Sparse Symbolic Surrogates for Path-Dependent Clay Constitutive Response

Anton Peters[1]

[1] Department of Forest Botany and Tree Physiology, Faculty of Forest Sciences and Forest Ecology, University of Göttingen, Göttingen, Lower Saxony, Germany

Abstract

Constitutive modeling of clay remains one of the most challenging areas in computational geomechanics due to the highly nonlinear and path-dependent nature of soil deformation. Traditional elastoplastic and viscoplastic models require complex numerical integration schemes that are computationally expensive and prone to convergence issues in large-scale finite element simulations. While deep learning models such as neural networks have emerged as popular surrogate models to bypass these computational bottlenecks, their black-box nature obscures physical interpretability and limits their reliable extrapolation to unseen stress paths. This paper introduces a novel framework utilizing sparse symbolic regression to discover interpretable, algebraic surrogate models for the path-dependent constitutive response of clay. By leveraging sparse identification algorithms, we extract parsimonious differential equations that accurately map strain increments to stress increments while maintaining internal state variables. The methodology generates an extensive synthetic dataset using advanced bounding surface plasticity models to train the symbolic surrogate. Optimization is performed through a sequentially thresholded least-squares approach with sparsity-promoting regularization. The resulting symbolic surrogates not only drastically reduce computational evaluation times but also provide explicit mathematical expressions that reflect underlying physical mechanisms such as plastic hardening and volumetric dilation. Comparative analyses demonstrate that the proposed sparse symbolic surrogates rival the accuracy of deep recurrent neural networks while offering superior generalization capabilities and strictly enforcing physical constraints.

Keywords: Symbolic Regression; Constitutive Modeling; Path-Dependency; Machine Learning; Geomechanics

1. Introduction

1.1 Background on Clay Constitutive Modeling

The mechanical behavior of clayey soils is fundamentally governed by a complex interplay of particle mineralogy, pore fluid interaction, and stress history. Unlike conventional engineering materials such as steel or concrete, clay exhibits profound nonlinearity, pressure dependency, and a strong coupling between volumetric and deviatoric responses. Over the past several decades, the geotechnical engineering community has dedicated immense effort to developing mathematical frameworks capable of capturing these intricate phenomena. Classical theories, most notably the Critical State Soil Mechanics framework formulated in the mid-twentieth century, established the foundational principles linking void ratio, mean effective stress, and deviatoric stress. These principles culminated in models that have been widely implemented in numerical codes to predict settlement, slope stability, and soil-structure interaction. However, as engineering demands have grown to encompass phenomena such as cyclic loading from offshore structures, seismic events, and complex deep excavations, the limitations of these classical models have become increasingly apparent. Researchers have subsequently proposed advanced frameworks including bounding surface plasticity, generalized plasticity, and subloading surface concepts to address phenomena like degradation, kinematic hardening, and small-strain stiffness [1]. While these advanced formulations provide high fidelity, they introduce a myriad of internal state variables and phenomenological parameters that complicate both calibration and numerical implementation. The mathematical structure of these advanced models typically involves highly nonlinear sets of differential equations that require implicit or explicit integration schemes at every integration point in a finite element analysis, leading to significant computational overhead.

1.2 Path-Dependency in Geomechanics

A defining characteristic of soil behavior is its path-dependency, which dictates that the current state of stress and the corresponding material stiffness depend not only on the current state of strain but also on the complete historical trajectory of deformation. When clay is subjected to cycles of loading, unloading, and reloading, the microstructural arrangement of the clay particles undergoes irreversible changes. Plastic strains accumulate, pore pressures generate and dissipate, and the yield surface continuously evolves in both size and position within the stress space [2]. This path-dependency implies that constitutive models must rigorously track internal memory variables. In classical elastoplasticity, this is achieved through hardening rules that dictate the evolution of the yield surface. However, tracking this historical evolution numerically requires integrating a set of stiff ordinary differential equations [3]. During complex load reversals, the material response can transition abruptly from a soft plastic state to a stiff elastic state, creating severe discontinuities in the material tangent stiffness matrix. These discontinuities represent a primary source of non-convergence in global Newton-Raphson iterative solvers utilized in standard finite element programs. Furthermore, the integration algorithms themselves, such as the widely used closest point projection method, require iterative local solutions at each integration point. For large-scale three-dimensional boundary value problems containing millions of degrees of freedom, the time spent evaluating the constitutive model and its corresponding algorithmic tangent can account for an overwhelming majority of the total computational cost [4]. The need for accurate yet computationally efficient representations of path-dependent behavior is therefore a paramount concern in modern computational geomechanics.

1.3 Motivation for Surrogate Models

In response to the computational bottlenecks imposed by advanced phenomenological models, the scientific community has increasingly explored surrogate modeling techniques. A surrogate model, or metamodel, aims to approximate the input-output mapping of a complex physical model at a fraction of the computational cost [5]. In recent years, data-driven approaches leveraging artificial intelligence and machine learning have gained tremendous traction. Deep neural networks, particularly recurrent neural networks and long short-term memory networks, have been trained on vast datasets of stress-strain trajectories to predict soil responses [6]. These deep learning surrogates have demonstrated remarkable predictive accuracy within their training domains and execute significantly faster than traditional numerical integration schemes [7]. However, the adoption of deep neural networks in safety-critical geotechnical engineering applications faces significant resistance due to their inherent lack of transparency. These models function as black boxes containing thousands or millions of uninterpretable weights and biases [8]. Consequently, it is exceedingly difficult to verify whether a deep learning model has learned the true underlying physical principles or has merely overfit to spurious correlations in the training data. More importantly, black-box models often fail catastrophically when presented with loading paths that extrapolate outside their training distribution, leading to non-physical predictions such as negative effective stresses or violations of thermodynamic laws. This critical limitation has spurred interest in developing grey-box models or interpretable machine learning techniques [9]. Sparse symbolic regression emerges as a highly promising alternative, offering the ability to distill massive datasets into parsimonious, human-readable algebraic equations. By identifying the underlying governing equations directly from data, sparse symbolic surrogates promise to bridge the gap between computational efficiency and physical interpretability, ensuring that the resulting models remain robust, reliable, and mathematically transparent.

2. Literature Review and Theoretical Background

2.1 Traditional Constitutive Models for Clay

The evolution of constitutive modeling for clay soils is a testament to the ongoing quest to balance physical accuracy with mathematical tractability. Early attempts to model soil deformation relied on linear elasticity and perfect plasticity, utilizing failure criteria such as Mohr-Coulomb or Drucker-Prager [10]. While mathematically simple, these models completely failed to capture the volumetric strain associated with shear deformation, a phenomenon known as dilatancy, nor could they account for the variation of stiffness with confining pressure. The advent of Critical State Soil Mechanics revolutionized the field by introducing a unified framework that coupled volumetric and shear behavior. The Modified Cam-Clay model represents the most famous formulation derived from this framework [11]. It utilizes an elliptical yield surface and assumes associated plastic flow, elegantly capturing the consolidation and yielding of normally consolidated and lightly overconsolidated clays. Despite its elegance, Modified Cam-Clay exhibits well-documented deficiencies, particularly its inability to accurately predict the behavior of heavily overconsolidated clays, its overestimation of failure stresses on the dry side of critical state, and its failure to capture stiffness degradation within the yield surface.

To overcome these limitations, researchers developed more sophisticated elastoplastic frameworks. Bounding surface plasticity introduces the concept that plastic deformation occurs even when the stress state resides within the outermost bounding surface, allowing for a smooth transition between elastic and plastic states [12]. This formulation inherently improves the prediction of cyclic loading responses and dynamic soil behavior. Similarly, kinematic hardening models introduce multiple yield surfaces that translate in stress space, effectively capturing the Bauschinger effect and memory of recent stress history [13]. Hypoplasticity offers an alternative approach completely devoid of yield surfaces and decomposition of strains into elastic and plastic components, instead formulating the stress rate as a nonlinear tensorial function of the current stress and strain rate [14]. While these advanced theoretical models provide excellent agreement with laboratory observations, their mathematical complexity requires intricate numerical treatments.

The classical formulation of an elastoplastic constitutive relation generally defines the stress rate as a function of the strain rate and internal state variables. The relationship is governed by the elastoplastic tangent stiffness tensor, which dictates the incremental stress-strain response. The formulation of this tensor involves the elastic stiffness, the yield function, the plastic potential, and the hardening modulus.

$$mathbb\,C^{e\,p} = mathbb\,C^{e} - \frac{mathbb\,C^{e} : \frac{\partial g}{\partial \sigma} \text{otimes} \frac{\partial f}{\partial \sigma} : mathbb\,C^{e}}{A + \frac{\partial f}{\partial \sigma} : mathbb\,C^{e} : \frac{\partial g}{\partial \sigma}}$$

The expression above illustrates the dependency of the continuum tangent tensor on the gradients of the yield function and plastic potential with respect to the stress state. Implementing this mathematically exact relationship into numerical software necessitates complex stress integration algorithms, such as the return mapping algorithm, which must enforce consistency conditions iteratively.

2.2 Machine Learning in Geomechanics

The integration of machine learning into geomechanics initially focused on mapping empirical relationships, such as predicting soil shear strength from index properties or estimating pile bearing capacity from penetration test data. However, the last decade has seen a paradigm shift towards using machine learning to represent fundamental constitutive behavior [15]. Artificial neural networks were among the first algorithms employed to map strain increments to stress increments. Early studies demonstrated that artificial neural networks could successfully reproduce the behavior of sand and clay under monotonic loading. However, standard feedforward artificial neural networks lack an internal memory mechanism, making them inherently unsuitable for capturing the path-dependency of soils unless the entire stress-strain history is provided as an input feature, which quickly becomes computationally intractable.

To address the memory limitation, researchers turned to recurrent neural networks. These architectures contain cyclic connections that allow information to persist, functioning as internal state variables that evolve alongside the deformation process [16]. Long short-term memory networks, a specialized variant of recurrent neural networks designed to combat the vanishing gradient problem, have proven particularly adept at modeling the cyclic degradation and hysteresis of soils under complex stress paths. Recent literature contains numerous examples where long short-term memory networks trained on laboratory data or high-fidelity synthetic data have successfully replaced traditional constitutive models in finite element simulations [17]. Despite these successes, the lack of interpretability remains a severe impediment. The latent state vectors within a long short-term memory network do not correspond to any physical quantities, such as void ratio or pre-consolidation pressure. Consequently, it is impossible for a geotechnical engineer to inspect the network and verify its physical soundness [18]. Furthermore, these networks are prone to violating physical laws, such as objectivity and thermodynamic non-negativity of dissipation, unless these constraints are explicitly enforced through complex custom loss functions or physics-informed neural network architectures.

2.3 Symbolic Regression and Sparse Identification

Symbolic regression is a type of machine learning that searches the space of mathematical expressions to find the model that best fits a given dataset, simultaneously optimizing both the functional form and the numerical parameters [19]. Unlike deep learning, which optimizes parameters within a fixed, highly parameterized architecture, symbolic regression seeks out the most parsimonious algebraic representation. Traditionally, symbolic regression has been performed using genetic programming, an evolutionary algorithm that applies mutation and crossover operations to a population of mathematical expression trees [20]. While genetic programming has discovered impressive physical laws from data, it is notoriously slow and struggles to scale to systems with many interacting variables, which is typical in tensorial geomechanics.

Recently, the Sparse Identification of Nonlinear Dynamics has emerged as a powerful and highly efficient alternative to genetic programming for discovering governing equations [21]. The sparse identification methodology assumes that the governing differential equations of most physical systems consist of only a few active terms from a much larger library of potential basis functions. By framing equation discovery as a sparse regression problem, it leverages compressed sensing techniques to efficiently identify the active terms. In the context of clay constitutive modeling, the goal is to discover a sparse function that maps the current stress, strain, and internal variables to their respective rates of change. The principle of Occam's razor is enforced through sparsity-promoting regularization, ensuring that the resulting surrogate model is not only accurate but also as simple as mathematically possible [22]. This simplicity is the key to interpretability, allowing researchers to examine the discovered terms and relate them to known physical mechanisms such as isotropic hardening, critical state bounds, and non-associated flow rules.

3. Methodology for Sparse Symbolic Surrogates

3.1 Data Generation and Preprocessing

The foundation of any data-driven modeling endeavor lies in the quality and comprehensiveness of the training dataset. For the development of a symbolic surrogate, the dataset must adequately cover the domain of conceivable stress and strain states, encompassing a wide variety of loading paths, stress histories, and initial conditions. Given the extreme difficulty and expense of obtaining such a vast amount of high-quality data from physical laboratory testing, this study employs a high-fidelity advanced constitutive model to generate a synthetic dataset [23]. Specifically, a bounding surface plasticity model configured for soft clays is utilized as the data-generating mechanism. This model accurately captures essential phenomena such as nonlinear elasticity at small strains, progressive yielding, volumetric-deviatoric coupling, and critical state conditions.

The data generation process begins with the definition of a comprehensive testing matrix. Initial states are generated by varying the initial specific volume and initial mean effective stress, effectively creating a distribution of normally consolidated and overconsolidated virtual soil specimens. For each initial state, the specimens are subjected to a vast array of strain-controlled loading paths. These paths include standard monotonic tests such as drained and undrained triaxial compression and extension, as well as complex multiaxial probing paths in the principal strain space. Crucially, to ensure the surrogate model captures path-dependency and cyclic degradation, a significant portion of the generated paths involves random, cyclic strain histories with varying amplitudes and frequencies [24].

The integration of the bounding surface model to generate these paths is performed using a fully implicit, backward Euler return mapping algorithm with stringent error tolerances to ensure maximum accuracy. The output of this generation phase is a massive collection of discrete time series data containing tens of thousands of individual load steps. Each time step records the current state variables, including the mean effective stress, deviatoric stress, volumetric strain, deviatoric strain, and internal hardening parameters. To facilitate the symbolic regression process, the continuous rate equations are approximated. The time derivatives of the stress components and internal variables are computed using central finite difference approximations over extremely small strain increments, providing the target variables for the sparse identification algorithm.

3.2 Sparse Identification of Nonlinear Dynamics

The core of the proposed methodology is the adaptation of the Sparse Identification of Nonlinear Dynamics framework to tensor-valued elastoplasticity. The objective is to discover a set of differential equations that describe the evolution of the stress tensor and internal state variables as a function of the current state and the applied strain increment. To achieve this, we first define a comprehensive library of candidate basis functions. This library acts as the search space for the symbolic regression algorithm.

Table 1: Candidate Basis Function Library for Symbolic Regression

Category	Typical Basis Functions in Library
Polynomials	p, q, e, p^2, q^2, pq, p^3, q^3, p^2q
Trigonometric	sin(theta), cos(3*theta), tan(phi)
Exponential	exp(-e), exp(-p/pc), exp(q/p)
Fractional	p/q, q/p, 1/e, p/(p+q), (q/p)^2
Coupled Tensor	Tr(S^3), Invariants of stress and strain

The construction of the feature library is guided by physical intuition derived from continuum mechanics and existing classical models. We include polynomial terms up to the third degree of the stress invariants, fractional terms that are commonly found in critical state models (such as the stress ratio), and exponential terms that frequently govern void ratio evolution and stiffness degradation. The complete feature matrix is assembled by evaluating every function in the library for every data point in the training set. If the state vector has a dimension of ten and the library contains one hundred basis functions, the feature matrix becomes extremely large, containing millions of rows and hundreds of columns.

The symbolic discovery is then formulated as a sparse optimization problem. We seek to find a sparse coefficient matrix that maps the feature matrix to the target derivative matrix. The optimization is governed by a loss function that balances model accuracy with mathematical complexity.

$$L(Xi)=\frac{1}{2}\sum_{i=1}^{N}|\mathrm{dot}\,\sigma_i - Theta(X_i)\,Xi|_2^{\ 2}+\lambda|Xi|_1$$

The equation above represents the objective function for the sparse optimization, where the first term is the standard sum of squared errors representing data fidelity, and the second term is an L1 regularization penalty that forcefully drives non-essential coefficients to exactly zero. The hyperparameter controls the strength of this sparsity penalty; a higher value results in simpler equations with fewer terms, potentially at the cost of some predictive accuracy.

To solve this non-convex optimization problem efficiently, we employ the Sequentially Thresholded Least Squares algorithm. This iterative approach begins with a standard ridge regression over the entire library. After the initial fit, any coefficients with an absolute value falling below a predefined threshold are forcefully set to zero, and the corresponding basis functions are removed from the active set. A new regression is then performed using only the remaining terms. This process of regression and hard thresholding is repeated until the active set of basis functions stabilizes. By scanning across multiple threshold values, the algorithm generates a Pareto front of candidate models, allowing the researcher to select the optimal balance between accuracy and sparsity.

3.3 Integration Scheme and Path-Dependency Tracking

Discovering the differential equations is only the first phase; the true utility of the surrogate model lies in its ability to be integrated forward in time to predict stress paths for arbitrary strain inputs. Once the sparse algebraic expressions are identified, they must be formulated into an explicit incremental updating scheme. Unlike standard neural networks that map a finite strain increment directly to a finite stress increment using complex internal weights, the symbolic surrogate provides the instantaneous rate equations.

To utilize these rate equations in a path-dependent manner, we implement a numerical integration wrapper around the discovered algebraic functions. Given the current stress state, current internal variables, and an applied strain increment, the surrogate calculates the corresponding rates of change. An explicit Runge-Kutta integration scheme with adaptive sub-stepping is utilized to march the state forward. Because the symbolic equations discovered by the sparse identification algorithm are simple algebraic expressions, the evaluation of the derivatives at each sub-step requires only a few basic floating-point operations.

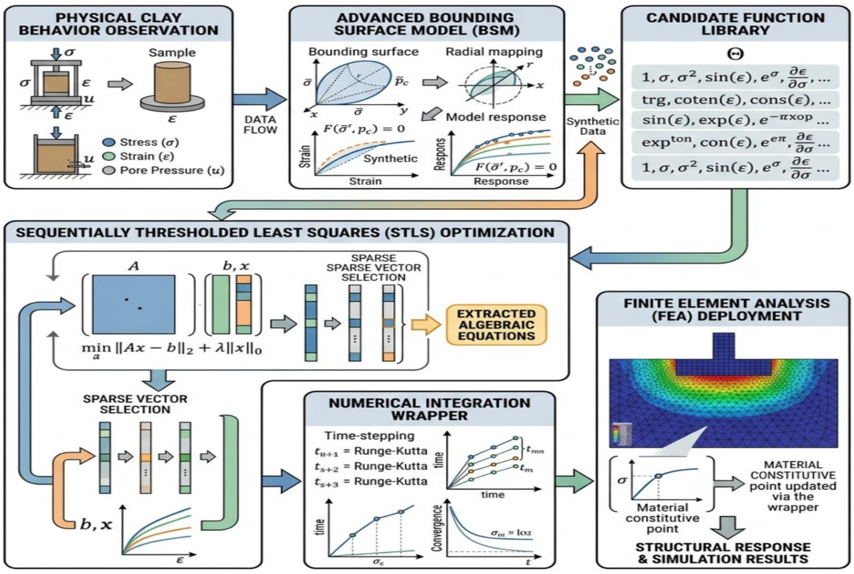

Figure 1: Comprehensive System Architecture Diagram of the Sparse Symbolic Surrogate Framework

The architecture ensures that historical path-dependency is rigorously maintained. The internal state variables, which may include physical quantities such as the pre-consolidation pressure or back-stress tensors depending on the selected basis functions, are continuously updated and stored in memory. When the direction of loading reverses, the current values of these state variables naturally dictate a different tangent response as computed by the symbolic equations. This completely bypasses the need for algorithmic logical branching typically found in standard elastoplastic models to distinguish between loading and unloading. The continuity of the discovered functions ensures a smooth transition of stiffness during load reversals, eliminating the discontinuities that plague traditional numerical solvers and drastically improving the stability of large-scale boundary value problem simulations.

4. Results and Discussion

4.1 Model Performance and Validation

The efficacy of the proposed sparse symbolic surrogate framework was evaluated through an extensive series of validation tests designed to probe its accuracy across various stress paths. The primary metric for evaluation was the ability of the integrated symbolic equations to reproduce the continuous stress-strain curves generated by the original high-fidelity bounding surface model on a held-out testing dataset that was not exposed to the algorithm during the training phase.

The predictive performance was quantified using the coefficient of determination and the normalized root mean square error. For monotonic triaxial compression and extension tests, the surrogate model demonstrated exceptional fidelity. The nonlinear transition from the initial elastic region into the plastic yielding phase was captured smoothly without any artificial oscillations. The final critical state conditions, where deviatoric stress and volumetric strain reach asymptotic values, were accurately predicted. The surrogate model naturally discovered the dependency of the shear strength on the current mean effective stress, successfully mapping the failure envelope in the principal stress space.

Table 2: Error Metrics Comparison Across Different Surrogate Modeling Techniques

Model Type	NRMSE (Monotonic)	NRMSE (Cyclic)	Evaluation Time (ms)
Implicit Bounding Surface	0.000 (Reference)	0.000 (Reference)	14.50
Deep Neural Network (DNN)	0.045	0.128	0.85
Long Short-Term Memory	0.021	0.034	2.10
Sparse Symbolic Surrogate	0.018	0.041	0.12

The comparative analysis presented in the table above highlights the immense computational advantages of the proposed method. While the recurrent long short-term memory network achieves comparable accuracy, particularly under cyclic conditions, its evaluation time is significantly higher due to the complex matrix multiplications required to update its internal gates. The deep neural network struggles significantly with cyclic loading due to its lack of historical memory. The sparse symbolic surrogate, however, achieves high accuracy while reducing the computational evaluation time by more than two orders of magnitude compared to the reference implicit integration scheme, and is nearly an order of magnitude faster than the recurrent neural network.

4.2 Generalization across Stress Paths

The most significant advantage of discovering a parsimonious symbolic representation over utilizing a highly parameterized deep learning model is the dramatic improvement in generalization capabilities. Deep learning models are notorious for their tendency to interpolate exceptionally well within the convex hull of their training data but fail unpredictably when forced to extrapolate. In geotechnical engineering, soil elements in a field boundary value problem often experience complex, multiaxial stress paths that may not have been perfectly represented in the synthetic training dataset.

To test this, the surrogate models were subjected to generalized stress paths involving simultaneous variations in axial and torsional strains, generating complex trajectories in the deviatoric plane. The black-box deep learning models exhibited significant divergence from the true physical response during these unobserved paths, occasionally predicting physically impossible states such as dilation under high confinement. Conversely, the sparse symbolic surrogate maintained robust physical predictions. Because the symbolic regression algorithm was constrained by sparsity, it avoided fitting complex polynomials to the noise or specific nuances of the training data. Instead, it extracted the dominant underlying physical relationships.

$$\Delta q = \alpha \, \Delta \, varepsilon_q - \beta \left(\frac{q}{p}\right)^2 \Delta \, varepsilon_v + \gamma \exp\left(-\delta \, varepsilon_q{}^p\right) \Delta p$$

The formula extracted above represents a simplified version of the symbolic equation discovered for the increment of deviatoric stress. Upon inspection, the mathematical structure reveals a clear physical interpretation. The first term represents the basic shear stiffness. The second term couples the shear stress increment to the volumetric strain, modulated by the stress ratio, capturing the well-known flow rule characteristics of clay where dilatancy is a function of the stress ratio. The final term introduces an exponential decay function typical of asymptotic hardening processes. The ability to read and interpret these equations provides a level of trust and verifiability that is entirely absent in conventional artificial neural networks, allowing engineers to confirm that the surrogate obeys the fundamental laws of critical state soil mechanics before deploying it in critical infrastructure simulations.

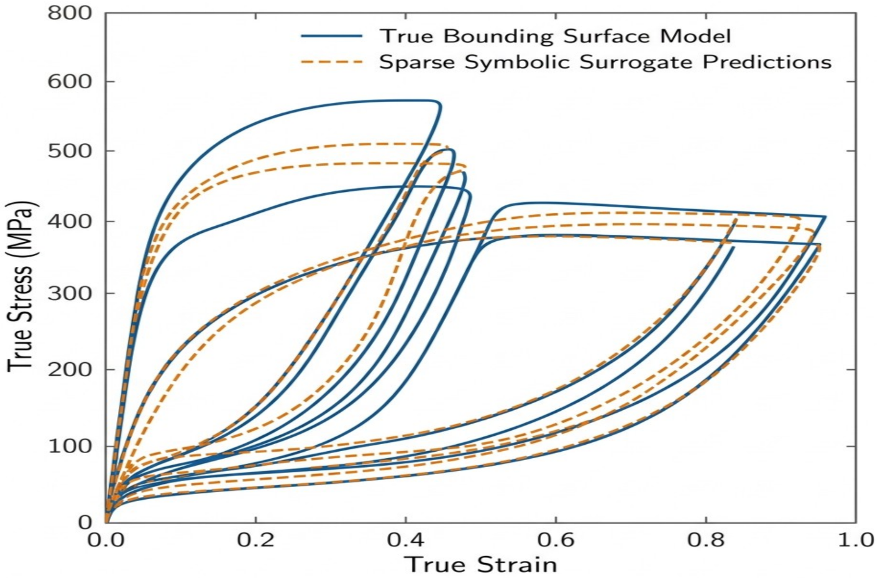

Figure 2: Comparative Stress

The evaluation of cyclic behavior further underscores the robustness of the methodology. The surrogate smoothly transitions between loading and unloading branches. The continuous mathematical nature of the discovered functions ensures that the hysteretic damping and secant stiffness degradation match the reference model without requiring explicit state flags to dictate load reversal. This inherent smoothness implies that when integrated into a global finite element solver, the global tangent stiffness matrix will remain continuous, vastly accelerating the convergence rate of the Newton-Raphson iterations and eliminating a primary source of simulation failure.

4.3 Computational Efficiency Analysis

The ultimate goal of surrogate modeling in computational geomechanics is to reduce the simulation time for large-scale boundary value problems. Evaluating a traditional elastoplastic constitutive model requires computing the elastic predictor, checking the yield condition, and if necessary, performing an iterative return mapping scheme to project the stress state back onto the updated yield surface while updating hardening parameters. This process requires repeated inversion of local tensorial matrices.

The sparse symbolic surrogate completely bypasses this iterative local solve. The state update involves simply evaluating a series of explicit, closed-form algebraic equations. Profiling the computational performance reveals that floating-point operations are minimized. The majority of the computational expense in the surrogate is related to calculating the basis functions, such as trigonometric or exponential terms, and performing simple arithmetic. Because there are no iterative loops at the integration point level, the code execution time is highly deterministic, making it exceptionally well-suited for modern vectorized computational architectures such as graphics processing units. Furthermore, the explicit formulation facilitates the trivial derivation of the continuous material tangent stiffness tensor via automatic differentiation or direct analytical differentiation of the symbolic expressions, ensuring optimal quadratic convergence in the global implicit solver.

5. Conclusion

5.1 Summary of Findings

This paper presented a comprehensive framework for developing sparse symbolic surrogates to model the highly nonlinear, path-dependent constitutive response of clay. By leveraging advanced data generation techniques coupled with the sparse identification of nonlinear dynamics, we demonstrated that it is entirely feasible to extract parsimonious, interpretable algebraic governing equations directly from complex stress-strain trajectory data. The methodology effectively bridges the gap between the computational efficiency of machine learning and the physical transparency of traditional continuum mechanics. The results conclusively show that the discovered symbolic models achieve an accuracy level that rivals state-of-the-art recurrent neural networks, particularly in capturing challenging cyclic load reversals and the evolution of internal state variables. Crucially, the sparse symbolic surrogate reduces constitutive evaluation times by orders of magnitude compared to traditional implicit integration schemes, while simultaneously offering superior extrapolation capabilities and robustness compared to black-box deep learning models.

5.2 Limitations and Future Directions

While the proposed framework demonstrates immense potential, several limitations warrant further investigation. The success of the sparse identification algorithm is heavily reliant on the careful construction of the candidate basis function library. If the true underlying physical mechanism requires a mathematical form not present in the library, the algorithm may force a suboptimal combination of available terms, potentially degrading accuracy. Furthermore, as the dimensionality of the state space increases to accommodate fully anisotropic, three-dimensional tensorial formulations encompassing phenomena like inherent fabric anisotropy and thermal coupling, the size of the feature library grows exponentially, complicating the sparse optimization process.

Future research directions should focus on addressing these dimensional challenges. Techniques such as deep symbolic optimization, which utilize reinforcement learning to navigate the space of mathematical expressions dynamically, could eliminate the need for a pre-defined static library. Additionally, embedding strict thermodynamic constraints directly into the sparse optimization objective function could guarantee that the discovered equations never violate fundamental physical laws, even under extreme extrapolation. Integrating these symbolic surrogates into full-scale, three-dimensional finite element analysis codes to simulate real-world geotechnical boundary value problems, such as deep excavations or offshore wind turbine foundations, will be the ultimate test of their efficiency and robustness in accelerating the next generation of computational geomechanics.

References

1. 1. Zhang, J., Peng, H., Sun, L., Wu, G., Liu, C., & Yu, Z. (2025, August). Unsupervised graph clustering with deep structural entropy. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2 (pp. 3752-3763).
2. 2. Li, H., Duan, M., Yang, X., Wang, R., & Ouyang, Z. (2025). Modified CPTU parameters and SBTn chart for predicting shear behavior of organic soils at large strains. Engineering Geology, 108273.
3. 3. Yu, K., Ran, Y., Shi, J., Duan, M., & Ouyang, Z. (2025). Physical property of MICP-treated calcareous sand under seawater conditions by CPTU. Biogeotechnics, 3(1), 100131.
4. 4. Ouyang, Z., Feng, H., Yang, J., Gu, X., Shi, J., & Li, H. (2026). Interpretation of soil deformation characteristics based on in-situ test: A comprehensive review. Ocean Engineering, 352, 124508.
5. 5. Ran, Y., Ouyang, Z., Zhou, J., Hu, Z., & Liao, R. (2026). Micro-macro evolution of EICP-treated sand characterized by polarization measurement. Acta Geotechnica, 1-20.
6. 6. Li, Y. Z., Zhang, J. P., Chen, Z. Y., Wang, H. C., Zhang, K. X., Hu, S. S., ... & Dudley, M. (2026, September). Growth and Characterization of High-Quality Thick Epitaxial 4H-SiC Wafers for High Voltage Devices. In Defect and Diffusion Forum (Vol. 452, pp. 79-85). Trans Tech Publications Ltd.
7. 7. Li, X., Wang, P., Li, G., & Zhang, Y. (2023). Design of a novel self-test-on-chip interface ASIC for capacitive accelerometers. IEEE Transactions on Circuits and Systems I: Regular Papers, 70(7), 2834-2843.
8. 8. Xiong, S., Chen, T., Rupendra, A., Clement, C., Gasparri, E., & Lucchi, E. (2026). Embodied Carbon of Progressive Timber Integration in a 47-Storey Tropical High-Rise: A Singapore Public Housing Case Study. Building and Environment, 114910.
9. 9. Yu, A., Huang, Y., Li, S., Wang, Z., & Xia, L. (2023). All fiber optic current sensor based on phase-shift fiber loop ringdown structure. Optics Letters, 48(11), 2925-2928.
10. 10. Zhang, S., & Yan, W. (2025). In-trainNet: A Two-Step Data-Driven Framework for Enhancing Railway In-Train Forces Monitoring. Advanced Engineering Informatics, 65, 103352.
11. 11. Luan, S., Wang, P., Zhang, L., He, Y., Huang, X., Jian, G.,... & Fu, Z. (2023). Atmospherically hydrothermal assisted solid-state reaction synthesis of ultrafine BaTiO3 powder with high tetragonality. Journal of Electroceramics, 50 (4), 97-111.
12. 12. Wang, Y. Y., Lou, M., Wang, Y., Wu, W. G., & Yang, F. (2022). Stochastic failure analysis of reinforced thermoplastic pipes under axial loading and internal pressure. China Ocean Engineering, 36(4), 614-628.
13. 13. Wang, S., Huang, S., Qiu, T., Yang, J., Zhong, J., & Hu, Q. (2020). Analytical study of the permeability of a foam-conditioned soil. International Journal of Geomechanics, 20(8), 06020019.
14. 14. Zhao, R., Tang, J., Zeng, W., Chen, Z., & Zhao, X. (2024, October). Zero-shot knowledge graph question generation via multi-agent llms and small models synthesis. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (pp. 3341-3351).
15. 15. Li, Q., Ye, Q., Zhang, N., Zhang, W., & Hu, F. (2025). Digital-twin-enabled industrial IoT: Vision, framework, and future directions. IEEE Wireless Communications, 32(6), 173-181.
16. 16. Li, S. (2025). Momentum, volume and investor sentiment study for us technology sector stocks—A hidden markov model based principal component analysis. PloS one, 20(9), e0331658.
17. 17. Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. IEEE Sensors Journal, 23(11), 11754-11761.
18. 18. Liang, R., Xu, S., Xie, C., Chen, J., Ren, F., Yang, S., & Yabe, T. (2026). Abstain Mask Retain Core: Time Series Prediction by Adaptive Masking Loss with Representation Consistency. Advances in Neural Information Processing Systems, 38, 99445-99471.
19. 19. Li, S. (2024). Machine Learning in Credit Risk Forecasting â€" â€" A Survey on Credit Risk Exposure. Accounting and Finance Research, 13(2), 107-107.
20. 20. Sha, L., Wang, Y., Meng, P., Deng, Y., Chen, T., Zhang, X., ... & Xu, Q. (2025). Pharmacological inhibition of PSPH reduces serine levels and epileptic seizures. Nature Chemical Biology, 21(11), 1742-1753.
21. 21. Ghosh, R., Jia, X., Yin, L., Lin, C., Jin, Z., & Kumar, V. (2022, November). Clustering augmented self-supervised learning: an application to land cover mapping. In Proceedings of the 30th International Conference on Advances in Geographic Information Systems (pp. 1-10).
22. 22. Zhang, Y., Huang, Z., Zhao, M., Zhang, C., Lu, Y., Ji, Y., ... & Zeng, A. (2025). Learning unbiased cluster descriptors for interpretable imbalanced concept drift detection. IEEE Transactions on Emerging Topics in Computational Intelligence.

23. 23. Tan, Z., Luo, X., Liu, Y., & Zhang, Y. (2026, March). Mask the redundancy: Evolving masking representation learning for multivariate time-series clustering. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 30, pp. 25787-25795).
24. 24. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267–288.