Scilight

Bulletin of Computational Intelligence
https://www.sciltp.com/journals/bci



Article

Check for updates

Credibility Assessment of Machine Learning Models for Slotted Cold-Formed Steel Beams

Lenganji Simwanda [1,2]

[1] Klokner Institute, Czech Technical University in Prague, 160 00 Prague, Czech Republic; lenganji.simwanda@northrise.net
[2] Faculty of Engineering and Technology, Northrise University, Ndola P.O. Box 24027, Zambia

How To Cite: Simwanda, L. Credibility Assessment of Machine Learning Models for Slotted Cold-Formed Steel Beams. *Bulletin of Computational Intelligence* **2026**, *2*(4), 386–407. https://doi.org/10.53941/bci.2026.100020

Received: 3 October 2026
Revised: 16 November 2026
Accepted: 23 November 2026
Published: 26 November 2026

Abstract: Machine-learning surrogates can accelerate structural assessment, but accuracy on a random split provides limited evidence for use outside the training domain. This study evaluates 14 regression methods on 432 paired numerical configurations of slotted cold-formed steel beams with mode-specific local- and distortional-buckling capacities. Geometry, steel-grade and slot-length hold-outs are combined with size-matched controls, matched-configuration transfer, training-only validation, split-conformal intervals, estimator repetitions, applicability screening and SHAP sensitivity. XGBoost attains in-distribution mean absolute errors of 0.064 and 0.105 kN m for the two targets, yet its geometry-shift errors rise to 13.813 and 9.150 kN m. Perforation transfer is strongly model-dependent: Gaussian-process errors reach 8.898 and 6.463 kN m, while leading tree models approximately match a baseline that transfers the corresponding known slot-length response. Reduced training size explains only a small part of the observed deterioration. Nominal 90% intervals also lose coverage under shifts; mean XGBoost geometry-shift coverage across 30 calibration allocations falls to 0.37% and 2.85%. A training-specific combined range and covariance screen rejects all shifted queries in this benchmark, while retaining 88.5% of the fixed in-distribution test set. These findings support a conditional credibility audit with explicit abstention and uncertainty checks rather than deployment based on aggregate accuracy alone.

Keywords: machine learning credibility; cold-formed steel; slotted beams; distribution shift; conformal prediction; applicability screening

1. Introduction

Machine-learning (ML) surrogates are increasingly used to accelerate structural analyses that would otherwise require large numbers of nonlinear finite-element (FE) simulations or experiments. In cold-formed steel (CFS) members, this opportunity is attractive because buckling resistance depends on strongly interacting geometric, material, and detailing variables, and high-fidelity nonlinear analyses can be expensive when repeated during parametric design or reliability studies [1,2]. Recent studies have demonstrated very high predictive accuracy for local- and distortional-buckling resistance of perforated CFS members using boosted and ensemble learners [3,4]. The engineering question, however, is not only whether a model reproduces held-out samples drawn from the same design grid, but whether its predictions remain defensible when a future query differs from the data used to train it.

Slotted CFS beams provide a demanding case study because perforations are introduced for functional and thermal reasons but create local stress redistribution and interact with plate slenderness and buckling mode. Recent work on perforated cold-formed members has shown that opening shape, opening length, edge stiffening, and section geometry can materially change bending resistance, and that validated FE parametric studies remain central to developing reliable design guidance [5–7]. These studies reinforce two features that make the present benchmark useful for credibility research: the response is highly nonlinear, and the design space is structured rather than continuously and uniformly populated.

Copyright: © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
Publisher's Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Applications of computational intelligence to structural resistance now extend across materials and response mechanisms. Silewu et al. [8] benchmarked ensemble models for the bending strength of ultra-high-performance concrete beams, while Subedi et al. [9] combined nested validation with SHAP analysis for geopolymer strength prediction. These studies illustrate the importance of separating predictive evaluation, model selection and interpretation. For deployment, an additional question is whether performance remains credible when the training evidence omits the regime represented by a new query.

The same question arises in monitoring and material modelling. Digital-twin approaches to tunnels connect sensor observations, virtual representations and predictive models [10]. Tree-based coal-abrasiveness models demonstrate that mineralogical descriptors, calorific properties and data quality influence prediction [11]. The literature on machine learning for 3D-printed soil–cement materials provides a further context in which material and process descriptors define the evidence available to a model [12]. These applications motivate explicit checks of training support and feature changes; they do not establish transferable performance for the present steel-beam benchmark.

Credibility is treated here as evidence of suitability for a stated context of use, consistent with verification, validation and uncertainty-quantification perspectives [13–15]. A random in-distribution (ID) hold-out measures performance within the sampled design grid. Out-of-distribution (OOD) hold-outs probe missing geometric, perforation and material regimes. Prediction intervals and estimator repetitions provide complementary evidence, but neither establishes acceptable extrapolation in the absence of relevant tests [16,17].

The contribution is an integrated structural case study and a reproducible decision audit, rather than a new OOD, conformal or SHAP algorithm. The geometric configurations and numerical capacities are reused from the source FE studies [18,19] and the associated prediction studies [3,4]. The present analysis adds reconciled case-level predictions, reverse slot-length transfer, a matched-configuration baseline, repeated size controls, OOD interval evaluation, training-specific support checks and linked-feature sensitivity. These analyses establish when strong random-split accuracy fails to justify a prediction and when the available evidence instead supports abstention.

2. Data and Benchmarks

2.1. Dataset Overview

The database contains 432 configurations of slotted cold-formed steel (CFS) lipped-channel beams and two paired targets, M_{LB} and M_{DB}. These are ultimate bending capacities obtained from separate nonlinear FE analyses for local- and distortional-buckling conditions. They are not elastic critical moments. The local-buckling models restrained distortional deformation, including compression-flange restraint in the pure-bending region; the distortional models allowed the compression flange to translate laterally in that region. The targets therefore describe two mode-specific, restraint-conditioned responses for a common geometric and material input vector [18,19].

Both source studies used ANSYS with SHELL181 elements, material and geometric nonlinearity, displacement-controlled four-point bending and a symmetric half-model representation. The selected parametric span was 2600 mm, with slotted perforations confined to the central pure-bending region and solid end shear spans. The distortional study identifies a 1000 mm central region. Web-side plates transferred the support and loading actions. The source studies compared 2600 and 4800 mm arrangements before selecting the shorter model. Solid regions were generally meshed at 5 × 5 mm and slotted regions at approximately 1.5 × 5 mm. The adopted elastic modulus and Poisson ratio were 200 GPa and 0.3, respectively [18,19].

The local study introduced the lowest local eigenmode with a positive amplitude of 0.15 times the wall thickness in its parametric analysis. Its validation considered both signs. The distortional study used distortional eigenmodes and assessed positive and negative amplitudes of 0.15, 0.64 and 0.94 times the thickness. Both studies included nonlinear material response and omitted the combined effects of residual stresses and corner-strength enhancement in accordance with their stated modelling assumptions. The capacities used here remain conditional on those modelling choices.

The original experimental comparisons are summarized in Table 1. The local verification used six solid-web lipped channels, with measured depths 153.43–203.74 mm, thicknesses 1.5–2.4 mm, flange widths 63.58–79.27 mm, lip widths 15.02–21.26 mm and yield stresses 483.49–541.13 MPa. Their test capacities ranged from 10.43 to 31.39 kN m. For the positive 0.15t analysis, the absolute test-to-FE differences calculated from the published rounded capacities range from 0.98 to 4.05 kN m, with an MAE of 2.17 kN m. The distortional verification included solid, rectangular-opening and slotted channels, so its validation evidence covers a broader set of opening configurations. These historical comparisons characterize the numerical evidence base; the present study evaluates ML agreement with its numerical labels.

Each target column was matched against all 432 case labels in the source parametric tables. Every capacity agrees within 0.005 kN m, consistent with the two-decimal values printed in those tables. The higher-precision spreadsheet values are retained for analysis. The input vector is

$$\mathbf{x} = \{D, B_f, B_l, t, L_{sl}, W_{sl}, S_{sl}, B_{sl}, N, n, F_y, r\}. \quad (1)$$

Here N is the number of slot-row groups and n is the total number of slot rows. Figure 1 distinguishes these descriptors. The database encodes $r = 2t$, while $B_l = 13$ mm, $S_{sl} = 100$ mm and $B_{sl} = 9.5$ mm are constant. These three constant inputs are retained in the reference input schema but excluded from covariance and correlation calculations. No missing entries, duplicate rows or duplicate input combinations were found.

Table 1. Experimental verification reported by the source FE studies. Ratios are Test divided by FE capacity.

Study	Verification Evidence	Imperfection Cases	Reported Test/FE Statistics
Local buckling [18]	6 solid-web lipped channels	+0.15t and −0.15t	Means 1.13 and 1.13; COV 0.030 and 0.035; ratio ranges 1.08–1.17 and 1.07–1.18
Distortional buckling [19]	20 specimens: 9 solid, 6 rectangular-opening and 5 slotted	±0.15t, ±0.64t and ±0.94t	Across the six assumptions: mean 0.98–1.06 and COV 0.06–0.08. At +0.94t: mean 1.03, COV 0.07, range 0.89–1.11

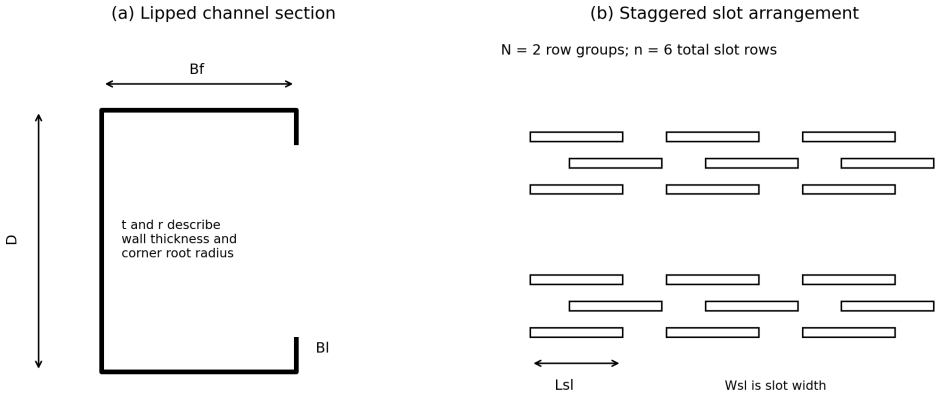


Figure 1. Schematic input definitions for the lipped channel and staggered slots. The row-group sketch illustrates N = 2 and n = 6; it is not drawn to scale.

2.2. Domain of Applicability and Dataset Characterization

The sampled domain is a discrete design grid. Thickness, radius and material grade have 144 cases at each of their three levels. Section depth and row count are unbalanced because permitted combinations depend on the section. A coordinate-wise range therefore describes an enclosing box, not a continuously populated design space. Table 2 summarizes the input schema and observed levels. Figure 2 shows the joint support and Figure 3 the correlations among the varying descriptors. The Pearson correlation coefficient is denoted by ρ to distinguish it from the root radius r. Perfect thickness–radius linkage makes separate causal interpretations of their model attributions inappropriate.

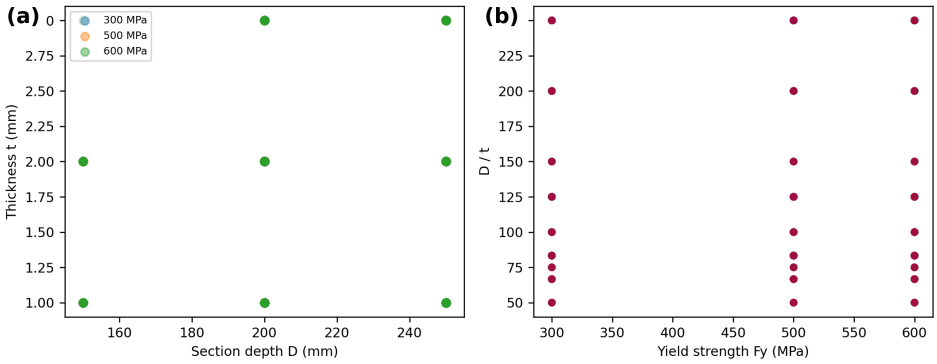


Figure 2. Joint support of the numerical database: (**a**) section depth D versus wall thickness t; (**b**) yield strength F_y versus the slenderness ratio D/t. Repeated points represent other combinations of the remaining descriptors.

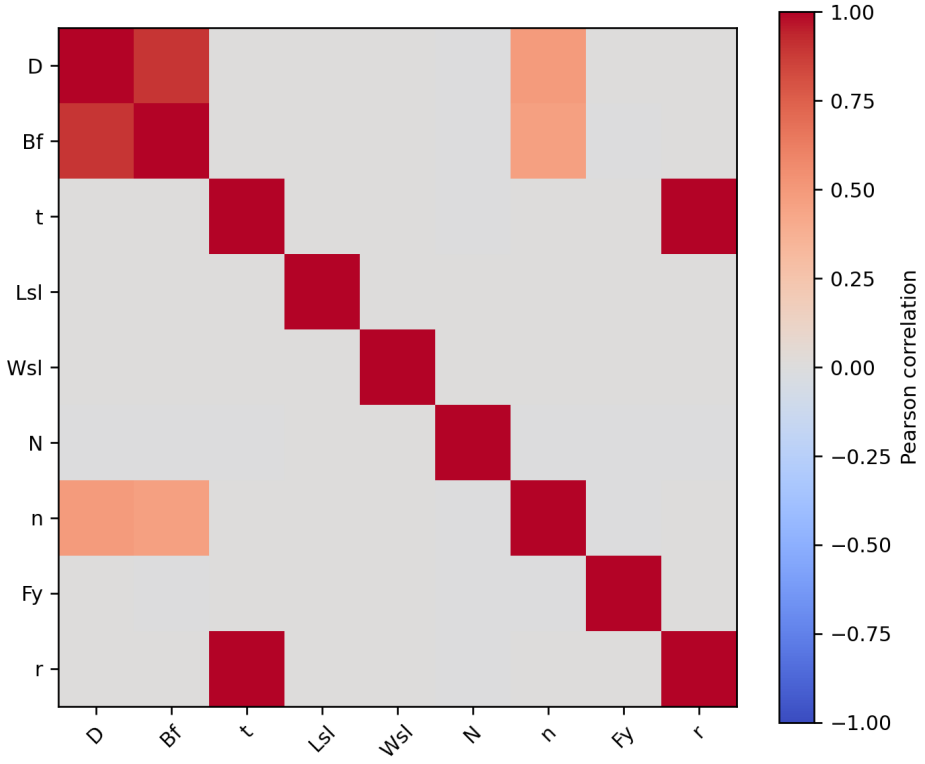


Figure 3. Pearson correlation coefficients for the nine varying input descriptors. The three constant descriptors are omitted.

Table 2. Input schema and observed levels for the 432 paired numerical configurations.

Variable	Meaning	Unit	Observed Levels or Constraint
D	Section depth	mm	150, 200, 250
Bf	Flange width	mm	45, 65; associated with section depth
Bl	Lip length	mm	13
t	Wall thickness	mm	1, 2, 3
Lsl	Slot length	mm	60, 75
Wsl	Slot width	mm	3, 5
Ssl	Slot-spacing descriptor	mm	100
Bsl	Slot-border descriptor	mm	9.5
N	Number of slot-row groups	—	1, 2
n	Total number of slot rows	—	6, 8, 12; constrained by depth
Fy	Yield strength	MPa	300, 500, 600
r	Root radius	mm	2, 4, 6; $r = 2t$

3. Credibility Assessment Framework

The framework separates predictive error, interval behaviour, estimator repeatability and training support. These quantities answer different questions and are evaluated using independent fitting and allocation rules. Figure 4 shows how the experiments branch from the original database. No withheld response is used to select the applicability threshold, fit preprocessing, choose hyperparameters or calibrate an interval.

3.1. Out of Distribution Stress Tests

Let the database be $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{432}$. A deterministic scenario rule $s(\mathbf{x})$ defines the withheld regime; the training partition contains all remaining cases:

$$\mathcal{D}_{\text{te}} = \{(\mathbf{x}_i, y_i) \in \mathcal{D} : s(\mathbf{x}_i) = 1\}. \tag{2}$$

The principal OOD scenarios omit the lower and upper tails of D/t, the 60 mm slot-length level or the 300 MPa

grade. A reverse slot-length experiment omits 75 mm. Each model is refitted from a fresh estimator on the corresponding training partition. Performance deterioration is summarized by the MAE inflation factor, defined once as

$$I_{m,s} = \frac{\mathrm{MAE}_{m,s}}{\mathrm{MAE}_{m,\mathrm{ID}}}. \tag{3}$$

The heatmaps display $\log_{10} I_{m,s}$ using unrounded errors. A value of 2 represents a 100-fold increase, while a negative value denotes a smaller scenario MAE than the ID MAE. Inflation is interpreted together with absolute error, training size and test-response variation; a small denominator can otherwise exaggerate the apparent severity of an increase.

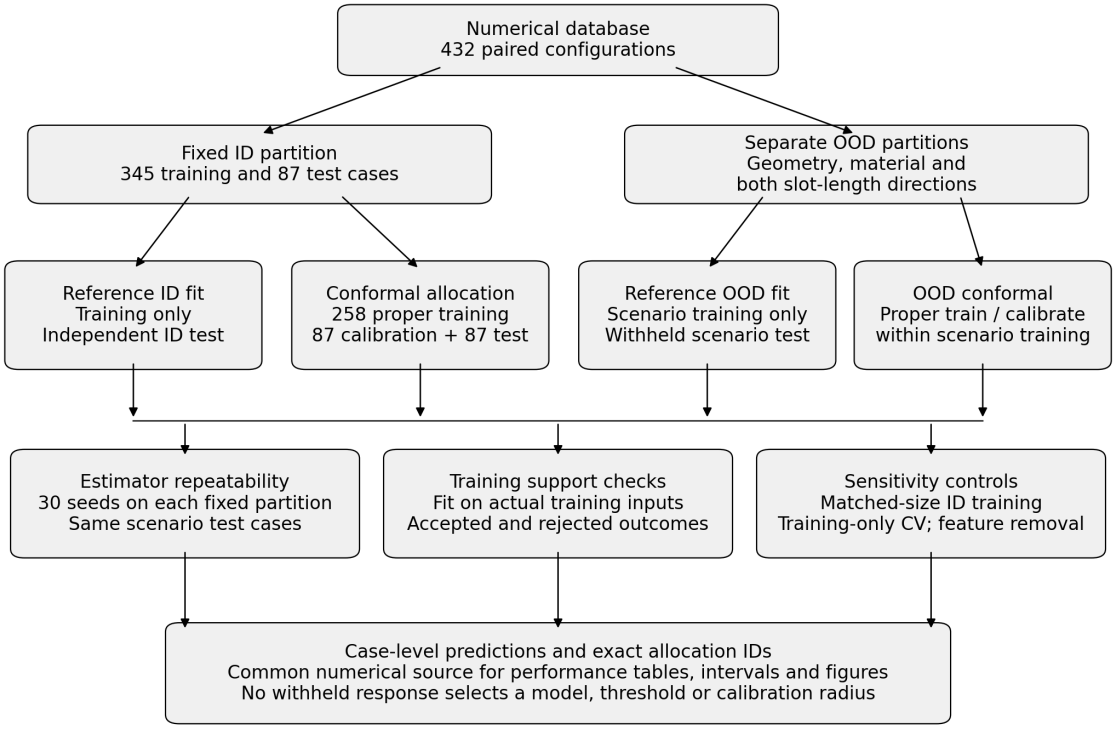


Figure 4. Experimental workflow. OOD experiments are separate partitions of the original database. Proper-training and calibration subsets are disjoint children of each applicable training domain; estimator repetitions retain the corresponding fixed test partition.

3.2. Conformal Prediction Intervals

Split conformal prediction uses a proper-training subset for fitting and a disjoint calibration subset for residual construction [20,21]. For calibration case j,

$$a_j = |y_j - \hat{f}(\mathbf{x}_j)|, \qquad j = 1, \ldots, n_{\mathrm{cal}}. \tag{4}$$

For nominal coverage $1 - \alpha$, sort the residuals in nondecreasing order, retain ties and select the finite-sample order statistic

$$k = \lceil (n_{\mathrm{cal}} + 1)(1 - \alpha) \rceil, \qquad q_{1-\alpha} = a_{(k)}. \tag{5}$$

If k exceeds the calibration size, the interval is unbounded rather than using a clamped residual rank. No percentile interpolation is used. With 87 calibration cases at 90% nominal coverage, $k = 80$. The symmetric interval is

$$\mathcal{I}_{1-\alpha}(\mathbf{x}) = [\hat{f}(\mathbf{x}) - q_{1-\alpha}, \hat{f}(\mathbf{x}) + q_{1-\alpha}]. \tag{6}$$

$$\widehat{C} = \frac{1}{n_{\mathrm{te}}} \sum_{i=1}^{n_{\mathrm{te}}} \mathbf{1}\{y_i \in \mathcal{I}_{1-\alpha}(\mathbf{x}_i)\}. \tag{7}$$

Endpoints count as covered. Coverage is reported with covered-case counts, 95% Wilson reference bounds and average interval width in kN m. The Wilson bounds describe finite test-count uncertainty under a binomial reference model; the structured design grid is not a random operational population. Thirty calibration allocations quantify split sensitivity without pooling their overlapping test observations as independent cases. OOD intervals are calibrated only on the corresponding training domain, so achieved OOD coverage is measured rather than inferred from nominal ID guarantees.

3.3. Domain of Applicability and Prediction Screening

Each fitted predictor receives its own training-support checks. The marginal screen uses the component-wise minima and maxima of that predictor's training inputs:

$$\Omega_{\mathrm{box}} = \{\mathbf{x} : x_j^{\min} \leq x_j \leq x_j^{\max} \, \forall j\}. \tag{8}$$

For the covariance screen, constant features in that training partition are removed, the remaining inputs are standardized on the same partition and a Ledoit–Wolf shrinkage covariance is fitted. The diagnostic distance is

$$d_M^2(\mathbf{x}) = (\mathbf{z} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}). \tag{9}$$

The combined rule requires both marginal acceptance and a squared distance below the training-distance 95th percentile. This percentile is a prespecified support heuristic, not a calibrated engineering safety limit. Threshold sensitivity is evaluated at the 90th, 95th, 97.5th and 99th percentiles without choosing a threshold from held-out errors. An enhanced rule additionally checks observed descriptor levels, positive dimensions, integer counts, the benchmark relation $r = 2t$ and nearest-training-combination distance against a leave-one-out training-distance threshold. A schema-valid but untested combination is distinguished from a count or benchmark-constraint violation; low covariance distance alone does not establish populated local support.

For interval predictors, support statistics are fitted on proper-training inputs, matching the actual interval model. For reference point predictors, they use the full corresponding training partition. The screen is assessed by retained fraction, accepted- and rejected-case errors, severe overprediction and accepted-case interval coverage. A rejected query leads to abstention and additional analysis; an accepted query remains subject to application-specific validation.

4. Investigation Methodology

4.1. Data Preprocessing and Candidate Models

The analysis uses Python 3.12.16, NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0 and scikit-learn 1.8.0, with XGBoost 3.4.1, LightGBM 4.7.0, CatBoost 1.2.10 and SHAP 0.54.1. Scikit-learn pipelines fit median imputation and any standardization only on the applicable training data [22]. Standardization is applied to Ridge, Lasso, Elastic Net, kNN, SVR and GPR. It is omitted for the tree-based learners because monotone rescaling preserves split ordering. Table 3 reports the reference configurations, with library defaults applying to unspecified settings [23–25].

A fixed model configuration is distinguished from parameters learned during fitting. GPR optimizes the constant multiplier, RBF length scale and white-noise level separately for every target and partition. Their initial bounds are 0.001–1000 for the multiplier and 0.00001–100000 for the length scale and noise level. Initial and fitted kernels, marginal likelihoods and fit warnings are exported for every reference fit in the numerical supplement. Thus the kernel starting values in Table 3 are not asserted to remain fixed after training.

A separate sensitivity branch uses shuffled 10-fold training-only validation, with identical fold IDs across learners within each scenario. It compares the reference configuration with up to seven prespecified alternatives from a compact grid, selects by mean CV RMSE, refits on the complete scenario training set and evaluates the held-out cases once. All transformations remain inside each fold. This controlled sensitivity branch replaces no reference result and is not an exhaustive hyperparameter search. Warning logs distinguish unsuccessful convergence in some Elastic Net CV candidates and GPR kernel-bound contacts from the selected model configurations. The selected Elastic Net configurations were additionally refitted and converged in all ten target–scenario combinations.

MAE, RMSE, MAPE and R squared are supplemented with signed error, defined as predicted minus numerical capacity, and the frequency and magnitude of overprediction. MAPE receives particular caution at small capacities; all labels here are positive. R squared depends on test-response variance and is not compared across scenarios without the accompanying target-distribution summaries. The provisional reporting tolerance requires both absolute error no greater than 1 kN m and relative error no greater than 10%. Severe overprediction denotes more than 20% relative overprediction. These prespecified values organize error reporting and do not represent code-calibrated resistance limits.

Table 3. Reference model configurations. Stochastic fits use seed 42 unless estimator seeds are deliberately varied.

Model	Reference Configuration
Ridge	alpha = 1
Lasso	alpha = 0.001; maximum 8000 iterations
Elastic Net	alpha = 0.001; L1 ratio = 0.5; maximum 8000 iterations
Decision Tree	Library defaults; seed 42
kNN	7 neighbours; distance weights
SVR	RBF kernel; C = 100; epsilon = 0.01; gamma = scale
GPR	Constant(1) × RBF(1) + WhiteKernel(0.001); normalize_y = True; L-BFGS-B optimizer; zero restarts
Random Forest	600 trees; max_features = sqrt
GBRT	100 estimators; other library defaults
HistGBR	100 maximum iterations; other library defaults
AdaBoost	50 estimators; other library defaults
XGBoost	800 trees; depth 6; learning rate 0.05; row and column subsampling 0.9; histogram tree method
LightGBM	800 trees; learning rate 0.05; subsample and column subsampling 0.9; other defaults
CatBoost	800 iterations; depth 8; learning rate 0.05; RMSE loss

4.2. In Distribution and Out of Distribution Evaluation

An 80/20 split with random seed 42 provides 345 ID training cases and 87 held-out cases. Table 4 defines the separate OOD partitions. Geometry testing withholds D/t $\leq$ 66.67 or D/t $\geq$ 250; material testing withholds 300 MPa; the two slot-length experiments train at one observed length and test at the other. Reusing a reference setting across partitions isolates a distribution change without assuming that it isolates training size or target composition.

Table 4. Analysis partitions and within-training conformal allocations. Calibration uses 25% of the applicable training partition.

Scenario	Train	Test	Proper Train	Calib.	Withheld Regime
ID	345	87	258	87	Random 20% hold-out
Geometry	288	144	216	72	Lower and upper D/t tails
Perforation	216	216	162	54	60 mm; training uses 75 mm
Material	288	144	216	72	300 MPa; training uses 500 and 600 MPa
Reverse perforation	216	216	162	54	75 mm; training uses 60 mm

Training-size controls retain the same 87 ID test cases and use nested random subsets of 216, 288 and 345 cases from the ID training pool. Thirty allocation seeds change subset membership while preserving the original row order and holding estimator seeds fixed. The full-size control is identical to the ID reference. Training and test response distributions are exported for every scenario, because size controls estimate an ID learning-size effect without making a complete causal decomposition of shift and target composition.

For the matched-configuration baseline, each held-out slot-length case is paired with the case at the other slot length having identical values of all remaining descriptors. There are 216 exact pairs. The observed source-length numerical capacity is transferred directly to its partner, without learning slot-length dependence. The experiment is repeated in both directions. This baseline tests whether low slot-transfer errors reflect a weak response change within this grid rather than recovery of a functional dependence from a single training level.

4.3. Uncertainty Quantification and Estimator Repeatability

Conformal evaluation includes all 14 methods independently of plotting layout, at 90% and 95% nominal coverage. Thirty allocations comprise the original calibration seed 42 and seeds 0–28. Each allocation refits the predictor on proper training, computes residuals only on its training-domain calibration subset and retains the fixed scenario test set. Coverage counts, widths, Wilson bounds and split summaries are stored separately from point-model results. Accepted-case interval statistics use the support screen fitted to the corresponding proper-training subset.

Estimator repeatability is evaluated within every ID and OOD scenario using seeds 0–29 for Decision Tree, Random Forest, GBRT, HistGBR, AdaBoost, XGBoost, LightGBM and CatBoost. All partition memberships and

reference settings are fixed. Ridge, cyclic Lasso and Elastic Net, kNN, SVR and zero-restart GPR receive one deterministic reference fit; repeating identical deterministic calculations is not treated as evidence of robustness. Some tree settings can also be insensitive to seeds on this structured dataset. Estimator variation, calibration allocation and training-subset variation are therefore reported as separate experiments.

5. Results

5.1. In Distribution Performance

Table 5 reports all reference MAEs from the same saved prediction records. The ID results in Figures 5 and 6 show very small errors for several flexible models. XGBoost gives MAE = 0.064 kN m for local buckling and 0.105 kN m for distortional buckling, compared with 0.083 and 0.200 kN m for GPR. Decision Tree, LightGBM and other boosted learners also fit the sampled grid closely. The regularized linear models have local errors near 2.77 kN m and distortional errors near 2.25 kN m. These differences establish an ID accuracy ranking for the reference configurations, not a universal ranking of algorithm families.

Table 5. Reference MAE in kN m for all methods and scenarios. LB and DB denote the two mode-specific targets.

Mode	Model	ID	Geometry	Slots 75→60	Material	Slots 60→75
LB	Ridge	2.774	4.908	2.714	4.063	2.729
LB	Lasso	2.773	4.908	2.718	4.077	2.736
LB	Elastic Net	2.773	4.908	2.717	4.075	2.735
LB	Decision Tree	0.117	14.718	0.163	5.868	0.163
LB	kNN	2.668	6.744	2.458	5.872	2.420
LB	SVR	0.155	4.425	9.035	8.179	8.990
LB	GPR	0.083	2.593	8.898	2.123	8.851
LB	Random Forest	1.011	7.811	0.453	6.004	0.416
LB	GBRT	0.178	10.314	0.188	5.878	0.185
LB	HistGBR	0.342	14.085	0.283	5.873	0.279
LB	AdaBoost	2.093	14.160	1.877	6.611	2.034
LB	XGBoost	0.064	13.813	0.163	5.868	0.163
LB	LightGBM	0.095	14.246	0.182	5.869	0.178
LB	CatBoost	0.203	6.722	0.163	5.868	0.163
DB	Ridge	2.247	4.108	2.234	3.074	2.217
DB	Lasso	2.245	4.108	2.236	3.085	2.219
DB	Elastic Net	2.246	4.108	2.236	3.084	2.219
DB	Decision Tree	0.201	12.354	0.192	4.275	0.192
DB	kNN	2.106	5.939	1.980	4.294	2.006
DB	SVR	0.154	3.826	7.513	6.425	7.515
DB	GPR	0.200	2.106	6.463	1.300	6.502
DB	Random Forest	0.809	6.377	0.363	4.382	0.396
DB	GBRT	0.242	9.515	0.222	4.288	0.242
DB	HistGBR	0.361	11.886	0.264	4.277	0.267
DB	AdaBoost	1.539	8.065	1.389	4.771	1.409
DB	XGBoost	0.105	9.150	0.192	4.275	0.192
DB	LightGBM	0.169	11.813	0.216	4.276	0.223
DB	CatBoost	0.135	5.515	0.192	4.275	0.192

Signed errors provide an additional view of these close fits. For ID XGBoost, the mean signed errors are +0.008 and +0.021 kN m, and the largest relative overpredictions are 3.26% and 6.38% for local and distortional capacities, respectively. The random hold-out therefore supports accurate prediction of configurations sampled from this evidence base. It does not directly test a missing material grade, geometry regime or slot-length level.

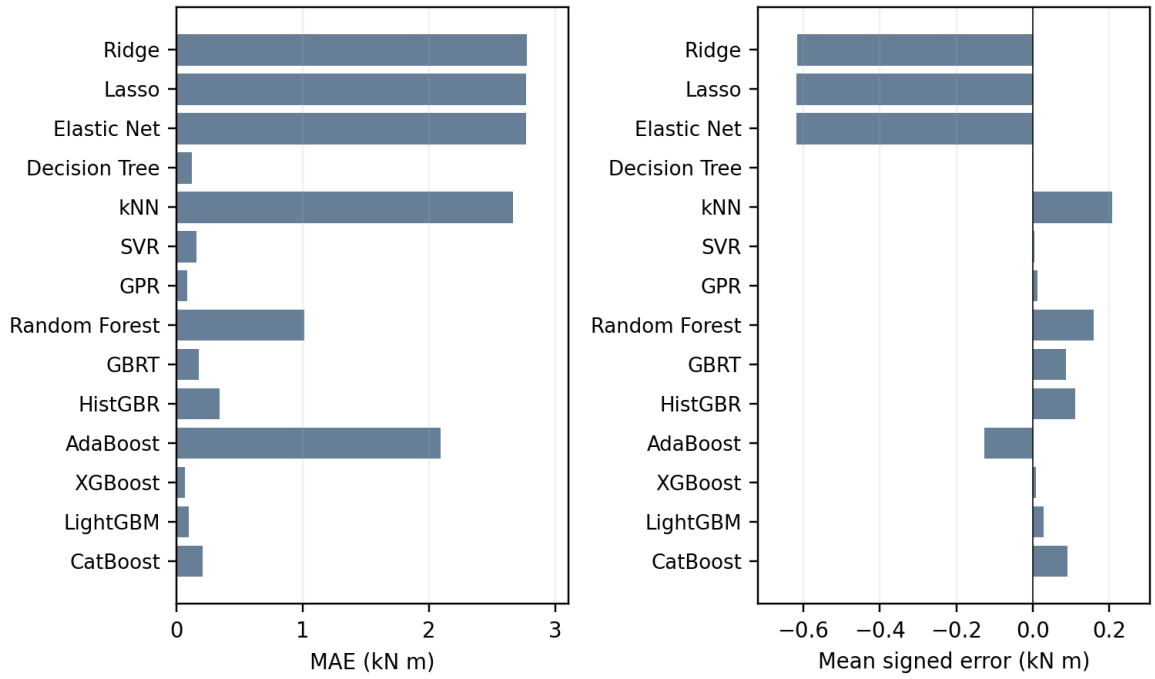


Figure 5. Reference in-distribution performance for the local-buckling target.

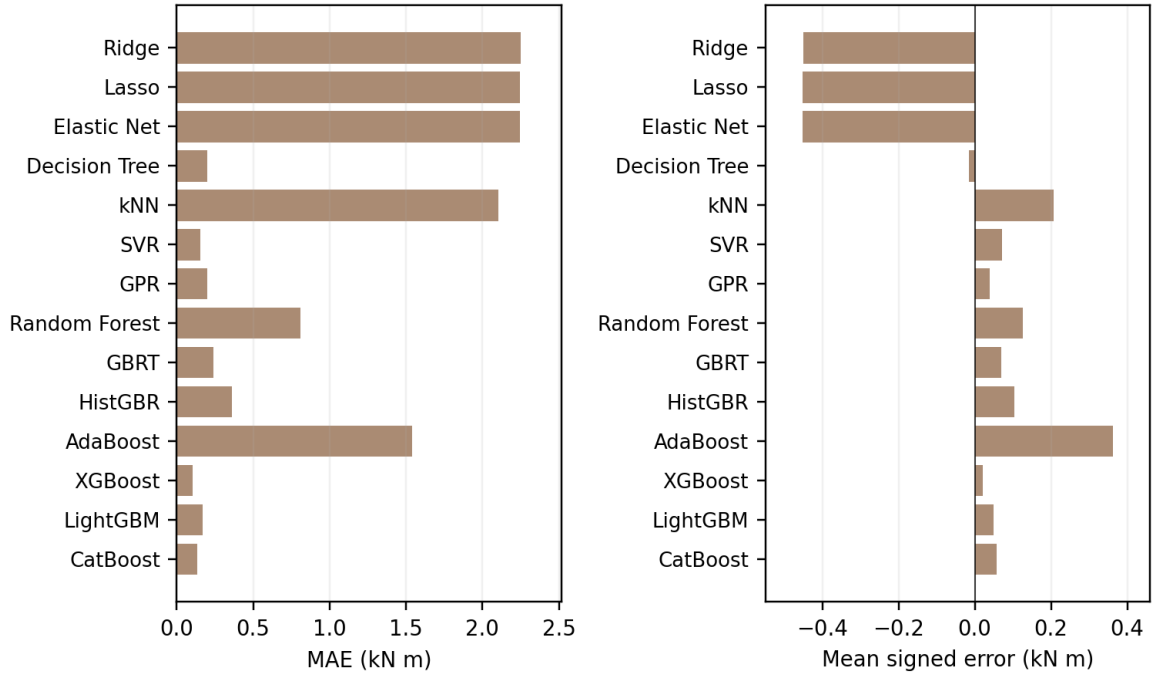


Figure 6. Reference in-distribution performance for the distortional-buckling target.

5.2. Out of Distribution Performance and Controls

5.2.1. Geometry Shift

Geometry omission produces large and predominantly positive resistance errors for many models (Figures 7 and 8). XGBoost MAEs increase to 13.813 and 9.150 kN m, with R squared values of −2.272 and −1.245. Every XGBoost prediction in this geometry test overpredicts its numerical target; the largest relative overpredictions are 204.45% and 246.08%. Decision Tree, LightGBM and several ensemble methods also deteriorate substantially. GPR remains comparatively more accurate, at 2.593 and 2.106 kN m, but overpredicts all geometry-test cases and reaches maximum relative overpredictions of 40.36% and 46.08%. Relative superiority within this stress test therefore does not imply conservative resistance prediction.

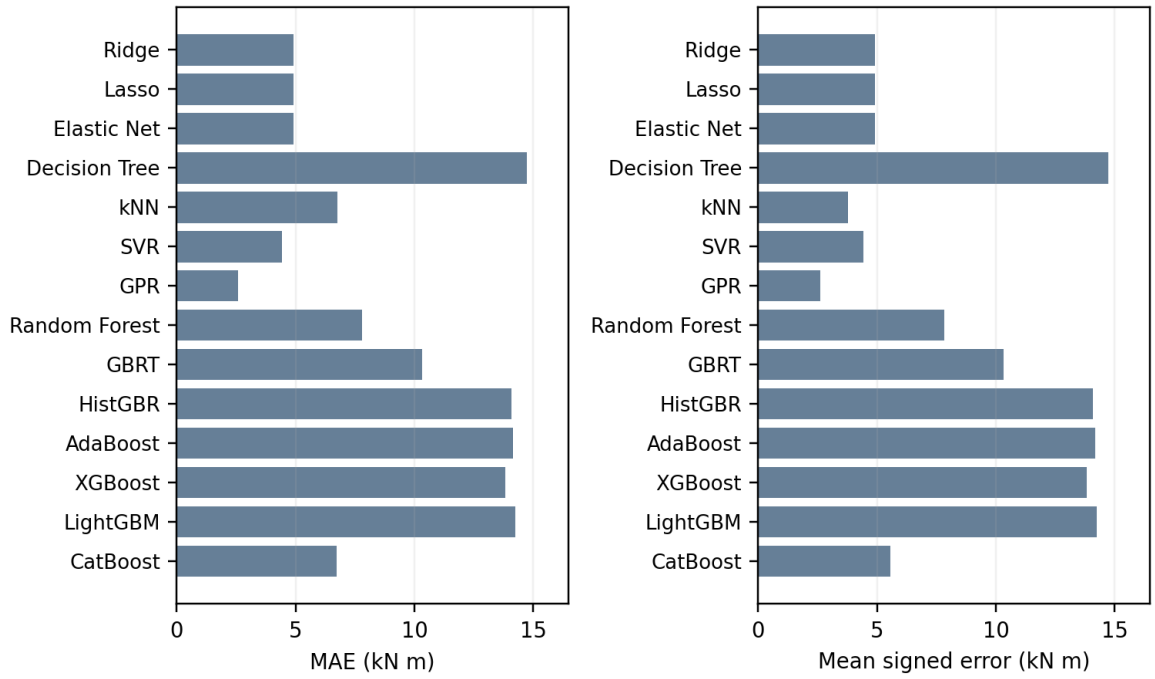


Figure 7. Geometry-shift performance for the local-buckling target.

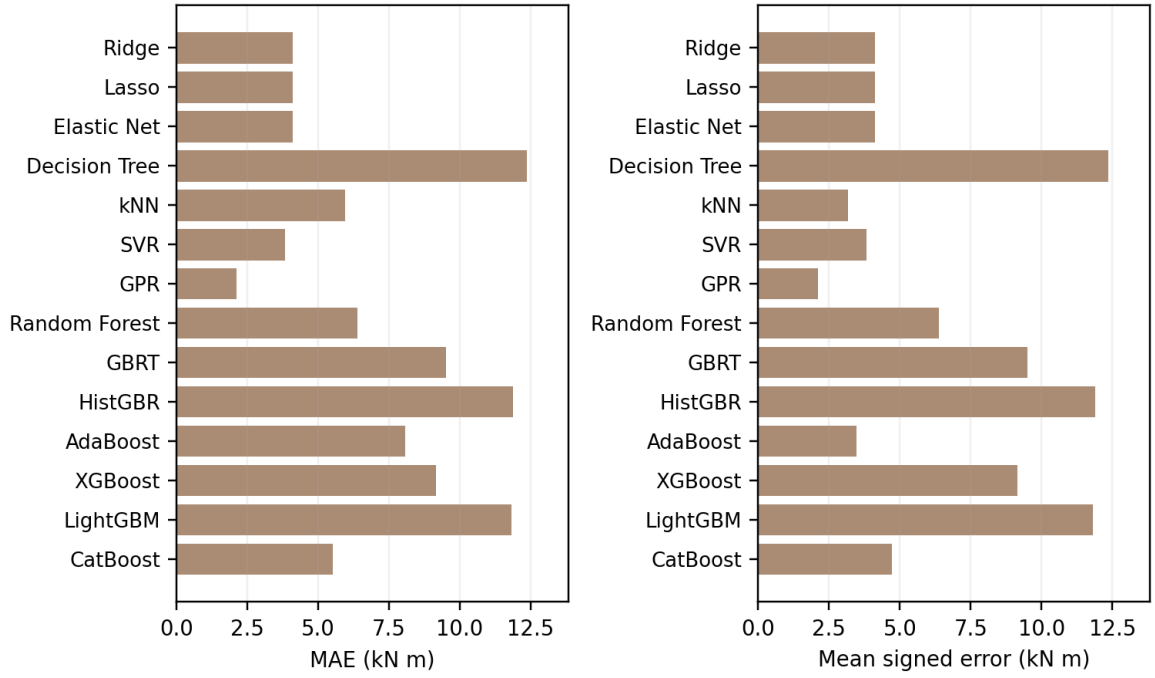


Figure 8. Geometry-shift performance for the distortional-buckling target.

5.2.2. Perforation Shift

Slot-length transfer is strongly model-dependent (Figures 9 and 10). Training at 75 mm and testing at 60 mm gives GPR MAEs of 8.898 and 6.463 kN m, with R squared values 0.050 and 0.262. SVR errors are 9.035 and 7.513 kN m, with R squared values close to zero. By contrast, XGBoost, Decision Tree and CatBoost give local errors near 0.163 kN m and distortional errors near 0.192 kN m. Random Forest also transfers more accurately than its own ID reference in this particular partition. Thus the slot-length experiment cannot be characterized as uniformly benign across the model set.

Only one slot-length level remains in training, so the learner cannot identify a slot-length response function from these data. The withheld 60 mm coordinate is outside the marginal training support. The matched-configuration

baseline in Table 6 attains MAEs of 0.163173 and 0.192148 kN m by transferring the known 75 mm response directly. Its agreement with the strongest tree predictions shows that the numerical response change between the two lengths is small for many matched configurations. This is evidence of effective transfer across a weakly influential coordinate within the sampled grid, rather than evidence of interpolation in an unseen slot length.

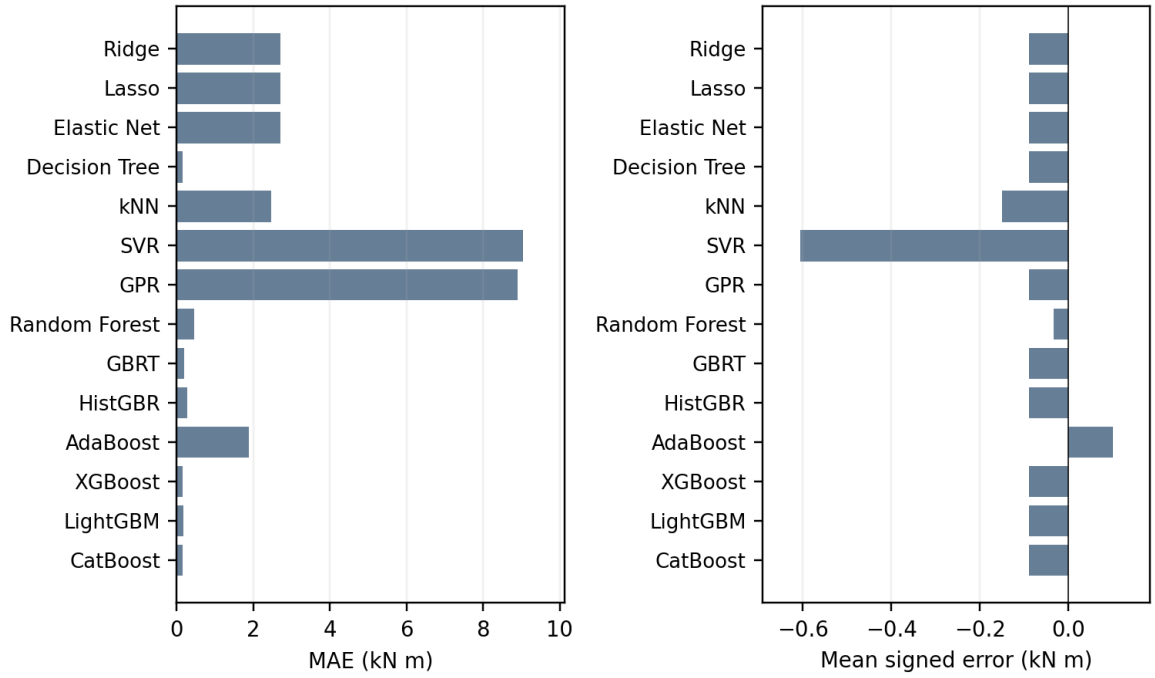


Figure 9. Transfer from 75 mm training slots to 60 mm test slots for the local-buckling target.

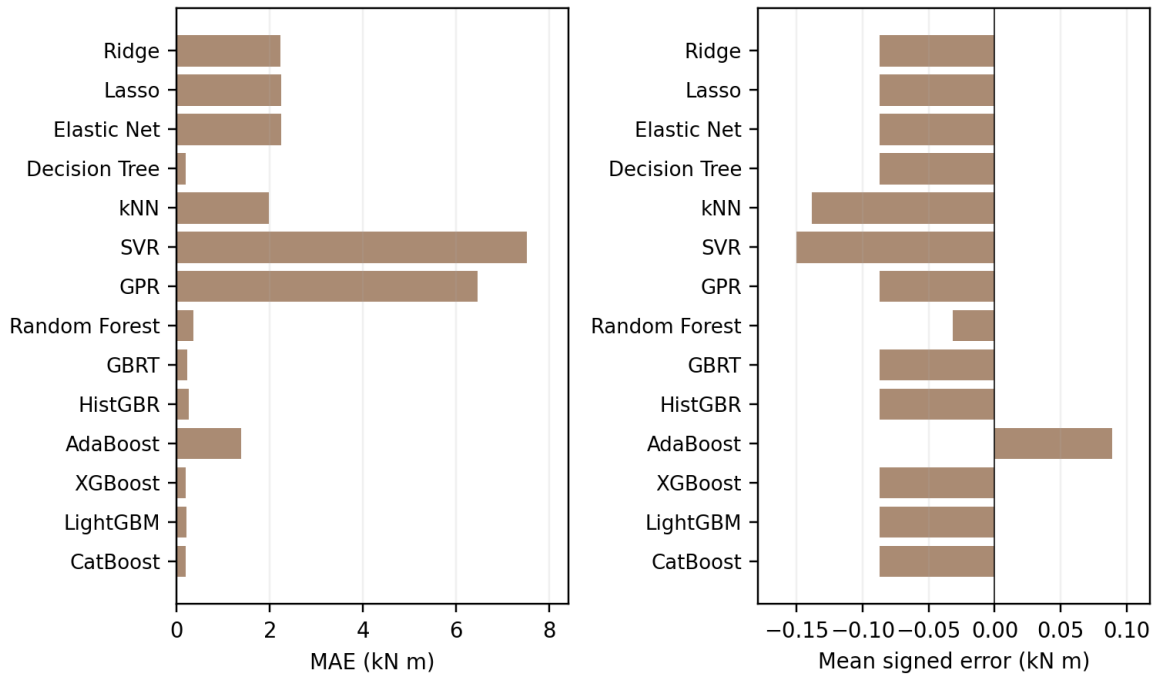


Figure 10. Transfer from 75 mm training slots to 60 mm test slots for the distortional-buckling target.

5.2.3. Material Shift

Withholding 300 MPa also produces substantial deterioration (Figures 11 and 12). GPR gives MAEs of 2.123 and 1.300 kN m, whereas SVR gives 8.179 and 6.425 kN m. XGBoost errors are 5.868 and 4.275 kN m, close to those of several other trees. All XGBoost material-test predictions overpredict the numerical target, with maxima of 61.55% and 51.30% relative overprediction. GPR overpredicts all local material-test cases and approximately

90.3% of distortional cases. Its maximum relative overpredictions of 105.64% and 94.26% occur at low capacities, illustrating why a moderate aggregate MAE can coexist with important resistance overestimation.

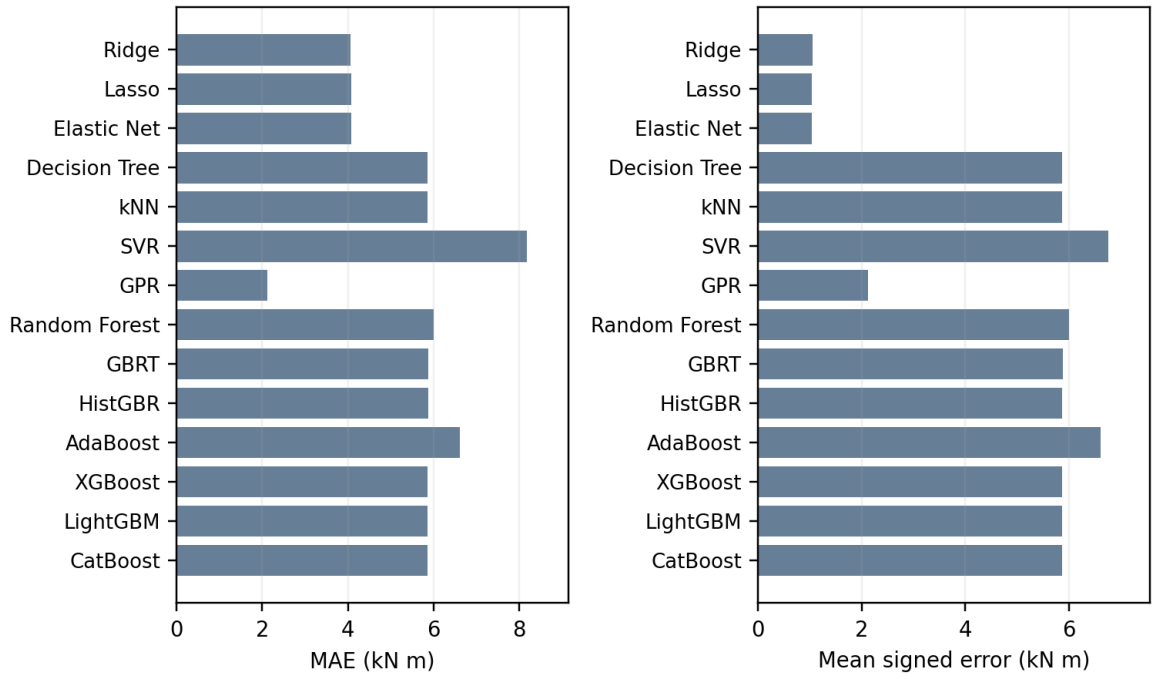


Figure 11. Material-shift performance for the local-buckling target with the 300 MPa group withheld.

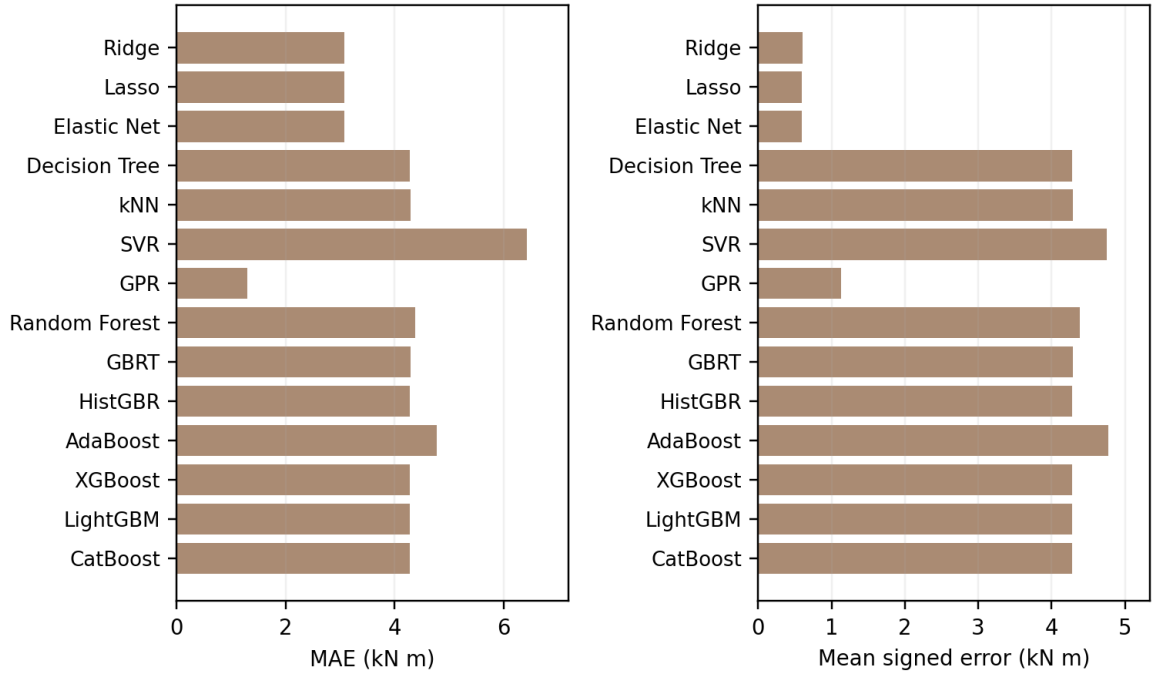


Figure 12. Material-shift performance for the distortional-buckling target with the 300 MPa group withheld.

5.2.4. Training Size Controls and Reverse Slot Transfer

Table 7 shows that reducing the random ID training pool increases error, but by much less than the principal OOD failures. For local XGBoost, mean ID MAE rises from 0.064 kN m at 345 cases to 0.080 at 288 and 0.134 at 216. Its geometry and material errors at a training size of 288 are 13.813 and 5.868 kN m. For local GPR, the 216-case ID control has mean MAE 0.122 kN m, compared with 8.898 kN m under the equally sized slot-length omission. The discrepancy therefore persists after a direct training-size comparison. Test-response composition remains different across OOD scenarios and is reported in the numerical supplement.

Reversing the slot-length split produces the same qualitative contrast. GPR errors remain high at 8.851 and 6.502 kN m; XGBoost gives 0.162911 and 0.192074 kN m. The matched-transfer baseline has the same MAE in both directions because each absolute paired response difference is unchanged by reversal. Its signed error reverses, so underprediction in one direction can become overprediction in the other. Reverse transfer therefore supports the importance of the matched baseline without establishing a general slot-length extrapolation capability.

Table 6. Matched-configuration baseline and representative model errors for both slot-length directions. All errors are in kN m.

Mode	Direction	Method	MAE	RMSE
LB	75→60	Matched transfer	0.163173	0.210994
LB	75→60	XGBoost	0.163161	0.210891
LB	75→60	GPR	8.898331	10.899538
LB	60→75	Matched transfer	0.163173	0.210994
LB	60→75	XGBoost	0.162911	0.210897
LB	60→75	GPR	8.850771	10.817241
DB	75→60	Matched transfer	0.192148	0.274913
DB	75→60	XGBoost	0.191995	0.274898
DB	75→60	GPR	6.462782	7.903566
DB	60→75	Matched transfer	0.192148	0.274913
DB	60→75	XGBoost	0.192074	0.274752
DB	60→75	GPR	6.502340	7.904819

Table 7. Size-matched ID controls over 30 subset allocations. The 345-case membership is identical to the reference; estimator seeds are fixed.

Mode	Model	Training Cases	Mean MAE	SD of MAE
LB	GPR	216	0.122	0.015
LB	GPR	288	0.097	0.007
LB	GPR	345	0.083	0.000
LB	XGBoost	216	0.134	0.028
LB	XGBoost	288	0.080	0.012
LB	XGBoost	345	0.064	0.000
DB	GPR	216	0.215	0.034
DB	GPR	288	0.204	0.020
DB	GPR	345	0.200	0.000
DB	XGBoost	216	0.173	0.030
DB	XGBoost	288	0.122	0.013
DB	XGBoost	345	0.105	0.000

The training-only CV sensitivity preserves the main failure patterns. Local GPR slot-transfer MAE remains approximately 8.898 kN m, and local XGBoost geometry MAE remains 13.813 kN m. Distortional SVR slot-transfer MAE changes from 7.513 to 6.871 kN m but remains much larger than its ID error. Some rankings and magnitudes change, which limits broad algorithm-family claims. The candidate grid, validation scores, selected parameters and fit warnings are supplied with the numerical results; no OOD response was consulted during this selection.

5.3. Prediction Interval Performance

Table 8 reports the original calibration-seed 42 ID counts and Wilson bounds for all 14 methods, together with mean coverage over 30 allocations. The two targets do not have identical calibration behaviour. At nominal 90% coverage, mean ID coverage across methods ranges from 83.72% to 91.84% for local buckling and from 83.41% to 91.95% for distortional buckling. For local GPR, the original 81/87 covered count gives 93.10%, whereas the

repeated-allocation mean is 88.39%. The difference shows why one favourable calibration allocation should not be treated as stable coverage evidence.

Distribution shifts cause much greater coverage loss than the finite-allocation changes within ID data (Table 9). Local XGBoost geometry coverage averages only 0.37%, while its material coverage is zero for every allocation. Corresponding distortional means are 2.85% and zero. GPR local means are zero for geometry and material shifts and 1.94% for the forward slot-length shift. Even where point errors are small, interval calibration can remain inadequate: local XGBoost slot-transfer coverage averages 82.65%, with individual allocations ranging from 62.96% to 98.15%. These intervals reflect residual behaviour within the retained training domain and cannot be assumed to cover a missing response regime.

Table 8. Nominal 90% ID interval performance. Counts, Wilson bounds and widths use calibration seed 42; the final column summarizes 30 allocations. Coverage values are percentages and widths are kN m.

Mode	Model	Covered	Coverage	95% Bounds	Width	Mean ± SD
LB	Ridge	79/87	90.8	82.9–95.3	12.715	89.9 ± 3.4
LB	Lasso	79/87	90.8	82.9–95.3	12.655	89.9 ± 3.4
LB	Elastic Net	79/87	90.8	82.9–95.3	12.658	89.9 ± 3.4
LB	Decision Tree	77/87	88.5	80.1–93.6	0.849	90.6 ± 4.5
LB	kNN	73/87	83.9	74.8–90.2	9.535	88.4 ± 3.2
LB	SVR	73/87	83.9	74.8–90.2	0.842	83.7 ± 4.5
LB	GPR	81/87	93.1	85.8–96.8	0.472	88.4 ± 3.9
LB	Random Forest	78/87	89.7	81.5–94.5	5.185	91.3 ± 3.6
LB	GBRT	74/87	85.1	76.1–91.1	0.800	91.3 ± 3.4
LB	HistGBR	80/87	92.0	84.3–96.0	1.592	90.3 ± 3.6
LB	AdaBoost	81/87	93.1	85.8–96.8	7.260	91.0 ± 4.3
LB	XGBoost	75/87	86.2	77.4–91.9	0.336	91.0 ± 3.8
LB	LightGBM	82/87	94.3	87.2–97.5	0.908	91.8 ± 3.7
LB	CatBoost	71/87	81.6	72.2–88.4	0.920	84.0 ± 3.7
DB	Ridge	79/87	90.8	82.9–95.3	10.359	91.0 ± 2.4
DB	Lasso	79/87	90.8	82.9–95.3	10.296	91.1 ± 2.3
DB	Elastic Net	79/87	90.8	82.9–95.3	10.301	91.1 ± 2.3
DB	Decision Tree	80/87	92.0	84.3–96.0	1.011	89.0 ± 2.1
DB	kNN	76/87	87.4	78.8–92.8	8.293	89.2 ± 2.5
DB	SVR	73/87	83.9	74.8–90.2	0.794	86.5 ± 5.3
DB	GPR	72/87	82.8	73.5–89.3	0.788	83.4 ± 5.3
DB	Random Forest	80/87	92.0	84.3–96.0	3.628	92.0 ± 3.0
DB	GBRT	73/87	83.9	74.8–90.2	0.958	87.5 ± 3.6
DB	HistGBR	80/87	92.0	84.3–96.0	1.697	87.1 ± 4.2
DB	AdaBoost	77/87	88.5	80.1–93.6	7.937	89.6 ± 4.3
DB	XGBoost	74/87	85.1	76.1–91.1	0.374	88.3 ± 3.6
DB	LightGBM	79/87	90.8	82.9–95.3	1.336	87.6 ± 3.3
DB	CatBoost	72/87	82.8	73.5–89.3	0.889	87.3 ± 4.2

The complete supplement includes 95% nominal results, covered and missed counts, residual ranks, widths and support-conditioned results for each allocation. Increasing nominal coverage changes interval width but does not restore exchangeability across an omitted regime. Coverage deficits denote empirical misses of numerical capacities; they are not structural failure probabilities. The evaluation supports reporting interval behaviour together with the deployment domain rather than using nominal coverage as an engineering acceptance certificate.

Table 9. Nominal 90% OOD coverage over 30 calibration allocations for every model. Entries are mean ± SD in percentage points; calibration is confined to the corresponding training domain.

Mode	Model	Geometry	Slots 75→60	Material	Slots 60→75
LB	Ridge	54.7 ± 8.1	92.5 ± 2.5	64.8 ± 7.4	93.2 ± 2.8
LB	Lasso	54.5 ± 7.6	92.5 ± 2.4	64.3 ± 7.5	93.3 ± 2.6
LB	Elastic Net	54.6 ± 7.7	92.5 ± 2.4	64.3 ± 7.5	93.3 ± 2.7
LB	Decision Tree	0.0 ± 0.0	92.5 ± 4.1	0.0 ± 0.0	89.3 ± 5.3
LB	kNN	34.9 ± 2.9	88.9 ± 5.2	53.3 ± 12.4	88.9 ± 5.3
LB	SVR	0.0 ± 0.0	5.9 ± 4.2	0.0 ± 0.0	6.8 ± 3.8
LB	GPR	0.0 ± 0.0	1.9 ± 1.6	0.0 ± 0.0	1.7 ± 1.5
LB	Random Forest	10.0 ± 10.2	97.4 ± 0.7	17.8 ± 11.2	98.2 ± 0.5
LB	GBRT	1.8 ± 7.0	90.6 ± 5.4	0.0 ± 0.0	88.6 ± 6.1
LB	HistGBR	0.7 ± 3.2	93.2 ± 3.0	0.1 ± 0.6	94.8 ± 3.3
LB	AdaBoost	10.2 ± 9.1	90.9 ± 3.3	30.4 ± 6.3	92.1 ± 3.6
LB	XGBoost	0.4 ± 2.0	82.7 ± 9.8	0.0 ± 0.0	81.0 ± 10.6
LB	LightGBM	0.1 ± 0.6	92.9 ± 4.0	0.0 ± 0.0	93.7 ± 4.3
LB	CatBoost	1.9 ± 3.4	96.0 ± 2.3	0.0 ± 0.0	96.8 ± 3.6
DB	Ridge	49.2 ± 6.9	93.0 ± 2.6	69.6 ± 3.7	93.3 ± 2.8
DB	Lasso	49.2 ± 6.9	93.0 ± 2.5	69.5 ± 3.8	93.3 ± 2.8
DB	Elastic Net	49.2 ± 6.9	93.0 ± 2.5	69.5 ± 3.8	93.3 ± 2.8
DB	Decision Tree	0.0 ± 0.0	93.8 ± 2.8	0.4 ± 0.9	93.8 ± 4.2
DB	kNN	31.3 ± 3.9	88.8 ± 4.5	63.4 ± 11.5	88.3 ± 5.3
DB	SVR	0.0 ± 0.0	11.6 ± 3.0	0.0 ± 0.0	10.5 ± 1.9
DB	GPR	0.9 ± 1.1	8.1 ± 4.0	8.4 ± 2.1	4.2 ± 2.9
DB	Random Forest	14.8 ± 11.9	97.8 ± 0.7	29.9 ± 4.9	97.7 ± 0.5
DB	GBRT	0.0 ± 0.0	93.5 ± 3.0	0.9 ± 1.2	88.2 ± 4.5
DB	HistGBR	6.8 ± 9.7	93.4 ± 3.3	5.1 ± 5.0	93.5 ± 3.3
DB	AdaBoost	11.1 ± 10.2	89.6 ± 3.9	31.9 ± 2.7	92.5 ± 3.1
DB	XGBoost	2.8 ± 6.4	88.3 ± 6.7	0.0 ± 0.0	88.9 ± 5.6
DB	LightGBM	1.3 ± 3.0	92.4 ± 4.5	0.3 ± 1.0	92.7 ± 2.3
DB	CatBoost	3.4 ± 3.5	96.1 ± 1.8	0.1 ± 0.5	95.4 ± 1.9

5.4. Error Inflation under Distribution Shift

Figures 13 and 14 display the verified inflation factors defined in Equation (3). The local GPR slot-length value is $\log_{10}(8.898331/0.082704)$, approximately 2.032, equivalent to a 107.6-fold increase. The local Random Forest value is approximately -0.348 because its slot-transfer MAE, 0.453 kN m, is below its 1.011 kN m ID error. Thus the direction of a change can differ across models within the same scenario. For XGBoost, local geometry inflation is approximately 215 times, compared with approximately 2.54 times for forward slot transfer. All displayed ratios use the unrounded reference errors; the displayed decimal values are rounded only for presentation.

An inflation map identifies a relative change, not an absolute safety margin. A method with a larger ID error can have a modest ratio and still produce unacceptable resistance overprediction. Conversely, a small absolute transfer error can have a visible ratio when the ID denominator is exceptionally small. Absolute errors in Table 5, size controls in Table 7 and signed-error records therefore remain necessary for interpretation.



Figure 13. Local-target MAE inflation computed from the common verified prediction records. The symmetric colour scale is centred on zero.

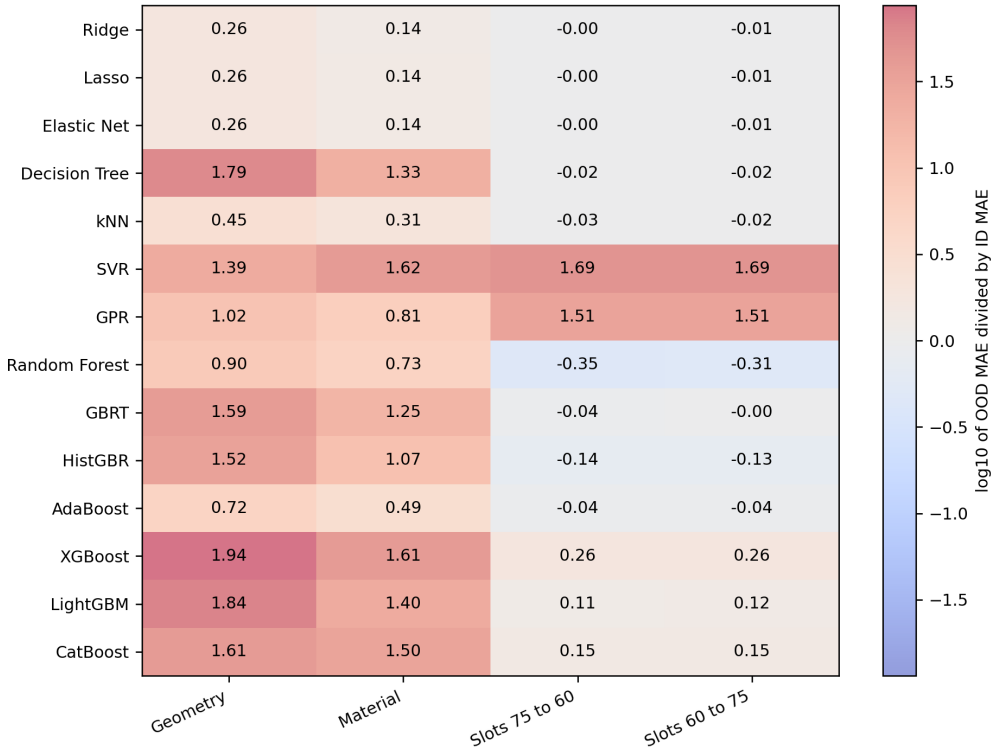


Figure 14. Distortional-target MAE inflation computed from the common verified prediction records. Negative values describe lower error on the particular scenario partition.

5.5. Estimator Repeatability within Each Scenario

The 30-seed results show persistent OOD failures for the stochastic learners rather than a failure confined to the original seed. Table 10 reports representative methods; complete results for all learners are supplied separately. Local XGBoost geometry MAE ranges from 8.797 to 13.806 kN m, with mean 12.925 and SD 0.959 kN m. Its distortional

range is 6.948–10.613 kN m. These values remain far above the corresponding ID errors. Material and slot-transfer results vary much less under estimator seeds. The evidence supports a scenario-specific statement about estimator randomness, while calibration allocation and training-sample selection remain separate sources of variation.

Table 10. Estimator-seed repeatability for representative stochastic learners over 30 seeds. MAE statistics are in kN m.

Mode	Model	Scenario	Mean	SD	Range
LB	XGBoost	ID	0.068	0.010	0.058–0.109
LB	XGBoost	Geometry	12.925	0.959	8.797–13.806
LB	XGBoost	Slots 75 to 60	0.163	0.000	0.163–0.163
LB	XGBoost	Material	5.868	0.000	5.868–5.868
LB	Random Forest	ID	1.009	0.023	0.959–1.047
LB	Random Forest	Geometry	7.678	0.197	7.331–8.188
LB	Random Forest	Slots 75 to 60	0.457	0.013	0.435–0.491
LB	Random Forest	Material	6.000	0.009	5.985–6.016
LB	Decision Tree	ID	0.118	0.001	0.116–0.120
LB	Decision Tree	Geometry	14.718	0.000	14.718–14.718
LB	Decision Tree	Slots 75 to 60	0.163	0.000	0.163–0.163
LB	Decision Tree	Material	5.868	0.000	5.868–5.868
DB	XGBoost	ID	0.106	0.007	0.096–0.136
DB	XGBoost	Geometry	9.207	0.896	6.948–10.613
DB	XGBoost	Slots 75 to 60	0.192	0.000	0.192–0.192
DB	XGBoost	Material	4.275	0.000	4.275–4.275
DB	Random Forest	ID	0.797	0.025	0.760–0.862
DB	Random Forest	Geometry	6.387	0.179	6.083–6.922
DB	Random Forest	Slots 75 to 60	0.371	0.008	0.353–0.388
DB	Random Forest	Material	4.379	0.008	4.365–4.399
DB	Decision Tree	ID	0.203	0.004	0.196–0.207
DB	Decision Tree	Geometry	12.354	0.000	12.354–12.354
DB	Decision Tree	Slots 75 to 60	0.192	0.000	0.192–0.192
DB	Decision Tree	Material	4.275	0.000	4.275–4.275

5.6. Feature Influence and Linked Predictor Sensitivity

Figures 15 and 16 explain separately fitted XGBoost reference models for each target and scenario. Each model uses its own training-only background, sampled to at most 100 rows, and up to 216 held-out evaluation rows. Background and evaluation IDs are fixed across the full, dropped-radius, dropped-thickness and dropped-constant variants. Interventional Tree SHAP is used with raw model output, and numerical additivity is checked. These primary plots combine a change in fitted model with a change in evaluation population. A separate supplement holds the ID-fitted model and its background fixed across populations to isolate the population component; overlap with ID training is explicitly recorded for that descriptive diagnostic.

Thickness has the largest individual attribution in the full reference model, but its exact linkage with root radius prevents an independent physical attribution to either predictor. For ID local capacity, mean absolute thickness and radius contributions are 6.276 and 1.352 kN m; the combined contribution, computed as mean absolute value of the case-level sum, is 7.628 kN m. The corresponding distortional combined contribution is 6.425 kN m. The linked group is a more defensible description of model reliance than an interpretation of radius as a separately identified secondary mechanism.

Removing either linked predictor changes both ID and geometry-transfer errors (Table 11). For local XGBoost, geometry MAE falls from 13.813 kN m in the full model to 8.027 without radius and 9.309 without thickness, although both remain large. The corresponding distortional errors are 9.150, 6.299 and 7.108 kN m. Material-transfer failures remain essentially unchanged. The sensitivity therefore confirms that redundant representation can influence a fitted model's allocation of importance and its geometry extrapolation, while providing no basis for causal interpretation of the separate full-model SHAP values.

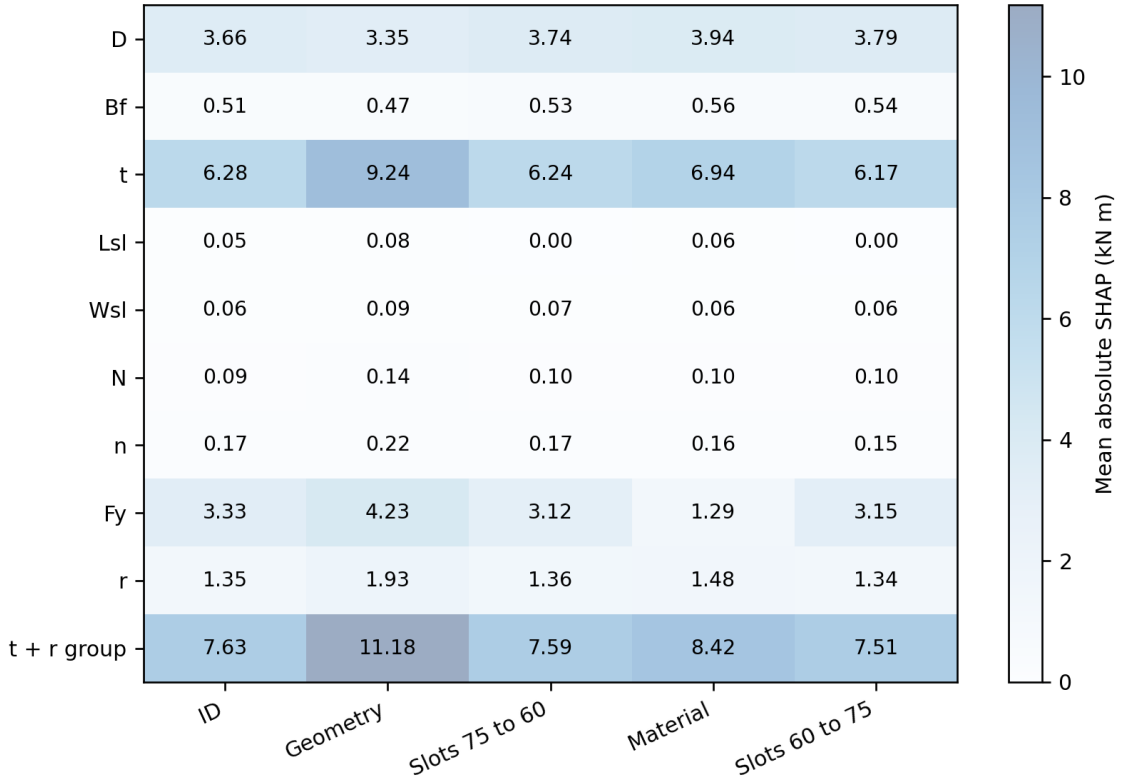


Figure 15. Mean absolute SHAP contributions for separately fitted local-target XGBoost models on each corresponding held-out population. The last row groups thickness and radius; it is not an additional independent feature.

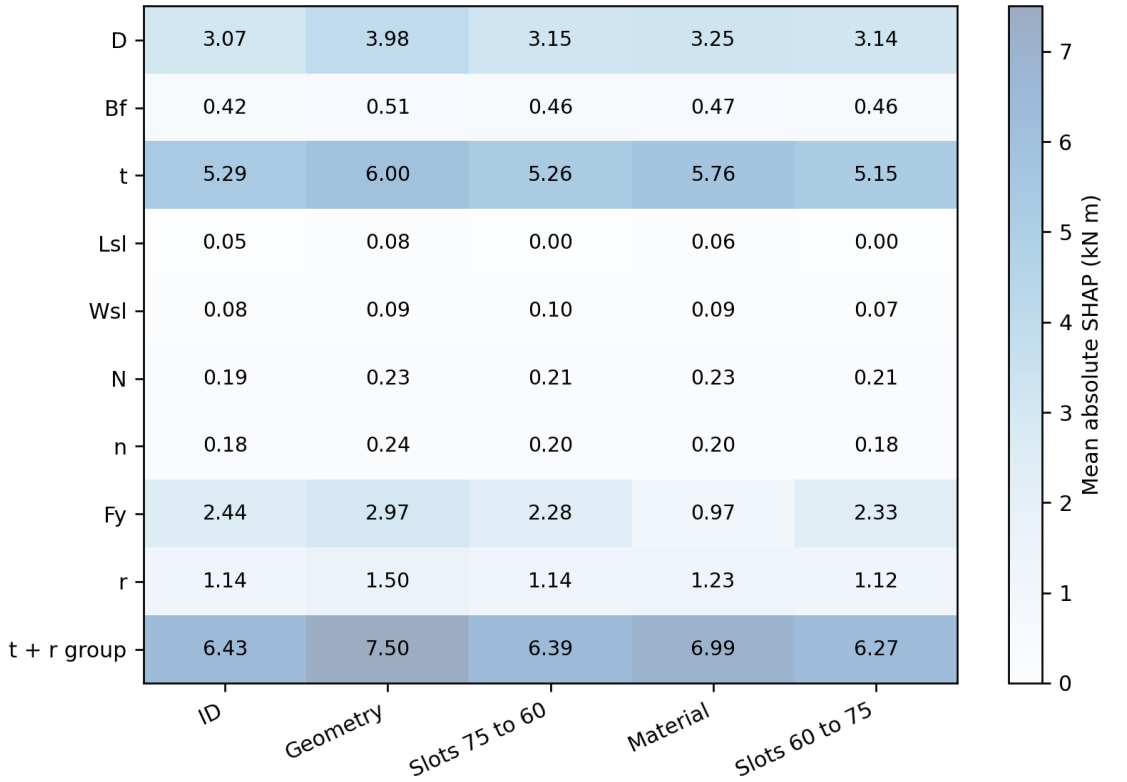


Figure 16. Mean absolute SHAP contributions for separately fitted distortional-target XGBoost models on each corresponding held-out population. The last row groups thickness and radius; it is not an additional independent feature.

Table 11. XGBoost feature-removal sensitivity. Each variant is refitted on the same scenario training cases; MAEs are kN m. GPR and reverse-direction results are in the supplement.

Mode	Feature set	ID	Geometry	Slots 75→60	Material
LB	All features	0.064	13.813	0.163	5.868
LB	Without radius	0.178	8.027	0.163	5.868
LB	Without thickness	0.146	9.309	0.163	5.868
LB	Without constants	0.059	14.236	0.163	5.868
DB	All features	0.105	9.150	0.192	4.275
DB	Without radius	0.155	6.299	0.192	4.275
DB	Without thickness	0.137	7.108	0.192	4.275
DB	Without constants	0.097	9.043	0.192	4.275

5.7. Applicability Screening and Accepted Case Outcomes

A scenario-specific screen gives a different assessment from a screen trained once on the ID pool. The marginal range screen accepts all 144 geometry-test inputs because their individual feature values remain represented, but the combined range and covariance screen rejects all 144. Each slot-length and material hold-out is already rejected by the marginal screen because its held-out coordinate is absent from training. The reverse slot-length result behaves similarly. Thus the combined screen rejects every shifted test query at the prespecified 95th-percentile rule in this benchmark (Table 12). Accepted-case OOD errors and interval coverage are undefined because retention is zero; these rejections establish abstention behaviour, not validated OOD prediction accuracy.

For the fixed ID test set, marginal ranges retain 87/87 cases, the combined rule retains 77/87 and the enhanced rule retains 74/87. For distortional XGBoost, combining the checks reduces accepted-case MAE from 0.105 to 0.074 kN m and the largest relative overprediction from 6.38% to 1.42%, at 88.5% retention. For local XGBoost, MAE changes from 0.064 to 0.059 kN m and maximum relative overprediction from 3.26% to 2.00%. The enhanced rule reduces retention further without improving every local error statistic. Threshold sensitivity and proper-training support-conditioned interval coverage are provided for all methods and allocations. A stronger exclusion rule therefore has a measurable retention cost and should be judged against accepted-case outcomes.

Table 12. Retention at the prespecified 95th-percentile support rule and accepted-case XGBoost MAE. Support counts are common to targets; a dash denotes an empty accepted subset.

Scenario	Rule	Accepted	Retained %	LB MAE	DB MAE
ID	Marginal	87/87	100.0	0.064	0.105
ID	Combined	77/87	88.5	0.059	0.074
ID	Enhanced	74/87	85.1	0.060	0.071
Geometry	Marginal	144/144	100.0	13.813	9.150
Geometry	Combined	0/144	0.0	—	—
Geometry	Enhanced	0/144	0.0	—	—
Slots 75 to 60	Marginal	0/216	0.0	—	—
Slots 75 to 60	Combined	0/216	0.0	—	—
Slots 75 to 60	Enhanced	0/216	0.0	—	—
Material	Marginal	0/144	0.0	—	—
Material	Combined	0/144	0.0	—	—
Material	Enhanced	0/144	0.0	—	—
Slots 60 to 75	Marginal	0/216	0.0	—	—
Slots 60 to 75	Combined	0/216	0.0	—	—
Slots 60 to 75	Enhanced	0/216	0.0	—	—

6. Discussion

The main result is the separation between accurate prediction inside a structured grid and credible use when a regime is missing. Several learners have extremely small random-hold-out errors, yet geometry and material

omissions produce systematic resistance overprediction. Size-matched controls demonstrate that the loss cannot be attributed solely to a smaller training set. The target distributions and absolute errors remain necessary because the withheld regimes also change response composition and make ratios and R squared values scenario-dependent.

The two slot-length experiments sharpen this interpretation. Small tree-model errors coexist with severe GPR and SVR failures even though the same dominant geometry and strength descriptors are available. A direct transfer between matched configurations already attains errors near those of the leading trees, so the low errors do not show that a slot-length dependence has been learned from a single observed level. The numerical evidence supports conditional transfer within the present grid, while any additional length level or combined feature change requires its own test.

Prediction intervals reveal a further limitation of deploying a model on the strength of a random hold-out. Residuals from the retained training domain can be small even when an omitted regime shifts the response substantially. Consequently, nominal 90% intervals can have almost no OOD coverage. Repeating calibration allocations distinguishes this persistent loss from one arbitrary allocation. The same deployment issue arises in monitoring systems and digital twins, where changing sensor, environment or material populations can alter the relationship learned by the predictive component [10]. Material-property and printing-process studies likewise place emphasis on the descriptors and data used to construct a model [11,12]. Their relevance here is to the need for a specified evidence domain, not to a claim that one model family is generally reliable across applications.

Training-specific applicability checks are useful when interpreted as abstention rules with measured retention and accepted-case errors. In this benchmark, the combined rule excludes all intentionally shifted cases and moderately improves the retained ID XGBoost predictions. That evidence supports flagging unsupported queries for higher-fidelity analysis. It does not establish the accuracy of accepted OOD predictions because none are retained. The 95th percentile and provisional error tolerances must therefore be distinguished from engineering acceptance criteria calibrated to a particular design decision.

SHAP analysis complements these checks by describing model reliance. The thickness–radius pair is the dominant linked group, but separate feature rankings change with the representation, fitting domain and evaluation population. The feature-removal results show that redundant descriptors can materially affect geometry transfer. This supports reporting model provenance, grouped attributions and predictive sensitivity together. Plausible-looking feature importance does not repair a measured error or interval-coverage failure.

The study is limited to numerical capacities and three principal shift mechanisms, with a reverse test for the two-level slot descriptor. The FE source studies supply experimental verification evidence, but the data do not constitute a new experimental campaign. Fabrication variation, alternative restraints, load patterns, additional steel grades, continuous slot lengths and compound shifts are not represented. The training-only sensitivity grid is compact, and some candidate fits encounter optimization warnings. The framework should therefore be transferred to another application by redefining its context, training support and tolerances and then evaluating the relevant response regimes, rather than transferring the present numerical rankings unchanged.

7. Conclusions

A common case-level analysis of 14 learners on 432 paired slotted-beam configurations shows that near-perfect ID accuracy can coexist with severe OOD resistance overprediction. XGBoost's ID MAEs of 0.064 and 0.105 kN m increase to 13.813 and 9.150 kN m under the geometry omission. Stochastic repetitions preserve the principal failures, and size-matched ID controls remain much more accurate.

Perforation transfer is model-dependent. GPR and SVR fail substantially when one slot-length level is withheld, whereas leading tree models approximately reproduce the matched-configuration transfer baseline. The reverse hold-out confirms this contrast. The result supports a description of transfer across a weak response change in the sampled coordinate, rather than interpolation or generally learned slot-length dependence.

Conformal intervals require direct assessment within the deployment domain. Across 30 calibration allocations, nominal 90% geometry-shift coverage for XGBoost averages only 0.37% for local and 2.85% for distortional capacity. Reporting covered counts, allocation variability, widths and training-domain provenance makes this loss visible and prevents nominal coverage from being interpreted as a physical safety guarantee.

The combined applicability screen provides a transparent abstention mechanism. It rejects all shifted test queries and retains 88.5% of the fixed ID set, with improved accepted-case XGBoost errors. An enhanced schema and nearest-combination check further reduces retention. Domain acceptance must consequently be reported together with retention, accepted-case error and interval behaviour, and cannot authorize an unqualified structural prediction by itself.

The resulting workflow is an evidence audit for numerical surrogates: define the structural response and training domain, assess relevant missing regimes, quantify uncertainty and fitting variability, and abstain where support or demonstrated performance is inadequate. Its practical use is to identify where additional analysis or data are required before a surrogate is used in a resistance decision.

Supplementary Materials

The following supporting information can be downloaded at: https://media.sciltp.com/articles/others/261126 1337270661/BCI-26100022-SM.zip.

Funding

This research was supported by the Czech Science Foundation (GAČR) under Grants No. 23-06222S and No. 24-10892S.

Institutional Review Board Statement

Not applicable. This study did not involve human participants or animals.

Informed Consent Statement

Not applicable. This study did not involve human participants.

Data Availability Statement

The original project repository is https://github.com/Simwanda/stress_testing_ml_surrogates. The submission supplement provides the reanalysis notebook and Python module, the 432-case spreadsheet, environment requirements, exact split IDs, reference case-level predictions, model settings, numerical summaries, fit warnings and SHAP provenance. These files identify the completed full-profile analysis through source and data hashes. The larger case-level conformal, estimator-seed and size-control records are retained in the computational results archive and can be supplied with the numerical submission files. No previously serialized fitted estimator is used as an input to the reanalysis.

Conflicts of Interest

The author declares no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the author used OpenAI ChatGPT to assist with language editing, organization of reviewer and proof responses, and refinement of manuscript and analysis code. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

1. Iyappan, G.R.; Sangeetha, P. Machine Learning Methods for Predicting the Performance of Cold-Formed Steel Sections: A State-of-the-Art Review. *Struct. Des. Tall Spec. Build.* **2025**, *34*, e70079. https://doi.org/10.1002/tal.70079.
2. Mittelstedt, C. Generalized Beam Theory for the Analysis of Thin-Walled Structures—A State-of-the-Art Survey. *Thin-Walled Struct.* **2024**, *200*, 111849. https://doi.org/10.1016/j.tws.2024.111849.
3. Simwanda, L.; Gatheeshgar, P.; Ilunga, F.M.; et al. Explainable Machine Learning Models for Predicting the Ultimate Bending Capacity of Slotted Perforated Cold-Formed Steel Beams under Distortional Buckling. *Thin-Walled Struct.* **2024**, *205*, 112587. https://doi.org/10.1016/j.tws.2024.112587.
4. Simwanda, L.; Gatheeshgar, P.; Sykora, M.; et al. Local Buckling Strength Prediction of Slotted Cold-Formed Steel Beams Using Ensemble Learning. *J. Constr. Steel Res.* **2026**, *236*, 110024. https://doi.org/10.1016/j.jcsr.2025.110024.
5. Chandramohan, D.L.; Roy, K.; Fang, Z.; et al. Moment Capacity of Cold-Formed Steel Channel Beams with Edge-Stiffened and Unstiffened Elongated Web Holes. *Thin-Walled Struct.* **2025**, *206*, 112605. https://doi.org/10.1016/j.tws.2024.112605.
6. Ananthi, G.B.G.; Chandramohan, D.L.; Mandal, D.; et al. Flexural Strength of Cold-Formed Steel Unstiffened and Edge-Stiffened Hexagonal Perforated Channel Sections. *Buildings* **2025**, *15*, 2679. https://doi.org/10.3390/buildings15152679.
7. Paul, B.; Roy, K.; Lakshmanan, D.C.; et al. Flexural Behaviour and Design of Cold-Formed Aluminium Channel Sections with Elongated and Circular Web Holes. *Structures* **2026**, *86*, 111212. https://doi.org/10.1016/j.istruc.2026.111212.
8. Silewu, K.; Kahanji, C.; Simwanda, L.; et al. Intelligent Data Driven Ensemble Approaches for Bending Strength Prediction of Ultra-High Performance Concrete Beams. *Bull. Comput. Intell.* **2025**, *1*, 31–52. https://doi.org/10.53941/bci.2025.100003.

9. Subedi, A.; Subedi, S.; Adhikari, S.; et al. Explainable Hybridized Machine Learning for Prediction of Compressive Strength of Fly-Ash Based Geopolymer Concrete. *Bull. Comput. Intell.* **2026**, *2*, 213–234. https://doi.org/10.53941/bci.2026.100012.
10. Afrazi, M.; Jahed Armaghani, D.; Afrazi, H.; et al. Real-Time Monitoring of Tunnel Structures Using Digital Twin and Artificial Intelligence: A Short Overview. *Deep Undergr. Sci. Eng.* **2026**, *5*, 315–330. https://doi.org/10.1002/dug2.70029.
11. Afrazi, M.; Huat, C.Y.; Onifade, M.; et al. Modeling the Abrasive Index from Mineralogical and Calorific Properties Using Tree-Based Machine Learning: A Case Study on the KwaZulu-Natal Coalfield. *Mining* **2025**, *5*, 48. https://doi.org/10.3390/mining5030048.
12. Fareghian, M.; Afrazi, M.; Jahed Armaghani, D.; et al. Machine Learning in the Modeling of 3D Printed Soil-Cement Materials: A Short Overview. *J. Soft Comput. Civ. Eng.* **2026**, *10*, e1966.
13. Ben-Haim, Y.; Hemez, F.M. Robustness, Fidelity and Prediction-Looseness of Models. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **2012**, *468*, 227–244. https://doi.org/10.1098/rspa.2011.0050.
14. Parvinian, B.; Bighamian, R.; Scully, C.G.; et al. Credibility Assessment of a Subject-Specific Mathematical Model of Blood Volume Kinetics for Prediction of Physiological Response to Hemorrhagic Shock and Fluid Resuscitation. *Front. Physiol.* **2021**, *12*, 705222. https://doi.org/10.3389/fphys.2021.705222.
15. Scuoppo, R.; Catalano, C.; Crascì, F.; et al. Credibility Assessment of the Patient-Specific Modeling of the Aneurysmal Ascending Thoracic Aorta: Verification, Validation and Uncertainty Quantification. *Cardiovasc. Eng. Technol.* **2025**, *16*, 663–677. https://doi.org/10.1007/s13239-025-00801-1.
16. Müller, A.; Taras, A.; Kraus, M.A. Scientific Machine and Deep Learning Investigations of the Local Buckling Behaviour of Hollow Sections. *ce/papers* **2022**, *5*, 1034–1042. https://doi.org/10.1002/cepa.1848.
17. Hu, X.; Wang, R.; Liang, X.; et al. A Comprehensive Validation Framework for Numerical Simulation Based on Uncertainty Quantification. *Eng. Comput.* **2025**, *42*, 1302–1315. https://doi.org/10.1108/ec-07-2024-0648.
18. Degtyareva, N.; Gatheeshgar, P.; Poologanathan, K.; et al. Local Buckling Strength and Design of Cold-Formed Steel Beams with Slotted Perforations. *Thin-Walled Struct.* **2020**, *156*, 106951. https://doi.org/10.1016/j.tws.2020.106951.
19. Degtyareva, N.; Gatheeshgar, P.; Poologanathan, K.; et al. New Distortional Buckling Design Rules for Slotted Perforated Cold-Formed Steel Beams. *J. Constr. Steel Res.* **2020**, *168*, 106006. https://doi.org/10.1016/j.jcsr.2020.106006.
20. Angelopoulos, A.N.; Bates, S. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv* **2021**, arXiv:2107.07511. http://arxiv.org/abs/2107.07511v6.
21. Romano, Y.; Patterson, E.; Candès, E.J. Conformalized Quantile Regression. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 3543–3553.
22. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
23. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. https://doi.org/10.1145/2939672.2939785.
24. Ke, G.; Meng, Q.; Finley, T.; et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 3146–3154.
25. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; et al. CatBoost: Unbiased Boosting with Categorical Features. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018; Volume 31, pp. 6639–6649.