The 9th European Congress on Computational Methods in Applied Sciences and Engineering
ECCOMAS Congress 2024
3–7 June 2024, Lisboa, Portugal

MULTI-FIDELITY SURROGATE MODEL FOR EFFICIENT AERODYNAMIC PREDICTIONS

J. NIETO-CENTENERO[1,2], A. MARTÍNEZ-CAVA[1,3] AND E. ANDRÉS[2]

[1] Universidad Politécnica de Madrid (UPM)
Plaza Cardenal Cisneros 3, E-28040 Madrid, Spain

[2] Spanish National Institute for Aerospace Technology (INTA)
Theoretical and Computational Aerodynamics Branch, Flight Physics Department
Institute for Aerospace Technology (INTA), Torrejón de Ardoz, Spain

[3] Instituto Universitario "Ignacio Da Riva" (IDR/UPM)
Universidad Politécnica de Madrid, Plaza Cardenal Cisneros 3, E-28040 Madrid, Spain

Key words: Multi-Fidelity, Surrogate Modelling, Reduced Order Model, Machine Learning

Summary. This study presents a Multi-Fidelity surrogate model designed to improve the accuracy and efficiency of aerodynamic data prediction, particularly for the pressure coefficient (C_p) distribution over airfoils. The methodology begins by increasing both the dimensionality and fidelity of baseline data through the integration of Low-Fidelity data from inviscid panel method and High-Fidelity data from a RANS-based CFD database for a set of NACA 4-digit airfoils. Multi-Fidelity Gaussian Process Regression (MFGPR) is used to combine a high-dimensional, Low-Fidelity C_p vector with a sparse, High-Fidelity C_p vector, where the sparsity simulates the limited pressure tap measurements typically available in wind tunnel experiments. The next phase is to develop a regression model that predicts the C_p distribution with the improved fidelity provided by the new Multi-Fidelity database, using the airfoil geometry as input. To improve computational efficiency, Locally Linear Embedding (LLE) is used to map the C_p data into a low-dimensional space that retains the essential physical features of the system, allowing a neural network to be trained at a significantly reduced computational cost. This approach highlights the potential of Multi-Fidelity databases to provide a more accurate and efficient framework for aerodynamic prediction.

1 INTRODUCTION

In the field of aerodynamics, the calculation of flow parameters relies on multiple sources of information, each with distinct strengths and limitations: Computational Fluid Dynamics (CFD), Wind Tunnel Tests (WTT), and Flight Test Data [1]. In industry, CFD is widely used, starting with panel methods for preliminary analysis, followed by more accurate but computationally intensive Reynolds Average Navier-Stokes (RANS) simulations. CFD provides comprehensive pressure distributions over the entire surface under study. In contrast, WTT offer localized measurements constrained by the number of sensors [2], but deliver high accuracy. Flight Test Data offers the highest fidelity by reflecting real-world performance, but its high cost limits its use to the final stages of design.

In light of the inherent limitations of aerodynamic data sources in terms of precision, resolution, and cost, it is essential to integrate these sources through Multi-Fidelity approaches. The most commonly used data-fusion techniques in aerodynamics include Gappy Proper Orthogonal Decomposition (GPOD) [3, 13] and Multi-Fidelity Gaussian Process Regression (MFGPR) [14, 9, 11]. This study employs MFGPR, a statistical method that combines data from multiple levels of fidelity by modeling their relationships through Gaussian processes, thereby enhancing prediction accuracy and optimizing computational efficiency for more robust aerodynamic analysis.

Furthermore, manifold learning has emerged as a promising tool for dimensionality reduction and the extraction of relevant features in high-dimensional aerodynamic datasets. The most frequently used non-linear techniques in aerodynamics are Locally Linear Embedding (LLE) [15] and Isometric Feature Mapping (Isomap) [19]. For this study, we employ LLE due to its effectiveness in preserving local structures within the data while reducing dimensionality. Recent applications have demonstrated that integrating manifold learning can significantly enhance the efficiency of aerodynamic regression models without compromising accuracy [4, 6].

The framework presented in this study introduces a novel Multi-Fidelity approach to predict aerodynamic performance, comprising two principal stages: data generation and regression. A key innovation of this work lies in its integration of sparse High-Fidelity and dense Low-Fidelity aerodynamic data, forming a comprehensive dataset of pressure coefficient (C_p) vectors. The regression stage begins with the application of LLE, which is employed to reduce the dimensionality of the enhanced dataset while ensuring the retention of critical aerodynamic features. Subsequently, the LLE-derived embedding is employed to train a Multilayer Perceptron (MLP) to map NACA 4-digit airfoil parameters, while a k-nearest neighbors (k-NN) decoder reconstructs the C_p distributions. Ultimately, this approach aims to create a surrogate model capable of accurately predicting the C_p distribution for airfoils outside the training set. This methodology not only ensures highly accurate aerodynamic predictions, but also provides a robust and computationally efficient solution for complex aerodynamic analyses.

The remainder of this paper is organized as follows: Section 2 details the methodology, including the construction of the Multi-Fidelity database, the use of manifold learning techniques, and the application of the Deep Neural Network (DNN) regressor. Section 3 presents the results of the aerodynamic predictions, evaluating the performance of the proposed framework. Finally, the conclusions are drawn in Section 4.

2 METHODOLOGY

This section outlines the methodology, structured according to the pipeline depicted in Figure 1. The first stage of the pipeline is dedicated to generating the Multi-Fidelity database, with a comprehensive description of the databases provided in 2.1. Subsequently, a concise overview of the MFGPR model, used to enhance the database, is presented in 2.2. The second stage of the pipeline focuses on the construction of a predictor based on manifold learning, 2.3. The Multi-Fidelity database is integrated into the predictor through the application of the LLE method, as detailed in 2.3.1. The backmapping procedure, which recovers high-dimensional data from the latent space of the LLE, is detailed in 2.3.2. To complete the pipeline, in 2.4, we discuss the neural network that performs regression between the geometric parameters of a NACA 4-digit airfoil and the latent space variables.

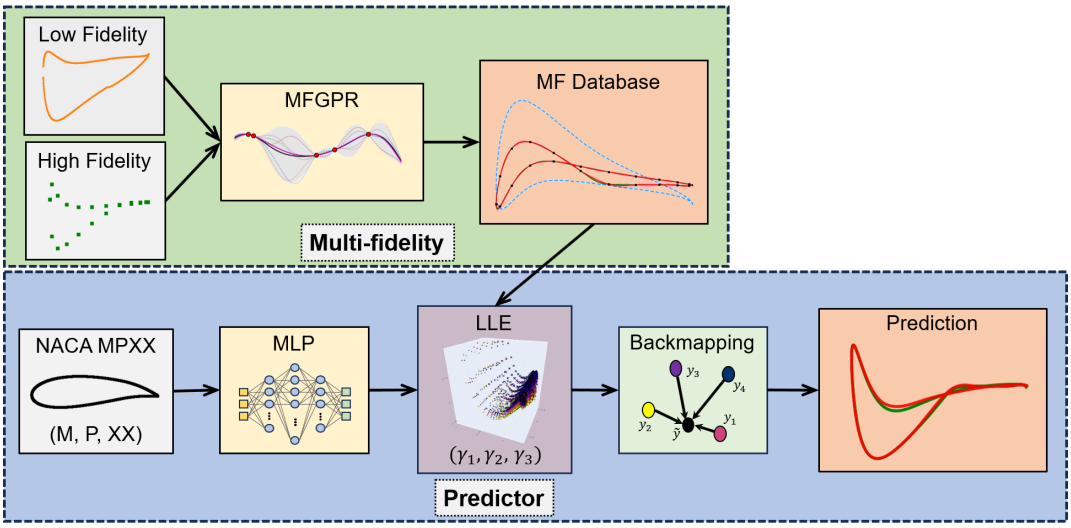

Figure 1: Schematic of the Multi-Fidelity pipeline, combining Low- and High-Fidelity data through MFGPR to create a database. This database is then used in a manifold learning-based predictor for airfoil performance estimation.

2.1 Database

This study examines the efficacy of a Multi-Fidelity regression methodology utilizing two datasets with distinct levels of fidelity. The datasets comprise Cp vectors gathered from CFD simulations of 1809 NACA 4-digit airfoils, all evaluated at a constant angle of attack of 10 degrees. The airfoils are characterized by three parameters: M, P, and XX. The parameter M represents the maximum camber, expressed as a percentage of the chord length, with values ranging from 2 to 9. Similary, parameter P denotes the maximum camber position, expressed as a tenth of the chord, with values ranging from 4 to 8. The parameter XX indicates the thickness of the airfoil, expressed as a percentage, with values ranging from 05 to 50. These parameters collectively define the geometric characteristics of the airfoils and provide a robust dataset for evaluating the Multi-Fidelity framework.

2.1.1 High-Fidelity

The High-Fidelity dataset is composed of RANS simulations of turbulent airflow, conducted at a Reynolds number of 3×10^6. The dataset is openly available on Zenodo [16, 17]. Technical details of the dataset generation may be consulted in Schillaci et al [18].

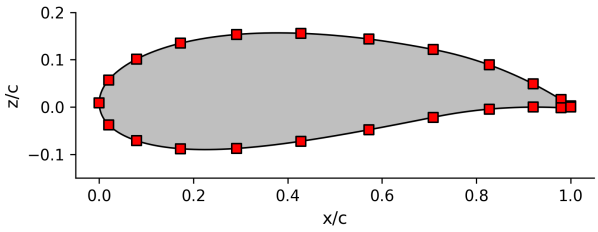

Figure 2: Virtual pressure tap locations on a NACA 5724 airfoil.

Each simulation provides 1,500 C_p values along the airfoil surface. In order to emulate WTT conditions, where the number of pressure tap measurements is more restricted, a selection of 24 virtual pressure taps was implemented based on an extensive literature review, as illustrated in 2. The taps are arranged in a cosine distribution along the chord of the airfoil, thereby providing a higher resolution in the regions near leading and trailing edges.

2.1.2 Low-Fidelity

The Low-Fidelity dataset is generated through inviscid simulations using the panel method implemented in XFOIL [7]. These simulations were conducted for the same 1809 airfoil geometries as in the High-Fidelity dataset, producing a vector of 400 C_p values for each airfoil. Although the inviscid panel method is generally successful in capturing the overall aerodynamic trends, it does exhibit considerable deviations from the High-Fidelity results due to the absence of viscous effects.

Despite the availability of relatively inexpensive viscous panel method simulations, the inviscid approach was deliberately chosen for this study. The rationale behind this choice is to demonstrate that even when Low-Fidelity data shows significant discrepancies from High-Fidelity results, it can still offer valuable insights if it accurately captures the fundamental physical trends. This approach not only tests the robustness of the Multi-Fidelity regression framework, but also underscores the significant role that Low-Fidelity models can play in enhancing overall predictive accuracy.

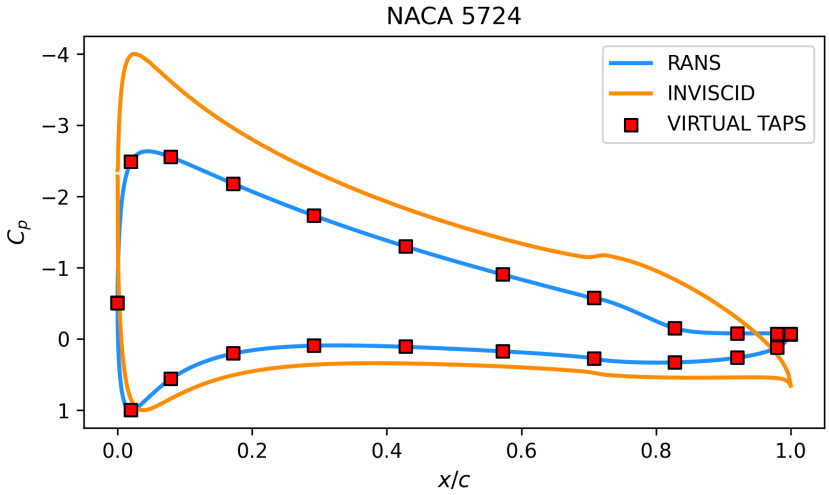

Figure 3: Distribution of C_p along the chord length (x/c) for the NACA 5724 airfoil. The **RANS** (—) data are obtained from RANS simulations, while the **INVISCID** (—) data are derived from the inviscid panel method. **VIRTUAL TAPS** (■) mark the locations of virtual pressure taps.

Figure 3 illustrates the inputs for the MFGPR method for a particular airfoil. The orange line shows Low-Fidelity data from the inviscid panel method, while the red points represent High-Fidelity virtual pressure measurements from RANS simulations. The RANS data, shown as the blue line, are considered the ground truth, and the C_p values from these simulations are used to assess the accuracy of the Multi-Fidelity framework.

2.2 Multi-Fidelity Gaussian Process Regression

2.2.1 Gaussian Process Regression

Gaussian Process Regression (GPR) is a non-parametric Bayesian approach used for regression tasks. It defines a distribution over functions, where any finite subset of function values follows a joint Gaussian distribution. A Gaussian process is fully specified by its mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$. For simplicity, the mean function is often assumed to be zero, leading to the notation $f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$.

Given training data $\mathbf{y}$ and input points $\mathbf{X}$, the joint distribution of the observed values $\mathbf{y}$ and the function values $\mathbf{f}_*$ at new points $\mathbf{X}_*$ is:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix}\right),$$

where $\mathbf{K}$ denotes covariance matrices computed using a kernel function $k(\mathbf{x}, \mathbf{x}')$.

By deriving the conditional distribution, we arrive at the predictive equations for GPR as $\bar{\mathbf{f}}_*|\mathbf{X}, \mathbf{y}, \mathbf{X}_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \text{cov}(\mathbf{f}_*))$, where the predictive mean and covariance at new points $\mathbf{X}_*$ are given by:

$$\bar{\mathbf{f}}_* = K(\mathbf{X}_*, \mathbf{X})[K(\mathbf{X}, \mathbf{X})]^{-1}\mathbf{y}, \tag{1}$$

$$\text{cov}(\mathbf{f}_*) = K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X})[K(\mathbf{X}, \mathbf{X})]^{-1}K(\mathbf{X}, \mathbf{X}_*). \tag{2}$$

Hyperparameters, including kernel parameters, are optimized by maximizing the log marginal likelihood:

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^\mathsf{T}[K(\mathbf{X}, \mathbf{X})]^{-1}\mathbf{y} - \frac{1}{2}\log|K(\mathbf{X}, \mathbf{X})| - \frac{n}{2}\log(2\pi). \tag{3}$$

For a detailed mathematical treatment and further insights into GPR, refer to Williams and Rasmussen [20] and Gramacy [8].

2.2.2 Linear Autoregressive Gaussian Process

GPR can be extended to construct probabilistic models that allows the combination of variable fidelity information sources. In the Linear Autoregressive Gaussian Process model [10], with s fidelity levels, the data can be sequentially organized by fidelity level t, represented as $D_t = \mathbf{x_t}, \mathbf{y_t}$ for $t = 1, \ldots, s$. The model that relates two consecutive fidelity levels is expressed as:

$$f_t(\boldsymbol{x}) = \rho_t f_{t-1}(\boldsymbol{x}) + \delta_t(\boldsymbol{x}) \ , \tag{4}$$

where $\delta_t(\mathbf{x})$ is a Gaussian process independent of $f_{t-1}(\mathbf{x}), \ldots, f_0(\mathbf{x})$ with mean μ_{δ_t} and covariance function k_{δ_t}, and ρ_t represents a scale factor between $f_t(\mathbf{x})$ and $f_{t-1}(\mathbf{x})$.

By adopting the recursive inference scheme proposed by Le Gratiet and Garnier [12], the inference problem is decoupled into s standard GPR problems, yielding the Multi-Fidelity posterior distribution with the predictive mean and variance at each level given by:

$$\bar{\mathbf{f}}_{*\mathbf{t}} = \rho_t \bar{\mathbf{f}}_{*\mathbf{t-1}} + \mu_{\delta_t} + K\left(\mathbf{X}_*, \mathbf{X_t}\right)\left[K(\mathbf{X_t}, \mathbf{X_t})\right]^{-1}\left[\mathbf{y_t} - \rho_t \bar{\mathbf{f}}_{*\mathbf{t-1}} - \mu_{\delta_t}\right] \ , \tag{5}$$

$$\operatorname{cov}(\mathbf{f}_{*\mathbf{t}}) = \rho_t^2 \operatorname{cov}(\mathbf{f}_{*\mathbf{t-1}}) + K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X_t}) \left[K(\mathbf{X_t}, \mathbf{X_t})\right]^{-1} K(\mathbf{X_t}, \mathbf{X}_*) \ . \tag{6}$$

The Matérn 3/2 kernel is used across all fidelities due to its effectiveness in modeling smooth functions while accommodating non-linear behavior, which is essential for capturing the significant pressure drops in the C_p profile. The Matérn 3/2 kernel is defined as $k_{\text{Matérn } 3/2}(\mathbf{x}, \mathbf{x}') = \sigma^2 \left(1 + \sqrt{3}r/\ell\right) \exp\left(-\sqrt{3}r/\ell\right)$, where $r = \|\mathbf{x} - \mathbf{x}'\|$ is the Euclidean distance, ℓ is the length-scale parameter and σ^2 is the variance.

This study employs Automatic Relevance Determination (ARD) to adjust scaling for each input variable, specifically the airfoil geometry components (x and y coordinates). ARD fine-tunes kernel hyperparameters for each variable, enhancing the GPR model's ability to capture specific feature characteristics and improve accuracy and robustness.

Hyperparameters are determined by maximizing the log-likelihood (Equation 3). Due to the non-convex nature of the log-likelihood, optimization is initialized with multiple values and restarted five times to ensure robust results. It is important to note that this optimization process is prone to overfitting. To mitigate this, constraints are imposed on the kernel hyperparameters.

2.3 Manifold Learning via LLE

Locally Linear Embedding [15] is a non-linear dimensionality reduction technique designed to preserve the local geometric structure of high-dimensional data. In the context presented in this work, LLE is employed to embed C_p of the Multi-Fidelity dataset into a low-dimensional space. This serves two primary objectives: first, to visualize the correlation between the parameters of the NACA 4-digit airfoil and the latent space variables; and second, to use these new variables for training a cost-effective surrogate model.

2.3.1 Locally Linear Embedding

LLE is applied to a dataset $X \in \mathbb{R}^{P \times N}$, where each data point $\mathbf{x}_i \in \mathbb{R}^P$ represents the pressure coefficients of a NACA 4-digit airfoil, and N is the total number of airfoils in the training dataset. The algorithm first identifies the k nearest neighbors for each $\mathbf{x}_i$, assuming that the data points and their nearest neighbors lie on a locally linear manifold. LLE then computes the weights that best reconstruct each data point as a linear combination of its neighbors, minimizing the reconstruction error under the constraint that the weights sum to one. The low-dimensional embedding is obtained by finding the coordinates that minimize the same reconstruction error, ensuring that the local relationships captured by the weights are preserved. This involves solving an eigenvalue problem, where the embedding corresponds to the eigenvectors associated with the smallest non-zero eigenvalues.

The LLE algorithm is conditioned by two fundamental parameters: embedding dimensionality, d, and the number of neighbors, k. In this work, the latent space is predefined to have three dimensions (γ_1, γ_2, γ_3). This selection is based on the assumption that, under fixed flow conditions, the C_p on the airfoil surface is determined exclusively by its geometry. For NACA 4-digit airfoils, this geometry is characterized by three variables (M, P, XX), which makes a three-dimensional latent space suitable for capturing the relevant physical phenomena. Figure 4 demonstrates that these LLE components are highly correlated with the geometric attributes of the airfoil. The symbols in the figure represent the distribution of the data for varying the maximum camber (M), showing a pronounced negative correlation with γ_1. The color gradient

indicates that the position of maximum camber (P) decreases as γ_3 increases. Furthermore, the size of the symbols reflects the thickness (XX), which increases along the γ_2 axis within each group of airfoils sharing the same M.

To ensure robustness against noise, the selection of the number of nearest neighbors k is critical. This study uses the Residual Variance (RV) algorithm [19] to determine the optimal k, which quantifies the amount of information that remains unexplained after reducing the data to a lower-dimensional embedding. The goal is to minimize the RV while maintaining the integrity and smoothness of the manifold.

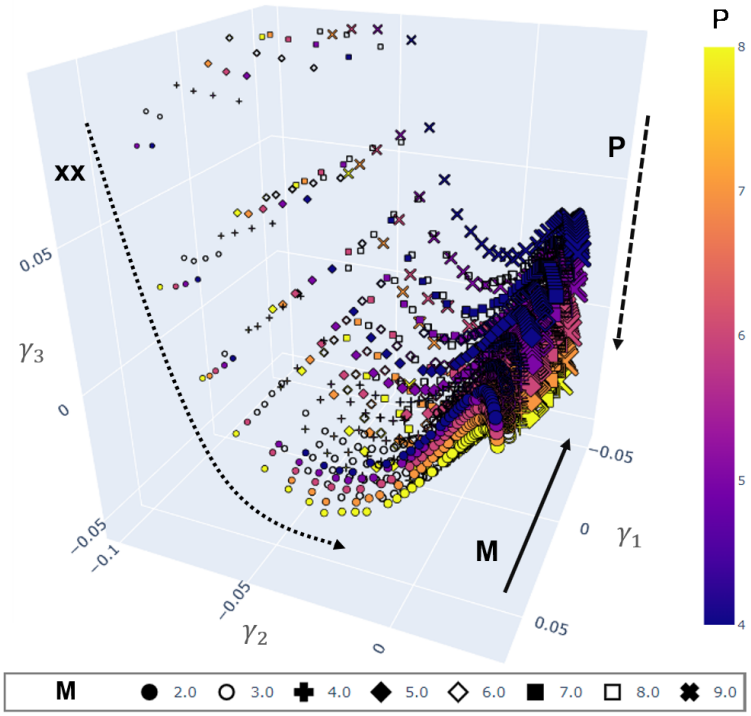

Figure 4: LLE embedding. Representation of the pressure coefficient C_p ($\mathbb{R}^{1500}$) in the low-dimensional ($\mathbb{R}^3$) embedding computed with LLE for the Multi-Fidelity database. The directions for the growing M (⟶), P (- ->) and XX (⋯>) are highlighted. The color scale represents P, the symbol style represents M and the symbol size represents the XX.

2.3.2 Backmapping

LLE method lacks an inherent decoding mechanism, preventing the direct reconstruction of high-dimensional data from its latent space representation. To address this, a backmapping process [15] is required to project points from the low-dimensional embedding space back to the high-dimensional space.

For an arbitrary low-dimensional point $\mathbf{y}$, the k nearest neighbors in the embedding space, denoted as $\mathcal{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_k]$, correspond isometrically to the nearest neighbors $\mathcal{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_k]$ in the high-dimensional space. To reconstruct the high-dimensional point $\mathbf{x}$ we use a first-order Taylor expansion starting from the nearest neighbors to be mapped back to the original space:

$$\mathbf{x} \approx \mathbf{x}_1 + (\mathbf{y} - \mathbf{y}_1)\nabla f(\mathbf{y}_1)^\top,$$

where $\nabla f(\mathbf{y}_1)$ represents the gradient tensor. This tensor is estimated by projecting the local differences observed in the high-dimensional space onto the local differences in the low-dimensional space using an orthogonal projection matrix.

The performance of this method is affected by the number of k neighbors. To maintain consistency, we select the same number of neighbors for the backmapping process as used in the LLE, ensuring that the local structure of the high-dimensional data is accurately captured.

2.4 Deep Neural Network Regressor

The ultimate goal of our model is to develop a surrogate model that accurately predicts the C_p distribution for airfoils not included in the training set. To accomplish this, a DNN is trained to take the three geometric variables of the NACA 4-digit airfoil as input and predict their corresponding components in the LLE embedding.

The chosen architecture is an MLP consisting of multiple layers of linear neurons with ReLU activation function. This regressor model is selected for its proven efficacy in capturing complex non-linear relationships, as evidenced by its extensive use in similar studies [4, 5]. The input layer of the DNN comprises 3 neurons, each corresponding to a design feature of the NACA 4-digit airfoil (M, P, XX). The output layer is designed to predict the components of the latent space obtained from LLE, thus having 3 neurons to match the embedding dimensions (γ_1, γ_2, γ_3). The network architecture includes 3 hidden layers that follow an increasing and then decreasing pattern, specifically configured as [3, 32, 256, 32, 3]. The Adam optimizer is used for training, with the mean squared error (MSE) loss function employed for optimization. To mitigate overfitting, early stopping is applied. The model converges in less than 500 epochs, achieving an approximate MSE of 8×10^{-7}.

3 RESULTS

This section evaluates the effectiveness of the Multi-Fidelity framework in forecasting the aerodynamic performance of NACA 4-digit airfoils. The framework was trained using 80% of the datasets, with the remaining 20% designated for testing. This approach permits a thorough assessment of the framework's capacity to generalize across diverse airfoil geometries, thereby ensuring the model's predictive performance is both robust and reliable.

The performance analysis is comprised of two principal aspects. Firstly, the global errors are quantified for each stage of the pipeline using the testing dataset. Table 1 presents the root mean square error (RMSE) and R^2 values for each regression method, ultimately providing an overall error comparison between the RANS simulation (ground truth) and the framework's C_p predictions. Secondly, a local error analysis is conducted, examining the predictions of C_p for four distinct airfoils, as illustrated in Figure 5. This detailed analysis identifies regions with significant deviations and investigates potential causes for these errors, providing insight into the predictive accuracy of the framework.

3.1 Global Analysis

The performance metrics presented in Table 1 illustrate the efficacy of each component within the Multi-Fidelity framework. The R^2 values are consistently close to 1 across all stages, indicating that each model component captures nearly all the variability in the test data, underscoring the robustness of the overall framework in predicting the C_p distribution.

Table 1: Performance metrics for the stages of the Multi-Fidelity framework. These include RMSE and R^2 score, which have been computed using the testing dataset. The MFGPR column shows the metrics between the generated Multi-Fidelity database and the RANS data, the MLP + Backmapping column shows the metrics between the regressor output and the Multi-Fidelity database, and the Full Model column shows the metrics between the pipeline output and the RANS data.

	MFGPR	**MLP + Backmapping**	**Full Model**
RMSE	0.0315	0.0222	0.0375
$\mathbf{R^2}$	0.999	0.999	0.998

Upon examination of the RMSE values across different stages, it becomes evident that the MFGPR introduces the highest error, which is only 16% lower than the error of the entire pipeline. This outcome is expected given the complexity of the MFGPR task, which involves increasing the dimensionality from the 24 virtual pressure taps to the 1500 C_p values from the RANS simulations, while attempting to maintain fidelity. In contrast, the MLP+Backmapping stage, which performs regression from the NACA 4-digit airfoil parameters to the LLE embedding variables, exhibits a lower error. This is due to the relatively simpler nature of this task, where the MLP leverages the strong correlations between the airfoil parameters and the LLE variables, resulting in precise and efficient predictions. Finally, given that the magnitude of the C_p values is of the order of unity, an RMSE of 0.0375 for the full model indicates a high level of accuracy. Therefore, this indicates that the model is well suited for its intended application, providing a reliable and effective solution for aerodynamic performance analysis.

3.2 Local Analysis

Figure 5 highlights local errors for selected cases of study, to illustrate the performance of the methodology. As suggested by the global error metrics, the predictions of the complete framework are in proximity to the MFGPR solution. The main local error introduced by the MLP+backmapping stage is a slight bias in the C_p distribution, mainly due to the backmapping process. The primary source of error, as detailed in Table 1, is the Multi-Fidelity data generation. Three main local errors are observed with the MFGPR. Firstly, there are small oscillations starting at $x/c = 0.7$, which are typical of Gaussian process regression. In addition, the suction peak pressure for the NACA 4824 and NACA 5415 is underestimated, suggesting the need for more pressure taps in this region to better capture the pressure distribution. Finally, for NACA 4824, there is a noticeable trend shift in C_p around $x/c = 0.8$. This occurs because the linear nature of the Multi-Fidelity model means that any step or discontinuity present in the Low-Fidelity data will be directly reflected in the regression output, as the model does not account for non-linear relationships between the Low-Fidelity and High-Fidelity data. Overall, despite some discrepancies, the model is robust and provides a reliable and effective tool for aerodynamic

prediction.

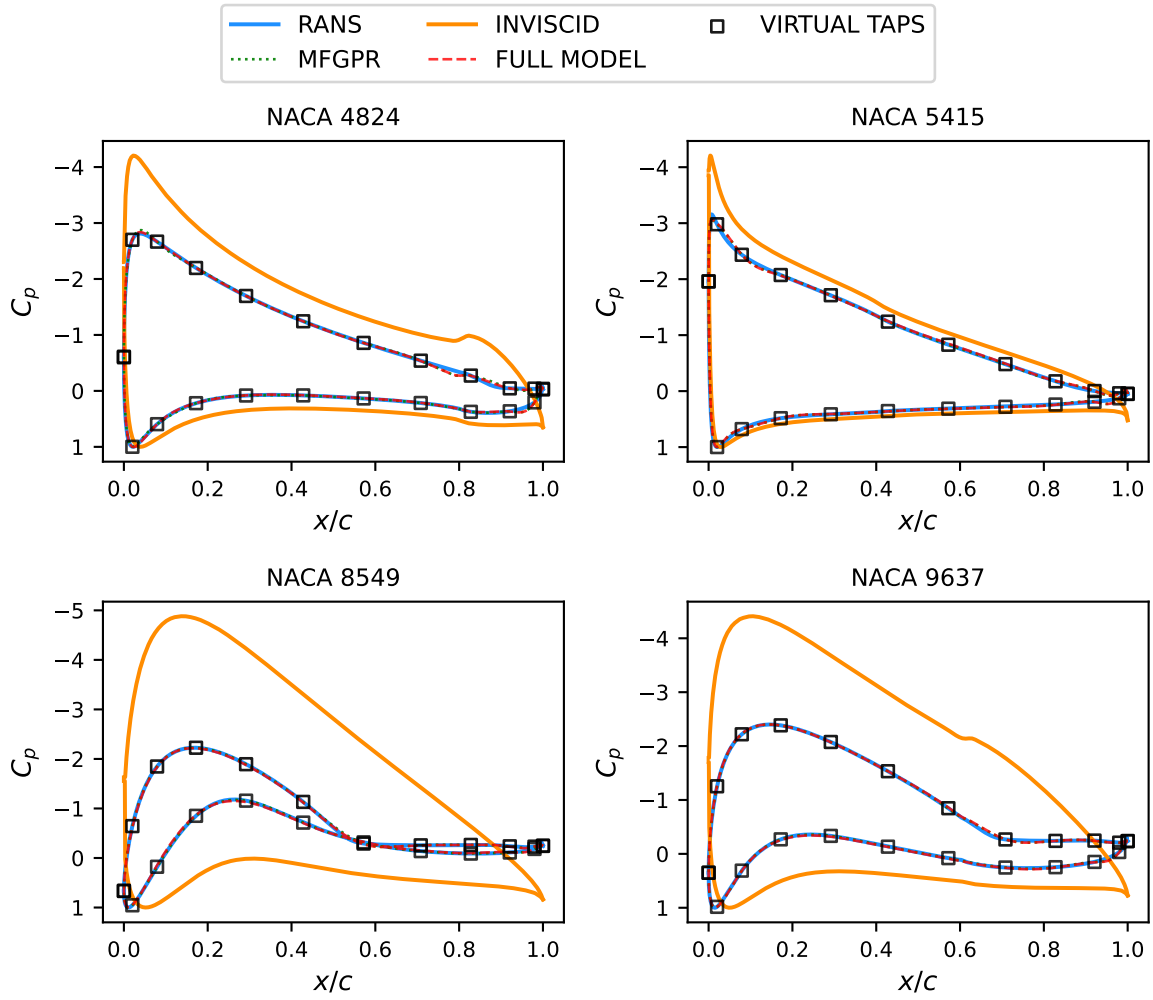

Figure 5: Distribution of C_p along the chord length (x/c) for four test airfoils. The **RANS** (—) data represent values obtained from RANS simulations. The **INVISCID** (—) data are derived from the inviscid panel method. **VIRTUAL TAPS** (□) denote the locations of virtual pressure taps. The **MFGPR** (···) line corresponds to regression made using the MFGPR method. Finally, the **FULL MODEL** (- -) line represents values obtained from the complete pipeline.

4 CONCLUSIONS

This study presents a Multi-Fidelity framework to predict the pressure coefficient distribution over NACA 4-digit airfoils. The framework consists of two main stages: data generation and regression.

During the data generation stage, the framework integrates data from simulations of varying fidelities. These include a dense, Low-Fidelity dataset from inviscid panel method simulations and a sparse, High-Fidelity dataset from RANS simulations, from which a limited number of

virtual pressure taps strategically placed on the airfoil surface have been taken to emulate real-world sensor measurements. This Multi-Fidelity data is combined using a Multi-Fidelity Gaussian Process Regression model, which effectively leverages the strengths of both datasets to create a comprehensive and reliable Multi-Fidelity dataset.

In the regression stage, the enhanced Multi-Fidelity dataset is processed through Locally Linear Embedding for dimensionality reduction. LLE compresses the dataset from 1500 to 3 dimensions, simplifying the data and improving model interpretability by revealing parameter correlations and potential outliers. This dimensionality reduction facilitates a more efficient training of the Multilayer Perceptron model, decreasing computational complexity. The resulting model achieves a R^2 value of 0.998, which shows excellent agreement with the actual data. Localized errors are within the expected ranges for contemporary aerodynamic regression models.

This study highlights the potential of combining Low- and High-Fidelity data to provide a robust and efficient framework for complex regressions, suggesting a promising direction for future research in data-driven methods for aerodynamics.

ACKNOWLEDGEMENTS

This work has been supported by projects TIFON (ref. PLEC2023-010251/ MCIN/ AEI/ 10.13039/ 501100011033, funded by the Spanish State Research Agency) and CETACEO (PTAG-20231008, funded by AIRBUS D&S through the Aeronautical Technology Programme).

REFERENCES

[1] Mehdi Anhichem, Sebastian Timme, Jony Castagna, Andrew Peace, and Moira Maina. Multifidelity data fusion applied to aircraft wing pressure distribution. In *AIAA AVIATION 2022 Forum*, page 3526, 2022.

[2] Rubén Conde Arenzana, Andrés F López-Lopera, Sylvain Mouton, Nathalie Bartoli, and Thierry Lefebvre. Multi-fidelity gaussian process model for cfd and wind tunnel data fusion. In *AeroBest 2021*, 2021.

[3] Tan Bui-Thanh, Murali Damodaran, and Karen Willcox. Aerodynamic data reconstruction and inverse design using proper orthogonal decomposition. *AIAA journal*, 42(8):1505–1516, 2004.

[4] R. Castellanos, J. Nieto-Centenero, A. Gorgues, S. Discetti, A. Ianiro, and E. Andrés. Towards aerodynamic shape optimisation by manifold learning and neural networks. In *EUROGEN 2023*. ISAAR-NTUA, 2023. doi:10.7712/140123.10190.18887.

[5] Rodrigo Castellanos, Jaime Bowen Varela, Alejandro Gorgues, and Esther Andrés. An assessment of reduced-order and machine learning models for steady transonic flow prediction on wings. *ICAS 2022*, 2022.

[6] Kenneth Decker, Henry D Schwartz, and Dimitri Mavris. Dimensionality reduction techniques applied to the design of hypersonic aerial systems. In *AIAA Aviation 2020 Forum*, page 3003, 2020.

[7] Mark Drela. Xfoil: An analysis and design system for low reynolds number airfoils. In *Low Reynolds Number Aerodynamics: Proceedings of the Conference Notre Dame, Indiana, USA, 5–7 June 1989*, pages 1–12. Springer, 1989.

[8] Robert B Gramacy. *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. Chapman and Hall/CRC, 2020.

[9] Zhong-Hua Han and Stefan Görtz. Hierarchical kriging model for variable-fidelity surrogate modeling. *AIAA journal*, 50(9):1885–1896, 2012.

[10] Marc C Kennedy and Anthony O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

[11] Yuichi Kuya, Kenji Takeda, Xin Zhang, and Alexander IJ Forrester. Multifidelity surrogate modeling of experimental and computational aerodynamic data sets. *AIAA journal*, 49(2):289–298, 2011.

[12] Loic Le Gratiet and Josselin Garnier. Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *International Journal for Uncertainty Quantification*, 4(5), 2014.

[13] Michael Mifsud, Alexander Vendl, Lars-Uwe Hansen, and Stefan Görtz. Fusing wind-tunnel measurements and cfd data using constrained gappy proper orthogonal decomposition. *Aerospace Science and Technology*, 86:312–326, 2019.

[14] J. Nieto-Centenero, R. Castellanos, A. Gorgues, and E. Andrés. Fusing aerodynamic data using multi-fidelity gaussian process regression. In *EUROGEN 2023*. ISAAR-NTUA, 2023. doi:10.7712/140123.10191.18890.

[15] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[16] A. Schillaci, M. Quadrio, and G. Boracchi. A database of cfd-computed flow fields around airfoils for machine-learning applications [data set], 2021. URL: https://doi.org/10.5281/zenodo.4106752.

[17] A. Schillaci, M. Quadrio, and G. Boracchi. A database of cfd-computed flow fields around airfoils for machine-learning applications (part 2) [data set], 2021. URL: https://doi.org/10.5281/zenodo.4638071.

[18] A. Schillaci, M. Quadrio, C. Pipolo, M. Restelli, and G. Boracchi. Inferring functional properties from fluid dynamics features. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*, January 2020.

[19] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi:10.1126/science.290.5500.2319.

[20] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.