):180, 2022.

Jian Weng, Rebecka Lindvall, Kejia Zhuang, Jan-Eric Ståhl, Han Ding, and Jinming Zhou. A machine learning based approach for determining the stress-strain relation of grey cast iron from nanoindentation. *Mechanics of Materials*, 148:103522, 2020.

Xiao-Cheng Zhang, Jian-Guo Gong, and Fu-Zhen Xuan. A physics-informed neural network for creep-fatigue life prediction of components at elevated temperatures. *Engineering Fracture Mechanics*, 258:108130, 2021.

A APPENDIX. NEURAL NETWORKS ARCHITECTURE

To enable direct comparison with the published benchmark, the neural network architecture was adopted without modification from the baseline model described in (Agha et al., 2025). The objective of this work was not to improve performance through architectural changes, but rather to investigate the impact of physics-informed learning objectives and reliability-aware assessment.

A.1 INPUTS AND OUTPUTS

Each MAPS simulation generated force, displacement, and temperature-strain measurements from 12 predefined regions of interest (ROI). The original feature set therefore consisted of 12 ROI strains, 12 ROI temperatures, force, and displacement. Because all tests were conducted at the same displacement rate, displacement did not provide additional discriminatory information and was excluded from the model inputs. In addition to the raw measurements, centered-difference features were computed for force and the 12 ROI strain signals to capture local deformation trends during loading. The final model therefore utilized a 38-dimensional input vector consisting of force, 12 ROI temperatures, 12 ROI strains, and 13 centered-difference features. The network output consisted of 12 stress values corresponding to the constitutive response at each ROI. For each displacement increment, the centered difference was calculated within an individual MAPS test using a finite-difference approximation,

$$\Delta x_i = \frac{x_{i+2} - x_{i-2}}{4}$$

where x_i represents the value of a force or strain feature at displacement increment i, and Δx_i denotes the corresponding centered-difference approximation. This formulation captures the local rate of change of the underlying response while remaining less sensitive to measurement noise than simple forward or backward differences. The resulting features provide information regarding deformation history and constitutive evolution in addition to the instantaneous force, strain, and temperature measurements.

A.2 ARCHITECTURE

The multilayer perceptron (MLP) architecture consists of an input layer with 25 neurons, followed by three fully connected hidden layers containing 256, 128, and 64 neurons, respectively. Rectified Linear Unit (ReLU) activation functions were used for all hidden layers, while a linear activation function was used in the output layer. Consistent with the benchmark implementation, all input features were standardized to zero mean and unit variance prior to training

A.3 TRAINING CONFIGURATION

Model training was performed using an 80/20 train-validation split. A maximum of 500 epochs was specified, with early stopping applied based on validation loss using a patience of 20 epochs. The model parameters corresponding to the lowest validation loss were retained for subsequent evaluation, preventing overfitting and ensuring consistent model selection across experiments. During development, several optimization strategies were evaluated, including Adam, AdamW, AdamW with gradient clipping, and multiple shape-loss weighting configurations. AdamW-based approaches frequently exhibited staircase-like optimization behavior and did not provide a consistent improvement in leave-one-material-out performance. Consequently, the Adam optimizer was retained for all experiments reported in this work. Additional optimizer and loss-function ablation studies are summarized in Appendix E.

A.4 PHYSICS-INFORMED OBJECTIVES

The physics-informed objectives introduced in Section 3.3 require preservation of the spatial relationships between neighboring regions of interest (ROIs) within an individual MAPS test. Conventional row-wise batching randomly mixes observations originating from different materials, temperatures, and deformation states, making it difficult to apply losses that depend on relationships between neighboring ROIs. To address this limitation, complete MAPS test trajectories were preserved during batching rather than randomly shuffling individual observations. This trajectory-aware

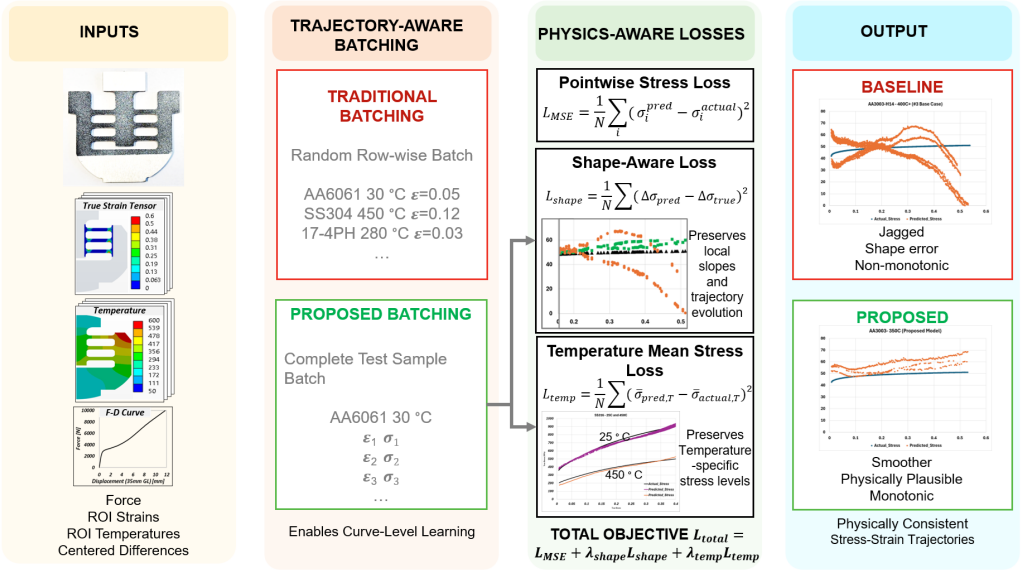

Figure 6: Physics-informed training objectives used to improve constitutive prediction accuracy and physical consistency.

batching strategy maintains the physical relationships between neighboring ROIs and enables the use of shape- and temperature-based objectives.

A.4.1 Shape Loss

The shape loss operates on stress differences between neighboring ROI locations rather than absolute stress values. The predicted and actual stress differences are defined as

$$\Delta\sigma_i^{\text{pred}} = \sigma_{i+1}^{\text{pred}} - \sigma_i^{\text{pred}}$$

and

$$\Delta\sigma_i^{\text{actual}} = \sigma_{i+1}^{\text{actual}} - \sigma_i^{\text{actual}}$$

respectively. By matching these local stress differences, the network is encouraged to preserve constitutive trajectory shape and local slope information across neighboring ROI locations.

A.4.2 Per-Temperature Mean Stress Loss

The shape loss preserves local constitutive trends but does not explicitly constrain the average stress level associated with a given temperature. To incorporate this additional physics-based information, observations were grouped according to temperature and the mean stress response was evaluated within each group. The predicted and actual mean stresses for temperature group t are denoted by

$$\bar{\sigma}_t^{\text{pred}}$$

and

$$\bar{\sigma}_t^{\text{actual}}$$

respectively. This objective preserves temperature-dependent stress levels across the dataset and discourages systematic prediction bias at individual temperatures.

A.4.3 Hyperparameter Selection

The loss weights λ_{shape} and λ_{temp} were selected through a two-stage hyperparameter optimization study. An initial coarse grid search was performed over candidate values for both parameters,

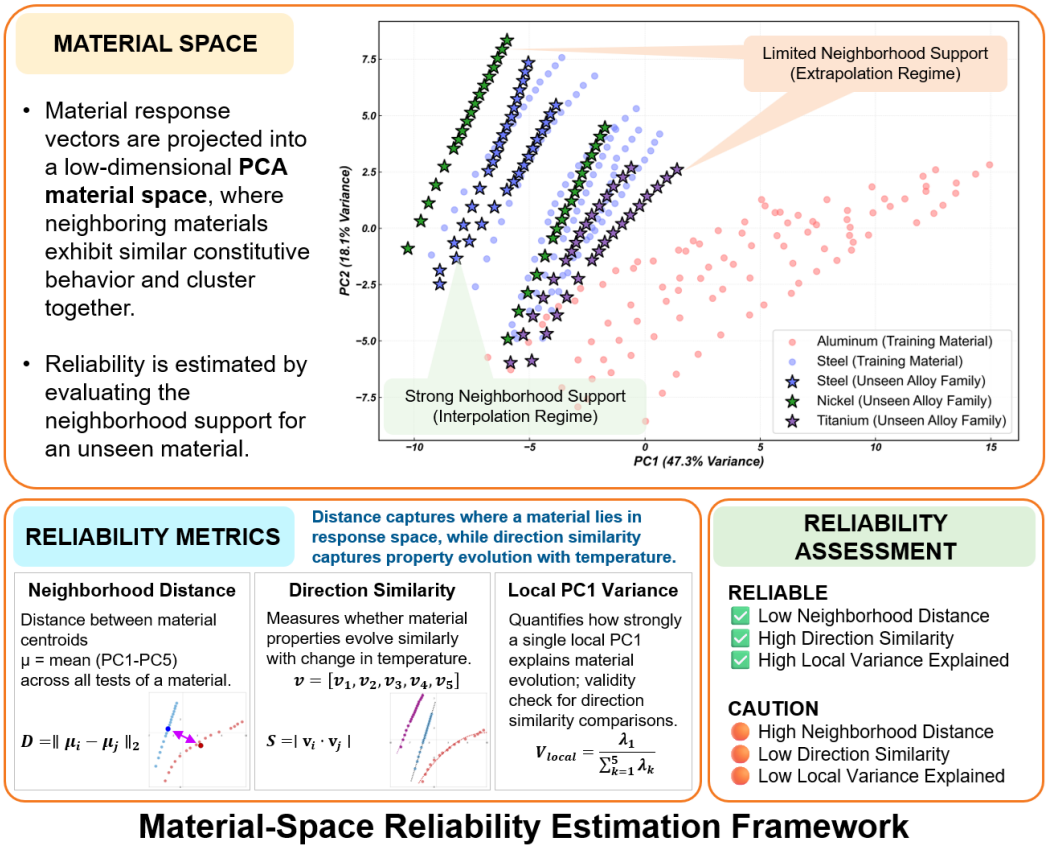

Figure 7: Materials-space reliability framework.

after which the most promising configurations were evaluated using full 10-seed leave-one-material-out (LOMO) experiments across all 12 training materials. Candidate models were assessed using prediction accuracy (MAPE), constitutive shape fidelity (Shape RMSE), and physical consistency (Monotonicity Violation). The final model used

$$\lambda_{\text{shape}} = 800$$

and

$$\lambda_{\text{temp}} = 1.5$$

which provided the best overall balance between prediction accuracy, constitutive fidelity, and training stability. Additional loss-function ablation results are provided in Appendix E.

A.5 Materials-Space Construction and Reliability Threshold Selection

Figure 7 illustrates the construction of the materials-space reliability framework.

A.5.1 Material Descriptor Construction

Each MAPS test contains hundreds of displacement increments and thousands of raw measurements distributed across multiple regions of interest (ROIs). To enable material-level comparison, each test was converted into a fixed-length 63-dimensional material descriptor designed to capture its mechanical and thermal behavior.

The descriptor consists of:

- 11 force-history features sampled at predefined displacement values;

- 5 global force descriptors, including stiffness, peak force, displacement at peak force, energy shape factor, and hardening factor;
- 36 strain features derived from six ROI strain histories sampled at six displacement levels;
- 6 ROI temperature measurements; and
- 5 global thermal descriptors describing the thermal state of the sample.

Consequently, each MAPS test is represented as a single point within a 63-dimensional response space that captures force-displacement behavior, strain evolution, and temperature distribution.

A.5.2 Principal Component Analysis

Because many descriptor components are strongly correlated, Principal Component Analysis (PCA) was applied to obtain a compact representation of material behavior. The first principal component (PC1) explains approximately 47% of the total variance, while the second principal component (PC2) explains approximately 18%. Together, PC1 and PC2 account for approximately 65% of the variance, while the first five principal components explain approximately 97% of the total variance.

Analysis of the PCA loadings revealed physically interpretable latent dimensions. PC1 is dominated by force-history and strain-related features and primarily captures constitutive response magnitude and deformation behavior. PC2 is dominated by thermal descriptors and ROI temperature features and therefore primarily captures thermal state and temperature evolution. Higher-order components capture more specialized effects, including strain localization, hardening behavior, and thermo-mechanical coupling.

These results indicate that material behavior within the MAPS dataset is intrinsically low-dimensional and can be represented effectively using a small number of latent variables.

A.5.3 Material Centroid Representation

Each material is represented by the centroid of all associated tests within the five-dimensional PCA space. The centroid coordinates are computed as the mean values of PC1–PC5 across all simulated temperatures for a given material. These centroids serve as the basis for the neighborhood distance calculations described in Section 3.4.

A.5.4 Reliability Threshold Selection

The thresholds used in Equation 9 were established using the distribution of the training materials within the PCA space and subsequently held fixed throughout all leave-one-material-out (LOMO) and external validation studies.

A material was classified as Reliable when it exhibited:

- low normalized neighborhood distance,
- strong local constitutive alignment as measured by local PC1 variance, and
- high agreement in temperature-dependent evolution as measured by direction similarity.

Materials failing any of these conditions were assigned a Caution classification. Importantly, a Caution designation does not indicate prediction failure. Rather, it indicates that the corresponding prediction relies on greater extrapolation beyond the material relationships represented within the available training dataset.

A.6 Coverage-Disagreement Framework Details

A.6.1 Coverage Metric

The thermo-mechanical space was discretized by partitioning temperatures into 50°C intervals. Coverage was computed as the number of unique Sample-ROI combinations contributing at least one observation within a given temperature bucket. Higher coverage indicates stronger observational support arising from multiple independent measurements, while lower coverage reflects sparse sampling and reduced experimental support.

A.6.2 DISAGREEMENT METRIC

Coverage alone does not indicate whether observations occupying similar thermo-mechanical states produce consistent responses. To quantify local variability, temperature was discretized into 25°C intervals and strain was discretized into bins of 0.005 strain. All observations within a thermo-mechanical cell were grouped together and disagreement was computed using Equation 11. Larger disagreement values indicate reduced consistency among observations occupying similar thermo-mechanical states and therefore greater local variability.

A.6.3 INTERPRETATION OF COVERAGE-DISAGREEMENT REGIMES

The joint interpretation of coverage and disagreement partitions the thermo-mechanical space into four characteristic regimes.

- **High Coverage / Low Disagreement (Well Constrained):** This region contains a large number of independent observations that produce consistent responses. The combination of strong experimental support and low variability suggests that the underlying constitutive behavior is well characterized, and additional testing is unlikely to provide substantial new information.
- **Low Coverage / Low Disagreement (Sparse Data; Consistent Predictions):** This region contains relatively few observations, but the available measurements exhibit consistent behavior. Although additional experiments may further refine confidence in the response, the current observations suggest that the constitutive behavior is relatively stable and predictable within the sampled range.
- **High Coverage / High Disagreement (Disagreement Despite Coverage):** This region contains substantial experimental support, yet considerable variability remains between observations occupying similar thermo-mechanical states. Such behavior suggests that the observed variability cannot be explained by data sparsity alone and may instead reflect material complexity, sensitivity to local thermo-mechanical conditions, or limitations in the predictive model.
- **Low Coverage / High Disagreement (Highest Priority for New Tests):** This region combines limited experimental support with elevated variability between observations. Because both uncertainty and data sparsity are present simultaneously, they represent the most important candidates for adding new experiments. Additional measurements yielding data in these temperature buckets are expected to provide the greatest benefit for improving model reliability and reducing uncertainty.

These regime classifications provide a practical framework for prioritizing future data acquisition and identifying thermo-mechanical regions where additional characterization is expected to provide the greatest benefit.

B APPENDIX. EVALUATION PROTOCOL AND METRICS

B.1 EVALUATION PROTOCOL

Model performance was evaluated using the leave-one-material-out (LOMO) protocol described in 3.1. Each experiment was repeated using 10 independent random seeds, and reported results correspond to the mean and standard deviation across runs.

B.2 PERFORMANCE METRICS

Prediction accuracy was evaluated using the Mean Absolute Percentage Error (MAPE),

$$\mathrm{MAPE} = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{\sigma_i^{\mathrm{pred}} - \sigma_i^{\mathrm{actual}}}{\sigma_i^{\mathrm{actual}}} \right|$$

where σ_i^{pred} and $\sigma_i^{\mathrm{actual}}$ represent the predicted and ground-truth stress values, respectively.

In addition to prediction accuracy, two supplementary metrics were used to assess constitutive response quality. Shape RMSE measures agreement between predicted and experimental constitutive trajectories, providing a quantitative assessment of stress-strain curve fidelity. Monotonicity Violation quantifies the fraction of strain increments exhibiting non-physical stress reduction with increasing strain. Lower values of both metrics indicate improved constitutive fidelity and physical consistency. Together, MAPE, Shape RMSE, and Monotonicity Violation provide complementary measures of predictive accuracy, curve quality, and physical realism.

C APPENDIX. EFFECT OF CURVE-WISE BATCHING

The proposed physics-informed objectives require preservation of the relationships between neighboring ROIs within a MAPS test. As a result, complete constitutive trajectories must be processed together during training rather than treating individual rows of data as independent samples. Prior to introducing the physics-informed loss functions, a batch-size study was performed using the baseline MSE-only objective to determine an appropriate curve-wise batching strategy.

Table 3 summarizes the effect of increasing the number of constitutive trajectories processed simultaneously during each optimization step. Increasing the batch size from one to four curves improved average leave-one-material-out (LOMO) performance, reducing the mean prediction error from 9.95% to 9.28%. The largest benefit was observed for the most challenging extrapolation cases like AA3003-H14, where the mean error decreased from 37.6% to 28.9%.

In addition to improving prediction accuracy, larger curve-wise batches produced smoother training and validation loss trajectories, indicating more stable optimization behavior. The reduction in variance suggests that additional constitutive trajectories within each batch provide a more representative estimate of the loss landscape during training.

Based on these results, a batch size of four complete constitutive trajectories was selected for all subsequent experiments. This configuration provided the best balance between optimization stability and generalization performance while remaining computationally efficient.

D APPENDIX. PCA COMPONENTS AS INPUT FEATURES

The materials-space reliability framework described in 3.4 represents each MAPS test using a low-dimensional PCA embedding derived from the 63-dimensional material response vector. Because these principal components capture material-level information regarding constitutive behavior and thermal response, additional experiments were performed to investigate whether such information could improve constitutive prediction when included directly as model inputs.

Table 4 summarizes the resulting leave-one-material-out performance as PCA coordinates and material-signature features were progressively added to the baseline feature set. Contrary to expectations, prediction accuracy consistently deteriorated as additional material-level descriptors were introduced. Adding only the first principal component increased the mean MAPE from 6.5% to 8.5%, while incorporating all five principal components increased the error to 24.8%. Similar degradation was observed when material-signature features derived from the mean and standard deviations of PC1-PC5 components across all tests belonging to a material were included.

The observed behavior suggests that the network learned material-identifying shortcuts that improved fit to the training materials while reducing its ability to generalize to unseen materials. As more explicit material-level descriptors were introduced, the model increasingly relied on these identifiers instead of learning transferable constitutive relationships between force, strain, temperature, and stress response.

These results indicate that PCA-derived descriptors are valuable for post-hoc characterization of material similarity and reliability assessment but are detrimental when used directly as predictive model inputs. Consequently, PCA coordinates were used exclusively within the materials-space reliability framework and were deliberately excluded from the constitutive prediction model.

Table 3: Effect of batch composition on model performance. Increasing the number of constitutive curves processed simultaneously results in smoother optimization dynamics and modestly improved generalization performance.

Curves per Batch	LOMO Mean MAPE (%)	LOMO Std MAPE (%)	AA3003-H14 MAPE (%)	AA3003-H14 Std MAPE (%)
1	9.95	1.54	37.6	6.69
2	9.54	1.44	35.5	7.03
4	9.28	1.54	28.9	5.71

Table 4: Effect of adding PCA-derived features to the input representation.

Input Features	# Features	Mean MAPE (%)
Baseline Features	25	6.5
+ PC1	26	8.5
+ PC1 + PC2	27	11.6
+ PC1–PC5	30	24.8
Material Signature	35	24.5
PC1–PC5 + Material Signature	40	21.8

E APPENDIX. ADDITIONAL PHYSICS-INFORMED LOSS ABLATIONS

Several physics-informed loss formulations were investigated during model development to improve constitutive response quality while maintaining strong generalization performance. Table 5 summarizes the impact of the final physics-informed objectives relative to the MSE-only baseline. While the baseline achieved the lowest prediction error (6.43% MAPE), the full model substantially improved constitutive fidelity and physical consistency, reducing Shape RMSE from 1.36 to 0.97 and Monotonicity Violation from 0.20 to 0.12 while maintaining comparable predictive accuracy.

To understand the contribution of individual loss components and training strategies, a broader ablation study was performed. Table 6 summarizes representative configurations evaluated using leave-one-material-out (LOMO) experiments.

When combined with the curve-wise batching strategy described in Appendix C, the shape loss produced substantially more stable training behavior, reducing the average standard deviation of MAPE from 1.34% to 0.96% while preserving the intended shape-aware constraints. This result demonstrates that the effectiveness of shape-aware objectives depends strongly on maintaining the physical relationships between neighboring points within a constitutive trajectory.

Additional experiments were performed using a cubic monotonicity penalty designed to discourage non-physical stress reduction with increasing strain. Although physically motivated, the monotonicity loss consistently increased prediction error and training variability. Similar behavior was observed when the monotonicity loss was combined with the per-temperature mean stress objective, suggesting that explicit monotonicity enforcement provided limited benefit beyond the constraints already imposed by the shape-aware formulation.

Overall, the best balance of prediction accuracy, constitutive fidelity, and training stability was achieved by combining the shape loss with the per-temperature mean stress loss. These results highlight that not all physically motivated constraints improve constitutive learning performance; increasingly restrictive loss formulations often degraded generalization performance without providing additional practical benefit.

Table 5: Effect of physics-informed losses on constitutive response prediction.

Model	Mean MAPE (%)	Std MAPE (%)	Shape RMSE	Mono. Violation
MSE Only	**6.43**	1.18	1.36	0.20
MSE + Shape Loss	6.91	0.96	1.21	0.17
Full Model	6.90	**0.86**	**0.97**	**0.12**

Table 6: Effect of physics-based loss functions on model performance. Results show the mean MAPE and mean standard deviation over a 10-seed leave-one-material-out (LOMO) study.

Loss Configuration	MAPE (%)	Std. (%)	Key Observation
MSE Only (Batch 32)	6.4	1.20	Best MAPE; no physics constraints.
Shape Loss (Batch 1)	7.2	1.34	Higher error and training variability.
Shape Loss (Batch 4)	6.9	0.96	Improved stability through larger batch composition.
Shape + Cubic Monotonicity (Batch 4)	7.3	1.31	Monotonicity penalty increased error.
Shape + Per-Temperature Mean Stress (Batch 4)	6.8	0.84	Best overall balance; selected final model.
Shape + Cubic Monotonicity + Per-Temperature Mean Stress	7.2	1.09	Additional regularization provided no benefit.