Civil and Environmental Engineering
Vol. 0, Issue 0, DOI: 10.2478/cee-2027-0026

Cross-database validation and calibrated uncertainty for machine learning models of the shear capacity of FRP-strengthened RC beams

Sajjad E. RASHEED[1,*], Hasan A. ZAINY[2]

[1] Civil Engineering Department, University of Karbala, Karbala 56001, Iraq.

[2] Civil Engineer, General Secretariat of the Holy Imam Hussain Shrine, Karbala, Iraq.

* corresponding author: sajjad.e@uokerbala.edu.iq

Date of Submission: 6 August 2026
Revision Date: 2 September 2026
Date of Acceptance: 2 September 2026

Civil and Environmental Engineering
Journal of the Faculty of Civil Engineering | University of Žilina

Abstract

Machine learning models for the shear contribution of externally bonded fibre-reinforced polymer (FRP) to reinforced concrete beams are validated almost exclusively on a random hold-out split of a single self-compiled database. Because such databases are assembled from a few dozen experimental programmes, each contributing several near-identical specimens, a random split leaves siblings of nearly every test specimen in the training set. This study provides a cross-database assessment of that practice. Two independently compiled databases (442 and 455 specimens) were harmonised into a provenance-traced dataset of 828 records from 167 programmes spanning 1992-2024, and six learners were evaluated under five regimes of increasing severity. For gradient boosting the coefficient of determination fell from 0.67 and 0.69 under random splitting to 0.39 and 0.22 under study-grouped validation, and to 0.21 under cross-database transfer. A control in which the same physical specimens were scored under each compilation in turn changed little (0.35 versus 0.37), showing that clustering by programme, not curation disagreement, is responsible. A hierarchical errors-in-variables model attributes 38% of the residual variance to programme identity and delivers calibrated predictive intervals (91.8% coverage at a nominal 90%), which span a factor of about twelve. Design provisions remain less accurate than the fitted models, but the reported margin roughly halves under honest validation. A reporting protocol is proposed, requiring publication of the specimen-to-study mapping, validation grouped on it, disclosure of out-of-range evaluation cases, and reporting of interval coverage and a characteristic value.

Keywords

FRP shear strengthening; Machine learning; Cross-database validation; Data leakage; Hierarchical Bayesian model; Predictive uncertainty.

1. Introduction

Externally bonded fibre-reinforced polymer (FRP) is an established means of increasing the shear capacity of deficient reinforced concrete (RC) beams, but the mechanics of the FRP contribution remain contested. The contribution depends on a bond-limited effective strain that is itself a function of the geometry, the concrete strength, the anchorage

© Author(s) 2026. This work is distributed under the Creative Commons BY 4.0 license (https://creativecommons.org/licenses/by/4.0/).

condition and the interaction with internal stirrups, and no mechanically derived expression commands consensus (Chen & Teng, 2003a, 2003b; Teng et al., 2004; Monti & Liotta, 2007; Gonzalez-Libreros et al., 2017). This has made the problem attractive to data-driven modelling. Machine learning (ML) is now applied widely across structural engineering (Sun et al., 2021; Naser, 2021; Thai, 2022), including to shear problems that resist closed-form treatment (Feng et al., 2021), and a substantial body of work reports ML models of the FRP shear contribution trained on compiled experimental databases (Perera et al., 2010; Marani & Nehdi, 2022; Wang et al., 2023; Benzaamia et al., 2024; Ke et al., 2024; Ombres et al., 2024; Liu et al., 2024; Zhao et al., 2025).

These studies share a validation protocol. A database is compiled from the literature and a fraction of the records is withheld at random. A model is then fitted to the remainder, and the coefficient of determination on the withheld fraction is reported as evidence of predictive capability. This is typically accompanied by a comparison against design equations, which the ML model invariably wins. The protocol is inherited from mainstream supervised learning, where it is appropriate because observations are exchangeable. Compiled structural test databases are not exchangeable. In the merged dataset assembled here, 828 specimens originate from 167 experimental programmes, contributing a median of four specimens each and up to eighteen, with two thirds of programmes contributing three or more. Specimens within a programme share a concrete batch, a fabrication protocol, an instrumentation scheme, a single laboratory's systematic measurement bias, and usually a deliberate parametric design in which only one variable is varied. Under random splitting, siblings of nearly every withheld specimen remain in the training set. What is reported as generalisation is therefore closer to interpolation within programmes that the model has already seen. This is a recognised failure mode across data-driven science, where it is one of the leading causes of irreproducible performance claims (Kapoor & Narayanan, 2023), but it has not been examined in the FRP shear literature.

A stronger test is available but has not, to the authors' knowledge, been performed: train a model on one independently compiled database and evaluate it on another. Independent compilation matters because two research groups draw on partly disjoint source literature, apply different inclusion criteria, and transcribe the same primary papers in different ways. A model that survives that transfer has demonstrated something that no random hold-out can demonstrate. The obstacle has been that such an experiment requires two databases of the same phenomenon that were genuinely built independently, together with a means of identifying the specimens they share, since leaving shared specimens in place would leak the test set into training.

This study performs that experiment. It uses the databases published with two recent and independent ML studies of FRP shear strengthening (Wang et al., 2023; Zhao et al., 2025), identifies their common specimens through provenance rather than through numerical similarity, and evaluates six learners under a graded sequence of validation regimes of increasing severity. It then asks what the correct statistical treatment of such data would be. The answer developed here is a hierarchical errors-in-variables model, in which the clustering by experimental programme is represented explicitly. Its covariates are treated as quantities measured with error, and the magnitude of that error is estimated from the specimens the two compilations share rather than assumed.

The contributions are fourfold. First, to the authors' knowledge the first cross-database generalisation assessment for ML models of FRP shear capacity, with shared specimens removed. Second, a decomposition of the generalisation gap into a component attributable to clustering by experimental programme and a component attributable to curation, using a control in which the same physical specimens are scored under each compilation in turn. Third, a quantified measurement of how far two expert compilations of the same primary literature actually agree. Fourth, a hierarchical errors-in-variables model that returns calibrated predictive intervals, and hence a characteristic value, for a beam belonging to an experimental programme the model has not seen.

2. Aim and objectives

The aim is to establish whether the predictive accuracy reported for ML models of the FRP shear contribution survives evaluation on independently compiled data, and to identify the mechanism responsible for any loss. The objectives are:

- to harmonise two independently compiled databases into a single provenance-traced dataset, documenting every unit conversion, placeholder and taxonomy reconciliation required;
- to quantify the agreement between the two compilations on the specimens they share, and thereby separate target transcription noise from covariate transcription noise;
- to evaluate a common set of learners under random, study-grouped, cross-database, pooled and curation-controlled validation regimes;
- to determine whether the degradation observed under transfer is caused by clustering within experimental programmes or by disagreement between compilations; and
- to develop and validate a hierarchical errors-in-variables model that yields calibrated predictive uncertainty for an unseen experimental programme.

3. Databases and harmonisation

Figure 1 sets out the workflow. Two independently compiled databases are harmonised onto a common feature definition, the specimens they share are identified by provenance and removed from the transfer experiment, and the merged dataset is then used to evaluate learners under five validation regimes and to fit the hierarchical errors-in-variables model. Each stage is described in turn in Sections 3 to 6.

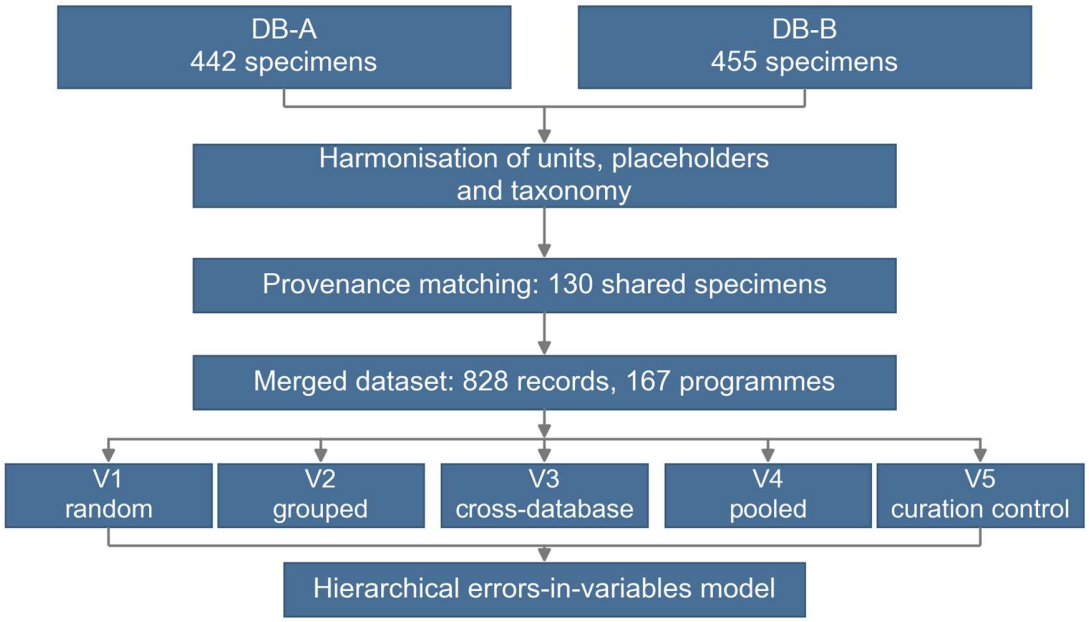


Figure 1: Workflow of the study, from the two independently compiled source databases through harmonisation and provenance-based identification of shared specimens to the five validation regimes and the hierarchical errors-in-variables model

3.1. Sources

Two databases were used. DB-A is the compilation published with Wang et al. (2023), supplied as a raw sheet of 442 specimens together with a filtered subset of 315 specimens obtained by the original authors through unsupervised anomaly detection. The raw sheet was adopted here, because a study concerned with compilation practice should not inherit another group's outlier decisions. DB-B is Table S1 of the Supplementary Materials of Zhao et al. (2025), containing 455 specimens. The two were assembled independently, from partly disjoint primary literature: DB-A cites 108 references, of which nine are duplicate citations of the same study under two numbers, giving 99 distinct source studies and spanning 1992–2019; DB-B cites 59 source studies and spans 2002–2024, extending to large-rupture-strain, basalt, natural-fibre and polyethylene terephthalate systems that DB-A does not label. Table 1 summarises the two sources.

Table 1: Specification of the two independently compiled databases as supplied

Attribute	DB-A (Wang et al., 2023)	DB-B (Zhao et al., 2025)
Records as supplied [-]	442	455
Distinct source studies [-]	99 (108 citations, 9 duplicated)	59
Publication span of sources [year]	1992–2019	2002–2024
Target quantity	V_f	V_u and V_f
Elastic modulus unit	[MPa] (raw sheet)	[GPa]
FRP ratio unit	decimal fraction [-] (raw sheet)	[%]
Concrete strength	f_c	f_{ck}
Configuration coding	numeric code, undocumented	SB / U / FW
Missing-value encoding	blank	the string "/"
Fields unique to the source	h, flange geometry, failure mode, h_{fe}	longitudinal steel, FRP material type

3.2. Reconciliation protocol

Merging the two sources required the reconciliations listed in Table 2, each of which would silently corrupt a naive merge. Three deserve comment. First, both databases encode a continuously bonded sheet by setting the strip width and the strip spacing to unity. This is a placeholder rather than a one-millimetre strip, and it affects 164 of 442 records in DB-A and 153 of 455 in DB-B, so a model that consumes the width-to-spacing ratio literally is corrupted on approximately one third of both databases. The ratio was therefore recovered from the FRP ratio, which both databases define self-consistently as $\rho f = 2\ tf\ (wf/sf)/bw$. In DB-A the recovery is exact to machine precision, confirming the definition; in DB-B it departs from the tabulated columns by up to 0.17, and 15 DB-B records imply a width-to-spacing ratio exceeding unity, which corresponds to overlapping strips and is physically impossible. Second, DB-B encodes missing values as a solidus character rather than as a blank, which coerces the affected columns to text on loading and silently disables numerical handling. Third, the configuration taxonomy had to be reconciled across an undocumented numeric code, which is addressed in Section 3.3.

Table 2: Reconciliations applied when merging the two databases, with the number of records affected

Issue	Treatment	Records affected [-]
E_f in [MPa] (DB-A) versus [GPa] (DB-B)	converted to [GPa]	442
ρ_f as fraction [-] (DB-A) versus per cent [%] (DB-B)	converted to [%]	442
Continuous-sheet placeholder $w_f = s_f = 1$	ratio recovered from ρ_f; binary indicator added	164 (A) + 153 (B)
$w_f/s_f > 1$ (overlapping strips, impossible)	retained and flagged, not altered	15 (B)
Missing values encoded as "/"	coerced to missing	57 in V_f, 192 in f_{yv} (B)
Undocumented configuration code (DB-A)	recovered empirically (Section 3.3)	442
f_{yv} absent where no stirrups are present	set to zero where $\rho_{sv} = 0$	—
Records with no reported target	excluded	57
Exact numeric duplicates within a database	excluded	12

The merged dataset contains 828 records (432 from DB-A and 396 from DB-B) drawn from 167 experimental programmes. The harmonised feature set is the intersection of the two sources: web width, effective depth, shear span-to-depth ratio, concrete strength, transverse steel ratio and yield strength, FRP modulus, tensile strength, thickness, width-to-spacing ratio, FRP ratio, fibre orientation and strengthening configuration. Fields recorded by only one source — flange geometry, failure mode and effective FRP height in DB-A, longitudinal steel and FRP material type in DB-B — cannot enter a cross-database model and are excluded.

3.3. Empirical recovery of the configuration taxonomy

DB-A encodes the strengthening configuration as a three-level numeric code whose meaning is not documented in the supplied file, whereas DB-B states the configuration as side-bonded (SB), U-jacketed (U) or fully wrapped (FW). The mapping was recovered empirically from the specimens common to both databases, for which DB-B supplies the configuration explicitly. The cross-tabulation assigns code 1 to U-jacketing, code 2 to side bonding and code 3 to full wrapping, with majority purities of 98%, 89% and 88% respectively. The recovered mapping is corroborated independently by the fitted model of Section 6, in which full wrapping carries the largest positive coefficient and side bonding a negative one, reproducing the ordering of effectiveness established experimentally for these three configurations.

3.4. Distributional comparison

Figure 2 compares the marginal distributions of the harmonised variables. The two compilations differ significantly on six of the eight quantities shown, with two-sample Kolmogorov–Smirnov tests rejecting equality at the one per cent level for web width, effective depth, concrete strength, FRP modulus, FRP ratio and FRP thickness. They do not differ significantly in shear span-to-depth ratio ($D = 0.067$, $p = 0.30$) or, notably, in the target itself ($D = 0.087$, $p = 0.082$). The transfer task posed in Section 5 is therefore a genuine covariate-shift problem rather than a difference in the response being modelled.

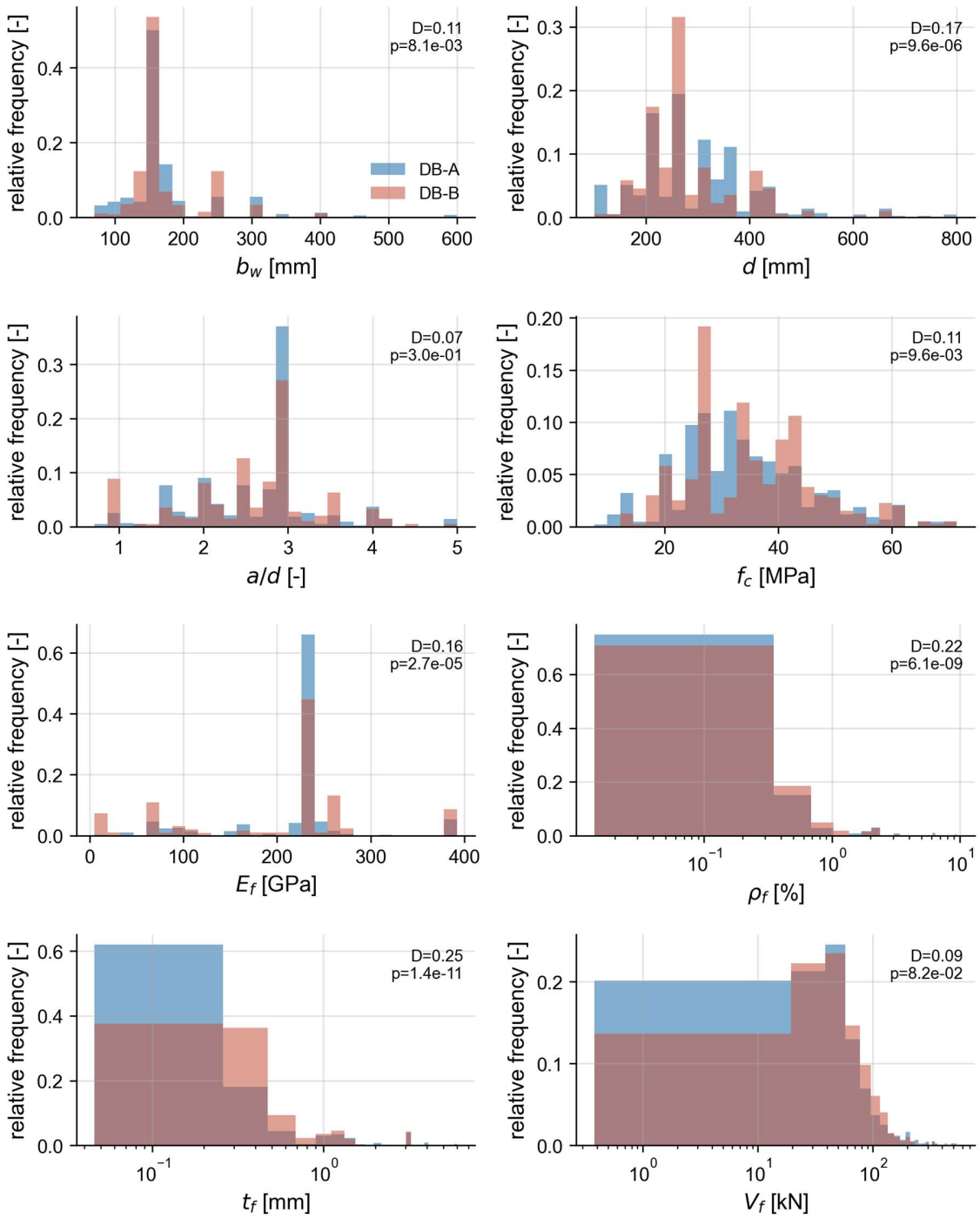


Figure 2: Marginal distributions of the harmonised variables in the two compilations, with two-sample Kolmogorov–Smirnov statistics and p-values; the target and the shear span-to-depth ratio are the only quantities not differing significantly

4. How far do two independent compilations agree?

Because both databases draw on the same primary literature, some physical specimens appear in both. Identifying them is a prerequisite for an unbiased transfer experiment, and the agreement between the two transcriptions is informative in its own right.

Specimen labels alone are not a valid matching key. Matching on labels across the whole of both databases produces pairs that are demonstrably different beams sharing a generic designation: the label "S4", for example, occurs in DB-A with a web width of 600 mm and in DB-B with a web width of 120 mm. Matching on measured values is also unsuitable, because any quantity used as a key is forced into agreement and the resulting statistics are circular. Matching was therefore performed on provenance. The two bibliographies were reconciled on first author and year, yielding 35 source studies present in both databases; specimen labels were then matched only within an already-matched pair of studies. No measured quantity enters the key. This procedure identifies 130 specimens common to both compilations.

Table 3 reports the agreement. FRP thickness, web width and FRP tensile strength are transcribed almost identically. Agreement deteriorates for quantities that require interpretation of the source paper: the two teams record different concrete strengths for 23% of shared specimens, differing by up to 20% at the ninetieth percentile — partly a genuine ambiguity, since DB-A records a cylinder strength and DB-B a characteristic strength. Transverse steel ratio and effective depth also disagree materially. Specimen identity is established by provenance, so that no measured quantity is used as a matching key.

Table 3: Agreement between the two compilations on the 130 specimens common to both

Quantity	n [-]	Identical [%]	Median rel. diff. [%]	90th pct rel. diff. [%]	log-scale SD [-]
t_f [mm]	130	100.0	0.00	0.00	0.000
b_w [mm]	130	96.9	0.00	0.00	0.001
f_{fu} [MPa]	130	96.9	0.00	0.00	0.000
E_f [GPa]	130	89.2	0.00	2.13	0.241
V_f (target) [kN]	128	84.4	0.00	0.60	0.121
d [mm]	130	80.8	0.00	3.07	0.016
f_c vs f_{ck} [MPa]	130	76.9	0.00	20.49	0.093
a/d [-]	130	73.1	0.00	4.29	0.069
ρ_{sv} [%]	52	63.5	0.00	15.72	0.053
ρ_f [%]	130	58.5	0.00	0.01	0.059

The target itself is reproduced in Figure 3. Overall the two compilations agree on the FRP contribution with a coefficient of determination of 0.57 to 0.77, the range reflecting which compilation is taken as the reference, since the statistic is not symmetric. That figure is dominated by a minority of records. Among the 109 pairs on which the two databases also agree on the specimen's identifying geometry, agreement on the target is near-perfect (R^2 = 0.997). The 19 pairs on which the compilations disagree about the beam's own dimensions are a different matter, and include discrepancies in the recorded contribution of up to 63%. No attempt is made here to adjudicate which compilation is correct in those cases; the relevant point is that approximately 15% of shared records are transcribed inconsistently enough by expert compilers to change the recorded capacity materially. This sets a ceiling on what any cross-database validation can achieve and supplies the measurement-error magnitudes used in Section 6.

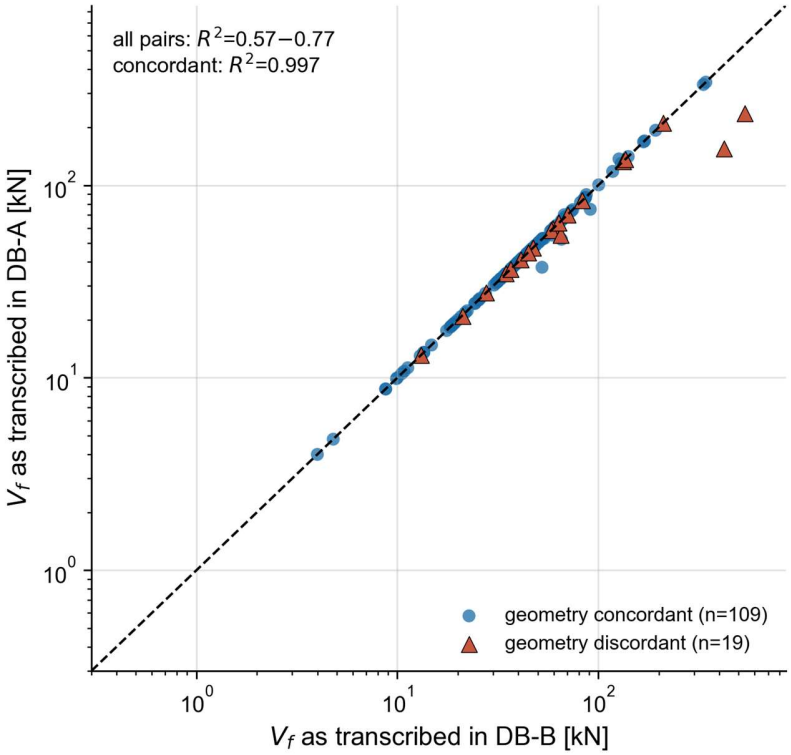


Figure 3: The FRP shear contribution as transcribed in DB-A against the same physical specimen as transcribed in DB-B, separated by whether the two sources also agree on the identifying geometry; the broken line is equality

5. Validation protocols, learners and metrics

5.1. The five regimes

Six learners were evaluated under five regimes of increasing severity, holding the feature set and the tuning procedure fixed so that differences between regimes reflect the protocol rather than the algorithm.

- V1, random five-fold cross-validation within each database. This reproduces prevailing practice and establishes the optimistic baseline against which published figures should be read.
- V2, study-grouped cross-validation within each database, in which all specimens from a given experimental programme fall on the same side of every split. This is the established remedy for dependence structure in cross-validation (Roberts et al., 2017), and it isolates the effect of clustering while holding the compilation constant.
- V3, cross-database transfer. A model is trained on one database and evaluated on the other, after removing every specimen identified in Section 4 as common to both. Retaining shared specimens would leak the test set into training.
- V4, pooled training on the merged dataset with study-grouped validation. This is the deliverable predictive model and tests whether pooling recovers the loss.
- V5, the curation control. A model trained on DB-A excluding shared specimens is evaluated on the shared specimens twice: once with the values as transcribed by DB-A, and once with the values for the same physical beams as transcribed by DB-B. Because the specimens are identical, any difference isolates curation from population shift.

5.2. Learners and tuning

The roster was chosen to span inductive biases while remaining conventional, since the contribution concerns the protocol rather than the algorithm. It comprises ridge regression, support vector regression with a radial basis kernel, a multilayer perceptron with two hidden layers, a random forest (Breiman, 2001), gradient boosting (Chen & Guestrin, 2016), and gradient boosting with monotonicity constraints imposed on web width, effective depth, FRP thickness, FRP ratio and FRP modulus. The constrained variant follows the physically motivated approach of Benzaamia et al. (2024); the transverse steel ratio is deliberately left unconstrained, because the interaction between internal stirrups and the FRP contribution is documented as negative and imposing an increasing constraint would encode incorrect mechanics. Gradient boosting and the neural network are included because they are the learners on which the two source studies based their conclusions (Wang et al., 2023; Zhao et al., 2025). Ridge regression is included as a regularised linear floor, a comparison this literature almost never reports.

The response is modelled as the natural logarithm of the FRP contribution. Hyperparameters were selected by grid search using study-grouped cross-validation conducted strictly inside each training partition, so that model selection cannot bias the reported error (Varma & Simon, 2006); no preprocessing statistic, imputation rule or hyperparameter was ever estimated on data used for evaluation. In V3 the tuning loop never sees the target database in any form. Missing transverse yield strengths were set to zero where the transverse steel ratio is zero, which is physically correct rather than an imputation, and median-imputed otherwise from the training partition.

5.3. Metrics

Performance is reported as the coefficient of determination on the untransformed contribution, together with the mean and coefficient of variation of the ratio of predicted to measured capacity and the proportion of unconservative predictions, which are the quantities relevant to design and which the coefficient of determination conceals. Ratio statistics are computed only on specimens with a measured contribution of at least 5 kN: the merged dataset contains recorded contributions as small as 0.38 kN, for which the ratio is numerically unstable and would dominate any dispersion measure. This screen is stated rather than applied silently.

6. A hierarchical errors-in-variables model

The regimes of Section 5 diagnose the problem but do not repair it. They establish that a model fitted to compiled test data cannot be evaluated as though the specimens were independent, and Section 4 establishes that the covariates carry a measurable transcription error; neither conclusion changes the model being fitted. A learner trained under grouped validation is still a learner that treats every specimen as an independent draw with exactly known inputs, so it can be scored honestly but cannot express what it does not know. The remainder of this paper therefore turns from diagnosis to construction, and asks what model form the two diagnosed properties actually imply.

Both properties have a standard statistical remedy. Clustering within experimental programmes is a random-effects structure; covariates observed with error are a latent-variable structure. Combining them gives a hierarchical errors-in-variables model, which is the statistically appropriate response to the two findings rather than an additional algorithm added for comparison. Its practical value is that it returns a predictive distribution for a beam belonging to a programme it has never seen, and hence the characteristic value that a design provision requires and that no point-prediction model in this literature can supply.

Let i index specimens and s(i) the source experimental programme. Writing z for the latent true values of the log-transformed covariates and e for the remaining covariates, the model is

$$\ln V_{f,i} = b0 + z_i'\beta + e_i'\gamma + u_s(i) + \varepsilon_i, \quad u_s \sim N(0, \tau^2), \quad \varepsilon_i \sim N(0, \sigma^2) \tag{1}$$

Where:
$ln\ V_{f,i}$ – natural logarithm of the FRP shear contribution of specimen i [ln kN],
b_0 – intercept of the model [-],
z_i – vector of latent true values of the log-transformed covariates of specimen i [-],
β – vector of coefficients on the latent log-covariates [-],
e_i – vector of the remaining, directly observed covariates of specimen i [-],
γ – vector of coefficients on those covariates [-],
$u_{s(i)}$ – random effect of the source experimental programme s(i) [-],
ε_i – within-programme residual of specimen i [-],
τ – standard deviation between experimental programmes [-],
σ – standard deviation within an experimental programme [-],
w_i – vector of covariate values as recorded in the database [-],
m_i – transcription error of specimen i [-],
S_m – covariance matrix of the transcription error, fixed to the values measured in Section 4 [-].

The observed covariate record w_i is related to the latent truth by $w_i = z_i + m_i$. The essential feature of the formulation is that S_m and the target error variance are not free parameters and are not assumed: they are fixed to the transcription discrepancies measured in Section 4 on the 130 specimens common to both databases, given as log-scale standard deviations in the final column of Table 3. The model therefore carries an empirically measured estimate of how accurately the literature has been transcribed, rather than a nominal one.

6.1. Assumptions, measurement-error variances and priors

Three modelling choices carry the weight of this section and are stated here with their justification. The first is the log transformation. The FRP contribution spans nearly three orders of magnitude in the merged dataset, its dispersion grows with its level, and the mechanical quantities that govern it enter multiplicatively. A model additive in logarithms is therefore the natural functional form, and it also makes the random-effect and measurement-error structures additive on the same scale.

The second is the treatment of the measurement-error variances. These are fixed rather than estimated, because a model in which both the latent covariates and their error variances are free is only weakly identified from a single observation per specimen. Fixing them is possible here only because the shared specimens of Section 4 provide a direct estimate: the log-scale standard deviations in the final column of Table 3 are the dispersion of the discrepancy between two independent transcriptions of the same physical test. This is an empirical estimate of transcription error, and it is deliberately conservative in two respects. It is available only for covariates that both databases record, so the remaining covariates are treated as exact, which understates rather than overstates the correction. It also measures disagreement between compilers rather than error in the original laboratory measurement, so the true covariate error is at least this large.

The third is the prior specification. Coefficients are standardised before fitting, so a standard normal prior on each is weakly informative by construction. It places most of its mass on effects smaller than one standard deviation of the response per standard deviation of the covariate, which is generous for a mechanical relationship, while still regularising the poorly identified directions. The intercept receives a deliberately diffuse N(0, 25) prior because its scale is set by the units of the response. Inverse-gamma(2, 0.5) priors on the variance components are proper, place negligible mass near zero, and have a mean of the same order as the observed dispersion of the data. Section 7.5 reports a companion fit without the measurement-error layer, and the closeness of the two results is itself evidence that the conclusions are not being driven by these choices.

Every full conditional of this model is Gaussian or inverse-gamma, so the posterior was sampled with a purpose-written Gibbs sampler; the conditional distributions, the sampling scheme and the prior specification are given in Appendix B. Four chains of 4,000 retained draws after 1,500 burn-in iterations were run, giving 16,000 draws for each of the 13 monitored parameters. Convergence was assessed with the rank-normalised split statistic and the bulk and tail effective

sample sizes of Vehtari et al. (2021). The largest value of the statistic over all monitored parameters was 1.0025, the smallest bulk effective sample size was 1977 and the smallest tail effective sample size 3397, and the Monte Carlo standard error of every posterior mean was at most 2.2 per cent of the corresponding posterior standard deviation. These are comfortably inside the usual thresholds of 1.01 and 400 respectively. Per-parameter diagnostics, together with trace and rank plots, are given in Appendix B. A companion model without the measurement-error layer was fitted identically for comparison.

7. Results and discussion

7.1. The cost of the splitting protocol

Table 4 and Figure 4 report the coefficient of determination for every learner under every regime. The pattern is consistent and large. Taking gradient boosting as the reference learner, random splitting yields 0.672 on DB-A and 0.694 on DB-B, values comparable with those reported in the literature. Merely preventing specimens from the same experimental programme from straddling the split, changing nothing else, reduces these to 0.391 and 0.216. Cross-database transfer gives 0.213 from DB-A to DB-B and 0.211 in the reverse direction. Roughly two thirds of the apparent predictive accuracy of these models is therefore an artefact of the validation protocol rather than a property of the models themselves. V1 reproduces prevailing practice, V2 removes sibling leakage, and V3 is transfer between independently compiled databases with the shared specimens removed from training.

Table 4: Coefficient of determination on the FRP shear contribution for each learner under each validation regime

Regime [-]	Subset [-]	Ridge [-]	SVR [-]	MLP [-]	RF [-]	XGBoost [-]	XGB-mono [-]
V1	DB-A	0.540	0.549	0.558	0.591	0.672	0.665
V1	DB-B	0.418	0.560	0.443	0.474	0.694	0.613
V2	DB-A	0.433	0.270	0.511	0.451	0.391	0.379
V2	DB-B	0.073	0.230	0.067	0.080	0.216	0.228
V3	A->B	0.021	0.231	0.100	0.208	0.213	0.279
V3	B->A	-95.726	0.173	-0.122	0.329	0.211	0.108
V4	merged	0.411	0.377	0.446	0.457	0.449	0.465
V5	curation A	0.013	0.398	0.370	0.316	0.348	0.454
V5	curation B	0.223	0.273	0.302	0.321	0.372	0.417

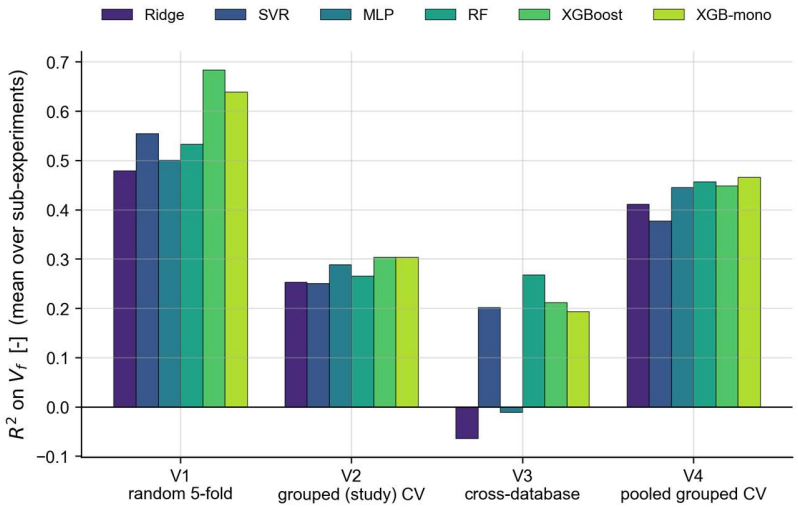


Figure 4: Coefficient of determination by learner and validation regime, averaged over the sub-experiments of each regime and truncated below at -0.15 for legibility

The degradation is not confined to one model family. Every learner loses accuracy in the same order, which rules out an explanation based on the inductive bias of a particular algorithm. Two learners behave anomalously and instructively: ridge regression and the multilayer perceptron both return negative coefficients of determination in the DB-B to DB-A direction. Both are unbounded extrapolators, and both fail when asked to predict outside the region occupied by their training data. Section 7.4 examines this directly. The two scales in Table 4 also guard against a metric artefact: original-scale values are weighted towards the largest specimens, and on the log scale the gradient-boosted models fall below zero under V2 on DB-B, so the collapse there is deeper than the original-scale figure alone conveys.

7.2. Interpreting the negative and small coefficients of determination

A negative coefficient of determination means only that a model predicts the evaluation set less well than that set's own mean, which is an ordinary outcome when a model is scored on a population different from the one it was fitted to. Two entries in Table 4 are large enough in magnitude to need comment, because they are as much a property of the metric as of the model. The response is modelled in logarithms and returned to the original scale by exponentiation, so a large error in log space becomes an enormous error in kilonewtons and the sum of squares is then dominated by a few specimens. For ridge regression transferred from DB-B to DB-A the value is -95.7 on the original scale but -0.223 on the log scale, and the median ratio of predicted to measured capacity is 1.07; two specimens carry 88 per cent of the total squared error. The failure is nevertheless real rather than cosmetic, since 13 of 262 ridge predictions exceed ten times the measured value, which is the unbounded extrapolation quantified in Section 7.4. The small positive values elsewhere in the table carry the study's central message: once sibling specimens are separated, the explainable share of specimen-to-specimen variation is modest, and reporting it honestly is the point.

The ratio statistics in Table 5 show that the loss is not merely a loss of correlation. Under random splitting the mean ratio of predicted to measured capacity for gradient boosting is 1.136 on DB-A with a coefficient of variation of 0.912; under cross-database transfer from DB-B to DB-A the mean ratio rises to 1.396 and the proportion of unconservative predictions reaches 51.4 per cent. A model transferred to an independently compiled population is not simply less precise; it becomes systematically unconservative, which is the failure mode that matters for design. The ratio statistics of Table 5 are computed on specimens with a measured contribution of at least 5 kN.

Table 5: Design-relevant statistics for the gradient boosting model by regime

Regime [-]	Subset [-]	R^2 [-]	Mean ratio [-]	CoV [-]	Unconservative [%]	n [-]
V1	DB-A	0.672	1.136	0.912	43.9	432
V1	DB-B	0.694	1.108	0.623	45.6	396
V2	DB-A	0.391	1.176	0.785	44.6	432
V2	DB-B	0.216	1.236	0.997	41.8	396
V3	A->B	0.213	1.015	0.918	32.9	231
V3	B->A	0.211	1.396	0.787	51.4	262
V4	merged	0.449	1.162	0.770	46.2	828

7.3. Clustering, not curation, is the mechanism

The V5 control isolates the cause. A model trained on DB-A excluding shared specimens was evaluated on the shared specimens as transcribed by DB-A, and then, for the same physical beams, as transcribed by DB-B. For gradient boosting the two evaluations give 0.348 and 0.372; for the random forest, 0.316 and 0.321. Substituting one compilation's transcription for the other's therefore changes performance very little, even though Section 4 established that the two disagree on a meaningful minority of records. The degradation seen in V2 and V3 is consequently attributable to clustering by experimental programme rather than to disagreement between compilers.

This is the most immediately actionable finding of the study, because study-grouped validation costs nothing. Any future paper in this field can adopt it, provided the specimen-to-study mapping is retained, which is precisely the field that compiled databases most often discard. The finding also explains why cross-database transfer is not much worse than grouped validation within a single database: once sibling leakage is removed, a model is already being asked to generalise to unseen programmes, and crossing a database boundary adds comparatively little.

Figure 5 summarises the progression, with the hierarchical model of Section 6 superimposed. It also shows that pooling helps. Training on the merged dataset under grouped validation recovers to 0.449 for gradient boosting and 0.465 for its monotonic variant, the best values obtained under any honest regime. Breadth of provenance, not volume of records, is what buys generalisation, and the monotonicity constraints of Benzaamia et al. (2024) help most exactly where the data are thinnest.

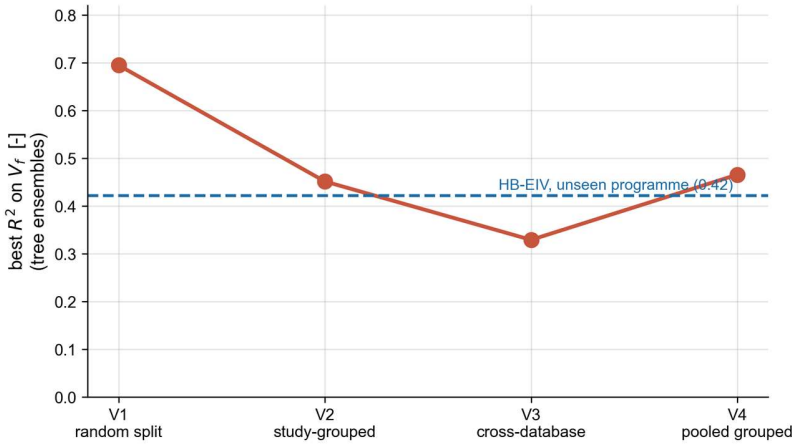


Figure 5: Best coefficient of determination among the tree ensembles under each regime, with the hierarchical errors-in-variables model evaluated on unseen programmes shown for reference

7.4. Extrapolation behaviour

Because the two compilations differ significantly in six of eight covariates (Figure 2), a substantial part of the transfer task lies outside the region spanned by the training data. Of the 231 DB-B specimens used to evaluate transfer from DB-A, 150 lie inside the training range on every covariate and 81 lie outside it on at least one. Figure 6 separates performance accordingly, and the result contradicts a plausible prior expectation. Tree ensembles cannot extrapolate; outside the training hull they return the boundary constant. That inability is usually described as a limitation, but here it acts as a safeguard: gradient boosting retains a positive coefficient of determination of 0.132 outside the hull, and the monotonic variant 0.221. Ridge regression, which extrapolates freely, falls to -0.431 with a mean ratio of 1.56, that is, it overpredicts by more than half on average.

The practical implication is that unbounded learners should not be deployed on beams whose parameters lie outside the range of the training database without an explicit domain check, and that reporting the proportion of a test set lying outside the training hull should accompany any claim of generalisation in this field.

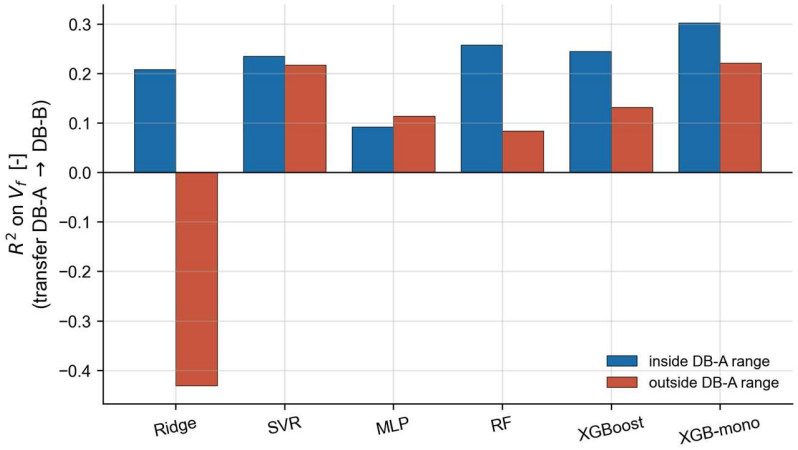


Figure 6: Transfer performance from DB-A to DB-B separated by whether the test specimen lies inside the range of the training database on every covariate; values truncated below at -0.6 for legibility

7.5. Variance decomposition and physical consistency

The hierarchical errors-in-variables model quantifies directly what Sections 7.1 and 7.3 establish indirectly. The between-programme standard deviation of the log contribution is 0.444 against a within-programme value of 0.567, so the intraclass correlation is 0.380 with a 95 per cent credible interval of 0.294 to 0.469. In other words, approximately 38 per cent of the residual variation in the FRP shear contribution, after the mechanical covariates have been accounted for, is associated with the identity of the experimental programme rather than with the mechanics of the specimen. This is the quantity that random splitting implicitly allows a model to memorise, and it explains the magnitude of the drop between V1 and V2 without reference to any particular algorithm.

The posterior coefficients in Figure 7 are physically coherent, which supports the harmonisation as well as the model. The strongest positive effects are the effective depth and the FRP ratio, with standardised coefficients of +0.288 and +0.288; web width, FRP modulus and concrete strength follow at +0.140, +0.133 and +0.124. Three results are worth emphasising. First, the transverse steel ratio carries a negative coefficient (-0.067), reproducing the interaction whereby internal stirrups reduce the contribution mobilised by the external FRP, which is well documented experimentally (Cao et al., 2005; Grande et al., 2009; Mofidi & Chaallal, 2011); this was left unconstrained by design and the model recovers it from the data. Second, full wrapping carries the largest configuration effect (+0.295) and side bonding a negative one (-0.121), independently corroborating the configuration mapping recovered empirically in Section 3.3. Third, the shear span-to-depth ratio is effectively null (+0.004), consistent with its being the one covariate whose distribution does not differ between the two compilations.

The measurement-error layer behaves as theory predicts but with modest magnitude. Correcting for covariate error reduces the within-programme standard deviation from 0.582 to 0.567 and raises the FRP modulus coefficient from +0.125 to +0.133, the direction expected from the attenuation bias that measurement error induces. The correction is small because the measured transcription noise, although non-negligible, is small relative to the spread of the covariates themselves. This is a useful negative result: it shows that the disagreement documented in Section 4, while real, is not what limits predictive performance, and it corroborates the V5 control from an entirely different direction.

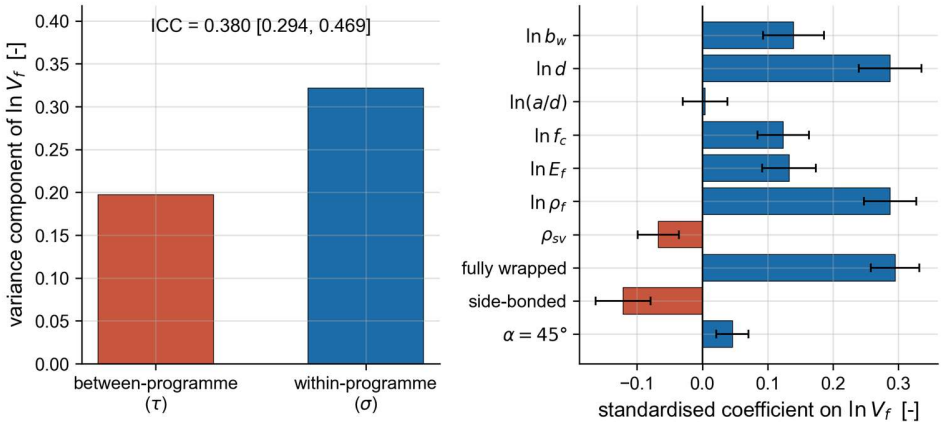


Figure 7: Variance decomposition of the log FRP contribution into between-programme and within-programme components, and standardised posterior coefficients with one standard deviation; the intraclass correlation is given with its 95 per cent credible interval

7.6. Calibrated prediction for an unseen experimental programme

Point accuracy is not the only requirement, and for design it is not the primary one. Evaluated under grouped cross-validation, so that every prediction concerns a programme the model has never seen, the hierarchical model attains a coefficient of determination of 0.422, comparable with the best tree ensemble under the equivalent regime. Its distinctive property is that its uncertainty statements are honest. Empirical coverage of the nominal 90 per cent predictive interval is 91.8 per cent, and of the nominal 95 per cent interval 96.4 per cent (Figure 8). The fifth percentile of the predictive distribution, which is the natural candidate for a characteristic value, is exceeded from below by 6.3 per cent of specimens against a target of 5 per cent.

The width of those intervals is the substantive result. The median 90 per cent predictive interval spans a factor of approximately 11.7 between its bounds. This is the honest predictive uncertainty for the FRP shear contribution of a beam from a laboratory whose systematic offset is unknown, and it is concealed entirely by the point estimates and single-number coefficients of determination that dominate this literature. It follows that the practical output of data-driven modelling in this field should be a calibrated characteristic value accompanied by its interval, not a mean prediction; the present model supplies one, and Figure 9 shows the predictions with their intervals against the measured values.

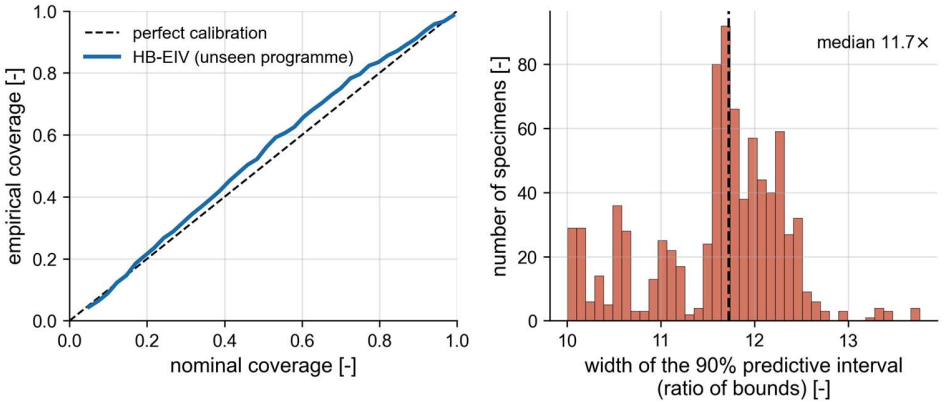


Figure 8: Calibration of the hierarchical errors-in-variables model for unseen experimental programmes: empirical against nominal coverage, and the distribution of the width of the 90 per cent predictive interval expressed as the ratio of its bounds

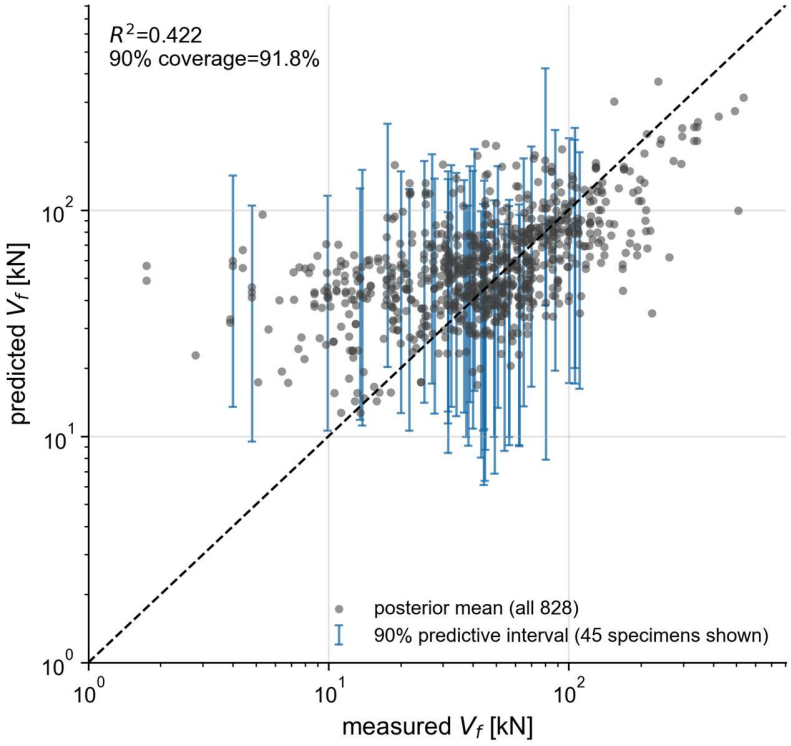


Figure 9: Posterior mean prediction against measured FRP shear contribution for unseen experimental programmes, with the pointwise 90 per cent predictive interval; the broken line is equality

7.7. Comparison with design provisions

The comparison that motivates most of this literature is between data-driven models and design equations, and it is normally made under the random-split protocol that Section 7.1 shows to be optimistic. Because design equations are not fitted to the data, their performance is invariant to the splitting protocol; they therefore provide a fixed reference against which the fitted models can be read as the protocol tightens. ACI 440.2R-17 and the effective-strain formulation of fib Bulletin 14 were evaluated on every specimen of the merged database. Neither provision defines the effective depth of the FRP shear reinforcement in a way that DB-B records, so that depth is taken as the beam effective depth throughout; the convention is applied identically to both compilations, so it cannot favour either side of the transfer experiment.

Table 6 and Figure 10 give the result. At the median, ACI 440.2R-17 is close to unbiased on this population, with a median ratio of predicted to measured contribution of 1.03, but its dispersion is very large (coefficient of variation 1.17) and 52 per cent of its predictions are unconservative. The fib Bulletin 14 formulation is systematically less conservative still, with a median ratio of 1.49 and 72 per cent of predictions exceeding the measured capacity. Both return negative coefficients of determination on every evaluation set, meaning that neither explains the specimen-to-specimen variation better than the mean of the data does — a known limitation of these provisions, which are calibrated for design safety rather than for mean-value prediction. The design values in Table 6 change between regimes only because the evaluation sets differ, not because the provisions are refitted.

Table 6: Coefficient of determination for the best fitted learner and for two unfitted design provisions

Regime [-]	Best fitted learner [-]	ACI 440.2R-17 [-]	fib Bulletin 14 [-]
V1	0.694	-0.601	-0.240
V2	0.451	-0.601	-0.240
V3	0.329	-0.881	-0.547
V4	0.465	-0.678	-0.261

The engineering conclusion is that the advantage of data-driven models over design provisions survives honest validation, but is much smaller than the literature implies. Under random splitting the best fitted learner reaches 0.694 against -0.601 for ACI 440.2R-17. Under study-grouped validation the fitted value falls to 0.451, and under cross-database transfer to 0.329, while the design values are unchanged apart from the change of evaluation set. A reader of the existing literature would infer a margin roughly twice the one that survives when sibling specimens are prevented from straddling the split. The provisions remain materially less accurate than the fitted models, so the case for data-driven prediction is not overturned; what is overturned is the size of the reported margin.

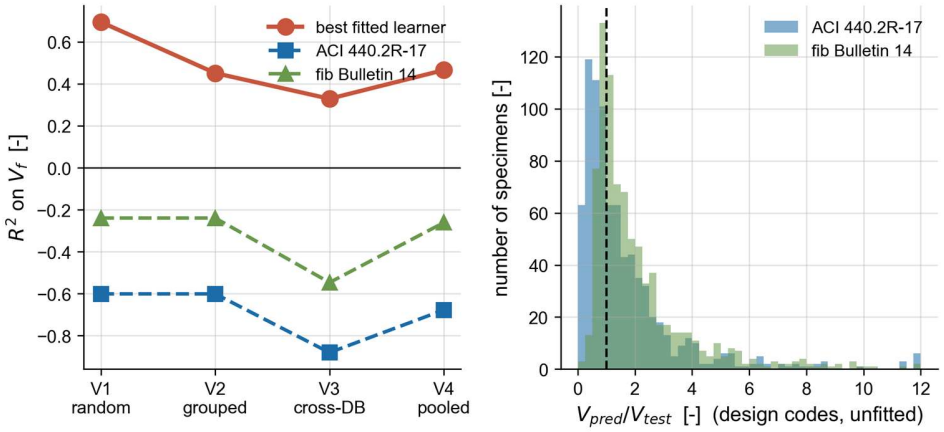


Figure 10: Left: coefficient of determination for the best fitted learner and for two unfitted design provisions as the validation protocol tightens. Right: distribution of the ratio of predicted to measured contribution for the two provisions, truncated at twelve

7.8. Explainability and its stability across compilations

Parameter-importance analysis is now a standard component of this literature, and its conclusions are reported as physical insight. Because Section 7.1 shows that the accuracy of these models is protocol-dependent, it is reasonable to ask whether their explanations are compilation-dependent. Two analyses were run: SHAP values for the pooled model evaluated under study-grouped validation, and a comparison of the importance ordering obtained when the same learner is fitted on each database separately.

Under grouped validation the pooled model attributes most influence to the effective depth, to full wrapping, and to the strip-width-to-spacing ratio, followed by the web width and the FRP ratio (Figure 11, left). The ordering is mechanically coherent. The four leading quantities are those that set the bonded area and the anchorage condition, which is what the FRP contribution is expected to depend upon. The configuration indicator for full wrapping ranks second, consistent with the coefficient recovered independently by the hierarchical model in Section 7.5.

The stability analysis is the more informative one. Fitting the same learner to DB-A and to DB-B separately, and ranking the sixteen inputs by mean absolute SHAP value in each case, gives a Spearman rank correlation of 0.80 and a Kendall correlation of 0.67 between the two orderings (Figure 11, right). The agreement is therefore substantial but not complete, and Table 7 gives the full ordering under each compilation. The two leading parameters are identical, and 4 of the top five are shared, so the headline conclusions an author would draw from either compilation would broadly coincide. The middle of the ranking is far less stable: the largest displacement is 8 places, for the transverse reinforcement yield strength, which ranks sixth on DB-A and fourteenth on DB-B, while the concrete strength and the shear span-to-depth ratio each move by four to five places in the opposite direction.

Table 7: Importance rank of each input when the same gradient boosting model is fitted to each database separately, ordered by rank on DB-A. A positive shift indicates a parameter ranked more highly on DB-B

Input [-]	Rank (DB-A) [-]	Rank (DB-B) [-]	Shift [-]
d	1	1	+0
fully wrapped	2	2	+0
ρ_f	3	4	-1
b_w	4	7	-3
w_f/s_f	5	5	+0
f_{yv}	6	14	-8
a/d	7	3	+4
ρ_{sv}	8	8	+0
t_f	9	9	+0
f_{fu}	10	11	-1
f_c	11	6	+5
E_f	12	12	+0
α	13	13	+0
side-bonded	14	10	+4
U-jacket	15	16	-1
continuous sheet	16	15	+1

The practical implication is that importance rankings from a single compiled database should be reported with the same caution as accuracy figures from a single random split. Statements about the leading two or three parameters are reproducible across independent compilations. Statements about the relative importance of mid-ranked quantities are not: a study using DB-A alone would report the transverse steel yield strength as the sixth most influential input, while a study using DB-B alone would place it fourteenth. Interpretability claims in this field should therefore be accompanied by evidence that the ordering is stable under resampling of the source literature, not merely under resampling of specimens.

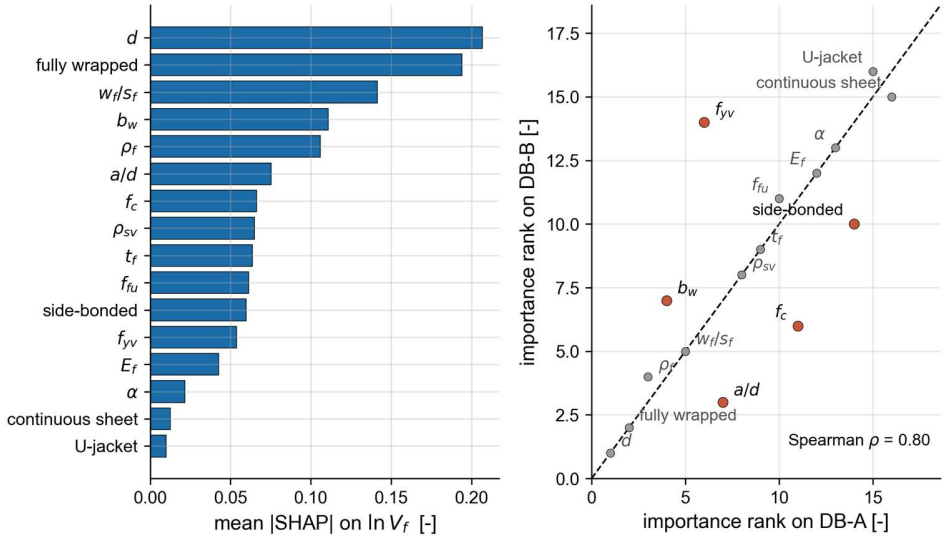


Figure 11: Left: mean absolute SHAP value for the pooled model under study-grouped validation. Right: importance rank of each input when the same learner is fitted to each database separately, rank 1 being the most influential; points displaced by three or more places are highlighted and the broken line is equality

7.9. Implications and recommended practice

The recommendations that follow from Sections 7.1 to 7.8 are collected in Table 8 rather than restated here. Two of them deserve emphasis because they cost nothing. Publishing the specimen-to-study mapping, and grouping validation on it, accounts for most of the difference between optimistic and honest performance. Documenting unit conventions, placeholder encodings and missing-value codes prevents the silent corruption that the reconciliations of Table 2 would otherwise cause.

A broader implication concerns how this literature should be read. The comparison between ML models and design equations, which is the standard concluding argument of these papers, is made under a protocol that flatters the ML model and leaves the design equation unaffected, since the latter is not fitted and therefore cannot benefit from leakage. Section 7.7 quantifies the consequence: the advantage survives, but at roughly half the apparent size.

Table 8 sets out a reporting protocol that would make results in this field comparable and auditable. None of its items requires additional experimental work; all are properties of how an analysis is conducted and reported.

Table 8: Proposed reporting protocol for data-driven studies of FRP shear strengthening. Each item addresses a failure mode demonstrated in this paper

Requirement	Purpose
Publish the specimen-to-study mapping	Enables grouped validation and re-analysis by others
Report study-grouped CV alongside any random-split figure	Removes leakage between sibling specimens of one programme
State the proportion of the evaluation set outside the training range	Model behaviour differs qualitatively across that boundary (Section 7.4)
Report predictive intervals and their empirical coverage	A point estimate conceals the uncertainty relevant to design
Report a characteristic (5th percentile) value	The quantity a design provision requires, unobtainable from a mean prediction
Benchmark against at least one unfitted design provision	Provides a protocol-invariant reference (Section 7.7)
Document units, placeholder codes and missing-value encodings	Each was capable of corrupting the merge silently (Table 2)
Demonstrate that importance rankings are stable across source literature	Mid-ranked parameters are compilation-dependent (Section 7.8)

The experiments reported here concern one strengthening problem, and the quantitative results should not be transferred to others without repeating the analysis. The mechanism, however, is not specific to FRP shear strengthening. It requires only that the training data be pooled from published laboratory programmes, each contributing several specimens that share a batch, a protocol and a laboratory. Several other structural problems are modelled from databases assembled in that way, among them the shear strength of deep beams, punching of slabs, confined-concrete strength, bond of ribbed bars and the behaviour of corroded members. Where that condition holds, the effective number of independent units is closer to the number of programmes than to the number of specimens, and a random split would be expected to overstate generalisation. How large that effect is in any particular field is an empirical question that this study does not answer, and the protocol in Table 8 is offered as a means of measuring it rather than as evidence that it is equally severe elsewhere.

8. Limitations

Four limitations qualify these results. First, the two databases, although compiled independently, draw on a partly common body of published experiments. Reconciling their bibliographies identifies 35 source programmes present in both, against a combined total of roughly 123 distinct programmes, so about seventy per cent of the source literature is unique to one compilation or the other. The transfer experiment is therefore substantially, but not completely, external. Validation against a database assembled from experiments unseen by either compilation would be stronger still, and is the natural next step. Second, the design provisions of Section 7.7 require an effective depth of FRP shear reinforcement that DB-B does not record, so the beam effective depth was substituted throughout. That convention is applied identically to both compilations and therefore cannot bias the comparison between them. It does mean, however, that the absolute

accuracy reported for those provisions is approximate, and it should not be read as a validation of the provisions themselves. Third, the measurement-error magnitudes are available only for covariates that both databases record. The remaining covariates are treated as exact, which is conservative in the sense that it understates rather than overstates the correction. Finally, the specimen-to-study mapping in DB-A is reconstructed by forward-filling a sparsely populated reference column. That procedure is faithful to the structure of the file, but it cannot be verified independently for every block.

9. Conclusions

- Roughly two thirds of the predictive accuracy reported for machine learning models of the FRP shear contribution is an artefact of random splitting. For gradient boosting the coefficient of determination falls from 0.67 and 0.69 under random splitting to 0.39 and 0.22 under study-grouped validation, and to 0.21 under cross-database transfer.
- The cause is clustering within experimental programmes, not disagreement between compilers. Scoring the same physical specimens under each compilation in turn changes performance negligibly (0.35 against 0.37), whereas grouping on the source programme reproduces almost the whole of the transfer penalty.
- A hierarchical model attributes 38 per cent of the residual variance of the log contribution to the identity of the experimental programme, with a 95 per cent credible interval of 29 to 47 per cent. This is the quantity that random splitting allows a model to memorise.
- Two independent compilations of the same primary literature agree on the FRP contribution with a coefficient of determination of 0.57 to 0.77 across the 130 specimens they share, rising to 0.997 on the subset where they also agree on the specimen geometry. Approximately 15 per cent of shared records are transcribed inconsistently enough to change the recorded capacity materially.
- Prediction for an unseen experimental programme can be made honestly but not precisely. The hierarchical errors-in-variables model achieves empirical coverage of 91.8 per cent at a nominal 90 per cent, and its median predictive interval spans a factor of approximately 11.7 between bounds, which is the uncertainty that point predictions in this field conceal.
- Databases should publish the specimen-to-study mapping, validation should be grouped on it, and the proportion of the evaluation set lying outside the training range should be reported alongside any claim of generalisation.

Acknowledgements

The authors gratefully acknowledge the research groups whose experimental programmes constitute the databases analysed here, and the authors of the two source compilations for releasing their datasets as supplementary material, without which a cross-database assessment would not have been possible.

Author Contributions

S.E.R. conceived and designed the study, harmonised the two databases, developed the validation regimes and the hierarchical errors-in-variables model, and prepared the first draft of the manuscript. **H.A.Z.** contributed to the reconciliation of the source bibliographies, the verification of the merged dataset and the interpretation of the results. Both authors critically reviewed and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Disclosure of Interest

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

Data Availability Statement

The harmonised and provenance-traced merged dataset, the reconciliation script, the validation code and the Gibbs sampler are provided as supplementary material. The two source databases are the supplementary files of Wang et al. (2023) and Zhao et al. (2025).

Abbreviations

Abbreviation	Definition
CoV	Coefficient of variation
CV	Cross-validation
DB-A	Database of Wang et al. (2023)
DB-B	Database of Zhao et al. (2025)
EIV	Errors-in-variables
FRP	Fibre-reinforced polymer
FW	Fully wrapped
HB-EIV	Hierarchical errors-in-variables model
ICC	Intraclass correlation coefficient
KS	Kolmogorov-Smirnov
MLP	Multilayer perceptron
ML	Machine learning
RC	Reinforced concrete
RF	Random forest
SB	Side-bonded
SVR	Support vector regression
U	U-jacketed
XGB	Extreme gradient boosting

References

Abdel-Jaber, M. S. (2003). Shear strengthening of reinforced concrete beams using different configurations of externally bonded carbon fibre reinforced plates. *Materials and Structures, 36*(259), 291–301. https://doi.org/10.1617/13845

Abdul Samad, A. A., Ali, N., Mohamad, N., Jayaprakash, J., Tee, K. F., & Mendis, P. (2017). Shear strengthening and shear repair of 2-span continuous RC beams with CFRP strips. *Journal of Composites for Construction, 21*(3), Article 04016099. https://doi.org/10.1061/(asce)cc.1943-5614.0000756

ACI Committee 440. (2017). *Guide for the design and construction of externally bonded FRP systems for strengthening concrete structures* (ACI 440.2R-17). American Concrete Institute.

Adhikary, B. B., & Mutsuyoshi, H. (2004). Behavior of concrete beams strengthened in shear with carbon-fiber sheets. *Journal of Composites for Construction, 8*(3), 258–264. https://doi.org/10.1061/(asce)1090-0268(2004)8:3(258)

Adhikary, B. B., Mutsuyoshi, H., & Ashraf, M. (2003). Effective shear strengthening of concrete beams using FRP sheets with bonded anchorage. *Fibre-Reinforced Polymer Reinforcement for Concrete Structures*, 457–466. https://doi.org/10.1142/9789812704863_0042

Akkaya, H. C., Aydemir, C., & Arslan, G. (2022a). Investigation on shear behavior of reinforced concrete deep beams without shear reinforcement strengthened with fiber reinforced polymers. *Case Studies in Construction Materials, 17*, Article e01392. https://doi.org/10.1016/j.cscm.2022.e01392

Akkaya, H. C., Aydemir, C., & Arslan, G. (2022b). An experimental research on reinforced concrete deep beams fully wrapped with fiber reinforced polymers against shear. *Case Studies in Construction Materials, 17*, Article e01198. https://doi.org/10.1016/j.cscm.2022.e01198

Akkaya, H. C., Aydemir, C., & Arslan, G. (2024). Evaluation of shear behavior of short-span reinforced concrete deep beams strengthened with fiber reinforced polymer strips. *Engineering Structures, 299*, Article 117145. https://doi.org/10.1016/j.engstruct.2023.117145

Alam, M. A., & Al Riyami, K. (2018). Shear strengthening of reinforced concrete beam using natural fibre reinforced polymer laminates. *Construction and Building Materials, 162*, 683–696. https://doi.org/10.1016/j.conbuildmat.2017.12.011

Al-Lebban, Y. F., & Mackie, K. R. (2019). Polyurethane-FRP external strengthening of RC beams with no steel stirrups. *Journal of Composites for Construction, 23*(1), Article 04018074. https://doi.org/10.1061/(asce)cc.1943-5614.0000916

Al-Rousan, R. Z., Haddad, R. H., & Swesi, A. O. (2015). Repair of shear-deficient normal weight concrete beams damaged by thermal shock using advanced composite materials. *Composites Part B: Engineering, 70*, 20–34. https://doi.org/10.1016/j.compositesb.2014.10.032

Al-Rousan, R. Z., & Issa, M. A. (2016). The effect of beam depth on the shear behavior of reinforced concrete beams externally strengthened with carbon fiber–reinforced polymer composites. *Advances in Structural Engineering, 19*(11), 1769–1779. https://doi.org/10.1177/1369433216649386

Altin, S., Anil, Ö., Kopraman, Y., Mertoğlu, Ç., & Kara, M. E. (2010). Improving shear capacity and ductility of shear-deficient RC beams using CFRP strips. *Journal of Reinforced Plastics and Composites, 29*(19), 2975–2991. https://doi.org/10.1177/0731684410363182

Bae, S. W., Tann, B. D., & Belarbi, A. (2012). Size effect of reinforced concrete beams strengthened in shear with externally bonded CFRP sheets. *Proceedings of the 6th International Conference on FRP Composites in Civil Engineering (CICE 2012)*, Rome, Italy.

Barros, J., Dias, S., & Lima, J. (2007). Efficacy of CFRP-based techniques for the flexural and shear strengthening of concrete beams. *Cement and Concrete Composites, 29*(3), 203–217. https://doi.org/10.1016/j.cemconcomp.2006.09.001

Benzaamia, A., Ghrici, M., Rebouh, R., Zygouris, N., & Asteris, P. G. (2024). Predicting the shear strength of rectangular RC beams strengthened with externally-bonded FRP composites using constrained monotonic neural networks. *Engineering Structures, 313*, Article 118192. https://doi.org/10.1016/j.engstruct.2024.118192

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Bukhari, I. A., Vollum, R., Ahmad, S., & Sagaseta, J. (2012). Shear strengthening of short span reinforced concrete beams with CFRP sheets. *Arabian Journal for Science and Engineering, 38*(3), 523–536. https://doi.org/10.1007/s13369-012-0333-z

Cao, S. Y., Chen, J. F., Teng, J. G., Hao, Z., & Chen, J. (2005). Debonding in RC beams shear strengthened with complete FRP wraps. *Journal of Composites for Construction, 9*(5), 417–428. https://doi.org/10.1061/(asce)1090-0268(2005)9:5(417)

Carolin, A., & Täljsten, B. (2005). Experimental study of strengthening for increased shear bearing capacity. *Journal of Composites for Construction, 9*(6), 488–496. https://doi.org/10.1061/(asce)1090-0268(2005)9:6(488)

Chaallal, O., Nollet, M. J., & Perraton, D. (1998). Shear strengthening of RC beams by externally bonded side CFRP strips. *Journal of Composites for Construction, 2*(2), 111–113. https://doi.org/10.1061/(asce)1090-0268(1998)2:2(111)

Chen, J., & Teng, J. (2003b). Shear capacity of FRP-strengthened RC beams: FRP debonding. *Construction and Building Materials, 17*(1), 27–41. https://doi.org/10.1016/s0950-0618(02)00091-0

Chen, J. F., & Teng, J. G. (2003a). Shear capacity of fiber-reinforced polymer-strengthened reinforced concrete beams: Fiber reinforced polymer rupture. *Journal of Structural Engineering, 129*(5), 615–625. https://doi.org/10.1061/(asce)0733-9445(2003)129:5(615)

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

CNR. (2013). *Guide for the design and construction of externally bonded FRP systems for strengthening existing structures* (CNR-DT 200 R1/2013). National Research Council.

Colalillo, M. A., & Sheikh, S. A. (2012). Seismic retrofit of shear-critical reinforced concrete beams using CFRP. *Construction and Building Materials, 32*, 99–109. https://doi.org/10.1016/j.conbuildmat.2010.12.065

Colalillo, M. A., & Sheikh, S. A. (2014). Behavior of shear-critical reinforced concrete beams strengthened with fiber-reinforced polymer—experimentation. *ACI Structural Journal, 111*(6), 1373–1384. https://doi.org/10.14359/51687035

Diagana, C., Li, A., Gedalia, B., & Delmas, Y. (2003). Shear strengthening effectiveness with CFF strips. *Engineering Structures, 25*(4), 507–516. https://doi.org/10.1016/s0141-0296(02)00208-0

Dias, S. J. E., & Barros, J. A. O. (2005). Experimental research of a new CFRP-based shear strengthening technique for reinforced concrete beams. *Revista IBRACON de Estruturas, 1*(2), 103–126.

Dias, S. J. E., & Barros, J. A. O. (2009). Performance of reinforced concrete T beams strengthened in shear with NSM CFRP laminates. *Engineering Structures, 32*(2), 373–384. https://doi.org/10.1016/j.engstruct.2009.10.001

Dong, J. F., Wang, Q. Y., & Zhu, Y. M. (2010). Experimental study on precracked RC beams strengthened with externally bonded CFRP sheets. *Advanced Materials Research, 150–151*, 842–846. https://doi.org/10.4028/www.scientific.net/amr.150-151.842

Ebead, U., & Saeed, H. (2013). Hybrid shear strengthening system for reinforced concrete beams: An experimental study. *Engineering Structures, 49*, 421–433. https://doi.org/10.1016/j.engstruct.2012.11.039

El-Ghandour, A. (2011). Experimental and analytical investigation of CFRP flexural and shear strengthening efficiencies of RC beams. *Construction and Building Materials, 25*(3), 1419–1429. https://doi.org/10.1016/j.conbuildmat.2010.09.001

El-Saikaly, G., & Chaallal, O. (2015). Extending the fatigue life of reinforced concrete T-beams strengthened in shear with externally bonded FRP: Upgrading versus repairing. *Journal of Composites for Construction, 19*(1), Article 04014027. https://doi.org/10.1061/(asce)cc.1943-5614.0000492

Feng, D. C., Wang, W. J., Mangalathu, S., Hu, G., & Wu, T. (2021). Implementing ensemble learning methods to predict the shear strength of RC deep beams with/without web reinforcements. *Engineering Structures, 235*, Article 111979. https://doi.org/10.1016/j.engstruct.2021.111979

fib. (2001). *Externally bonded FRP reinforcement for RC structures* (fib Bulletin 14). International Federation for Structural Concrete.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Godat, A., & Chaallal, O. (2013). Strut-and-tie method for externally bonded FRP shear-strengthened large-scale RC beams. *Composite Structures, 99*, 327–338. https://doi.org/10.1016/j.compstruct.2012.11.034

Gonzalez-Libreros, J., Sneed, L., D'Antino, T., & Pellegrino, C. (2017). Behavior of RC beams strengthened in shear with FRP and FRCM composites. *Engineering Structures, 150*, 830–842. https://doi.org/10.1016/j.engstruct.2017.07.084

Grande, E. (2013). Experimental response of RC beams strengthened in shear by FRP sheets. *The Open Civil Engineering Journal, 7*(1), 126–134. https://doi.org/10.2174/1874149501307010126

Grande, E., Imbimbo, M., & Rasulo, A. (2009). Effect of transverse steel on the response of RC beams strengthened in shear by FRP: Experimental study. *Journal of Composites for Construction, 13*(5), 405–414. https://doi.org/10.1061/(asce)1090-0268(2009)13:5(405)

Hu, Z., Zhou, X., Guo, M., Huang, X., & Hu, B. (2020). Enhancing the performance of CFRP shear-strengthened RC beams using “ductile” anchoring devices. *Frontiers in Materials, 7*, Article 292. https://doi.org/10.3389/fmats.2020.00292

Ianniruberto, U., & Imbimbo, M. (2004). Role of fiber reinforced plastic sheets in shear response of reinforced concrete beams: Experimental and analytical results. *Journal of Composites for Construction, 8*(5), 415–424. https://doi.org/10.1061/(asce)1090-0268(2004)8:5(415)

Islam, M., Mansur, M., & Maalej, M. (2005). Shear strengthening of RC deep beams using externally bonded FRP systems. *Cement and Concrete Composites, 27*(3), 413–420. https://doi.org/10.1016/j.cemconcomp.2004.04.002

Jayaprakash, J., Abdul Samad, A. A., Anvar Abbasovich, A., & Abang Ali, A. A. (2008). Shear capacity of precracked and non-precracked reinforced concrete shear beams with externally bonded bi-directional CFRP strips. *Construction and Building Materials, 22*(6), 1148–1165. https://doi.org/10.1016/j.conbuildmat.2007.02.008

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4*(9), Article 100804. https://doi.org/10.1016/j.patter.2023.100804

Kar, S., & Biswal, K. (2021). External shear strengthening of RC beams with basalt fiber sheets: An experimental study. *Structures, 31*, 305–315. https://doi.org/10.1016/j.istruc.2021.01.094

Karzad, A. S., Leblouba, M., Toubat, S. A., & Maalej, M. (2019). Repair and strengthening of shear-deficient reinforced concrete beams using carbon fiber reinforced polymer. *Composite Structures, 223*, Article 110963. https://doi.org/10.1016/j.compstruct.2019.110963

Ke, Y., Zhang, S., Jedrzejko, M., Lin, G., Li, W., & Nie, X. (2024). Strength models of near-surface mounted (NSM) fibre-reinforced polymer (FRP) shear-strengthened RC beams based on machine learning approaches. *Composite Structures, 337*, Article 118045. https://doi.org/10.1016/j.compstruct.2024.118045

Khalifa, A., & Nanni, A. (2002). Rehabilitation of rectangular simply supported RC beams with shear deficiencies using CFRP composites. *Construction and Building Materials, 16*(3), 135–146. https://doi.org/10.1016/s0950-0618(02)00002-8

Khalifa, A. M., Alkhrdaji, T., Nanni, A., & Lansburg, S. (1999). Anchorage of surface mounted FRP reinforcement. *Concrete International, 21*(10), 49–54.

Kim, G., Sim, J., & Oh, H. (2008). Shear strength of strengthened RC beams with FRPs in shear. *Construction and Building Materials, 22*(6), 1261–1270. https://doi.org/10.1016/j.conbuildmat.2007.01.021

Koutas, L., & Triantafillou, T. C. (2013). Use of anchors in shear strengthening of reinforced concrete T-beams with FRP. *Journal of Composites for Construction, 17*(1), 101–107. https://doi.org/10.1061/(asce)cc.1943-5614.0000316

Lee, J. Y., Hwang, H. B., & Doh, J. H. (2012). Effective strain of RC beams strengthened in shear with FRP. *Composites Part B: Engineering, 43*(2), 754–765. https://doi.org/10.1016/j.compositesb.2011.11.050

Leung, C. K. Y., Chen, Z., Lee, S., Ng, M., Xu, M., & Tang, J. (2007). Effect of size on the failure of geometrically similar concrete beams strengthened in shear with FRP strips. *Journal of Composites for Construction, 11*(5), 487–496. https://doi.org/10.1061/(asce)1090-0268(2007)11:5(487)

Li, A., Diagana, C., & Delmas, Y. (2002). Shear strengthening effect by bonded composite fabrics on RC beams. *Composites Part B: Engineering, 33*(3), 225–239. https://doi.org/10.1016/s1359-8368(02)00003-3

Li, A., Diagana, C., Delmas, Y., & Gedalia, B. (2003). Shear performance with externally bonded carbon fibre fabrics. *Fibre-Reinforced Polymer Reinforcement for Concrete Structures*, 497–506. https://doi.org/10.1142/9789812704863_0046

Li, W., & Leung, C. K. (2017). Effect of shear span-depth ratio on mechanical performance of RC beams strengthened in shear with U-wrapping FRP strips. *Composite Structures, 177*, 141–157. https://doi.org/10.1016/j.compstruct.2017.06.059

Lim, D. H., & Kwon, Y. S. (2009). An experimental study on the shear behavior of RC beams strengthened with near surface mounted and externally bonded CFRP strips. *Journal of the Korea Concrete Institute, 21*(3), 337–345. https://doi.org/10.4334/jkci.2009.21.3.337

Liu, X., Figueredo, G., Gordon, G. S., & Thermou, G. (2024). Data-driven shear strength prediction of RC beams strengthened with FRCM jackets using machine learning approach. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4898878

Manos, G., Theofanous, M., & Katakalos, K. (2013). Numerical simulation of the shear behaviour of reinforced concrete rectangular beam specimens with or without FRP-strip shear reinforcement. *Advances in Engineering Software, 67*, 47–56. https://doi.org/10.1016/j.advengsoft.2013.08.001

Marani, A., & Nehdi, M. L. (2022). Predicting shear strength of FRP-reinforced concrete beams using novel synthetic data driven deep learning. *Engineering Structures, 257*, Article 114083. https://doi.org/10.1016/j.engstruct.2022.114083

Mitsui, Y., Murakami, K., Takeda, K., & Sakai, H. (1997). A study on shear reinforcement of reinforced concrete beams externally bonded with carbon fiber sheets. *Composite Interfaces, 5*(4), 285–295. https://doi.org/10.1163/156855498x00081

Mofidi, A., & Chaallal, O. (2011). Shear strengthening of RC beams with externally bonded FRP composites: Effect of strip-width-to-strip-spacing ratio. *Journal of Composites for Construction, 15*(5), 732–742. https://doi.org/10.1061/(asce)cc.1943-5614.0000219

Mofidi, A., & Chaallal, O. (2014). Tests and design provisions for reinforced-concrete beams strengthened in shear using FRP sheets and strips. *International Journal of Concrete Structures and Materials, 8*(2), 117–128. https://doi.org/10.1007/s40069-013-0060-1

Mohammad, A., Sarsam, K., & Al-Bayati, N. (2018). Shear strength enhancement of lightweight aggregate reinforced concrete deep beams by using CFRP strips. *MATEC Web of Conferences, 162*, Article 04011. https://doi.org/10.1051/matecconf/201816204011

Monti, G., & Liotta, M. (2007). Tests and design equations for FRP-strengthening in shear. *Construction and Building Materials, 21*(4), 799–809. https://doi.org/10.1016/j.conbuildmat.2006.06.023

Moradi, E., Naderpour, H., & Kheyroddin, A. (2020). An experimental approach for shear strengthening of RC beams using a proposed technique by embedded through-section FRP sheets. *Composite Structures, 238*, Article 111988. https://doi.org/10.1016/j.compstruct.2020.111988

Mosallam, A. S., & Banerjee, S. (2007). Shear enhancement of reinforced concrete beams strengthened with FRP composite laminates. *Composites Part B: Engineering, 38*(5–6), 781–793. https://doi.org/10.1016/j.compositesb.2006.10.002

Mostofinejad, D., Hosseini, S. A., & Razavi, S. B. (2016). Influence of different bonding and wrapping techniques on performance of beams strengthened in shear using CFRP reinforcement. *Construction and Building Materials, 116*, 310–320. https://doi.org/10.1016/j.conbuildmat.2016.04.113

Naser, M. Z. (2021). Mechanistically informed machine learning and artificial intelligence in fire engineering and sciences. *Fire Technology, 57*(6), 2741–2784. https://doi.org/10.1007/s10694-020-01069-8

Nguyen-Minh, L., & Rovňák, M. (2015). Size effect in uncracked and pre-cracked reinforced concrete beams shear-strengthened with composite jackets. *Composites Part B: Engineering, 78*, 361–376. https://doi.org/10.1016/j.compositesb.2015.02.035

Niu, P., Li, S., & Li, S. (2011). Researches on the flexing resistance of RC defective beams of hybrid-bonding of FRP. *Design, Construction, Rehabilitation, and Maintenance of Bridges*, 175–183. https://doi.org/10.1061/47630(409)23

Norris, T., Saadatmanesh, H., & Ehsani, M. R. (1997). Shear and flexural strengthening of R/C beams with carbon fiber sheets. *Journal of Structural Engineering, 123*(7), 903–911. https://doi.org/10.1061/(asce)0733-9445(1997)123:7(903)

Ombres, L., Aiello, M. A., Cascardi, A., & Verre, S. (2024). Modeling of steel-reinforced grout composite system-to-concrete bond capacity using artificial neural networks. *Journal of Composites for Construction, 28*(5), Article 04024034. https://doi.org/10.1061/jccof2.cceng-4453

Ozden, S., Atalay, H. M., Akpinar, E., Erdogan, H., & Vulaş, Y. Z. (2014). Shear strengthening of reinforced concrete T-beams with fully or partially bonded fibre-reinforced polymer composites. *Structural Concrete, 15*(2), 229–239. https://doi.org/10.1002/suco.201300031

Pellegrino, C., & Modena, C. (2002). Fiber reinforced polymer shear strengthening of reinforced concrete beams with transverse steel reinforcement. *Journal of Composites for Construction, 6*(2), 104–111. https://doi.org/10.1061/(asce)1090-0268(2002)6:2(104)

Pellegrino, C., & Modena, C. (2008). An experimentally based analytical model for the shear capacity of FRP-strengthened reinforced concrete beams. *Mechanics of Composite Materials, 44*(3), 231–244. https://doi.org/10.1007/s11029-008-9016-y

Perera, R., Barchín, M., Arteaga, A., & Diego, A. D. (2010). Prediction of the ultimate strength of reinforced concrete beams FRP-strengthened in shear using neural networks. *Composites Part B: Engineering, 41*(4), 287–298. https://doi.org/10.1016/j.compositesb.2010.03.003

Pham, T. M., & Hao, H. (2016). Impact behavior of FRP-strengthened RC beams without stirrups. *Journal of Composites for Construction, 20*(4), Article 04016011. https://doi.org/10.1061/(asce)cc.1943-5614.0000671

Qu, Z., Lu, X., Ye, L., Chen, J. F., & Rotter, J. (2006). Numerical modeling of FRP shear strengthened RC beams using compression field theory. *Proceedings of the Third International Conference on FRP Composites in Civil Engineering (CICE 2006)*, 391–394, Miami, Florida, USA.

Qu, Z., Lu, X., & Ye, L. P. (2005). Size effect of shear contribution of externally bonded FRP U-jackets for RC beams. *Proceedings of the International Symposium on Bond Behaviour of FRP in Structures (BBFS 2005)*, Hong Kong, China.

Rizzo, A., & De Lorenzis, L. (2009). Behavior and capacity of RC beams strengthened in shear with NSM FRP reinforcement. *Construction and Building Materials, 23*(4), 1555–1567. https://doi.org/10.1016/j.conbuildmat.2007.08.014

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., & Wintle, B. A. (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

Saribiyik, A., Abodan, B., & Balci, M. T. (2021). Experimental study on shear strengthening of RC beams with basalt FRP strips using different wrapping methods. *Engineering Science and Technology, an International Journal, 24*(1), 192–204. https://doi.org/10.1016/j.jestch.2020.06.003

Sengun, K., & Arslan, G. (2022). Investigation of the parameters affecting the behavior of RC beams strengthened with FRP. *Frontiers of Structural and Civil Engineering, 16*(6), 729–743. https://doi.org/10.1007/s11709-022-0854-9

Sengun, K., & Arslan, G. (2023). Performance of RC beams strengthened in flexure and shear with CFRP and GFRP. *Iranian Journal of Science and Technology, Transactions of Civil Engineering, 48*(1), 117–130. https://doi.org/10.1007/s40996-023-01305-5

Sun, H., Burton, H. V., & Huang, H. (2021). Machine learning applications for building structural design and performance assessment: State-of-the-art review. *Journal of Building Engineering, 33*, Article 101816. https://doi.org/10.1016/j.jobe.2020.101816

Täljsten, B. (2003). Strengthening concrete beams for shear with CFRP sheets. *Construction and Building Materials, 17*(1), 15–26. https://doi.org/10.1016/s0950-0618(02)00088-0

Teng, J. G., Chen, G. M., Chen, J. F., Rosenboom, O. A., & Lam, L. (2009). Behavior of RC beams shear strengthened with bonded or unbonded FRP wraps. *Journal of Composites for Construction, 13*(5), 394–404. https://doi.org/10.1061/(asce)cc.1943-5614.0000040

Teng, J. G., Lam, L., & Chen, J. F. (2004). Shear strengthening of RC beams with FRP composites. *Progress in Structural Engineering and Materials, 6*(3), 173–184. https://doi.org/10.1002/pse.179

Thai, H. T. (2022). Machine learning for structural engineering: A state-of-the-art review. *Structures, 38*, 448–491. https://doi.org/10.1016/j.istruc.2022.02.003

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*(1), Article 91. https://doi.org/10.1186/1471-2105-7-91

Vehtari, A., Gelman, A., & Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P. C. (2021). Rank-normalization, folding, and localization: An improved Rˆ for assessing convergence of MCMC (with discussion). *Bayesian Analysis, 16*(2), 667–718. https://doi.org/10.1214/20-ba1221

Wang, C., Zou, X., Sneed, L. H., Zhang, F., Zheng, K., Xu, H., & Li, G. (2023). Shear strength prediction of FRP-strengthened concrete beams using interpretable machine learning. *Construction and Building Materials, 407*, Article 133553. https://doi.org/10.1016/j.conbuildmat.2023.133553

Yu, F., Guo, S., Wang, S., & Fang, Y. (2019). Experimental study on high pre-cracked RC beams shear-strengthened with CFRP strips. *Composite Structures, 225*, Article 111163. https://doi.org/10.1016/j.compstruct.2019.111163

Zhang, Z., Hsu, C. T. T., & Moren, J. (2004). Shear strengthening of reinforced concrete deep beams using carbon fiber reinforced polymer laminates. *Journal of Composites for Construction, 8*(5), 403–414. https://doi.org/10.1061/(asce)1090-0268(2004)8:5(403)

Zhao, J., Zhu, M., Xu, L., Chen, M., & Shi, M. (2025). Prediction of shear capacity of fiber-reinforced polymer-reinforced concrete beams based on machine learning. *Buildings, 15*(11), Article 1908. https://doi.org/10.3390/buildings15111908

How to Cite This Article

Rasheed, S. E., & Zainy, H. A. (2026). Cross-database validation and calibrated uncertainty for machine learning models of the shear capacity of FRP-strengthened RC beams. Civil and Environmental Engineering, 0 (0). https://doi.org/10.2478/cee-2027-0026

Appendix A. Source experimental programmes

The merged database draws on the primary experimental literature compiled by the two source studies. Table A1 lists every source programme whose bibliographic record could be independently verified against Crossref or OpenAlex, together with the database in which it appears. Records that could not be verified against an indexed bibliographic entry — predominantly Chinese-language journal articles, university theses and older conference proceedings that are not indexed — are retained in the dataset but are not listed here, and are flagged in the supplementary reconciliation log. One entry in the DB-A bibliography is annotated by its own compilers as "not finally confirmed" and is excluded from the reference list.

Table A1: Independently verified source experimental programmes contributing specimens to the merged database. Full bibliographic details appear in the reference list

#	Source programme	Database
1	Abdel-Jaber (2003)	A/B
2	Abdul Samad et al. (2017)	DB-B
3	Adhikary et al. (2003)	DB-A
4	Adhikary & Mutsuyoshi (2004)	DB-B
5	Akkaya et al. (2022a)	DB-B
6	Akkaya et al. (2022b)	DB-B
7	Akkaya et al. (2024)	DB-B
8	Al-Lebban & Mackie (2019)	A/B
9	Al-Rousan et al. (2015)	DB-A
10	Al-Rousan & Issa (2016)	A/B
11	Alam & Al Riyami (2018)	DB-B
12	Altin et al. (2010)	DB-A
13	Bae et al. (2012)	DB-A
14	Barros et al. (2007)	A/B
15	Bukhari et al. (2012)	DB-B
16	Carolin & Täljsten (2005)	A/B
17	Chaallal et al. (1998)	DB-A
18	Colalillo & Sheikh (2012)	A/B
19	Colalillo & Sheikh (2014)	DB-A
20	Diagana et al. (2003)	DB-A

#	Source programme	Database
21	Dias & Barros (2005)	DB-A
22	Dias & Barros (2009)	DB-A
23	Dong et al. (2010)	DB-A
24	Ebead & Saeed (2013)	A/B
25	El-Ghandour (2011)	DB-A
26	El-Saikaly & Chaallal (2015)	DB-A
27	Godat & Chaallal (2013)	DB-A
28	Grande (2013)	DB-B
29	Hu et al. (2020)	DB-B
30	Ianniruberto & Imbimbo (2004)	DB-B
31	Islam et al. (2005)	DB-A
32	Jayaprakash et al. (2008)	DB-B
33	Kar & Biswal (2021)	DB-B
34	Karzad et al. (2019)	DB-B
35	Khalifa et al. (1999)	DB-A
36	Khalifa & Nanni (2002)	DB-B
37	Kim et al. (2008)	A/B
38	Koutas & Triantafillou (2013)	DB-A
39	Lee et al. (2012)	DB-B
40	Leung et al. (2007)	A/B
41	Li et al. (2002)	DB-A
42	Li et al. (2003)	DB-A
43	Li & Leung (2017)	A/B
44	Lim & Kwon (2009)	DB-A
45	Manos et al. (2013)	DB-A
46	Mitsui et al. (1997)	DB-A
47	Mofidi & Chaallal (2014)	DB-A
48	Mohammad et al. (2018)	DB-B
49	Moradi et al. (2020)	DB-B
50	Mosallam & Banerjee (2007)	DB-A
51	Mostofinejad et al. (2016)	DB-B
52	Nguyen-Minh & Rovňák (2015)	DB-A
53	Niu et al. (2011)	DB-A
54	Norris et al. (1997)	DB-A
55	Ozden et al. (2014)	DB-A
56	Pellegrino & Modena (2002)	DB-B
57	Pellegrino & Modena (2008)	DB-B
58	Pham & Hao (2016)	DB-A
59	Qu et al. (2005)	DB-A
60	Qu et al. (2006)	DB-A
61	Rizzo & De Lorenzis (2009)	DB-A
62	Saribiyik et al. (2021)	DB-B
63	Sengun & Arslan (2022)	DB-B
64	Sengun & Arslan (2023)	DB-B
65	Teng et al. (2009)	A/B
66	Täljsten (2003)	A/B
67	Yu et al. (2019)	DB-B
68	Zhang et al. (2004)	A/B

Appendix B. Sampling scheme for the hierarchical errors-in-variables model

Write $X = [1, z, e]$ for the design matrix formed from the latent log-covariates z and the remaining covariates e. Let $\theta = [b_0, \beta, \gamma]$ denote the coefficient vector, u the vector of programme random effects, and S_m the diagonal measurement-

error covariance, fixed to the values measured in Section 4. With normal priors on θ and on the latent covariates, and inverse-gamma priors on the variance components, the full conditionals are:

- θ | · is normal, with precision $X'X/\sigma^2 + P_0$ and mean obtained by solving that system against $X'(y - u)/\sigma^2$, where P_0 is the prior precision;
- u_s | · is normal for each programme independently, with precision $n_s/\sigma^2 + 1/\tau^2$ and mean equal to the programme residual sum divided by σ^2 and by that precision;
- z_i | · is normal with precision $S_m^{-1} + I/s_0^2 + \beta\beta'/\sigma^2$. Because S_m and β are shared across specimens this precision is identical for every i, so one Cholesky factorisation per sweep serves the whole sample and the latent block is drawn jointly;
- σ^2 and τ^2 | · are inverse-gamma, with shape and scale updated by the residual and random-effect sums of squares respectively.

Priors were weakly informative throughout: standard normal on standardised coefficients, N(0, 25) on the intercept, $N(0, 1.5^2)$ on the latent standardised covariates, and inverse-gamma(2, 0.5) on both variance components. The random effects are drawn in non-centred form, which is necessary because many programmes contribute only three or four specimens and a centred parameterisation mixes poorly in that regime. The target transcription variance measured in Section 4 enters as a fixed additive component of the observation variance, so the reported within-programme standard deviation is net of transcription noise.

Convergence was monitored for every regression coefficient, the intercept and both variance components. Table B1 reports, for each, the posterior mean and standard deviation, the rank-normalised split statistic, the bulk and tail effective sample sizes, and the Monte Carlo standard error of the posterior mean. Figure 12 shows trace plots and rank histograms for a representative subset; well-mixed traces and uniform rank histograms across chains indicate that the chains are exploring the same distribution.

Table B1: Convergence diagnostics for the hierarchical errors-in-variables sampler, from four chains of 4,000 retained draws after 1,500 burn-in iterations

Parameter [-]	Mean [-]	SD [-]	R-hat [-]	ESS bulk [-]	ESS tail [-]	MCSE [-]
ln b_w	+0.140	0.047	1.0009	3436	6001	0.0008
ln d	+0.288	0.048	1.0004	3461	6747	0.0008
ln a_d	+0.004	0.034	1.0008	3877	6565	0.0005
ln f_c	+0.124	0.039	1.0015	2658	4661	0.0008
ln E_f	+0.133	0.041	1.0025	2570	5839	0.0008
ln rho_f	+0.288	0.040	1.0007	3576	6730	0.0007
rho_{sv}	-0.067	0.032	1.0013	5029	8836	0.0004
cfg_{FW}	+0.295	0.037	1.0005	4155	6978	0.0006
cfg_{SB}	-0.121	0.042	1.0002	3896	6335	0.0007
alpha=45	+0.046	0.025	1.0001	10035	12976	0.0002
tau	+0.444	0.038	1.0011	3372	6807	0.0007
sigma	+0.567	0.017	1.0004	9001	13119	0.0002
b0	+3.766	0.043	1.0008	1977	3397	0.0010

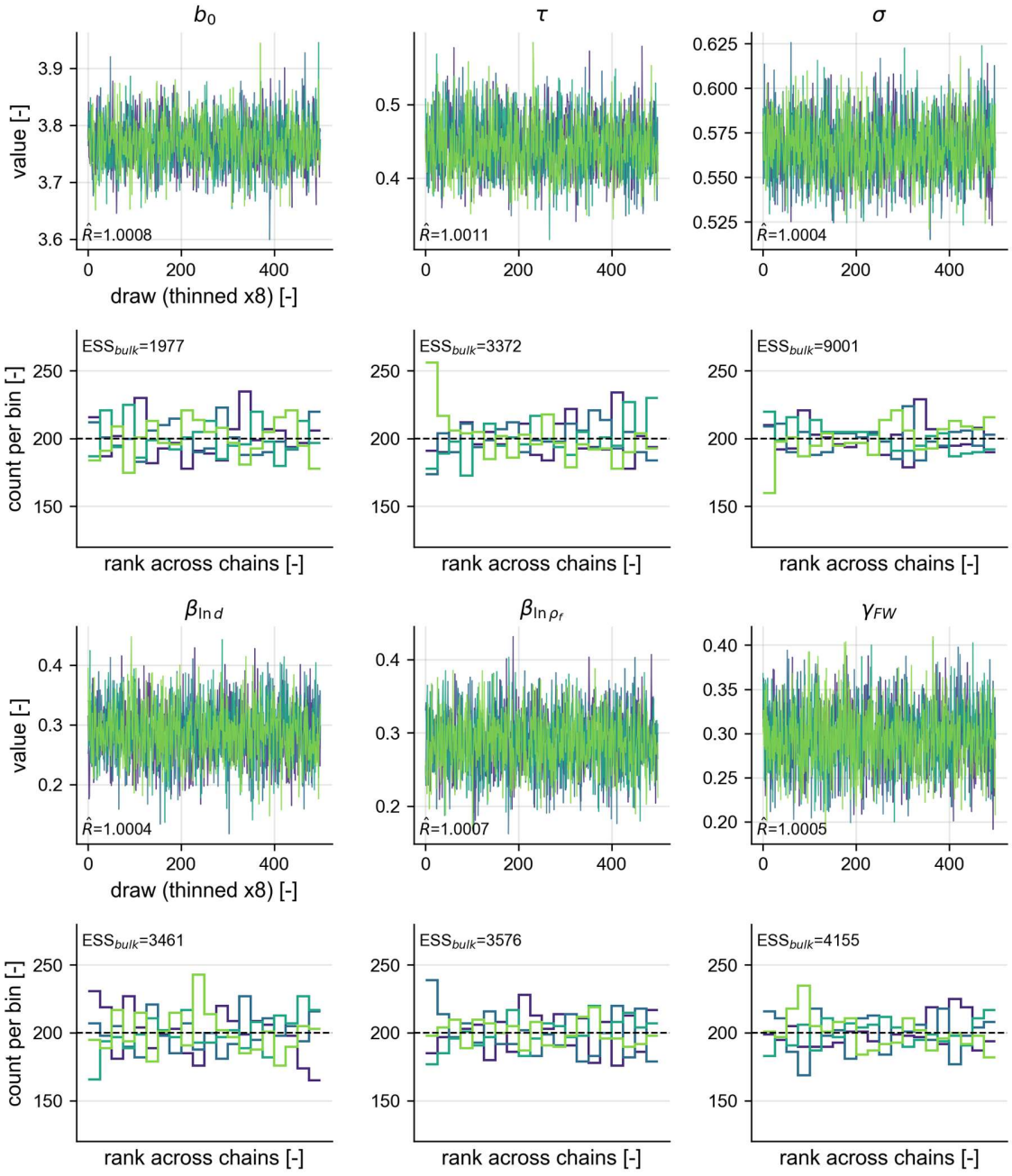


Figure 12: Trace plot (upper panel of each pair) and rank histogram (lower panel of each pair) for the intercept, the two variance components and three representative coefficients; the broken line marks the count expected under uniformity

Appendix C. Results on the logarithmic scale

Table 4 reports the coefficient of determination on the untransformed contribution, which is the scale used in the literature this paper examines. Because the models are fitted to the logarithm of the contribution, the same quantity computed on that scale is given in Table C1. The ordering of the regimes is unchanged, but two differences are worth noting. The extreme negative entry of Table 4 becomes an ordinary small negative number, confirming that its magnitude is an artefact of exponentiation. Conversely, the gradient-boosted models fall below zero under V2 on DB-B, so the collapse there is deeper than the original-scale figure alone conveys; original-scale values are weighted towards the largest specimens, whereas the log scale weights every specimen equally.

Table C1: Coefficient of determination on the logarithmic scale, on which every model is fitted, for each learner under each validation regime

Regime [-]	Subset [-]	Ridge [-]	SVR [-]	MLP [-]	RF [-]	XGBoost [-]	XGB-mono [-]
V1	DB-A	0.407	0.480	0.442	0.550	0.571	0.575
V1	DB-B	0.250	0.371	0.269	0.373	0.382	0.363
V2	DB-A	0.263	0.255	0.380	0.388	0.360	0.359
V2	DB-B	0.063	0.083	-0.036	-0.100	-0.465	-0.551
V3	A->B	0.150	0.182	0.200	0.141	0.128	0.121
V3	B->A	-0.223	0.304	0.044	0.290	0.256	0.200
V4	merged	0.292	0.320	0.320	0.257	0.272	0.202
V5	curation A	-0.083	0.236	0.286	0.299	0.226	0.276
V5	curation B	-0.093	0.199	0.261	0.281	0.265	0.278