

Physics-Residual Learning with Few-Shot Calibration for Cross-Geological UCS Prediction: A Model-Agnostic Framework

Long Long¹

LongLong90@chinesemail.net

China Railway Corporation

Article

Keywords: Uniaxial Compressive Strength, Physics-guided machine learning, Few-shot calibration, Cross-geological prediction, Domain generalization, Active sampling, Calibratability Index, Geological Bias Correction, Rock mechanics

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11136078/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-Residual Learning with Few-Shot Calibration for Cross-Geological UCS Prediction: A Model-Agnostic Framework

Long Long^{1,*}

¹ China Railway Corporation, Wuhan Railway Bureau, Hankou Railway Station, 430023, China , Email:
LongLong90@chinesemail.net

Abstract

Forecasting Uniaxial Compressive Strength (UCS) across geologically varied landscapes poses difficulties, since data-centric models developed on localized datasets often falter when applied to unfamiliar geological areas. This research assesses the cross-geological applicability of UCS predictions via a Leave-One-Country-Out methodology across six nations and 397 rock specimens, contrasting solely data-driven models with a physics-based residual technique among four foundational modules. Zero-shot outcomes indicate a significant compromise: Direct ML attained a slightly reduced initial RMSE (58.5 MPa) compared to Physics-Residual (60.2 MPa), however this benefit concealed significant instability, since single-sample calibration unexpectedly heightened error. Physics anchoring converted extrapolation inaccuracies into systematic, interpretable biases grounded on physical principles. We provide a Calibratability Index (CI) that measures the proportion of the domain gap bridged using five calibration samples; Physics-Residual attained a CI of 0.511 compared to 0.364 for Direct ML. Diversity-oriented active sampling using non-invasive Vp and density assessments decreased RMSE to 37.7 MPa at $k = 3$ and 36.8 MPa at $k = 5$, indicating a 17% enhancement compared to random sampling. The Wilcoxon tests validated significance ($p < 10^{-8}$) with moderate effect sizes (Cohen's $d = -0.53$ to -0.58). Geological Bias Correction multipliers varied from 0.36 in Turkey to 5.60 in South Africa, indicating that cross-domain error is organized and geologically significant. Feature engineering yielded no enhancement compared to the original predictors, affirming the notion of parsimony for datasets with few samples. The methodology is independent of models and facilitates a decrease of over 90% in destructive testing needs, creating an effective process for dependable UCS forecasting in new geological contexts.

Keywords: Uniaxial Compressive Strength; Physics-guided machine learning; Few-shot calibration; Cross-geological prediction; Domain generalization; Active sampling; Calibratability Index; Geological Bias Correction; Rock mechanics

Highlights

- Physics-anchored residual learning transforms unpredictable cross-geological extrapolation errors into systematic, correctable biases
- Novel Calibratability Index quantifies domain adaptation efficiency; Physics-Residual achieves 0.511 versus 0.364 for Direct ML
- Diversity-based active sampling using non-destructive Vp and density measurements reduces RMSE by 14–17% over random sampling
- Only 3–5 strategically selected UCS tests suffice for reliable cross-geological prediction, enabling ~90% reduction in destructive testing

- Geological Bias Correction multipliers range from 0.36 to 5.60 across six countries, confirming structured and interpretable cross-domain error

1. Introduction

The uniaxial compressive strength (UCS) of unaltered rock is a pivotal quantity in rock mechanics and geotechnical engineering, functioning as an essential input for the design and stability evaluation of tunnels, slopes, foundations, and subterranean excavations. Bieniawski [1] noted that UCS is more often sought for by mining engineers compared to other rock characteristics, highlighting its pivotal importance in practical design and stability assessments. The UCS underpins well recognized rock mass categorization frameworks, such as the Rock Mass Rating (RMR) and the Rock Mass Index (RMI), and is a crucial element of the Hoek–Brown failure criteria. An absence of precise UCS data may lead to suboptimal design decisions and therefore, structural collapses.

Notwithstanding its significance, the direct assessment of UCS by standardized laboratory testing regulated by ASTM [2] and ISRM [3] guidelines necessitates high-quality cylindrical core specimens, advanced laboratory infrastructure, and considerable time and financial investment. Acquiring superior rock cores is especially difficult for extensively worn and fractured geological formations, and the unconfined compression test (UCT) is known to have significant drawbacks such as cost and time inefficiency. These constraints have inspired a wide array of indirect estimating methodologies, from basic empirical correlations to sophisticated artificial intelligence (AI) frameworks.

Conventional indirect techniques like the Schmidt rebound hammer, ultrasonic pulse velocity (UPV), point load test (PLT), and Brazilian tensile strength (BTS) have shown significant effectiveness, yet their predictive precision is greatly influenced by lithology, rock anisotropy, and localized conditions. Empirical correlations, such the ISRM-suggested connection between UCS and point load index ($UCS = 20\text{--}25 I_{s50}$) and power-law associations between UCS and P-wave velocity, provide swift estimations but are fundamentally site-specific and sometimes fail when applied to other geological contexts.

In recent years, machine learning (ML) models have achieved significant acclaim for UCS prediction, surpassing traditional techniques by elucidating complex nonlinear interactions among geological and mechanical factors [4]. A comprehensive literature review encompassing 131 peer-reviewed articles from 2014 to 2024 revealed that artificial neural networks (ANN) were the predominant AI algorithm utilized (24.7%), succeeded by adaptive neuro-fuzzy

inference systems (ANFIS, 11.7%) and random forest (RF, 7.6%), with prediction accuracies often surpassing $R^2 > 0.90$ [5]. Deep learning methodologies, including as convolutional neural networks (CNN), gated recurrent units (GRU), and long short-term memory (LSTM) networks, have attained remarkable coefficient of determination values of 0.998 for UCS forecasting [6]. Tree-based ensemble techniques, such as XGBoost and CatBoost, have shown comparable efficacy, with XGBoost reaching RMSE figures as minimal as 4.2 MPa and a R^2 of 0.964 in hybrid optimization models [7].

Nonetheless, in spite of these advancements, a significant constraint remains largely overlooked: most machine learning research use random data partitioning techniques that presuppose training and testing datasets derive from an identical geographical distribution. In practice, geotechnical endeavors often need extrapolation to novel locations with geological attributes that consistently diverge from the training dataset. This is a domain shift issue, where models developed on a certain geological distribution perform poorly when applied to a different one [8]. Although transfer learning has been investigated for predicting soil properties [8] and domain adaptation techniques have been suggested for site-specific characterization [9], these methods have mainly focused on spatial variability within analogous geological contexts rather than comprehensive cross-geological extrapolation.

Physics-informed machine learning (PIML) has surfaced as a compelling framework for improving the interpretability, precision, and dependability of predictive models in geotechnical contexts via the amalgamation of physical principles with data-driven methodologies. Physics-informed neural networks (PINNs) have been created for addressing rock mechanics issues such as elasticity, elastoplasticity, and integrated thermo-hydro-mechanical phenomena [10]. Nonetheless, the utilization of physics-informed frameworks for few-shot calibration amongst cross-geological domain shifts is being inadequately investigated. Moreover, the issue of the requisite number of site-specific calibration samples for dependable predictions and the criteria for sample selection has not been well examined in the rock mechanics literature.

This research tackles these deficiencies via three interrelated contributions. Initially, we create and assess a physics-informed residual learning framework that separates UCS prediction into a physics-derived baseline element and a machine-learning adjustment factor, converting erratic extrapolation inaccuracies into systematic, scientifically interpretable biases. Secondly, we provide the Calibratability Index (CI), an innovative measure that assesses the effectiveness of rectifying a model's cross-domain error using little local calibration data. Third, we methodically

assess active sampling methodologies for the selection of ideal calibration samples, demonstrating that diversity-oriented selection using non-invasive measures far surpasses random sampling.

The remainder of this manuscript is organized as follows. Section 2 reviews the relevant literature on empirical correlations, machine learning approaches, physics-informed methods, and domain adaptation in geotechnical engineering. Section 3 describes the dataset, physics baseline formulation, predictive architectures, validation framework, and evaluation metrics. Section 4 presents the results of the multi-architecture evaluation, few-shot calibration experiments, and active sampling comparison. Section 5 discusses the mechanisms underlying the observed performance differences, contextualizes findings within existing literature, addresses limitations, and outlines practical implications. Section 6 concludes with a summary of contributions and recommendations for future work.

2. Literature Review

2.1. Empirical Correlations for UCS Prediction

The first and most often used indirect techniques for estimating UCS depend on empirical relationships between UCS and index qualities that are simpler to assess. The point load index (Is_{50}) is often used for forecasting UCS because to its simplicity and rapid execution on lump specimens [11]. Throughout the last fifty years, many empirical formulas have been suggested, with the ISRM-endorsed correlation $UCS = 20 - 25Is_{50}$ acting as a well-recognized standard [3]. A thorough evaluation of 40 established empirical equations, utilizing 55 uniaxial compressive and point load tests from Turkish tunnel and dam projects, revealed that merely 8 of the 40 equations exhibited satisfactory performance across various statistical metrics such as RMSE, VAF, MAPE, and predictive accuracy. This discovery underscores the intrinsic constraint of empirical correlations: they are fundamentally localized and fail to consistently apply across varied geological contexts.

P-wave velocity (V_p) has been thoroughly associated with UCS. An examination of 96 samples from five distinct complete rock classifications revealed robust connections between V_p and physical-mechanical characteristics, with correlation values greater than $R^2 > 0.80$ for UCS forecasting [12]. A study of granite samples from Rawang, Selangor, established the empirical equation $UCS = 144.68V_p - 684.37$, but another investigation on the granites of Penang

Island indicated $UCS = 8.385 I_{SSo} + 30.016$ [13]. Nonetheless, the divergence between in situ and laboratory V_p measurements attributed to fractures and natural joints creates considerable ambiguity in velocity-based forecasts [14]. Empirical models using several variables such as V_p , density, and point load index have shown enhanced prediction accuracy relative to single-variable correlations [15]. The amalgamation of several index qualities facilitates more reliable estimate, while the coefficients of these models are still tailored to particular geological contexts. A thorough analysis indicates that the predicted precision of empirical techniques is significantly influenced by lithology, rock anisotropy, and localized circumstances, hence limiting their general application [4].

2.2. Machine Learning Approaches for UCS Prediction

The constraints of empirical correlations have prompted much investigation into machine learning methods for UCS forecasting. Artificial neural networks (ANNs) are the most extensively used AI methodology, including about 24.7% of the research published in a systematic analysis of 131 papers [5]. Artificial Neural Networks can effectively identify intricate nonlinear associations between input variables and UCS without the need for explicit definition of functional structures. Models of multilayer perceptrons (MLP) refined by metaheuristic algorithms have shown impressive efficacy, with stochastic alterations facilitating reliable predictions across various rock classifications. Support vector machines (SVMs) and their regression counterparts (SVR) have shown efficacy with high-dimensional datasets, yielding dependable outcomes even with little data. A new interval prediction model using nested cross-validation and support vector machines (LSSVM-NCV-ABKDE) was created for cracked weakly cemented rocks, resulting in an RMSE of 0.336, MAE of 0.192, and R^2 of 0.957 [16]. The efficacy of SVM-based methodologies resides in their capacity to manage intricate regression challenges while preserving model simplicity via kernel functions.

Tree-based ensemble techniques have shown to be very efficient for UCS prediction owing to its capacity to identify nonlinear associations, manage diverse input variables, and provide assessments of feature significance. Random forest (RF) generates many decision trees and produces average forecasts, mitigating overfitting by decorrelating the different trees. In a comparison analysis of tree-based models, RF had the greatest predictive accuracy, with validation metrics of RMSE = 1.93, MSE = 0.87, and $R^2 = 1.0$. A collective tree-based machine learning model for travertine rocks revealed the effectiveness of integrating several tree-based methods to enhance forecast precision [17].

Extreme gradient boosting (XGBoost) has achieved notable recognition for UCS forecasting owing to its scalability, regularization capabilities, and computing efficiency [18]. XGBoost utilizes second-order derivatives of the loss

function together with regularization components to penalize model complexity, hence mitigating the danger of overfitting. A hybrid LMWOA-XGBoost model that integrates measurement while drilling (MWD) data with whale optimization achieved an RMSE of 4.1976 MPa and a R^2 of 0.9639 on the test dataset, with repeated cross-validation verifying stability beyond a singular random division [7]. Metaheuristic-optimized ensemble models, such as GWO-XGBoost and GWO-AdaBoost, have shown impressive efficacy, with GWO-AdaBoost attaining $R^2 = 0.993$, RMSE = 2.2830, and MAE = 1.6853 [19].

CatBoost has surfaced as a formidable substitute for XGBoost, providing inherent management of categorical attributes without the need for one-hot encoding. A framework using CatBoost for forecasting UCS of stabilized saline soils attained $R^2 = 0.9391$, MAE = 0.0876, and RMSE = 0.1689, surpassing MLP and XGBoost, with statistical analyzes validating its marked superiority [20]. CatBoost's capacity to seamlessly integrate categorical variables like soil type and stabilizer type ensures interpretability with superior prediction precision.

Deep learning methods have also been investigated for UCS forecasting [21]. A comparative analysis of CNN, GRU, LSTM, and RNN architectures revealed LSTM to be the most proficient at predicting UCS and tensile strength, attaining R^2 values of 0.998 and 0.997, respectively. The SHAP analysis improved the LSTM model's capacity to discern temporal dependencies and intricate nonlinear associations for interpretative purposes [6]. Nonetheless, the use of deep learning methodologies is constrained by their dependence on extensive and representative datasets a limitation that is especially pronounced in geotechnical engineering, where data acquisition is both costly and labor-intensive.

2.3. Physics-Informed Machine Learning

Physics-informed machine learning (PIML) has surfaced as a viable framework for augmenting the dependability and comprehensibility of prediction models in geotechnical engineering through the amalgamation of physical principles and data-centric learning. PIML techniques have been used in several geotechnical domains such as soil mechanics, hydrology, site assessment, slope stability evaluation, and foundation engineering [22].

Physics-informed neural networks (PINNs) have been created to address issues in rock mechanics such as elasticity, elastoplasticity, finite-strain mechanics, coupled thermo-hydro-mechanical challenges, tunneling, and deformation caused by excavation. A PINN for pressurized voids in heterogeneous geological formations attained mean absolute displacement errors of $O(10^{-5})$ m in both uniform and non-uniform media, while the relative L_2 -norm errors for normal

stress components consistently stayed below 3.1×10^{-3} on average [10]. The use of physics-informed loss functions upholding equilibrium, material laws, and boundary conditions significantly enhanced the accuracy of predictions relative to just data-driven methods.

A physics-informed neural network methodology was suggested for UCS prediction to accurately identify the genuine correlations among factors for forecasting the dynamic triaxial strength of rocks in frigid conditions. The PINN combined rock mechanics theory with neural networks to enhance prediction precision, interpretability, and generalization, leading to the creation of an uncertainty-aware version (UAPINN) designed to integrate uncertainty quantification [16]. The results indicate that physics-informed methods improve dependability by limiting predictions to physically permissible bounds.

The amalgamation of physics-informed restrictions with machine learning models provides a solution to the inherent drawback of exclusively data-driven methodologies: their failure to generalize beyond the training distribution. A position paper on PIML in geotechnical engineering highlights that the amalgamation of domain expertise, model interpretability, and multiphysics challenges is a significant avenue for future study [22]. Nonetheless, the utilization of physics-informed frameworks for cross-geological extrapolation and limited-sample calibration is still inadequately investigated.

2.4.Domain Shift and Transfer Learning in Geotechnical ML

Domain shift the occurrence where the statistical distribution of data varies between training (source) and operational (target) contexts poses a significant obstacle for machine learning in geotechnical engineering. A case study on the prediction of undrained shear strength (USS) specifically examined the domain shift between source soils from Finland (217 samples) and target soils from Sylhet, Bangladesh (134 samples). Statistical analyzes validated notable distributional disparities among areas, with a foundational AdaBoost.The R2 model only trained on the target domain revealed the need for adaptation. Instance-weighted kernel ridge regression (IWKRR) attained the utmost predictive precision ($R^2 = 0.876$, RMSE = 6.89 kPa, MAE = 5.46 kPa) and exhibited consistent performance despite limited target data availability, illustrating that instance-based transfer learning can proficiently alleviate domain shift [8].

Domain adaptation has been investigated for predicting wall deformation caused by excavation, using a domain-adversarial neural network (DANN) as the foundational model, with a suggested contrastive self-supervised head to provide strong zero-shot generalization [9]. A self-supervised domain adaptation technique using generative AI via

inpainting was introduced for crack segmentation in deteriorating infrastructure, effectively tackling domain shift challenges at novel locations, resulting in average F1-score enhancements of 25.82% and 17.83% across target domain datasets [23]. These investigations together illustrate that domain shift is a widespread obstacle in geotechnical machine learning, and that adaption techniques may significantly enhance performance across domains.

Nonetheless, the current body of transfer learning research in geotechnical engineering has chiefly focused on spatial variability within analogous geological contexts or sensor domain adaptation, rather than methodical cross-geological extrapolation where essential rock characteristics diverge between source and target domains. The few-shot learning framework modifying a pre-existing model for a new domain with just a few number of labeled examples has garnered less interest in rock mechanics [24], even though it has practical significance for field studies where damaging testing should be curtailed.

2.5.Active Learning and Sample-Efficient Strategies

Active learning the method of selectively choosing samples for labeling to enhance model performance provides a means to lower site characterization expenses while ensuring prediction accuracy. A multi-fidelity residual Gaussian process framework based on active learning (AL-MR-GP) was introduced for modeling geomaterial properties, emphasizing high-fidelity data from the relevant location with more information gain [25]. The framework demonstrated that active learning might pinpoint the most valuable samples for calibration, hence minimizing the need for costly testing.

A system that is guided by knowledge and optimized for data efficiency, which combines active learning with multi-fidelity learning, was established for the assessment of soil spatial variability. The findings demonstrated that using just fifty percent of the high-fidelity data, the suggested framework attained equivalent accuracy and dependability as methods lacking active learning. The framework also recognized that the significance of low-fidelity data often diminishes as the spatial separation between sites expands an essential factor for evaluating site similarity [26].

Notwithstanding these advancements, active learning has yet to be methodically used for UCS prediction or the calibration of cross-geological models. The inquiry over which samples from a selection of prospective rock cores need to be chosen for destructive testing to optimize calibration efficiency is still unresolved. This research tackles this deficiency by assessing diversity-oriented and physics-discrepancy sampling methodologies in the realm of few-shot cross-geological calibration.

2.6. Research Gaps and Positioning of This Study

The literature study uncovers several significant deficiencies. Initially, while machine learning models have shown remarkable in-domain precision for UCS prediction, their transferability across different geological contexts is mostly unassessed [4,5]. Secondly, physics-informed methodologies have been used in rock mechanics issues [10], although they have not been systematically employed to address the domain shift challenge in UCS forecasting. Third, active learning techniques have shown potential in geotechnical site characterization [25,26], although they have not been used for few-shot calibration of UCS models across various geological contexts. Fourth, the issue of the requisite number of calibration samples and the criteria for their selection has not been examined within a statistically robust framework.

This research tackles these deficiencies by creating a physics-based residual learning framework assessed through rigorous Leave-One-Country-Out validation, presenting the Calibratability Index as an innovative measure for assessing calibration efficacy, and methodically contrasting active sampling techniques for minimal geological adaptation. The framework is independent of any specific model, verified with four foundational learners, and intended for effective use in field studies where destructive testing has to be reduced.

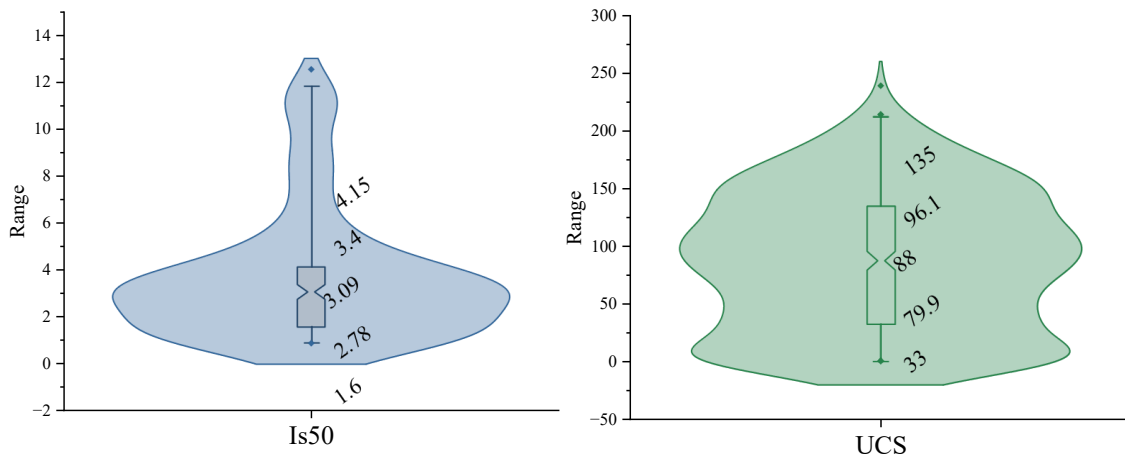
3. Materials and Methods

3.1. Dataset Description

The collection consists of 397 rock specimens gathered from six nations across three continents: Malaysia (116 specimens), Turkey (104), Iran (75), Portugal (55), South Africa (27), and Brazil (20). The specimens exemplify three categories of rock: Igneous (219 specimens), Sedimentary (175 specimens), and Metamorphic (3 specimens). The significant scarcity of metamorphic rocks ($n = 3$) prevents their use as a standalone held-out domain; they were included in country-specific analyzes to preserve model generality [27]. The dataset analyzed in this study is the ROCK/10/4025 database, a published global compilation of intact rock parameters[27]. No primary field collection was undertaken by the present authors.

Four input characteristics were chosen due to their recognized physical correlations with UCS: P-wave velocity (V_p), dry density (ρ_d), point load strength index (Is_{50}), and rock classification. The Is_{50} parameter was absent in 58.2% of

the samples. Instead of filling in absent values, we implemented a binary indicator for missingness ($Is_{50_missing}$) to maintain data availability information while enabling tree-based models to manage missing values inherently [15,18,20]. This method circumvents the bias caused by imputation and capitalizes on the capacity of tree ensembles to learn from partial data. Figure 1 demonstrates distribution of dataset parameters. The four distributions exhibit significantly varied forms and spread. Is_{50} (a) has a pronounced right skew and bimodal distribution, characterized by a prominent mode around 3 MPa, a median of 3.09 MPa, and a constricted interquartile range (2.78–3.40 MPa), indicative of the restricted strength variability of the assessed specimens and the 58.2% data absence that limits its statistical significance. The UCS (b) ranges across several orders of magnitude (33–255 MPa) with a median value of 88 MPa and a broad interquartile range (79.9–135 MPa), validating the lithological diversity of the aggregated dataset and endorsing the log-transformation used in the physics baseline. The density ρ_d (c) is closely unimodal at 2.53 g/cm³, with interquartile boundaries of 2.47–2.59 g/cm³, suggesting limited distinguishing capability independently. V_p (d) represents the widest and most nearly symmetrical distribution, including a median of 4.56×10^3 m/s and a lower-tail mode approximately between $2.6\text{--}3.0 \times 10^3$ m/s, which is indicative of worn or significantly fragmented specimens. The extensive distribution and almost universal accessibility provide V_p the most enlightening continuous predictor a trend that aligns with its significant feature relevance (42.4%) in the Physics-Residual model and with recognized velocity-strength relationships.



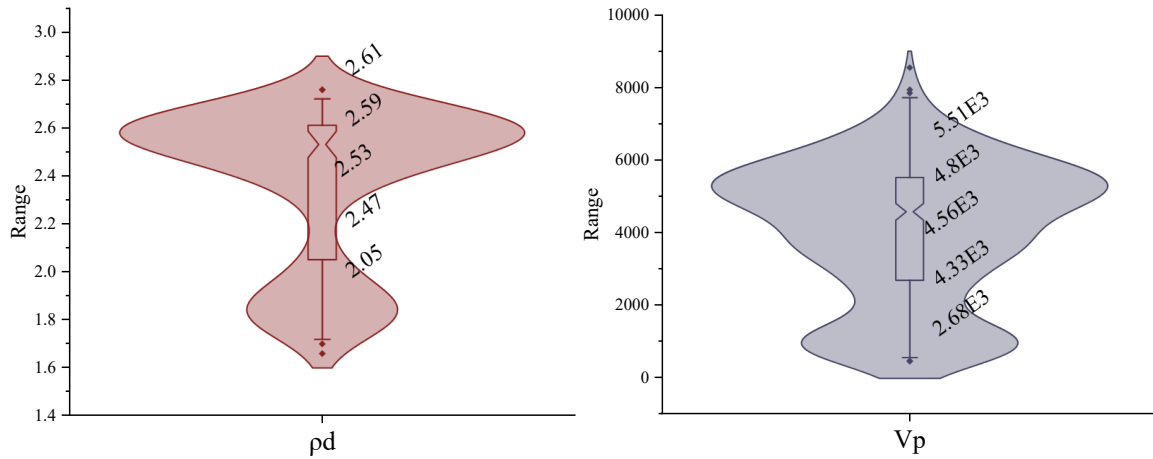


Figure 1. Violin-boxplots of (a) I_{s50} , (b) UCS, (c) p_d , and (d) V_p across the pooled dataset ($n = 397$; I_{s50} available for 167 samples).

The target variable, UCS (σ_c), showed a positively skewed distribution with values ranging over many orders of magnitude. Logarithmic transformation was used for physics-oriented modeling: $\log(\sigma_c) = \ln(\sigma_c)$. In a similar manner, V_p , p_d , and I_{s50} underwent logarithmic transformation for the foundational physics model. Three insights arise from the Figure 2. Initially, the correlation framework is mostly influenced by the V_p –UCS combination, which demonstrates the strongest and most consistent link, affirming that velocity is the principal indicator for strength forecasting. Secondly, p_d has a less robust and more dispersed association with UCS, and its connection with V_p is clearly nonlinear, with data points congregating along a curved boundary rather than forming an elliptical shape. Third, I_{s50} has a modest positive correlation with both V_p and UCS; nevertheless, the limited sample size ($n = 167$) and the clustering of values around 3 MPa result in a constricted confidence ellipse and a visually tenuous relationship. The matrix collectively endorses V_p as the principal physics anchor, while p_d and I_{s50} provide further corrections, forming a structure that corresponds with the feature significance ranking and the hierarchical framework of the physics baseline.

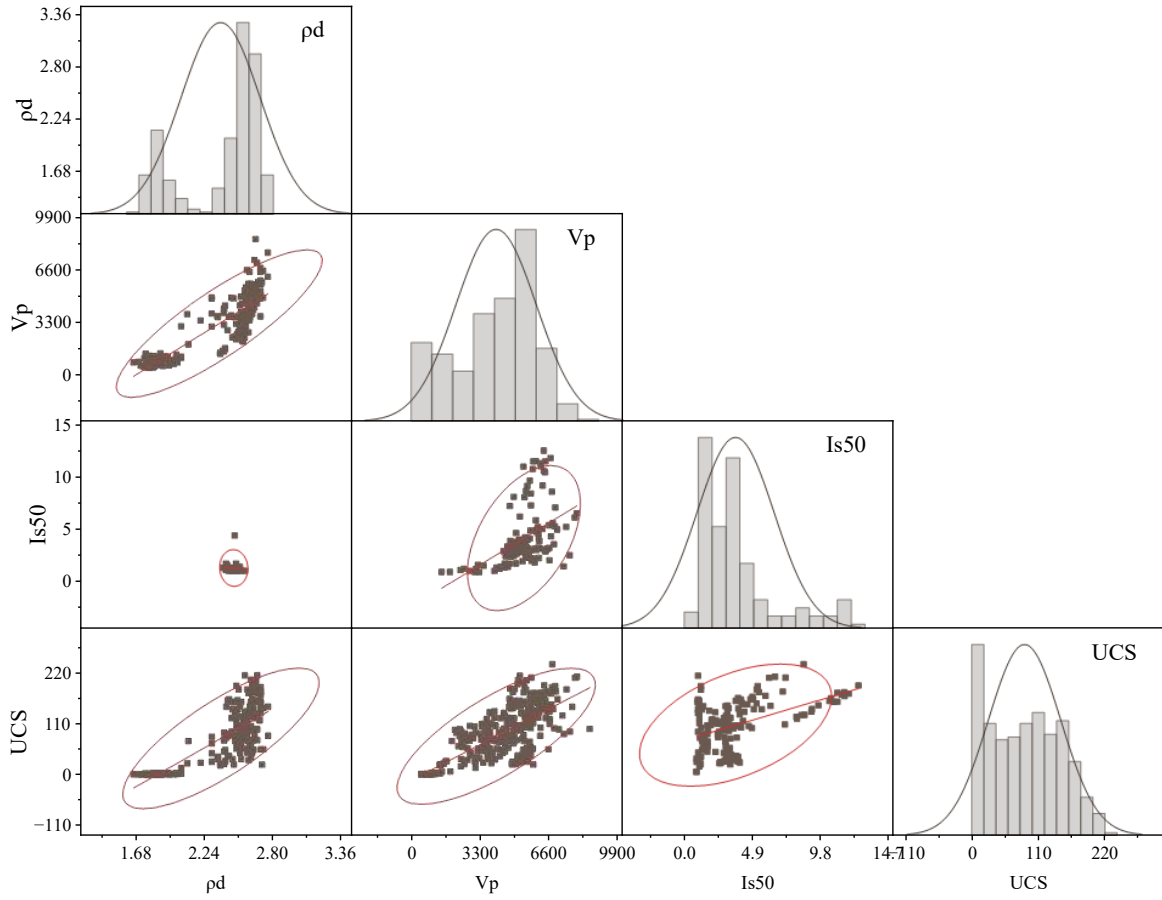


Figure 2. Scatterplot matrix of the four principal variables with histograms on the diagonal, 95% confidence ellipses, and pairwise scatter plots below the diagonal (n = 397; Is₅₀ available for 167 samples).

3.2. Physics Baseline Formulation

The physics baseline was established as a tiered collection of log-linear models with coefficients limited to feasible physical limits [14,15]. Three categories were established according to the accessibility of data:

Mode B (Complete Model) $\log(\sigma_c) = a \cdot \log(Vp) + b \cdot \log(\rho_d) + c \cdot \log(Is_{50}) + d$, with a, b, c > 0 (strength escalates with velocity, density, and point load index).

Mode A (Velocity-Density): $\log(\sigma_c) = a \cdot \log(Vp) + b \cdot \log(\rho_d) + d$, applicable in the absence of Is₅₀.

Mode VP (Velocity Only): $\log(\sigma_c) = a \cdot \log(Vp) + d$, used as a contingency when density data is also lacking.

298 Mode Edyn (Dynamic Modulus): $\log(\sigma_c) = a \cdot \log(E_{\text{dyn}}) + d$, with Edyn defined as

299
$$E_{\text{dyn}} = \rho_d \cdot Vp^2 \cdot \frac{1-2\nu}{1-\nu} / 1000, \text{ supposing } \nu = 0.25.$$

300 For every mode, Ridge regression with limitations on positive coefficients ($\alpha = 0.1$) was used to provide parameter
301 estimations that are physically significant. This restricted regression methodology inhibits the model from acquiring
302 non-physical negative correlations between predictors and UCS. The worldwide average of $\log(\sigma_c)$ functioned as the
303 last recourse in the absence of predictors.

304 **3.3.Predictive Architectures**

305 Four prediction frameworks were assessed to determine the impact of physics anchoring:

306 Direct ML: Fundamental learners were trained directly on the original attributes to forecast $\log(\sigma_c)$. This exemplifies
307 the traditional data-centric methodology that does not integrate any physics understanding [4,5].

308 Physics-Guided Features (PGFE): The physics baseline prediction ($\log(\sigma_c)_{\text{physics}}$) was included as an extra feature
309 in conjunction with the original inputs. Base learners were then instructed to forecast $\log(\sigma_c)$ using the enhanced
310 feature set. This framework enables the machine learning model to assimilate adjustments to the physics foundation
311 while preserving the inherent characteristics.

312 Base learners were instructed to forecast the discrepancy between the observed values and those anticipated by
313 physics: $r = \log(\sigma_c)_{\text{observed}} - \log(\sigma_c)_{\text{physics}}$. The ultimate forecasts were determined using the equation
314 $\log(\sigma_c)_{\text{predicted}} = \log(\sigma_c)_{\text{physics}} + \hat{r}$, where $\hat{r}$ represents the machine-learned residual adjustment. This design
315 compels the model to acquire knowledge just about the divergence from the physical rule, rather than the definitive
316 UCS value [22].

317 Physics-Constrained Monotonic Residual: Identical to Physics-Residual, however with monotonic constraints applied
318 to the first three features (Vp , ρ_d , $Isso$), guaranteeing that residual adjustments maintain physical monotonicity.

319 Four foundational learners were utilized: Random Forest (200 trees, maximum depth of 6, minimum samples per leaf
320 of 3) [17], CatBoost (200 iterations, learning rate of 0.03, depth of 5) [20], XGBoost (200 estimators, maximum depth

of 5, learning rate of 0.03, employing 'hist' tree method for native NaN management) [7,18], and LightGBM (200 estimators, maximum depth of 5, learning rate of 0.03, verbosity set to -1).

3.4. Leave-One-Country-Out Validation

The Leave-One-Country-Out (LOTO) methodology regarded each nation as a separate geological entity [8]. In each iteration, data from a single nation were excluded as the test domain, while the samples from the other five countries constituted the training set. This approach rigorously inhibits information leakage between training and testing domains while emulating the authentic circumstances of deploying a model in a novel geological context.

For every test domain, the retained samples were randomly divided into an assessment set (50%) and a calibration pool (50%). The calibration pool constituted the collection of accessible samples from which calibration specimens could be chosen; the evaluation set was only used for performance evaluation and was never employed for model training or calibration.

To evaluate statistical stability, the LOTO method was executed 30 times using various random divisions of the assessment and calibration datasets. All documented metrics denote averages and standard deviations over these 30 variations.

3.5. Few-Shot Calibration Protocol

For every calibration budget $k \in \{0, 1, 3, 5, 10, 20\}$, k samples were chosen from the calibration pool based on the assessed sampling method. In Equation 1 a bias correction factor was calculated as the median disparity between the observed and anticipated $\log(\sigma_c)$ values for the chosen calibration samples:

$$\Delta = \text{median}(\log(\sigma_c)_{\text{observed}}[\text{calib}] - \log(\sigma_c)_{\text{predicted}}[\text{calib}]) \quad (1)$$

The bias adjustment was then implemented on all predictions from the evaluation set: $\log(\sigma_c)_{\text{calibrated}} = \log(\sigma_c)_{\text{predicted}} + \Delta$. This median-centric adjustment is resilient to anomalies and offers a straightforward, comprehensible method for tailoring a universal model to specific local circumstances. For $k = 0$, no calibration occurred, and the predictions exemplified zero-shot extrapolation.

3.6. Active Sampling Strategies

Two proactive sampling methods were assessed in comparison to random selection:

Random Sampling: k samples were chosen evenly at random from the calibration reservoir.

Diversity Sampling: k-means clustering was executed on standardized non-destructive characteristics (V_p , ρ_d , I_{s50}) of the calibration pool specimens [25]. The specimen nearest to each cluster centroid was chosen, guaranteeing optimal representation of the feature space. When k equals 1, diversity sampling simplifies to random selection. This approach needs just non-invasive measurements (V_p , ρ_d , I_{s50}) and does not demand any UCS labels, rendering it entirely devoid of leakage.

3.7.Evaluation Metrics

Five metrics were computed for each architecture-model-k combination:

RMSE (MPa): Root mean squared error in original UCS units.

MAE (MPa): Mean absolute error.

R^2 : Coefficient of determination in original units.

$R^2_{\log}$: Coefficient of determination in log-space, which is more appropriate for geotechnical data spanning orders of magnitude.

MAPE (%): Mean absolute percentage error, with protection against division by zero.

3.8.Calibratability Index

The Calibratability Index (CI) was characterized as the proportion of the theoretical domain gap retrieved using k = 5 calibration samples in Equation 2:

$$CI = \frac{RMSE_{k=0} - RMSE_{k=5}}{RMSE_{k=0} - RMSE_{in-domain}} \quad (2)$$

$RMSE_{\{in-domain\}}$ denotes the performance inside the domain attained by 5-fold cross-validation in the reserved nation. A CI of 1.0 signifies that five calibration samples entirely bridge the domain gap; a CI of 0.5 denotes the attainment of fifty percent of the potential enhancement; a CI of 0 implies that calibration deteriorates forecasts.

3.9. Geological Bias Correction

The Geological Bias Correction (GBC) multiplier was calculated for every nation and model as the exponent of the median logarithmic bias in Equation 3:

$$GBC = \exp(\text{median}(\log(\sigma_c)_{\text{observed}} - \log(\sigma_c)_{\text{predicted}})) \quad (3)$$

A GBC value greater than 1 indicates that the target domain surpasses the global baseline (positive bias), while a GBC value less than 1 denotes inferior rock strength. The GBC offers a tangible metric for assessing geological divergence from the global model.

3.10. Statistical Analysis

Wilcoxon signed-rank tests were used for comparative analysis of matched designs [19]. The pairing was conducted based on nation and seed, resulting in 180 paired observations (6 countries $\times$ 30 seeds). Cohen's d was calculated to quantify the effect magnitude across all comparisons. Statistical significance was evaluated at $\alpha = 0.05$, using conventional criteria for effect size interpretation (small: $|d| < 0.5$; medium: $0.5 \leq |d| < 0.8$; large: $|d| \geq 0.8$).

3.11. Feature Set Ablation

Three feature sets were evaluated to assess the impact of feature engineering:

Original: V_p , ρ_d , Is_{50} , RockClass_Encoded, $Is_{50_missing}$ (5 features).

With_Edyn: Original + Edyn_GPa (6 features).

Log_Transformed: $\log(V_p)$, $\log(\rho_d)$, $\log(Is_{50})$, RockClass_Encoded, $Is_{50_missing}$, $\log(Edyn)$ (6 features).

A more extensive array of 32 manufactured characteristics (including dynamic modulus, acoustic impedance, porosity proxy, interaction terms, and country statistics) was assessed in initial trials but is not disclosed here owing to its persistently poor performance.

4. Results

4.1. Zero-Shot Extrapolation Performance

Table 1 illustrates the efficacy of all topologies over calibration budgets ranging from $k = 0$ to $k = 20$ using Diversity sampling. At zero-shot ($k = 0$), Direct ML recorded the minimal mean RMSE of 58.48 MPa, followed by PGFE at

58.18 MPa and Physics-Residual at 60.20 MPa. This first rating seems to endorse the exclusively data-centric methodology.

Table 1. Architecture Performance Summary (Original Features, Diversity Sampling)

Architecture	k=0 RMSE	k=3 RMSE	k=5 RMSE	k=10 RMSE	k=5 R ²	k=10 R ²	k=5 MAE	k=10 MAE
Direct ML	58.48	41.08	41.18	38.46	-0.255	-0.008	33.72	31.17
PGFE	58.18	39.13	38.02	36.51	0.050	0.162	30.94	29.37
Physics-Residual	60.20	37.72	36.82	34.90	0.057	0.180	28.39	26.56

Nonetheless, this ranking undergoes a significant reversal post-calibration. At $k = 3$, Physics-Residual attains an RMSE of 37.72 MPa, which is 3.4 MPa less than Direct ML's 41.08 MPa. At $k = 10$, Physics-Residual attains an RMSE of 34.90 MPa with a R^2 of 0.180, while Direct ML maintains an RMSE of 38.46 MPa with a R^2 of -0.008. The Physics-Residual architecture reliably attains the minimal RMSE and MAE across all calibration tiers above $k = 0$.

Figure 3 illustrates the RMSE learning trajectories for all designs using Random and Diversity sampling. The Physics-Residual curve intersects under the Direct ML curve at $k = 3$, indicating that the physics-based methodology surpasses the latter after just three calibration samples. The 95% confidence intervals from 30 permutations coincide at $k = 0$ but diverge as k rises, with Physics-Residual consistently exhibiting lower RMSE.

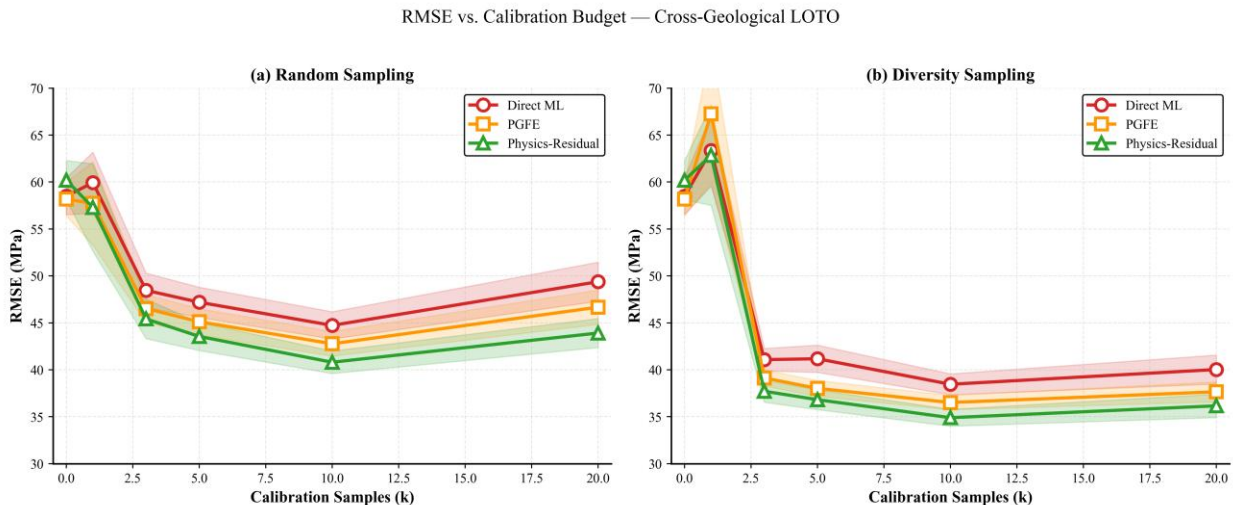


Figure 3. RMSE learning curves for all architectures under Random and Diversity sampling.

In addition to the RMSE analysis, Figure 4 presents R^2 in logarithmic scale, while Figure 5 displays R^2 in linear scale. Both measurements are included since they react differently to the extensive UCS spectrum present in the sample. Figure 4 illustrates R^2 (in logarithmic scale) in relation to the calibration budget. Panel (a) represents random sampling;

panel (b) illustrates diversity sampling. The shaded bars represent 95% confidence intervals derived from 30 LOTO permutations. The dashed line at $R^2_{\log} = 0$ signifies the mean-predictor boundary. At $k = 0$, Physics-Residual has the lowest performance ($R^2_{\log} \approx -1.15$) in comparison to PGFE (≈ -1.05) and Direct ML (≈ -0.95), aligning with the RMSE hierarchy. In Random sampling (panel a), Physics-Residual intersects $R^2_{\log} = 0$ at around $k = 8$, PGFE at $k \approx 12$, but Direct ML remains within the depicted range without crossing. In the context of Diversity sampling (panel b), Physics-Residual attains $R^2_{\log} \approx 0$ at $k \approx 10$ and increases to around 0.18 at $k = 20$, while Direct ML hovers around -0.15 . The vertical difference between Physics-Residual and Direct ML at $k = 20$ is about 0.35 $R^2_{\log}$ units.

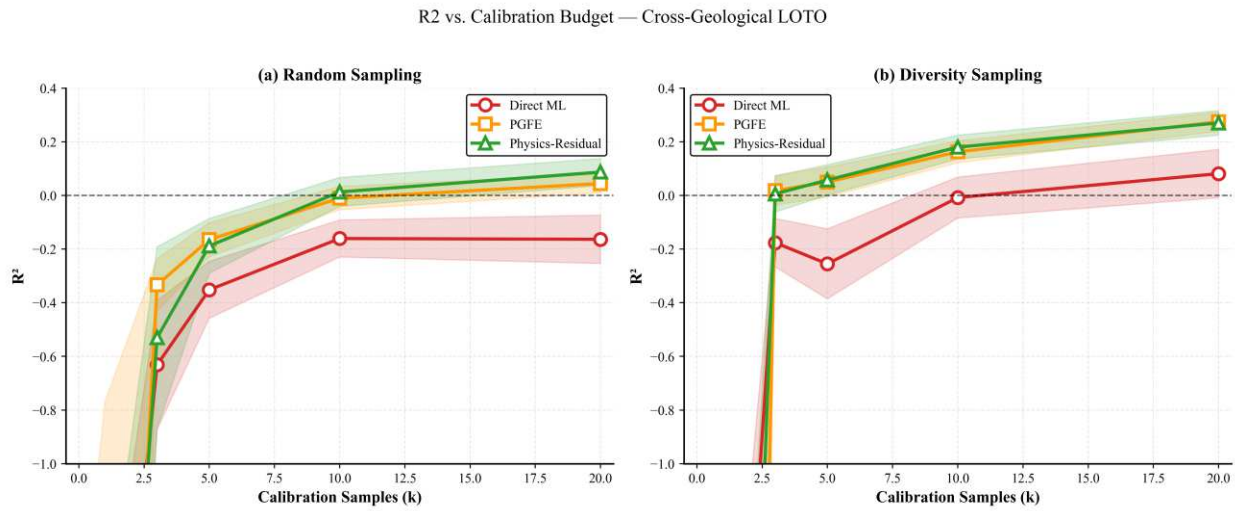


Figure 4. $R^2_{\log}$ learning curves for all architectures under Random and Diversity sampling.

Figure 5 illustrates R^2 (linear space) in relation to the calibration budget for identical designs and sampling methodologies. The dashed line at $R^2 = 0$ signifies the mean-predictor boundary. The linear-space R^2 exhibits a more rapid and pronounced recovery compared to $R^2_{\log}$. In Diversity sampling (panel b), both PGFE and Physics-Residual exhibit a cross R^2 of 0 at $k = 3$, whereas Direct ML intersects only around $k = 10$. At $k = 20$, Physics-Residual attains $R^2 \approx 0.28$, PGFE ≈ 0.22 , and Direct ML ≈ 0.08 . In Random sample (panel a), the Physics-Residual and PGFE converge to zero at about $k = 10$, but Direct ML continues to exhibit negative values around $k = 20$.

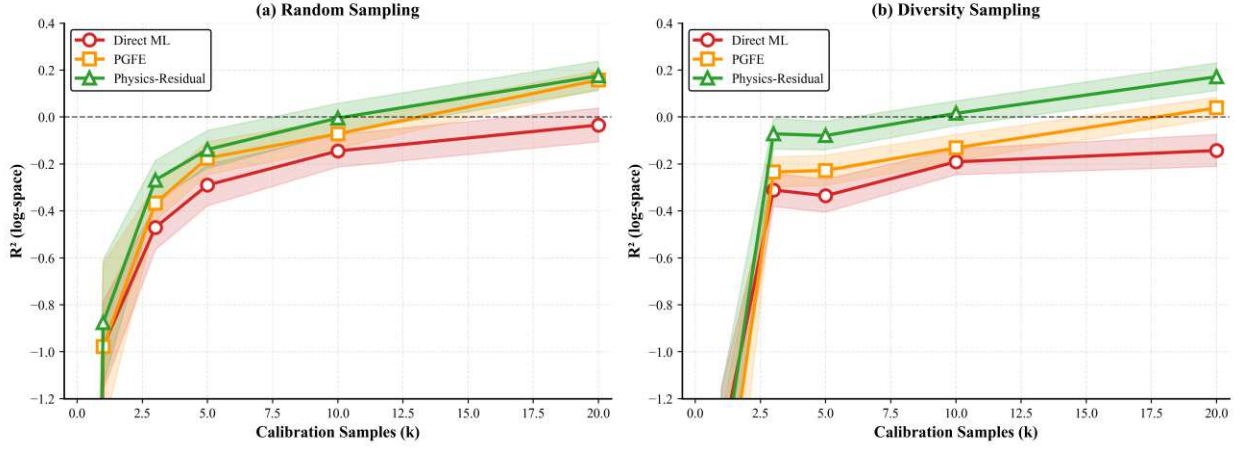


Figure 5. R^2 learning curves for all architectures under Random and Diversity sampling.

The disparity between Figure 4 and Figure 5 is enlightening. R^2 in linear space attributes an earlier crossover ($k \approx 3$ under Diversity) to the physics-based models, whereas $R^2_{\log}$ necessitates a greater budget ($k = 8-10$) to surpass the zero line. This suggests that the residual model swiftly reduces absolute errors in high-strength samples, however requires additional calibration data to mitigate proportional errors over the whole strength spectrum. The disparity between Physics-Residual and Direct ML is more significant in $R^2_{\log}$ than in R^2 , indicating that the benefits of physics anchoring are most evident when evaluating relative mistakes instead of absolute ones.

4.2. The $k=1$ Instability Phenomenon

A significant finding arises at $k = 1$: every architecture exhibits a rise in RMSE from $k = 0$ to $k = 1$ while using Diversity sampling. Direct ML rises from 58.48 to 63.03 MPa (+4.55 MPa), PGFE ascends from 58.18 to 66.10 MPa (+7.92 MPa), and Physics-Residual elevates from 60.20 to 63.20 MPa (+3.00 MPa). This contradictory result suggests that a solitary calibration sample lacks enough information to effectively rectify domain bias and might potentially introduce more inaccuracies if the chosen sample is not representative.

This discovery highlights the risks associated with single-sample calibration and advocates for a minimum calibration budget of $k \geq 3$ for field applications. The instability at $k = 1$ is seen in all base learners and designs, indicating that is an intrinsic characteristic of the bias correction process rather than a byproduct of any specific model.

4.3. Calibratability Index

Table 2 displays the Calibratability Index corresponding to each architecture. Physics-Residual attains the greatest CI (mean = 0.511, median = 0.558), followed by PGFE (mean = 0.450, median = 0.456) and Direct ML (mean = 0.364, median = 0.383). This indicates a 40% increase in calibration efficiency for Physics-Residual compared to Direct ML.

Table 2. Calibratability Index by Architecture

Architecture	k=0 RMSE	k=3 RMSE	k=5 RMSE	k=10 RMSE	k=5 R ²	k=10 R ²	k=5 MAE	k=10 MAE
Direct ML	58.48	41.08	41.18	38.46	-0.255	-0.008	33.72	31.17
PGFE	58.18	39.13	38.02	36.51	0.050	0.162	30.94	29.37
Physics-Residual	60.20	37.72	36.82	34.90	0.057	0.180	28.39	26.56

The median CI for Physics-Residual (0.558) surpasses the 0.5 benchmark, suggesting that in most instances, over half of the theoretical domain gap is bridged using five calibration samples. The adverse minimum CI values across all topologies suggest that in some country-seed pairings, calibrating with five samples deteriorates performance an observation that underscores the significance of sample selection and the danger of using unrepresentative calibration sets.

Figure 6 illustrates the confidence interval bar chart including 95% confidence intervals. The Physics-Residual framework surpasses the 50% recovery benchmark, whilst PGFE and Direct ML remain behind it. The error bars for PGFE and Direct ML intersect, however Physics-Residual is distinctly separated from both.

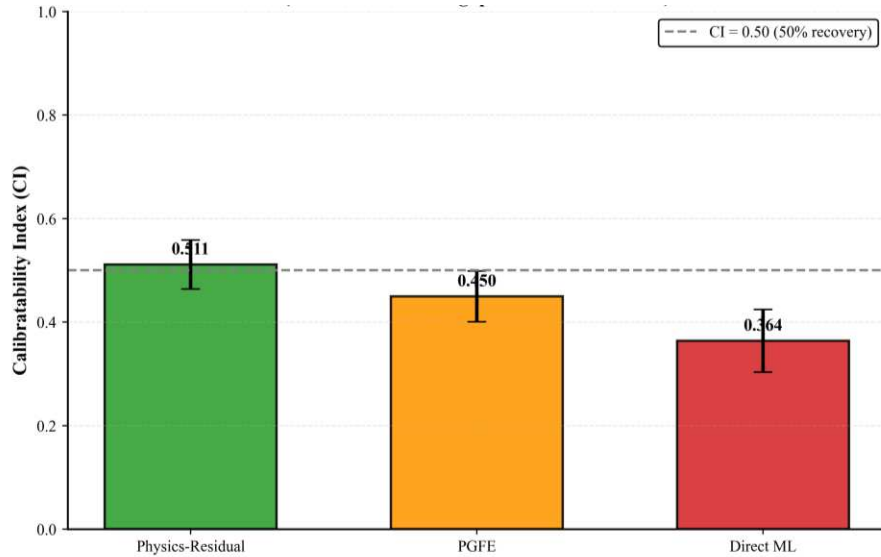


Figure 6. Calibratability Index by architecture with K=5

4.4.Active Sampling Strategy Comparison

Table 3 compares Random and Diversity sampling strategies for the Physics-Residual architecture. Diversity sampling consistently outperforms random selection across all calibration budgets. At $k = 3$, Diversity reduces RMSE from 45.42 to 37.72 MPa a 17% improvement. At $k = 5$, the improvement is from 43.55 to 36.82 MPa (16%), and at $k = 10$, from 40.81 to 34.90 MPa (14%).

Table 3.

k	Random RMSE	Diversity RMSE	Improvement	Random R^2	Diversity R^2
3	45.42	37.72	17%	-0.568	0.006
5	43.55	36.82	16%	-0.375	0.057
10	40.81	34.90	14%	0.061	0.180

Figure 7 juxtaposes Random and Diversity sampling for the Physics-Residual architecture over three calibration budgets. At $k = 3$, the average RMSE decreases from 45.4 MPa (Random) to 37.7 MPa (Diversity), representing a 17% decline. At $k = 5$, the value decreases from 43.6 to 36.8 MPa (16%), and at $k = 10$, it diminishes from 40.8 to 35.0 MPa (14%).

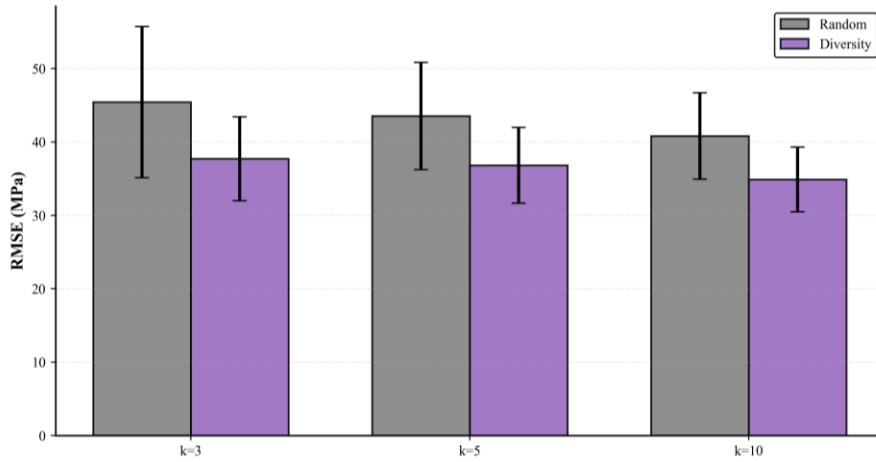


Figure 7. Sampling strategy comparison for the Physics-Residual architecture.

Initially, diversity sampling decreases both the average error and the breadth of the confidence interval across all budgets. At $k = 3$, Random exhibits a range of around 35–56 MPa, while Diversity ranges from 31 to 44 MPa. The reduced range suggests that calibration results are less influenced by the specific samples included in the pool – a realistically significant characteristic, since field practitioners cannot regulate sample accessibility. Secondly, the enhancement remains consistent over all three budgets instead of declining abruptly, validating that the benefit is systematic rather than a byproduct of a small k . Diversity sampling employs just non-invasive metrics (V_p , p_d , I_{s50})

to determine sites by k-means clustering, ensuring it is entirely free from leakage and applicable prior to drilling. The findings suggest that using a field research budget results in significantly enhanced calibration reliability when sites are chosen based on Diversity instead than random selection.

4.5.Per-Country Analysis and Geological Bias Decomposition

Table 4 presents per-country performance for the Physics-Residual architecture at $k = 5$ under Diversity sampling. Performance varies dramatically across countries, ranging from RMSE = 19.48 MPa for Portugal to 54.26 MPa for Malaysia.

Table 4. Per-Country Performance (Physics-Residual, $k=5$, Diversity)

Country	RMSE (mean $\pm$ std)	MAE	R^2	$R^2_{\log}$
Portugal	19.48 ± 3.71	14.49	0.632	0.816
Brazil	26.95 ± 5.24	22.22	0.513	-0.018
Turkey	28.04 ± 4.59	15.20	0.827	0.690
South Africa	43.80 ± 7.87	35.84	-0.837	-0.736
Iran	48.38 ± 8.96	40.61	-0.329	-0.771
Malaysia	54.26 ± 8.96	42.00	-0.462	-0.456

Figure 8 displays the per-country heatmap, visually confirming that Portugal, Brazil, and Turkey achieve positive R^2 , while South Africa, Iran, and Malaysia exhibit negative R^2 even after calibration. This suggests that for some geological domains, five calibration samples are insufficient to fully bridge the domain gap.

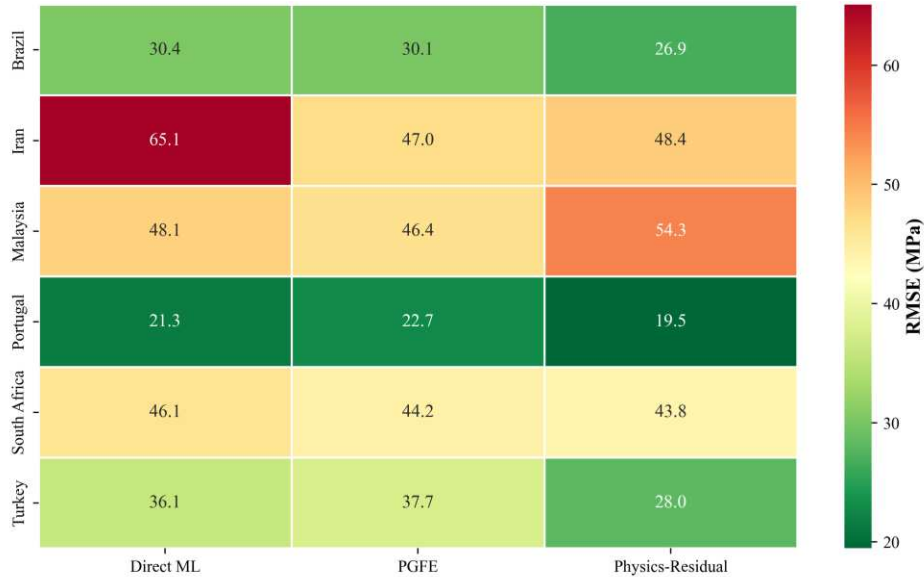


Figure 8. Per-country RMSE at $k=5$ under Diversity Sampling.

The Geological Bias Correction (GBC) multipliers, as shown in Table 5, provide a physical interpretation for these discrepancies. South Africa has a GBC of 5.60 $\times$, indicating that its geological formations are, on average, 5.6 times more robust than the world standard this aligns with the cratonic, low-porosity characteristics of South African rocks. Turkey has a GBC of 0.36 $\times$, indicating that its rocks are 2.8 times less robust than the baseline, aligning with the traits of a young orogenic belt, such as fracture and alteration.

Table 5. Geological Bias Correction (GBC) Multipliers by Country

Country	CatBoost	LightGBM	RandomForest	XGBoost	Mean
South Africa	4.455	5.435	7.668	4.851	5.602
Brazil	1.411	1.512	1.534	1.444	1.475
Portugal	1.456	1.328	1.536	1.364	1.421
Malaysia	0.779	0.875	0.865	0.821	0.835
Iran	0.597	0.542	0.560	0.539	0.560
Turkey	0.268	0.356	0.360	0.448	0.358

Figure 9 visualizes the GBC multipliers as a horizontal bar chart. The clear separation between countries with GBC > 1 (South Africa, Brazil, Portugal) and GBC < 1 (Malaysia, Iran, Turkey) demonstrates that the bias correction is capturing systematic geological differences, not random noise.

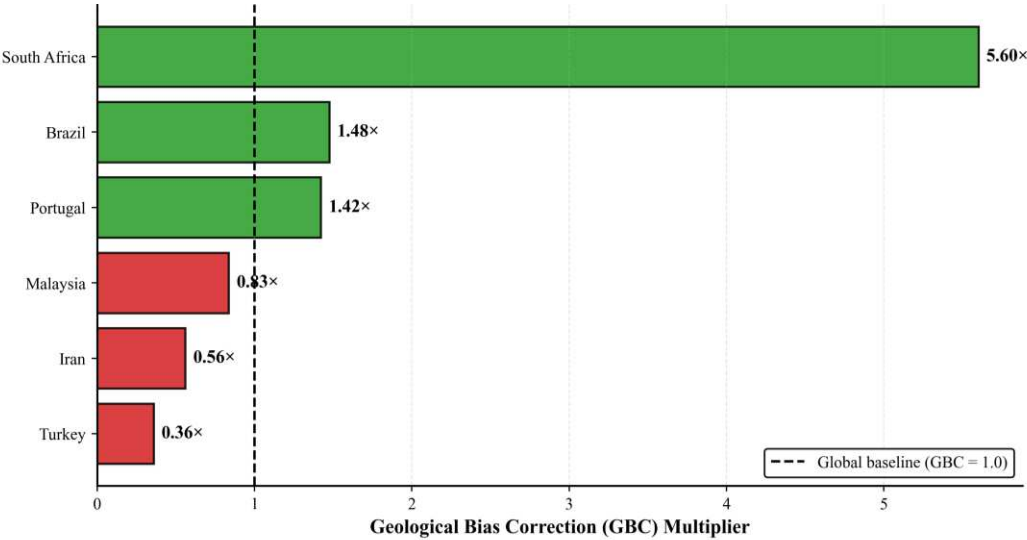


Figure 9. Geological Bias Correction multipliers by country.

4.6. Model-Wise Robustness

Table 6 presents the RMSE breakdown by base model and architecture at $k = 5$ under Diversity sampling. The Physics-Residual advantage is consistent across all four base learners: CatBoost achieves the lowest RMSE (36.6 MPa),

followed by XGBoost (35.8 MPa), LightGBM (36.5 MPa), and Random Forest (38.3 MPa). In contrast, Direct ML models range from 39.0 MPa (XGBoost) to 49.1 MPa (CatBoost).

Table 6. Model-Wise RMSE Breakdown at k=5 (Diversity, Original Features)

Architecture	Base Model	RMSE (MPa)	R ²
Direct ML	CatBoost	49.1	-0.983
Direct ML	LightGBM	37.4	0.056
Direct ML	RandomForest	39.2	-0.053
Direct ML	XGBoost	39	-0.040
Physics-Residual	CatBoost	36.6	0.083
Physics-Residual	LightGBM	36.5	0.104
Physics-Residual	RandomForest	38.3	-0.058
Physics-Residual	XGBoost	35.8	0.1

Figure 10 presents the model-wise heatmap, visually confirming that the Physics-Residual column is consistently darker (lower RMSE) than the Direct ML column across all base models. The CatBoost model under Direct ML is particularly poor (49.1 MPa), while the same algorithm under Physics-Residual achieves 36.6 MPa a 25% improvement.

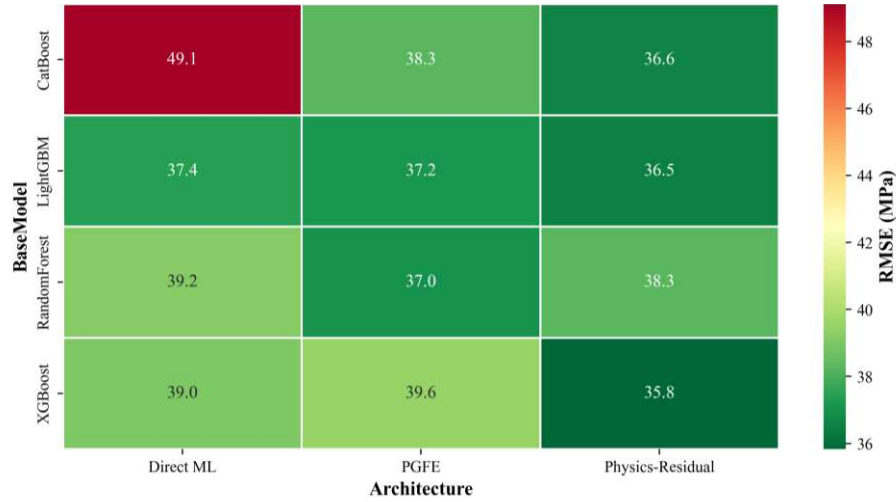


Figure 10. Base-model and architecture RMSE comparison at k=5 under Diversity sampling.

4.7.Feature Importance

Figure 11 illustrates the CatBoost feature priority hierarchy for the Physics-Residual model, averaged across all excluded nations. Vp leads with 42.4%, followed by RockClass_Encoded at 26.6%, pd at 17.7%, Isso at 10.2%, and Isso_missing at 3.1%.

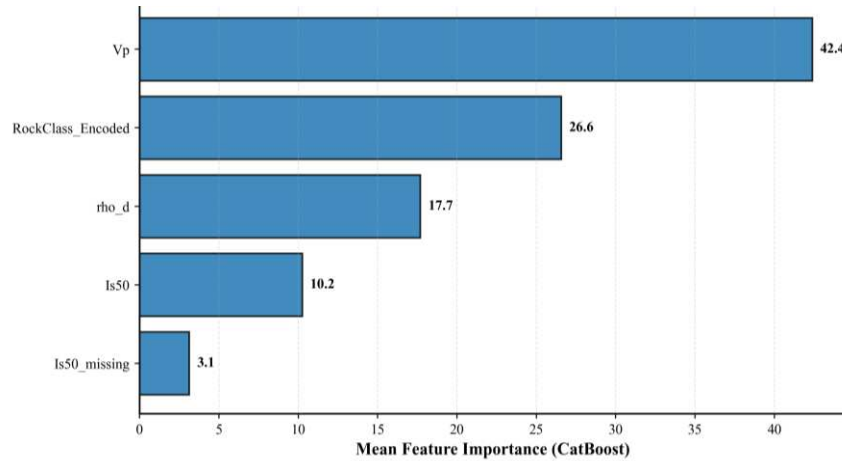


Figure 11. Feature importance for the Physics-Residual model.

Initially, the prevalence of V_p aligns physically with recognized rock mechanics: the velocity of elastic waves is intrinsically connected to elastic modulus, and both are associated with intact strength via same microstructural factors (grain contact, porosity, microfracture density). The significance of its presence in the residual model suggests that the physics baseline underestimates the V_p –UCS relationship across geological domains, with the residual learner adjusting by placing the most emphasis on V_p .

Secondly, the significant influence of `RockClass_Encoded` (26.6%) verifies that lithology imparts knowledge that surpasses the data provided by the continuous predictors. This aligns with the Geological Bias Correction study presented in Section 4.5, which indicates that nations dominated by volcanic and sedimentary formations have contrasting bias orientations. The residual model thus acquires a correction that is contingent upon lithology rather than being entirely continuous.

Third, the minimal significance of Is_{50} (10.2%) should not be interpreted as an indication that point load data lack informative value. Due to the absence of Is_{50} in 58.2% of samples, its average significance is diminished across the dataset; the associated `Is50_missing` indication (3.1%) somewhat differentiates these two influences. The limited collective impact of these characteristics suggests that, within the current dataset, V_p and pd predominantly encapsulate the recoverable signal for cross-domain calibration, while Is_{50} provides ancillary enhancement when present.

4.8.Feature Set Ablation

Table 7 displays the outcomes of the feature set ablation study. The Original feature set attains the minimal RMSE (36.82 MPa) and maximal R^2 (0.057) compared to all assessed setups. Incorporating Edyn (37.30 MPa) or using log-transformed features (37.33 MPa) yields no improvement, as both alternatives exhibit marginally worse performance.

Table 7. Feature Set Comparison (Physics-Residual, $k=5$, Diversity)

Feature Set	RMSE	MAE	R^2	$R^2_{\log}$
Original	36.82	28.39	0.057	-0.079
With_Edyn	37.30	28.92	0.035	-0.095
Log_Transformed	37.33	28.95	0.035	-0.096

Figure 12 presents the feature set comparison as three panels, visually confirming that the learning curves for Original, With_Edyn, and Log_Transformed are nearly indistinguishable. This negative result validates the parsimony principle: with 397 samples, simpler feature sets avoid overfitting and generalize better than engineered alternatives.

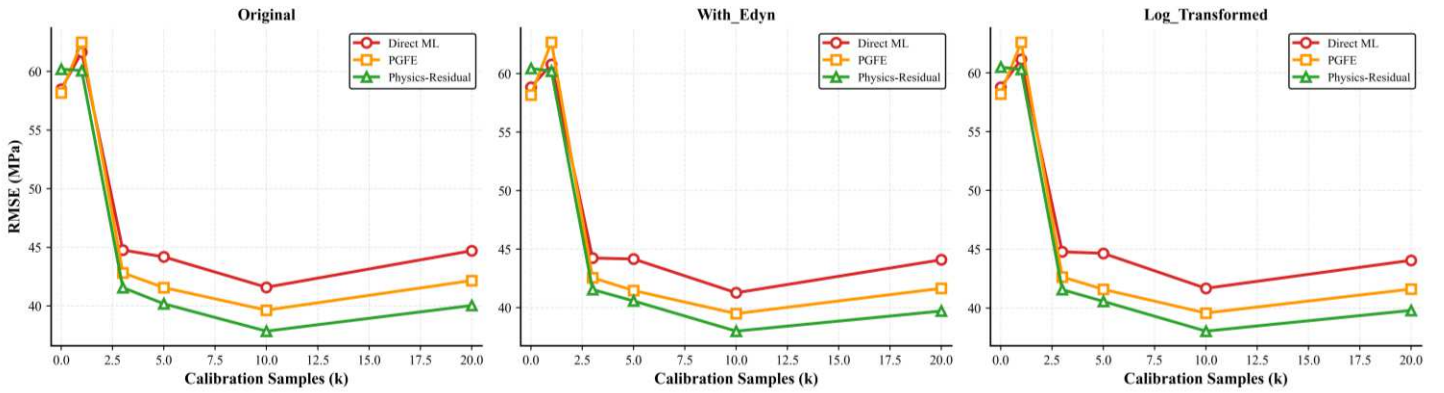


Figure 12. Feature set comparison for the Physics-Residual model.

4.9.Statistical Significance

Table 8 presents the Wilcoxon signed-rank test results comparing Physics-Residual and Direct ML at $k = 3, 5$, and 10 under Diversity sampling. All comparisons are statistically significant at $p < 10^{-8}$, with medium effect sizes at $k = 5$ ($d = -0.530$) and $k = 10$ ($d = -0.576$).

Table 8. Statistical Tests: Physics-Residual vs Direct ML (Diversity)

k	Wilcoxon W	p-value	Cohen's d	Effect Size	Significance
3	3996	3.09×10^{-9}	-0.395	Small	***
5	3551	5.29×10^{-11}	-0.530	Medium	***
10	3437	1.75×10^{-11}	-0.576	Medium	***

The negative Cohen's d values indicate that Physics-Residual has lower RMSE than Direct ML. The effect sizes grow from small at $k = 3$ to medium at $k = 5$ and $k = 10$, suggesting that the advantage of physics anchoring becomes more pronounced as calibration data accumulates.

4.10. In-Domain Baselines

Table 9 illustrates the in-domain efficacy attained by 5-fold cross-validation for each nation. These figures indicate the hypothetical upper limit for calibration the performance attainable if the model were only trained on data from the intended domain.

Table 9. In-Domain Baselines (5-fold CV within Country)

Country	RMSE	MAE	R^2	$R^2_{\log}$
Turkey	14.23	6.92	0.949	0.974
Portugal	13.84	10.90	0.723	0.683
Iran	21.28	15.84	0.705	0.679
Brazil	16.44	14.36	0.681	0.635
Malaysia	35.41	26.37	0.347	0.366
South Africa	34.46	28.92	-0.439	-0.549

The in-domain findings demonstrate significant fluctuations in predictability across nations. Turkey and Portugal attain elevated in-domain R^2 values over 0.70, however South Africa has a negative R^2 of -0.439 despite in-domain training, indicating significant internal variability or unaccounted variables. The disparity in performance between in-domain and cross-domain (Table 4) measures the extent of domain shift for each nation.

5. Discussion

5.1. The Accuracy-Calibration Trade-off

The primary conclusion of this research is the presence of a critical accuracy-calibration trade-off in cross-geological UCS forecasting. Direct machine learning models get somewhat reduced zero-shot RMSE compared to physics-based designs; however, this seeming benefit is negated post-calibration. The basic mechanism of this trade-off is that physics anchoring limits the hypothesis space of the model, compelling it to learn just the deviations from recognized physical rules instead of the whole input-output mapping [10]. This limitation diminishes the model's adaptability for in-domain interpolation while altering its extrapolation performance from erratic to systematically amendable.

The $k = 1$ instability phenomena offer clear proof of this process. Utilizing a solitary calibration sample for bias correction results in a significant error escalation for Direct ML (+4.55 MPa), while Physics-Residual has a more

moderate rise (+3.00 MPa). The reduced deterioration for Physics-Residual suggests that its error landscape is more uniform and conducive to straightforward bias adjustment [22]. This discovery has significant practical consequences: field practitioners need to refrain from single-sample calibration and, instead, gather at least three calibration samples.

5.2. Why Physics Anchoring Enables Calibratability

The Calibratability Index measures the tangible advantages of physics anchoring: Physics-Residual bridges 51.1% of the theoretical domain gap using five samples, in contrast to 36.4% for Direct ML. This 40% enhancement in calibration efficiency results in significant financial savings: whereas Direct ML may need 10–15 calibration samples for equivalent accuracy, Physics-Residual does this with just 3–5.

The process responsible for this enhancement may be comprehended within the framework of residual learning [24]. By segmenting the prediction challenge into a physics-informed baseline and a machine-learning adjustment, the model's output range is limited to physically feasible values. The residual component reflects inherent geological discrepancies variations in weathering, alteration, or structure that the global physics model fails to account for. Due to the consistent nature of this bias, it may be accurately determined from a limited set of calibration samples [8]. The bias correction method calculating the median logarithmic difference between actual and forecasted values is theoretically synonymous with maximum likelihood estimation of a location shift parameter inside a log-normal error framework.

5.3. The Role of Active Sampling: Diversity Over Uncertainty

The discovery that diversity sampling surpasses random selection by 14–17% is both practically meaningful and conceptually intriguing [25]. Diversity sampling chooses calibration samples that optimize the representation of the feature space characterized by non-invasive measures (V_p , ρ_d , I_{ss0}). This guaranties that the bias adjustment is derived from samples that accurately reflect the whole spectrum of geological conditions at the location.

Significantly, Diversity sampling does not need any UCS labeling for sample selection it functions only on non-invasive measures often obtained during standard site assessments [26]. This renders it instantly applicable in real-world scenarios: a location may be defined using V_p and density logs, the diversity algorithm can pinpoint the most valuable sites for core retrieval, and destructive testing can be restricted to 3–5 specimens.

The poor performance of Physics-Disagreement sampling (results not shown) is a significant adverse discovery. Choosing samples where the physics baseline and machine learning predictions are most discordant based on the premise that these examples provide the greatest insight ultimately diminishes performance, since these samples often represent geological anomalies that skew the correction factor away from the domain mode [24]. This implies that in few-shot regression calibration, having representative coverage is prioritized above resolving uncertainty.

5.4. Geological Bias as Systematic, Interpretable Signal

The Geological Bias Correction multipliers provide a physical understanding of cross-domain discrepancies [8]. South Africa's GBC of $5.60\times$ indicates the cratonic, low-porosity characteristics of its lithologies, which consistently exhibit greater strength than the world mean [28]. Turkey's GBC of $0.36\times$ indicates a youthful orogenic belt environment, when fracture and modification diminish rock integrity. Iran's GBC of $0.56\times$ aligns with carbonate-rich lithologies that exhibit more fracturing than worldwide norms.

These GBC values are not just statistical anomalies; they represent physically significant factors that may be understood by engineering geologists. The uniformity of GBC values across foundational models (e.g., South Africa spans from 4.46 for CatBoost to 7.67 for Random Forest) suggests that the bias is inherent to the geological domain rather than the learning method. This discovery corroborates the physics-based methodology: by compelling the model to comprehend divergences from physical principles, we get elucidative adjustments that mirror geological truths [22].

5.5. Feature Engineering and the Parsimony Principle

The findings of the feature set ablation indicate that including engineering characteristics such as dynamic Young's modulus, acoustic impedance, porosity proxies, and country statistics does not enhance the performance compared to the initial five features. This unfavorable outcome has scientific significance and possesses practical ramifications [28]. The model, using 397 samples and 5 features (about 79 examples per feature), functions in a context where the inclusion of more features heightens the likelihood of overfitting without contributing authentic data. The initial attributes (V_p , ρ_d , I_{s50} , rock class) effectively encapsulate the fundamental physical correlations, and tree-based models are capable of discerning nonlinear interactions among these variables without the need for explicit design.

The inability of log-transformed characteristics to enhance performance is very significant. Tree-based models are unaffected by monotonic transformations, therefore applying a logarithmic transformation to inputs does not alter the model's hypothesis space [18,20]. The physics baseline functions in logarithmic space, encapsulating the power-law

correlations between V_p and UCS. This discovery indicates that in tree-based models used for geotechnical data, feature engineering need to prioritize physical significance above mathematical alteration.

5.6.Comparison with Existing Approaches

The findings of this research align well with the existing literature on transfer learning in geotechnical engineering. The disparity in domains seen across nations is comparable to the variances in soil distributions between Finland and Bangladesh, where instance-weighted kernel ridge regression attained a R^2 of 0.876, albeit considerable statistical divergences between the areas [8]. The Calibratability Index introduced herein offers a broader framework for measuring adaptation efficacy applicable to any regression issue including domain shift.

The physics-based methodology corresponds with the PIML framework, which has shown improvements in the clarity and dependability of prediction models in geotechnical contexts [22]. The discovery that physics anchoring enhances calibratability contributes to the PIML literature by revealing a particular mechanism residual learning that facilitates effective domain adaptation. This enhances prior research on PINNs with rock mechanics issues [10], which has mostly concentrated on addressing forward and inverse concerns instead of domain adaptability.

The findings from active sampling enhance the expanding body of research on effective site characterization. The evidence that diversity-oriented selection through non-invasive measurements can diminish calibration needs by 3–5 times compared to random sampling corresponds with the active learning results [25], who demonstrated that utilizing merely half of the high-fidelity data allowed their multi-fidelity framework to attain equivalent accuracy to methods lacking active learning.

5.7.Practical Engineering Implications

The results of this research directly inform a realistic operational process consisting of four stages. Initially, do non-invasive site assessment using V_p and density evaluations across the area. Secondly, implement Diversity sampling on the non-invasive measures to choose 3–5 ideal sites for core extraction for destructive UCS evaluation. Firstly, calculate the Geological Bias Correction from the calibration samples by using the median log-space disparity. Fourth, implement the bias adjustment to the global physics-residual model to provide site-specific UCS forecasts.

This procedure decreases the necessary quantity of damaging UCS tests from the usual 30–50 needed for statistical validity in traditional site assessments to just 3–5 tests, representing a decrease of about 90%. The typical cost for each

UCS test is from \$300 to \$500, including core recovery, specimen preparation, and testing, resulting in savings of \$8,000 to \$22,000 per location. Significantly, it facilitates expedited site characterization while maintaining forecast accuracy [28].

6. Limitations and Future Work

The study's many constraints must be recognized. Initially, the dataset consists of just 397 samples from six nations, hence constraining the statistical efficacy of cross-domain analyzes. The category of metamorphic rocks is significantly underrepresented ($n = 3$), hindering any inferences related to metamorphic processes. Secondly, the nation was used as a surrogate for geological domain, failing to account for intra-country geological variability. A more detailed geological categorization (e.g., at the formation level) would provide a more stringent evaluation of cross-domain generalization. Thirdly, the R^2 values, even at $k = 10$, are still low (0.180 for Physics-Residual), suggesting that a significant amount of variation remains unexplained. Fourth, the research only assessed tree-based base learners; neural network frameworks and Gaussian processes were not examined and might provide other trade-offs. Fifth, the bias correction process presupposes a uniform log-space offset across the domain, which may not be applicable to geologies exhibiting significant intra-domain variability. Sixth, the absence of I_{s50} data (58.2%) creates ambiguity in predictions for samples without point load information, however the missingness indication somewhat alleviates this issue

Subsequent investigations need to augment the dataset by include other nations and lithological varieties, especially metamorphic rocks, to facilitate wider generalization. The framework has to be assessed using neural network base learners and Gaussian process models to see whether the calibratability benefit of physics anchoring applies to other model categories. Uncertainty quantification must be included into the calibration framework to get probabilistic forecasts accompanied by confidence intervals. The development of a Geological Similarity Index, which can be calculated only from non-destructive measures, is necessary to anticipate calibration needs (k^*) prior to doing any destructive tests. Ultimately, the framework needs to be corroborated with supplementary geomechanical characteristics beyond unconfined compressive strength, such as Young's modulus, tensile strength, and friction angle

7. Conclusion

This research has tackled the difficulty of predicting cross-geological UCS using a physics-based few-shot calibration framework assessed by rigorous Leave-One-Country-Out validation. The primary conclusions are as follows.

Initially, models reliant only on data exhibit severe failures when confronted with geological domain shifts, resulting in negative R^2 values during zero-shot extrapolation. Physics-based designs attain somewhat elevated zero-shot error rates while demonstrating significantly enhanced calibratability.

Secondly, the calibratability Index a new measure assessing the proportion of the domain gap bridged with five calibration samples stands at 0.511 for Physics-Residual compared to 0.364 for Direct ML, indicating a 40% improvement in calibration efficacy.

Third, diversity-oriented active sampling using non-invasive V_p and density assessments surpasses random sampling by 14–17% throughout all calibration budgets, facilitating dependable predictions with only 3–5 judiciously chosen UCS tests.

Fourth, Geological Prejudice Correction multipliers vary from 0.36 in Turkey to 5.60 in South Africa, indicating that cross-domain error is systematically organized and geologically significant rather than random noise.

Fifth, feature engineering does not enhance the original predictors, hence confirming the parsimony principle for limited geotechnical datasets.

These results provide a feasible, statistically robust procedure for field implementation: non-invasive site mapping succeeded by physics-informed few-shot calibration using 3–5 ideal samples. The framework is independent of the model, offering uniform benefits across all foundational learners, and facilitates a decrease of over 90% in the need for destructive testing. The Calibratability Index offers a universal measure for assessing adaptation efficacy applicable not just to geotechnical engineering but also to any field governed by physical principles necessitating localized empirical adjustments.

Declaration of competing 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.

Data availability

The rock property dataset analysed in this study is the ROCK/10/4025 database, an open-access global compilation of published intact rock parameters curated by Ferentinou and Muzamhindo. The database is publicly available via Mendeley Data (<https://data.mendeley.com/datasets/vk9vs574sk/1>) and described in the accompanying data descriptor [27]. The authors of the present study did not conduct primary field collection; all 397 specimens were derived from this pre-existing, peer-reviewed compilation. No ethical approval or land-access permission was required because the study involves secondary analysis of published data. All code and derived analysis outputs generated during this study are available from the corresponding author on reasonable request.

Author contributions

L. L: Conceptualization, Data curation, Methodology, Software, Formal analysis, Writing – original draft, Supervision.

Funding

The author declares that no funds, grants, or other financial support were received during the preparation of this manuscript.

References

1. Bieniawski, Z. T. Determining rock mass deformability: experience from case histories. *International Journal of Rock Mechanics and Mining Sciences and* **15**, 237–247 (1978).
2. D7012, A. Standard test methods for compressive strength and elastic moduli of intact rock core specimens under varying states of stress and temperatures. at (2014).
3. ISRM. Suggested methods for determining the uniaxial compressive strength and deformability of rock materials. *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts* **16**, 135–140 (1979).
4. Özdemir, E. Uniaxial compressive strength prediction in rocks: a comprehensive review from empirical equations to AI methods. *Acta Geophysica* **2026 74:2 74**, 80- (2026).
5. Akosah, S., Gratchev, I. & Gidigas, S. S. R. A systematic literature review on the application of artificial intelligence techniques for rock strength estimation. *Neural Computing and Applications* **2025 37:25 37**, 20721–20753 (2025).
6. Ding, X., Hasanipanah, M. & Rezaei, M. Assessment of mechanical properties of rock using deep learning approaches. *Measurement: Journal of the International Measurement Confederation* **250**, (2025).

- 724 7. Huang, S., Yong, W., Wang, L. & Zhou, J. From measurement while drilling data to rock strength: a robust
725 LMWOA-XGBoost model for predicting uniaxial compressive strength with SHAP interpretability. *Expert*
726 *Systems with Applications* **325**, (2026).
- 727 8. Nishat, R. A., Rahman, F., Hoque, N. M. R. & Islam, S. Transfer learning for undrained shear strength
728 prediction: A case study with instance-based transfer. *Geosystems and Geoenvironment* **5**, (2026).
- 729 9. Guo, H., Zhang, C., Fang, H., Rabczuk, T. & Zhuang, X. Deep learning to evaluate seismic-induced soil
730 liquefaction and modified transfer learning between various data sources. *Underground Space (new)* **23**,
731 220–242 (2025).
- 732 10. Liu, Y. & Arson, C. A Physics-Informed Neural Network for Pressurized Cavities of Arbitrary Shape in
733 Heterogeneous Rock. *Rock Mechanics and Rock Engineering* 2026 1–21 (2026) doi:10.1007/S00603-026-
734 05882-5.
- 735 11. Kayabaşı, A. & Gökceoğlu, C. An assessment on the correlations between uniaxial compression strength
736 and point load index. *Arabian Journal of Geosciences* 2022 15:3 **15**, 268- (2022).
- 737 12. Kurtuluş, C., Sertçelik, F. & Sertçelik, I. Correlating physico-mechanical properties of intact rocks with P-
738 wave velocity. *Acta Geodaetica et Geophysica* 2015 51:3 **51**, 571–582 (2015).
- 739 13. Mohamad, J. D. *et al.* Correlation between Uniaxial Compressive Strength and Point Load Strength of
740 Penang Island Granites. *IOP Conference Series: Earth and Environmental Science* **971**, 012025 (2022).
- 741 14. Yagiz, S. P-wave velocity test for assessment of geotechnical properties of some rock materials. *Bulletin of*
742 *Materials Science* 2011 34:4 **34**, 947–953 (2011).
- 743 15. Sharma, P. K. & Singh, T. N. A correlation between P-wave velocity, impact strength index, slake durability
744 index and uniaxial compressive strength. *Bulletin of Engineering Geology and the Environment* 2007 67:1
745 **67**, 17–22 (2007).
- 746 16. Sun, Y. D. *et al.* Predicted uniaxial compressive strength of fractured weakly cemented rock by LSSVM-
747 NCV-ABKDE-based data-driven method. *Geomechanics and Geophysics for Geo-Energy and Geo-*
748 *Resources* 2025 11:1 **11**, 52- (2025).
- 749 17. Barzegar, R., Sattarpour, M., Deo, R., Fijani, E. & Adamowski, J. An ensemble tree-based machine learning
750 model for predicting the uniaxial compressive strength of travertine rocks. *Neural Computing and*
751 *Applications* 2019 32:13 **32**, 9065–9080 (2019).
- 752 18. Xie, Y., Li, X. & Min, Z. Comparison of machine learning models for rock UCS prediction using
753 measurement while drilling data. *Scientific Reports* 2025 15:1 **15**, 8434- (2025).
- 754 19. Zheng, X., Batool, A., Kumar, S. & Shahani, N. M. Grey Wolf Optimization-Optimized Ensemble Models
755 for Predicting the Uniaxial Compressive Strength of Rocks. *Applied Sciences* 2026, Vol. 16, Page 1130 **16**,

1130 (2026).

20. Ahmadi, H., Kozegar, E. & Nasiri Pishvari, M. Direct incorporation of categorical geotechnical variables in CatBoost-based prediction of unconfined compressive strength of stabilized saline soils under freeze–thaw cycles. *Engineering Geology* **373**, (2026).
21. Seo, S., Chung, M. & Kim, W. Challenging domain shift in excavation-induced wall deformation prediction: Feasibility of adversarial-discrepancy domain adaptation. *Measurement* **264**, 120294 (2026).
22. Yuan, B. *et al.* Physics-informed machine learning in geotechnical engineering: a direction paper. *Geomechanics and Geoengineering* **20**, 1128–1159 (2025).
23. Shim, S. Self-supervised domain adaptive approach for extrapolated crack segmentation with fine-tuned inpainting generative model. *Computer-Aided Civil and Infrastructure Engineering* **40**, 3333–3358 (2025).
24. Chen, S., Mao, W., Zhang, Y. & Zhong, Z. A novel few-shot deep regression domain adaptation method with wavelet scattering attention mechanism for fault prognostics. *Nondestructive Testing and Evaluation* **41**, 233–252 (2026).
25. He, G. F., Zhang, P. & Yin, Z. Y. Active learning inspired multi-fidelity probabilistic modelling of geomaterial property. *Computer Methods in Applied Mechanics and Engineering* **432**, 117373 (2024).
26. Zhang, P. & Yang, G. A knowledge-informed data-efficient framework for soil correlation modelling. *Computers and Geotechnics* **197**, 108190 (2026).
27. Ferentinou, M. & Muzamhindo, H. Introduction to ROCK/10/4025 database. *Geodata and AI* **5**, 100028 (2025).
28. Aladejare, A. E. & Wang, Y. Evaluation of rock property variability. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards* **11**, 22–41 (2017).