Res Intell Manuf Assem, 2026, 5(2): 423-433
DOI: 10.25082/RIMA.2026.02.004

RESEARCH ARTICLE

Validation Under Archetype Shift for Data-Driven Bio-Inspired Lightweight Design: Optimism, Conformal Calibration and Density-Matched Auditing

Aswin Karkadakattil

Independent Researcher, Kasaragod, Kerala 671314, India

Check for updates

Correspondence to: Aswin Karkadakattil, Independent Researcher, Kasaragod, Kerala 671314, India; Email: ashwinharik20000@gmail.com

Received: September 19, 2026;
Accepted: November 22, 2026;
Published: November 26, 2026.

Citation: Karkadakattil A. Validation Under Archetype Shift for Data-Driven Bio-Inspired Lightweight Design: Optimism, Conformal Calibration and Density-Matched Auditing. *Res Intell Manuf Assem*, 2026, **5**(2): 423-433.
https://doi.org/10.25082/RIMA.2026.02.004

Copyright: © 2026 Aswin Karkadakattil. This is an open access article distributed under the terms of the Creative Commons Attribution-Noncommercial 4.0 International License, which permits all noncommercial use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: Data-driven design rules for bio-inspired lightweight structures are almost always validated with random train–test splits. That tells a designer little about the question that matters: will a rule learned from known architectures hold for a new one? We examine this on two controlled benchmarks with a known governing law, one inside the model library and one outside it (15 draws each), and organise the findings into a four-part archetype-shift protocol. Holding out whole archetypes raises prediction error by 11–36%. The excess squared error is additive in archetype-offset and paper-level variance (leave-one-setting-out R^2 of 0.79–0.92), and flexible models pay an extra extrapolation cost. Conformal intervals calibrated on random folds cover only 80–85% of unseen archetypes at a nominal 90%; calibrating on leave-one-archetype-out residuals keeps coverage within three points of nominal from 50 to 95%, at the price of intervals 18–34% wider, and descriptor-distance flags do not show where the failures occur. In a sparse Gibson–Ashby-type law, only the density exponent and its stretching-fraction interaction are recovered in every refit; secondary terms are selected in 38–90% of refits whether or not they exist. Finally, a break-even density exponent applied to a published study of eight honeycomb designs shows that reported energy-absorption gains of 22–47% came with 1.2- to 3.3-fold more mass, and that only three of seven gains exceed 10% once the study's model error is allowed for, even if absorption were independent of density. Together, these results turn a general caution about random splits into a concrete validation and reporting protocol for data-driven bio-inspired design.

Keywords: bio-inspired structures, lightweight design, grouped validation, conformal prediction, sparse regression, density-matched comparison

1 Introduction

Nacre toughens a brittle mineral with staggered platelets, bamboo grades its fibre density across the wall, pomelo peel absorbs impact in a hierarchical foam, and Bouligand layering deflects cracks away from the loading direction [1,2]. Engineered analogues of these structures now fill the literature, and a growing number of studies mine it for general design rules by fitting regression or tree-ensemble models to reported properties [3,4].

The weak point of this approach is validation. Entries from one paper share a fabrication route, a test frame and often a systematic offset, so a random split puts near-duplicates on both sides of the partition and rewards a model for recognising the paper rather than the architecture [5]. A designer faces a different problem. The rule must work for an architecture that is absent from the training data, which means holding out an archetype, not a row. Grouped validation is the accepted remedy for dependent data in ecology and materials informatics [6–8], and area-of-applicability scores help flag unreliable extrapolation [9]. For design rules learned from lightweight-architecture data, four questions are still open: how badly random splits mislead and why; whether uncertainty intervals can stay valid for a new archetype; which terms of a discovered law survive the shift; and whether the gains reported in the underlying studies are architectural at all.

We take up these questions on controlled benchmarks, where the governing law is known and every claim can be checked against ground truth, and then on one published dataset. The answers fit together as a single archetype-shift protocol (Figure 1), which is the main contribution of this paper. It has four components:

(1) Optimism, quantified. The excess squared error under held-out-archetype validation is additive in archetype-offset variance and paper-level variance, with an extra extrapolation term for flexible models. The model predicts the optimism of settings it was not fitted on (leave-one-setting-out R^2 of 0.79 to 0.92).

(2) Archetype-calibrated intervals. Conformal intervals calibrated on random folds undercover unseen archetypes. Calibrating on leave-one-archetype-out residuals restores near-nominal coverage at every confidence level, something descriptor-distance risk scores cannot do, and it exposes the archetypes that still fall short.

(3) An identifiability map. Selection frequencies under archetype-grouped refitting show which terms of a Gibson–Ashby-type law survive archetype shift and which do not.

(4) A density-matched audit. A break-even density exponent states how strongly specific energy absorption must depend on density before a reported advantage disappears. We apply it, with an allowance for model error, to a published study of eight honeycomb designs.

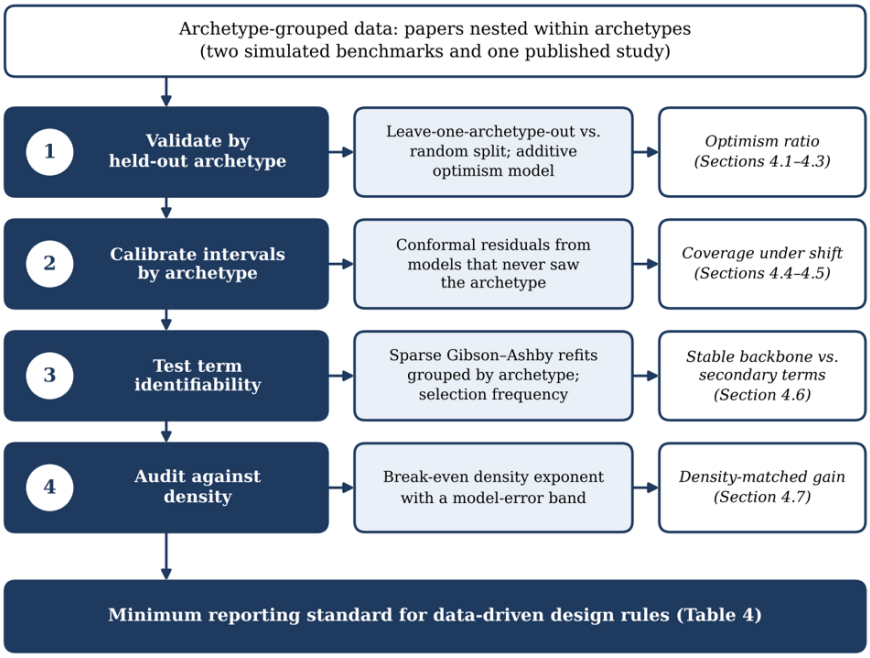


Figure 1 The archetype-shift protocol. Data in which papers are nested within archetypes pass through four steps, each yielding a quantity that is reported alongside the result: the optimism ratio, archetype-calibrated coverage, selection frequencies of law terms and density-matched gains. Together they define the minimum reporting standard in Table 4.

2 Background

For cellular solids the relative stiffness scales as $E/E_s \propto \bar{\rho}^n$, where $\bar{\rho}$ is the relative density and the exponent n runs from 1 for stretching-dominated to 2 for bending-dominated topologies [10]. Hierarchy, grading and anisotropy shift the prefactor and sometimes the exponent [1, 2], and learned surrogates for lattices and metamaterials are now routine [3, 4].

Conformal prediction offers finite-sample coverage under exchangeability [11, 12], an assumption that breaks down when test points come from an unseen group. Weighted conformal methods correct covariate shift but require the shift to be specified [13]; cross-validation-based variants such as the jackknife+ avoid a separate calibration set [14, 15]; and hierarchical conformal methods treat groups as the exchangeable unit [16]. In design applications the shift is known in advance, because the target is a new archetype. Calibration can therefore reproduce it directly, using residuals from models that never saw the calibration archetype.

Sparse regression over a candidate library, solved by sequential thresholded least squares [17], offers a lightweight alternative to free-form symbolic search [18], and selection frequencies under resampling are the basis of stability selection [19]. For energy absorption, specific metrics divide by mass, which removes size but not the dependence of the property on relative density [20]. For thin-walled honeycombs crushed along the cell axis, the mean crush stress scales as $\rho^{-5/3}$ [21].

3 Methods

The protocol has four steps (Figure 1). The first three are evaluated on two simulated benchmarks in which the governing law is known; the fourth is applied to published data. The sections below describe the benchmarks, the models and each step in turn.

3.1 Benchmarks

Both benchmarks contain eight archetypes (nacre, bamboo, pomelo peel, honeycomb, sponge, Bouligand, cuttlebone, elytra). Each has its own mean stretching fraction s, hierarchy order h and gradient index g, and contributes five to eight simulated papers of three to six entries, giving about 240 entries per draw. Base materials are polymer, metal or ceramic. Each paper carries an offset $b_{paper} \sim \mathrm{N}(0, {\sigma_p}^2)$ that stands in for fabrication and test differences, with $\sigma_p = 0.10$ unless varied. The target is $\log_{10}$ of the effective stiffness E.

B1 (law inside the library). The stiffness follows:

$$\log E = \log E_s - 0.097 + (2 - s)\log\bar{\rho} + \Delta_h + \Delta_g + 0.02\log\dot{\varepsilon} + a_{\mathrm{arch}} + b_{\mathrm{paper}} + \varepsilon \quad (1)$$

$$\Delta_h = 0.07h(1 - \bar{\rho}), \quad \Delta_g = 0.10g(1 - s) \quad (2)$$

where a_{arch} is an archetype offset, $\dot{\epsilon}$ is the strain rate and $\epsilon \sim N(0,\ 0.07^2)$ is the residual noise.

B2 (law outside the library). The exponent becomes 2 - s + δ_{arch} with $\delta_{arch} \sim \mathrm{N}(0, 0.08^2)$, and the hierarchy term Δ_h in Equation (1) is replaced by

$$\Delta_h' = 0.14\tanh(h/1.2)(1 - \bar{\rho}) + 0.06hg(1 - \bar{\rho}) + 0.4\left[\max(0, \log\bar{\rho} + 0.7)\right]^2 \quad (3)$$

and the noise is heteroscedastic, $\sigma_\varepsilon = 0.04 + 0.15\bar{\rho}$. The saturating, interaction, densification and archetype-specific exponent terms cannot be expressed in the candidate library below, so B2 tests the protocol when the model is misspecified. Each benchmark was run for 15 independent draws (seeds 0–14). We report means and standard deviations across draws, with 95% t-intervals and sign tests for paired comparisons.

3.2 Models

The baseline is a linear regression of log E on log ρ^- and log E_s, the Gibson–Ashby form. Random forest (100 trees, minimum leaf 2) and gradient boosting (100 trees, depth 3, learning rate 0.08, subsample 0.8) use all descriptors with fixed hyperparameters, so the advantage of the sparse model reported below holds against untuned ensembles. The sparse model starts from 13 candidate terms (main effects, products of stretching fraction with density, hierarchy and gradient, and quadratic terms), keeps the base modulus by construction, and prunes by sequential thresholded least squares [17] on standardised coefficients (threshold 0.03). It is a sparse linear regression over a fixed library, not a free-form symbolic search such as PySR [18].

3.3 Validation, optimism and its drivers

RMSE was computed on the $\log_{10}$ scale under random five-fold, leave-one-paper-out (eight grouped folds) and leave-one-archetype-out (LOAO, eight folds) validation. Validation optimism is the ratio of LOAO to random-split RMSE within a draw. To identify its drivers, the study was repeated over paper-level noise $\sigma_p \in \{0, 0.1, 0.2, 0.3\}$ and archetype-offset scale $s_a \in \{0, 1, 2\}$ (eight draws per cell, B1 generator). For each model, the additive model

$$\mathrm{RMSE}^2_{\mathrm{LOAO}} - \mathrm{RMSE}^2_{\mathrm{random}} = c + a\,s_a^2 + b\,\sigma_p^2 \quad (4)$$

was fitted to the twelve cell means by least squares, validated by leave-one-cell-out prediction, and used to convert predictions to optimism ratios. To test the effect of data volume, we retained 25, 50, 75 or 100% of the papers in each archetype (12 draws per level, both benchmarks).

3.4 Conformal intervals under archetype shift

For each held-out archetype, the other seven form the training set T. Absolute out-of-fold residuals on T are pooled either from random five-fold splits (random-fold calibration, the usual practice) or from seven-fold leave-one-archetype-out splits within T (archetype calibration), so that each residual comes from a model that has not seen the archetype it is evaluated on. The interval half-width q is the $\lceil (n + 1)(1 - \alpha) \rceil / n$ empirical quantile of those n residuals (miscoverage level α), and the final model f_T is refitted on all of T, giving the interval $f_T(x) \pm q$. This is

the cross-validation member of the jackknife family [14] with archetypes in place of rows. Its coverage guarantee requires the held-out archetype to be exchangeable with the calibration archetypes, and with seven calibration groups that guarantee is weak, so coverage is assessed empirically. Coverage is reported for α from 0.05 to 0.5, together with interval width, averaged over eight folds and 15 draws, and separately for each archetype.

3.5 Identifiability, ablation and risk flags

Term selection frequencies of the sparse law were counted over the eight LOAO refits in every draw, giving 120 refits per benchmark; the base modulus is excluded because it is retained by construction. In B1 the generator terms beyond the backbone are known (hierarchy, hierarchy $\times$ density, gradient $\times$ (1 - s) and strain rate), so the area under the ROC curve measures how well selection frequency separates them from the seven absent candidate terms. A descriptor ablation refitted gradient boosting without one descriptor at a time under LOAO. As a risk flag we used the mean distance from each held-out entry to its five nearest training entries in standardised descriptor space [9], and correlated it with the absolute LOAO error.

3.6 Density-matched audit of published data

Data were taken from Table 2 of a recent numerical study of eight honeycomb designs in aluminium alloy AA3003-H18, compressed along the cell axis to a strain of 0.8 [22]: a regular hexagon and seven bio-inspired variants, each reported with mass, absorbed energy and specific energy absorption (SEA). The table was transcribed twice, once into the analysis script and once by automated retrieval from the publisher's PDF. All 24 values agreed, and absorbed energy divided by mass reproduces the tabulated SEA for all eight entries. The reference cell has a 9 mm edge, 0.75 mm single-wall thickness and 15 mm height, and the designs are modifications of it, so the mass ratio D to the regular hexagon is taken as the relative-density ratio. The paper does not state that the footprint is identical for every design, so this is an explicit assumption of the audit.

A design with SEA ratio S and mass ratio D should be compared with the reference thickened to the same mass, whose SEA is that of the reference multiplied by D^k if SEA scales as ρ^{-k} within a family of similar topologies. The density-matched ratio and the break-even exponent are then

$$S \cdot D^{-k}, \quad k^* = \frac{\ln S}{\ln D} \tag{5}$$

where k^* is the value of k at which the advantage disappears. Thin-walled honeycombs crushed along the cell axis have a mean crush stress proportional to $\rho^{-5/3}$ [21], which gives k = 2/3 for SEA. As an empirical alternative, the slope of ln SEA against ln mass across the eight designs can serve as k. The paper reports a 23.25% deviation in absorbed energy between simulation and experiment for the reference. To allow for design-specific model error, we therefore recomputed k^* after multiplying and dividing each SEA ratio by 1.2325, a conservative band, and evaluated the matched gain at k = 0 (SEA independent of density), the case most favourable to any design.

4 Results

4.1 Validation optimism

Table 1 and Figure 2 give the cross-validated errors. The ordering of the models was the same under every scheme, but the absolute error was not. On B1, LOAO error exceeded random-split error by 26% for gradient boosting (0.157 to 0.198; 95% interval of the ratio 1.21–1.31), 31% for random forest, 14% for the sparse law (0.135 to 0.154; 1.09–1.19) and 11% for the baseline. On B2 the increases were 32%, 36%, 27% and 12%. Leave-one-paper-out fell between the two, so grouping by paper alone does not solve the problem.

Table 1 RMSE (log10 units), mean $\pm$ SD over 15 draws

Benchmark / Model	Random	LOPO	LOAO	Opt.
B1 Gibson–Ashby	0.292 $\pm$ 0.016	0.302 $\pm$ 0.018	0.326 $\pm$ 0.019	1.11
B1 Random forest	0.173 $\pm$ 0.012	0.192 $\pm$ 0.017	0.226 $\pm$ 0.020	1.31
B1 Grad. boosting	0.157 $\pm$ 0.012	0.171 $\pm$ 0.014	0.198 $\pm$ 0.021	1.26
B1 Sparse law	0.135 $\pm$ 0.010	0.141 $\pm$ 0.012	0.154 $\pm$ 0.019	1.14
B2 Gibson–Ashby	0.298 $\pm$ 0.030	0.308 $\pm$ 0.032	0.333 $\pm$ 0.037	1.12
B2 Random forest	0.179 $\pm$ 0.020	0.200 $\pm$ 0.023	0.244 $\pm$ 0.038	1.36
B2 Grad. boosting	0.164 $\pm$ 0.020	0.180 $\pm$ 0.024	0.217 $\pm$ 0.043	1.32
B2 Sparse law	0.144 $\pm$ 0.020	0.153 $\pm$ 0.020	0.186 $\pm$ 0.049	1.27

Note: LOPO: leave-one-paper-out; LOAO: leave-one-archetype-out; Opt.: LOAO/random ratio.

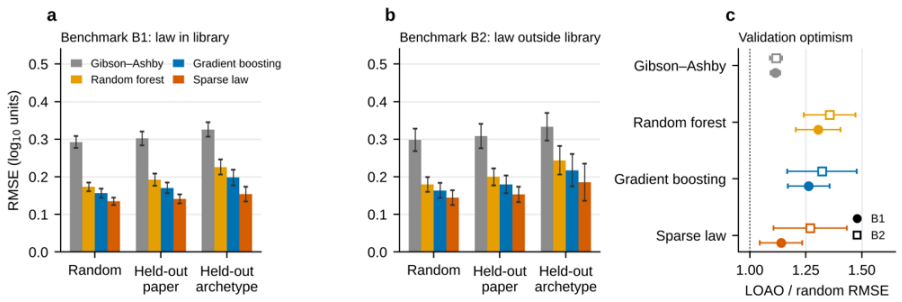


Figure 2 Cross-validated error and validation optimism. (a,b) RMSE under random, leave-one-paper-out and leave-one-archetype-out validation; bars show means and error bars the SD across 15 draws. (c) Ratio of held-out-archetype to random-split RMSE; filled circles B1, open squares B2.

The sparse law gave the lowest LOAO error in all 15 draws against gradient boosting and the baseline on both benchmarks, and against random forest in 15 of 15 (B1) and 14 of 15 (B2) draws (sign tests, $p < 0.001$). Its mean lead over gradient boosting was 0.044 log units on B1 (95% interval 0.036–0.052) and 0.032 on B2 (0.025–0.039). The lead persists on B2, where the library cannot express the true law. With about 240 noisy entries, then, the variance of a flexible model costs more than the bias of a misspecified library. Section 4.3 shows how this changes with data volume.

4.2 What controls the optimism

Optimism depends on the data as much as on the method (Figure 3a and 3b). With no archetype offsets and no paper noise, the sparse law showed none (ratio 1.00), as expected when held-out and training groups are exchangeable, whereas gradient boosting still lost 12% when an archetype was removed from its training set. Doubling the archetype offsets raised the ratio to 1.55 (sparse law) and 1.52 (boosting) at $\sigma_p = 0$. Paper-level noise lowered the ratio at large offsets, to 1.14 and 1.20 at $\sigma_p = 0.3$, because it inflates the random-split error in the denominator. It also raised the excess error itself. With no offsets, σ_p of 0.1 to 0.3 gave ratios of 1.07 to 1.11 for the sparse law, because a random split shares paper offsets between training and test entries whereas leave-one-archetype-out removes them.

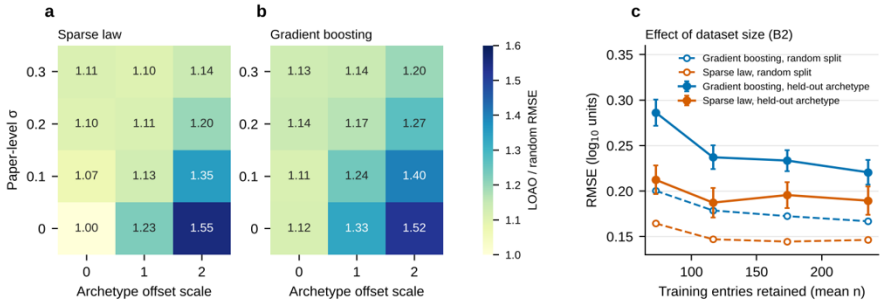


Figure 3 Drivers of validation optimism. (a,b) LOAO/random RMSE ratio over paper-level noise (rows) and archetype-offset scale (columns) for the sparse law and gradient boosting (B1 generator, eight draws per cell). (c) Learning curves on B2: RMSE against the number of training entries retained, held-out-archetype (solid) and random split (dashed); error bars show the standard error over 12 draws.

The additive model of Equation (4) captures this. Fitted to the twelve settings, it explained 91% (sparse law) and 97% (boosting) of the variance of the excess error, and 79% and 92% in leave-one-setting-out prediction, which corresponds to an RMSE of 0.08 and 0.05 in the predicted optimism ratio (Figure 4a). The extrapolation term c was indistinguishable from zero for the sparse law (0.0015 ± 0.0017) and reached 0.0050 ± 0.0013 for boosting; this is the part of optimism that belongs to the model class rather than the data. The archetype-offset coefficient was similar for the two models ($a = 0.0045$ and 0.0059), and the paper-level coefficient b was 0.17 and 0.21, so about a fifth of the paper-level variance reappears as excess error when whole archetypes are held out. Optimism therefore tracks between-archetype against within-archetype variability. It is largest where archetypes differ and measurements are clean, as in controlled experimental or finite-element datasets, which are exactly the datasets on which random splits mislead most.

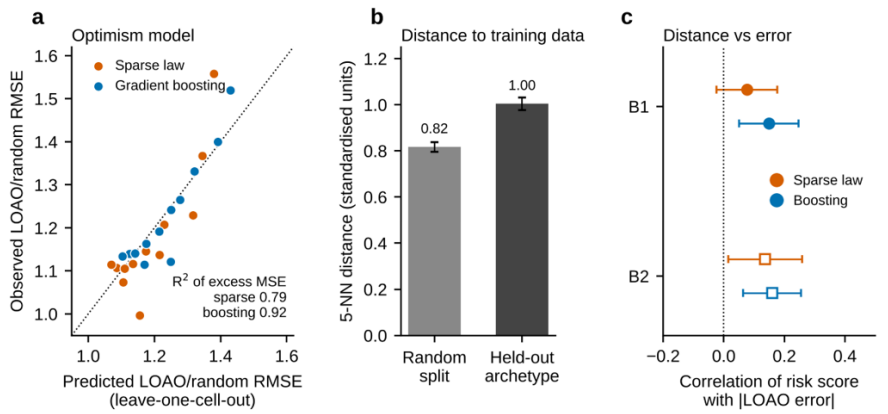


Figure 4 Additive optimism model and distance-based risk flags. (a) Optimism ratio predicted by the additive model (each point left out of the fit in turn) against the observed ratio for the twelve noise and offset settings; dotted line: identity. (b) Mean distance to the five nearest training entries (standardised descriptor units) for random-split test entries and for held-out archetypes; error bars show the SD over 15 draws. (c) Correlation between this distance and the absolute LOAO error within the held-out set; error bars show the SD over 15 draws; filled circles B1, open squares B2.

4.3 Dataset size

The learning curve (Figure 3c) separates the two models. Gradient boosting kept improving with more data: on B2, its LOAO RMSE fell from 0.286 at 73 entries to 0.221 at 236, and its optimism from 1.45 to 1.31. The sparse law was already near its plateau at 116 entries (0.187 against 0.189 at 236), and its optimism stayed close to 1.3 throughout. As a result, its lead over boosting narrowed from 0.074 to 0.031 log units on B2 and from 0.093 to 0.045 on B1.

4.4 Conformal coverage under archetype shift

At a nominal 90%, random-fold calibration covered 84.5% (sparse law) and 81.2% (boosting) of held-out entries on B1, and 80.1% and 79.6% on B2 (Table 2, Figure 5a). Calibrating on leave-one-archetype-out residuals raised coverage to 90.3% and 90.6% on B1 and to 88.5% and 89.7% on B2, with much less spread between draws. The paired gains were 5.7 (3.7–7.7) and 9.3 (7.2–11.5) percentage points on B1, and 8.4 (5.6–11.1) and 10.0 (7.4–12.6) on B2, for the sparse law and boosting respectively (95% intervals in brackets). The reliability diagram (Figure 5b) shows that the failure of random-fold calibration is systematic. On B2 it under-covered at every nominal level from 50 to 95%, by 7 to 13 points, while archetype calibration stayed within two points of nominal for both models (within three on B1). The cost is wider intervals, by 18% (sparse law, B1) to 34% (boosting, B2). At equal coverage, the sparse law's intervals were also the narrower ones (0.53 against 0.68 $\log_{10}$ units on B1).

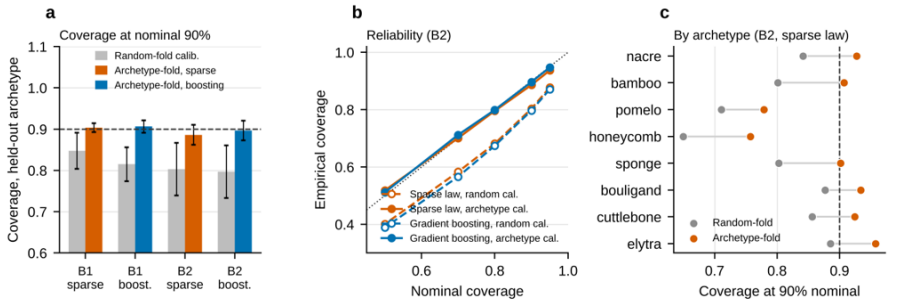


Figure 5 Conformal coverage on held-out archetypes. (a) Coverage at a nominal 90%; bars show means and error bars the SD over 15 draws. (b) Reliability diagram on B2 (dotted line: perfect calibration). (c) Coverage by held-out archetype for the sparse law on B2, random-fold versus archetype-fold calibration.

Coverage is only a marginal guarantee, and individual archetypes can still fall short (Figure 5c). On B2, archetype calibration lifted the sparse law's coverage for honeycomb from 0.65 to 0.76 and for pomelo peel from 0.71 to 0.78, yet these remained the weakest cases. They have the lowest stretching fractions, so whenever they are held out they sit at the edge of the training range. Intervals for such an archetype are optimistic and should be flagged as such.

Table 2 Coverage of nominal 90% intervals on held-out archetypes and mean interval width (log10 units), mean over 15 draws (SD of coverage in brackets)

Model, calibration	B1 coverage	B1 width	B2 coverage	B2 width
Sparse, random folds	0.845 (0.041)	0.45	0.801 (0.066)	0.48
Sparse, archetype folds	0.903 (0.011)	0.53	0.885 (0.026)	0.62
Boosting, random folds	0.812 (0.043)	0.52	0.796 (0.064)	0.55
Boosting, archetype folds	0.906 (0.013)	0.68	0.897 (0.022)	0.74

4.5 Distance-based risk flags

Applicability-domain scores are the usual way to flag unreliable predictions. Held-out archetypes lay 23% farther from the training data than random-split test entries (mean five-neighbour distance 1.00 against 0.82 standardised units; ratio 1.23, 95% interval 1.21–1.25), so the score does detect that a shift has occurred (Figure 4b). Within the held-out set, however, it ranked the errors poorly (Figure 4c). Its correlation with the absolute LOAO error was 0.08 $\pm$ 0.10 for the sparse law and 0.15 $\pm$ 0.10 for boosting on B1, and 0.14 $\pm$ 0.12 and 0.16 $\pm$ 0.10 on B2. In these benchmarks the error is driven by archetype-level offsets, not descriptor distance, so the score cannot substitute for archetype-calibrated intervals.

4.6 Which terms of the law are identifiable

Two terms were selected in all 120 refits of each benchmark: the density term and the stretching–density interaction (Figure 6a). The base modulus is retained by construction. Together they form the Gibson–Ashby backbone with a topology-dependent exponent. The fitted exponent at the bending limit was 2.06 $\pm$ 0.25 (B1) and 2.11 $\pm$ 0.27 (B2) against a generating value of 2, and the stretching coefficient was -1.00 $\pm$ 0.17 and -1.01 $\pm$ 0.19 against -1. Hierarchy was selected in 90% and 85% of refits and strain rate in 88% and 78%. At a selection frequency of 0.9, the upper end of the cut-offs recommended for stability selection [19], hierarchy reaches the threshold on B1 but not on B2. Every other term appeared in 38 to 72% of refits, and two quadratic terms that do not exist in B1 were selected in 60 to 66% of them.

In B1, where the generator terms are known, selection frequency only partly separated the four secondary terms that exist (44 to 90%) from the seven that do not (38 to 66%). The area under the ROC curve was 0.82, and the gradient–stretching term that does exist in the generator ranked below five absent terms. Secondary terms should therefore be reported with their selection frequency, and interpreted mechanistically only when they prove stable.

The ablation points the same way (Figure 6b). Removing the stretching fraction raised held-out RMSE of gradient boosting by 64% on B1 and 54% on B2; removing hierarchy raised it by 10% and 6%, and removing strain rate by 2% on both. Removing the gradient index actually lowered RMSE, by 5% on B1 and 1% on B2, so this descriptor did not transfer between archetypes even though it enters the B1 generator. Stretching versus bending dominance carries the physics in these benchmarks, and it is the descriptor that transfers.

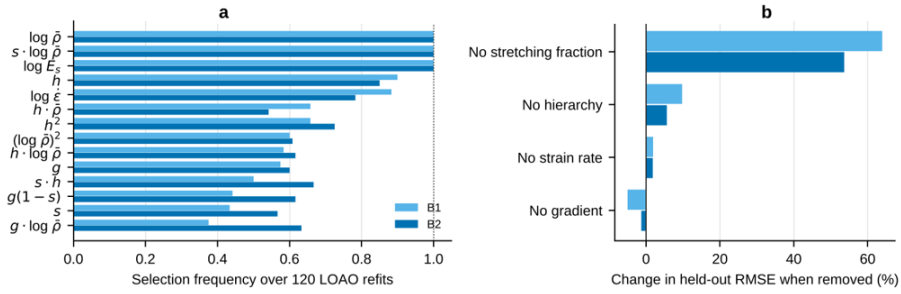


Figure 6 Identifiability under archetype shift. (a) Fraction of the 120 leave-one-archetype-out refits in which each candidate term was retained by the sparse law. (b) Increase in held-out RMSE of gradient boosting when one descriptor is removed. Light blue: B1; dark blue: B2.

4.7 Density-matched audit of a published study

The seven bio-inspired honeycombs weighed 1.24 to 3.25 times as much as the regular hexagon and reported SEA gains of 22 to 47% (Table 3, Figure 7). For the snail, spider, hierarchical, pomelo-peel and grass-stem designs the break-even exponents were 0.18 to 0.38,

so the gain disappears if SEA depends on density more strongly than that. At the thin-wall exponent of $k = 2/3$, the matched ratios for those five fell to 0.59–0.78. The bamboo-inspired design sat at break-even ($k^* = 0.68$), and the wavy design reached break-even only at $k^* = 0.99$, because its mass rose by just 24%.

Table 3 Density-matched audit of the eight designs in Solak [22]

Design	S	D	k^*	k^* range	Share	Matched, $k = 0.19$	Matched, $k = 2/3$
Spider	1.29	2.69	0.26	0.05–0.47	0.73	1.07	0.67
Snail	1.22	2.94	0.18	−0.01–0.38	1.05	0.99	0.59
Wavy	1.24	1.24	0.99	0.02–1.96	0.19	1.19	1.07
Bamboo	1.27	1.42	0.68	0.08–1.28	0.28	1.19	1.00
Pomelo peel	1.47	3.25	0.33	0.15–0.51	0.58	1.18	0.67
Grass stem	1.40	2.42	0.38	0.15–0.62	0.50	1.18	0.78
Hierarchical	1.38	3.25	0.27	0.10–0.45	0.70	1.10	0.63

Note: S: SEA ratio to the regular hexagon; D: mass ratio; k^*: break-even exponent (range for ±23.25% error in S); Share: fraction of ln S explained by mass at the empirical exponent 0.19; last two columns: $S \cdot D^{-k}$ at $k = 0.19$ and $k = 2/3$.

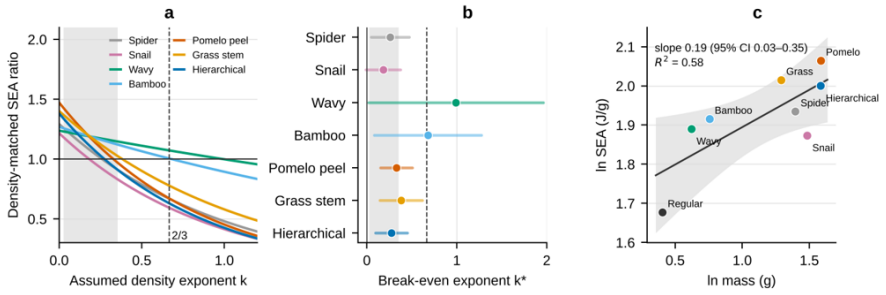


Figure 7 Audit of Solak [22]. (a) Density-matched SEA ratio as a function of the assumed density exponent; the grey band is the 95% CI of the empirical exponent and the dashed line marks 2/3. (b) Break-even exponent k^* with the range for ±23.25% model error. (c) ln SEA against ln mass for the eight designs with the fitted trend and its 95% confidence band.

The empirical slope gives a milder picture. Across the eight designs, ln SEA rose with ln mass with a slope of 0.19 (95% CI 0.03–0.35, $R^2 = 0.58$, $p = 0.03$), and leaving out any one design moved it only between 0.10 and 0.22. Absorbed energy itself was almost proportional to mass (slope 1.19, $R^2 = 0.98$). At $k = 0.19$ the matched ratio stays above one for six of the seven designs (1.07 to 1.19) and falls to 0.99 for the snail design. Mass accounts for more than half of the log gain for the snail (105%), spider (73%), hierarchical (70%) and pomelo-peel (58%) designs, for half of it in the grass-stem design, and for less than 30% in the wavy and bamboo designs, which keep matched gains of 19%. Because the pooled slope mixes topology with density, it summarises the data but is not a clean exponent.

What survives the audit depends on k, and the model uncertainty is as large as several of the gains. With the ±23.25% allowance, the gain remaining at $k = 0$ is 5% (spider), -1% (snail), 0.4% (wavy), 3% (bamboo), 12% (hierarchical), 14% (grass stem) and 20% (pomelo peel). Only three designs therefore exceed 10% even in the case most favourable to them, and for those three mass still explains 50 to 70% of the log gain at the empirical exponent. One conclusion holds across all the exponents considered: SEA comparisons in this literature cannot be interpreted without mass or relative density. Where a study fixes mass or relative density by design, as in Liu et al. (2025) [23] and He et al. (2025) [24], the question does not arise.

5 Discussion

Validation optimism is not a fixed property of a method. It combines a data component, set by archetype-offset and paper-level variance, with a model component, which is zero for the sparse law in the absence of offsets and about 12% for gradient boosting. Because the data component is judged against the measurement scatter in the denominator, a pooled R^2 cannot be compared across datasets unless the regime is stated.

Uncertainty has to be calibrated for the way it will be used. A designer who needs a bound on stiffness for a new architecture needs coverage under archetype shift, and random-fold intervals

overpromise by about ten points. Archetype-calibrated intervals are wider, and that extra width is the real price of covering a new archetype. A distance-based risk score is no substitute: it detects that a shift has occurred, but not where the error will be large. And because the guarantee is marginal, an archetype at the edge of the training range, such as honeycomb in these benchmarks, still warrants caution.

A recovered law is best presented as a backbone with selection frequencies. The Gibson–Ashby term with a topology-dependent exponent was found every time, hierarchy was borderline, and secondary terms were only moderately more likely to be selected when present than when absent. The audit of the published study suggests a matching rule for experiments and simulations alike. Report mass or relative density next to every specific metric and, where a gain is claimed, show a density-scaled reference, since a gain that mostly reflects added material is weak evidence for a better architecture. Table 4 collects these requirements as a minimum reporting standard.

Table 4 Minimum reporting for data-driven design rules for lightweight architectures

Item	What to report
Validation	Error under leave-one-archetype-out validation next to the random-split error, and their ratio.
Uncertainty	Intervals calibrated on leave-one-archetype-out residuals, with coverage reported per archetype.
Discovered laws	Selection frequencies under grouped refitting; mechanistic interpretation only for terms at or above 0.9.
Specific metrics	Mass or relative density beside every specific value, and a density-scaled reference or the break-even – exponent for each claimed gain.
Data extraction	Documented search, two independent extractors on at least a fifth of the entries, and the agreement rate.

Taken together, the four steps in Figure 1 change what a reader can conclude from a data-driven design study. An accuracy figure arrives with its archetype-shift penalty, a stiffness bound with honest coverage, a discovered law with its stable core, and an efficiency gain with its density-matched value.

6 Limitations and Outlook

The validation, conformal and identifiability results come from simulation. Both generators contain a Gibson–Ashby backbone, which makes recovery easier than it would be with real data, and the optimism and coverage gaps depend on the offsets and noise chosen; the sweeps in Sections 4.2 and 4.3 map that dependence. The additive optimism model was fitted to twelve settings of one generator, so its coefficients are specific to it and what transfers is its additive structure. Real datasets bring missing descriptors, inconsistent definitions of relative density, nonlinear collapse and publication bias. Hyperparameters were fixed, so the lead of the sparse law applies to untuned ensembles, and the coverage results are empirical for seven calibration archetypes, with no finite-sample guarantee.

The real-data audit rests on eight entries from one finite-element study with a single run per design. Most of the other open-access papers on bio-inspired honeycombs, lattices and tubes that we screened report energy absorption or stiffness only in figures, or omit mass and density, so they could not be extracted reliably. The decisive test is a dataset of 200 to 400 entries from a documented search, extracted independently by two people for at least a fifth of the entries with a reported agreement rate, and processed with the pipeline presented here. If the optimism ratio for held-out archetypes on such a dataset came out close to one, that would show the descriptors capture the transferable physics.

7 Conclusions

Random-split validation overstates how well design rules for bio-inspired architectures transfer, and the overstatement can be measured, predicted and corrected. On controlled benchmarks, holding out whole archetypes raised prediction error by 11–36%. The excess error was additive in archetype-offset and paper-level variance, with an additional extrapolation cost for flexible models. Conformal intervals calibrated to the same shift restored coverage to within three points of nominal at the cost of 18–34% wider intervals; descriptor distance did not locate the failures, and archetypes at the edge of the training range remained under-covered. A sparse extension of the Gibson–Ashby law outperformed gradient boosting in every draw at this sample size and yielded a stable backbone of density and stretching–density terms, while secondary terms were not identifiable. In the published study we audited, much of the reported energy-absorption advantage of bio-inspired honeycombs coincided with added mass,

and only three of seven gains exceeded 10% after allowing for model error, even if SEA were density-independent. Held-out-archetype validation and archetype-calibrated intervals should therefore become standard for data-driven design rules, and mass or relative density should accompany every specific metric. The next step is to apply the protocol to a larger, documented extraction of the literature.

Data Availability Statement

The data supporting the findings of this study are available from the author upon reasonable request.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Meyers MA, Chen P-Y, Lin AY-M, et al. Biological materials: Structure and mechanical properties. Progress in Materials Science. 2008, 53(1): 1-206. https://doi.org/10.1016/j.pmatsci.2007.05.002

[2] Wegst UGK, Bai H, Saiz E, et al. Bioinspired structural materials. Nature Materials. 2014, 14(1): 23-36. https://doi.org/10.1038/nmat4089

[3] Bessa MA, Bostanabad R, Liu Z, et al. A framework for data-driven analysis of materials under uncertainty: Countering the curse of dimensionality. Computer Methods in Applied Mechanics and Engineering. 2017, 320: 633-667. https://doi.org/10.1016/j.cma.2017.03.037

[4] Bastek J-H, Kumar S, Telgen B, et al. Inverting the structure–property map of truss metamaterials by deep learning. Proceedings of the National Academy of Sciences. 2021, 119(1). https://doi.org/10.1073/pnas.2111505119

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

[6] Roberts DR, Bahn V, Ciuti S, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017, 40(8): 913-929. https://doi.org/10.1111/ecog.02881

[7] Meredig B, Antono E, Church C, et al. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. Molecular Systems Design & Engineering. 2018, 3(5): 819-825. https://doi.org/10.1039/c8me00012c

[8] Ploton P, Mortier F, Réjou-Méchain M, et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature Communications. 2020, 11(1). https://doi.org/10.1038/s41467-020-18321-y

[9] Meyer H, Pebesma E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods in Ecology and Evolution. 2021, 12(9): 1620-1633. https://doi.org/10.1111/2041-210x.13650

[10] Gibson LJ, Ashby MF. Cellular Solids. Published online May 1, 1997. https://doi.org/10.1017/cbo9781139878326

[11] Vovk V, Gammerman A, Shafer, G. Algorithmic learning in a random world. Springer, New York, 2005. https://doi.org/10.1007/b106715

[12] Angelopoulos AN, Bates S. Conformal Prediction: A Gentle Introduction. Foundations and Trends® in Machine Learning. 2023, 16(4): 494-591. https://doi.org/10.1561/2200000101

[13] Tibshirani RJ, Barber RF, Candès EJ, et al. (2019). Conformal prediction under covariate shift. Advances in Neural Information Processing Systems. 2019, 32: 2530-2540. https://dl.acm.org/doi/abs/10.5555/3454287.3454514

[14] Barber RF, Candès EJ, Ramdas A, et al. Predictive inference with the jackknife+. The Annals of Statistics. 2021, 49(1): 486-507. https://doi.org/10.1214/20-aos1965

[15] Lei J, G'Sell M, Rinaldo A, et al. Distribution-Free Predictive Inference for Regression. Journal of the American Statistical Association. 2018, 113(523): 1094-1111. https://doi.org/10.1080/01621459.2017.1307116

[16] Dunn R, Wasserman L, Ramdas A. Distribution-Free Prediction Sets for Two-Layer Hierarchical Models. Journal of the American Statistical Association. 2022, 118(544): 2491-2502. https://doi.org/10.1080/01621459.2022.2060112

[17] Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proceedings of the National Academy of Sciences. 2016, 113(15): 3932-3937. https://doi.org/10.1073/pnas.1517384113

[18] Cranmer M. Interpretable machine learning for science with PySR and SymbolicRegression.jl. arXiv preprint, 2023. https://doi.org/10.48550/arXiv.2305.01582

[19] Meinshausen N, Bühlmann P. Stability Selection. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2010, 72(4): 417-473. https://doi.org/10.1111/j.1467-9868.2010.00740.x

[20] Ashby MF. The properties of foams and lattices. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2005, 364(1838): 15-30. https://doi.org/10.1098/rsta.2005.1678

[21] Wierzbicki T. Crushing analysis of metal honeycombs. International Journal of Impact Engineering. 1983, 1(2): 157-174. https://doi.org/10.1016/0734-743x(83)90004-0

[22] Solak A. Numerical Investigation of Compression Behavior of Bio-Inspired Honeycomb Designs. Firat University Journal of Experimental and Computational Engineering. 2026, 5(1): 250-261. https://doi.org/10.62520/fujece.1796508

[23] Liu C, Tsang H-H, Xu S, et al. Lateral dynamic crushing of skeletal muscle-inspired hierarchical celled structures. Bioinspiration & Biomimetics. 2025, 20(4): 46003. https://doi.org/10.1088/1748-3190/addc25

[24] He Z, Yang Y, Hu L, et al. On the Design of Bionic Hierarchical H-Type Whip Restraints for Nuclear Power Plants. Applied Sciences. 2025, 15(10): 5507. https://doi.org/10.3390/app15105507