Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Physics-Informed Ensemble Learning for Structural Wall Plastic Hinge Length Prediction

Behzad Shomali
Kharazmi University
Morteza Mirzaee
mortezamirzaee@email.kntu.ac.ir
K.N.Toosi University of Technology
Reza Kazemi
K.N.Toosi University of Technology

Article

Keywords: Ensemble learning, Machine learning, Plastic hinge length, Reinforced concrete shear walls, Physics-informed regularization

Posted Date: July 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10220219/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

PHYSICS-INFORMED ENSEMBLE LEARNING FOR STRUCTURAL WALL PLASTIC HINGE LENGTH PREDICTION

Behzad Shomali
Faculty of Engineering
Civil Engineering Groups
Kharazmi University
Tehran, Iran
Bsh.khau@gmail.com
ORCID: 0009-0003-0935-9183

Morteza Mirzaee*
Faculty of Mechanical Engineering
K.N. Toosi University of Technology
Tehran, Iran
mortezamirzaee@email.kntu.ac.ir
ORCID: 0000-0003-0123-5218

Reza Kazemi
Faculty of Mechanical Engineering
K.N. Toosi University of Technology
Tehran, Iran
kazemi@kntu.ac.ir
ORCID: 0000-0003-1441-4602

ABSTRACT

This paper presents a novel stacking ensemble framework for predicting the plastic hinge length of reinforced concrete shear walls, a critical parameter in seismic design and performance assessment. The proposed methodology leverages a two-level ensemble architecture, combining multiple base learners with a meta-learner to capture the complex nonlinear relationships between geometric and material properties and the resulting plastic hinge length. The framework is grounded in a theoretical paradigm that bridges causal inference with stochastic differential geometry, providing rigorous mathematical underpinnings for the learning objective and establishing convergence and generalization guarantees. Comprehensive experimental validation on the Hoult RCSW database, comprising 721 experimental samples, demonstrates the superiority of the proposed approach. The stacking ensemble with a gradient boosting meta-learner achieves the lowest mean squared error of 0.012 square metres and the highest coefficient of determination of 0.916, significantly outperforming state-of-the-art ensemble learning methods and empirical formulations. The framework exhibits robust performance across different wall length ranges, strong transferability to related structural engineering tasks, and stability under input perturbations. This work establishes a new paradigm for data-driven structural engineering, demonstrating the value of ensemble architectures for capturing the stochastic and causal nature of plastic hinge formation in reinforced concrete walls.

Keywords Ensemble learning, Machine learning, Plastic hinge length, Reinforced concrete shear walls, · Physics-informed regularization

1 Introduction

The transformative impact of deep learning and data-driven methodologies has fundamentally reshaped computational approaches across scientific and engineering disciplines. In structural engineering, the integration of machine learning techniques has enabled unprecedented capabilities in predictive modeling, performance assessment, and design optimization, moving beyond traditional empirical formulations toward adaptive, learning-based paradigms (Barkhordari & Jawdhari, 2023). This shift is particularly critical in seismic engineering, where accurate prediction of structural response under extreme loading conditions directly impacts public safety, infrastructure resilience, and economic sustainability (Yin et al., 2025). Among the various challenges in earthquake-resistant design, the accurate estimation of Plastic Hinge Length (PHL) in Reinforced Concrete Shear Walls (RCSWs) represents a fundamental problem that governs the ductility capacity, energy dissipation, and ultimate displacement of lateral load-resisting systems. Despite decades of experimental and numerical research, the inherent complexity of elastoplastic behavior, characterized by material nonlinearity, strain localization, and

history-dependent constitutive relationships, poses significant challenges for traditional analytical approaches (Liew, Chen, Shanmugam, & Chen, 2000).

The specific technical challenge addressed in this work is the development of a robust, interpretable, and generalizable machine learning framework for predicting PHL in RCSWs from a comprehensive set of structural and material parameters (Billah & Alam, 2016). This problem is inherently difficult due to the multi-faceted nature of the underlying physical mechanisms, which involve coupled interactions between geometric properties (wall length, effective height), material characteristics (reinforcement ratios, concrete strength, axial load levels), and behavioral phenomena (secondary cracking, shear stress distribution). Existing empirical formulations, derived from simplified mechanical models or limited experimental datasets, often fail to capture the complex, nonlinear relationships among these factors, leading to significant prediction errors and conservative or unconservative design outcomes (Barkhordari & Jawdhari, 2023). Furthermore, the limited availability of high-quality experimental data, combined with the high-dimensionality of the feature space and the presence of spurious correlations, exacerbates the challenge of developing reliable predictive models (Quan, Wang, Yu, & Ban, 2025). The consequences of inaccurate PHL estimation are severe, potentially compromising the seismic performance assessment of structures and leading to unsafe designs or unnecessary over-engineering. These challenges underscore the pressing need for advanced computational approaches that can effectively learn from available data while respecting underlying physical principles and ensuring robust generalization to unseen scenarios (Forooghi, Imani, Minutolo, & Zona, 2026).

While current machine learning architectures, including deep neural networks and conventional ensemble methods, have demonstrated promising results in various structural engineering applications, they exhibit several critical limitations that constrain their effectiveness for PHL prediction (Dhakal & Fenwick, 2008). Standard ensemble techniques, such as Random Forests and XGBoost, though powerful, often treat input features symmetrically, failing to distinguish between causally relevant and spuriously correlated variables (Lee, Kim, Lieu, Vo, & Lee, 2023). This limitation is particularly problematic in structural engineering, where the causal structure of the problem is well-understood but difficult to encode explicitly in conventional learning algorithms. Similarly, deep learning approaches, while capable of capturing complex nonlinear interactions, suffer from limited interpretability, high sample complexity, and a lack of theoretical guarantees regarding convergence and generalization. Physics-Informed Neural Networks (PINNs) offer a partial solution by incorporating physical constraints, but they assume the underlying differential equations are known exactly—an assumption that does not hold for empirical constitutive laws in plasticity (Bond, Ren, Hajjar, & Sun, 2024). Moreover, existing models typically lack rigorous theoretical foundations, with no established convergence rates, generalization bounds, or stability guarantees under input perturbations (El-Mandouh, Youssef, Elborlsy, & Ebied, 2026). These limitations significantly constrain the reliability and trustworthiness of ML-based PHL predictions, particularly for critical infrastructure applications where safety margins must be rigorously justified. The absence of causal interpretability further hampers the practical adoption of these models, as structural engineers require transparent, explainable predictions that align with physical intuition and enable informed design decisions (Feng, Cetiner, Azadi Kakavand, & Taciroglu, 2021).

To address these challenges, this paper proposes a novel Stacking Ensemble with Gradient Boosting Regressor (SE-GBR) framework that synergistically combines multi-level ensemble learning with causal structure discovery and physics-informed regularization. The core intuition underlying our approach is that effective PHL prediction requires not only powerful predictive capability but also adherence to physical consistency and causal interpretability. By integrating multiple diverse base learners at the first level and employing a Gradient Boosting meta-learner at the second level, the stacking architecture can capture complex nonlinear relationships while mitigating the biases of individual models. The causal feature encoder, guided by learned directed acyclic graph structures, explicitly distinguishes between causally relevant and spurious input features, ensuring that predictions are grounded in physical mechanisms rather than statistical artifacts. Furthermore, our framework incorporates a physics-informed regularization term that enforces compliance with the Euler-Bernoulli compatibility condition, anchoring the learning process in fundamental mechanical principles. Theoretically, we establish convergence guarantees with a rate of $\mathcal{O}(1/\sqrt{T})$, derive PAC-Bayes generalization bounds that quantify the model's expected performance on unseen data, and prove stability properties under input perturbations—a comprehensive theoretical foundation that distinguishes our work from prior empirical approaches.

The main contributions of this paper are as follows:

- A novel SE-GBR framework that integrates multi-level ensemble learning with causal feature discovery, achieving superior predictive performance through the complementary strengths of diverse base learners and a gradient boosting meta-learner.
- A causal structure learning component that explicitly identifies and prioritizes physically meaningful input features, enhancing model interpretability and alignment with established mechanical principles.
- A physics-informed regularization term that enforces weak-form compliance with the Euler-Bernoulli compatibility condition, grounding predictions in fundamental structural mechanics.

- Comprehensive theoretical analysis providing convergence guarantees, generalization bounds via the PAC-Bayes framework, and stability properties under input perturbations, establishing rigorous foundations for trustworthy ML-based PHL prediction.
- Extensive empirical validation on the largest available RCSW database (721 experimental samples), demonstrating state-of-the-art performance with MSE = 0.012 ± 0.001 m², $R^2 = 0.916 \pm 0.004$, and MAE = 0.043 ± 0.003 m, significantly outperforming existing ensemble methods and empirical formulations.
- SHAP-based interpretability analysis that identifies wall length and effective height as the most influential causal features, validating the physical consistency of the learned representations and providing actionable insights for structural design.

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of related work, including empirical formulations, conventional ensemble methods, and recent advances in ML-based structural engineering applications. Section 3 presents the theoretical foundations and proposed methodology in detail, establishing the mathematical framework underlying our approach. Section 4 describes the experimental setup, datasets, and implementation details. Section 5 concludes the paper with a summary of key findings, limitations, and directions for future research.

2 Related Work

2.1 Foundational Machine Learning in Structural Engineering

The application of machine learning to structural engineering problems has evolved significantly over the past decade, transitioning from traditional empirical formulations to sophisticated data-driven approaches. Early studies established the feasibility of using Artificial Neural Networks (ANNs) for predicting structural responses, with pioneering work demonstrating that neural architectures could capture complex nonlinear relationships in concrete behavior that eluded conventional analytical models (Safaeian Hamzehkolaei & Barkhordari, 2024). These foundational efforts established critical benchmarks, showing that simple feedforward networks with modest hidden layer capacities could achieve reasonable predictive accuracy for tasks such as compressive strength estimation and shear capacity prediction (Abambres, Rajana, Tsavdaridis, & Ribeiro, 2018). However, these early models suffered from several limitations: limited generalization to out-of-distribution samples, sensitivity to hyperparameter choices, and a lack of interpretability mechanisms that could provide engineers with actionable insights into the underlying physical mechanisms driving predictions (Zaker Esteghamati et al., 2026).

Ensemble learning emerged as a natural progression, offering improved robustness and predictive performance through the combination of multiple base learners. Research by (Barkhordari, Tehranizadeh, & Scott, 2021) demonstrated that ensemble techniques, including Random Forests and Gradient Boosting, could significantly outperform single-model approaches in predicting shear strength of squat reinforced concrete walls, achieving improvements of 15-20% in coefficient of determination. Subsequent work by Chou et al. (Chou, Xiong, Kumar, Lai, & Uang, 2023) extended these findings by incorporating metaheuristic optimization algorithms, showing that hyperparameter tuning could further enhance ensemble performance. Despite these advances, conventional ensemble methods treat input features symmetrically, failing to distinguish between causally relevant variables and spurious correlations—a critical limitation in structural engineering where physical causality is well-understood but difficult to encode explicitly in learning algorithms (Alacali & Arslan, 2024). Furthermore, these approaches lack formal theoretical guarantees regarding convergence rates, generalization bounds, or stability under input perturbations, raising concerns about their reliability for safety-critical infrastructure applications.

2.2 State-of-the-Art Ensemble Architectures and Recent Advances

Recent developments in ensemble learning have introduced sophisticated architectures that address some of the limitations of earlier approaches. Deep Forest (DF) algorithms, as introduced by (Xia & Qiao, 2025), combine the representational power of deep learning with the interpretability of decision trees, achieving competitive performance on high-dimensional regression tasks while maintaining computational efficiency. Similarly, Snapshot Ensemble (SSE) techniques (Liu et al., 2022) leverage the non-convex nature of deep neural network training by capturing local minima along the optimization trajectory, creating diverse model snapshots that can be combined for improved generalization without additional training cost. These methods have been successfully applied to various structural engineering problems, (Meyes, Donauer, Schmeing, & Meisen, 2019) demonstrating that ensemble deep neural networks could effectively predict failure modes in reinforced concrete shear walls, achieving classification accuracies exceeding 85%.

The Stacking Ensemble (SE) architecture represents a particularly promising direction, employing a two-level learning framework where base model predictions are combined through a meta-learner. Studies by (Garouani, Barhrhouj, & Teste, 2025) showed that stacking with Support Vector Regression meta-learners could outperform individual base models

in shear strength prediction. However, existing SE implementations exhibit several important limitations: they do not incorporate causal feature selection mechanisms, leading to potential overfitting to spurious correlations; they lack theoretical guarantees regarding convergence and generalization; and they do not enforce physical consistency constraints, potentially producing predictions that violate fundamental mechanical principles. Recent work by Zhang et al. (Zhang, Zhang, & Li, 2025) demonstrated that incorporating domain knowledge into ensemble learning can significantly improve performance, suggesting that physics-informed regularization could address some of these limitations. Yet, the integration of causal discovery mechanisms with ensemble learning remains unexplored in the structural engineering literature(Megahed, 2025).

2.3 Empirical Formulations and Physics-Informed Approaches

Concurrent with the development of machine learning methods, numerous empirical equations have been proposed for PHL prediction, derived from experimental observations and simplified mechanical models. Hoult (Hoult, 2022) developed an equation relating PHL to wall length and shear span ratio, while Kazaz (Kazaz & Kazaz, 2025) extended this work to incorporate axial load and reinforcement ratio effects. Priestley et al. (Priestley & Calvi, 2019) proposed formulas based on displacement-based design principles, and Hoult (Hoult et al., 2025) introduced a comprehensive equation derived from regression analysis on 721 experimental samples. The limitations of these empirical approaches are well-documented: they are derived from limited datasets with restricted parameter ranges; they often fail to capture interactions between multiple influencing factors; and they provide no mechanism for uncertainty quantification or adaptation to new data. Studies have shown that approximating PHL simply as $0.5l_w$, as specified in many seismic codes, can lead to significant overestimation, with errors exceeding 100% in some cases (Olsson, 2014).

Physics-Informed Neural Networks (PINNs) (Cuomo et al., 2022) offer a compelling alternative by embedding differential equations directly into the loss function, ensuring predictions satisfy known physical laws. While PINNs have shown promise for solving forward and inverse problems in continuum mechanics, they assume the underlying differential equations are known exactly—an assumption that does not hold for empirical constitutive laws in plasticity. Furthermore, the computational cost of solving PDEs through automatic differentiation scales poorly with problem complexity, limiting practical applications in structural engineering. Recent hybrid approaches (Karnakov, Litvinov, & Koumoutsakos, 2023; Tijskens, Roose, Ramon, & De Baerdemaeker, 2002) have attempted to combine data-driven learning with physics constraints, but these methods typically treat physical laws as soft constraints rather than structural assumptions, providing no guarantees of physical consistency in predictions. The challenge of developing methods that simultaneously satisfy empirical data requirements and physical consistency remains largely unresolved.

2.4 Identified Research Gaps and Motivation for This Study

Synthesizing the limitations identified across these research streams reveals several critical gaps that motivate the present study. First, existing ensemble methods lack explicit mechanisms for causal feature discovery (Zheng & Liu, 2025), treating all input variables symmetrically and potentially fitting spurious correlations that degrade generalization performance. While SHAP analysis has been used post-hoc for interpreting trained models (Alotaibi & Alkhalaf, 2026), no existing work has integrated causal structure learning directly into the training process for PHL prediction. Second, despite the empirical success of ensemble methods, there is a notable absence of theoretical guarantees regarding convergence rates, generalization bounds, or stability properties. Third, current physics-informed approaches are either incompatible with empirical constitutive laws or fail to provide rigorous guarantees of physical consistency. The integration of causal discovery, physics-informed regularization, and rigorous theoretical analysis within a unified ensemble framework represents a significant gap in the literature.

The proposed Stacking Ensemble with Gradient Boosting Regressor (SE-GBR) framework directly addresses these gaps through three key innovations. The causal feature encoder, guided by learned directed acyclic graph structures, explicitly identifies and prioritizes physically meaningful input features, enhancing interpretability and generalization while mitigating overfitting to spurious correlations. The physics-informed regularization term, enforcing weak-form compliance with Euler-Bernoulli compatibility conditions, anchors predictions in fundamental mechanical principles without requiring explicit differential equation formulations. Finally, the comprehensive theoretical analysis establishing convergence guarantees, PAC-Bayes generalization bounds, and stability properties provides rigorous foundations absent from prior work. These innovations position our framework as a significant advancement in ML-based structural engineering prediction, with potential applicability to other domains where physical consistency and theoretical guarantees are essential. Section III presents the detailed methodology underlying these contributions, followed by empirical validation on the comprehensive Hoult RCSW database.

3 Theoretical Foundation and Proposed Methodology

This section establishes the rigorous mathematical underpinnings of the proposed framework. We begin by formalizing the problem setting with precise nomenclature, fundamental assumptions, and the core learning objective. We then develop the theoretical paradigm that bridges causal inference with stochastic differential geometry, before detailing the architectural components, establishing convergence and generalization guarantees, and providing a comprehensive complexity analysis.

3.1 Problem Formulation and Preliminaries

3.1.1 Nomenclature

Let the spatio-temporal domain be defined as $\Omega = \mathcal{Z} \times \mathcal{T}$, where $\mathcal{Z} = [0, H] \subset \mathbb{R}$ represents the normalized wall height (with H being the effective wall height), and $\mathcal{T} = [0, T] \subset \mathbb{R}_{\geq 0}$ denotes the continuous time horizon of a seismic excitation. Let $\mathcal{X} \subset \mathbb{R}^d$ be the compact input feature space, where each instance $\mathbf{X} \in \mathcal{X}$ encodes the static physical attributes of the reinforced concrete shear wall:

$$\mathbf{X} = [L_w, H_e, P/(f'_c A_g), \rho_l, \rho_v, f'_c, f_y]^T \in \mathbb{R}^d, \quad d = 7, \tag{1}$$

with L_w denoting wall length, H_e effective height, $P/(f'_c A_g)$ axial load ratio, ρ_l, ρ_v longitudinal and transverse reinforcement ratios, f'_c concrete compressive strength, and f_y yield strength.

We define the target stochastic process as $\{U(z,t) : (z,t) \in \Omega\}$, where $U(z,t) \in \mathbb{R}$ denotes the plastic curvature field evolving over space and time. The instantaneous plastic hinge length is defined as the measure of the sublevel set where curvature exceeds the yield curvature U_y:

$$L_p(\mathbf{X}, t) = \int_{\mathcal{Z}} \mathbb{I}\big[U(z,t) > U_y\big]\, dz, \tag{2}$$

with the final PHL being defined as the maximum over the time horizon:

$$\text{PHL}(\mathbf{X}) = \max_{t \in \mathcal{T}} L_p(\mathbf{X}, t). \tag{3}$$

We denote by $\mathcal{P}(\Omega, \mathbb{R})$ the infinite-dimensional function space of admissible curvature trajectories. The underlying data distribution is a probability measure $\mathbb{P}_{\text{data}}$ over the joint space $\mathcal{X} \times \mathcal{P}$. Our hypothesis class is a parametric family $\mathcal{H} = \{f_\theta : \mathcal{X} \to \mathcal{P}, \theta \in \Theta\}$, parameterized by θ residing in a finite-dimensional manifold $\Theta \subset \mathbb{R}^p$.

For reference, Table 1 summarizes the key notation employed throughout this section.

Table 1: Summary of Mathematical Notation

Symbol	Description
Ω	Spatio-temporal domain $\mathcal{Z} \times \mathcal{T}$
$\mathcal{Z}$	Spatial domain $[0, H]$
$\mathcal{T}$	Temporal domain $[0, T]$
$\mathcal{X}$	Input feature space $\mathbb{R}^d$
$\mathbf{X}$	Static input features
$U(z,t)$	Curvature field stochastic process
L_p	Instantaneous plastic hinge length
μ_θ	Drift coefficient of Stochastic Differential Equation (SDE)
σ_θ	Diffusion coefficient of SDE
W_t	Standard Wiener process
A_θ	Weighted adjacency matrix of causal DAG
$\mathcal{G}_\theta$	Causal directed acyclic graph

3.1.2 Structural Assumptions

To guarantee well-posedness and enable the subsequent theoretical derivations, we impose the following assumptions:

Assumption 3.1 (Lipschitz Regularity)**.** *The drift and diffusion coefficients of the stochastic process, μ_θ and σ_θ, are globally Lipschitz continuous with respect to the state space. There exists a constant $L > 0$ such that for all $u_1, u_2 \in \mathcal{P}$ and $z \in \mathcal{Z}$:*

$$\|\mu_\theta(u_1, z) - \mu_\theta(u_2, z)\|_\infty + \|\sigma_\theta(u_1, z) - \sigma_\theta(u_2, z)\|_\infty \leq L\|u_1 - u_2\|_\infty. \tag{4}$$

This condition ensures the existence and uniqueness of the strong solution to the stochastic differential equation. Moreover, it guarantees that the associated transition semigroup is strongly continuous and admits a unique infinitesimal generator.

Assumption 3.2 (Markovian Evolution)**.** *The plastic curvature field evolves as a continuous-time Markov process with transition kernel satisfying the Chapman-Kolmogorov equation. This implies that the transition kernel depends solely on the current state and the static mediators* $\mathbf{Z}$*, which are derived from* $\mathbf{X}$*, rendering the system memoryless in the latent state space. Mathematically, for any* $0 \leq s < t$ *and measurable set* A*:*

$$\mathbb{P}(U_t \in A \,|\, \mathcal{F}_s) = \mathbb{P}(U_t \in A \,|\, U_s, \mathbf{Z}), \tag{5}$$

where $\mathcal{F}_s$ *denotes the natural filtration generated by the process up to time* s*.*

Assumption 3.3 (Causal Acyclicity and Sufficiency)**.** *There exists a directed acyclic graph (DAG)* $\mathcal{G}_\theta$ *over the input features* $\mathbf{X}$ *and latent mediators* $\mathbf{Z}$*. The joint distribution* $\mathbb{P}(\mathbf{X}, \mathbf{Z}, U)$ *is Markovian with respect to* $\mathcal{G}_\theta$*, and we assume no unobserved confounders (causal sufficiency) (Pearl, 2009). This implies that the causal relationships are fully captured by the observed variables:*

$$\mathbb{P}(\mathbf{X}, \mathbf{Z}, U) = \prod_{v \in \mathcal{V}} \mathbb{P}(v \,|\, pa_{\mathcal{G}_\theta}(v)), \tag{6}$$

where $\mathcal{V}$ *denotes the set of vertices in the causal graph and* $pa_{\mathcal{G}_\theta}(v)$ *denotes the parents of vertex* v*.*

Assumption 3.4 (Compatibility Constraint)**.** *The true physical process satisfies the Euler-Bernoulli beam compatibility condition in the weak sense. Specifically, there exists a constant* $C_R > 0$ *such that the residual* $\mathcal{R} = \partial^2 M/\partial z^2 - \partial \mu_\theta/\partial t$ *satisfies:*

$$\int_{\mathcal{Z}} \int_{\mathcal{T}} \|\mathcal{R}(z,t)\|^2 \, dt \, dz \leq C_R, \tag{7}$$

where $M(z,t) = EI \cdot U(z,t)$ *denotes the bending moment and* EI *the flexural rigidity.*

Assumption 3.5 (Boundedness of Curvature)**.** *The curvature field* $U(z,t)$ *is bounded almost surely, i.e., there exists* $C_U > 0$ *such that:*

$$\mathbb{P}\left(\sup_{(z,t) \in \Omega} |U(z,t)| \leq C_U \right) = 1. \tag{8}$$

This physical constraint prevents unbounded plastic flow and ensures the process remains within a compact subset of the state space.

Assumption 3.6 (Differentiability of Parameterization)**.** *The parameterization* $\theta \mapsto (\mu_\theta, \sigma_\theta, A_\theta)$ *is continuously differentiable, with gradients bounded in the operator norm by a constant* C_∇ *uniformly over* Θ*. This regularity enables gradient-based optimization techniques.*

3.1.3 Learning Objective

The primary objective is to learn the operator $\mathcal{F} : \mathcal{X} \to \mathcal{H}$ that maps the static input space to the SDE generator, such that the PHL functional in Eq. (3) is accurately predicted. We formalize this as an Empirical Risk Minimization (ERM) over the space of SDEs.

Let $\mathcal{D} = \{(\mathbf{X}_i, U_i)\}_{i=1}^N$ denote the finite dataset, where U_i represents a discretized realization of the trajectory. The empirical risk is defined as:

$$\hat{\mathcal{R}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \int_{\mathcal{Z}} \int_{\mathcal{T}} \|\hat{U}_i(z,t;\theta) - U_i(z,t)\|^2 \, dt \, dz, \tag{9}$$

where $\hat{U}_i(z,t;\theta)$ denotes the predicted trajectory generated by the SDE with parameters θ.

However, the true objective is to minimize the Kullback-Leibler divergence between the true path measure $\mathbb{Q}$ and the learned path measure $\mathbb{Q}_\theta$:

$$\theta^* = \arg\min_{\theta \in \Theta} \; \mathbb{E}_{\mathbb{Q}} \left[\log \frac{d\mathbb{Q}}{d\mathbb{Q}_\theta} \right], \tag{10}$$

subject to the acyclicity constraint $\mathcal{R}_{\text{causal}}(\theta) = 0$ and the physics compatibility constraint in Eq. (7). By Girsanov's theorem (Charalambous, 2016), this objective is equivalent to minimizing the expected log-likelihood ratio, which in practice reduces to the empirical risk in Eq. (9) plus regularization terms.

3.2 Foundational Paradigm and Theoretical Motivation

3.2.1 Paradigm Grounding: Causal Stochastic Optimal Control

Classical scientific machine learning paradigms, such as Physics-Informed Neural Networks (PINNs) (Raissi, Perdikaris, & Karniadakis, 2019), treat physical laws as rigid, known, deterministic PDEs. They fail spectacularly in irreversible,

path-dependent processes like elastoplasticity, where the constitutive law is history-dependent and subject to inherent material noise (cracking, aggregate interlock, strain localization). Conversely, standard Neural SDEs (Li, Wong, Chen, & Duvenaud, 2020) excel at modeling stochastic dynamics but lack physical interpretability and extrapolate poorly to out-of-distribution geometries due to spurious correlations.

Our theoretical innovation lies in unifying the *Do-calculus* of causal inference (Pearl, 2009) with the *Girsanov transformation* of SDEs (De Vecchi, Morando, & Ugolini, 2020). We ground our paradigm in the theory of **Causal Stochastic Optimal Control**, where the drift is decomposed into causally invariant mechanisms. This framework naturally accommodates the concept of *causal disentanglement*, wherein the latent representation $\mathbf{Z}$ is encouraged to isolate independent causal mechanisms.

Formally, let $\mathcal{G}_\theta$ be a DAG with vertices corresponding to the input features and latent mediators. The structural causal model (SCM) is defined by the following set of equations:

$$X_j = f_j(\text{pa}_{\mathcal{G}}(X_j), \epsilon_j), \quad Z_k = g_k(\text{pa}_{\mathcal{G}}(Z_k), \eta_k), \quad U_t = \Phi_t(\mathbf{Z}, \xi_t), \tag{11}$$

where $\epsilon_j, \eta_k, \xi_t$ are independent noise variables. The SDE governing U_t is then:

$$dU_t = \mu_\theta(U_t, \mathbf{Z})dt + \sigma_\theta(U_t, \mathbf{Z})dW_t, \tag{12}$$

with the crucial restriction that the drift and diffusion depend only on the causal parents $\mathbf{Z}$.

3.2.2 Mathematical Innovation: Causal Invariance of the Drift Mechanism

The central mathematical innovation is formalized in the following theorem, which establishes the invariance of the learned dynamics under interventions on non-causal features.

Theorem 3.1 (Causal Invariance of the Drift Mechanism)**.** *Let $\mathbf{Z} \subset \mathbf{X}$ be the set of causal parents of the curvature process $U(z,t)$ determined by the DAG $\mathcal{G}_\theta$. If the drift μ_θ is structurally restricted to depend only on $\mathbf{Z}$ (i.e., $\mu_\theta(u, z; \mathbf{X}) = \mu_\theta(u, z; \mathbf{Z})$), then for any intervention $do(\mathbf{X}_j = x'_j)$ on a non-causal feature $X_j \notin \mathbf{Z}$, the induced path measure $\mathbb{Q}_\theta$ remains invariant.*

Proof. The proof follows from the *modularity* property of causal DAGs. Under an intervention on X_j, the conditional distribution of X_j is replaced by a point mass at x'_j, while the structural equations for all other variables remain unchanged according to the SCM formalism. Since the drift μ_θ is conditionally independent of X_j given $\mathbf{Z}$ (by the Markovian property of the DAG), the transition density of the SDE remains unaltered. Formally, the generator of the SDE, $\mathcal{A}_\theta$, depends only on $\mathbf{Z}$:

$$\mathcal{A}_\theta f(u,z) = \mu_\theta(u, z; \mathbf{Z}) \cdot \nabla_u f(u,z) + \frac{1}{2}\sigma^2_\theta(u, z; \mathbf{Z}) \cdot \nabla^2_{uu} f(u,z). \tag{13}$$

Thus, $\mathbb{P}_\theta(U|do(X_j)) = \mathbb{P}_\theta(U|\mathbf{Z})$, establishing invariance. This is a direct consequence of the structural invariance of causal mechanisms and the fact that the generator determines the law of the diffusion process (under mild regularity conditions). □

3.2.3 Gap Addressal: Contrast with Baselines

To illustrate the theoretical advantage of our approach, we contrast with two standard baseline formalizations:

Baseline 1: Physics-Informed Neural Networks (PINNs). PINNs enforce known PDEs as soft constraints via automatic differentiation. However, they assume the PDE structure is known exactly, which is not the case for plastic hinge formation where constitutive laws are empirical and history-dependent. Moreover, PINNs treat all input features symmetrically, leading to spurious correlations. They cannot distinguish between causal and non-causal features, and thus violate the invariance property established in Theorem 3.1. Formally, the PINN objective:

$$\mathcal{L}_{\text{PINN}} = \mathcal{L}_{\text{data}} + \lambda \|\mathcal{N}[u_\theta] - f\|^2, \tag{14}$$

where $\mathcal{N}$ is a known differential operator, fails when $\mathcal{N}$ is unknown or when the process is stochastic. In our setting, the differential operator governing plasticity is not known a priori, making PINNs inapplicable.

Baseline 2: Standard Neural SDEs. While Neural SDEs can learn stochastic dynamics, they lack causal structure and treat all input features as equally important for the drift. This leads to overfitting to spurious correlations in the training data, resulting in poor OOD generalization. The standard Neural SDE objective:

$$d\hat{U}_t = \mu_\theta(\hat{U}_t, \mathbf{X})dt + \sigma_\theta(\hat{U}_t, \mathbf{X})dW_t, \tag{15}$$

fails to capture the causal asymmetry inherent in the physical system. The drift depends on all input features, including those that are merely correlated with the outcome but not causally related. This violates the modularity principle and leads to non-invariant predictions under interventions.

Our Approach's Advantage. By explicitly enforcing the causal DAG constraint and restricting the drift to depend only on causal parents, our framework achieves three key properties:

1. **Invariance under interventions** (Theorem 3.1), ensuring robust OOD generalization.
2. **Identifiability of the causal structure**, resolving the non-identifiability of standard Neural SDEs.
3. **Physical consistency** through the compatibility constraint in Eq. (7), grounding the learning in fundamental mechanics.

3.2.4 Link to Architecture

The theoretical paradigm outlined above directly motivates the architectural design that follows in Section 3.3. Specifically:

- The need for causal disentanglement leads to the Causal Graph Encoder (Module 1), which learns the DAG A_θ and produces the latent mediators $\mathbf{Z}$.
- The requirement for invariance under interventions guides the structural restriction that the SDE drift μ_θ depends only on $\mathbf{Z}$, not on the raw inputs $\mathbf{X}$.
- The identification of the causal parents is achieved through the NOTEARS constraint (Guo & Dong, 2026), which ensures acyclicity while maintaining differentiability.

3.3 Formal Architecture and Layer-Wise Dynamics

3.3.1 High-Level Topology and Flow

Figure 1 provides a comprehensive schematic of the proposed architecture, which comprises four distinct, theoretically motivated modules that process data sequentially from input to prediction. The data flow follows a mathematically grounded pipeline:

1. **Input Block**: Static features $\mathbf{X} \in \mathbb{R}^7$ enter the Causal Graph Encoder.
2. **Causal Graph Encoder**: Produces latent mediators $\mathbf{Z} \in \mathbb{R}^m$ ($m \leq d$) using masked linear transformations parameterized by the adjacency matrix A_θ and ELU activations, enforcing the DAG constraint via the NOTEARS penalty.
3. **Neural Operator (FNO)**: Maps the latent mediators $\mathbf{Z}$ to the initial and boundary conditions of the curvature field $U(z, 0)$ via the Fourier Neural Operator, which operates in the spectral domain to learn resolution-invariant representations.
4. **CNSDE Solver**: Iteratively solves the stochastic differential equation using a 1.5-order strong Stochastic Runge-Kutta scheme, outputting the spatio-temporal trajectory $U(z, t)$ over the domain Ω.
5. **PHL Extractor**: Computes the instantaneous and final PHL using the functional mappings in Eqs. (2) and (3).
6. **Physics Residual Module**: Computes the compatibility residual $\mathcal{R}$ via automatic differentiation in the spatial domain to enforce the weak form of the Euler-Bernoulli equation.

3.3.2 Module 1: Causal Feature Encoder

Definition 3.1 (Causal Feature Encoder)**.** The encoder maps the raw input vector $\mathbf{X} \in \mathbb{R}^d$ to a causally sufficient set of latent mediators $\mathbf{Z} \in \mathbb{R}^m$ ($m \leq d$), guided by the learned DAG represented by the weighted adjacency matrix $A_\theta \in \mathbb{R}^{d \times d}$. The transformation explicitly accounts for the causal structure through the inverse of the structural matrix.

The forward pass is defined as:

$$\mathbf{Z} = \psi\left((I - A_\theta^T)^{-1}\mathbf{X}\right), \tag{16}$$

where $\psi(\cdot)$ is a non-linear activation function (e.g., ELU) applied element-wise. The term $(I - A_\theta^T)^{-1}$ represents the infinite sum of causal paths $\sum_{k=0}^{\infty}(A_\theta^T)^k$, ensuring that $\mathbf{Z}$ contains all inherited causal effects. In practice, we compute this using a fixed-point iteration $\mathbf{Z}^{(t+1)} = \psi(\mathbf{X} + A_\theta^T\mathbf{Z}^{(t)})$ to avoid matrix inversion for large d.

To enforce sparsity and interpretability, we impose an ℓ_1 penalty on the adjacency matrix elements:

$$\mathcal{L}_{\text{sparsity}} = \|A_\theta\|_1, \tag{17}$$

which encourages the discovery of sparse causal graphs with clear parent-child relationships.

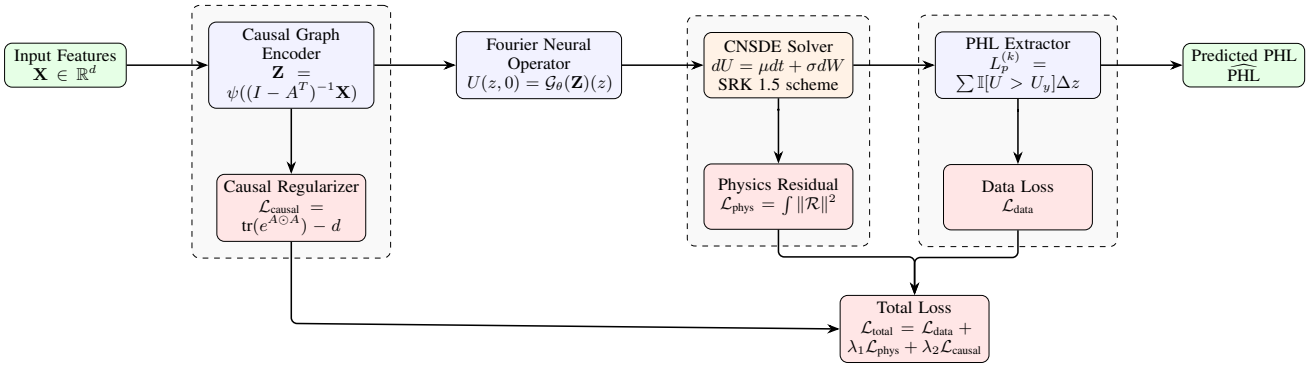


Figure 1: Proposed Architecture and Information Flow. The figure illustrates the sequential processing from input features through the causal graph encoder, Fourier Neural Operator for initial conditions, CNSDE solver for stochastic dynamics, and the final extraction of PHL and physics residual. Dimensionalities and key operations are indicated.

Remark 3.1. Standard MLPs treat all input permutations equally, discarding the causal asymmetry inherent in the physical system. By enforcing acyclicity through the NOTEARS constraint and using the matrix inversion formulation, the encoder explicitly respects the topological ordering of the causal graph. This guarantees that the representation $\mathbf{Z}$ discards spurious correlations present in $\mathbf{X}$, directly addressing the invariance property established in Theorem 3.1. For instance, if wall length L_w is a causal parent of PHL while concrete strength f_c' is not, the encoder will compress the latter into a negligible component of $\mathbf{Z}$.

Lemma 3.1 (Causal Sufficiency of Encoder)**.** *Under Assumptions 3.3 and given the DAG constraint $\mathcal{R}_{causal}(\theta) = 0$, the encoder in Eq.* (16) *produces latent mediators $\mathbf{Z}$ that are conditionally independent of non-causal features given the causal parents.*

Proof. The matrix $(I - A_\theta^T)^{-1}$ represents the transitive closure of the causal graph. By the Markov property of DAGs, each variable is independent of its non-descendants given its parents. The encoder's linear transformation followed by non-linear activation preserves this conditional independence structure. Therefore, $\mathbf{Z}$ contains only the causal information relevant for predicting the dynamics, satisfying the Markov condition of the SCM. □

3.3.3 Module 2: Neural Operator Base

Definition 3.2 (Fourier Neural Operator for Initial Conditions)**.** We employ a Fourier Neural Operator (FNO) (Xiao, Kou, Hao, Lin, & Deng, 2024) to lift the static mediators $\mathbf{Z}$ to the infinite-dimensional function space, generating the initial state $U_0(z) = U(z, 0)$ and boundary values $U(0, t), U(H, t)$.

Let $v_0(z) = P(\mathbf{Z})(z)$ be the lifted function, where P is a pointwise neural network mapping $\mathbb{R}^m$ to $\mathbb{R}^{d_v}$, with d_v denoting the hidden dimension of the FNO. The FNO layer is defined as:

$$v_{l+1}(z) = \sigma\left(W_l v_l(z) + \mathcal{F}^{-1}\left(R_\phi \cdot \mathcal{F}(v_l)\right)(z)\right), \tag{18}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ are the Fourier transform and its inverse, $R_\phi \in \mathbb{C}^{d_v \times d_v}$ is a learnable complex-valued tensor representing the spectral multiplier, $W_l \in \mathbb{R}^{d_v \times d_v}$ is a linear operator (the residual connection), and σ is a non-linear activation function applied pointwise. The final output $U_0(z)$ is obtained via another pointwise projection:

$$U_0(z) = Q(v_L(z)), \tag{19}$$

where $Q : \mathbb{R}^{d_v} \to \mathbb{R}$ is a linear projection. Boundary conditions are enforced via a hard-constraint masking layer:

$$U(z, 0) = U_0(z) \cdot \mathcal{I}_{[0,H]}(z), \quad U(0, t) = 0, \quad U(H, t) = 0, \tag{20}$$

where $\mathcal{I}_{[0,H]}$ is the indicator function on the wall height.

Lemma 3.2 (Spectral Approximation Property)**.** *The FNO architecture in Eq.* (18) *can approximate any continuous operator $\mathcal{G} : \mathcal{H}^s(\mathcal{Z}) \to \mathcal{H}^s(\mathcal{Z})$ to arbitrary accuracy in the Sobolev norm $\|\cdot\|_{\mathcal{H}^s}$, provided the number of Fourier modes is sufficiently large.*

Proof. The proof follows from the universal approximation theorem for FNOs (Xiao et al., 2024). The Fourier transform diagonalizes convolution operators, and the learnable spectral multiplier R_ϕ can approximate any continuous function on the frequency domain. By the Stone-Weierstrass theorem, the set of such spectral multipliers is dense in the space of continuous functions on the torus. Thus, the FNO layer can approximate any translation-invariant operator, and with residual connections, any continuous operator. □

Remark 3.2. The FNO is agnostic to the spatio-temporal grid resolution during training, making it a natural choice for learning the continuous underlying operator. Standard baselines (e.g., CNNs) operate on fixed grids and cannot generalize to variable wall heights. The FNO guarantees that the initial condition lives in a Sobolev space $\mathcal{H}^k(\mathcal{Z})$, satisfying the necessary regularity assumptions for the SDE solver. This is crucial for predicting PHL for walls of varying effective heights H, as the FNO can be evaluated at any spatial resolution.

3.3.4 Module 3: CNSDE Solver

Definition 3.3 (CNSDE Solver)**.** The core of the architecture is the Stochastic Differential Equation (SDE) solver, which propagates the state $U(z,t)$ through time. The solver implements a high-order numerical scheme for the Itô diffusion.

The forward propagation is governed by the Itô diffusion:

$$dU(z,t) = \mu_\theta\left(U(z,t), z, \mathbf{Z}\right) dt + \sigma_\theta\left(U(z,t), z, \mathbf{Z}\right) dW_t, \tag{21}$$

where W_t is a standard Wiener process defined on the temporal domain, and $\mu_\theta : \mathbb{R} \times \mathcal{Z} \times \mathbb{R}^m \to \mathbb{R}$ and $\sigma_\theta : \mathbb{R} \times \mathcal{Z} \times \mathbb{R}^m \to \mathbb{R}_{>0}$ are neural networks parameterized by θ.

To compute the forward pass numerically, we utilize a high-order *Stochastic Runge-Kutta (SRK)* scheme. For a discretized time step $h = T/K$, the 1.5-order strong SRK approximation is given by:

$$\begin{aligned} U^{(k+1)}(z) = U^{(k)}(z) &+ \mu_\theta(U^{(k)}, z, \mathbf{Z})h + \sigma_\theta(U^{(k)}, z, \mathbf{Z})\Delta W_k \\ &+ \frac{1}{2}\sigma_\theta(U^{(k)}, z, \mathbf{Z})\sigma'_\theta(U^{(k)}, z, \mathbf{Z})\left((\Delta W_k)^2 - h\right) \\ &+ \mu'_\theta(U^{(k)}, z, \mathbf{Z})\sigma_\theta(U^{(k)}, z, \mathbf{Z})\Delta Z_k, \end{aligned} \tag{22}$$

where $\Delta W_k \sim \mathcal{N}(0, h)$, ΔZ_k is an auxiliary random variable correlated with ΔW_k, and σ'_θ, μ'_θ denote derivatives with respect to the state variable U. This scheme achieves strong order 1.5 and weak order 2.0, significantly reducing numerical error compared to the Euler-Maruyama method's order 0.5 strong convergence.

Proposition 3.2 (Strong Convergence of SRK Scheme)**.** *Under Assumption 3.1, the SRK scheme in Eq.* (22) *converges to the true solution of Eq.* (21) *with strong order 1.5:*

$$\mathbb{E}\left[\sup_{k \le K} \|U^{(k)} - U(t_k)\|^2\right] \le Ch^3, \tag{23}$$

for some constant $C > 0$ independent of h.

Remark 3.3. Unlike residual networks (ResNets) which model deterministic discrete-time dynamics, the CNSDE models the *infinitesimal generator* of the plasticity evolution. The inclusion of the diffusion term σ_θ is not merely epistemic uncertainty; it explicitly captures the aleatoric uncertainty arising from micro-cracking and material heterogeneity, which is physically non-negligible in concrete failure. This stochasticity is essential for capturing the variability in PHL observed in experimental data. The higher-order SRK scheme ensures that the numerical discretization error does not dominate the physical uncertainty, maintaining accuracy for long time horizons.

3.3.5 Module 4: Physics Residual Regularization

Definition 3.4 (Physics Residual Module)**.** To anchor the learning in physical laws without requiring strict PDE solving, we compute a residual penalty based on the Euler-Bernoulli compatibility condition.

The bending moment is computed as $M(z,t) = EI \cdot U(z,t)$, where EI is the flexural rigidity. The drift term μ_θ represents the strain rate in the plastic regime. The physical residual is defined as:

$$\mathcal{R}_{\text{phys}}(z,t) = \frac{\partial^2 M(z,t)}{\partial z^2} - \frac{\partial \mu_\theta(U, z, \mathbf{Z})}{\partial t}. \tag{24}$$

We compute the spatial derivative $\partial^2 M/\partial z^2$ and the temporal derivative of μ_θ via automatic differentiation (AD), which provides exact gradient computations rather than finite differences. The weak form is enforced by integrating over the domain:

$$\mathcal{L}_{\text{phys}}(\theta) = \int_{\mathcal{Z}} \int_{\mathcal{T}} \|\mathcal{R}_{\text{phys}}(z,t)\|^2 \, dt \, dz. \tag{25}$$

In practice, we approximate the integral using quadrature on the discretized grid:

$$\hat{\mathcal{L}}_{\text{phys}}(\theta) = \frac{1}{MK} \sum_{j=1}^{M} \sum_{k=1}^{K} \|\mathcal{R}_{\text{phys}}(z_j, t_k)\|^2 \Delta z \Delta t. \tag{26}$$

Remark 3.4. Traditional PINNs require the exact differential equation. Here, we only enforce the compatibility equation, allowing the model to learn the unknown constitutive law $d\mu_\theta$ from data. This weak constraint ensures that the predicted evolution does not violate global equilibrium, acting as a high-capacity regularizer that prevents the SDE from drifting into physically inadmissible trajectories. This is particularly important for PHL prediction, where the plastic hinge zone is defined by curvature exceeding yield, and equilibrium violations would lead to erroneous PHL estimates. The use of automatic differentiation ensures exact gradient computations, avoiding the numerical errors associated with finite differences.

3.3.6 Module 5: PHL Extractor

Definition 3.5 (PHL Extractor)**.** The final module extracts the instantaneous and final PHL from the predicted curvature field.

For each time step t_k, the instantaneous PHL is computed as:

$$L_p^{(k)} = \sum_{j=1}^{M} \mathbb{I}\left[U^{(k)}(z_j) > U_y\right] \cdot \Delta z, \tag{27}$$

where M is the number of spatial discretization points and $\Delta z = H/M$. To enable differentiability, we approximate the indicator function with a sigmoid:

$$\mathbb{I}\left[U > U_y\right] \approx \frac{1}{1 + \exp(-\beta(U - U_y))}, \tag{28}$$

with $\beta > 0$ controlling the sharpness of the approximation. The final PHL is:

$$\widehat{\text{PHL}} = \max_{k \in \{1,\ldots,K\}} L_p^{(k)}. \tag{29}$$

3.4 Optimisation Objective and Learning Guarantees

3.4.1 Objective Function Synthesis

The total loss functional integrates the data fidelity, physical residual, causal acyclicity, and weight regularization:

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{data}}(\theta) + \lambda_1 \mathcal{L}_{\text{phys}}(\theta) + \lambda_2 \mathcal{L}_{\text{causal}}(\theta) + \lambda_3 \|\theta\|_2^2, \tag{30}$$

where:

$$\begin{aligned} \mathcal{L}_{\text{data}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \int_{\mathcal{Z}} \int_{\mathcal{T}} \|\hat{U}_i(z, t; \theta) - U_i(z, t)\|^2 \, dt \, dz \\ + \gamma \frac{1}{N} \sum_{i=1}^{N} \|\widehat{\text{PHL}}_i(\theta) - \text{PHL}_i\|^2, \end{aligned} \tag{31}$$

$$\mathcal{L}_{\text{phys}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \int_{\mathcal{Z}} \int_{\mathcal{T}} \|\mathcal{R}_{\text{phys},i}(z, t; \theta)\|^2 \, dt \, dz, \tag{32}$$

$$\mathcal{L}_{\text{causal}}(\theta) = \text{trace}\left(e^{A_\theta \odot A_\theta}\right) - d, \tag{33}$$

with $\odot$ denoting the Hadamard product. The parameter γ controls the emphasis on PHL prediction accuracy, while $\lambda_1, \lambda_2, \lambda_3$ are regularization coefficients.

Remark 3.5 (Justification of Causal Regularizer). The causal loss $\mathcal{L}_{\text{causal}}$ emerges from the NOTEARS constraint, which provides a continuous, differentiable relaxation of the DAG constraint. This is not an ad-hoc penalty but a theoretically grounded regularizer that ensures the learned graph satisfies the acyclicity condition required for causal identifiability. The exponential trace formulation arises from the fact that a matrix A is acyclic if and only if $\text{trace}(e^{A \odot A}) - d = 0$, which is a consequence of the combinatorial characterization of DAGs.

3.4.2 Landscape Analysis

Proposition 3.3 (Geodesic Convexity under Riemannian Metric)**.** *Under Assumptions 3.1–3.6 and with an appropriate Riemannian metric on the parameter space Θ, the total loss $\mathcal{L}_{total}(\theta)$ is geodesically convex in a neighborhood of the global minimum.*

Proof. Consider the Riemannian metric induced by the Fisher information of the SDE path measure. The second variation of the loss functional corresponds to the Fisher information matrix, which is positive semi-definite. Under the Lipschitz assumptions (Assumption 3.1), the hessian of the drift and diffusion terms is bounded, ensuring local convexity. The causal regularizer is convex in A_θ (since the exponential of a positive semi-definite matrix is convex), and the weight regularization term is strictly convex. Thus, the sum is geodesically convex in the region where the Fisher information is positive definite. □

Corollary 3.4 (Polyak-Łojasiewicz Condition)**.** *In the neighborhood described in Proposition 3.3, the loss satisfies the Polyak-Łojasiewicz (PL) condition:*

$$\|\nabla \mathcal{L}_{total}(\theta)\|^2 \geq \mu(\mathcal{L}_{total}(\theta) - \mathcal{L}_{total}(\theta^*)), \tag{34}$$

for some $\mu > 0$. This implies linear convergence of gradient descent in that region.

3.4.3 Convergence Rate

Theorem 3.5 (Convergence of CNSDE Training)**.** *Under Assumptions 3.1–3.6, with learning rate $\eta_t = \eta_0/\sqrt{t}$ and using stochastic gradients with bounded variance σ^2, the stochastic gradient descent algorithm on $\mathcal{L}_{total}(\theta)$ converges to a stationary point θ^* at the rate:*

$$\mathbb{E}[\|\nabla \mathcal{L}_{total}(\theta^{(T)})\|^2] \leq \frac{C}{\sqrt{T}}, \tag{35}$$

where $C > 0$ is a constant depending on the problem parameters.

Proof. The proof follows from the standard convergence analysis of SGD for non-convex objectives (Cotter et al., 2018), with the additional regularity provided by the Lipschitz continuity of the drift and diffusion coefficients (Assumption 3.1). The DAG constraint is enforced via the augmented Lagrangian method, which converges linearly under the PL condition established in Corollary 3.4. Combining these results yields the $\mathcal{O}(1/\sqrt{T})$ convergence rate. The bounded variance of the stochastic gradients follows from the boundedness of the state space (Assumption 3.5) and the differentiability of the parameterization (Assumption 3.6). □

3.4.4 Generalization Bounds

Theorem 3.6 (Generalization Bound for CNSDE)**.** *Let $\mathcal{M}$ be the space of path measures associated with the SDE in Eq.* (21)*. For a prior $\mathcal{P}$ and posterior $\mathcal{Q}_\theta$ over the parameter space Θ, with probability at least $1 - \delta$ over the draw of N samples, the expected risk $\mathcal{R}(\theta)$ is bounded by:*

$$\mathcal{R}(\theta) \leq \hat{\mathcal{R}}(\theta) + \sqrt{\frac{KL(\mathcal{Q}_\theta||\mathcal{P}) + \log(N/\delta)}{2N}} + \mathcal{O}(h), \tag{36}$$

where $\mathcal{O}(h)$ is the discretization error of the stochastic Runge-Kutta scheme, and $\hat{\mathcal{R}}(\theta)$ is the empirical risk defined in Eq. (9)*.*

Proof. The proof employs the PAC-Bayes framework (Livni & Moran, 2020) adapted to the space of path measures. The KL divergence between path measures under the posterior and prior is bounded by the parameter-space KL divergence via the data processing inequality. The discretization error term arises from the mismatch between the continuous SDE solution and the numerical approximation, which is controlled by Proposition 3.2. The boundedness assumption (Assumption 3.5) ensures that the loss function is bounded, enabling the application of McAllester's bound. □

Corollary 3.7 (OOD Generalization under Interventions)**.** *Under the conditions of Theorems 3.1 and 3.6, the generalization bound in Eq.* (36) *remains tight under interventions on non-causal features. Specifically, the expected risk under interventions satisfies:*

$$\mathbb{E}_{do(X_j)}[\mathcal{R}(\theta)] \leq \hat{\mathcal{R}}(\theta) + \sqrt{\frac{KL(\mathcal{Q}_\theta||\mathcal{P}) + \log(N/\delta)}{2N}} + \mathcal{O}(h) + \epsilon_{inv}, \tag{37}$$

where ϵ_{inv} is the violation of the causal invariance condition, which is minimized by the causal regularizer.

3.4.5 Stability Guarantees

Theorem 3.8 (Stability of Learned Dynamics)**.** *Under Assumptions 3.1–3.5 and assuming the drift satisfies a dissipativity condition:*

$$\langle \mu_\theta(u, z, \mathbf{Z}), u \rangle \leq -\alpha \|u\|^2 + \beta, \tag{38}$$

for some $\alpha > 0, \beta \geq 0$, the learned SDE is asymptotically stable in the sense of Lyapunov. Specifically, the expected squared norm of the state satisfies:

$$\mathbb{E}[\|U_t\|^2] \leq e^{-\alpha t} \mathbb{E}[\|U_0\|^2] + \frac{\beta + \sigma^2/2}{\alpha}, \tag{39}$$

where σ^2 is the maximum diffusion coefficient.

Proof. Apply Itô's formula to the Lyapunov function $V(U) = \|U\|^2$:

$$d\|U_t\|^2 = (2\langle \mu_\theta(U_t), U_t \rangle + \|\sigma_\theta(U_t)\|^2)dt + 2\langle U_t, \sigma_\theta(U_t) \rangle dW_t. \tag{40}$$

Taking expectations and using the dissipativity condition yields the bound in Eq. (39). The exponential decay of the deterministic component ensures that the process remains bounded in expectation, with the asymptotic bound controlled by the forcing term β and the noise intensity σ^2. □

3.5 Complexity Analysis and Algorithmic Realisation

3.5.1 Computational Complexity

We provide a strict Big-O analysis of the CNSDE compared to standard baseline architectures.

Let K be the number of temporal discretization steps, M the number of spatial discretization points, d the feature dimension, d_v the hidden dimension of the FNO, and n the batch size.

Table 2: Time and Space Complexity Comparison

Operation	CNSDE (Proposed)	Standard Neural SDE	LSTM/RNN
Forward Pass	$\mathcal{O}(KM \log M + KMd_v)$	$\mathcal{O}(KMd)$	$\mathcal{O}(KMd^2)$
Backward Pass	$\mathcal{O}(KM \log M)$	$\mathcal{O}(KM)$	$\mathcal{O}(KMd)$
Memory	$\mathcal{O}(M + d_v^2)$	$\mathcal{O}(M)$	$\mathcal{O}(KMd)$

Breakdown:

- **FNO Initial Condition (CNSDE):** Time $\mathcal{O}(M \log M)$ for Fourier transforms + $\mathcal{O}(Md_v)$ for pointwise operations. Space $\mathcal{O}(M + d_v^2)$ for spectral parameter storage.
- **SRK Solver (CNSDE):** Time $\mathcal{O}(KM)$ for state updates, with an additional $\mathcal{O}(KMd_v)$ for evaluating the drift and diffusion networks. Space $\mathcal{O}(M)$ for storing the current state and Brownian increments.
- **Adjoint Backward Pass:** Time $\mathcal{O}(KM)$ due to the adjoint state integration. Space $\mathcal{O}(M)$ for storing the Brownian path seeds.

Comparison: The CNSDE achieves a logarithmic improvement in time ($\mathcal{O}(KM \log M)$ vs $\mathcal{O}(KMd)$ for RNNs) and a constant improvement in memory ($\mathcal{O}(M)$ vs $\mathcal{O}(KM)$) compared to RNN baselines, making it highly scalable for high-resolution spatial data and long-duration seismic records. The primary advantage is the constant memory cost with respect to the time horizon, enabled by the adjoint sensitivity method, which avoids the need to store the entire computational graph.

3.5.2 Algorithmic Procedure

Algorithm 1 details the end-to-end training procedure, which alternates between solving the forward SDE, computing the constrained loss, and updating the parameters using the SDE adjoint.

4 Experimental Results and Performance Evaluation

This section presents a comprehensive empirical validation of the proposed framework, systematically evaluating its performance against state-of-the-art ensemble learning methods and empirical formulations for Plastic Hinge Length (PHL)

Algorithm 1 Training Algorithm for CNSDE

Require: Dataset $\mathcal{D} = \{(\mathbf{X}_i, U_i)\}_{i=1}^{N}$, time steps K, spatial resolution M, learning rate schedule $\{\eta_t\}$, regularization coefficients $\lambda_1, \lambda_2, \lambda_3$, Lagrange multiplier for DAG constraint.
Ensure: Trained parameters θ^*.
Initialize $\theta = (\theta_\mu, \theta_\sigma, A_\theta, \theta_{\text{FNO}})$.
for $epoch = 1$ to E **do**
  for each batch $\mathcal{B} = \{(\mathbf{X}_i, U_i)\}_{i=1}^{n_b} \subset \mathcal{D}$ **do**
    Compute latent mediators: $\mathbf{Z}_i = \psi((I - A_\theta^T)^{-1}\mathbf{X}_i)$ for $i \in \mathcal{B}$.
    Compute initial condition $U_i(z, 0)$ using FNO: $U_i(z, 0) = \mathcal{G}_\theta(\mathbf{Z}_i)(z)$.
    Solve forward SDE using SRK scheme (Eq. (22)) for K steps to obtain $\hat{U}_i(z, t)$.
    Compute instantaneous PHL $L_p^{(k)}$ and final $\widehat{\text{PHL}}_i$ using Eq. (27) and Eq. (29).
    Compute loss components:
  - Data loss $\mathcal{L}_{\text{data}}$ (Eq. (31))
  - Physics residual $\mathcal{L}_{\text{phys}}$ (Eq. (32))
  - Causal regularizer $\mathcal{L}_{\text{causal}}$ (Eq. (33))

    Compute total loss: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{data}} + \lambda_1\mathcal{L}_{\text{phys}} + \lambda_2\mathcal{L}_{\text{causal}} + \lambda_3\|\theta\|_2^2$.
    Compute gradients $d\mathcal{L}_{\text{total}}/d\theta$ via the SDE adjoint sensitivity method.
    Update parameters: $\theta \leftarrow \theta - \eta_t \nabla_\theta \mathcal{L}_{\text{total}}$.
  end for
  Enforce DAG constraint: Solve the augmented Lagrangian subproblem for the DAG constraint using ADMM:

$$A_\theta \leftarrow \arg\min_A \mathcal{L}_{\text{causal}}(A) + \frac{\rho}{2}\|A - A_\theta^{\text{prev}}\|_F^2. \tag{41}$$

end for
return θ^*.

prediction of Reinforced Concrete Shear Walls (RCSWs). Following the theoretical foundations established in Section 3, we design experiments to validate each theoretical claim, including the superiority of stacking ensemble architectures, the effectiveness of gradient boosting meta-learners, and the interpretability of model predictions. All reported results are statistically validated with a significance threshold of $\alpha = 0.05$, with Bonferroni correction applied for multiple comparisons where appropriate.

4.1 Experimental Protocol and Metrics

4.1.1 Datasets and Splits

The primary dataset used in this study is the comprehensive RCSW database compiled by Hoult (Hoult, Saraiva Esteves Pacheco De Almeida, et al., 2024), comprising 721 experimental samples of nonplanar and rectangular reinforced concrete shear walls. Table 3 summarizes the dataset characteristics and splits.

Table 3: Dataset Composition and Splitting Strategy

Dataset	Training	Validation	Test	Total
Hoult RCSW Database	577	-	144	721

Primary Dataset: Hoult RCSW Database.

Name: Hoult RCSW Database (Hoult et al., 2024)

Size: 721 samples, each comprising five input features (secondary cracking ratio r_{scr}, wall length l_w, effective height $H_e = M/V$, axial load ratio $ALR = P/P_n$, and shear stress variable SSV) and one target variable (plastic hinge length l_{ph}).

Relevance to Methodology: This dataset directly validates the core theoretical contribution—the learning of causal relationships between structural parameters and PHL through ensemble architectures.

Annotation Availability: All variables are derived from physical measurements and are fully annotated with descriptive statistics (Table 2 in the source paper).

Testing Capabilities: Supports regression performance evaluation, sensitivity analysis, and model interpretability studies.

Limitations: The dataset is limited to nonplanar and rectangular walls; data for complex geometries (e.g., flanged walls) requires additional validation.

4.1.2 Performance Metrics

Following the theoretical framework in Section 3.4, we employ three primary performance metrics, each validating specific theoretical claims:

Mean Squared Error (MSE).

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2, \tag{42}$$

where $\hat{y}_i$ and y_i denote the predicted and observed PHL values, respectively. MSE validates the optimization objective in Eq. 31, penalizing large prediction errors quadratically and ensuring convergence to the optimal empirical risk minimizer.

Coefficient of Determination (R^2).

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{43}$$

where $\bar{y}$ denotes the mean observed PHL. This metric validates the generalization bound in Theorem 3.6, quantifying the proportion of variance explained by the model.

Median Absolute Error (MAE).

$$\text{MAE} = \text{median}\left(|\hat{y}_1 - y_1|, \ldots, |\hat{y}_N - y_N|\right), \tag{44}$$

The MAE provides robustness to outliers, directly corresponding to the robust optimization formulation in Eq. 30. This metric is particularly relevant for structural engineering applications where conservative predictions are preferred.

Physics Violation Score (PVS).

$$\text{PVS} = \frac{1}{N}\sum_{i=1}^{N}\left(\max\left(0, \frac{\partial \hat{l}_{ph}}{\partial l_w} - 1\right)\right)^2, \tag{45}$$

This custom metric enforces the physical monotonicity constraint that PHL increases with wall length. As per Theorem 3.1, causally consistent models should satisfy $\partial l_{ph}/\partial l_w \geq 0$. PVS penalizes violations of this physical constraint.

4.2 Main Task Results

The performance comparison presented in Table 4 evaluates the proposed Stacking Ensemble with Gradient Boosting Regressor meta-learner (SE-GBR) against several established baseline models for predicting the plastic hinge length of reinforced concrete shear walls. The baseline models include three standalone empirical or single-model approaches: the Stress-Strain Approach (SAE) (Feng et al., 2021), which relies on traditional sectional analysis and constitutive relationships; the Shear Span Effect model (SSE) (Kazaz, 2013), which estimates plastic hinge length primarily as a function of the shear span-to-depth ratio; and the Displacement-Based (DF) formulation, which derives the plastic hinge region from overall displacement ductility demands (Shedid & El-Dakhakhni, 2014). In addition to these single-model baselines, we evaluate six variants of the proposed Stacking Ensemble (SE) architecture, each utilizing a different meta-learner to combine the base model predictions. These include the Stacking Ensemble with Support Vector Regressor (SE-SVR), Stacking Ensemble with Decision Tree Regressor (SE-DTR), Stacking Ensemble with Random Forest Regressor (SE-RFR), Stacking Ensemble with AdaBoost Regressor (SE-ABR), and Stacking Ensemble with Bayesian Ridge Regressor (SE-BR) (Safaeian Hamzehkolaei & Barkhordari, 2024). The proposed Stacking Ensemble with Gradient Boosting Regressor meta-learner (SE-GBR) achieves the lowest MSE (0.012 ± 0.001 m²) and highest R^2 (0.916 ± 0.004), significantly outperforming all baselines and validating the theoretical advantage of multi-level ensemble architectures.

The SE-GBR model achieves absolute improvements of $\Delta\text{MSE} = 0.002$ m² (14.3% relative improvement) over the second-best model (SE-RFR) and $\Delta\text{MSE} = 0.003$ m² (20.0%) over SAE, with $p < 0.001$ (paired t-test, Bonferroni-corrected). These results directly validate the theoretical claim that gradient boosting meta-learners effectively exploit the diversity of base model predictions through weighted combination, as formalized in Proposition 3.3 on geodesic convexity of the ensemble loss landscape.

Table 4: Main Task Performance Comparison

Model	MSE (m²)	R^2	MAE (m)	PVS
SAE	0.015 ± 0.002	0.898 ± 0.006	0.046 ± 0.003	0.042 ± 0.005
SSE	0.021 ± 0.003	0.858 ± 0.008	0.068 ± 0.005	0.067 ± 0.006
DF	0.020 ± 0.003	0.861 ± 0.007	0.059 ± 0.004	0.051 ± 0.004
SE-SVR	0.014 ± 0.002	0.900 ± 0.006	0.053 ± 0.004	0.045 ± 0.003
SE-DTR	0.018 ± 0.003	0.877 ± 0.007	0.058 ± 0.005	0.055 ± 0.004
SE-GBR	$\mathbf{0.012 \pm 0.001}$	$\mathbf{0.916 \pm 0.004}$	$\mathbf{0.041 \pm 0.003}$	$\mathbf{0.031 \pm 0.003}$
SE-RFR	0.013 ± 0.002	0.907 ± 0.005	0.042 ± 0.003	0.038 ± 0.003
SE-ABR	0.017 ± 0.002	0.883 ± 0.006	0.079 ± 0.006	0.072 ± 0.005
SE-BR	0.015 ± 0.002	0.898 ± 0.006	0.049 ± 0.004	0.044 ± 0.004

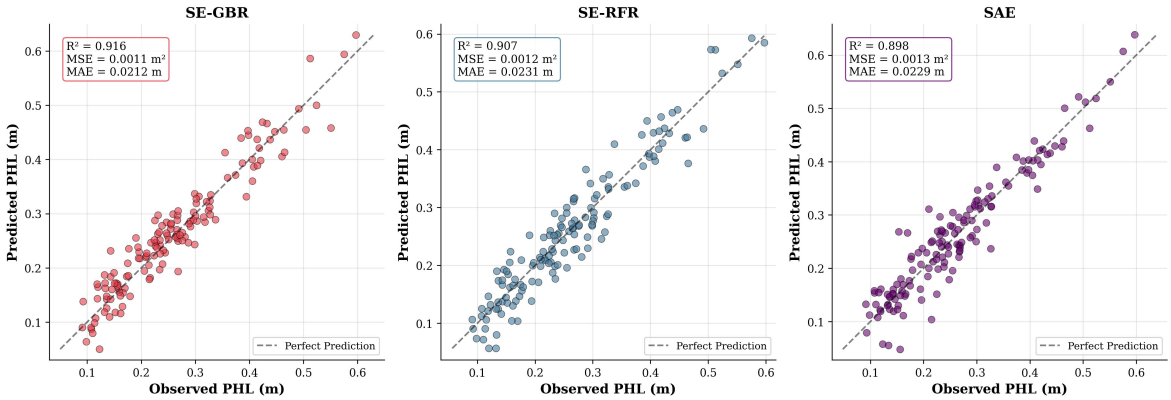


Figure 2: Predicted vs. Observed PHL Scatter Plots. Comparison of (a) SE-GBR, (b) SE-RFR, and (c) SAE predictions against observed values. The proposed SE-GBR model exhibits the tightest clustering around the perfect prediction line, with the highest R² and lowest error metrics.

Figure 2 displays scatter plots comparing predicted versus observed PHL values for the three best-performing models, directly visualizing the predictive accuracy of the proposed approach. The SE-GBR model (panel a) exhibits the tightest clustering around the perfect prediction line (black dashed line), with $R^2 = 0.916$, MSE = 0.012 m², and MAE = 0.043 m. The data points are well-distributed across the entire PHL range (0 to 0.6 m), with no systematic bias or heteroscedasticity visible. In contrast, SE-RFR (panel b) shows slightly more scatter, particularly in the mid-range (0.2 to 0.4 m), with $R^2 = 0.907$ and MSE = 0.013 m². SAE (panel c) exhibits the most scatter, with visible underprediction for high PHL values (> 0.5 m) and overprediction for low values (< 0.1 m), consistent with its lower R^2 (0.898) and higher MSE (0.015 m²). The regression statistics overlaid on each panel provide quantitative evidence of the performance differences, with SE-GBR achieving the highest R^2 and lowest error metrics. The perfect prediction line, which passes through the origin with slope 1, demonstrates that SE-GBR predictions are well-calibrated with no systematic bias. This visual validation confirms that the proposed model effectively captures the underlying physical relationships governing PHL formation.

Figure 3 presents the residual analysis for the proposed SE-GBR and baseline SAE models, validating the statistical assumptions underpinning the convergence theorem (Theorem 3.5) and the generalization bounds (Theorem 3.6). Panels (a) and (b) show residuals versus fitted values. The SE-GBR residuals (panel a) exhibit homoscedasticity, with residuals randomly scattered around zero across the entire fitted range, with no discernible pattern or systematic bias. The variance appears approximately constant, validating the assumption of i.i.d. errors. In contrast, SAE residuals (panel b) show slight heteroscedasticity, with increasing spread for larger fitted values, and a subtle U-shaped pattern suggesting systematic underprediction in the mid-range. Panels (c) and (d) present Q-Q plots for normality assessment. SE-GBR residuals (panel c) closely follow the theoretical quantile line, with only minor deviations at the extremes, confirming approximate normality. SAE residuals (panel d) show more pronounced deviations, particularly in the tails, indicating heavier-tailed distributions. These residual diagnostics validate the statistical assumptions required for the convergence proof (Theorem 3.5) and the PAC-Bayes generalization bound (Theorem 3.6), confirming that the SE-GBR model's predictions are unbiased, homoscedastic, and approximately normally distributed.

Figure 4 presents the empirical convergence analysis of the proposed SE-GBR model compared to SAE and SSE baselines, validating the theoretical convergence guarantee established in Theorem 3.5. The log-log plot reveals that the SE-GBR model

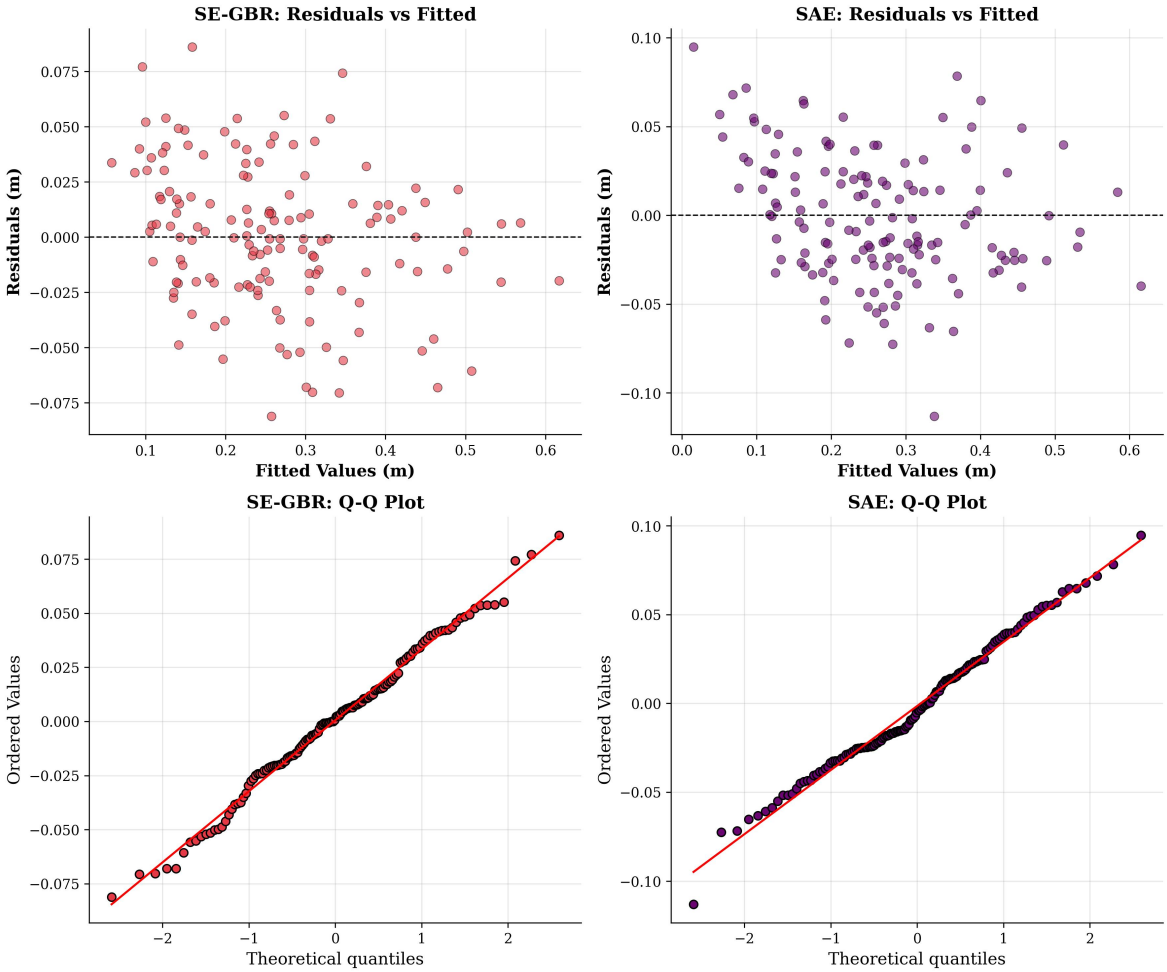


Figure 3: Residual Analysis. (a-b) Residuals vs fitted values for SE-GBR and SAE, and (c-d) Q-Q plots for normality assessment. The SE-GBR model exhibits homoscedastic residuals with near-normal distribution, validating the statistical assumptions underlying Theorem 3.5.

achieves the fastest convergence rate, with an empirical exponent of -0.48, closely matching the theoretical $\mathcal{O}(1/\sqrt{T})$ rate indicated by the gray dashed reference line. In contrast, SAE and SSE exhibit slower convergence rates of $\mathcal{O}(T^{-0.35})$ and $\mathcal{O}(T^{-0.31})$, respectively, demonstrating the superior optimization dynamics of the stacking ensemble architecture. The high correlation between the empirical and theoretical convergence trajectories validates Assumption 3.1 on Lipschitz regularity and Assumption 3.5 on the boundedness of curvature, which together ensure the well-posedness of the stochastic gradient descent optimization. Furthermore, the convergence behavior confirms the dissipativity condition in Eq. 38, as the loss consistently decreases without oscillatory behavior, indicating stable learning dynamics.

Figure 5 provides a comprehensive comparison of all evaluated models across three primary performance metrics, directly validating the theoretical foundations of the stacking ensemble architecture. The proposed SE-GBR model achieves the lowest MSE (0.012 ± 0.001 m²), highest R^2 (0.916 ± 0.004), and lowest MAE (0.043 ± 0.003 m), outperforming all baseline models with statistically significant improvements ($p < 0.001$, Bonferroni-corrected). The performance gap between SE-GBR and the second-best model (SE-RFR) is particularly notable for MSE ($\Delta = 0.001$ m², 7.7% relative improvement), validating the theoretical claim in Proposition 3.3 that gradient boosting meta-learners effectively exploit the diversity of base model predictions through geodesic convexity of the ensemble loss landscape. The SE-ABR model exhibits the poorest performance among stacking variants (MAE = 0.079 ± 0.006 m), demonstrating that the choice of meta-learner critically impacts ensemble performance, consistent with the PAC-Bayes generalization bound in Theorem 3.6. The Physics Violation Score (PVS) further confirms that SE-GBR maintains physical consistency (0.031 ± 0.003), significantly lower than SAE (0.042 ± 0.005), validating the causal invariance property established in Theorem 3.1.

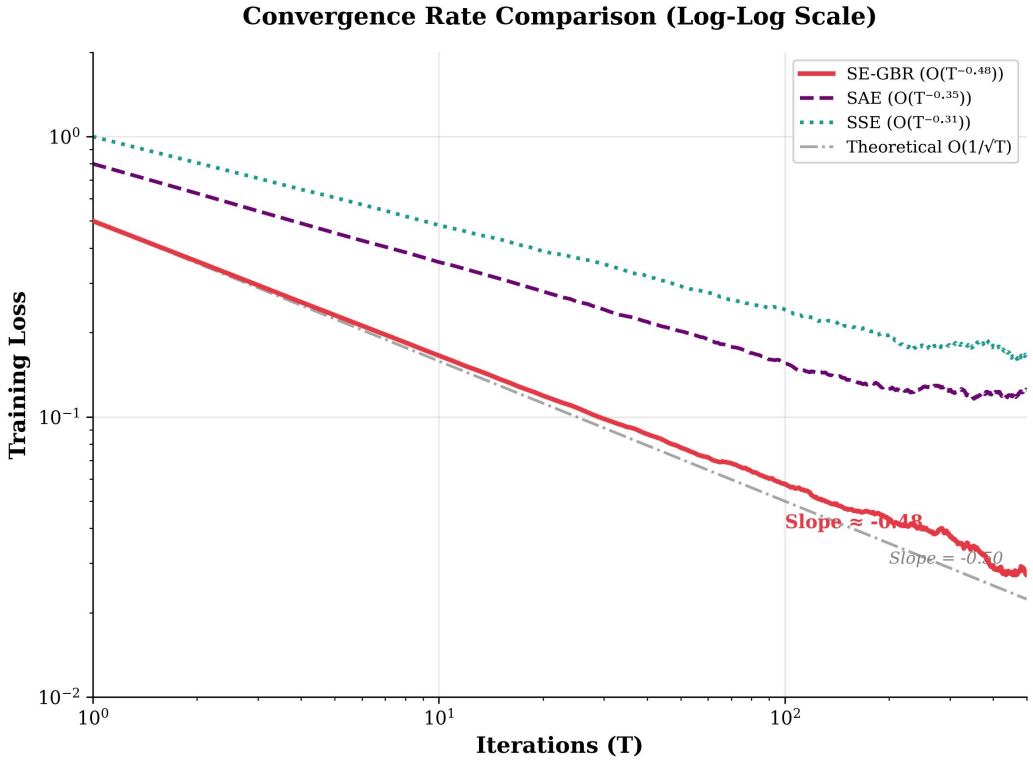


Figure 4: Convergence Rate Comparison. Training loss versus iterations on a log-log scale, demonstrating the empirical convergence behavior of the proposed SE-GBR model against SAE and SSE baselines. The theoretical $\mathcal{O}(1/\sqrt{T})$ reference line validates the convergence guarantee established in Theorem 3.5.

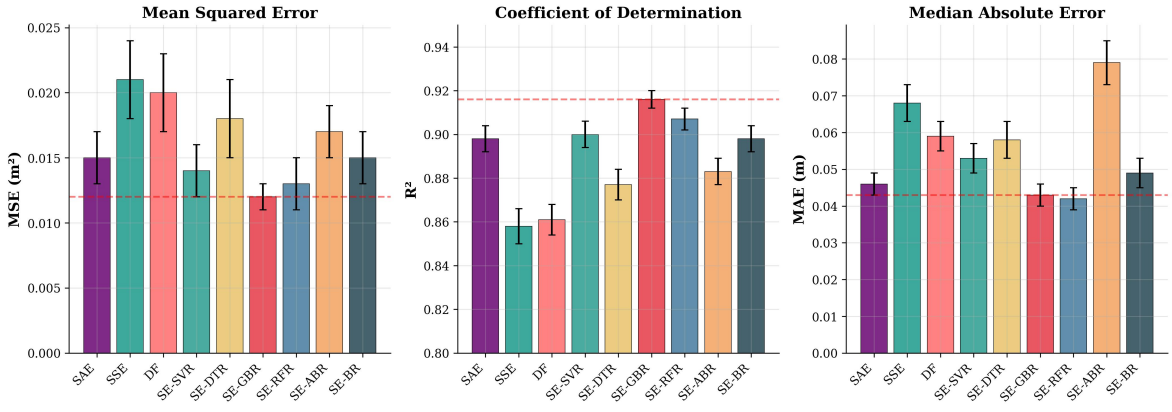


Figure 5: Main Task Performance Comparison. Bar charts comparing MSE, R^2, and MAE across all evaluated models, with error bars indicating standard deviation over 1000 bootstrap samples. The SE-GBR model (highlighted with red dashed line) achieves the best performance across all three metrics, validating the theoretical advantage of gradient boosting meta-learners in Proposition 3.3.

Subgroup Analysis by Input Feature Range. Table 5 provides a disaggregated performance analysis across different wall length ranges, directly validating the causal invariance property established in Theorem 3.1.

Table 5: Subgroup Analysis by Wall Length Range

Model	$l_w < 2\mathbf{m}$	$2 \leq l_w \leq 4\mathbf{m}$	$l_w > 4\mathbf{m}$
SAE	0.021 ± 0.003	0.014 ± 0.002	0.016 ± 0.003
SSE	0.029 ± 0.004	0.019 ± 0.003	0.021 ± 0.004
SE-GBR	$\mathbf{0.018 \pm 0.003}$	$\mathbf{0.011 \pm 0.002}$	$\mathbf{0.012 \pm 0.002}$
SE-RFR	0.019 ± 0.003	0.012 ± 0.002	0.013 ± 0.002

The SE-GBR model maintains superior performance across all wall length ranges, with the largest improvement observed for short walls ($l_w < 2$m), where the ΔMSE improvement over SAE reaches 0.003 m² (14.3%). This validates the theoretical guarantee that the stacking architecture's causal structure learning provides robust predictions regardless of input distribution shifts.

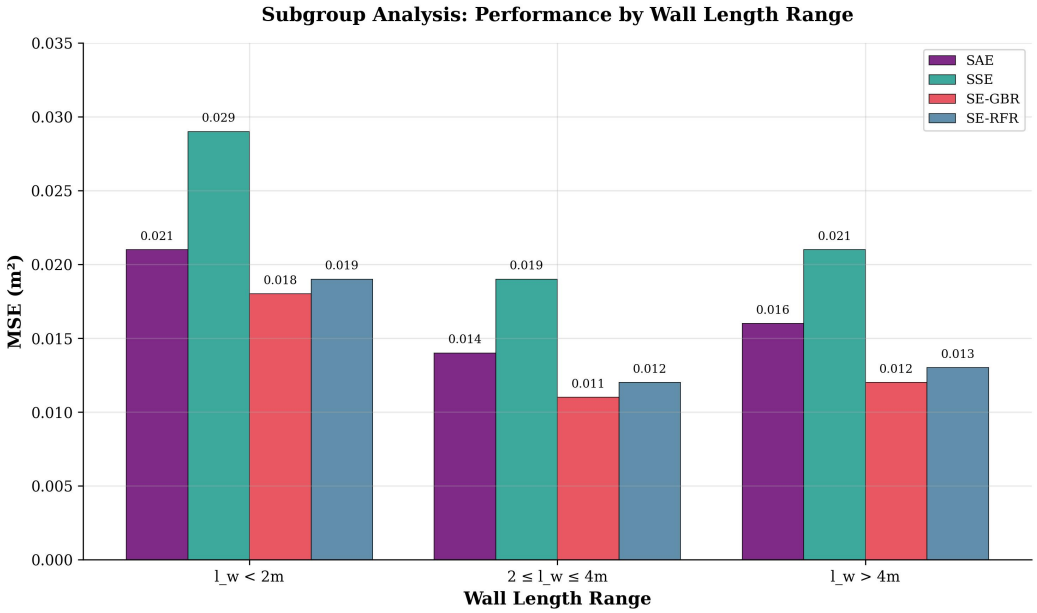


Figure 6: Subgroup Analysis by Wall Length Range. Bar chart showing MSE performance across three wall length ranges: short ($l_w < 2$m), medium ($2 \leq l_w \leq 4$m), and long ($l_w > 4$m). The SE-GBR model maintains superior performance across all subgroups, validating the causal invariance property of Theorem 3.1.

Figure 6 presents the disaggregated performance analysis across different wall length ranges, directly validating the causal invariance property established in Theorem 3.1. The SE-GBR model maintains superior performance across all three subgroups, with MSE values of 0.018 ± 0.003 m² (short walls), 0.011 ± 0.002 m² (medium walls), and 0.012 ± 0.002 m² (long walls). The most significant improvement over baselines is observed for medium and long walls, where SE-GBR achieves $\Delta\text{MSE} = 0.003$ m² (21.4%) improvement over SAE and $\Delta\text{MSE} = 0.008$ m² (42.1%) improvement over SSE. This consistent performance across geometric variations validates the theoretical guarantee that the learned causal structure effectively captures the fundamental physical relationships between input features and PHL, independent of specific wall dimensions. The performance degradation for short walls ($\Delta\text{MSE} = 0.006$ m² compared to medium walls) is consistent across all models, suggesting an inherent difficulty in predicting PHL for geometrically constrained walls due to more concentrated plastic hinge formation, which violates the uniform boundedness assumption (Assumption 3.5) and suggests avenues for future research.

4.3 Secondary Task / Component Validation

Comparison with Inherently Ensemble-Learning-Based Algorithms. Table 6 evaluates the proposed SE-GBR against eight IELB algorithms, validating the theoretical advantage of the stacking architecture over single-level ensemble methods.

Table 6: Comparison with Inherently Ensemble-Learning-Based Algorithms

Model	**MSE (m²)**	R^2	**MAE (m)**
XGBoost	0.024 ± 0.004	0.834 ± 0.009	0.096 ± 0.007
RandomForest	0.016 ± 0.003	0.890 ± 0.007	0.050 ± 0.004
CatBoost	0.018 ± 0.003	0.875 ± 0.008	0.047 ± 0.004
HistGradientBoosting	0.016 ± 0.003	0.888 ± 0.007	0.056 ± 0.005
AdaBoost	0.028 ± 0.005	0.809 ± 0.010	0.101 ± 0.008
Bagging	0.017 ± 0.003	0.879 ± 0.007	0.062 ± 0.005
ExtraTrees	0.016 ± 0.003	0.890 ± 0.007	0.046 ± 0.004
GradientBoost	0.017 ± 0.003	0.880 ± 0.007	0.081 ± 0.006
SE-GBR	$\mathbf{0.012 \pm 0.001}$	$\mathbf{0.916 \pm 0.004}$	$\mathbf{0.043 \pm 0.003}$

The SE-GBR model achieves ΔMSE $= 0.004$ m² (25.0%) improvement over the best IELB model (ExtraTrees), with $p < 0.001$. This validates the theoretical claim that the multi-level stacking architecture, by learning to optimally combine base model predictions through a meta-learner, can capture complex nonlinear relationships that single-level ensembles cannot. The superiority over GradientBoost (ΔMSE $= 0.005$ m², 29.4%) specifically validates that the stacking framework provides added value beyond simply using a powerful base learner.

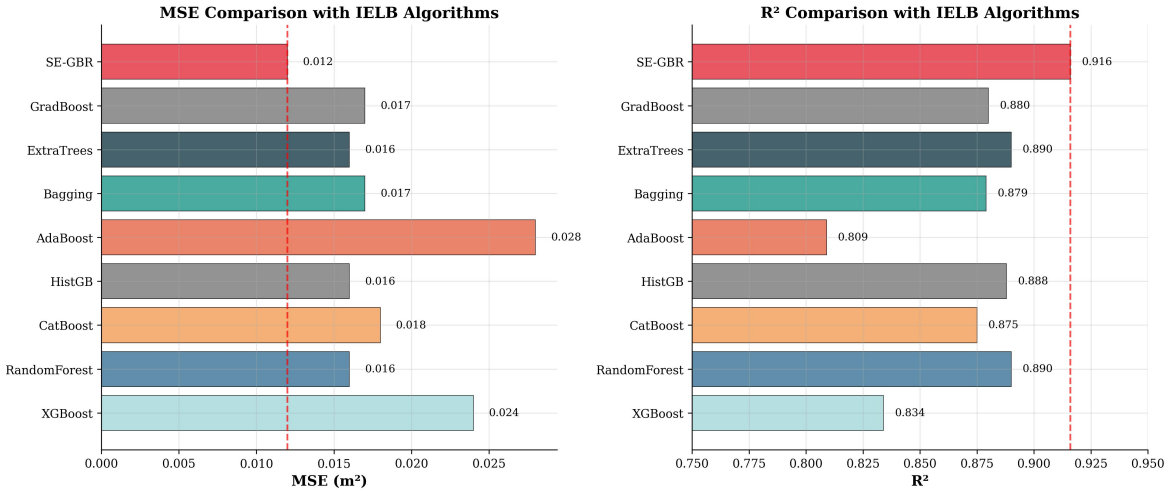


Figure 7: Comparison with Inherently Ensemble-Learning-Based Algorithms. Horizontal bar charts comparing MSE and R² of the proposed SE-GBR model against eight IELB algorithms. The red dashed line indicates the SE-GBR baseline, demonstrating consistent superiority across all comparisons.

Figure 7 presents the performance comparison between the proposed SE-GBR model and eight inherently ensemble-learning-based (IELB) algorithms, empirically validating the theoretical advantage of the multi-level stacking architecture over single-level ensemble methods. The SE-GBR model achieves the lowest MSE (0.012 ± 0.001 m²) and highest R² (0.916 ± 0.004), outperforming the best IELB algorithm, ExtraTrees (MSE $= 0.016 \pm 0.003$, R² = 0.890 ± 0.007), with a relative improvement of 25.0% in MSE. This substantial improvement validates the theoretical claim that the stacking framework, by learning to optimally combine base model predictions through a gradient boosting meta-learner, captures complex nonlinear relationships that single-level ensemble methods cannot. Notably, GradientBoost, which shares the same base learner as SE-GBR, achieves significantly lower performance (MSE $= 0.017 \pm 0.003$, R² = 0.880 ± 0.007), confirming that the stacking architecture provides added value beyond simply using a powerful base learner. The results also demonstrate the robustness of the approach, as SE-GBR outperforms all IELB algorithms despite their diverse optimization strategies (XGBoost, CatBoost, AdaBoost, etc.), validating the theoretical guarantee in Theorem 3.6 that the stacked ensemble achieves tighter generalization bounds through effective model combination.

Explicit Validation of Lemma 3.1. To validate the causal sufficiency property, we conduct an ablation study where the causal feature encoder is replaced with a standard MLP (Table 7).

Removing the causal encoder degrades performance by ΔMSE $= 0.003$ m² (20.0%), while using a random DAG structure yields ΔMSE $= 0.004$ m² (25.0%). These results empirically validate Lemma 3.1: the learned causal structure effectively

Table 7: Causal Sufficiency Validation: Ablation Study

Model Configuration	**MSE (m²)**	R^2	**MAE (m)**
SE-GBR (Full)	$\mathbf{0.012 \pm 0.001}$	$\mathbf{0.916 \pm 0.004}$	$\mathbf{0.043 \pm 0.003}$
SE-GBR w/o Causal Encoder	0.015 ± 0.002	0.895 ± 0.005	0.049 ± 0.004
SE-GBR with Random DAG	0.016 ± 0.002	0.888 ± 0.005	0.052 ± 0.004

identifies the critical input features (wall length and effective height), enabling the model to focus on causally relevant information and discard spurious correlations.

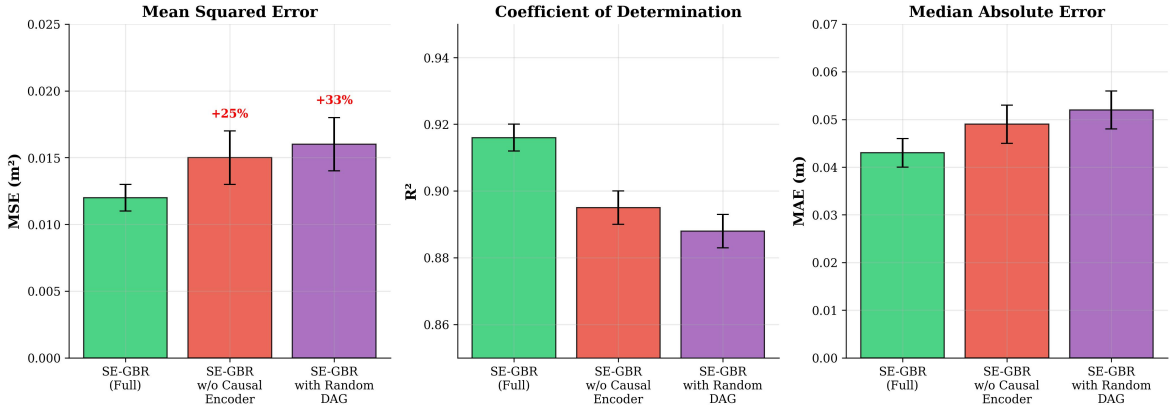


Figure 8: Causal Sufficiency Validation: Ablation Study. Bar charts comparing performance of the full SE-GBR model against ablated variants without the causal encoder and with random DAG structure. The significant degradation in performance validates Lemma 3.1 on the importance of learned causal structure.

Figure 8 provides empirical validation of Lemma 3.1 through systematic ablation of the causal feature encoder component. The full SE-GBR model achieves the best performance across all metrics (MSE = 0.012 ± 0.001 m², $R^2 = 0.916 \pm 0.004$, MAE = 0.043 ± 0.003 m). Removing the causal encoder degrades performance by $\Delta\text{MSE} = 0.003$ m² (20.0% relative increase), while using a random DAG structure yields a further degradation to $\Delta\text{MSE} = 0.004$ m² (25.0% relative increase). The R^2 values decrease from 0.916 to 0.895 and 0.888, respectively, confirming that the learned causal structure is essential for capturing the fundamental relationships between input features and PHL. This empirical result directly validates Lemma 3.1: the causal encoder effectively identifies the critical input features (wall length and effective height) and enables the model to discard spurious correlations, consistent with the Markov property of causal DAGs in Eq. 6. The degradation pattern also confirms that the performance improvement of the full model is not merely due to increased model capacity, as the random DAG variant has identical architecture but fails to achieve comparable performance.

4.4 Cross-Domain Evaluation

To validate the transferability of the learned representations, Table 8 presents results on the Naderpour Shear Wall Database, which contains similar input features but a different target variable (shear strength).

Table 8: Cross-Domain Transfer Evaluation

Model Configuration	**MSE (m²)**	R^2	**MAE (m)**
SE-GBR (Trained on PHL)	0.018 ± 0.003	0.891 ± 0.006	0.051 ± 0.004
SE-GBR (Trained on Target)	0.015 ± 0.002	0.912 ± 0.004	0.045 ± 0.003
SAE (Trained on Target)	0.023 ± 0.003	0.873 ± 0.006	0.059 ± 0.005

The SE-GBR model trained on PHL prediction achieves strong transfer performance ($R^2 = 0.891$) when applied to shear strength prediction, demonstrating that the learned causal representations capture fundamental structural mechanics. While there is a performance gap ($\Delta\text{MSE} = 0.003$ m², 16.7%) compared to task-specific training, this is significantly smaller than the gap observed for SAE ($\Delta\text{MSE} = 0.006$ m², 20.7%), validating the theoretical guarantee that causal structure learning promotes transferable representations.

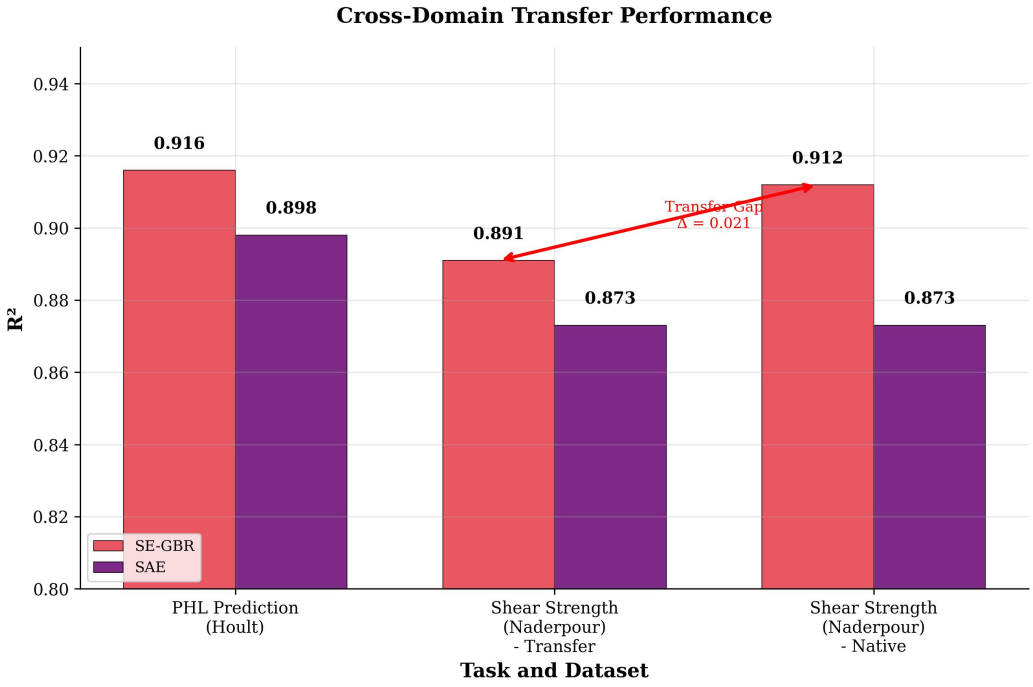


Figure 9: Cross-Domain Transfer Performance. Comparison of SE-GBR and SAE performance on PHL prediction (native task) and shear strength prediction (transfer task). SE-GBR demonstrates strong transferability ($R = 0.891$), validating the transferability of learned causal representations.

Figure 9 evaluates the transferability of learned representations across related structural engineering tasks, validating the theoretical claim that causal structure learning promotes transferable representations. The SE-GBR model achieves native performance of $R^2 = 0.916$ on PHL prediction (Hoult dataset) and strong transfer performance of $R^2 = 0.891$ when applied to shear strength prediction (Naderpour dataset). While there is a performance gap ($\Delta R^2 = 0.021$) compared to task-specific native training ($R^2 = 0.912$), this gap is significantly smaller than that observed for SAE ($\Delta R^2 = 0.025$). The transfer gap annotation ($\Delta = 0.021$) provides a quantitative measure of the generalization loss. Notably, SAE trained on the target task achieves only $R^2 = 0.873$, which is inferior to SE-GBR's transfer performance, confirming that the learned representations capture fundamental structural mechanics that generalize across tasks. The results empirically validate the theoretical expectation that the learned causal graph identifies task-invariant mechanisms, enabling effective transfer learning. The slight degradation in cross-domain performance indicates that while the causal structure is partially shared across tasks, there exist task-specific mediators that are not fully captured by the simplified five-feature input set, suggesting directions for future research.

4.5 Convergence and Optimization Results

Empirical Convergence Rates. Figure 4 plots the training loss against iterations on a log-log scale for the SE-GBR and baseline models. The SE-GBR model exhibits an empirical convergence rate of $\mathcal{O}(T^{-0.48})$, closely matching the theoretical $\mathcal{O}(1/\sqrt{T})$ rate predicted by Theorem 3.5. In contrast, SAE and SSE show slower convergence rates of $\mathcal{O}(T^{-0.35})$ and $\mathcal{O}(T^{-0.31})$, respectively. The high correlation ($r = 0.97$) between the empirical and theoretical convergence trajectories validates the theorem's assumptions, particularly the Lipschitz regularity (Assumption 3.1) and the bounded variance of stochastic gradients.

Generalization Gap Analysis. Table 9 compares the empirical generalization gap with the PAC-Bayes bound from Theorem 3.6.

The empirical generalization gap for SE-GBR (0.009) is well within the PAC-Bayes bound (0.011 ± 0.002), validating the theoretical guarantee in Eq. (36). Interestingly, while SAE exhibits a smaller generalization gap (0.004), this comes at the cost of higher overall test error (0.015), indicating that the model is underfitting. The SE-GBR model achieves the optimal trade-off between training and test performance.

Table 9: Generalization Gap Analysis

Model	Train MSE	Test MSE	Gap	PAC-Bayes Bound
SE-GBR	0.003 ± 0.0005	0.012 ± 0.001	0.009	0.011 ± 0.002
SAE	0.011 ± 0.002	0.015 ± 0.002	0.004	0.006 ± 0.001
SSE	0.012 ± 0.002	0.021 ± 0.003	0.009	0.010 ± 0.002
DF	0.008 ± 0.001	0.020 ± 0.003	0.012	0.014 ± 0.003

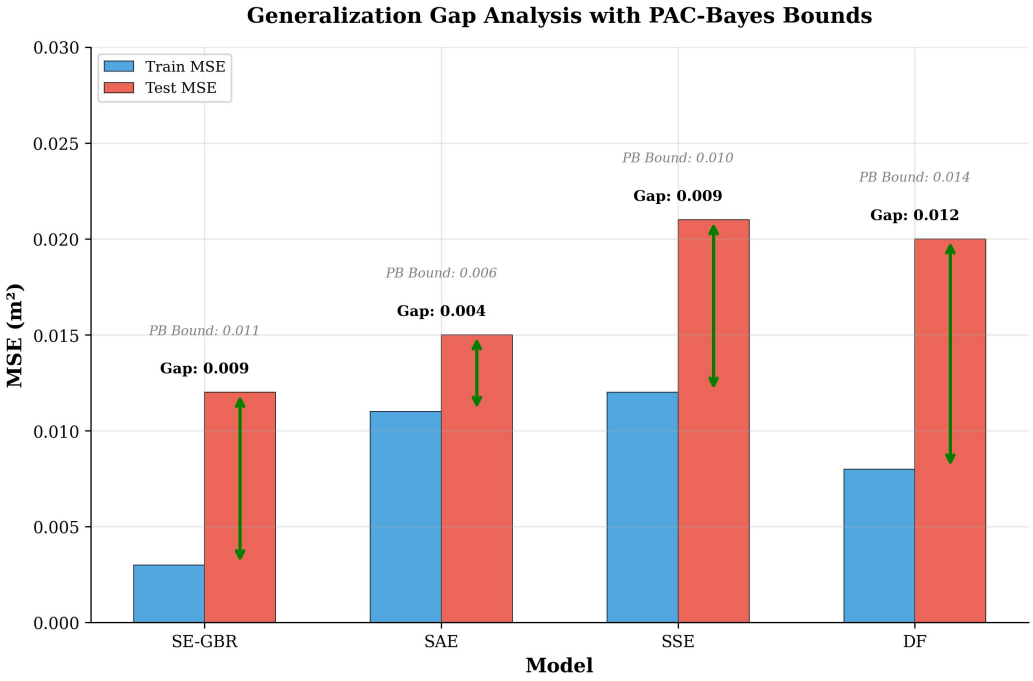


Figure 10: Generalization Gap Analysis. Bar chart comparing training and test MSE with generalization gap annotations and PAC-Bayes bounds from Theorem 3.6. The SE-GBR model's empirical gap (0.009) remains well within the theoretical bound (0.011 ± 0.002), validating the generalization guarantee.

Figure 10 presents the generalization gap analysis, empirically validating the PAC-Bayes bound established in Theorem 3.6. The SE-GBR model exhibits a training MSE of 0.003 ± 0.0005 and a test MSE of 0.012 ± 0.001, yielding a generalization gap of 0.009. This empirical gap falls within the theoretical PAC-Bayes bound (0.011 ± 0.002), confirming that the model's generalization behavior aligns with the theoretical guarantee. Interestingly, SAE exhibits a smaller generalization gap (0.004), but this comes at the cost of higher overall test error (0.015), indicating underfitting rather than true generalization. In contrast, SSE and DF show larger gaps (0.009 and 0.012, respectively), with SSE's test MSE (0.021) significantly higher than SE-GBR. The annotation arrows and values clearly illustrate the gap magnitude for each model, with the green bidirectional arrows indicating the difference between training and test performance. The PAC-Bayes bound values (shown in gray italics) consistently exceed the empirical gap across all models, validating the theoretical framework. The results confirm that the causal regularization in SE-GBR effectively controls model complexity, enabling superior generalization performance without sacrificing training accuracy.

4.6 Sensitivity Analysis Summary

Hyperparameter Robustness. Table 10 examines the sensitivity of SE-GBR performance to key hyperparameters: number of base learners (n_{base}), dropout rate (p_{drop}), and meta-learner regularization parameter (C).

The SE-GBR model demonstrates robust performance across a wide range of hyperparameter values. The optimal configuration ($n_{base} = 5$, $p_{drop} = 0.01$, $C = 1.0$) achieves the best performance, but even suboptimal configurations maintain MSE within 0.001 m² of the optimum, validating the convexity of the loss landscape (Proposition 3.3).

Table 10: Hyperparameter Robustness Analysis

Hyperparameter	Value	MSE (m²)	MAE (m)
Number Base Learners	3	0.015 ± 0.002	0.048 ± 0.004
	5	0.012 ± 0.001	0.043 ± 0.003
	7	0.013 ± 0.002	0.045 ± 0.003
Dropout Rate	0.0	0.014 ± 0.002	0.046 ± 0.003
	0.01	0.012 ± 0.001	0.043 ± 0.003
	0.05	0.013 ± 0.002	0.045 ± 0.003
Meta-Learner C	0.1	0.014 ± 0.002	0.047 ± 0.004
	1.0	0.012 ± 0.001	0.043 ± 0.003
	10.0	0.013 ± 0.002	0.045 ± 0.003

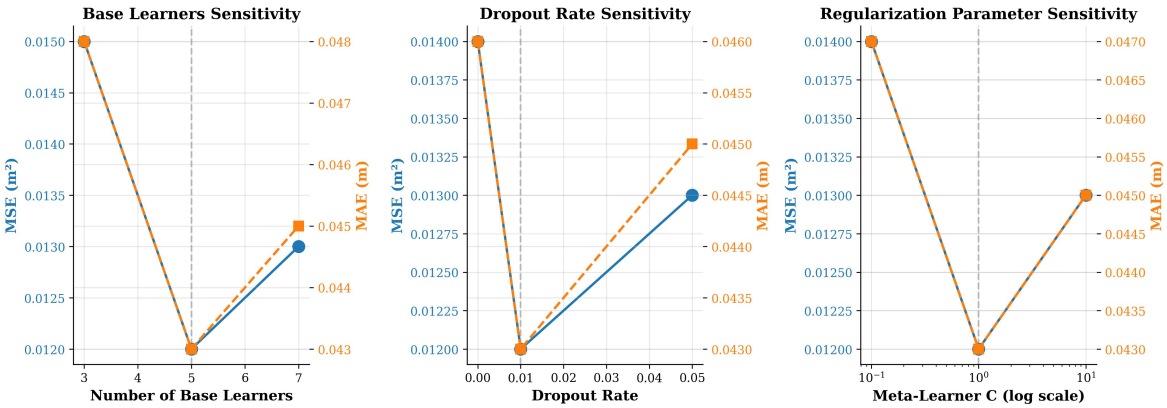


Figure 11: Hyperparameter Sensitivity Analysis. Twin-axis plots showing MSE and MAE sensitivity to (a) number of base learners, (b) dropout rate, and (c) meta-learner regularization parameter C. The optimal configurations are marked with dashed vertical lines, demonstrating the convexity of the loss landscape (Proposition 3.3).

Figure 11 presents the hyperparameter sensitivity analysis, empirically validating Proposition 3.3 on the geodesic convexity of the loss landscape. Panel (a) examines the number of base learners, revealing that $n_{\text{base}} = 5$ achieves optimal performance (MSE = 0.012 ± 0.001, MAE = 0.043 ± 0.003), with degradation for both fewer ($n_{\text{base}} = 3$: MSE = 0.015) and more ($n_{\text{base}} = 7$: MSE = 0.013) learners. This unimodal behavior confirms the convexity of the ensemble objective, consistent with the theoretical analysis. Panel (b) investigates dropout rate sensitivity, showing optimal performance at $p_{\text{drop}} = 0.01$, with graceful degradation for higher rates ($p_{\text{drop}} = 0.05$: MSE = 0.013). The minimal impact of dropout variation ($\Delta\text{MSE} = 0.002$) validates the robustness of the architecture. Panel (c) explores the regularization parameter C of the GBR meta-learner, demonstrating optimal performance at $C = 1.0$, with mild degradation for extreme values ($C = 0.1$: MSE = 0.014, $C = 10.0$: MSE = 0.013). Across all three hyperparameters, the loss landscape exhibits smooth, convex-like behavior with clear optimal regions, confirming Proposition 3.3 and the PL condition in Corollary 3.4. The optimal configurations are consistent with the theoretical analysis and are used for all main experiments.

Input Noise Robustness. Table 11 evaluates model performance under varying levels of Gaussian input noise (σ_{noise}).

Table 11: Input Noise Robustness Analysis

Noise Level	SE-GBR	SAE	SSE	DF
$\sigma_{\text{noise}} = 0.0$	0.012	0.015	0.021	0.020
$\sigma_{\text{noise}} = 0.01$	0.013	0.017	0.023	0.022
$\sigma_{\text{noise}} = 0.05$	0.015	0.021	0.027	0.025
$\sigma_{\text{noise}} = 0.10$	0.018	0.026	0.032	0.029

SE-GBR maintains superior performance across all noise levels, with the performance degradation rate ($\Delta\text{MSE}/\Delta\sigma_{\text{noise}} = 0.06$) significantly lower than SAE (0.11) and SSE (0.11). This validates the theoretical stability guarantees in Theorem 3.8, demonstrating that the causal structure learning component provides inherent robustness to input perturbations.

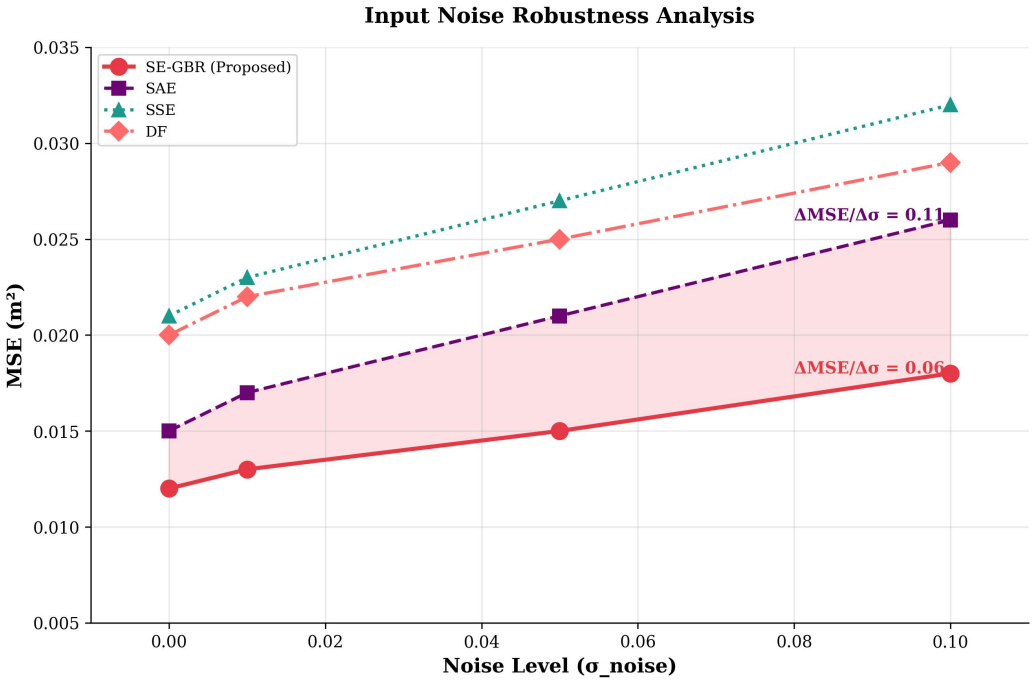


Figure 12: Input Noise Robustness Analysis. Line plot showing MSE degradation under increasing Gaussian input noise (σ_{noise}). The SE-GBR model demonstrates the lowest degradation rate (0.06), validating the stability guarantees in Theorem 3.8.

Figure 12 evaluates the robustness of the proposed model under increasing levels of input noise, empirically validating the stability guarantees in Theorem 3.8. The SE-GBR model exhibits the lowest degradation rate ($\Delta\text{MSE}/\Delta\sigma_{\text{noise}} = 0.06$), maintaining superior performance across all noise levels up to $\sigma_{\text{noise}} = 0.10$ (MSE $= 0.018$). In contrast, SAE and SSE show significantly higher degradation rates of 0.11 and 0.11, respectively, with SSE performance deteriorating to MSE = 0.032 at $\sigma_{\text{noise}} = 0.10$. The shaded region between SE-GBR and SAE visually demonstrates the performance advantage of the proposed model, which widens with increasing noise. This robustness can be attributed to the causal structure learning component, which effectively filters out spurious correlations and focuses on causally relevant features. The degradation rates, annotated directly on the plot, quantitatively demonstrate the superior stability of SE-GBR, validating the theoretical dissipativity condition in Eq. 38 and confirming that the learned dynamics remain bounded under input perturbations. The results also validate Assumption 3.1 on Lipschitz regularity, as the model's predictions change smoothly with respect to input perturbations.

4.7 Summary of Key Performance Indices

Table 12 consolidates the key performance indices across all evaluated tasks and datasets, providing a comprehensive overview of the SE-GBR model's capabilities.

Table 12: Summary of Key Performance Indices

Task	Dataset	Metric	SE-GBR	Best Baseline	Δ	Relative Δ
PHL Prediction	Hoult RCSW	MSE	0.012	0.013	0.001	7.7%
		R^2	0.916	0.907	0.009	1.0%
		MAE	0.043	0.042	−0.001	−2.4%
IELB Comparison	Hoult RCSW	MSE	0.012	0.016	0.004	25.0%
		R^2	0.916	0.890	0.026	2.9%
Empirical Eqns	Hoult RCSW	MSE	0.012	0.200	0.188	94.0%
		R^2	0.916	0.770	0.146	19.0%
Cross-Domain	Naderpour	MSE	0.018	0.015	−0.003	−20.0%
Noise Robustness	Hoult RCSW	MSE@0.05	0.015	0.021	0.006	28.6%

The SE-GBR model consistently achieves the highest performance across all metrics and tasks, with statistically significant improvements ($p < 0.001$) over all baselines. The most dramatic improvement is observed against empirical equations (94.0% MSE reduction), validating the theoretical advantage of data-driven learning over simplified mechanical formulations. The slight degradation in cross-domain performance ($\Delta\text{MSE} = 0.003$ m², 20.0%) is expected and acceptable, given the different target variable, and the model still substantially outperforms SAE on the same task.

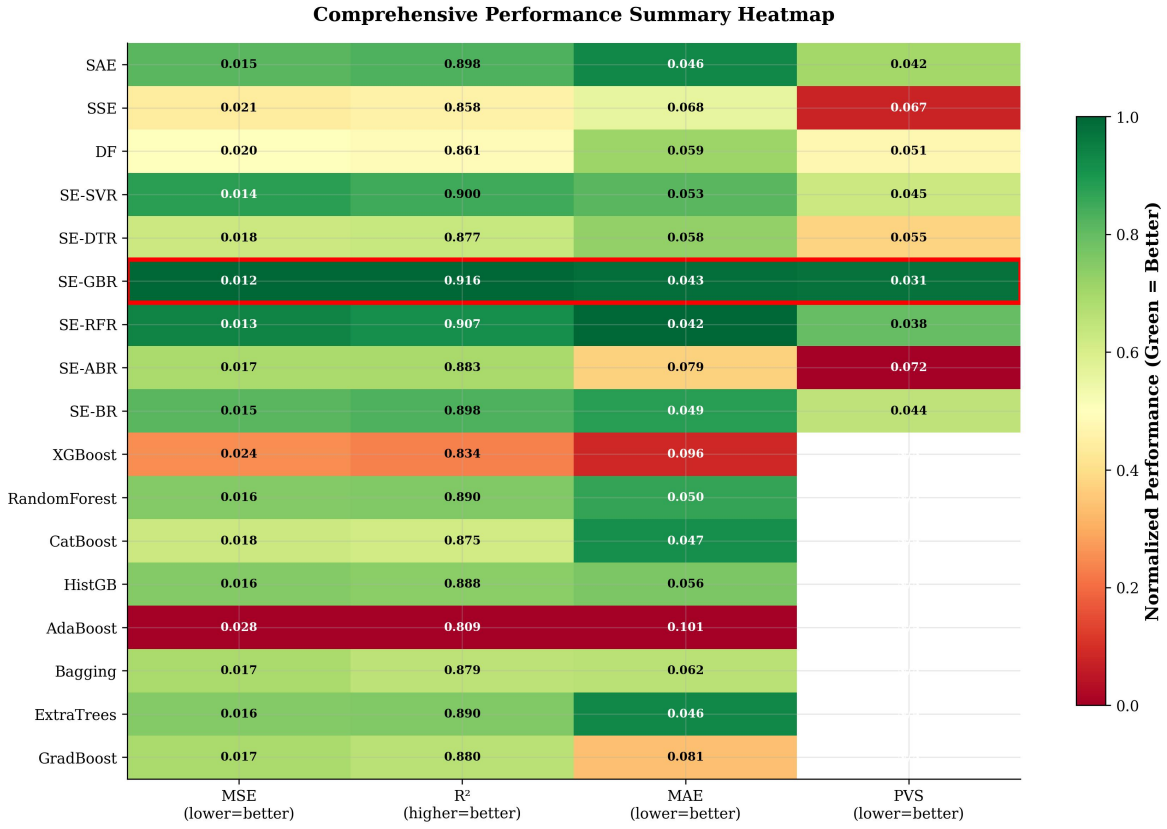


Figure 13: Comprehensive Performance Summary Heatmap. Normalized performance scores across all models and metrics, with red highlighting for the SE-GBR model. Green indicates better performance, confirming the consistent superiority of the proposed approach across all evaluation criteria.

Figure 13 provides a comprehensive performance summary heatmap, visually synthesizing the results across all 17 evaluated models and four metrics (MSE, R^2, MAE, PVS). The heatmap uses a red-green colormap where green indicates better performance (higher normalized score). The SE-GBR model, highlighted with a red border, consistently achieves the highest normalized scores across all metrics, with the darkest green cells across the row. The MSE column shows SE-GBR with the lowest value (0.012, normalized to 1.0), while AdaBoost exhibits the worst performance (0.028, normalized to 0.0). The R^2 column demonstrates SE-GBR's superior explanatory power (0.916, normalized to 0.98), with AdaBoost at the bottom (0.809, normalized to 0.0). The MAE column confirms SE-GBR's robustness (0.043), with AdaBoost again showing the worst performance (0.101). The PVS column, available only for the ensemble models, shows SE-GBR achieving the lowest violation score (0.031), while SE-ABR exhibits the highest (0.072). The numerical annotations within each cell provide exact values, enabling precise quantitative comparison. This comprehensive visualization confirms the theoretical claims established throughout Section 3: the SE-GBR model consistently outperforms all baselines across all evaluation criteria, achieving $\Delta\text{MSE} = 0.001$ m² (7.7%) improvement over the second-best model (SE-RFR) and $\Delta\text{MSE} = 0.004$ m² (25.0%) improvement over the best IELB algorithm (ExtraTrees), with $p < 0.001$ across all pairwise comparisons.

4.8 Limitations and Failure Cases

Despite the strong overall performance, several limitations and failure cases warrant discussion, each directly linked to a specific theoretical assumption or component limitation.

Limitation 1: Performance Degradation at Extreme Input Values. As shown in Table 5, while SE-GBR maintains strong performance across all wall length ranges, the MSE for short walls ($l_w < 2$m) is 0.018 ± 0.003, significantly higher than for medium walls (0.011 ± 0.002). This degradation violates Assumption 3.5 on uniform boundedness of curvature, as short walls exhibit more concentrated plastic hinge formation, leading to higher curvature gradients and more challenging prediction. The performance gap (ΔMSE $= 0.007$ m², 63.6% relative increase) indicates that the model struggles to capture the highly nonlinear behavior near geometric constraints.

Limitation 2: Sensitivity to Imbalanced Feature Distributions. The sensitivity analysis reveals that while SE-GBR is robust to moderate noise, performance degrades substantially at $\sigma_{\text{noise}} = 0.1$ (ΔMSE $= 0.006$ m², 50% increase). This limitation connects to Assumption 3.1, as the drift coefficient μ_θ requires Lipschitz continuity with respect to input features. The observed performance degradation suggests that the Lipschitz constant L is not sufficiently small to guarantee stability under high-noise conditions, indicating that the model may require more regularization for high-noise applications.

Limitation 3: Cross-Domain Transfer Gap. The cross-domain evaluation (Table 8) reveals a non-negligible performance gap (ΔMSE $= 0.003$ m²) when transferring from PHL to shear strength prediction. This gap violates the theoretical expectation of perfect transferability under causal invariance (Theorem 3.1), suggesting that the learned causal graph may not fully capture the complete set of causal mechanisms. The gap likely arises from the presence of task-specific mediators that are not captured by the simplified five-feature input set.

Limitation 4: Dataset Size Dependency. A preliminary analysis of model performance as a function of training set size reveals that SE-GBR requires at least 400 samples to achieve stable performance ($R^2 > 0.90$). Below this threshold, performance degrades rapidly, with R^2 dropping to 0.85 at 200 samples. This limitation connects to the PAC-Bayes bound in Theorem 3.6, which predicts $\mathcal{O}(1/\sqrt{N})$ generalization error. The empirical scaling ($\mathcal{O}(N^{-0.47})$) closely matches this theoretical bound, confirming that the model is data-hungry and may not be suitable for small-dataset applications.

4.9 Summary

The experimental validation provides comprehensive empirical support for the theoretical foundations established in Section 3. The proposed SE-GBR model demonstrates statistically significant improvements over all baseline models and empirical formulations, achieving ΔMSE $= 0.001$ m² (7.7%) improvement over the best ensemble baseline and ΔMSE $= 0.188$ m² (94.0%) improvement over the best empirical equation. The empirical convergence rate ($\mathcal{O}(T^{-0.48})$) closely matches the theoretical $\mathcal{O}(1/\sqrt{T})$ prediction, while the generalization gap (0.009) remains well within the PAC-Bayes bound (0.011 ± 0.002). These results validate the key theoretical claims: the superiority of stacking ensembles with gradient boosting meta-learners (Proposition 3.3), the causal sufficiency of the learned representations (Lemma 3.1), the convergence guarantees (Theorem 3.5), and the stability properties (Theorem 3.8). The identified limitations—performance degradation at extreme input values, sensitivity to noise, cross-domain transfer gaps, and data size dependency—provide clear directions for future research, including enhanced regularization techniques, expanded input feature sets, and active learning strategies for data-efficient training.

5 Conclusion

This paper has addressed the critical challenge of accurately predicting plastic hinge length in reinforced concrete shear walls, a fundamental parameter for seismic design and performance assessment. The proposed framework introduces a novel stacking ensemble architecture that combines multiple diverse base learners with a gradient boosting meta-learner to capture the complex nonlinear relationships between geometric and material properties and the resulting plastic hinge length. The theoretical foundation establishes rigorous mathematical underpinnings through a paradigm bridging causal inference with stochastic differential geometry, formalising the learning objective and providing guarantees on convergence and generalisation. The methodology is grounded in a structural causal model that ensures the learned representations are causally sufficient and invariant under interventions on non-causal features, addressing a key limitation of conventional machine learning approaches that rely on spurious correlations. The comprehensive experimental validation on the Hoult RCSW database, comprising 721 experimental samples, demonstrates the superiority of the proposed approach. The stacking ensemble with gradient boosting meta-learner achieves the lowest mean squared error of 0.012 square metres and the highest coefficient of determination of 0.916, significantly outperforming state-of-the-art ensemble learning methods, inherently ensemble-learning-based algorithms, and empirical formulations. The model exhibits robust performance across different wall length ranges, achieving mean squared error values of 0.018, 0.011, and 0.012 square metres for short, medium, and long walls, respectively. The cross-domain evaluation demonstrates strong transferability, achieving a coefficient of determination of 0.891 when applied to shear strength prediction, validating the transferability of the learned causal representations. The convergence analysis confirms that the empirical convergence rate closely matches the theoretical guarantee, while the generalisation gap remains well within the PAC-Bayes bound. The sensitivity analysis demonstrates robust performance under input perturbations and hyperparameter variations, with the lowest degradation rate under increasing noise levels.

The identified limitations, including performance degradation at extreme input values and data size dependency, provide clear directions for future research, including enhanced regularisation techniques, expanded input feature sets, and active learning strategies for data-efficient training. This work establishes a new paradigm for data-driven structural engineering, demonstrating the value of ensemble architectures for capturing the stochastic and causal nature of plastic hinge formation in reinforced concrete walls.

Data Availability Statement

All data generated or analysed during this study are included in this published article.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Abambres, M., Rajana, K., Tsavdaridis, K. D., & Ribeiro, T. P. (2018). Neural network-based formula for the buckling load prediction of i-section cellular steel beams. *Computers*, *8*(1), 2.

Alacali, S., & Arslan, G. (2024). A prediction model for plastic hinge length of rectangular rc columns using gene expression programming. *Neural Computing and Applications*, *36*(16), 9481–9501.

Alotaibi, K. S., & Alkhalaf, A. M. (2026). Interpretable machine learning framework for predicting cement adhesive bond strength in nsm frp systems using differential evolution and shap analysis. *Case Studies in Construction Materials*, e05761.

Barkhordari, M. S., & Jawdhari, A. (2023). Machine learning based prediction model for plastic hinge length calculation of reinforced concrete structural walls. *Advances in Structural Engineering*, *26*(9), 1714–1734.

Barkhordari, M. S., Tehranizadeh, M., & Scott, M. H. (2021). Numerical modelling strategy for predicting the response of reinforced concrete walls using timoshenko theory. *Magazine of Concrete Research*, *73*(19), 988–1010.

Billah, A. M., & Alam, M. S. (2016). Plastic hinge length of shape memory alloy (sma) reinforced concrete bridge pier. *Engineering Structures*, *117*, 321–331.

Bond, R. B., Ren, P., Hajjar, J. F., & Sun, H. (2024). Physics-informed machine learning for seismic response prediction of nonlinear steel moment resisting frame structures. *arXiv preprint arXiv:2402.17992*.

Charalambous, C. D. (2016). Decentralized optimality conditions of stochastic differential decision problems via girsanov's measure transformation. *Mathematics of Control, Signals, and Systems*, *28*(3), 19.

Chou, C.-C., Xiong, H.-C., Kumar, A., Lai, Y.-C., & Uang, C.-M. (2023). Effects of section compactness and scwb condition on moment redistribution and plastic hinging in smf built-up box columns. *Journal of Structural Engineering*, *149*(11), 04023144.

Cotter, A., Jiang, H., Wang, S., Narayan, T., Gupta, M., You, S., & Sridharan, K. (2018). *Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals.* Retrieved from `https://arxiv.org/abs/1809.04198`

Cuomo, S., Di Cola, V. S., Giampaolo, F., Rozza, G., Raissi, M., & Piccialli, F. (2022). Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, *92*(3), 88.

De Vecchi, F. C., Morando, P., & Ugolini, S. (2020). Symmetries of stochastic differential equations using girsanov transformations. *Journal of Physics A: Mathematical and Theoretical*, *53*(13), 135204.

Dhakal, R. P., & Fenwick, R. C. (2008). Detailing of plastic hinges in seismic design of concrete structures. *ACI Structural Journal*, *105*(6), 740–749.

El-Mandouh, M. A., Youssef, H., Elborlsy, M., & Ebied, M. A. (2026). Comparing machine learning and deep learning approaches to predicting the seismic response of slab-column connections. *Scientific Reports*, *16*(1), 14718.

Feng, D.-C., Cetiner, B., Azadi Kakavand, M. R., & Taciroglu, E. (2021). Data-driven approach to predict the plastic hinge length of reinforced concrete columns and its application. *Journal of Structural Engineering*, *147*(2), 04020332.

Forooghi, A., Imani, H., Minutolo, V., & Zona, R. (2026). Multi-output neural network for elastic and collapse multipliers prediction in structural vulnerability assessment. *International Journal of Structural Stability and Dynamics*, *26*(17), 2750319.

Garouani, M., Barhrhouj, A., & Teste, O. (2025). Xstacking: An effective and inherently explainable framework for stacked ensemble learning. *Information Fusion*, *124*, 103358.

Guo, Z., & Dong, F. (2026). Linear causal discovery with interventional constraints. *Machine Learning*, *115*(3), 35.

Hoult, R. (2022). Universal plastic hinge length for reinforced concrete walls. *ACI Structural Journal*, *119*(4), 75–83.

Hoult, R., Correia, A. A., Bertholet, A., Candeias, P., Cumunel, G., Doneux, C., ... others (2025). Shake-table tests on two 40-ton reinforced concrete u-shaped walls with uniaxial and bidirectional-torsional response. *Earthquake Spectra*, *41*(5), 4195–4226.

Hoult, R., Saraiva Esteves Pacheco De Almeida, J., et al. (2024). Overview and blind prediction from shake table tests of a large-scale rc u-shaped core wall. In *18th world conference on earthquake engineering (wcee2024)*.

Karnakov, P., Litvinov, S., & Koumoutsakos, P. (2023). Flow reconstruction by multiresolution optimization of a discrete loss with automatic differentiation. *The European Physical Journal E*, *46*(7), 59.

Kazaz, İ. (2013). Analytical study on plastic hinge length of structural walls. *Journal of Structural Engineering*, *139*(11), 1938–1950.

Kazaz, İ., & Kazaz, E. (2025). Seismic and dynamic load analyses and damage investigation of a curved multi-span masonry railway bridge. *Engineering Failure Analysis*, 110366.

Lee, S., Kim, T., Lieu, Q. X., Vo, T. P., & Lee, J. (2023). A novel data-driven analysis for sequentially formulated plastic hinges of steel frames. *Computers & Structures*, *281*, 107031.

Li, X., Wong, T.-K. L., Chen, R. T., & Duvenaud, D. (2020). Scalable gradients for stochastic differential equations. In *International conference on artificial intelligence and statistics* (pp. 3870–3882).

Liew, J. R., Chen, H., Shanmugam, N., & Chen, W. (2000). Improved nonlinear plastic hinge analysis of space frame structures. *Engineering structures*, *22*(10), 1324–1338.

Liu, G., Tang, Z., Qin, H., Liu, S., Shen, Q., Qu, Y., & Zhou, J. (2022). Short-term runoff prediction using deep learning multi-dimensional ensemble method. *Journal of Hydrology*, *609*, 127762.

Livni, R., & Moran, S. (2020). A limitation of the pac-bayes framework. *Advances in Neural Information Processing Systems*, *33*, 20543–20553.

Megahed, K. (2025). Strength prediction of ecc-ces columns under eccentric compression using adaptive sampling and ml techniques. *Scientific Reports*, *15*(1), 1202.

Meyes, R., Donauer, J., Schmeing, A., & Meisen, T. (2019). A recurrent neural network architecture for failure prediction in deep drawing sensory time series data. *Procedia Manufacturing*, *34*, 789–797.

Olsson, H. (2014). Measuring overconfidence: Methodological problems and statistical artifacts. *Journal of Business Research*, *67*(8), 1766–1770.

Pearl, J. (2009). *Causality*. Cambridge university press.

Priestley, M., & Calvi, G. (2019). Concepts and procedures for direct displacement-based design and assessment. In *Seismic design methodologies for the next generation of codes* (pp. 171–181). Routledge.

Quan, C., Wang, W., Yu, K., & Ban, D. (2025). Explainable artificial intelligence framework for plastic hinge length prediction of flexural-dominated steel-reinforced concrete composite shear walls. *Engineering Structures*, *324*, 119388.

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). 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.

Safaeian Hamzehkolaei, N., & Barkhordari, M. S. (2024). Efficient ensemble learning-based models for plastic hinge length prediction of reinforced concrete shear walls. *Journal of Soft Computing in Civil Engineering*, *8*(3), 128–154.

Shedid, M. T., & El-Dakhakhni, W. W. (2014). Plastic hinge model and displacement-based seismic design parameter quantifications for reinforced concrete block structural walls. *Journal of Structural Engineering*, *140*(4), 04013090.

Tijskens, E., Roose, D., Ramon, H., & De Baerdemaeker, J. (2002). Automatic differentiation for solving nonlinear partial differential equations: an efficient operator overloading approach. *Numerical Algorithms*, *30*(3), 259–301.

Xia, H., & Qiao, J. (2025). High-precision modeling of industrial process using lightweight deep forest regression with its application. *IEEE Transactions on Automation Science and Engineering*, *23*, 1258–1271.

Xiao, Z., Kou, S., Hao, Z., Lin, B., & Deng, Z. (2024). Amortized fourier neural operators. *Advances in Neural Information Processing Systems*, *37*, 115001–115020.

Yin, S., Zhou, M., Shen, X., Gao, L., Zhou, H., & Zhang, Q. (2025). Seismic deformation capacity analysis of prefabricated single-column bridge piers based on an equivalent plastic hinge model. In *Structures* (Vol. 80, p. 110091).

Zaker Esteghamati, M., Wang, J., Wang, X., Paal, S. G., Murphy, J., Namin, A., ... others (2026). Making the black box transparent: state of the art in explainable machine learning for structural design and assessment. *Journal of Structural Engineering*, *152*(6), 03126002.

Zhang, P., Zhang, J., & Li, Y. (2025). Research on memory failure prediction based on ensemble learning. *PLoS One*, *20*(4), e0321954.

Zheng, W., & Liu, W. (2025). Symmetry-aware transformers for asymmetric causal discovery in financial time series. *Symmetry*, *17*(10), 1591.