

Probabilistic failure prediction for early-stage design of novel composite wing structures via heteroscedastic Bayesian learning

Runze Li ^{*}, Mário Miranda , Silvestre T. Pinho

Department of Aeronautics, Imperial College London, Exhibition Rd, South Kensington, London, SW7 2AZ, UK

ARTICLE INFO

Communicated by Dr Kwok Kawai

Keywords:

Uncertainty quantification
Composite wing
Failure envelope
Bayesian neural networks
Early design stages
Novel structural concepts

ABSTRACT

Composite wing structures increasingly rely on advanced materials and novel configurations to meet future aviation performance and sustainability targets. However, incorporating failure constraints in early-stage design remains challenging, particularly for unconventional wing architectures beyond the traditional design space. This is mainly due to the high dimensionality of composite layups, which greatly increases the need for training data and computational resources in predictive models and optimisation design. This paper proposes a probabilistic framework for rapid failure envelope prediction of composite wing structural elements under varying geometries and loading conditions. Instead of explicitly modelling high-dimensional stacking sequences, the composite layup is treated as an unobserved stochastic variable, and the failure response is formulated as a conditional probability distribution over the population of engineering-rule-compliant layups. A heteroscedastic variational Bayesian model is developed to capture both the mean response and input-dependent uncertainty induced by layup variability. The proposed approach directly predicts probabilistic best-case and worst-case failure envelopes without requiring explicit layup optimisation, enabling efficient integration of failure constraints into early-stage design. The framework is demonstrated on a Z-beam wingbox concept, with failure indices evaluated using the LaRC05 criteria and a high-fidelity dataset. Results show that the model provides accurate and efficient predictions of the probability distributions of failure indices for different wingbox geometries under varying load conditions. The predicted envelopes offer valuable insights for structural assessment and highlight the need for further layup optimisation.

1. Introduction

1.1. Motivation of incorporating failure constraints in early design stages

Composite materials enable lightweight and high-performance aircraft structures due to their superior strength-to-weight ratios and design flexibility. In the pursuit of net-zero aviation by mid-century, emerging concepts such as high-aspect-ratio wings [1], novel structural configurations [2,3], and aeroelastic tailoring [4,5] are being actively explored to reduce weight and aerodynamic drag. However, the structural complexity associated with composite materials leads to intricate and highly coupled failure mechanisms, which are difficult to assess with traditional early-stage design tools [6,7]. As a result, failure considerations are often deferred to later design stages, where detailed analysis tools are available [4,8]. This practice can lead to costly redesign cycles and may restrict the exploration of innovative structural concepts.

Therefore, incorporating failure constraints from the early design stages is essential for enabling reliable design decisions. It not only mit-

igates the risk of late-stage design changes, but also facilitates a broader and more confident assessment of candidate designs, supporting the development of novel configurations while maintaining structural integrity and safety margins.

1.2. Challenges of incorporating composite failure constraints

There are two major challenges in incorporating failure constraints into the optimisation design of composite wings. The first is the **curse of dimensionality** arising from stacking sequence (composite layup). Optimising all composite components simultaneously leads to a highly complex problem [4,8,9]. The second challenge is the **lack of accurate and fast failure prediction tools**, especially for novel structural concepts. The main design strategy generally adopted by industry is to size structural components based on simplified and conservative assumptions, which predict the lower bound of structural strength. While effective for ensuring safety, this approach inevitably restricts the exploration of innovative structural designs [10].

^{*} Corresponding author.

E-mail address: runze.li@imperial.ac.uk (R. Li).

<https://doi.org/10.1016/j.ast.2026.113816>

Received 21 April 2026; Received in revised form 1 August 2026; Accepted 15 September 2026

Available online 17 September 2026

1270-9638/© 2026 The Author(s). Published by Elsevier Masson SAS. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1.3. Current approaches to failure consideration in early design stages

There have been studies that attempt to incorporate failure constraints into the optimisation design of composite wings; however, most of them fail to simultaneously address the aforementioned challenges. Existing approaches can be broadly categorised into four classes. The first class directly optimises simplified problems, either structures with a limited number of plies (e.g., six plies [11]) or models based on simplified theories [6]. The second class performs optimisation under predefined or fixed layup configurations [10,12,13]. The third class reduces the design space by using ply thickness (super layers) as design variables, followed by predefined or optimised layups for failure evaluation [14–16]. The fourth class employs lamination parameters as design variables, which yields a reduced and well-conditioned optimisation problem, with simplified and conservative criteria for failure evaluation [17–20].

1.4. Limitations in current fast failure prediction models for large composite structures

More recently, machine learning-based fast prediction models (surrogate models) have emerged as a promising approach for representing failure constraints in optimisation design. Some studies relied on fixed layups [21–23], predefined layup patterns [24,25], or fixed geometries [22,23,26]. More advanced models use stacking sequences and geometric parameters directly as inputs, but this increases training complexity and does not reduce the dimensionality of the design space for optimisation [27]. Beyond machine-learning surrogates, earlier two-level approaches predict the upper bound of structural strength using super layers and local loads as inputs, where the upper bound is obtained from separate optimisation procedures [8].

1.5. Our proposed framework

In this study, we propose a Bayesian framework for the rapid evaluation of failure behaviour in large composite structures under varying geometries and loading conditions. The framework predicts the best-case and worst-case failure envelopes probabilistically, that is, optimistic and conservative estimates of the failure boundary across different layup configurations, without relying on pre-defined layups, conservative simplifications, or additional sub-level optimisation procedures. By directly characterising the distribution of failure responses, the proposed model enables the integration of failure constraints into early-stage composite wing design while retaining a low-dimensional and tractable design space. This allows designers to assess conservative and reliable configurations, as well as to identify the potential performance improvements achievable through layup optimisation. Overall, the framework provides both robust safety assessment and enhanced design exploration capability, making it particularly suitable for early-stage design and for supporting the development of next-generation composite wing concepts.

An important characteristic of this problem is that the uncertainty of the failure response is expected to be input-dependent. The sensitivity of the failure behaviour to the layup configuration varies with the structural geometry and the loading conditions: for some geometries and load cases the failure response changes strongly with the layup, whereas for others it is comparatively insensitive. The variance of the failure response over the candidate layup population therefore varies across the design space, which calls for a heteroscedastic surrogate model that can capture this input-dependent uncertainty.

The objective of this work is to develop a probabilistic surrogate framework for predicting failure-index distributions and failure envelopes of composite wing structural elements under varying layup configurations. The main contributions of this work are summarised as follows.

1. A Bayesian framework for failure prediction of large composite structures, in which the composite layup is treated as an unobserved

stochastic variable. The framework predicts the distribution of failure responses conditional on an assumed candidate layup population, and it keeps the early-stage design space low-dimensional because explicit stacking-sequence design variables are avoided.

2. A heteroscedastic variational Bayesian last layer (HVBLL) surrogate model that captures the input-dependent uncertainty of the failure response. The model is validated on synthetic datasets with analytically known noise levels and on further datasets, namely a heteroscedastic image regression benchmark and an open-hole compression database of composite laminates (Appendix A).
3. The application of the proposed framework to a novel Z-beam representative structural element for composite wings. Probabilistic failure envelopes are constructed from finite-element failure index maps compressed by principal component analysis, and their use in early-stage failure assessment is illustrated on a representative gust-response strain trajectory.

The proposed framework is intended for early-stage design assessment, where it serves as a rapid screening tool for candidate designs. It does not replace detailed laminate optimisation or high-fidelity failure analysis at later design stages. In particular, configurations that lie near the predicted envelope boundaries still require detailed analysis and layup optimisation in the subsequent design stages.

The rest of the paper is structured as follows. The Bayesian framework and mathematical formulation are introduced in Section 2. The algorithm and verification of the proposed heteroscedastic Bayesian model are presented in Section 3. The proposed method is then applied to a novel composite wing structural concept in Section 4, and conclusions are drawn in Section 5.

2. Methodology

2.1. Classical failure analysis procedure and its limitations

A wingbox structure based on the Z-beam concept proposed by Airbus [28] is considered in this study. The structure features a series of span-wise zigzag members in the wing tip region, as illustrated in Fig. 1. The primary load-carrying components include spars and ribs (black lines), stringers (grey lines), and zigzag Z-beams (red lines). This structural configuration enables aeroelastic tailoring, where both the Z-beam geometry and composite layups act as design variables to control bending, torsional, and coupling stiffness characteristics [29].

The classical multi-scale finite element analysis (FEA) procedure for failure evaluation is illustrated in Fig. 2. First, a global finite element model of the aircraft is constructed and used for aeroelastic simulations. As shown in Fig. 1, the wing tip region is discretised into a sequence of Z-beam Representative Structural Elements (RSEs), which serve as the fundamental units for structural analysis. The homogenised stiffness matrix of each RSE can be obtained, providing a reduced-order representation of the local structural behaviour. Consequently, the Z-beam structure in the wing tip region can be approximated as a one-dimensional sequence of RSEs with varying geometric parameters Θ and composite layups Γ , enabling efficient structural modelling at the global level [31]. From this model, the strain state Δ of each Z-beam RSE is extracted, including axial strain ϵ_z , torsional curvature κ_z , and bending curvatures κ_x and κ_y .

Then, critical failure hotspots within the RSE are identified, and detailed local models are constructed using a global-local analysis approach. The maximum failure index (FI) of each RSE is then computed. By repeating this process over a range of strain states, a failure index map is generated. The isoline corresponding to $FI = 1$ defines the failure envelope of the RSE parametrised by (Θ, Γ) . In this study, this high-fidelity procedure is employed to generate training and validation data; further details can be found in [30].

Despite its high accuracy, this classical procedure is computationally expensive and therefore unsuitable for early-stage design. In addition, the dependence on composite layup Γ introduces a high-dimensional

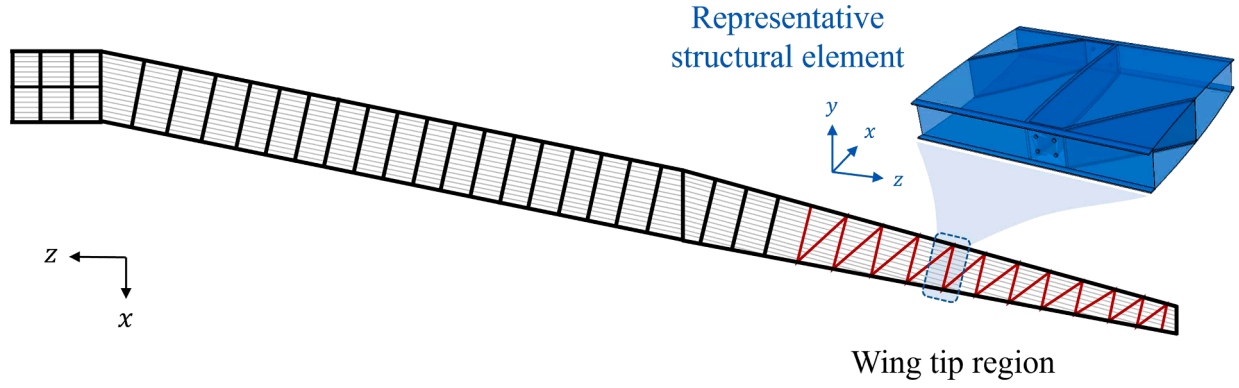


Fig. 1. Z-beam structures in the wing tip region of a high-aspect-ratio wing [30].

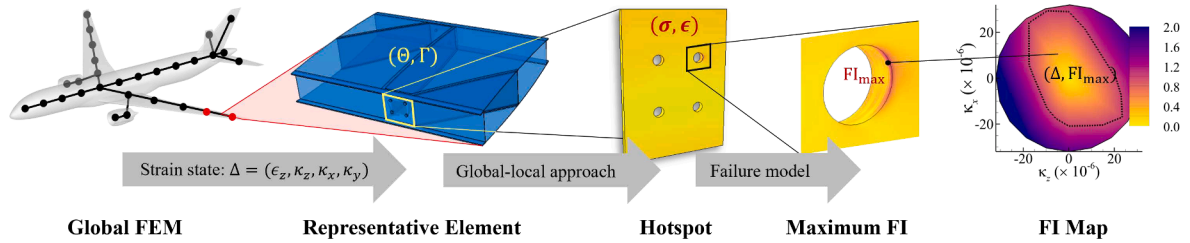


Fig. 2. Classical multi-scale finite element analysis procedure for evaluating the failure behaviour of a representative structural element in civil aircraft [30]. The strain state (Δ) is extracted from the global FEM model and applied to the detailed RSE model for failure analysis. Repeating this process across strain states yields a failure index map, representing the maximum FI of the RSE under any loading condition.

design space, which further limits its applicability in optimisation and rapid design exploration.

2.2. Bayesian framework for failure prediction of large composite structures

Building on existing industrial design strategies introduced in Section 1.3, we propose a novel Bayesian framework for fast failure prediction of large composite structures, as shown in Fig. 3, addressing the two key challenges of incorporating failure constraints into aircraft optimisation design:

1. **Curse of dimensionality.** Reducing the design space dimension is essential for optimisation. This is often achieved by replacing the composite layup Γ with super layers (ply thickness), lamination parameters, or even excluding the composite layup entirely from the design space. However, such simplifications inevitably discard structural information and introduce errors or uncertainty into the prediction.
2. **Fast and comprehensive failure prediction.** Previous approaches that omit composite layup in optimisation rely on additional assumptions, simplifications, or sub-level optimisations. These either lead to conservative predictions or search for the optimal layup with maximum strength at a lower level. In contrast, our Bayesian model directly predicts both limits probabilistically, as the best-case and worst-case failure envelopes, offering comprehensive insight into structural behaviour across different layups.

Novelty and contributions: To the best of our knowledge, this is the first work that treats the composite layup as an unobservable stochastic variable and accounts for it as a source of uncertainty, rather than as a quantity to be designed or prescribed, so that the failure envelopes of composite wing structural elements are predicted probabilistically through heteroscedastic Bayesian surrogate modelling. Unlike previous approaches that replace the composite layup with simplified substitutes such as super layers or lamination parameters, the proposed framework

removes the stacking-sequence variables from the design space altogether and accounts for their influence probabilistically, which reduces the effective design-space dimensionality to the geometric parameters during early-stage design. The main strength of this framework is therefore not only its computational speed, but its ability to provide a probabilistic description of the failure behaviour at the early design stage, when the detailed laminate layup is not yet fixed.

2.3. Modelling as partially observable systems

2.3.1. Mathematical formulation: applying uncertainty quantification to deterministic systems

The proposed Bayesian framework and the previous studies reduce the design space dimension by neglecting part or all of the information about the composite layup, which introduces uncertainty into the system. This problem can be mathematically described as partially observable systems:

$$\mathbf{y} = F(\mathbf{x}, \mathbf{e}), \mathbf{e} \sim p(\mathbf{e}), \quad (1)$$

where $\mathbf{y}$ are the quantities of interest, e.g., the maximum failure index in the Z-beam representative element; F is the accurate and deterministic description of the system, e.g., the FEM model of the composite structure; $\mathbf{e}$ are the unobservable random variables, e.g., the composite layups. The probability distribution $p(\mathbf{e})$ is usually unknown.

In uncertainty quantification theory, these unobservable random variables can be considered a form of input uncertainty, which ultimately contributes to the **aleatoric uncertainty** within the system [32]. Therefore, the prediction of $\mathbf{y}$ becomes the probabilistic modelling of the conditional probability distribution $p(\mathbf{y}|\mathbf{x})$.

2.3.2. Assumptions and scope

We make two assumptions in quantifying uncertainty introduced by unobservable variables:

1. **Assumption 1:** We assume the existence of a mapping from independent random variables $\delta \sim \mathcal{N}(0, I)$ to $\mathbf{e}$, such that $\mathbf{e} - \mathbf{e}_0 = G(\delta; \gamma)$,

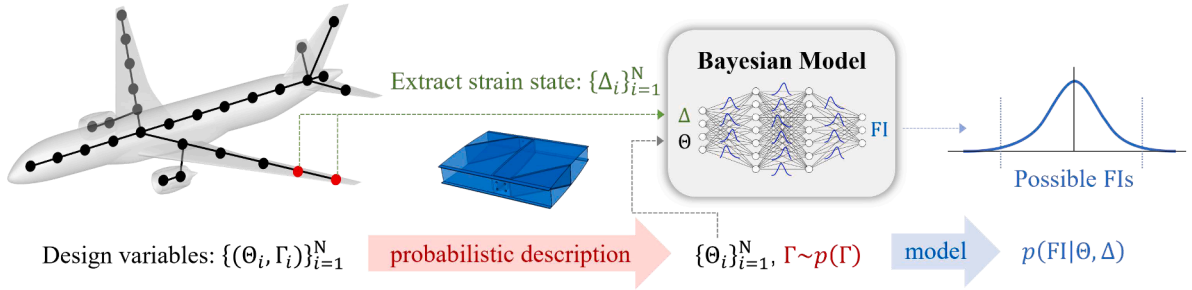


Fig. 3. Bayesian framework for failure prediction of large composite structures. The design variables are reduced to geometric parameters to simplify the optimisation problem. The composite layup is treated as an unobservable random variable with an unknown distribution, and a Bayesian model is used to estimate the probability distribution of the maximum failure index in the representative element. The unobserved layup variable is not modelled explicitly: the Bayesian model is trained on a dataset sampled from a layup population that satisfies engineering design guidelines, and the variability of the layup within this population is absorbed into the probabilistic prediction. The predicted distribution is therefore the conditional distribution of the maximum failure index given the geometric parameters and the strain state; its mean and spread represent the expected failure index and the layup-induced uncertainty over this layup population. The framework provides probabilistic failure envelopes for optimisation.

where $\mathbf{e}_0$ is the nominal value and γ are the parameters of the mapping. This assumption is commonly valid and underlies most generative models [33].

2. **Assumption 2:** In the absence of a better alternative, we assume that $\mathbf{y}$ can be **linearly approximated**, i.e.,

$$\mathbf{y} \approx F(\mathbf{x}, \mathbf{e}_0) + \frac{\partial F}{\partial \mathbf{e}} \frac{\partial G}{\partial \delta} \bigg|_{\mathbf{x}, \delta=0} \delta := f_1(\mathbf{x}) + f_2(\mathbf{x})\delta. \quad (2)$$

Under this approximation, the conditional probability becomes Gaussian: $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(f_1, f_2 f_2^T)$. This corresponds to a heteroscedastic problem, meaning that the variance of $\mathbf{y}$ depends on the input. For the discrete layup variable considered in this work, $\mathbf{e}$ is understood as a continuous latent representation of the layup population, and the linearisation in Assumption 2 is a first-order moment-matching ansatz: it is the simplest construction that yields an input-dependent mean and variance of the conditional distribution. Its adequacy is not assumed a priori but is assessed empirically through the calibration of the prediction intervals (Appendix B).

The scope of the framework follows from these assumptions. The Gaussian form of the conditional distribution is a consequence of the local linear approximation in Assumption 2; it is the maximum-entropy distribution for a given mean and variance and is therefore fully determined by the two quantities that the model predicts, but it may not capture strongly nonlinear or multimodal failure-index distributions, for example in the presence of distinctly different layup families. The predicted distribution is conditional on the layup population sampled in the training dataset, which consists of engineering-rule-compliant nominal and randomised layups, and does not cover all possible laminate configurations. The Gaussian predictive distribution is also examined empirically in the Z-beam application through the calibration of the prediction intervals, which supports its use in the model output space (Appendix B). These limitations are discussed in detail in Section 4.6.

The partially observable formulation therefore calls for a surrogate model that represents a Gaussian conditional distribution with an input-dependent variance, provides Bayesian uncertainty estimates, and remains computationally efficient and scalable to the limited and costly training datasets typical of engineering applications. The following section reviews existing uncertainty quantification methods against these requirements and introduces the Heteroscedastic Variational Bayesian Last Layers (HVBLL) model, which combines an efficient Bayesian last-layer treatment with an input-dependent noise model and is adopted for the probabilistic modelling of the conditional distribution $p(\text{FI}|\Theta, \Delta)$ of the maximum failure index in the Z-beam representative elements. In the surrogate, the geometric parameters Θ are the model input, whereas the dependence on the strain state Δ is carried by the output: the model predicts the whole failure index map over Δ for a given Θ .

3. Heteroscedastic uncertainty quantification

3.1. Existing methods

Uncertainty quantification is crucial for reliable optimisation and decision-making in scientific and engineering applications [32]. Several classes of methods have been proposed. Monte Carlo Dropout [34] interprets dropout as approximate Bayesian inference and estimates uncertainty via stochastic forward passes. Gaussian process regression [35] provides well-calibrated Bayesian uncertainty but scales cubically with data size. Bayesian Neural Networks (BNNs) [36] place distributions over network weights but often incur high computational costs. These methods typically require sampling-based inference or major architectural changes, limiting their practicality in large-scale and high-dimensional settings. The recently proposed Variational Bayesian Last Layers (VBLL) framework [37] overcomes these limitations by providing efficient, sampling-free uncertainty estimation with minimal overhead, and represents the state of the art of this lightweight model family. VBLL is therefore adopted as the reference model in this work: it shares the same network architecture as the proposed HVBLL and differs only in the noise model, so that the comparison between the two isolates the effect of the heteroscedastic noise model.

3.2. Heteroscedastic variational Bayesian last layers

We propose Heteroscedastic Variational Bayesian Last Layers (HVBLL) to relax the homoscedastic noise assumption while preserving the key advantages of VBLL. The model architectures of Bayesian Linear Regression (BLR), VBLL, and the proposed HVBLL are presented in Fig. 4. In VBLL, a Bayesian treatment is applied only to the last layer of the neural network, while the feature extractor remains deterministic. A Gaussian distribution is placed over the last-layer weights, yielding a lightweight and scalable framework. In HVBLL, we further replace the constant noise term with an input-dependent Gaussian distribution parametrised by an auxiliary neural network.

The proposed Heteroscedastic VBLL (HVBLL) is described as

$$\mathbf{y} = \mathbf{w}^T \phi_\theta(\mathbf{x}) + \epsilon(\mathbf{x}); \quad \epsilon \sim \mathcal{N}(0, \Sigma_\epsilon(\mathbf{x})), \quad (3)$$

where noise $\epsilon(\mathbf{x})$ is assumed to be independent of the stochastic weights $\mathbf{w}$. The noise variance (a scalar for the single output considered here) is modelled with a neural network, $\Sigma_\epsilon(\mathbf{x}) = \exp g_\beta(\mathbf{x})$, where β are the weights of the neural network g . Denote the approximate posterior of weights as $q_\eta(\mathbf{w}) = \mathcal{N}(\mu_w, S_w)$. Then, the predictive distribution of the model is

$$p(\mathbf{y}|\mathbf{x}, \theta, \eta, \beta) = \mathcal{N}(\mu_w^T \phi, \phi^T S_w \phi + \exp g_\beta(\mathbf{x})). \quad (4)$$

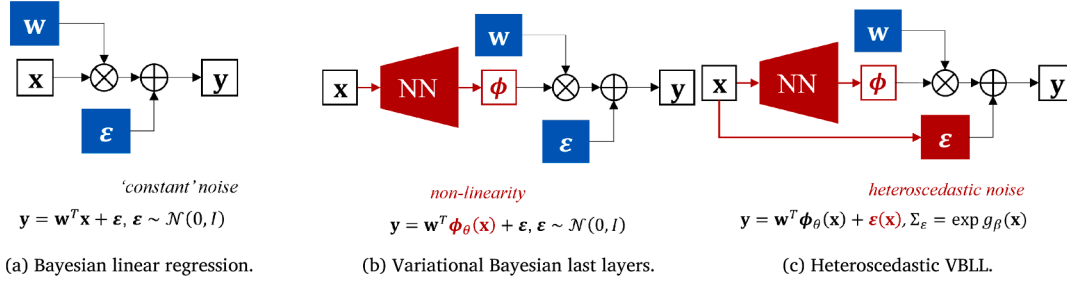


Fig. 4. Bayesian last-layer models. (a) BLR learns stochastic weights ($\mathbf{w}$) under the assumption of independent noise ϵ with constant variance. (b) VBLL introduces a feature network (ϕ_θ) to capture nonlinearity. (c) The proposed HVBLL further replaces the constant noise term with an auxiliary neural network to model the input-dependent covariance (Σ_ϵ).

Note that the $\phi^T S_w \phi$ term is the epistemic uncertainty, and the $\Sigma_\epsilon(\mathbf{x})$ describes the heteroscedastic aleatoric uncertainty. The Gaussian form of the predictive distribution in Eq. (4) is the modelling assumption motivated in Section 2.3.2; its adequacy for the present application is assessed by the verification experiments of Sections 3 and 4, which evaluate the full predictive distribution with proper scoring rules. HVBLL employs sampling-free stochastic variational inference [38]. The evidence lower bound (ELBO) can be derived from:

$$N_s^{-1} \log p(D_y | D_x, \theta, \beta) \geq \mathcal{L}(\theta, \eta, \beta) - N_s^{-1} \text{KL}[q_\eta(\mathbf{w}) || p(\mathbf{w})], \quad (5)$$

where D_y and D_x are the output and input data and N_s is the number of samples. The first term of Eq. (5) is the expected log-likelihood under the approximate posterior,

$$\mathcal{L}(\theta, \eta, \beta) = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathbb{E}_{q_\eta(\mathbf{w})} [\log \mathcal{N}(y_i | \mathbf{w}^T \phi_i, \exp g_\beta(\mathbf{x}_i))], \quad (6)$$

For a Gaussian posterior $q_\eta(\mathbf{w}) = \mathcal{N}(\mu_w, S_w)$ this expectation can be evaluated analytically, which gives

$$\mathcal{L}(\theta, \eta, \beta) = \frac{1}{N_s} \sum_{i=1}^{N_s} \left[\log \mathcal{N}(y_i | \mu_w^T \phi_i, \exp g_\beta(\mathbf{x}_i)) - \frac{1}{2 \exp g_\beta(\mathbf{x}_i)} \phi_i^T S_w \phi_i \right], \quad (7)$$

Then, the loss function is

$$\begin{aligned} \mathcal{J}_{\theta, \eta, \beta} &= -\mathcal{L}(\theta, \eta, \beta) + \lambda_w \text{KL}(q_\eta(\mathbf{w}) || p(\mathbf{w})) \\ &\quad - \lambda_\epsilon \frac{1}{N_s} \sum_{i=1}^{N_s} \left[\nu \log |\Lambda_\epsilon(\mathbf{x}_i)| - \frac{\omega}{2} \text{tr}(\Lambda_\epsilon(\mathbf{x}_i)) \right], \end{aligned} \quad (8)$$

where λ_w and λ_ϵ are the regularisation weights of the last-layer KL divergence and the noise prior term, respectively (Table E.9). Eq. (8) is therefore a reweighted form of the negative evidence lower bound: the two regularisation terms carry the tunable weights λ_w and λ_ϵ in place of the N_s^{-1} factor of Eq. (5), so that their strength can be controlled independently of the number of training samples. Following the VBLL formulation, a Wishart prior is placed on the noise precision matrix $\Lambda_\epsilon = \Sigma_\epsilon^{-1}$, $\log p(\Lambda_\epsilon) = \nu \log |\Lambda_\epsilon| - \frac{\omega}{2} \text{tr}(\Lambda_\epsilon) + \text{const}$, with degrees of freedom ν and scale ω . For the diagonal noise covariance used here, this corresponds to an independent gamma prior on each precision with shape $\nu + 1$ and rate $\omega/2$, or equivalently an inverse-gamma prior on each noise variance with prior mean $\omega/(2\nu)$. The scale is set to $\omega = 2\nu\sigma_0^2$, so that the prior mean of the noise variance equals the estimated noise level σ_0^2 . Since $\Lambda_\epsilon(\mathbf{x})$ is a deterministic function of the input through the noise network, the Wishart term acts in practice as a regulariser that pulls the modelled noise variance towards σ_0^2 , with a strength controlled by ν and λ_ϵ , rather than as a prior on a random quantity. An isotropic zero-mean Gaussian prior is placed on the last-layer weights $\mathbf{w}$. For Gaussian priors, the Kullback–Leibler divergence can be computed in closed form:

$$\begin{aligned} \text{KL}(q_\eta(\mathbf{w}) || p(\mathbf{w})) &= \frac{1}{2} \text{tr}(S_{w,0}^{-1} S_w) + \frac{1}{2} (\mu_w - \mu_{w,0})^T S_{w,0}^{-1} (\mu_w - \mu_{w,0}) \\ &\quad - \frac{n_f}{2} - \frac{1}{2} \log(|S_w|/|S_{w,0}|). \end{aligned} \quad (9)$$

The formulation above is written for a scalar output y and extends directly to a d -dimensional output vector $\mathbf{y}$. In the multi-output case, the last layer is a $d \times n_f$ weight matrix $\mathbf{W}$, and its approximate posterior is factorised over the rows: each row $\mathbf{w}_j$ has an independent Gaussian posterior $\mathcal{N}(\mu_{w,j}, S_{w,j})$ with a dense covariance matrix $S_{w,j}$ over the n_f features. The noise covariance is diagonal, $\Sigma_\epsilon(\mathbf{x}) = \text{diag}(\exp g_\beta(\mathbf{x}))$, where a single noise network g_β maps the input to the d log-variances, one for each output dimension, so that the noise $\epsilon(\mathbf{x})$ is modelled independently per output. The predictive distribution of the output vector therefore factorises into d independent Gaussian distributions of the form of Eq. (4), with marginal variances $\phi^T S_{w,j} \phi + \exp g_{\beta,j}(\mathbf{x})$, and the reconstruction and KL terms in Eqs. (7)–(9) are summed over the d output dimensions. The Z-beam surrogate in Section 4 uses this multi-output form with $d = 20$ PCA coefficients.

The noise prior variance σ_0^2 in Eq. (8) regularises the modelled noise covariance $\Sigma_\epsilon(\mathbf{x})$ and thereby controls the spread of the predictive distribution. Instead of fixing σ_0^2 arbitrarily, we estimate it from the data with a clustering-based noise estimation procedure, which is described in Appendix C.

3.3. Verification on synthetic datasets

The datasets used in the verification experiments of this section and in the applications of this work are summarised in Appendix D. Two synthetic toy functions are used to demonstrate the necessity of employing HVBLL for heteroscedastic problems, described as:

$$y = \sin(2\pi x) + 0.2 \sin(2\pi x) \cdot \epsilon; \quad (10)$$

$$y = x^2 \sin(5\pi x) + \max(0.3 - x, 0.05) \cdot \epsilon; \quad (11)$$

where $\epsilon \sim \mathcal{N}(0, 1)$. 50 and 500 samples are used for training and 200 for testing. For Eq. (10), training samples are taken from $[0, 1]$ and testing samples from $[-0.1, 1.1]$. For Eq. (11), training samples are drawn from $[-0.15, 0.85]$ and testing samples from $[-0.25, 0.95]$.

Both HVBLL and VBLL have three hidden layers with 128 neurons in the feature network ϕ_θ , the dimension of features is $n_f = 64$. The neural network of g_β in HVBLL contains two hidden layers with 128 neurons. The noise prior variance is set to $\sigma_0^2 = 0.05$, which is of the order of the average ground-truth noise variance of the two functions. The initial learning rate is 0.01 and is gradually reduced over the 3000 training epochs. Adam optimiser [39] is used for training. The complete set of training hyperparameters for this and the following experiments is summarised in Appendix E.

Prediction results are shown in Figs. 5 and 6. Training samples are plotted as scatter points and the predictive means as solid lines, the shaded regions show the 1σ bands of the ground-truth uncertainty, the predicted aleatoric uncertainty, and the predicted total predictive uncertainty. The results demonstrate that VBLL fails to capture the input-dependent noise, whereas HVBLL exhibits greater robustness and successfully models heteroscedasticity.

Model performance is evaluated using Negative Log-Likelihood (NLL), Mean Absolute Error (MAE), and Continuous Ranked Probabil-

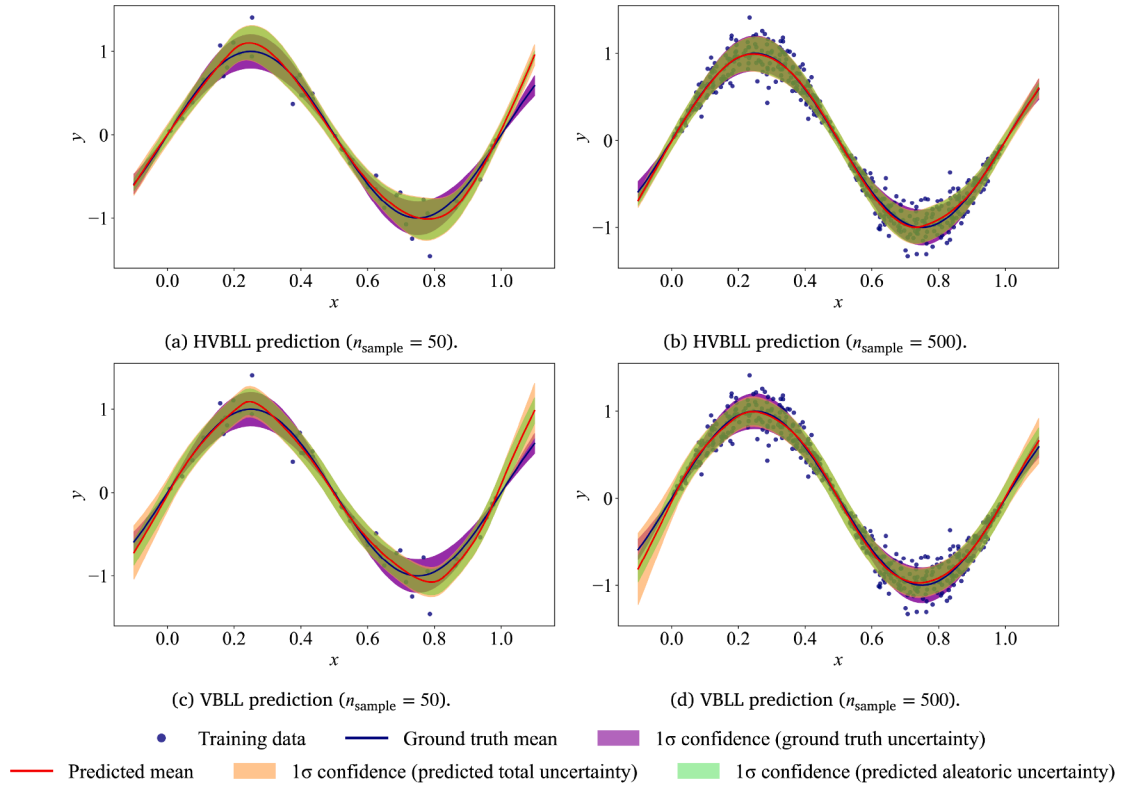


Fig. 5. Predictions of HVBLL and VBLL for Eq. (10) on different sample sizes. The shaded regions show the 1σ prediction bands of the ground-truth uncertainty, the predicted aleatoric uncertainty, and the predicted total predictive uncertainty, as indicated in the legend

ity Score (CRPS). For all metrics, smaller values correspond to better performance. Details of these metrics are provided in Appendix F. Each experiment is repeated ten times with different random samples: in each repetition, new random samples are drawn from the toy function and the network weights are randomly re-initialised, and the reported metrics are the average over the ten repetitions, with the $\pm$ values denoting the standard deviation across the repetitions. The comparative results are summarised in Table 1.

Overall, HVBLL consistently outperforms VBLL across all metrics on both synthetic datasets. HVBLL achieves lower NLL, MAE, and CRPS, with performance further improving as the training set increases from 50 to 500 samples. In particular, the improved NLL and CRPS values demonstrate a better probabilistic quality of the full predictive distribution, rather than only a better pointwise accuracy. VBLL, in contrast, exhibits large variance and higher error, failing to capture the input-dependent noise. VBLL also shows unstable CRPS values in the prediction of Eq. (11), indicating unreliable uncertainty estimation. These results confirm that HVBLL is more robust and capable of capturing heteroscedastic noise structures.

3.4. Verification on benchmark and engineering datasets

The proposed HVBLL model is further verified on two further datasets, as detailed in Appendix A. On the heteroscedastic image regression benchmark of Immer et al. [40], in which the noise level is a prescribed function of the image class, the advantage of HVBLL over VBLL increases monotonically with the degree of heteroscedasticity at a matched total noise variance, whereas the two models are statistically indistinguishable on the homoscedastic control. On the open-hole compression database, a debiased analysis of the heteroscedasticity gain shows that the failure index is nearly homoscedastic, and the two models achieve comparable accuracy; however, HVBLL trains robustly with the default configuration, whereas VBLL requires careful tuning of four hyperparameters to avoid divergence and reach the

Table 1

Performance comparison of HVBLL and VBLL on different toy functions under different sample sizes. The large standard deviations of the VBLL results on the X2Sine-Sine function, most visibly the CRPS values of 2.46 ± 6.82 , are caused by individual training repetitions with unreliable uncertainty estimates, which indicates the unstable uncertainty estimation of VBLL on this heteroscedastic case.

Dataset	Training samples	Model	NLL	MAE	CRPS
Sine-Sine	50	HVBLL	-0.519 ± 0.048	0.103 ± 0.006	0.078 ± 0.004
		VBLL	-0.125 ± 0.651	0.129 ± 0.047	0.130 ± 0.109
	500	HVBLL	-0.565 ± 0.036	0.096 ± 0.004	0.074 ± 0.003
		VBLL	-0.159 ± 0.662	0.121 ± 0.045	0.125 ± 0.110
X2Sine-Sine	50	HVBLL	0.834 ± 0.325	0.171 ± 0.022	0.208 ± 0.019
		VBLL	1.324 ± 1.586	0.216 ± 0.056	2.459 ± 6.819
	500	HVBLL	0.740 ± 0.247	0.142 ± 0.055	0.188 ± 0.037
		VBLL	1.280 ± 1.616	0.184 ± 0.056	2.446 ± 6.824

same accuracy. These results complete a progressive validation strategy: the model is first tested on controlled synthetic cases, then on datasets of graded heteroscedasticity, before being applied to the strongly heteroscedastic Z-beam problem in the following section, where the layout-induced variance accounts for about half of the total output variance (Appendix C).

4. Probabilistic failure prediction of composite wing structural elements

This section applies the proposed framework to the Z-beam representative structural element. The workflow consists of five steps. First, the element is parametrised by three geometric parameters, and a training dataset is generated by sampling the geometric design space and the layout population (Section 4.1). Second, high-fidelity multi-scale finite element analyses of the sampled configurations are performed to con-

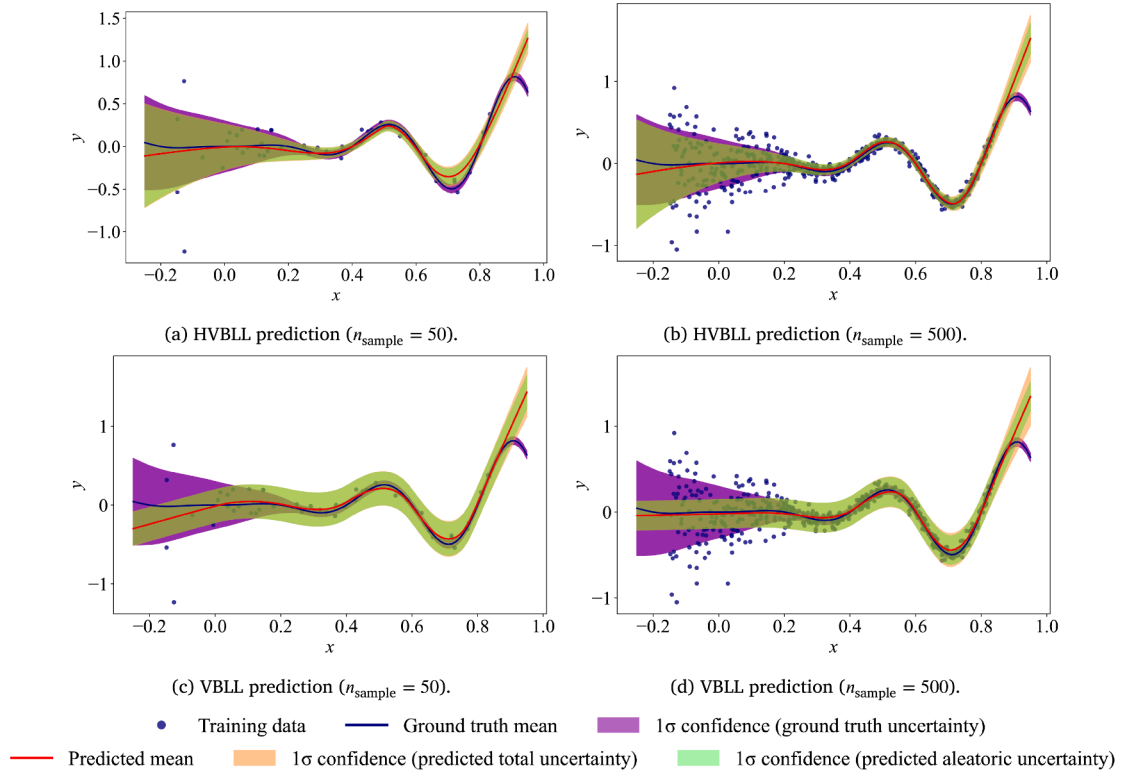


Fig. 6. Predictions of HVBLL and VBLL for Eq. (11) on different sample sizes. The shaded regions show the 1σ prediction bands of the ground-truth uncertainty, the predicted aleatoric uncertainty, and the predicted total predictive uncertainty, as indicated in the legend

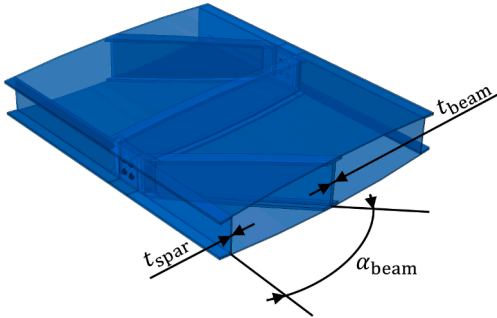


Fig. 7. Independent geometric parameters of Z-beam representative element [30].

struct failure index maps (Section 4.2). Third, the failure index maps are compressed by principal component analysis to reduce the output dimensionality (Section 4.3). Fourth, the HVBLL model is trained to predict the distribution of the failure index map from the geometric parameters, with the strain state spanning the output space of the map, and its performance is compared with the homoscedastic VBLL model (Section 4.4). Finally, the trained model is used to construct probabilistic failure envelopes, and their use in failure assessment is illustrated on a representative gust-response strain trajectory (Section 4.5). The analysis in this section is performed at the level of the representative structural element; system-level structural and manufacturing concerns of the full Z-beam wing concept are outside the scope of this study and are discussed in Section 4.6.

4.1. Parameters of the Z-beam representative element

The Z-beam representative element consists of cover skins, front and rear C-shaped spars, a zigzag I-shaped beam, and titanium bolts (Ti-6Al-4V). All composite components use IM7/8551-7 carbon/epoxy com-

Table 2

Independent geometric parameters and their ranges.

Parameter	Meaning	Range
t_{spar} (mm)	thickness of spars	[5, 25]
t_{beam} (mm)	thickness of beam	[10, 22]
α_{beam} ($^{\circ}$)	angle between diagonal member of the beam and the spar	[20, 60]

posite. The material properties are listed in Appendix G. The design is parametrised by three geometric parameters, $\Theta = (t_{\text{spar}}, t_{\text{beam}}, \alpha_{\text{beam}})$, as illustrated in Fig. 7. The definitions and ranges of these parameters are provided in Table 2.

In this study, the geometric parameters are discretised on a regular grid within the ranges specified in Table 2, with 11 values of t_{spar} , 5 values of t_{beam} , and 9 values of α_{beam} . Grid points for which the finite element analysis did not converge are excluded, leaving 377 valid geometry points. For each geometry point, the dataset contains one *nominal* layout and up to four *randomised* layouts. The nominal layout is a representative design that satisfies the engineering guidelines detailed in Appendix H, and it is selected following the recommendations of Airbus experts. The randomised layouts are random variations that satisfy the same engineering guidelines, so that all layouts in the dataset are practically manufacturable. The front and rear spars are assigned identical layouts, while the cover skins adopt nominal layouts fixed throughout this work. In total, the dataset consists of 1396 samples, comprising 377 samples with nominal layouts and 1019 samples with randomised layouts. The resulting layout population is not uniform over the candidate designs. For each ply count, the candidate layouts satisfying the engineering guidelines are enumerated in advance; the randomised layouts are drawn uniformly from the remaining candidates, whereas the nominal layout is included for every geometry point by construction, so that the nominal layout family is over-represented relative to a uniform sampling over the candidate set. Since the layout is treated as an unobserved stochastic variable in the proposed framework, the predicted failure-

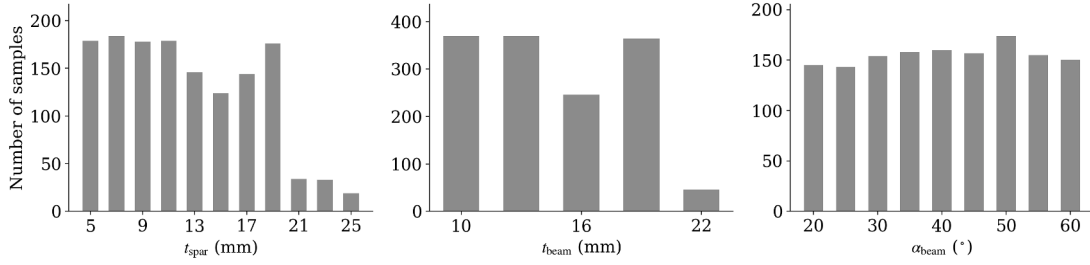


Fig. 8. Statistics of the dataset: distribution of wingbox samples over the three geometric parameters t_{spar} , t_{beam} , and α_{beam} .

index distribution is conditioned on this layup population. The statistics of geometric parameters are presented in Fig. 8.

4.2. Finite element analysis methods

The sampled Z-beam representative elements are simulated in Abaqus following the multi-scale finite element analysis procedure shown in Fig. 2. The complete zigzag beam in the wing tip region is approximated by a one-dimensional series of Z-beam representative elements, while the aeroelastic deformation of the aircraft is captured using reduced-order models [41]. For each representative element, strain states (Δ) are applied under periodic boundary conditions. At the hotspot level, stress (σ) and strain (ϵ) fields are resolved on a refined ply-by-ply mesh using a global-local approach, and failure indices (FI) are computed according to the LaRC05 failure criteria [42]. The periodic boundary conditions are enforced through constraint equations between corresponding nodes on the opposite boundary faces of the representative element, which prescribe the macroscopic axial strain, bending curvatures and twist rate. In the local hotspot models, the mesh is refined in the bolt-hole region, where stress concentrations govern failure initiation. A two-stage mesh convergence study was performed in the radial and circumferential directions of the bolt hole, and the relative difference in the predicted stresses between the two finest meshes is below 1%. The full mesh convergence study is presented in Appendix I. By analysing multiple strain states, a failure index map is generated from the maximum FI in the structure, taken over all LaRC05 failure modes and all material points of the local hotspot region. Then, the corresponding failure envelope can be obtained from the FI map.

Specifically, the FI map is constructed on a polar grid in the two-dimensional strain space spanned by the bending curvature κ_x and the torsional curvature κ_z , with the axial strain ϵ_z and the bending curvature κ_y set to zero. The grid consists of 41 loading directions uniformly distributed over $[0, 2\pi]$ and 9 radial strain scales, from 1×10^{-6} to 32×10^{-6} . For each grid point, the local stress and strain fields of the hotspot model are scaled linearly with the strain scale and the LaRC05 failure indices are evaluated, which exploits the linearity of the underlying finite element analysis and requires no additional simulations. The FI is found to be nearly proportional to the strain scale in each loading direction, and the location of the critical point is nearly independent of the strain scale, as verified on the dataset. The failure envelope (FI = 1) in each loading direction is therefore obtained by linear interpolation of the FI along the radial direction. Further details of this procedure can be found in [30].

4.3. Dimensionality reduction of failure index maps

To mitigate the curse of dimensionality in the model output, principal component analysis (PCA) is applied to compress the high-dimensional failure index (FI) map. The cumulative explained variance ratio as a function of the number of components is shown in Fig. 9a; the reconstruction mean squared error (MSE) is presented in Fig. 9b. Based on these two criteria, 20 principal components were selected to reduce the dimensionality while preserving the essential features of the

FI map. The local MSE distribution for the reconstructed FI map with 20 components is shown in Fig. 9c, demonstrating that the reconstruction error remains small in the critical region near the failure envelope, highlighted by the black dashed lines. To quantify the reconstruction accuracy in this critical region, we examined the error in two ways. First, on all grid points whose true FI lies in $[0.9, 1.1]$, the reconstruction error has a root mean square value of 0.016 and a signed mean of -6×10^{-4} . The reconstruction is therefore practically unbiased around FI = 1, so the failure envelope is not shifted systematically. Second, we measured directly how much the envelope itself moves. For each sample and each loading direction, the envelope is located at the strain scale where the FI map crosses FI = 1, and we found this crossing point by linear interpolation in both the true and the reconstructed maps. Fig. 9d shows the distribution of the absolute difference between the two crossing radii. The mean difference is 0.22×10^{-6} strain, which amounts to 1.07% of the true envelope radius on average and stays below 2.90% in 95% of all cases. The failure envelope is therefore reproduced by the PCA reconstruction with an accuracy of about one percent.

Specifically, each FI map is discretised on a polar grid of 9 strain scales and 41 strain directions and is flattened into a 369-dimensional vector $\mathbf{f}$. Each component of the vector is standardised to zero mean and unit variance across the dataset before PCA. With 20 principal components, 99.15% of the total variance is retained. The surrogate model predicts the 20 PCA coefficients $\mathbf{c}$, and the full FI map is reconstructed by

$$\hat{\mathbf{f}} = (\mathbf{c}\mathbf{V}) \odot \mathbf{s} + \boldsymbol{\mu}, \quad (12)$$

where $\mathbf{V}$ is the matrix whose rows are the principal components, $\boldsymbol{\mu}$ is the mean vector of the original data, and $\mathbf{s}$ is the vector of standardisation scales. The 20 PCA coefficients are predicted as independent Gaussian variables (Section 3.2), so the predictive covariance of the coefficient vector is diagonal, and the predictive standard deviation σ_c of the coefficients is mapped back to the FI map grid by exact variance propagation,

$$\sigma_{\hat{\mathbf{f}}} = \sqrt{(\sigma_c \odot \sigma_c)(\mathbf{V} \odot \mathbf{V}) \odot \mathbf{s}}, \quad (13)$$

where $\odot$ denotes element-wise multiplication and the square root is taken element-wise. This mapping is exact for the pointwise predictive variance under the independence of the predicted coefficients. It is a marginal statement at each grid point and does not by itself describe the joint predictive distribution over the whole failure index map; the calibration of the predicted intervals in both spaces is assessed in Appendix B.

4.4. Model training and performance

The models are further applied to the probabilistic failure prediction of composite Z-beam representative elements. Both HVBLL and VBLL have five hidden layers with 512 neurons in the feature network ϕ_θ , the dimension of features is $n_f = 128$. The neural network of g_β in HVBLL contains three hidden layers with 128 neurons.

The 20 PCA coefficients are standardised to unit variance before training. The total variance of the raw coefficients decreases from about

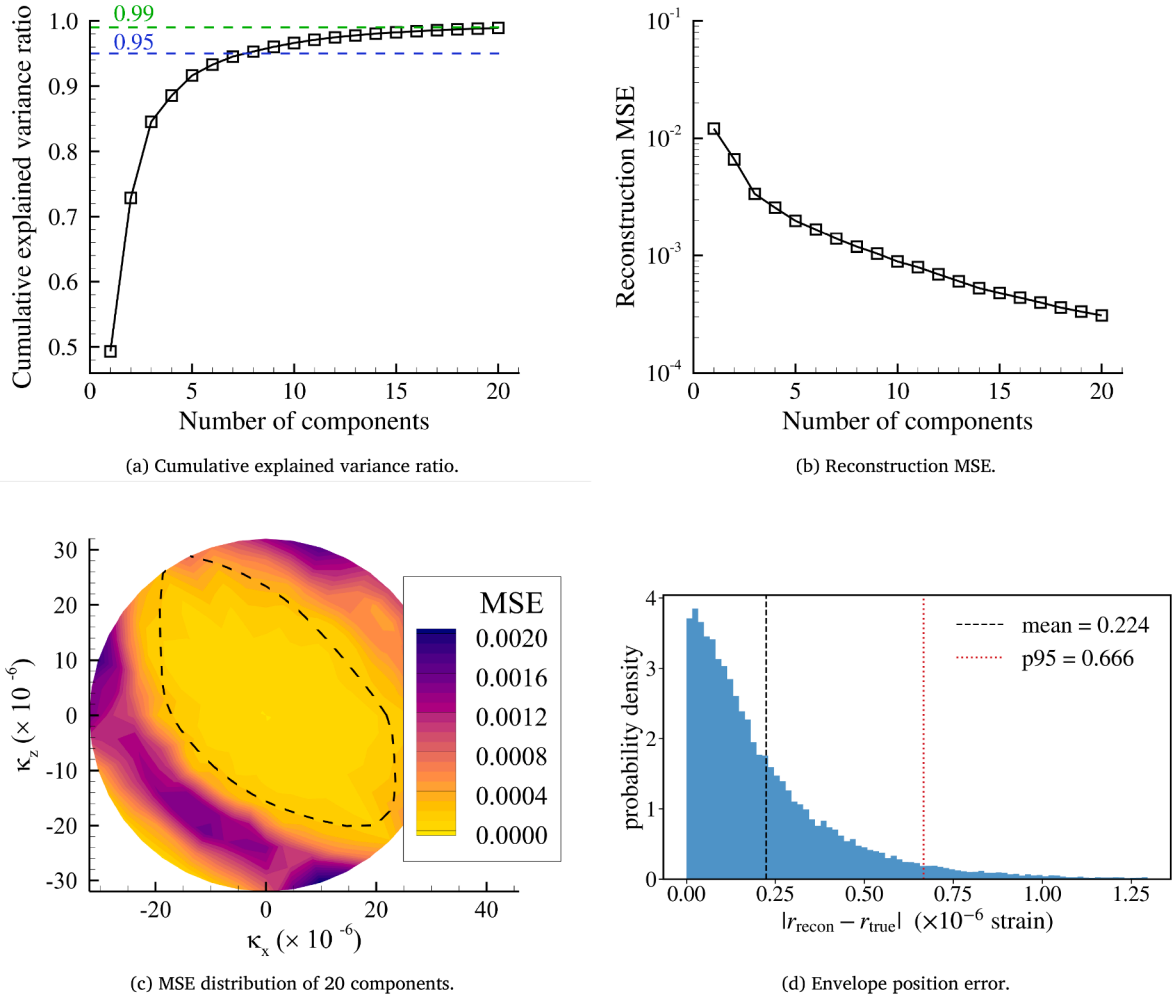


Fig. 9. PCA-based dimensionality reduction of the FI map. The cumulative explained variance ratio (a) and reconstruction error (b) indicate that 20 principal components are sufficient to accurately reconstruct the FI map. (c) shows the reconstruction MSE distribution of the FI map, the black dashed lines mark the average failure envelope (FI = 1). The MSE is small in the critical region near the failure envelope. (d) shows the distribution of the absolute error of the failure-envelope radius between the reconstructed and the true FI maps.

181 for the leading component to about 0.3 for the twentieth component, and the layout-induced noise variance spans a comparable range (Appendix C), so a single scalar noise prior cannot match all output dimensions in the raw coefficient space. After standardisation, the noise variance of every dimension is of order one, and a single scalar prior applies uniformly to all dimensions. The standardisation also makes the isotropic weight prior of the Bayesian last layer dimension-fair and improves the conditioning of the optimisation. Since the last layer is linear, the diagonal rescaling can be absorbed into its weights, so the standardisation does not restrict the mean function that the model can represent. The predicted mean and standard deviation are mapped back to the original coefficient scale before the PCA inverse transformation and the uncertainty propagation of Eq. (13), and all evaluation metrics are computed in the original coefficient space. The noise prior variance is set to $\sigma_0^2 = 0.55$ for both models, as estimated from the dataset in the standardised coefficient space by the procedure described in Section 3 and Appendix C. The Wishart prior on the noise precision uses $\nu = 10$ degrees of freedom with the scale $\omega = 2\nu\sigma_0^2$ (Section 3.2), and the noise prior term enters the loss function with a weight of 10^{-3} . The initial learning rate is 10^{-4} and is gradually reduced over the 5000 training epochs. Adam optimiser is used for training. The same splitting protocol is applied: for each sample size and each of the ten repetitions, an independent random subset is drawn from the 1396 samples and split sample-wise into 80%

training and 20% testing samples. This protocol matches the intended use of the surrogate, which is to replace the high-fidelity analysis within the design space that the training dataset is constructed to cover (Section 4.1), rather than to extrapolate to geometries outside that space. In the intended application at the full-wing level, a large number of candidate wingbox geometries within this covered space must be analysed and optimised rapidly to find the most suitable Z-beam geometry for the aircraft [31]. Each such query is an unseen combination of geometry and layup inside the covered space, which is exactly what the sample-wise split reproduces: the geometry of every test sample is represented in the training set through other layups, while the tested combination itself has never been simulated. The complete set of training hyperparameters is summarised in Appendix E.

The performance comparison between HVBLL and VBLL is summarised in Table 3. Each experiment is repeated ten times with different random samples, following the same seeding and re-initialisation protocol as in Appendix A.2. Both models use the identical training configuration, and the evaluation metrics are detailed in Appendix F. For all metrics, lower values indicate better performance. The results show that HVBLL consistently outperforms VBLL across all metrics and sample sizes. The calibration of the predictive distributions was assessed through the coverage of the $\mu \pm \sigma$ and $\mu \pm 2\sigma$ prediction intervals, and the full calibration results are reported in Appendix B.

Table 3

Performance comparison of HVBLL and VBLL on the composite wing structural element dataset. The noise prior variance $\sigma_0^2 = 0.55$ is estimated from the dataset in the standardised PCA-coefficient space by the procedure in [Appendix C](#); the smaller value $\sigma_0^2 = 0.01$ is reported for comparison to illustrate the sensitivity of the predictive distribution to the noise prior. All metrics are computed in the original PCA-coefficient space and reported as mean $\pm$ standard deviation over ten repetitions. The sample size denotes the number of samples drawn per repetition, which is split into 80% training and 20% testing samples.

Sample Size	Model	Noise prior σ_0^2	NLL	MAE	CRPS
100	HVBLL	0.55	1.913 ± 0.081	1.697 ± 0.092	1.222 ± 0.062
	HVBLL	0.01	1.971 ± 0.107	1.692 ± 0.138	1.225 ± 0.091
	VBLL	0.55	17.144 ± 3.908	2.159 ± 0.144	7.270 ± 4.973
	VBLL	0.01	9.815 ± 4.843	1.990 ± 0.167	3.134 ± 1.355
500	HVBLL	0.55	1.814 ± 0.041	1.600 ± 0.077	1.144 ± 0.060
	HVBLL	0.01	1.877 ± 0.054	1.610 ± 0.076	1.159 ± 0.064
	VBLL	0.55	20.558 ± 5.180	2.277 ± 0.126	7.321 ± 4.987
	VBLL	0.01	9.699 ± 4.150	2.026 ± 0.258	3.112 ± 1.362
1200	HVBLL	0.55	1.720 ± 0.031	1.571 ± 0.047	1.119 ± 0.037
	HVBLL	0.01	1.818 ± 0.048	1.564 ± 0.046	1.127 ± 0.039
	VBLL	0.55	19.332 ± 3.529	2.189 ± 0.085	7.269 ± 4.954
	VBLL	0.01	7.597 ± 3.502	1.932 ± 0.211	2.744 ± 1.073

The metrics of [Table 3](#) are defined in the PCA coefficient space, in which the model output is defined. Expressed instead in engineering terms, the mean absolute error of the predicted failure index is 0.110 in the critical region around $FI = 1$, and the predicted mean envelope deviates from the envelope of an individual test layup by 10.5% of the envelope radius, which is close to the 8.9% scatter that different layups of the same geometry show among themselves. The predictive uncertainty is in turn dominated by the layup-induced aleatoric term, which accounts for about 99% of the predictive variance, so the width of the predicted envelopes reflects the unknown layup rather than a lack of training data. Both analyses are reported in [Appendix K](#).

[Figs. 10 and 11](#) show the predicted mean and standard deviation maps of $FI_{\max}$, respectively. The failure envelope ($FI_{\max} = 1$) is indicated by the white dash-dotted line. Throughout this work, the predicted failure envelopes are defined as the level sets of the predicted FI maps: the mean envelope is the isoline $\mu_{FI_{\max}} = 1$, and the best-case and worst-case envelopes are the isolines $\mu_{FI_{\max}} - 2\sigma_{FI_{\max}} = 1$ and $\mu_{FI_{\max}} + 2\sigma_{FI_{\max}} = 1$, respectively. The mean envelope represents the expected failure boundary over the layup population, and the best-case and worst-case envelopes describe its optimistic and conservative deviations induced by the layup variability. Here $\sigma_{FI_{\max}}$ is the total predictive standard deviation defined in [Eq. \(4\)](#), which comprises the heteroscedastic aleatoric uncertainty $\Sigma_e(\mathbf{x})$ induced by the unobserved layup and the epistemic uncertainty $\phi^T S_w \phi$ of the Bayesian last layer, mapped back to the FI map grid through [Eq. \(13\)](#). Under the Gaussian predictive distribution, the true FI in a given strain direction falls within $\mu_{FI_{\max}} \pm 2\sigma_{FI_{\max}}$ with a

nominal probability of 95.4%, and exceeds each side with a probability of 2.3%. The best-case and worst-case envelopes are therefore pointwise probabilistic estimates in each strain direction, rather than strict bounds on the true envelope. The empirical probability levels of the predicted envelopes on the test set are assessed in [Appendix B](#).

The predicted envelopes do not correspond to one specific layup; identifying the layup that attains a desired failure behaviour is the task of the detailed layup optimisation in the later design stages. Two properties of this interpretation deserve emphasis. First, the predictive distribution is conditional on the sampled layup population, in which the nominal layup family is over-represented by construction ([Section 4.1](#)). The mean envelope is therefore anchored to the behaviour of nominal-like designs, and a differently sampled population would yield different envelopes. This conditioning is a deliberate choice, as the population represents the laminate designs that are considered relevant in practice, namely engineering-rule-compliant nominal designs and randomised variations around them.

Second, the best-case envelope is the pointwise 97.7% quantile of the failure boundary over the population, not the envelope of the best achievable layup, and the achievable optimum of the later detailed layup optimisation may lie beyond it.

The results reveal geometry-dependent variations in both the FI map and its associated total predictive uncertainty. Between the three geometries, the predicted standard deviation varies moderately: the geometry with $t_{\text{spar}} = 17$, $t_{\text{beam}} = 19$, $\alpha_{\text{beam}} = 60^\circ$ exhibits a standard deviation about 15–20% lower over the whole map than the other two geometries, which show similar levels. Within each map, the variation is much stronger and is modulated by the loading direction and the strain magnitude. The growth with the strain magnitude is consistent with the approximately linear scaling of the failure index with the strain level ([Section 4.2](#)), so that absolute differences between layups grow towards the failure envelope, and the angular modulation follows the directional anisotropy of the failure index map, which exhibits the highest failure resistance along a characteristic direction and the lowest resistance in the directions perpendicular to it [[43](#)]. The predicted uncertainty thus varies systematically with the geometry, the loading direction, and the strain magnitude, so a heteroscedastic uncertainty model is more appropriate for this problem than a constant-noise model. The quantitative comparison with the homoscedastic VBLL is given in [Table 3](#).

The computational cost of the proposed framework is dominated by the offline generation of the training data. A single high-fidelity multi-scale analysis of one representative element requires approximately 8 CPU-h [[43](#)], so the generation of the 1396 samples amounts to roughly 11,000 CPU-h in total, which can be parallelised across compute nodes. In contrast, training the surrogate for 5000 epochs takes only a few minutes on a NVIDIA A10 GPU, and a trained surrogate returns the complete probabilistic FI map of a new geometry in approximately 0.1 s. Once the offline database is available, the surrogate therefore reduces the cost of

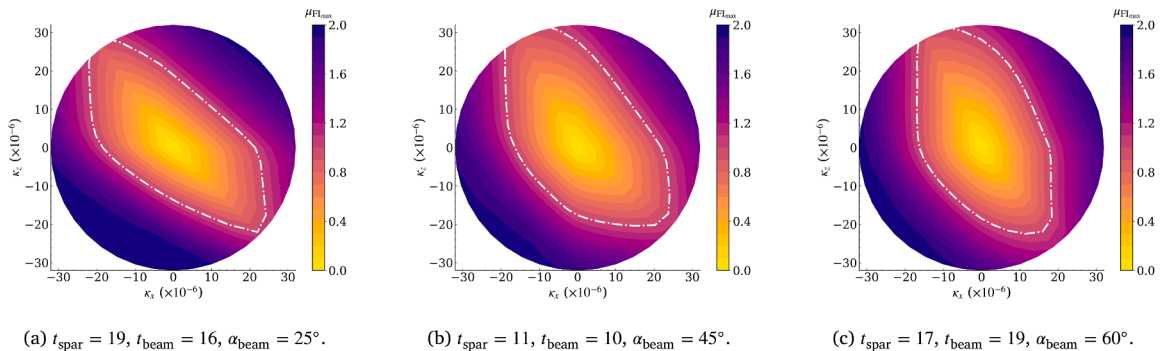


Fig. 10. Mean failure index maps predicted by HVBLL for three different Z-beam wing structural element geometries. The failure envelope ($\mu_{FI_{\max}} = 1$) is indicated by the white dash-dotted line. The predictive mean is shown; the corresponding total predictive standard deviation is shown in [Fig. 11](#).

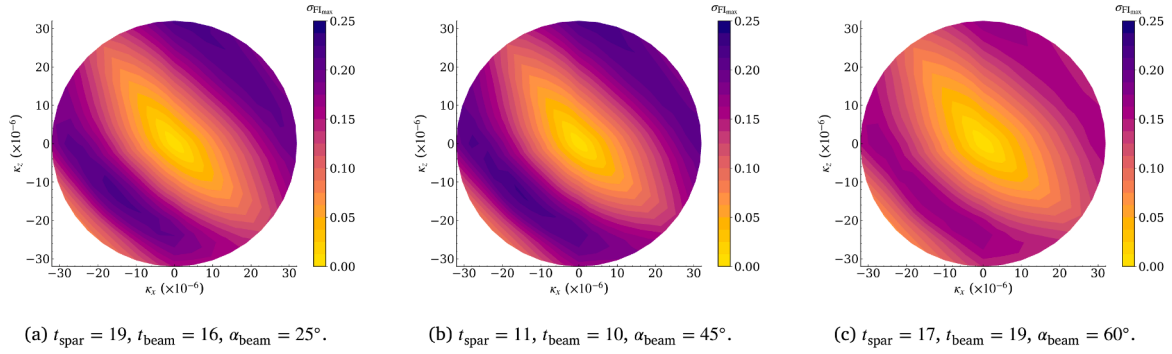


Fig. 11. The total predictive standard deviation of the maximum failure index ($\sigma_{FI_{max}}$) predicted by HVBLL for three different Z-beam wing structural element geometries, comprising the heteroscedastic aleatoric uncertainty induced by the unobserved layup and the epistemic uncertainty of the Bayesian last layer.

a single failure evaluation by several orders of magnitude. This advantage is significant in early-stage design exploration and optimisation, where a large number of candidate configurations must be evaluated, and where each configuration must be checked over the whole range of load cases encountered during structural sizing [31]. Repeating the high-fidelity analysis for every candidate configuration would be computationally prohibitive, whereas the offline cost of data generation and training is incurred only once and is amortised over all subsequent evaluations.

A comparison with explicit layup optimisation further illustrates this advantage. Optimising the layup explicitly for one geometry requires one high-fidelity analysis, at approximately 8 CPU-h per evaluation, for every candidate layup assessed by the optimiser, so that even a search over a few hundred candidates costs on the order of 10^3 CPU-h for that single geometry. Repeating such a search over the geometries of interest therefore exceeds the one-time offline cost of the proposed framework after only a handful of geometries. Once trained, the surrogate instead returns the probabilistic failure envelopes of any geometry in the sampled design space, without any per-geometry optimisation. Moreover, the proposed framework provides the full distribution of the failure response over the layup population rather than only the optimal design, and this distributional information is what supports robust decision-making at the early design stage.

4.5. Application to failure prediction under representative gust response

The predicted probabilistic failure index map of a Z-beam composite wing structural element ($t_{spar} = 19$, $t_{beam} = 16$, $\alpha_{beam} = 25^\circ$) is shown in Fig. 12. The mean failure envelope ($\mu_{FI_{max}} = 1$) is indicated by the white dash-dotted line, the best-case envelope ($\mu_{FI_{max}} - 2\sigma_{FI_{max}} = 1$) by the green dash-dotted line, and the worst-case envelope ($\mu_{FI_{max}} + 2\sigma_{FI_{max}} = 1$) by the blue dash-dotted line. In Fig. 12a, the predicted envelopes are overlaid with the ground-truth envelopes of four sampled layups of the same geometry, shown as dotted lines (black for the nominal layup; light blue, cyan, and light green for three randomised layups) for illustration. The nominal-layup sample is a test sample that was not used in training. The ground-truth envelopes scatter around the predicted mean envelope: all four shown ground-truth envelopes lie mostly or entirely within the best-case and worst-case envelopes, with two of them contained in all strain directions and the other two, including the nominal-layup sample, contained in about 93% of the directions. No individual ground-truth envelope is guaranteed to be contained by the predicted envelopes, because the envelopes are pointwise probabilistic estimates over the layup population rather than strict bounds. The statistical coverage of the predicted intervals is assessed in Appendix B.

An example trajectory of strain states under gust response is illustrated by the black dashed line in Fig. 12b. This trajectory is for demonstration purposes only and does not correspond to any actual structural element. The probabilistic envelopes support three classes of design de-

cisions: if the trajectory remains within the worst-case envelope (blue), the design is clearly safe; if it exceeds the best-case envelope (green), the design is clearly unsafe, since at most about 2.3% of the layups in the population survive in that strain direction and layup optimisation within the same engineering rules is therefore unlikely to rescue the configuration; and if it lies between, the design is intermediate and layup optimisation is required to ensure safety. Here, the mean envelope provides the expected failure response, and the predicted layup-induced variability indicates how sensitive the failure response is to the layup choices that will be made in the later design stages. The probability level associated with this design rule follows from the envelope definitions in Section 4.4: the probability that the true failure envelope lies inside the worst-case envelope is about 2.3% in each direction. These are the nominal levels of the Gaussian predictive distribution. The empirical levels measured on the test set (Appendix B) are close to these nominal levels: the true failure envelope lies inside the worst-case envelope in about 3% of the strain directions, and the best-case and worst-case envelopes together bracket the true envelope in about 94% of the directions. The three decision classes can therefore be interpreted at approximately the nominal probability levels, and a wider band with a recalibrated half-width is available when a specific application requires a strictly higher empirical coverage.

The strain trajectory in the (κ_x, κ_z) space is also the natural interface between the proposed framework and aeroelastic analysis and tailoring. Gust-response or manoeuvre simulations of the full wing provide the time history of the strain state of the element, and the predicted probabilistic failure index map then provides the failure probability along this trajectory. The same strain space carries the effect of aeroelastic tailoring: the bending-torsion coupling that the Z-beam concept exploits sets the direction of the strain state under lift-induced bending, while the directional anisotropy of the failure index map determines the failure resistance along that direction (Section 4.4). Evaluating the aeroelastic performance itself, such as the flutter speed, the divergence margin, the passive twist response, or the gust-load mitigation achieved by tailoring, requires a full aeroelastic model of the wing and is beyond the scope of the present work. Work in this direction is ongoing, building on the reduced-order gust-response model of the high-aspect-ratio wing reported in [31], which provides the strain states of the representative elements during gust encounters.

4.6. Limitations and future work

The proposed framework has several limitations that should be considered when interpreting the results, and these limitations also point to directions for future work.

As discussed in Section 2, the predicted failure-index distribution is conditional on the sampled layup population and does not represent all possible laminate configurations. The predictions also inherit the modelling assumptions of the underlying finite element simulations, and the

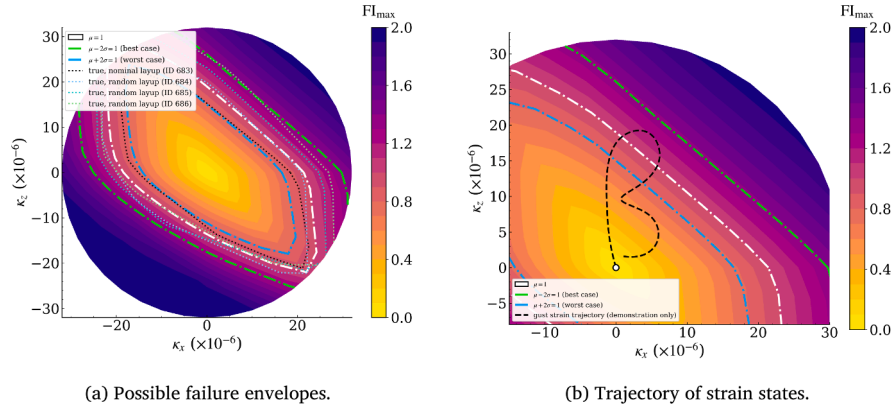


Fig. 12. Analysis of a Z-beam composite wing structural element ($t_{spar} = 19$, $t_{beam} = 16$, $\alpha_{beam} = 25^\circ$) under gust response. (a) shows the predicted mean failure envelope (white dash-dotted), the best-case envelope (green dash-dotted), and the worst-case envelope (blue dash-dotted), together with the ground-truth envelopes of four sampled layups. The mean, best-case, and worst-case envelopes are the isolines $\mu_{FI_{max}} = 1$, $\mu_{FI_{max}} - 2\sigma_{FI_{max}} = 1$, and $\mu_{FI_{max}} + 2\sigma_{FI_{max}} = 1$ of the predicted FI maps, respectively, corresponding to a pointwise probability level of approximately 95% in each strain direction under the Gaussian predictive distribution. (b) illustrates an example trajectory of strain states under gust response with black dashed lines.

assumed Gaussian form of the conditional response may not capture potential nonlinear or multimodal failure-index distributions. Finally, the generation of the high-fidelity training data remains expensive, at roughly 11,000 CPU-h for the present dataset.

The predicted envelopes are pointwise probabilistic estimates at the 95% level in each strain direction, rather than guaranteed bounds on the true failure envelope, and they do not identify the specific layup that attains a given failure boundary. Their reliability has been assessed through the coverage of the prediction intervals and directly in terms of the failure envelope (Appendix B).

The demonstrated scope of the framework is also limited. The validation is based on synthetic datasets, a heteroscedastic image regression benchmark, an open-hole compression case, and one Z-beam representative structural element with three geometric design variables. Demonstrations on additional wing structures and loading scenarios, as well as the scalability to larger design spaces and full-wing optimisation, remain to be shown. Within this scope, the framework is intended for the early-stage screening of candidate designs and does not replace detailed laminate optimisation, high-fidelity failure analysis, or certification-level structural verification. System-level concerns of the Z-beam wing concept, such as load paths in the wing tip region, joints and bolted connections, and assembly and tolerance issues, are not represented in the element-level analysis, and other design constraints, such as buckling, aeroelastic performance and stability (including flutter, divergence, and gust-load mitigation), and manufacturing constraints, and damage tolerance, are left for future work. A first step towards the aeroelastic aspects is reported in [31].

5. Conclusion

Composite materials are essential for achieving lightweight and high-performance aircraft structures. However, the incorporation of failure constraints in early-stage design remains challenging, particularly for novel structural concepts. This challenge is largely driven by the high-dimensional nature of composite layups, which renders conventional predictive modelling and optimisation approaches computationally prohibitive.

In this work, an uncertainty-aware surrogate framework is proposed for rapid failure-envelope prediction of composite wing structural elements under layup variability. By treating composite layups as stochastic variables, the framework avoids explicit exploration of the high-dimensional design space and directly predicts the distribution of failure responses. The predicted distributions are conditional on the sampled layup population (Section 4.1), and the framework does not claim to

cover all possible layup configurations. As a result, probabilistic best-case and worst-case failure envelopes can be efficiently obtained without relying on predefined layups, conservative assumptions, or nested optimisation procedures. This enables the seamless integration of failure constraints into early-stage design while maintaining a tractable and low-dimensional parameter space.

The main contributions and findings are summarised as follows:

- **Probabilistic failure envelope prediction for early-stage design:** The framework directly estimates failure envelopes across different layups, capturing both the mean trend and its variability to support reliable strength assessment. By avoiding the need for the detailed layup as a high-dimensional design variable, the framework remains applicable even at the early design stage, before the layup is determined. Calibration checks support this reliability: using a noise prior estimated from the dataset, the coverage of the $\mu \pm \sigma$ and $\mu \pm 2\sigma$ prediction intervals in the PCA coefficient space falls below the nominal levels by 0.06–0.11 and 0.04–0.08 and approaches them as the training sample grows (Section 4.4 and Appendix B).
- **Efficient heteroscedastic uncertainty modelling:** HVBLL captures input-dependent aleatoric uncertainty while staying computationally efficient and scalable. Its advantage over VBLL shows up on datasets with strong heteroscedasticity; on nearly homoscedastic problems, the added heteroscedastic noise model gives no benefit over the plain homoscedastic baseline.
- **Verification on a novel wing structural concept:** The validation follows a progressive strategy, in which the method is first verified on controlled synthetic cases and on datasets of graded heteroscedasticity (Appendix A) before being applied to the more complex Z-beam structure. Application to the Z-beam wing concept demonstrates that the proposed method accurately predicts failure index distributions and probabilistic failure envelopes, offering actionable insights for structural assessment and guiding subsequent layup optimisation. As a novel wing structural concept, the Z-beam application also demonstrates the practical applicability of the proposed method in real-world engineering scenarios.

Overall, the proposed framework shifts failure assessment from deterministic evaluation to probabilistic characterisation, providing a practical and scalable solution for early-stage composite wing design. The strength of the study lies in the specific combination of early-stage composite wing design, layup-induced uncertainty, and heteroscedastic Bayesian learning. As discussed in Section 4.6, the framework is intended for the early-stage screening of candidate designs and does not replace detailed layup optimisation or high-fidelity failure analysis in

the later design stages, but guides them. Beyond the Z-beam case, the methodology is potentially extendable to more complex structural configurations, which remains to be demonstrated in future work, supporting the development of next-generation aerospace structures.

These conclusions should be considered together with the limitations of the framework discussed in Section 4.6. Future work will extend the proposed framework to incorporate additional design criteria, including buckling, aeroelastic performance, and structural stability. In addition, the methodology will be integrated into a multidisciplinary design and optimisation framework to enable simultaneous consideration of multiple performance objectives in composite wing design.

CRedit authorship contribution statement

Runze Li: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Mário Miranda:** Writing – review & editing, Validation, Data curation, Conceptualization; **Silvestre T. Pinho:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

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

Acknowledgements

The authors would like to gratefully acknowledge the funding from Innovate UK under the Next Wing project No. 10002372 (<https://gtr.ukri.org/projects?ref=10002372>). The authors also gratefully acknowledge the valuable input and guidance provided by Airbus. In particular, Airbus contributed insightful suggestions on the wingbox structural configuration, structural design requirements, appropriate parameter ranges, and the practical context for the application of the proposed methodology. These inputs have been instrumental in ensuring the relevance and realism of this study.

Data availability

Data will be made available on request.

Appendix A. Verification on benchmark and engineering datasets

This appendix presents the verification of the proposed HVBLL model on two further datasets: a heteroscedastic image regression benchmark in which the degree of heteroscedasticity is controlled by construction, and an engineering open-hole compression database. The image regression benchmark shows how the advantage of HVBLL over VBLL depends on the degree of heteroscedasticity, and the open-hole compression tests show its behaviour on a nearly homoscedastic engineering problem.

A.1. Heteroscedastic image regression

Tabular regression benchmarks are of limited use for assessing heteroscedastic models, because the noise level of such datasets usually varies only weakly with the input. The same limitation is reported by Immer et al. [40], who consequently propose an image regression benchmark in which the noise level is a known function of the image content. That benchmark is adopted here. An observation is an image of the MNIST database [44] rotated by an angle ψ drawn uniformly from $[-90^\circ, 90^\circ]$, and the scalar target is

$$y = \psi + \sigma(c)\epsilon, \quad \epsilon \sim \mathcal{N}(0, 1), \quad (\text{A.1})$$

where $c \in \{0, 1, \dots, 9\}$ is the class of the image. Learning this task requires a mean function that recovers the rotation angle and, independently, a noise function that recognises the class of the image, so the

Table A.4

Ladder of noise levels of the image regression benchmark. The noise standard deviation is $\sigma(c) = \sigma_{\min} r^{c/9}$, with $\sigma_{\min}$ chosen such that the total noise variance $E[\sigma^2]$ is matched across the rungs to within 3%. The dynamic range r is therefore the only quantity that varies appreciably, and the heteroscedasticity gain H of Eq. (A.2) depends on r alone. The first rung is the homoscedastic control.

Rung	Dynamic range r	σ range	Heteroscedasticity gain H
1	1	59.6	0.000
2	3	30.6–91.8	0.118
3	10	12.0–119.6	0.457
4	30	4.6–137.2	0.870
5	100	1.5–150.8	1.379

noise network is exercised on a genuinely non-linear mapping rather than on a mild residual trend.

The degree of heteroscedasticity is quantified by the gain

$$H = \frac{1}{2} (\log E[\sigma^2] - E[\log \sigma^2]), \quad (\text{A.2})$$

which is the expected reduction in negative log-likelihood per sample that an ideal input-dependent variance model can achieve over the best constant-variance model. The same quantity is used in Appendix A.2 to show that the open-hole compression problem is nearly homoscedastic. Since $\sigma(c)$ is prescribed, H is known exactly rather than estimated.

In the original benchmark the noise standard deviation is the linear function $\sigma(c) = 11c + 1$, so that the degree of heteroscedasticity and the overall noise magnitude cannot be varied independently. To separate the two effects, the noise function is re-parametrised here in the geometric form $\sigma(c) = \sigma_{\min} r^{c/9}$, where $r = \sigma_{\max}/\sigma_{\min}$ is the dynamic range of the noise, and the benchmark is extended to a ladder of five rungs of increasing r . In this form the gain H depends on r alone, so that $\sigma_{\min}$ can be used to match the total noise variance $E[\sigma^2]$ across the rungs, which is done here to within 3%; the residual variation stems from the class frequencies of the dataset not being exactly uniform. The dynamic range is therefore the only quantity that changes appreciably along the ladder. The five rungs are listed in Table A.4. The first rung is homoscedastic and serves as the control.

The images are used in their raw form, each 28×28 pixel array being flattened into a 784-dimensional input vector, and both the feature network and the noise network are fully connected networks acting on this representation. The target is standardised with the mean and the standard deviation of the training subset, and all reported metrics are transformed back to the original angular units. Both models use a feature network of three hidden layers of width 500, a feature dimension of 64, and, for HVBLL, a noise network of two hidden layers of width 128. The KL regularisation weights are 0.1 for the last layer and 1.0 for the noise, and the Wishart noise prior uses $\nu = 1$ with the noise prior variance estimated from the training subset by Algorithm 1. The initial learning rate is 10^{-3} decayed by a factor of 0.8 every 500 steps, the batch size is 512, and both models are trained for a fixed budget of 80 epochs without early stopping. The training set contains 50,000 samples and the test set 10,000 samples, and every case is repeated ten times with different random samplings and initialisations. These settings are the default configuration of the model and were not tuned on this benchmark.

The results are summarised in Table A.5. On the homoscedastic control the two models are statistically indistinguishable on every metric, which confirms that the input-dependent noise model costs nothing when there is nothing to gain. As the heteroscedasticity gain increases, the advantage of HVBLL grows monotonically in both the CRPS and the root mean square error, and the Spearman rank correlation between H and the paired difference is unity for both metrics. The two models respond in opposite directions: the error of HVBLL against the true rotation angle decreases from 21.6 to 17.7 as the noise becomes more strongly input-dependent, whereas that of VBLL increases from 21.8 to 23.4. This is the expected behaviour, because a model that identifies which samples are noisy can discount them when fitting the mean, whereas a model

Table A.5

Performance of HVBLL and VBLL on the image regression benchmark, as a function of the heteroscedasticity gain H . Values are the mean $\pm$ standard deviation over ten repetitions, and the better mean of each pair is shown in bold. No values are highlighted on the homoscedastic control, where none of the paired differences is statistically significant. The root mean square error is measured against the true rotation angle. The total noise variance is matched across the rungs to within 3%, so H is the only quantity that changes appreciably.

H	NLL		CRPS		RMSE	
	HVBLL	VBLL	HVBLL	VBLL	HVBLL	VBLL
0.000	5.627 $\pm$ 0.014	5.626 $\pm$ 0.018	36.13 $\pm$ 0.29	36.16 $\pm$ 0.28	21.56 $\pm$ 0.77	21.76 $\pm$ 0.49
0.118	5.640 $\pm$ 0.013	5.627 $\pm$ 0.020	34.96 $\pm$ 0.24	35.09 $\pm$ 0.30	21.13 $\pm$ 0.90	22.19 $\pm$ 0.47
0.457	5.658 $\pm$ 0.013	5.634 $\pm$ 0.021	32.74 $\pm$ 0.21	33.22 $\pm$ 0.30	19.68 $\pm$ 0.72	22.72 $\pm$ 0.53
0.870	5.671 $\pm$ 0.016	5.639 $\pm$ 0.024	31.14 $\pm$ 0.21	31.90 $\pm$ 0.30	18.57 $\pm$ 0.81	23.02 $\pm$ 0.74
1.379	5.678 $\pm$ 0.019	5.644 $\pm$ 0.024	29.83 $\pm$ 0.24	30.79 $\pm$ 0.31	17.67 $\pm$ 0.88	23.43 $\pm$ 0.89

with a single global noise level cannot. In a paired comparison over the ten repetitions, all differences from the third rung onwards are significant at the 5% level, as is the difference in the root mean square error already at the second rung.

The negative log-likelihood behaves differently from the other two metrics and slightly favours VBLL, by 0.001 to 0.034 nats. The reason is that the negative log-likelihood penalises an underestimated variance through the term $(y - \mu)^2 / 2\sigma^2$, which is unbounded as σ decreases, so a small number of samples whose predicted noise level is too low can dominate the average. The CRPS penalty grows only linearly with the standardised residual and is not dominated by the same samples. Since the heteroscedastic model has to predict a range of noise levels while the homoscedastic model predicts a single conservative value, only the former is exposed to this penalty, and its magnitude grows with the width of the range that has to be predicted. The two models are therefore comparable in terms of the negative log-likelihood on this benchmark, and the difference between the metrics reflects their differing sensitivity to underestimated variances rather than a difference in the quality of the predictive distributions.

A further observation concerns the accuracy of the recovered noise map. The Spearman correlation between the predicted and the true noise level is between 0.68 and 0.76 across the four heteroscedastic rungs, so HVBLL recovers the ordering of the noise levels but not their magnitude. The mean absolute error of $\log \sigma$ grows along the ladder from 0.26 at the second rung to 1.43 at the fifth, that is, at the strongest rung the predicted noise level is off by a factor of about four, and its range is compressed to a small fraction of the two orders of magnitude that are prescribed. A separate control was run on the original noise function

$\sigma(c) = 11c + 1$ of [40], on which HVBLL reaches a mean absolute error of 0.79 in $\log \sigma$: training the same noise network directly on the known values of $\log \sigma^2$ by least squares reduces this error to 0.25, which indicates that the limitation lies in the training objective rather than in the capacity of the noise network. A plausible cause is the known difficulty of heteroscedastic maximum likelihood training [45], in which the gradient of the mean function is weighted by the inverse of the predicted variance, so that samples with a high noise level contribute less to the mean function, which in turn biases the estimated noise level. The Wishart prior counteracts this effect by shrinking the predicted noise level towards its prior mean, at the cost of a narrower predicted range. This suggests that the accuracy of the noise map could be improved further, and the results reported here are therefore a conservative estimate of the benefit of the heteroscedastic model.

A.2. Open-hole compression tests

The open-hole compression database used in this work contains 4000 finite-element samples at 168 geometric design points with different layouts at each point (Appendix J). Fig. A.13 shows the conditional mean and variance of the maximum failure index over the geometric design space, estimated by grouping the samples of each geometric point, so that no cross-geometry averaging is involved. The horizontal axis is the edge distance $d_{\text{edge}} = x_r l_x - r_{\text{hole}}$, where x_r is the relative hole position and $l_x = 50$ mm is the plate width transverse to the loading direction. The conditional mean shows the expected trend, with higher failure indices at small edge distances and large hole radii. The conditional variance varies only weakly over the design space, with a 90th-to-10th percentile ratio of about 1.5 and no visible spatial trend. To quantify this, the heteroscedasticity gain of Eq. (A.2) is evaluated by substituting the per-geometry conditional variances s^2 for σ^2 , with a digamma correction for the finite-group-size bias of $\log s^2$. The debiased gain is -0.014 nats per sample, which is essentially zero (the small negative value reflects the residual estimation bias) and indicates that the open-hole compression problem is nearly homoscedastic.

The performance of HVBLL and VBLL on this database is compared in Table A.6. For each sample size and each of the ten repetitions, an independent random subset is drawn from the full database and split sample-wise into 80% training and 20% testing samples, using different random seeds. In each repetition, the seed fixes the random subset selection and the train/test split, so that the data used in each repetition can be reproduced exactly. The network weights are randomly re-initialised in each repetition and are not fixed by the seed, so the ten repetitions differ in both the sampled data and the initialisation. The split is performed at the sample level rather than at the geometry level, so different layouts of the

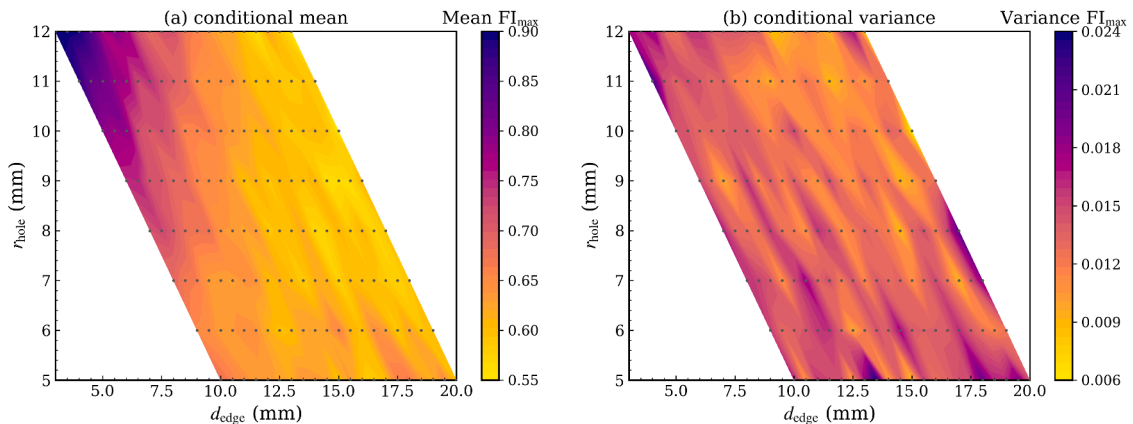


Fig. A.13. Conditional mean (a) and variance (b) of the maximum failure index in the open-hole compression database, estimated per geometric design point over its layouts. The grey dots mark the sampled geometric design points. These conditional statistics are finite-sample estimates from the layouts available at each design point, and they are used as a reference approximation of the conditional statistics rather than as exact ground truth.

Table A.6

Performance comparison of HVBLL and VBLL on the open-hole compression database (mean $\pm$ standard deviation over ten repetitions).

Sample Size	Model	NLL	MAE	CRPS
500	HVBLL	-0.716 ± 0.051	0.1029 ± 0.0058	0.0675 ± 0.0038
	VBLL	-0.729 ± 0.050	0.1024 ± 0.0054	0.0669 ± 0.0040
1000	HVBLL	-0.705 ± 0.054	0.1040 ± 0.0046	0.0691 ± 0.0035
	VBLL	-0.722 ± 0.034	0.1021 ± 0.0041	0.0671 ± 0.0027
4000	HVBLL	-0.733 ± 0.055	0.1033 ± 0.0038	0.0674 ± 0.0034
	VBLL	-0.734 ± 0.018	0.1020 ± 0.0022	0.0661 ± 0.0013

same geometry can appear in both the training and the testing set. The test set therefore contains unseen combinations of geometry and layout within the same design space, rather than entirely unseen geometries. The reported metrics are the average over the ten repetitions, with the $\pm$ values denoting the standard deviation across the repetitions. HVBLL trains stably with the configuration summarised in [Appendix E](#). VBLL, in contrast, diverges or collapses to a constant predictor under the default configuration and requires four simultaneous adjustments to train stably: a learning rate of 5×10^{-4} , a noise prior variance of 2.5×10^{-2} , a noise KL regularisation weight of 1.0, and 10,000 training epochs. With this configuration, the two models achieve comparable accuracy on all three metrics, and the paired differences over the ten repetitions are not statistically significant. The conditional distributions are well calibrated for both models, with coverages of 0.66–0.70 and 0.96–0.97 for HVBLL and 0.73–0.75 and 0.97–0.99 for VBLL for the $\mu \pm \sigma$ and $\mu \pm 2\sigma$ intervals. The open-hole compression problem is therefore a case of weak heteroscedasticity: the accuracy of the two models is comparable, and the practical advantage of HVBLL is its robustness to the training configuration, which is consistent with the behaviour observed on the image regression benchmark, where the advantage of HVBLL vanishes as the heteroscedasticity gain approaches zero.

Appendix B. Calibration of the predictive distributions

This appendix reports the calibration of the predictive distributions on the Z-beam dataset through the coverage of the prediction intervals, that is, the percentage of test outputs falling within the predicted $\mu \pm \sigma$ and $\mu \pm 2\sigma$ intervals, which should ideally match the nominal levels of 68.3% and 95.4% under a Gaussian predictive distribution. The coverage is evaluated in two spaces: in the PCA coefficient space, where the model output is defined, the coverage is averaged over the 20 coefficients, and in the FI map space, where the predictive standard deviation is mapped back through [Eq. \(13\)](#), the coverage is evaluated over all 369 grid points. The results are summarised in [Table B.7](#). With the estimated noise prior $\sigma_0^2 = 0.55$, the HVBLL coverage in the PCA space is 0.58–0.62 for the $\mu \pm \sigma$ interval and 0.87–0.92 for the $\mu \pm 2\sigma$ interval, which falls below the nominal levels by 0.06–0.11 and 0.04–0.08, respectively. The coverage increases monotonically with the sample size, from 0.576 to 0.624 for the $\mu \pm \sigma$ interval and from 0.871 to 0.915 for the $\mu \pm 2\sigma$ interval between the smallest and the largest sample size, approaching the nominal levels as more training data become available. The coverage obtained with the much smaller value $\sigma_0^2 = 0.01$ (cf. [Table 3](#)) is also reported: the resulting under-coverage of 0.54 and 0.84–0.85 in the PCA space shows that an arbitrarily chosen noise prior can miscalibrate the predictive distribution, which underlines the importance of the estimation procedure described in [Appendix C](#).

The coverage in the FI map space is slightly lower than in the PCA coefficient space, 0.57–0.61 and 0.86–0.89, for two reasons: the FI values are bounded below by zero, so their marginal distributions deviate from the assumed Gaussian form, and the coverage is averaged over all 369 grid points, including directions with near-zero variance where even small absolute deviations count as misses. The coverage of VBLL must be interpreted jointly with the accuracy metrics in [Table 3](#): the

Table B.7

Calibration of the predictive distributions on the composite wing structural element dataset: coverage of the $\mu \pm \sigma$ and $\mu \pm 2\sigma$ prediction intervals (nominal levels 0.683 and 0.954) in the PCA coefficient space and in the FI map space, mean $\pm$ standard deviation over ten repetitions.

Sample Size	Model	σ_0^2	PCA coefficient space		FI map space	
			$\mu \pm \sigma$	$\mu \pm 2\sigma$	$\mu \pm \sigma$	$\mu \pm 2\sigma$
100	HVBLL	0.55	0.576 ± 0.035	0.871 ± 0.021	0.565 ± 0.057	0.862 ± 0.032
	HVBLL	0.01	0.542 ± 0.037	0.838 ± 0.045	0.508 ± 0.059	0.815 ± 0.038
	VBLL	0.55	0.582 ± 0.079	0.686 ± 0.062	0.990 ± 0.011	0.998 ± 0.003
	VBLL	0.01	0.648 ± 0.065	0.774 ± 0.048	0.936 ± 0.080	0.985 ± 0.015
500	HVBLL	0.55	0.596 ± 0.014	0.896 ± 0.003	0.595 ± 0.016	0.879 ± 0.009
	HVBLL	0.01	0.538 ± 0.017	0.852 ± 0.008	0.518 ± 0.023	0.819 ± 0.016
	VBLL	0.55	0.579 ± 0.075	0.676 ± 0.062	0.985 ± 0.014	0.996 ± 0.005
	VBLL	0.01	0.645 ± 0.058	0.787 ± 0.049	0.938 ± 0.064	0.984 ± 0.015
1200	HVBLL	0.55	0.624 ± 0.008	0.915 ± 0.006	0.607 ± 0.018	0.889 ± 0.011
	HVBLL	0.01	0.537 ± 0.010	0.852 ± 0.011	0.507 ± 0.021	0.813 ± 0.016
	VBLL	0.55	0.583 ± 0.077	0.677 ± 0.061	0.987 ± 0.012	0.997 ± 0.003
	VBLL	0.01	0.655 ± 0.053	0.809 ± 0.051	0.914 ± 0.089	0.978 ± 0.022

over-coverage of VBLL in the FI map space, which reaches near-unit levels at the estimated noise prior, is achieved with NLL values several times above those of HVBLL, which indicates prediction intervals that are trivially wide rather than informative. The opposing trend of the VBLL coverage in the two spaces follows from the homoscedastic noise model. With an input-independent noise variance, VBLL cannot follow the input-dependent noise level of the individual coefficients: the predictive variance is too small where the true noise is large, which produces the under-coverage in the coefficient space and dominates the negative log-likelihood, and too large where the true noise is small. In the FI map space, the variance contributions of all 20 coefficients add in quadrature at every grid point through [Eq. \(13\)](#), so the overestimated variance components inflate the predictive intervals of the entire map, which leads to the over-covering but uninformative intervals. If coverage levels closer to the nominal values are required in a specific application, a post-hoc recalibration of the predictive variance on a held-out set can be applied.

The coverage in the FI map space averages the pointwise coverage over all grid points, whereas the design rule of [Section 4.5](#) acts on the failure envelope itself. The calibration of the predicted envelopes was therefore assessed directly in terms of the envelope radius. For each test sample and each of the 41 strain directions, the true envelope radius is the strain scale at which the true failure index map crosses $FI = 1$, found by linear interpolation, and the predicted band in the same direction is given by the crossing radii of the best-case and worst-case envelope isolines defined in [Section 4.4](#). Over the ten repetitions at the largest sample size, the true envelope radius falls within the predicted band in 0.938 ± 0.011 of the sample-direction cases, with a per-repetition range of 0.92–0.96; the true radius lies outside the best-case envelope in 0.033 ± 0.008 of the cases and inside the worst-case envelope in 0.029 ± 0.006 of the cases, both close to the nominal one-sided level of 2.3%.

A Monte Carlo cross-check, in which failure index maps are reconstructed from samples of the predicted PCA coefficient distribution, shows that the isoline band coincides with the 2.3%–97.7% percentile band of the predictive distribution of the envelope radius to within about one percentage point, so the predicted envelopes are a consistent pointwise summary of the predictive distribution of the failure envelope. The same assessment provides the empirical coverage of bands with other half-widths: widening the band to $\mu_{FI_{max}} \pm 2.5\sigma_{FI_{max}}$ raises the envelope-radius coverage to 0.973 ± 0.007 and reduces the probability of the true envelope falling inside the worst-case envelope to below 0.1%, whereas the 2σ band used throughout this work reaches an envelope-radius coverage of 0.938 ± 0.011 , close to but slightly below the nominal two-sided level of 95.4%.

In the protocol reported above, the PCA basis, the standardisation statistics, and the noise prior estimate are computed once on the full dataset and are then held fixed across all repetitions. The PCA basis is a fixed, unsupervised output representation rather than a fitted component of the surrogate. Training the models of Section 4.4 in the same coefficient space that is introduced in Section 4.3 therefore keeps the output representation identical across all repetitions and both models, so that their performance is directly comparable. As a control, the complete protocol was repeated with all preprocessing refit on the training subset of each repetition, including the PCA basis, the input and output scaling, and the noise-prior estimation; the re-estimated noise prior variance ranges from 0.57 to 0.68 across the sample sizes. The per-fold results are close to the reported ones: the HVBLL negative log-likelihood increases by 0.12 at the smallest sample size and by at most 0.01 at the larger ones, the interval coverages change by at most 0.017, and the VBLL results degrade further. All conclusions of this work are therefore unaffected by this design choice.

Appendix C. Noise prior estimation

The performance of probabilistic regression models is strongly affected by the noise prior σ_0^2 , especially when the training data is scarce, so a reasonable estimate of σ_0^2 should be obtained from the dataset itself. For a regression problem $y = f(\mathbf{x}) + \epsilon$ with input-dependent noise $\epsilon \sim \mathcal{N}(0, \sigma^2(\mathbf{x}))$, the law of total variance decomposes the total variance of the output as $V_{\text{total}} = E_{\text{noise}} + V_{\text{mean}}$, where $E_{\text{noise}} = E_{\mathbf{x} \sim p(\mathbf{x})}[\sigma^2(\mathbf{x})]$ is the average aleatoric uncertainty and $V_{\text{mean}} = \text{Var}_{\mathbf{x} \sim p(\mathbf{x})}[f(\mathbf{x})]$ is the variance of the mean function. The average conditional variance E_{noise} is therefore a natural estimate of σ_0^2 . In addition, the variance of the noise level $V_{\text{noise}} = \text{Var}_{\mathbf{x} \sim p(\mathbf{x})}[\sigma^2(\mathbf{x})]$ characterises the degree of heteroscedasticity of the dataset.

These quantities are estimated directly from the data by the clustering-based procedure summarised in Algorithm 1. The samples are partitioned into small groups of at most n_s samples with similar inputs $\mathbf{x}$. Within each group, a zero-order approximation of the regression problem is assumed, that is, the variation of the mean function within a group is negligible, so the within-group variance of the output can be interpreted as the local aleatoric uncertainty. The mean and the variance of the output are computed within each group, and E_{noise} , V_{mean} and V_{noise} follow as the mean of the within-group variances, the vari-

ance of the within-group means, and the variance of the within-group variances, respectively.

Algorithm 1 Clustering-based noise level estimation.

Require: Data set D , maximum number of samples per group n_s

- 1: Calculate the total variance V_{total} of y in D
- 2: Group the samples by similar $\mathbf{x}$ using nearest-neighbour clustering
- 3: $N_{\text{group}} \leftarrow$ number of groups
- 4: **for** $i = 1$ **to** N_{group} **do**
- 5: $m_i \leftarrow$ mean of y in group i
- 6: $v_i \leftarrow$ variance of y in group i
- 7: **end for**
- 8: $V_{\text{mean}} \leftarrow \text{Var}(m_i)$; $E_{\text{noise}} \leftarrow \text{Mean}(v_i)$; $V_{\text{noise}} \leftarrow \text{Var}(v_i)$

For the Z-beam surrogate, the procedure is applied to the 20 standardised PCA coefficients, with groups of $n_s = 5$ nearest neighbours identified in the standardised three-dimensional input space. The estimated noise variances range from 0.22 to 0.69 across the 20 output dimensions, with a mean of 0.55. Since the coefficients are standardised to unit variance, these values coincide with the ratio $E_{\text{noise}}/V_{\text{total}}$ of each dimension, so that a single scalar noise prior applies uniformly to all dimensions. In the raw, unstandardised coefficient space, the same estimates span more than two orders of magnitude, from 0.2 to 68, which motivates the standardisation of the coefficients described in Section 4.4. The noise prior variance is set to the mean estimate over the 20 dimensions, $\sigma_0^2 = 0.55$, for both HVBLL and VBLL, and the Wishart prior is configured with $\nu = 10$ degrees of freedom and the corresponding scale $\omega = 2\nu\sigma_0^2$ (Section 3.2). This estimate is computed once on the full dataset and is then held fixed across all sample sizes and repetitions, together with the PCA basis and the standardisation statistics; a control in which all of these are re-estimated on the training subset of each repetition is reported in Appendix B. The smaller value $\sigma_0^2 = 0.01$ is also reported in Table 3 and in Appendix B for comparison.

Appendix D. Datasets

Table D.8 summarises the datasets used in the verification experiments of Section 3 and in the Z-beam and open-hole compression applications; the image regression benchmark is described in Appendix A.1.

Table D.8
Summary of the datasets used in this work.

Dataset	Samples	Input variables	Output variables	Train/test split	Purpose
Synthetic, Eq. (10)	50 or 500 training, 200 testing	x	y	training from $[0, 1]$, testing from $[-0.1, 1.1]$	controlled verification of heteroscedastic noise modelling
Synthetic, Eq. (11)	50 or 500 training, 200 testing	x	y	training from $[-0.15, 0.85]$, testing from $[-0.25, 0.95]$	controlled verification of heteroscedastic noise modelling
Open-hole compression	4000 (subsets of 500, 1000, 4000)	$r_{\text{hole}}, d_{\text{edge}}$	FI_{max}	sample-wise 80%/20% split per repetition	verification on engineering finite-element data
Z-beam representative element	1396 (377 nominal-layup and 1019 randomised-layup; subsets of 100, 500, 1200)	$t_{\text{spar}}, t_{\text{beam}}, \alpha_{\text{beam}}$	FI map (20 PCA coefficients)	sample-wise 80%/20% split per repetition	target application: probabilistic failure envelopes for early-stage design

Appendix E. Training hyperparameters

Table E.9

Training hyperparameters of the four experiments. λ_w and λ_e are the regularisation weights of the last-layer and noise KL terms of Eq. (8). For the image regression and Z-beam experiments, the noise prior variance is estimated from the data as described in Appendix C.

	Toy (Section 3.3)	Image regression (Appendix A.1)	OHC (Appendix A.2)	Z-beam (Section 4.4)
Feature network	3×128	3×500	5×512	5×512
ϕ_θ				
Feature dimension n_f	64	64	128	128
Noise network g_θ (HVBLL)	2×128	2×128	3×128	3×128
Activation function	ELU	ELU	ELU	ELU
Noise prior σ_0^2	0.05	estimated	0.005	0.55
Wishart degrees of freedom ν	1	1	1	10
KL weights (λ_w, λ_e)	(0.01, 0.01)	(0.1, 1.0)	(0.1, 0.1)	(0.01, 0.001)
Batch size	500	512	500	500
Optimiser	Adam ($\beta_1 = 0.5, \beta_2 = 0.999$)			
Initial learning rate	0.01	0.001	0.001	10^{-4}
Learning rate decay ($\times 0.8$ every)	1000 epochs	500 steps	1000 epochs	500 epochs
Gradient clipping (norm)	1.0	1.0	1.0	1.0
Training epochs	3000	80	3000	5000
Stopping criterion	fixed number of epochs (no early stopping)			
Sample size N per repetition	50, 500	50,000	500, 1,000, 4000	100, 500, 1200
Repetitions per sample size	10			

Table E.9 summarises the training hyperparameters for the experiments in Sections 3.3 and 4.4, and for the image regression and open-hole compression experiments of Appendix A. The architecture notation $L \times N$ denotes a network with L hidden layers of N neurons. All models are trained for a fixed number of epochs without early stopping. In the last two rows, N denotes the sample size drawn for each repetition; for the toy and image regression experiments it is the number of training samples, whereas for the open-hole compression and Z-beam experiments it is the size of the random subset that is subsequently split into 80% training and 20% testing samples. All hyperparameters were fixed before the experiments and were kept identical across all sample sizes and repetitions. The only quantities determined from the data are the noise prior variances of the image regression and Z-beam experiments, which are estimated by the procedure of Appendix C rather than tuned on model performance; for the image regression benchmark the estimate is computed on the training subset of each repetition, and for the Z-beam experiment once on the full dataset. No separate validation set was used, and no hyperparameter tuning was performed on the test data. For the open-hole compression experiment, the values in the table correspond to HVBLL; VBLL requires a different configuration to train stably, namely a learning rate of 5×10^{-4} , a noise prior variance of 2.5×10^{-2} , a noise KL regularisation weight of 1.0, and 10,000 training epochs.

Appendix F. Metrics for evaluating model performance

We evaluate model performance using three metrics: Negative Log-Likelihood (NLL), Mean Absolute Error (MAE), and Continuous Ranked Probability Score (CRPS).

Assuming the predictive distribution of the model for a sample (x_i, y_i) is $p(y_i|x_i)$, the NLL is defined as

$$\text{NLL} = -\frac{1}{N} \sum_{i=1}^N \log p(y_i|x_i), \quad (\text{F.1})$$

which measures how well the predicted distribution assigns probability mass to the observed outcomes. Lower values indicate more accurate uncertainty estimation.

MAE evaluates the average absolute difference between predictions $\hat{y}_i$ and targets y_i :

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|, \quad (\text{F.2})$$

where $\hat{y}_i$ is typically the predictive mean. MAE captures pointwise predictive accuracy.

CRPS compares the entire predictive cumulative distribution function (CDF) F_i with the empirical CDF of the observation:

$$\text{CRPS}(F_i, y_i) = \int_{-\infty}^{\infty} (F_i(z) - \mathbf{1}\{y_i \leq z\})^2 dz, \quad (\text{F.3})$$

and the overall score is

$$\text{CRPS} = \frac{1}{N} \sum_{i=1}^N \text{CRPS}(F_i, y_i). \quad (\text{F.4})$$

CRPS generalises MAE to probabilistic forecasts by evaluating the quality of the full predictive distribution.

It should be noted that MAE only reflects point prediction accuracy and does not account for predictive uncertainty. When the ground truth noise level is high, observed values y_i tend to deviate more from the predictive mean $\hat{y}_i$, even if the model correctly captures the underlying distribution. As a result, the MAE naturally increases with larger noise variance. In contrast, metrics such as NLL and CRPS explicitly incorporate predictive uncertainty and therefore provide a more comprehensive evaluation under heteroscedastic settings.

Appendix G. Material properties

The material properties of components of the Z-beam representative element are listed in Tables G.10–G.12.

Table G.10
IM7/8551-7 properties [46].

Elastic properties	$E_{11} = 165 \text{ GPa}$ $E_{22} = E_{33} = 8.4 \text{ GPa}$ $\nu_{12} = \nu_{13} = 0.34$ $\nu_{23} = 0.5$ $G_{12} = G_{13} = 5.6 \text{ GPa}$ $G_{23} = 2.8 \text{ GPa}$
Strength	$X_T = 2560 \text{ MPa}$ $X_C = 1590 \text{ MPa}$ $Y_T = 73 \text{ MPa}$ $Y_C = 185 \text{ MPa}$ $S_{12} = S_{13} = 90 \text{ MPa}$ $ILSS = 90 \text{ MPa}$ $Z_T = 63 \text{ MPa}$
Other properties	$\alpha_0 = 53^\circ$ $\eta_L = 0.082$ $\eta_T = 0.287$

Table G.11
Ti-6Al-4V properties [47].

Elastic properties	$E = 129 \text{ GPa};$ $\nu = 0.33$
Yield strength	$\sigma_y = 1085 \text{ MPa}$

Table G.12
Ti-6Al-4V plastic behaviour
[47].

Plastic strain	True stress
0.00	1085 MPa
0.02	1093 MPa
0.04	1123 MPa
0.06	1155 MPa

Appendix H. Composite stacking sequence

For the Z-beam representative element, the ply thickness is 0.5 mm, and the number of plies is rounded to the nearest even integer. The laminates always contain about 40% of $\pm 45^\circ$ plies, 40% of 0° plies and 20% of 90° plies. The detailed stacking sequence is determined by the composite guidelines [48,49]:

- Use homogeneous laminate stacking sequence (LSS) for strength-controlled designs.
- Have at least four distinct ply angles (e.g., 0° , $\pm 45^\circ$, 90°) with a minimum of 10% of the plies oriented at each angle.
- Use a minimum of 40% of $\pm 45^\circ$ plies.
- LSS should be balanced and symmetric about the mid-plane. Quasi-isotropic patterns $(0^\circ / \pm 45^\circ / 90^\circ)_s$ and $(0^\circ / 45^\circ / 90^\circ / -45^\circ)_s$ are close to optimum.
- Minimise groupings of plies with the same orientation, e.g., stack no more than 4 plies of the same orientation together to reduce or eliminate fringe delamination.
- Alternate $\pm 45^\circ$ plies through the LSS except for the closest ply on either side of the symmetry plane. A $\pm 45^\circ$ pair of plies should be located as closely as possible while still meeting the other guidelines.
- Shield primary load carrying plies from exposed surfaces. For laminates loaded primarily in tension or compression in the 0° direction, the laminate should start with angle and transverse plies from the surface. For laminates loaded primarily in shear, should locate $\pm 45^\circ$ plies away from the surface.

Appendix I. Mesh convergence analysis

This appendix presents the mesh convergence study of the Z-beam representative element. Fig. I.14 shows the finite element mesh of the global wingbox RSE model. This model is used to compute the homogenised stiffness matrix and to provide boundary conditions for the refined local model within the global-local framework. Fig. I.15 shows the local finite element model of the connection plates.

To ensure the accuracy of the local finite element model used for failure prediction, a two-stage mesh convergence study was performed focusing on the bolt-hole region, where stress concentrations govern

failure initiation. The results and final selected mesh are presented in Fig. I.16.

First, a radial mesh refinement study was conducted using six different mesh densities by varying the element size and bias within the circular partition. All meshes consistently predicted the same critical failure location, indicating mesh-independent localisation. However, slight differences were observed in the predicted failure onset. To quantify convergence, the stress and matrix cracking failure index (FI_1) were extracted from the same element corresponding to the failure location in the most refined mesh. Extracting the values from the same element was enabled by keeping a constant number of circumferential elements (32). The results show clear convergence, with the relative stress difference between the two finest meshes below 1%. Considering computational efficiency, the second finest mesh was selected, reducing computation time by approximately 15% while maintaining comparable accuracy. Subsequently, a circumferential refinement study was performed based on the selected mesh. Increasing the number of elements along the circumference showed negligible influence on the predicted failure indices.

Overall, the adopted mesh provides a good balance between accuracy and computational cost, and is sufficient for reliable failure prediction in the proposed global-local framework.

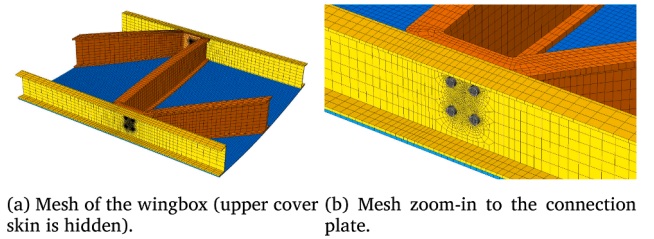


Fig. I.14. Finite element mesh of the global wingbox RSE [43].

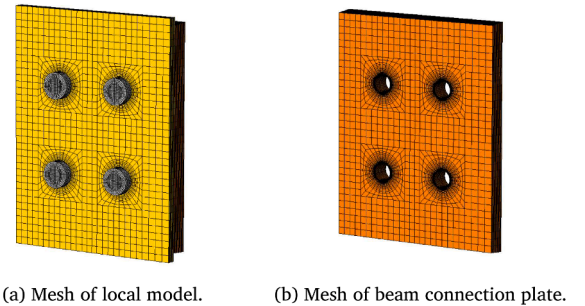


Fig. I.15. Finite element mesh of the local connection plate model [43].

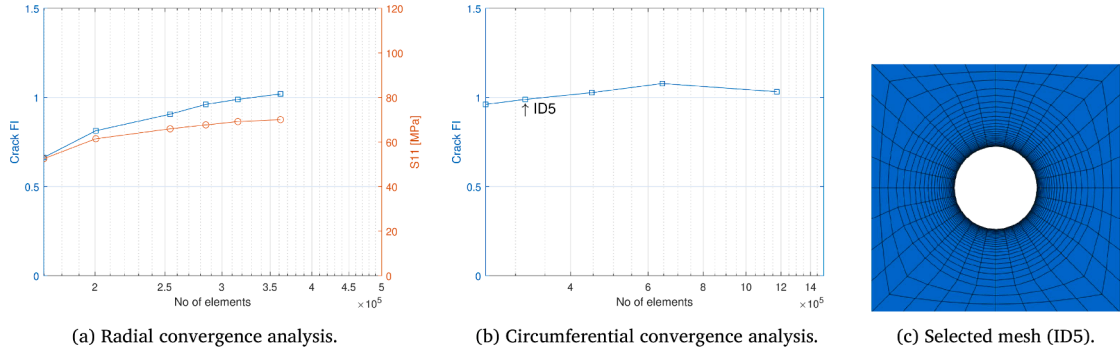


Fig. I.16. Mesh convergence analysis [43].

Appendix J. Open-hole compression test specimens

The specimen is a Carbon Fibre Reinforced Plastic (CFRP) plate with a central hole, designed for the open-hole compression (OHC) test. As shown in Fig. J.17, the plate has overall dimensions $l_x \times l_y \times t_z = 50 \text{ mm} \times 100 \text{ mm} \times 4 \text{ mm}$. The hole has radius r_{hole} , with its centre located at $(l_x x_r, l_y y_r)$ and $y_r = 0.5$. The boundary conditions are defined by clamping the face at $y = 0$ and applying a compressive displacement at the face $y = l_y$, corresponding to a nominal strain of 0.001.

The design variables are the relative x -coordinate of the hole centre $x_r \in [0.3, 0.5]$ and the hole radius $r_{\text{hole}} \in [5, 12] \text{ mm}$. The main source of uncertainty is the detailed stacking sequence of the laminate, consisting of 40 plies with ply thickness 0.1 mm. The output is the maximum failure index in the specimen, FI_{max} . The specimen uses IM7/8551-7 carbon/epoxy composite, it is simulated with a ply-by-ply mesh, and failure indices (FI) are computed according to the LaRC05 failure criteria [42].

The geometric design variables are uniformly sampled within their range, the composite layups are drawn from a predefined set of candidate configurations that satisfy engineering guidelines, as detailed in Appendix H. After simulation, the dataset of $\{(d_{\text{edge}}, r_{\text{hole}}), \text{FI}_{\text{max}}\}_i$ is obtained, where the edge distance is $d_{\text{edge}} = x_r l_x - r_{\text{hole}}$.

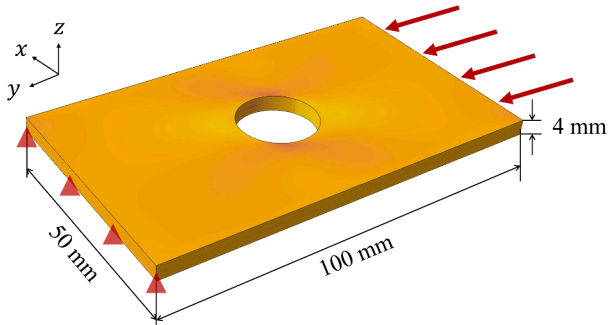


Fig. J.17. Open-hole compression specimen.

Appendix K. Prediction accuracy in engineering units and uncertainty decomposition

The metrics of Table 3 are computed in the PCA coefficient space, because that is where the model output is defined. Those numbers are convenient for comparing models, but they do not tell a designer how large the error is in the quantities that matter for structural assessment. This appendix therefore answers two practical questions: how accurate is the predicted failure envelope in physical terms, and where does the width of the predicted uncertainty band come from. All results below use the largest sample size ($N = 1200$, i.e. 960 training and 240 test samples) and the ten repetitions of Section 4.4, evaluated on the test samples of

each repetition. Values are given as mean $\pm$ standard deviation over the ten repetitions.

K.1. Accuracy of the predicted mean in engineering units

Two levels of accuracy are reported. The first is the error of the predicted failure index itself, obtained by reconstructing the predicted FI map from the predicted PCA coefficients and comparing it point by point with the true map. The second is the error of the failure envelope, which is the quantity the design rule of Section 4.5 actually acts on. For a given loading direction, the envelope radius is the strain scale at which the failure index reaches unity; it is found by linear interpolation along the radial direction, using the same procedure that produced the reconstruction error of Section 4.3. The results are collected in Table K.13.

A point that must be kept in mind when reading these numbers is that the surrogate does not predict the failure envelope of one particular laminate. It predicts the mean of the failure response over the layup population, whereas each test sample corresponds to one specific layup. The difference between the two therefore contains the layup-to-layup scatter by construction, and no mean function, however accurate, could remove it. To place the surrogate error against this floor, the same statistic was measured directly on the database: the samples were grouped by their exact geometry, and within each group the deviation of each layup from the mean of its own group was computed. This uses the 259 geometries for which the database contains at least two layups, with 4.9 layups per geometry on average.

The outcome is that the predicted mean envelope differs from the envelope of an individual test layup by 10.5% of the envelope radius on average, while different layups of the same geometry already differ from their own mean by 8.9%. Most of the apparent error is therefore the layup variability that the framework is designed to describe as a distribution, not a deficiency of the mean prediction. Expressing both quantities as root mean squares, which is the form in which independent contributions add, gives 13.1% and 11.2%; removing the layup scatter in quadrature leaves a residual of 6.8% attributable to the mean function itself. The PCA compression contributes a further 1.1% (Section 4.3) and is thus a minor term. The bias of the envelope prediction is small, +0.9%, so the predicted mean envelope is not systematically optimistic or conservative.

K.2. Decomposition of the predictive uncertainty

The predictive variance of Eq. (4) is the sum of two terms with different meanings. The epistemic term $\phi^T S_w \phi$ reflects what the model does not know because it has seen a finite amount of training data; it would shrink if more data were added. The aleatoric term $\Sigma_\epsilon(\mathbf{x})$ reflects the scatter of the failure response over the layup population at that input; it would not shrink with more data of the same kind, because it is a property of the problem rather than of the model. Since the framework rests

Table K.13

Accuracy of the predicted mean in engineering units, at $N = 1200$ over ten repetitions. The critical region collects the grid points whose true failure index lies in $[0.9, 1.1]$. The envelope-radius errors are relative to the true envelope radius and are evaluated per sample and loading direction. The layup-scatter row is not a model error: it is the spread of individual layups about the mean of their own geometry, measured directly on the database, and is the floor below which the comparison with an individual layup cannot go.

Quantity	Measure	Value
Failure index	mean absolute error, all grid points	0.085 ± 0.003
	mean absolute error, critical region	0.110 ± 0.004
	bias, critical region	$+0.022 \pm 0.008$
Envelope radius	mean relative error	$10.5\% \pm 0.3\%$
	relative bias	$+0.9\% \pm 0.8\%$
	95th percentile of relative error	$25.1\% \pm 0.7\%$
	root mean square relative error	13.1%
Reference levels	layup scatter within a geometry (mean / rms)	$8.9\% / 11.2\%$
	PCA compression alone (mean)	1.1%
Mean function	residual after removing layup scatter (rms)	6.8%

on the premise that the unknown layup is the dominant source of uncertainty, it is worth checking that the trained model actually behaves that way.

The two terms were extracted separately from the trained networks and compared. In the PCA coefficient space the aleatoric term accounts for $99.2\% \pm 0.1\%$ of the total predictive variance, and the ratio of the two standard deviations is $\sigma_\epsilon / \sigma_{\text{epistemic}} = 14.7 \pm 0.6$. Mapping both terms to the FI map grid through Eq. (13) gives the same picture, $99.2\% \pm 0.1\%$. The dominance is not an artefact of averaging: even at the individual output and sample where the epistemic term is largest, the aleatoric share is still about 75–79%.

The practical meaning is that the width of the predicted envelopes in Figs. 11 and 12 is set almost entirely by the fact that the layup is not yet chosen, and only marginally by the amount of training data. Collecting more samples of the same kind would sharpen the mean prediction but would leave the bands almost as wide; the bands would only narrow if the layup itself were constrained, which is precisely what the detailed layup optimisation of the later design stages does. This also confirms that interpreting the spread between the best-case and worst-case envelopes as layup-induced variability, as done throughout Section 4, is consistent with what the model has learned.

As a check on the decomposition, the two terms were verified to sum to the total predictive variance of Eq. (4) to within a relative error of 2×10^{-7} , and the runs used for the decomposition reproduce the predictions of the runs used for the calibration of Appendix B exactly, so that the decomposition refers to the same trained models.

References

- Y. Ma, A. Elham, Designing high aspect ratio wings: a review of concepts and approaches, *Prog. Aerosp. Sci.* 145 (2024) 100983.
- A. Kretov, D. Tiniakov, Evaluation of the mass and aerodynamic efficiency of a high aspect ratio wing for prospective passenger aircraft, *Aerospace* 9 (9) (2022) 497.
- J. Xiong, S. Yi, H. Wang, L. Shen, Gust alleviation of a high-aspect-ratio flexible wing by passive shape morphing via a multistable shell, *Aerosp. Sci. Technol.* 177 (2026) 112223.
- M. Ahmadi, T. Farsadi, Multidisciplinary optimization of high aspect ratio composite wings with geometrical nonlinearity and aeroelastic tailoring, *Aerosp. Sci. Technol.* 145 (2024) 108849.
- M.M. Navardi, H. Shahverdi, V. Kalafi, V.P. Zanussi, Nonlinear aeroelastic tailoring analysis of composite doubly-curved shallow panels with curvilinear fibers, *Aerosp. Sci. Technol.* (2026) 112391.
- M. Lillico, R. Butler, M. Holden, Conceptual design optimization of composite wings with aeroelastic and strength constraints, in: 6th Symposium on Multidisciplinary Analysis and Optimization, 1996, pp. 4004.
- F. Toffol, S. Ricci, A meta-model for composite wingbox sizing in aircraft conceptual design, *Compos. Struct.* 306 (2023) 116557.
- B. Liu, R.T. Haftka, M.A. Akgün, Two-level composite wing structural optimization using response surfaces, *Struct. Multidiscip. Optim.* 20 (2) (2000) 87–96.
- S. Guo, Aeroelastic optimization of an aerobatic aircraft wing structure, *Aerosp. Sci. Technol.* 11 (5) (2007) 396–404.
- C.B. York, Buckling analysis and minimum-mass design procedures for composite wing-box structures, *J. Aircr.* 43 (2) (2006) 528–536.
- M. McGurk, O. Stodieck, J. Yuan, Probabilistic aeroelastic analysis of high-fidelity composite aircraft wing with manufacturing variability, *Compos. Struct.* 329 (2024) 117794.
- A. Manan, J. Cooper, Design of composite wings including uncertainties: a probabilistic approach, *J. Aircr.* 46 (2) (2009) 601–607.
- Y. Liu, S. Date, T. Nagashima, T. Okabe, Y. Abe, Effects of aeroelastic coupling accuracy and geometrical nonlinearity on performances of optimized composite wings, *Aerosp. Sci. Technol.* 158 (2025) 109926.
- D. Li, S. Guo, Optimization of composite wing structure for a flying wing aircraft subject to multi constraints, in: 54th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, 2013, pp. 1934.
- T.R. Brooks, G. Kennedy, J.R. Martins, High-fidelity multipoint aerostructural optimization of a high aspect ratio tow-steered composite wing, in: 58th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, 2017, pp. 1350.
- S. Date, Y. Abe, T. Okabe, Effects of fiber properties on aerodynamic performance and structural sizing of composite aircraft wings, *Aerosp. Sci. Technol.* 124 (2022) 107565.
- S.T. IJsselmuiden, M.M. Abdalla, Z. Gürdal, Implementation of strength-based failure criteria in the lamination parameter design space, *AIAA J.* 46 (7) (2008) 1826–1834.
- A. Khani, S.T. IJsselmuiden, M.M. Abdalla, Z. Gürdal, Design of variable stiffness panels for maximum strength using lamination parameters, *Compos. B: Eng.* 42 (3) (2011) 546–552.
- J.K.S. Dillinger, T. Klimmek, M.M. Abdalla, Z. Gürdal, Stiffness optimization of composite wings with aeroelastic constraints, *J. Aircr.* 50 (4) (2013) 1159–1168.
- C. Scarth, J.E. Cooper, Reliability-based aeroelastic design of composite plate wings using a stability margin, *Struct. Multidiscip. Optim.* 57 (4) (2018) 1695–1709.
- Y. Dong, I. Hussain, S. He, Structural topology optimization of aircraft wing leading edge fabricated of multilayer composites, *Aerosp. Sci. Technol.* 159 (2025) 109993.
- G. Wang, L. Zhang, S. Xuan, X. Fan, B. Fu, X. Xue, X. Yao, An efficient surrogate model for damage forecasting of composite laminates based on deep learning, *Compos. Struct.* 331 (2024) 117863.
- Y.-W. Lu, Y. Wang, X. Liu, C. Zhang, C.-Y. Hsu, Physics-Informed machine learning for failure mode proportion prediction in composite adhesive joints, *J. Manuf. Sci. Eng.* 147 (8) (2025) 081007.
- A.A.F. Ogaili, M.N. Hamzah, A.A. Jaber, Integration of machine learning (ML) and finite element analysis (FEA) for predicting the failure modes of a small horizontal composite blade, *Int. J. Renew. Energy Res.* 12 (4) (2022) 2168–2179.
- Q. Wang, H. Qin, L. Jia, Z. Li, G. Zhang, Y. Li, Y. Liu, Failure prediction and optimization for composite pressure vessel combining FEM simulation and machine learning approach, *Compos. Struct.* 337 (2024) 118099.
- R.S. Chahar, T. Mukhopadhyay, Multi-fidelity machine learning based uncertainty quantification of progressive damage in composite laminates through optimal data fusion, *Eng. Appl. Artif. Intell.* 125 (2023) 106647.
- O.A.I. Azeem, S.T. Pinho, A machine learning assisted multifidelity modelling methodology to predict 3D stresses in the vicinity of design features in composite structures, *Int. J. Solids Struct.* 301 (2024) 112946.
- N.L. Iorga, Aerofoil structures with stiffness properties adjustment component, 2023, US Patent 11,772,780.
- M.H. Shirk, T.J. Hertz, T.A. Weisshaar, Aeroelastic tailoring-theory, practice, and promise, *J. Aircr.* 23 (1) (1986) 6–18.
- M. Miranda, R. Li, L. Iorga, S.T. Pinho, Multiscale modelling strategy for a novel wingbox structure with increased bend-twist coupling, *Compos. Struct.* 370 (2025) 119291.
- M. Miranda, A. Cea, R. Li, R. Palacios, S.T. Pinho, Efficient approach for early failure prediction in high-aspect-ratio wings during gust encounters, in: AIAA SCITECH 2026 Forum, 2026, pp. 2430.
- R.C. Smith, Uncertainty Quantification: Theory, Implementation, and Applications, SIAM, 2024.
- G. Papamakarios, E. Nalisnick, D.J. Rezende, S. Mohamed, B. Lakshminarayanan, Normalizing flows for probabilistic modeling and inference, *J. Mach. Learn. Res.* 22 (57) (2021) 1–64.
- Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: representing model uncertainty in deep learning, in: International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059.
- Q.V. Le, A.J. Smola, S. Canu, Heteroscedastic gaussian process regression, in: Proceedings of the 22nd International Conference on Machine Learning, 2005, pp. 489–496.
- R.M. Neal, Bayesian Learning for Neural Networks, Vol. 118, Springer Science & Business Media, 2012.
- J. Harrison, J. Willes, J. Snoek, Variational Bayesian last layers, arXiv:2404.11599 (2024).
- M.D. Hoffman, D.M. Blei, C. Wang, J. Paisley, Stochastic variational inference, *J. Mach. Learn. Res.* 14 (1) (2013) 1303–1347.
- D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, (2014). <https://doi.org/10.48550/arXiv.1412.6980>
- A. Immer, E. Palumbo, A. Marx, J. Vogt, Effective bayesian heteroscedastic regression with deep neural networks, *Adv. Neural Inf. Process. Syst.* 36 (2023) 53996–54019.
- H. Hesse, R. Palacios, Reduced-order aeroelastic models for dynamics of maneuvering flexible aircraft, *AIAA J.* 52 (8) (2014) 1717–1732.

- [42] S.T. Pinho, R. Darvizeh, P. Robinson, C. Schuecker, P.P. Camanho, Material and structural response of polymer-matrix fibre-reinforced composites, *J. Compos. Mater.* 46 (19–20) (2012) 2313–2341.
- [43] R. Li, M. Miranda, L. Iorga, S.T. Pinho, Physics-guided machine learning for the failure prediction of a novel Z-beam composite wing structural concept, *Compos. Struct.* 394 (2026) 120662.
- [44] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE* 86 (11) (1998) 2278–2324.
- [45] M. Seitzer, A. Tavakoli, D. Antic, G. Martius, On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks, (2022). [arXiv:2203.09168](https://arxiv.org/abs/2203.09168)
- [46] A.S. Kaddour, M.J. Hinton, Input data for test cases used in benchmarking triaxial failure theories of composites, *J. Compos. Mater.* 46 (19–20) (2012) 2295–2312. <https://doi.org/10.1177/0021998312449886>
- [47] X. Liu, X. Xi, C. Bai, J. Yang, X. Yang, Dynamic response and failure mechanism of Ti-6AL-4V hi-lock bolts under combined tensile-shear loading, *Int. J. Impact Eng.* 131 (2019) 140–151. <https://doi.org/https://doi.org/10.1016/j.ijimpeng.2019.04.025>
- [48] P.K. Mallick, *Composites Engineering Handbook* (1ed), CRC Press, 1997, Ch. 12. <https://doi.org/https://doi.org/10.1201/9781482277739>
- [49] M.C.-Y. Niu, *Composite Airframe Structures: Practical Design Information and Data*, Conmlit Press LTD, Hong Kong, 2005, Ch. 5. <https://doi.org/https://doi.org/10.1017/S0001924000012471>