

From Prediction Accuracy to Design Safety: Model Uncertainty, Interpretable Machine Learning and Resistance Factors for the Shear Strength of Reinforced Concrete Walls

Khan Muhammad Umar¹, Misha Hafza¹

¹ School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China ([✉ 202422040438@std.uestc.edu.cn](mailto:202422040438@std.uestc.edu.cn))

Received: 18 May 2025

Accepted: 21 June 2025

Available online: 22 Jun 2025

Abstract: Machine-learning models for the shear strength of reinforced concrete (RC) walls are usually validated on a random split of a test database and are reported to be far more accurate than code equations. This study re-examines that claim, and asks what such models are worth for design. A database of 208 wall tests, assembled from the ACI 445B repository and from the databases of two recent empirical models, was analysed. The ACI 318-14 in-plane shear provision was re-implemented and reproduced the tabulated predictions to within 2.1%. Its model uncertainty $\theta = V_{\text{exp}}/V_{\text{pred}}$ has a mean of 1.08 and a coefficient of variation (CoV) of 0.52; the truss model of Chandra et al. (2018) and the particle-swarm equation of Baghi et al. (2019) give CoVs of 0.46 and 0.33. A random-effects decomposition shows that 67% of the variance of $\ln \theta$ for ACI 318-14 lies between test programmes rather than within them. Gradient-boosting models were validated both with random ten-fold cross-validation and with leave-one-study-out validation. The random split gave RMSE = 118 kN ($R^2 = 0.93$); leaving a whole test programme out raised it to 298 kN, worse than the code equation (250 kN). A physics-guided model that learns the logarithm of the ACI model uncertainty from dimensionless features was the only data-driven model that generalised (219 kN, CoV of θ 0.39), and SHAP analysis showed why: ACI 318-14 overestimates the contribution of horizontal reinforcement and ignores the contribution of flanges. Reliability analysis with a first-order reliability method gave $\beta = 1.12$ for ACI 318-14 at $\phi = 0.75$ against $\beta = 1.57$ for the physics-guided model, under a gravity load model. Neither is close to the usual target, which shows that the scatter of wall shear databases, not the mean accuracy of the equation, is what limits the reliability that a shear-strength model can deliver.

Keywords: Reinforced concrete walls; shear strength; model uncertainty; machine learning; SHAP

1. Introduction

Reinforced concrete (RC) walls carry most of the lateral load in many buildings, and their shear strength is difficult to predict because it depends on the aspect ratio, on horizontal and vertical web reinforcement, on boundary elements and flanges, on axial load and on concrete strength. Code provisions such as ACI 318-14 (ACI Committee 318, 2014) use a simple additive expression with a concrete term and a horizontal reinforcement term; mechanics-based alternatives such as the truss model of Chandra et al. (2018) and optimisation-based empirical equations such as that of Baghi et al. (2019) aim at a better fit to test data. In parallel, machine-learning (ML) models have been trained on the same test databases: support vector regression combined with response surface models (Keshtegar et al., 2021, 2022), tree ensembles with SHAP interpretation (Feng et al., 2021) and classification models for failure mode (Mangalathu et al., 2020). Reported accuracies are impressive, with coefficients of determination above 0.95 being common.

Two issues make these comparisons difficult to use in design. First, accuracy statistics are not safety. A code calibration needs the model uncertainty $\theta = V_{\text{exp}}/V_{\text{pred}}$: its mean, its coefficient of variation and its distribution, because the resistance factor follows from these

(Ravindra and Galambos, 1978; Szerszen and Nowak, 2003; Nowak and Collins, 2013). Second, wall databases are strongly clustered. One test programme contributes between two and twenty nominally similar specimens that share a concrete batch, a detailing philosophy, a test rig and a loading protocol. When such a database is split at random, almost every test specimen has siblings in the training set, and the resulting error statistics measure interpolation inside a programme rather than prediction for a new one. Grouped cross-validation is the standard remedy in disciplines where clustered data are the norm (Roberts et al., 2017), but it is rarely used in structural-engineering model comparisons.

This study re-analyses a database of 208 wall tests that was used to develop a hybrid response-surface/support-vector model (Keshtegar et al., 2021) and asks three questions. How large is the model uncertainty of the code equation and of the two recent empirical models, and how is it structured? Does a modern tree ensemble still outperform them when whole test programmes are held out? And what resistance factor does each model imply? The answers are used to propose a design-oriented way of reporting data-driven shear models.

2. Database

The database contains 208 RC walls tested under in-plane shear, compiled from the ACI 445B shear wall database (Usta et al., 2017) and from the databases of Chandra et al. (2018) and Baghi et al. (2019), and used in that form by Keshtegar et al. (2021). Each record gives the concrete strength f'_c , the axial load P , the gross section area A_g , the wall height h_w , the web length l_w , the aspect ratio h_w/l_t , the web thickness t_w , the flange width b_f and thickness t_f , the boundary, vertical and horizontal reinforcement ratios and yield strengths, and the measured shear capacity V_{exp} . The tests come from 27 experimental programmes; the largest contributes 20 walls and six contribute two to four. 105 walls are rectangular ($b_f \leq t_w$) and 103 have flanges or barbell ends; 122 walls are squat ($h_w/l_t \leq 1$). Table 1 summarises the ranges and Figure 1 shows the coverage.

Table 1. Statistical properties of the 208-wall database

Variable	Minimum	Maximum	Mean	Std. dev.	CoV
f'_c (MPa)	16	138	46	26	0.57
P (kN)	0	2617	420	535	1.27
A_g (mm ²)	42250	369600	161768	81844	0.51
h_w (mm)	400	4572	1510	890	0.59
l_w (mm)	370	1905	1025	408	0.40
h_w/l_t	0.21	2.40	1.21	0.56	0.47
t_w (mm)	65	152	98	24	0.25
b_f (mm)	0	914	210	199	0.95
t_f (mm)	0	540	126	92	0.73
ρ_b	0.0000	0.9750	0.0385	0.0983	2.56
f_{yb} (MPa)	0	1009	403	235	0.58
ρ_v	0.0000	0.0290	0.0076	0.0064	0.85
f_{yv} (MPa)	0	1420	481	152	0.32
ρ_h	0.0000	0.0244	0.0066	0.0044	0.66
f_{yh} (MPa)	0	1420	478	164	0.34
V_{exp} (kN)	70	2483	601	436	0.72

l_t is the total wall length used in the source database; ρ_b , ρ_v and ρ_h are the boundary, vertical web and horizontal web reinforcement ratios.

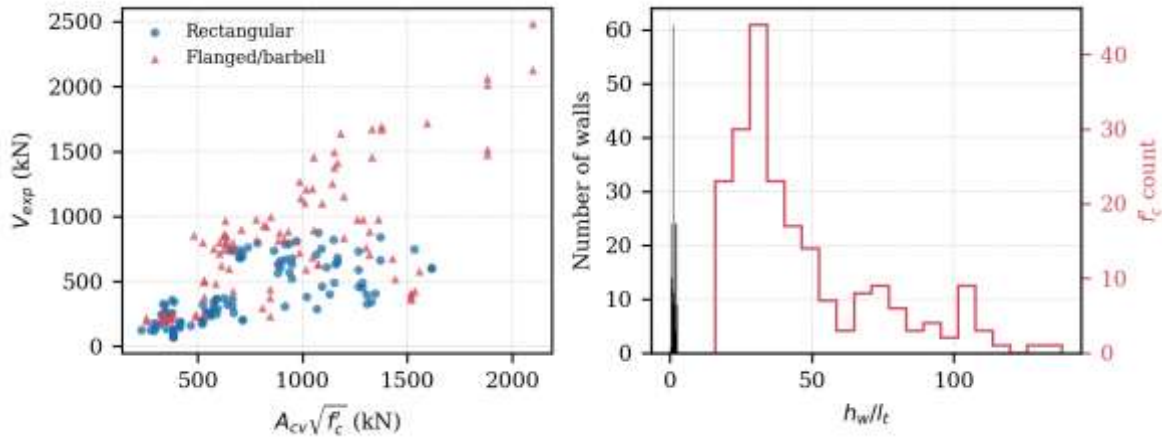


Figure 1. Database coverage: (a) measured shear capacity against the nominal shear area parameter $A_{cv}\sqrt{f'_c}$ for rectangular and flanged walls; (b) distribution of the aspect ratio (bars) and of the concrete strength (line).

Three records report a boundary reinforcement ratio above 0.2, which is physically implausible and is most probably a unit error in the source compilation; they were retained so that the analysis reproduces exactly the data used in the earlier studies, and the tree models used here are insensitive to such outliers in a single input.

3. Strength models

3.1. Code and empirical models

The in-plane shear strength of a wall in ACI 318-14 is

$$V_n = A_{cv}(\alpha_c \lambda \sqrt{f'_c} + \rho_t f_{yt}) \quad (1)$$

with A_{cv} the gross area of the web resisting shear, $\lambda = 1$ for normal-weight concrete, ρ_t and f_{yt} the ratio and yield strength of the horizontal (transverse) reinforcement, and $\alpha_c = 0.25$ for $h_w/l_w \leq 1.5$, $\alpha_c = 0.17$ for $h_w/l_w \geq 2.0$ and linear interpolation in between (SI units, MPa). This expression was re-implemented with $A_{cv} = t_w h$ and with α_c evaluated from the tabulated aspect ratio; the predictions reproduce the values tabulated in the source database with a median ratio of 1.000 and a

maximum deviation of 2.1%, which is the rounding of the published ratios. The code upper limit $V_n \leq 0.83 A_{cv}\sqrt{f'_c}$ for an individual wall segment governs 59 of the 208 walls and is reported separately, because it is a capacity limit rather than a strength prediction.

The truss model of Chandra et al. (2018) and the particle-swarm-calibrated equation of Baghi et al. (2019) are used through the predictions tabulated in the database; both are documented in the original publications and are not repeated here.

3.2. Data-driven models

Four data-driven models were trained. *XGB-raw* is a gradient-boosted tree ensemble (Friedman, 2001; Chen and Guestrin, 2016) trained on $\ln V_{exp}$ with the fifteen raw inputs of Table 1, i.e. the input set used in earlier ML studies of this database. *XGB-norm* is trained on the logarithm of the normalised shear stress $v = V_{exp}/(A_{cv}\sqrt{f'_c})$ with eight dimensionless features: h_w/l_t , $P/(A_g f'_c)$, $\rho_h f_{yh}/f'_c$, $\rho_v f_{yv}/f'_c$, $\rho_b f_{yb}/f'_c$, b_f/t_w , A_g/A_{cv} and f'_c . *XGB-ACIres* uses the same features but its target is $\ln \theta_{ACI} = \ln(V_{exp}/V_{ACI})$, so the code equation carries the

trend and the ensemble learns only a correction; this is a physics-guided or residual formulation. *Linear-mech* is a re-calibrated mechanical expression, $V = A_{cv}(a\sqrt{f'_c} + b\rho_h f_{yh} + c\rho_v f_{yv}) + eP$, fitted by non-negative least squares; it shows how much of the improvement is simply a re-calibration of the code format. Hyper-parameters (depth 2-3 and 300-700 trees, learning rate 0.03) were selected inside each training set by grouped cross-validation, so that no information from the held-out programme entered the selection. All models were implemented with scikit-learn (Pedregosa et al., 2011) and XGBoost.

3.3. Validation schemes

Two schemes were used: *random* ten-fold cross-validation, and *leave-one-study-out* (LOSO) validation in which all walls of a test programme are removed from training and predicted together. LOSO reproduces the situation of a designer applying the model to a wall that was not built in any of the laboratories represented in the database. The random scheme was repeated ten times with different fold assignments to quantify its variability.

4. Model uncertainty and reliability framework

The model uncertainty of each strength model is $\theta_i = V_{exp,i}/V_{pred,i}$. Its mean, CoV, quantiles, lognormality (Anderson-Darling and Kolmogorov-Smirnov tests on $\ln \theta$) and its trends with the main design parameters (Spearman rank correlations) were computed for the whole database and for subsets by section shape and aspect ratio. The mean-value correction b and the error CoV V_δ of EN 1990 Annex D (CEN, 2002) were also evaluated, together with the design factor

$$\psi_D = \frac{r_d}{r_t(X_m)} = b \exp(-k_{d,\infty} \alpha_{rt} Q_{rt} - k_{d,n} \alpha_\delta Q_\delta - 0.5Q^2) \quad (2)$$

with $k_{d,\infty} = k_{d,n} = \alpha_R \beta = 0.8 \times 3.8 = 3.04$ for a large number of tests, and V_{rt} the CoV of the basic variables, varied from 0.05 to 0.15.

Because walls are grouped by test programme, $\ln \theta$ was also decomposed with a one-way random-effects model, $\ln \theta_{ij} = \mu + u_i + e_{ij}$, whose intraclass correlation $ICC = \sigma_b^2/(\sigma_b^2 + \sigma_w^2)$ measures the share of the model uncertainty that belongs to the laboratory rather than to the specimen.

Resistance factors were obtained from a first-order reliability analysis of the limit state

$$g = \theta B_M V_n - D - L \quad (3)$$

in which V_n is the nominal shear strength from the model, B_M is a lognormal material and geometry factor with mean 1.0 and CoV $V_M = 0.10$, θ is lognormal with the mean and CoV measured on the database, D is a normal dead load effect with bias 1.05 and CoV 0.10 and L is a Gumbel 50-year maximum live load effect with

bias 1.0 and CoV 0.18, following the load model of Ellingwood et al. (1982) as used in the calibration of ACI 318 (Nowak and Szerszen, 2003; Szerszen and Nowak, 2003). The nominal strength follows from $\phi V_n = 1.2D_n + 1.6L_n$ with L_n/D_n between 0.5 and 3. The reliability index was computed with the Hasofer-Lind-Rackwitz-Fiessler algorithm and checked by importance sampling, and the resistance factor required for $\beta_T = 3.5$, the target of the ACI 318 calibration for members other than columns, was obtained by root-finding. The use of a gravity load combination is a deliberate simplification that isolates the effect of the resistance model; its implications are discussed in Section 6.

Finally, the residual model was interpreted with SHAP values (Lundberg and Lee, 2017), which attribute the predicted $\ln \theta_{ACI}$ of each wall to its features and therefore show *where* and *why* the code equation is biased.

5. Results

5.1. Model uncertainty of the existing models

Table 2 gives the statistics of θ . ACI 318-14 has a mean of 1.079 with a CoV of 0.516 and a median of 0.926: the equation is unbiased on average but unconservative for more than half of the walls (55%), with a 5% quantile of θ of 0.46. The truss model of Chandra et al. (2018) is conservative on average (mean 1.181) with a similar CoV (0.460), and the equation of Baghi et al. (2019) has the smallest scatter of the three (0.327) at the price of a mean of 1.130. Introducing the code upper limit lowers the mean prediction further (mean $\theta = 1.153$) without reducing the scatter.

Lognormality of θ is not rejected for any of the three models at the 5% level (Kolmogorov-Smirnov p between 0.20 and 0.75), which supports the lognormal assumption used in the reliability analysis.

5.2. Where the code equation is biased

Spearman correlations show strong, systematic trends for ACI 318-14: θ decreases with the horizontal reinforcement index $\rho_h f_{yh}/f'_c$ ($\rho_s = -0.42$, $p < 0.001$) and increases with the flange index A_g/A_{cv} ($\rho_s = +0.44$, $p < 0.001$) and with the concrete strength ($\rho_s = +0.23$, $p = 0.001$). In words, the code overestimates the benefit of heavy horizontal web reinforcement, an observation made experimentally by Wood (1990) and confirmed by later database studies (Gulec and Whittaker, 2011), and it neglects the strength provided by flanges and boundary elements, which it excludes from A_{cv} . The truss model is biased mainly in the aspect ratio ($\rho_s = -0.32$) and in the axial load ratio ($\rho_s = -0.34$). By contrast, the physics-guided ensemble leaves no significant trend ($|\rho_s| \leq 0.06$, all $p > 0.35$) even under leave-one-study-out prediction, which is the clearest evidence that the corrections it learns are real mechanics and not noise. Figure 2 shows the two patterns.

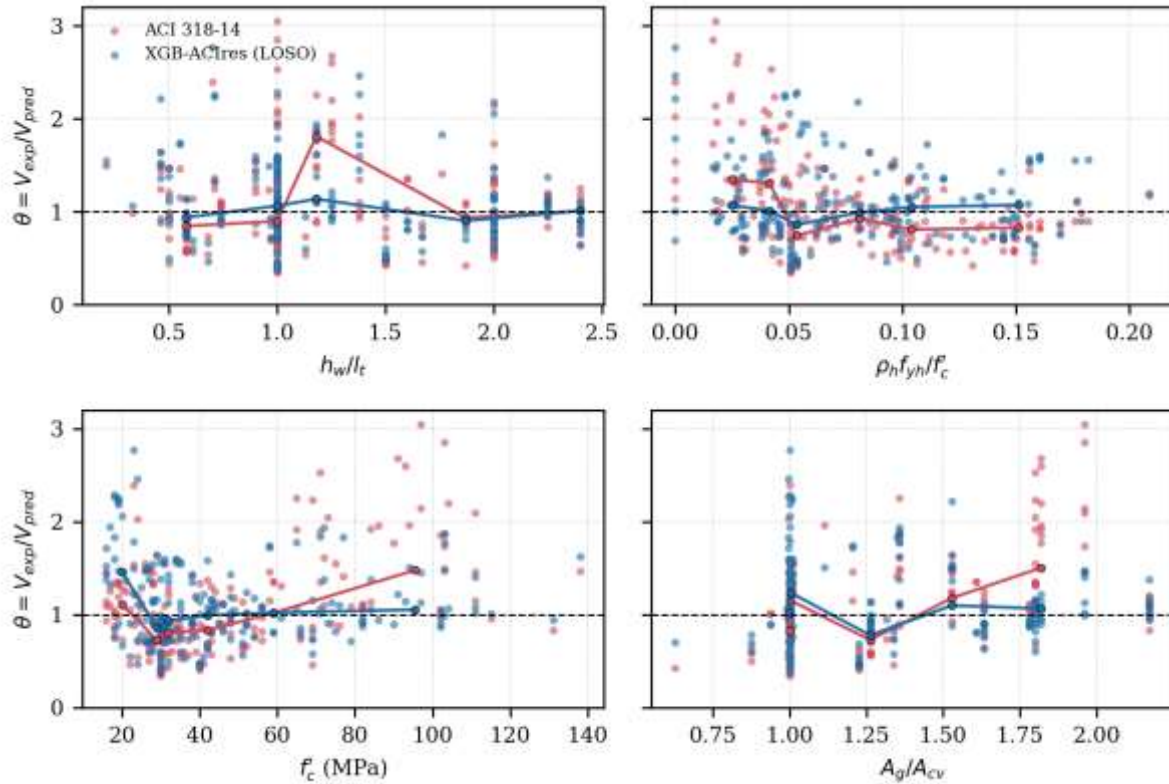


Figure 2. Model uncertainty against (a) aspect ratio, (b) horizontal reinforcement index, (c) concrete strength and (d) flange index, for ACI 318-14 and for the leave-one-study-out predictions of the physics-guided ensemble. Lines connect medians of six equal-count bins.

The subset statistics (Table 3) make the same point in design terms. For rectangular walls ACI 318-14 has a mean θ of 0.91, i.e. it is unconservative, while for flanged and barbell walls the mean is 1.25. For squat

walls ($h_w/l_t \leq 1$) the CoV reaches 0.56. A single resistance factor applied to this equation is therefore not equally safe across wall types (Figure 3).

Table 2. Model uncertainty and accuracy of all models on the 208-wall database

Model	Mean θ	CoV of θ	$\theta_{0.05}$	$\theta < 1$ (%)	RMSE (kN)	R^2	within $\pm 15\%$ (%)
Code and empirical models							
ACI 318-14	1.079	0.516	0.46	55	250	0.670	26
Chandra et al. (2018)	1.181	0.460	0.49	40	234	0.711	26
Baghi et al. (2019)	1.130	0.327	0.55	40	226	0.729	28
ACI 318-14 (0.83 limit)	1.153	0.482	0.46	45	270	0.613	25
Data-driven models, random 10-fold CV							
XGB-raw	1.027	0.247	0.74	50	118	0.926	71
XGB-norm	1.033	0.219	0.65	46	131	0.909	57
XGB-ACIres	1.033	0.224	0.63	48	137	0.901	53
Linear-mech	0.984	0.420	0.44	57	220	0.744	32
Data-driven models, leave-one-study-out							
XGB-raw	1.115	0.402	0.51	40	298	0.529	37
XGB-norm	1.115	0.386	0.47	46	260	0.643	26
XGB-ACIres	1.102	0.392	0.53	50	219	0.747	37
Linear-mech	0.992	0.448	0.43	57	240	0.694	28

Out-of-fold values for the data-driven models. $\theta_{0.05}$ is the empirical 5% quantile.

Table 3. Mean and CoV of θ by wall type

Subset	n	ACI 318-14	Chandra et al. (2018)	Baghi et al. (2019)	XGB-ACIres (LOSO)
Section shape					
Rectangular	105	0.91 / 0.37	1.16 / 0.56	1.12 / 0.31	1.15 / 0.43
Flanged/barbell	103	1.25 / 0.54	1.21 / 0.34	1.14 / 0.35	1.06 / 0.34
Aspect ratio					
$h_w/l_t \leq 1$	122	1.06 / 0.56	1.28 / 0.45	1.18 / 0.32	1.11 / 0.39
$h_w/l_t > 1$	86	1.11 / 0.46	1.04 / 0.45	1.06 / 0.33	1.09 / 0.40

Each cell gives mean θ / CoV of θ .

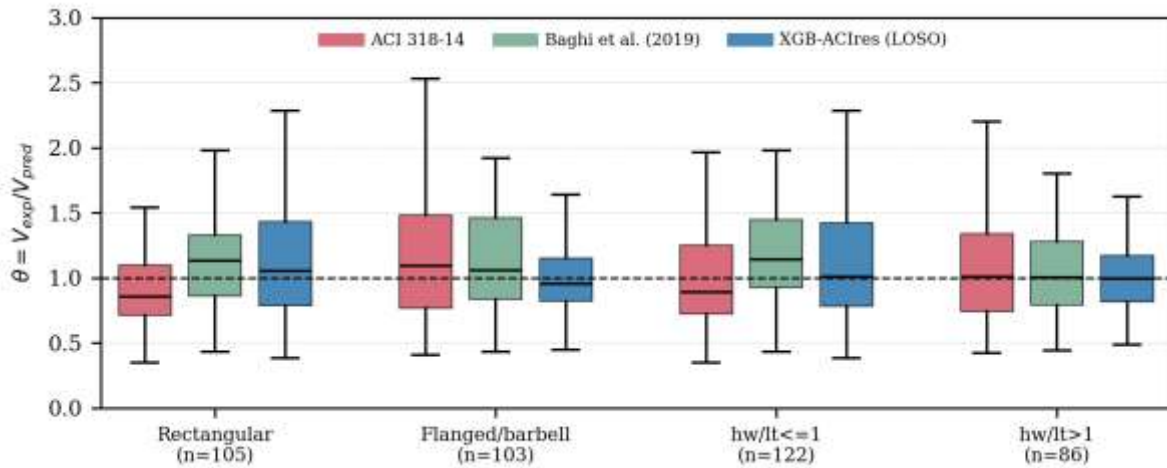


Figure 3. Model uncertainty by wall type for the code equation, the best empirical equation and the physics-guided ensemble (boxes: median, quartiles, 1.5 IQR whiskers).

5.3. How much of the scatter is a laboratory effect?

The random-effects decomposition (Table 4, Figure 4) gives ICC values between 0.62 and 0.81. For ACI 318-14 the between-programme standard deviation of $\ln \theta$ is 0.380 against 0.266 within a programme, so about 67% of the model uncertainty is a property of the test programme. Figure 5, which plots the median θ of each programme, makes the effect visible: whole programmes are predicted with a median θ as low as 0.44 or as high as 2.09, a spread far larger than the scatter inside most programmes. Part of this is real (different detailing, loading protocol, failure mode), and part is the usual laboratory bias, but for validation the consequence is the same: the effective sample size of this database is closer to its 27 programmes than to its 208 walls.

Table 4. Random-effects decomposition of $\ln \theta$ and lognormality test

Model	σ_{total}	σ_b	σ_w	ICC	KS p -value
ACI 318-14	0.448	0.380	0.266	0.67	0.75
Chandra et al. (2018)	0.449	0.394	0.227	0.75	0.66
Baghi et al. (2019)	0.349	0.283	0.223	0.62	0.42
XGB-ACIres (LOSO)	0.381	0.331	0.200	0.73	0.56
XGB-raw (LOSO)	0.399	0.365	0.229	0.72	0.20
Linear-mech (LOSO)	0.430	0.415	0.202	0.81	0.42

σ_b and σ_w are the between- and within-programme standard deviations (restricted maximum likelihood).

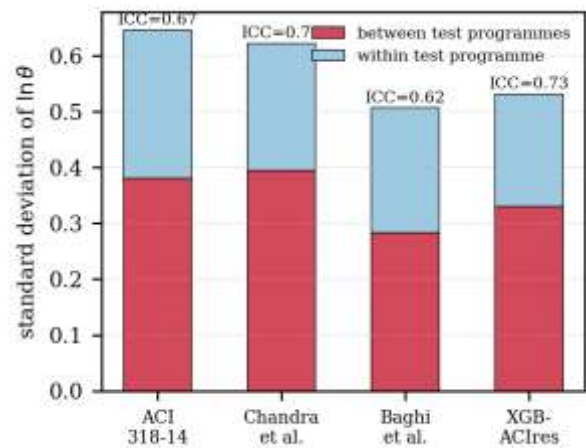


Figure 4. Between- and within-programme standard deviation of $\ln \theta$ with the intraclass correlation.

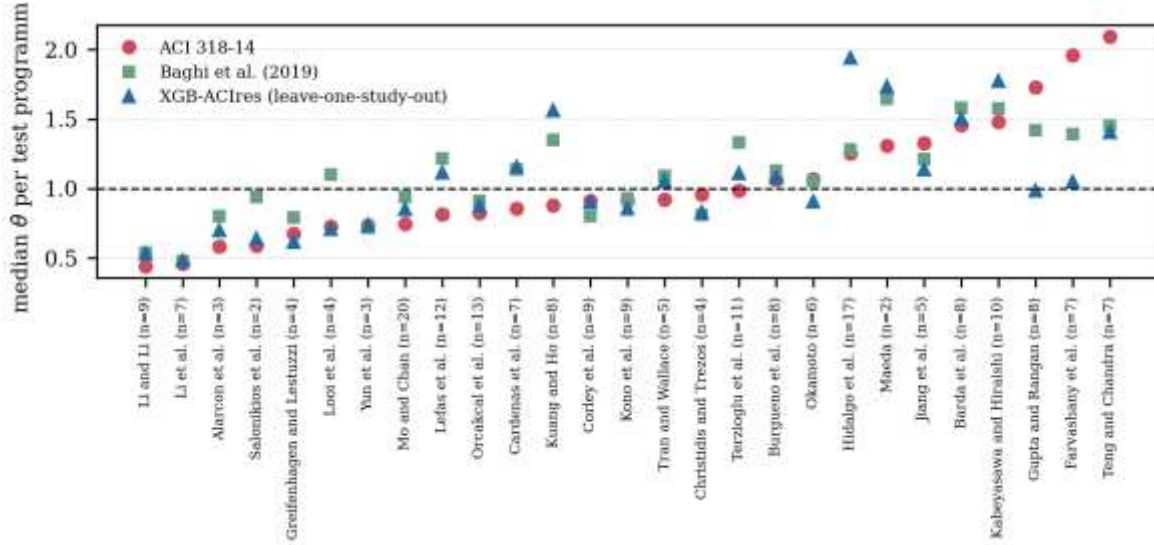


Figure 5. Median model uncertainty of each test programme for ACI 318-14, the equation of Baghi et al. (2019) and the leave-one-study-out predictions of the physics-guided ensemble; programmes sorted by the ACI value.

5.4. Random versus leave-one-study-out validation

Table 2 and Figure 6 give the central result. Under random ten-fold cross-validation the ensemble trained on the raw inputs reaches $RMSE = 118$ kN and $R^2 = 0.926$, with 71% of the walls predicted within $\pm 15\%$; this is the kind of performance reported in the ML literature for this database. Under leave-one-study-out validation the same model gives $RMSE = 298$ kN and $R^2 = 0.529$, which is worse than ACI 318-14 (250 kN) and than both empirical models. The repeated random cross-validation was stable (123 ± 13 kN over ten repetitions), so the gap is not sampling noise: it is the price of the clustered structure of the database.

The two physics-guided variants degrade much less. XGB-norm, which predicts a normalised shear stress, gives 260 kN, and XGB-ACIres, which corrects the code equation, gives 219 kN with a CoV of θ of 0.39, against 0.52 for the code equation itself and 0.33 for the best empirical equation. Its advantage over the code is a 12% reduction in RMSE and a 24% reduction in the CoV of the model uncertainty; this is a real but modest gain, an order of magnitude smaller than the gain suggested by a random split. The re-calibrated mechanical expression, whose fitted coefficients on the full database are $0.25\sqrt{f'_c}$, $0.64\rho_{fh}f_{yh}$ and $0.14\rho_{fv}f_{yv}$ (MPa) plus $0.24P$, is not competitive (240 kN), which shows that the additive code format itself is the limiting factor.

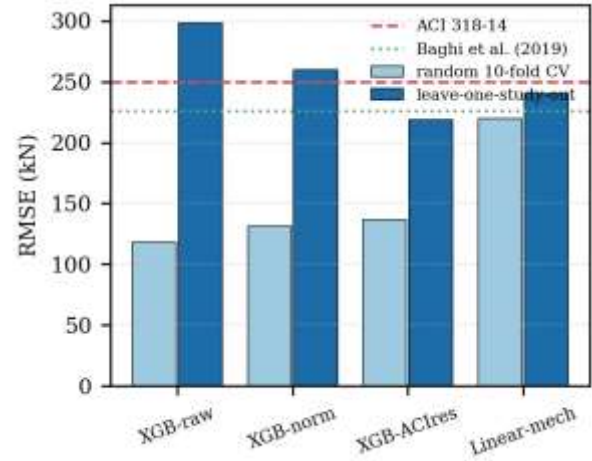


Figure 6. RMSE of the data-driven models under random ten-fold and leave-one-study-out validation, compared with the code and empirical equations (horizontal lines).

5.5. What the ensemble learned

Figure 7 shows the SHAP decomposition of the predicted $\ln \theta_{ACI}$. The dominant feature is the horizontal reinforcement index (mean $|SHAP| = 0.186$), with a clearly decreasing effect: walls with $\rho_{fh}f_{yh}/f'_c$ below about 0.03 need an upward correction of the code prediction of up to +60%, while heavily reinforced walls need a downward correction of about -20%. The second feature is the flange index A_g/A_{cv} (mean $|SHAP| = 0.145$): walls whose gross area exceeds the web area by more than about 50% are systematically stronger than the code prediction, because the flanges and boundary elements carry shear that the code ignores. The axial load ratio (mean $|SHAP| = 0.055$), the concrete strength (0.049) and the aspect ratio (0.045) follow. These are exactly the effects reported from experiments and from mechanical models, so the ensemble is not merely fitting noise; but the effects it captures are worth about a fifth

of the scatter, not the four fifths that a random-split comparison suggests.

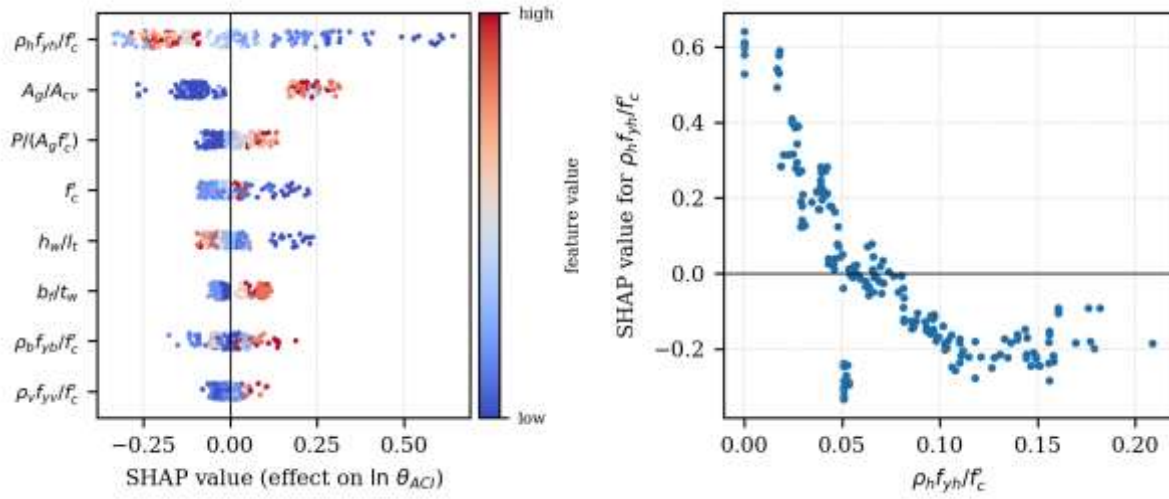


Figure 7. SHAP interpretation of the physics-guided ensemble: (a) contribution of each feature to the predicted logarithm of the ACI model uncertainty; (b) dependence on the horizontal reinforcement index.

5.6. Reliability and resistance factors

Table 5 and Figure 8 give the reliability results. With the gravity load model and $L_n/D_n = 1$, the reliability index of a wall designed with ACI 318-14 and $\phi = 0.75$ is $\beta = 1.12$; the empirical model of Baghi et al. (2019) gives $\beta = 1.95$ and the physics-guided ensemble $\beta = 1.57$. Importance sampling with 4×10^5 samples confirmed the FORM index for ACI 318-14 ($\beta_{\text{FORM}} = 1.12$, $\beta_{\text{IS}} = 1.10$). The model uncertainty dominates completely: its FORM importance factor is $\alpha^2 = 0.92$, against 0.04 for material and geometry and 0.03 for the live load.

None of the models reaches the target $\beta_T = 3.5$ at any realistic resistance factor: the factors that would be required are $\phi = 0.22$ for ACI 318-14, $\phi = 0.44$ for the best empirical equation and $\phi = 0.34$ for the ensemble. Two readings of this result are possible, and both are useful. Read literally, it says that a wall designed only for shear against a gravity-type load combination, with the scatter observed in this database, would not meet the

target reliability; the code reaches an acceptable level by other means, namely capacity design that forces flexural yielding before shear failure, the upper limit on V_n , and the much smaller variability of the walls actually built compared with a database assembled from research specimens driven deliberately to shear failure. Read comparatively, the β values rank the models in a way that error statistics cannot: moving from ACI 318-14 to the physics-guided ensemble buys $\Delta\beta = 0.45$ at the same resistance factor, or 53% more usable strength at the same reliability.

The EN 1990 Annex D design factors in Table 5 tell the same story in a code format: with $V_{rt} = 0.10$ the design value is $\psi_D = 0.23$ of the ACI mean prediction and 0.29 of the ensemble prediction. Subset calibration matters as well: the factor required for $\beta_T = 3.5$ with ACI 318-14 is 0.19 for squat walls and 0.28 for slender walls, while the ensemble needs 0.35 and 0.34, i.e. it is much more uniform across wall types.

Table 5. Reliability of a wall in shear ($V_M = 0.10$) and design factors

Model	Mean θ	CoV of θ	β at $\phi = 0.75$ ($L/D = 1$)	β at $\phi = 0.75$ ($L/D = 3$)	ϕ for $\beta_T = 3.5$	ψ_D ($V_{rt} = 0.05$)	ψ_D ($V_{rt} = 0.10$)
ACI 318-14	1.079	0.516	1.12	1.28	0.22	0.23	0.23
Chandra et al. (2018)	1.181	0.460	1.48	1.64	0.30	0.26	0.25
Baghi et al. (2019)	1.130	0.327	1.95	2.13	0.44	0.36	0.35
XGB-raw (LOSO)	1.115	0.402	1.56	1.74	0.34	0.32	0.31
XGB-norm (LOSO)	1.115	0.386	1.63	1.81	0.35	0.28	0.27
XGB-ACIres (LOSO)	1.102	0.392	1.57	1.75	0.34	0.30	0.29
Linear-mech (LOSO)	0.992	0.448	1.13	1.31	0.26	0.24	0.24

ψ_D is the EN 1990 Annex D design value as a fraction of the mean prediction.

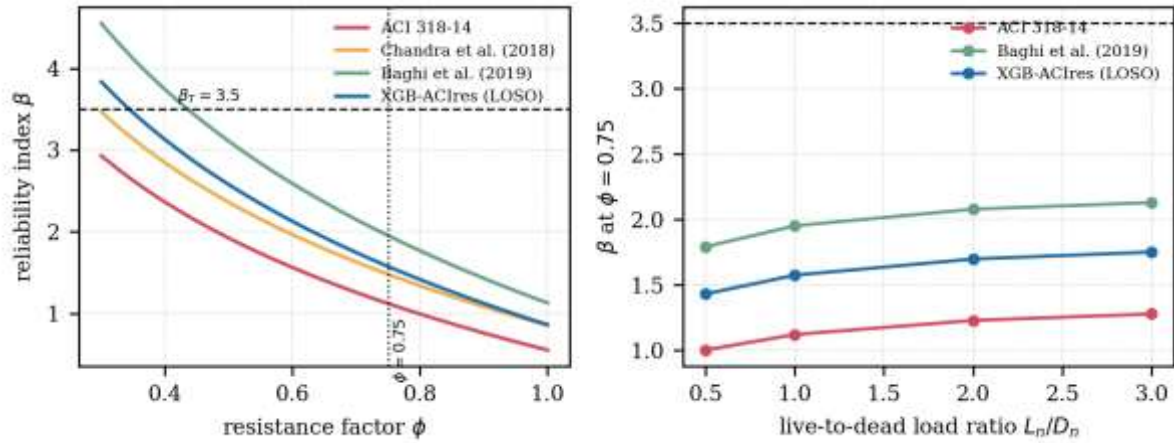


Figure 8. Reliability-based calibration: (a) reliability index against the resistance factor; (b) reliability index at $\phi = 0.75$ for different live-to-dead load ratios.

6. Discussion

The comparison of validation schemes is the practical message of this study. A tree ensemble trained on the raw inputs of a wall database can predict held-out specimens from the same laboratories almost perfectly, and this is what a random split measures; but it cannot predict a laboratory it has never seen any better than a code equation can. The reason is visible in the variance decomposition: two thirds of the model uncertainty is a programme effect that no set of input variables in the database can explain, because it is produced by factors that are not recorded — casting and curing, instrumentation, the loading protocol, the interpretation of "failure", the choice of specimens to publish. Any model, mechanical or data-driven, inherits this floor.

It does not follow that data-driven models are useless. The physics-guided residual model was the only one that both generalised across laboratories and removed the systematic trends of the code equation, and its SHAP decomposition reproduces known mechanics: the diminishing effectiveness of horizontal web reinforcement, which is the basis of the reduced steel contribution in the models of Wood (1990) and Gulec and Whittaker (2011), and the contribution of flanges, which ACI 318-14 excludes by definition. A correction of this kind, expressed as a function of two or three dimensionless quantities, is a plausible route towards improved code provisions, and it is testable: the correction predicted for a new test programme can be compared with the measured one, which is exactly the leave-one-study-out experiment reported here.

The reliability results must be read with their assumptions. The gravity load combination is a simplification: walls are designed for wind or seismic actions whose load statistics, load factors and target reliabilities differ, and in seismic design the shear demand follows from capacity design rather than from a load combination. The analysis also uses the model uncertainty of a research database, in which specimens are deliberately proportioned to fail in shear and cover

an extremely wide parameter range; the variability of a code-conforming wall in a real building is smaller. What the analysis does establish, and what does not depend on these assumptions, is the ranking of the models by the reliability they can support at a given resistance factor, the dominance of the model uncertainty over all other sources (α^2 above 0.9), and the fact that the reliability is limited by the scatter rather than by the mean bias of the equations.

Three limitations should be noted. First, the predictions of the truss model and of the particle-swarm equation were taken from the compiled database rather than re-implemented, so small rounding differences are possible; the ACI values were re-implemented and verified. Second, three walls carry an implausible boundary reinforcement ratio; removing them changes no conclusion but the database would benefit from a systematic audit. Third, the database is limited to walls with rectangular, flanged or barbell sections tested under in-plane shear, and to concrete strengths up to 138 MPa; the results should not be extrapolated to coupled walls, to walls with openings or to walls with fibre-reinforced concrete.

7. Conclusions

A database of 208 RC wall tests was re-analysed with the tools of code calibration rather than of prediction accuracy. The main conclusions are:

(1) The model uncertainty of ACI 318-14 in-plane shear has a mean of 1.08 and a CoV of 0.52 on this database; the truss model of Chandra et al. (2018) and the equation of Baghi et al. (2019) have CoVs of 0.46 and 0.33. The code equation is unconservative for rectangular and squat walls and conservative for flanged walls.

(2) Between 62% and 81% of the variance of $\ln \theta$ lies between test programmes. The effective sample size of the database is its 27 programmes, not its 208 walls.

(3) A gradient-boosting model on raw inputs reaches RMSE = 118 kN under random cross-validation but 298 kN when a whole test programme is held out, which is

worse than the code equation (250 kN). Published accuracies obtained from random splits of clustered wall databases should not be read as design-relevant.

(4) A physics-guided ensemble that learns the logarithm of the code model uncertainty from eight dimensionless features generalises across laboratories (219 kN, CoV of 0.39) and removes all significant residual trends. SHAP analysis attributes its corrections to the diminishing effectiveness of horizontal reinforcement and to the neglected contribution of flanges.

(5) Under a gravity load model, $\beta = 1.12$ at $\phi = 0.75$ for ACI 318-14 against 1.57 for the physics-guided ensemble, and the model uncertainty accounts for 92% of the FORM sensitivity. Reducing the scatter of shear models, not their mean error, is what would allow higher resistance factors.

Reporting the mean, the CoV and the grouped out-of-sample performance of θ , together with the resistance factor implied by the model, would make new shear-strength models directly usable by code committees.

Data and code availability

The cleaned database, the implementation of the code equation, the machine-learning, SHAP, variance-decomposition and reliability scripts, and the scripts that generate every figure and table are provided as supplementary material.

References

- ACI Committee 318 (2014). Building Code Requirements for Structural Concrete (ACI 318-14) and Commentary (ACI 318R-14). American Concrete Institute, Farmington Hills, MI.
- Baghi, H., Baghi, H., & Siavashi, S. (2019). Novel empirical expression to predict shear strength of reinforced concrete walls based on particle swarm optimization. *ACI Structural Journal*, 116(5), 247–260. <https://doi.org/10.14359/51716773>
- CEN (2002). EN 1990:2002 Eurocode – Basis of Structural Design, Annex D: Design assisted by testing. European Committee for Standardization, Brussels.
- Chandra, J., Chanthabouala, K., & Teng, S. (2018). Truss model for shear strength of structural concrete walls. *ACI Structural Journal*, 115(2), 323–335. <https://doi.org/10.14359/51701129>
- Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- Ellingwood, B., MacGregor, J.G., Galambos, T.V., & Cornell, C.A. (1982). Probability based load criteria: load factors and load combinations. *Journal of the Structural Division*, 108(5), 978–997. <https://doi.org/10.1061/JSDEAG.0005959>
- Feng, D.-C., Wang, W.-J., Mangalathu, S., & Taciroglu, E. (2021). Interpretable XGBoost-SHAP machine-learning model for shear strength prediction of squat RC walls. *Journal of Structural Engineering*, 147(11), 04021173. [https://doi.org/10.1061/\(ASCE\)ST.1943-541X.0003115](https://doi.org/10.1061/(ASCE)ST.1943-541X.0003115)
- Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5). <https://doi.org/10.1214/aos/1013203451>
- Gulec, C.K., & Whittaker, A.S. (2011). Empirical equations for peak shear strength of low aspect ratio reinforced concrete walls. *ACI Structural Journal*, 108(1), 80–89. <https://doi.org/10.14359/51664205>
- Keshtegar, B., Nehdi, M.L., Kolahchi, R., Trung, N.-T., & Bagheri, M. (2022). Novel hybrid machine learning model for predicting shear strength of reinforced concrete shear walls. *Engineering with Computers*, 38(Suppl. 5), 3915–3926. <https://doi.org/10.1007/s00366-021-01302-0>
- Keshtegar, B., Nehdi, M.L., Trung, N.-T., & Kolahchi, R. (2021). Predicting load capacity of shear walls using SVR–RSM model. *Applied Soft Computing*, 112, 107739. <https://doi.org/10.1016/j.asoc.2021.107739>
- Lundberg, S.M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).
- Mangalathu, S., Jang, H., Hwang, S.-H., & Jeon, J.-S. (2020). Data-driven machine-learning-based seismic failure mode identification of reinforced concrete shear walls. *Engineering Structures*, 208, 110331. <https://doi.org/10.1016/j.engstruct.2020.110331>
- Nowak, A.S., & Collins, K.R. (2013). *Reliability of Structures* (2nd ed.). CRC Press, Boca Raton, FL.
- Nowak, A.S., & Szerszen, M.M. (2003). Calibration of design code for buildings (ACI 318): Part 1 – statistical models for resistance. *ACI Structural Journal*, 100(3), 377–382. <https://doi.org/10.14359/12613>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Ravindra, M.K., & Galambos, T.V. (1978). Load and resistance factor design for steel. *Journal of the Structural Division*, 104(9), 1337–1353. <https://doi.org/10.1061/JSDEAG.0004981>
- Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillerá-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., & Dormann, C.F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929. <https://doi.org/10.1111/ecog.02881>
- Szerszen, M.M., & Nowak, A.S. (2003). Calibration of design code for buildings (ACI 318): Part 2 – reliability analysis and resistance factors. *ACI Structural Journal*, 100(3), 383–391. <https://doi.org/10.14359/12614>
- Usta, M., Pujol, S., ACI Subcommittee 445B, Puranam, A., Song, C., & Wang, Y. (2017). ACI 445B Shear

Wall Database (Version 1.0). Purdue University
Research Repository.
<https://doi.org/10.4231/R7HH6H39>

Wood, S.L. (1990). Shear strength of low-rise reinforced
concrete walls. *ACI Structural Journal*, 87(1), 99–107.
<https://doi.org/10.14359/2951>