; ; ; ;

FEDERAL UNIVERSITY OF ABC
GRADUATE COURSE IN ENERGY

Allan Moreira de Carvalho

REDUCED ORDER MODELS FOR PARAMETRIC DOMAINS: A DATA-DRIVEN METHODOLOGY FOR THE ACCELERATION OF ENGINEERING SIMULATIONS

Santo André, SP – Brazil
July 2025

Allan Moreira de Carvalho

REDUCED ORDER MODELS FOR PARAMETRIC DOMAINS: A DATA-DRIVEN METHODOLOGY FOR THE ACCELERATION OF ENGINEERING SIMULATIONS

Qualifying Exam presented to the Graduate course of Federal University of ABC, as a partial requirement for obtaining a Doctor's degree in Energy.

Advisor: Prof. Dr. Daniel Jonas Dezan
Coadvisor:Prof. Dr. Wallace Gusmão Ferreira

Santo André, SP – Brazil
July 2025

Allan Moreira de Carvalho,

Reduced Order Models for Parametric Domains: A Data-Driven Methodology for the Acceleration of Engineering Simulations/ Allan Moreira de Carvalho. – 2025.

53 p.: il.

Advisor: Prof. Dr. Daniel Jonas Dezan

Coadvisor: Prof. Dr. Wallace Gusmão Ferreira

Qualifying Exam (Doctor) – Federal University of ABC, Graduate Course in Energy, Santo André, 2025.

1. Reduced Order Model. 2. Machine Learning. 3. Proper Orthogonal Decomposition (POD). 4. Parametric Modeling. 5. Geometric Morphing. I. Jonas Dezan, Daniel II. Gusmão Ferreira, Wallace III. Graduate Course in Energy, 2025. IV. Title

This exemplar was revised and altered from the original version according to the observations done by the defense board at the defense day, under responsability of author and advisor's consent.

Santo André, _____ _____ , 20 _____

Author's signature: _____

Advisor's signature: _____

UFABC

GRADUATE COURSE IN ENERGIA

SIGN SHEET

Signatures of examining board members, which examinated and approved the defense of Doctoral Qualifying Exam of candidate Allan Moreira de Carvalho in July 10th, 2025:

Prof. Dr. Marcia Maria Penteado Marchesini - External Examiner

Prof. Dr. Wilson Carlos da Silva Júnior - External Examiner

Prof. Dr. Franciane Freitas Silveira - Internal Substitute Examiner

Prof. Dr. Rômulo Gonçalves Lins - Chair

I dedicate it...

Acknowledgement

I thank to...

Epigraph

"Learn from the mistakes of those who followed your advice"
(Unknown author)

Abstract

This dissertation introduces a comprehensive and adaptable framework for the construction of data-driven, parametric reduced-order models (ROMs) to accelerate the analysis of complex physical systems. High-fidelity numerical simulations, while accurate, impose a prohibitive computational cost in multi-evaluation scenarios such as design optimization, uncertainty quantification, and real-time control. To overcome this bottleneck, this work develops a unified pipeline that combines techniques in data preprocessing, dimensionality reduction, and machine learning. The core methodology employs Proper Orthogonal Decomposition (POD) to extract dominant, energy-optimal spatial modes from high-dimensional simulation snapshot data, which are then mapped from the design parameters using regression models like Gaussian Process Regression (GPR) or Artificial Neural Networks (ANNs). A critical innovation, developed to address challenges with complex geometries, is a sophisticated technique for handling geometric variations, which ensures topological consistency across differing parametric domains.

The framework's efficacy was first established in a 2D context by conducting a comparative analysis of POD-GPR and POD-ANN models for the reconstruction of supersonic nozzle flows, which are characterized by strong shock-wave/boundary-layer interactions. This initial study introduced a novel hybrid loss function for ANN training and employed SHapley Additive exPlanations (SHAP) to enhance model interpretability. However, the subsequent challenge of applying the framework to a complex 3D case—the parametric reconstruction of pressure and temperature fields on the blade surfaces of the NASA Rotor 37—revealed a fundamental limitation: meshes of varying sizes and inconsistent topologies prevented the direct application of POD. This problem necessitated the development of the aforementioned geometric mesh morphing technique, which proved crucial for enabling the methodology's extension to variable 3D shapes.

Results from both studies demonstrate exceptional predictive accuracy, with coefficients of determination (R^2) consistently exceeding 0.95, and computational speed-ups of several orders of magnitude (50x to over 10,000x) compared to full-fidelity simulations. This work not only provides a robust and validated methodology for accelerating simulation-based design and analysis cycles but also offers practical guidelines on model selection and training, establishing a foundation for the next generation of intelligent simulation tools.

Keywords:

Reduced Order Model; Machine Learning; Proper Orthogonal Decomposition (POD); Parametric Modeling; Geometric Morphing

Resumo

Esta dissertação introduz um framework abrangente e adaptável para a construção de modelos de ordem reduzida (ROMs) paramétricos e orientados por dados, visando acelerar a análise de sistemas físicos complexos. Simulações numéricas de alta fidelidade, embora precisas, impõem um custo computacional proibitivo em cenários de que necessitam de múltiplas avaliações, como otimização de forma, quantificação de incertezas e controle em tempo real. Para superar esse gargalo, este trabalho desenvolve um pipeline unificado que combina técnicas de pré-processamento de dados, redução de dimensionalidade e aprendizagem de máquina. A metodologia central emprega a Decomposição Ortogonal Própria (POD) para extrair modos espaciais dominantes, que são então mapeados a partir dos parâmetros de projeto usando modelos de regressão como a Regressão por Processo Gaussiano (GPR) ou Redes Neurais Artificiais (ANNs). Uma inovação crítica, desenvolvida para lidar com geometrias complexas, é uma técnica sofisticada para o tratamento de variações geométricas, que garante a consistência topológica entre diferentes domínios paramétricos.

A eficácia do framework foi primeiramente estabelecida em um contexto 2D, através de uma análise comparativa de modelos POD-GPR e POD-ANN para a reconstrução de escoamentos em bocais supersônicos. Este estudo inicial introduziu uma função de perda híbrida para o treinamento da ANN e empregou SHapley Additive exPlanations (SHAP) para aumentar a interpretabilidade do modelo. No entanto, o desafio subsequente de aplicar o framework a um caso 3D—a reconstrução paramétrica dos campos de pressão e temperatura nas pás do NASA Rotor 37—revelou uma limitação fundamental: malhas com tamanhos distintos e topologias inconsistentes impediram a aplicação direta do POD. Este problema tornou necessário o desenvolvimento da técnica de tratamento geométrico usando deformação de malha, que se mostrou crucial para permitir a extensão da metodologia a formas 3D variáveis.

Os resultados de ambos os estudos demonstram excepcional precisão preditiva, com coeficientes de determinação (R^2) consistentemente acima de 0.95, e uma aceleração computacional de várias ordens de magnitude (de 50x a mais de 10.000x) em comparação com as simulações de alta fidelidade. Este trabalho não apenas fornece uma metodologia robusta e validada para acelerar os ciclos de projeto e análise baseados em simulação, mas também oferece diretrizes práticas quanto a seleção e o treinamento de modelos, estabelecendo a base para a próxima geração de ferramentas de simulação inteligentes.

Palavras-chave: Modelo de Ordem Reduzida; Aprendizado de Máquina; Decomposição Ortogonal Própria (POD); Modelagem Paramétrica; Mapeamento de Malha

List of Figures

List of Tables

List of Algorithms

Contents

1 Introduction

1.1 The Grand Challenge: The Multi-Query Bottleneck in Computational Aerodynamics

In the modern era of aerospace and turbomachinery engineering, Computational Fluid Dynamics (CFD) has become an indispensable tool for analysis and design. High-fidelity numerical models, such as those based on the Reynolds-Averaged Navier-Stokes (RANS) equations, Large Eddy Simulation (LES), or even Direct Numerical Simulation (DNS), offer remarkable precision in predicting the behavior of fluid flows. This capability allows engineers to investigate complex physical phenomena, from the turbulent wake behind an aircraft to the intricate shock structures within a supersonic engine, with a level of detail that is often hard to achieve through physical experimentation alone.

However, the high fidelity of these simulations comes at a price: computational expense. Solving the governing equations of fluid motion across complex geometries discretized into millions or even billions of grid cells requires immense computational resources and can take hours, days, or even weeks on high-performance computing (HPC) clusters. While the cost of a single simulation may be justifiable for final design verification, it becomes prohibitive in the context of the modern engineering design cycle. The core challenge is not merely that a single CFD simulation is slow, but that contemporary design and analysis workflows are inherently ”multi-evaluation” in nature.

Tasks such as design space exploration, aerodynamic shape optimization, uncertainty quantification (UQ), and sensitivity analysis require the evaluation of hundreads, if not thousands, of design variations. For example, an optimization algorithm may need to iteratively adjust dozens of geometric parameters to maximize lift or minimize drag, with each iteration demanding a new CFD simulation. Similarly, a robust UQ analysis might involve propagating uncertainties from manufacturing tolerances or operational conditions through the model, a task that often relies on Monte Carlo methods requiring a vast number of simulations. When each model evaluation involves a computationally expensive CFD run, these essential engineering tasks become computationally intractable. This ”multi-evaluation bottleneck” represents a fundamental barrier to innovation, slowing down the design cycle and limiting the ability of engineers to explore novel concepts or quantify risks effectively.

To overcome this challenge, a paradigm shift is required, moving away from the direct,

repeated use of high-fidelity models. The strategic solution lies in the development of surrogate models, also known as metamodels or digital twins. A surrogate model is a computationally inexpensive, data-driven approximation of the complex, high-fidelity model. By learning the input-output relationship from a limited set of pre-computed high-fidelity simulations, a surrogate can provide near-instantaneous predictions for new design points, effectively replacing the expensive CFD solver within the multi-evaluation loop. This approach transforms an intractable problem into a feasible one, enabling fast design exploration and robust analysis without sacrificing the essential physical insights provided by the original high-fidelity data.

1.2 A Data-Driven Paradigm: Reduced-Order Models (ROMs)

Among the various classes of surrogate models, data-driven Reduced-Order Models (ROMs) have emerged as a particularly powerful paradigm for high-dimensional physical systems like fluid flows. The central philosophy of ROMs is an ”offline-online” computational strategy. In the offline, or ”training,” stage, a set of computationally expensive, high-fidelity simulations is performed to generate a database of ”snapshots” of the system's behavior across a range of parameters. This database is then used to train the ROM. Once trained, the ROM can be deployed in the online, or ”prediction,” stage, where it provides extremely fast approximations for new, unseen parameter inputs. This decouples the high computational cost of data generation from the rapid-query demands of the application.

The construction of a machine learning-based ROM (ML-ROM) for a parametric system typically follows a structured pipeline, which forms the backbone of this dissertation. This pipeline can be conceptualized in four primary stages:

- Data Generation: A Design of Experiments (DoE) is created to strategically sample the parametric design space. A high-fidelity CFD solver is then run for each sample point to generate a database of high-dimensional solution snapshots.

- Dimensionality Reduction: The immense dimensionality of the snapshot data (often millions of degrees of freedom per snapshot) is reduced to a very low-dimensional latent space. This is typically achieved using techniques like Proper Orthogonal Decomposition (POD), which extracts a small set of dominant, energy-optimal basis functions, or ”modes,” that capture the essential dynamics of the flow.

- Latent-Space Regression: A machine learning model (the surrogate regressor) is trained to learn the mapping between the low-dimensional input design parameters (e.g., blade angle, Mach number) and the low-dimensional latent-space representation (e.g., the POD mode coefficients) of the flow field.

- Field Reconstruction: During the online phase, the trained regressor predicts the latent-space coefficients for a new set of design parameters. These coefficients are then used to reconstruct the full, high-dimensional flow field through a linear combination of the pre-computed basis functions.

This structured approach allows for the systematic deconstruction of a complex modeling problem into a series of more manageable tasks, each addressable with specialized mathematical and computational tools.

1.3 Thesis Objectives and Contributions

The primary objective of this dissertation is to develop, validate, and analyze a unified, flexible, and robust framework for creating parametric reduced-order models for complex aerodynamic flows. This work aims to move beyond ad-hoc solutions for specific problems and establish a comprehensive methodology that can be adapted to a wide range of challenges in computational aerodynamics, from internal supersonic flows to external transonic turbomachinery.

The key contributions of this dissertation, which collectively advance the state-of-the-art in data-driven aerodynamic modeling, are as follows:

- A Unified Methodological Framework: The development of a comprehensive, end-to-end computational pipeline that synergistically integrates parametric geometry definition, high-fidelity data generation, advanced mesh processing, Proper Orthogonal Decomposition, and a selection of machine learning regressors (Gaussian Process Regression and Artificial Neural Networks). This unified structure provides a coherent and reproducible approach to ROM construction.

- Enabling Technology for 3D Parametric ROMs: The introduction and validation of a sophisticated mesh morphing technique, based on harmonic mapping, as a critical enabling technology. This method resolves the fundamental challenge of topological inconsistency in parametric studies, thereby allowing, for the first time in this context, the direct application of POD to complex 3D geometries with varying shapes, as demonstrated in the NASA Rotor 37 case study.

- Systematic Comparative Analysis of Surrogate Models: An in-depth, empirical comparison of Gaussian Process Regression (GPR) and Artificial Neural Network (ANN) regressors for latent-space mapping. This analysis, conducted through rigorous cross-validation and noise robustness studies, provides practical, evidence-based guidelines for model selection based on factors such as dataset size, data quality, and the underlying physics of the problem.

- Advancements in Model Training and Interpretability: The introduction of two novel techniques to enhance the fidelity and trustworthiness of ANN-based ROMs. First, a hybrid loss function is proposed that combines errors in both the latent and reconstructed physical spaces, improving the physical accuracy of the final predictions. Second, SHapley Additive exPlanations (SHAP) are employed to provide quantitative interpretability for the "black-box" models, linking their internal decision-making processes to fundamental physical principles.

- Validation Across Diverse Flow Regimes: The rigorous validation and demonstration of the framework's versatility across two distinct and challenging aerodynamic case studies: the 3D external transonic flow over the NASA Rotor 37 compressor blade, and the 2D internal supersonic flow within a de Laval nozzle, characterized by strong shock-wave/boundary-layer interactions.

1.4 Dissertation Outline

This dissertation is structured to guide the reader from foundational concepts to advanced applications and future possibilities.

- Chapter 2 provides a comprehensive review of the theoretical foundations of reduced-order modeling and machine learning as applied to fluid dynamics, establishing the context and key concepts for the work.

- Chapter 3 details the unified methodological framework developed in this thesis, presenting each stage of the computational pipeline, from data generation and mesh morphing to the hybrid POD-ML regression and advanced validation strategies.

- Chapter 4 presents the first major case study, applying the framework to the parametric reconstruction of 3D surface fields on the NASA Rotor 37, with a focus on the critical role of the mesh morphing technique.

- Chapter 5 presents the second case study, a comparative analysis of GPR and ANN surrogates for reconstructing 2D supersonic nozzle flows, delving into advanced topics of hyperparameter tuning, robustness, and model interpretability.

- Chapter 6 synthesizes the findings from both case studies, summarizes the key contributions of the dissertation, discusses its limitations, and proposes promising directions for future research.

2 Foundations of Reduced-Order Modeling and Machine Learning in Fluid Dynamics

2.1 The Imperative for Model Reduction in Computational Aerodynamics

The governing equations of fluid motion, the Navier-Stokes equations, are a set of coupled, nonlinear partial differential equations. Their numerical solution in a discretized domain representing a complex engineering geometry results in a system with an enormous number of degrees of freedom, often on the order of millions or billions. When considering parametric studies, where design variables such as geometry or boundary conditions are varied, the dimensionality of the problem space explodes further. This leads to the well-known "curse of dimensionality," where the number of samples required to adequately explore a design space grows exponentially with the number of parameters. A brute-force approach, involving the gridding of this high-dimensional parameter space and running a full-order simulation at each point, is computationally infeasible.

This reality underscores the fundamental imperative for model reduction. The challenge is not merely one of computational cost, but one of inherent complexity and high dimensionality of the solution manifold—the set of all possible solutions as the input parameters vary. The central hypothesis of reduced-order modeling is that despite the high dimensionality of the discretized system, the actual dynamics of many physical systems evolve on or near a much lower-dimensional, intrinsic manifold embedded within the high-dimensional state space. A successful ROM, therefore, is one that can discover and exploit this low-dimensional structure. The task of a ROM is not just to compress data, but to identify the underlying coherent patterns and structures that govern the system's behavior, thereby creating a compact, physically meaningful, and computationally tractable representation of the original complex system.

2.2 Dimensionality Reduction for Fluid Flows: Proper Orthogonal Decomposition (POD)

Proper Orthogonal Decomposition (POD) is arguably the most established and widely used technique for dimensionality reduction in fluid dynamics. Also known in other fields as

Principal Component Analysis (PCA) or the Karhunen-Loève expansion, POD provides a systematic method for finding the most efficient linear basis to represent a high-dimensional dataset.

2.2.1 Theoretical Formulation

The application of POD in CFD typically begins with the "method of snapshots," a technique pioneered for fluid dynamics applications. A set of N_s high-fidelity simulation results, or "snapshots," are collected. Each snapshot, representing a flow field at a specific parameter value, is reshaped into a column vector $u_j \in \mathbb{R}^{N_g}$, where N_g is the number of grid points in the mesh. These snapshot vectors are then assembled into a snapshot matrix $S = [u_1, u_2, ..., u_{N_s}] \in \mathbb{R}^{N_g \times N_s}$.

POD seeks a set of orthonormal basis vectors, or "modes," $\{\phi_k\}_{k=1}^{r}$, that are optimal in the sense that they maximize the projection of the snapshot data onto them. Mathematically, this is equivalent to finding the basis that minimizes the mean squared error of the projection for any given rank of approximation. The solution to this optimization problem is found by solving the eigenvalue problem of the data covariance matrix, $C = SS^T$. However, since N_g is typically much larger than N_s, it is computationally more efficient to solve the eigenvalue problem for the smaller matrix $K = S^T S \in \mathbb{R}^{N_s \times N_s}$. The POD modes ϕ_k are then recovered from the eigenvectors of K. In practice, this entire procedure is most robustly and efficiently performed using the Singular Value Decomposition (SVD) of the snapshot matrix, $S = U\Sigma V^T$, where the columns of the matrix U are the POD modes ϕ_k.

2.2.2 The "Energy-Optimal" Basis

A key property of POD is that the modes form an "energy-optimal" basis. The singular values σ_k (the diagonal entries of Σ) are related to the eigenvalues of the covariance matrix and represent the "energy" (or variance) captured by each corresponding mode ϕ_k. The modes are ordered hierarchically, such that the first mode ϕ_1 captures the most energy, ϕ_2 captures the most of the remaining energy, and so on. This hierarchy allows for a highly efficient low-rank approximation of the original data. Any snapshot u_j can be approximated as a linear combination of the first M modes:

$$u_j \approx \sum_{k=1}^{M} a_{jk}\phi_k$$

where $M \ll N_s$, and the coefficients $a_{jk} = u_j^T \phi_k$ are the projections of the snapshot onto the modes. By retaining only a small number of modes that capture a vast majority of the system's energy (e.g., 99.9%), the dimensionality of the problem is dramatically reduced from N_g to M.

2.2.3 Implementation and Preprocessing

Before applying POD, raw snapshot data must be preprocessed to ensure numerical stability and physical relevance. Two steps are crucial. First, the data is typically mean-centered by subtracting the ensemble-averaged field, $\bar{u} = \frac{1}{N_s}\sum_{j=1}^{N_s} u_j$, from each snapshot. POD is then performed on the fluctuation fields, $u'_j = u_j - \bar{u}$. This separates the mean behavior from the dynamic variations, which are often the primary interest of the model. Second, data scaling, for instance using a MinMaxScaler to bring all values into a uniform range like $[0, 1]$, is often applied. This prevents fields with large physical magnitudes (like pressure) from dominating the variance calculation over fields with smaller magnitudes (like temperature or Mach number), ensuring an equitable contribution from all physical quantities to the final modes.

2.2.4 Limitations and Future Directions

The power of POD lies in its simplicity and optimality for linear systems. However, its linearity is also its most fundamental weakness. POD finds the optimal linear subspace on which to project the data. Many important fluid dynamics phenomena, particularly those dominated by advection or featuring moving discontinuities like shock waves, are better described as evolving on a nonlinear manifold. Projecting a curved manifold onto a flat, linear subspace is an inherently inefficient representation. It often requires a large number of POD modes to accurately capture the nonlinear dynamics, diminishing the benefits of the model reduction.

This limitation has been a major driver of recent research in the field. The recognition that POD is suboptimal for strongly nonlinear systems has motivated the exploration of nonlinear dimensionality reduction techniques. Chief among these are methods based on deep learning, such as Convolutional Autoencoders (CAEs). A CAE uses a neural network (the encoder) to learn a nonlinear mapping from the high-dimensional input space to a low-dimensional latent space, and another network (the decoder) to map back. By training the network to minimize reconstruction error, the CAE can learn a compact, nonlinear representation that is potentially far more efficient than the linear subspace found by POD. The development and application of such nonlinear ROMs represent the next frontier in data-driven fluid dynamics, promising even greater efficiency and accuracy for the most challenging flow problems. This thesis, while focused on the robust application of POD, acknowledges this trajectory and provides a foundational framework upon which such future advancements can be built.

2.3 Surrogate Modeling for Latent Space Dynamics

Once dimensionality reduction has been performed, the original problem of mapping high-dimensional inputs to high-dimensional outputs is transformed into a more tractable one: mapping low-dimensional inputs (the design parameters) to a low-dimensional latent space (the POD coefficients). This is a classical regression problem, for which a variety of machine learning techniques can be employed. This work focuses on two of the most powerful and widely used methods: Gaussian Process Regression and Artificial Neural Networks.

2.3.1 Gaussian Process Regression (GPR)

Gaussian Process Regression is a non-parametric, Bayesian regression method that has gained significant popularity in engineering and machine learning. Unlike parametric models that fit a single function to the data, GPR defines a probability distribution over a space of functions. A Gaussian Process is a collection of random variables, any finite number of which have a joint Gaussian distribution. It is fully specified by a mean function and a covariance function, or kernel.

The kernel is the heart of a GPR model. It encodes the assumptions about the function being modeled, such as its smoothness and correlation structure. A common choice, used in this work, is the **Radial Basis Function (RBF) kernel**, also known as the squared exponential kernel. This kernel assumes that input points that are ”close” in the parameter space will have similar output values, with the notion of ”closeness” controlled by a length-scale hyperparameter.

The key advantages of GPR are twofold. First, as a Bayesian method, it provides not only a point prediction (the posterior mean) but also a measure of predictive uncertainty (the posterior variance). This is invaluable in engineering design, as it allows for principled uncertainty quantification and risk assessment. Second, the Bayesian framework provides a natural defense against overfitting, making GPR particularly robust for problems with small or sparse training datasets. However, GPR is not without its drawbacks. The primary limitation is its computational complexity, which scales as $O(N^3)$ with the number of training points N, due to the need to invert the covariance matrix. This makes standard GPR computationally challenging for very large datasets.

2.3.2 Artificial Neural Networks (ANNs)

Artificial Neural Networks, and specifically the Multi-Layer Perceptron (MLP), are powerful function approximators inspired by the structure of the biological brain. The universal approximation theorem states that a feedforward network with a single hidden layer containing a finite number of neurons can approximate any continuous function to

an arbitrary degree of accuracy, given enough neurons. This makes ANNs an extremely flexible and powerful tool for regression.

An ANN consists of interconnected layers of nodes, or "neurons." Each neuron performs a weighted sum of its inputs, adds a bias, and then passes the result through a nonlinear activation function (e.g., sigmoid, ReLU, tanh). By stacking multiple layers, ANNs can learn hierarchical representations of data and model highly complex, nonlinear relationships.

The primary advantages of ANNs are their scalability to very large datasets and their unparalleled flexibility in modeling complex functions. Unlike GPR, their training time can scale more favorably with the number of samples, especially when using modern GPU hardware and stochastic gradient descent-based optimizers. However, this flexibility comes at a cost. ANNs are often considered "black-box" models, making their internal reasoning difficult to interpret. They are also prone to overfitting if not properly regularized (e.g., using techniques like dropout or weight decay), and they typically have a large number of hyperparameters (e.g., number of layers, number of neurons, learning rate) that require careful and often extensive tuning to achieve optimal performance.

The choice between GPR and ANNs represents a fundamental trade-off in surrogate modeling. GPR offers probabilistic rigor, built-in uncertainty quantification, and robustness on small datasets, making it an excellent choice for well-posed problems where quantifying uncertainty is paramount. ANN, on the other hand, offers scalable flexibility and the raw power to model extremely complex, high-dimensional, and noisy phenomena, provided that sufficient data and computational resources are available for training and tuning. As demonstrated in Chapter 5 of this dissertation, the empirical evidence gathered from comparative studies is crucial for navigating this trade-off. The results suggest that there is no single "best" model; rather, the optimal choice depends on the specific characteristics of the problem at hand, including the size and quality of the available data and the nature of the underlying physics.

2.4 Toward Physically Consistent and Interpretable Models

As machine learning models become more integrated into safety-critical engineering workflows, two challenges have come to the forefront of the research community: physical consistency and interpretability. A model that is purely data-driven, with no knowledge of the underlying physics, may produce predictions that are highly accurate on average but violate fundamental physical laws, making them unreliable for engineering decisions. Similarly, a "black-box" model that provides accurate predictions without any explanation of its reasoning is difficult to trust, debug, or certify for use in critical applications.

2.4.1 The Rise of Physics-Informed Machine Learning (PIML)

To address the first challenge, the field of **Physics-Informed Machine Learning (PIML)** has emerged. The core idea of PIML is to embed physical domain knowledge, often in the form of governing partial differential equations (PDEs), directly into the machine learning algorithm. A common approach is to add the residuals of the governing PDEs as a penalty term in the model's loss function. This forces the model to learn not only from the data but also to satisfy the physical constraints, leading to better data efficiency, improved generalization, and more physically plausible solutions.

The work in this dissertation aligns with the PIML philosophy through a practical, "soft" implementation. The novel hybrid loss function introduced in Chapter 5 for training ANN surrogates includes a term, $L_{\text{reconstructed}}$, which directly measures the prediction error in the reconstructed physical space. By minimizing this term, the training process implicitly pushes the model to generate solutions that are physically accurate, as any significant deviation from the true physical field would incur a large penalty. This approach bridges the gap between the abstract latent space, where the model operates, and the physical space, which is the domain of interest for the engineer.

2.5 The Interpretability Imperative

To address the second challenge of trust, methods for model interpretability have become essential. It is no longer sufficient for a model to be accurate; it must also be explainable. Techniques like SHapley Additive exPlanations (SHAP) provide a powerful framework for peering inside the black box. Based on principles from cooperative game theory, SHAP assigns a quantitative "importance value" to each input feature, representing its contribution to a specific prediction. This allows one to understand which factors the model is weighing most heavily in its decision-making process.

The application of SHAP analysis in Chapter 5 is a direct response to this interpretability imperative. By analyzing the feature importances of the trained surrogate models, it is possible to verify that their behavior aligns with physical intuition. For instance, the finding that the model identifies inlet total pressure as the most critical parameter for nozzle flow builds confidence that the model has learned a physically meaningful relationship, rather than just spurious correlations in the data. Therefore, this dissertation does not simply apply machine learning as a black-box tool; it actively engages with two of the most critical and contemporary conversations in the field of scientific machine learning: the pursuit of physical consistency and the establishment of trust through interpretability.

3 A Unified Framework for Parametric Field Reconstruction

This chapter formalizes the comprehensive, end-to-end computational framework developed in this dissertation. The framework is designed to be modular and adaptable, providing a systematic pipeline for constructing parametric reduced-order models for a wide range of aerodynamic problems. The methodology is presented as a sequence of five stages, moving from initial problem definition and data generation to the final validation of a trained surrogate model. The subsequent chapters will demonstrate the application of this unified framework to two distinct and challenging case studies.

3.1 Overview of the End-to-End Pipeline

The proposed framework integrates several advanced computational techniques into a cohesive workflow. The process begins with the parameterization of a baseline geometry and the use of a Design of Experiments (DoE) methodology to sample the design space. High-fidelity CFD simulations are performed at these sample points to generate a snapshot database. A critical and innovative step, particularly for 3D applications, is the use of a mesh morphing pipeline to harmonize the topology of the disparate snapshot meshes, making them suitable for Proper Orthogonal Decomposition (POD). POD is then used to extract a low-dimensional basis from the high-dimensional field data. Finally, a machine learning regressor—either a Gaussian Process (GPR) or an Artificial Neural Network (ANN)—is trained to map the input design parameters to the low-dimensional POD coefficients. The trained model can then be used for rapid prediction of the full aerodynamic field for any new design within the parameterized space. This entire process is designed with reproducibility and rigor in mind, incorporating advanced strategies for model tuning and validation.

3.2 Stage 1: Parametric Data Generation

The foundation of any data-driven model is the data itself. The quality, quantity, and diversity of the training data directly determine the accuracy and robustness of the final surrogate model. This first stage focuses on the systematic generation of a high-quality snapshot database.

3.2.1 Geometry Parameterization

The process begins by defining a parametric design space. This involves identifying the key variables that control the geometry of the object under study and defining their range of variation. For aerodynamic applications, these variables often relate to the shape of an airfoil or blade. Techniques such as using control points to define Bézier splines or other curves are commonly employed to smoothly and intuitively vary features like blade thickness distributions, camber lines, or angle of attack schedules along the span. In the NASA Rotor 37 case study, for example, 28 design variables were used to control the blade angle and thickness distributions at the hub and shroud sections, fully defining the parametric space for the optimization study.

3.2.2 Design of Experiments (DoE)

Once the parameter space is defined, it must be sampled efficiently to select the points at which expensive high-fidelity simulations will be run. Given the high dimensionality of many design spaces, a simple gridding approach is infeasible. Instead, advanced DoE techniques are used to ensure that the sample points provide maximum information with a minimum of computational effort. Latin Hypercube Sampling (LHS) is a particularly effective statistical method for this purpose. LHS generates a set of sample points that are well-distributed across the parameter space, avoiding clustering and ensuring that the full range of each parameter is explored. This space-filling property is crucial for training a surrogate model that can generalize well to unseen regions of the design space.

3.2.3 High-Fidelity Simulation

With the sample points defined by the DoE, the final step in data generation is to execute the full-order numerical model (e.g., a RANS CFD solver) for each point. This is the most computationally expensive part of the entire offline phase. Each simulation produces a ”snapshot” of the steady-state solution, which includes the full-field data (e.g., pressure, temperature, velocity) across the entire computational mesh. The collection of all snapshots from the DoE constitutes the raw database that will be used to train the reduced-order model.

3.3 Stage 2: Topological Harmonization via Mesh Morphing

A fundamental prerequisite for the application of Proper Orthogonal Decomposition is that all snapshots must reside on a common grid with a consistent topology. This means that the number of points and their ordering must be identical across all snapshots, such that the i-th entry in each snapshot vector corresponds to the same spatial location or

entity. In parametric CFD studies, where the geometry is altered for each design point, this condition is inherently violated, as each simulation is performed on a unique, individually generated mesh. This topological inconsistency poses a major barrier to the application of POD for parametric ROMs, especially for complex 3D shapes.

To overcome this critical challenge, this framework incorporates a sophisticated mesh morphing pipeline. This pipeline acts as a preprocessing stage that transforms a collection of topologically inconsistent surface meshes into a set of regularized meshes that share a common structure. This process, detailed extensively in the NASA Rotor 37 case study , is not merely a data-cleaning step; it is the fundamental enabling technology that makes the subsequent application of POD possible for parametric 3D geometries. The process consists of three sequential steps:

- Harmonic Mapping (3D-to-2D Projection): The first step is to create a common reference frame for all the irregular 3D surface meshes. This is achieved by mapping each 3D mesh onto a canonical 2D parametric domain, such as a unit square $\times$. This is done by computing a harmonic map, Φ, which is the solution to Laplace's equation ($\Delta\Phi = 0$) with Dirichlet boundary conditions that map the boundary of the 3D mesh to the boundary of the 2D square. The solution to this elliptic PDE results in a smooth, one-to-one "flattening" of the 3D surface onto the 2D plane, preserving local angles and minimizing distortion.

- Interpolation onto a Regular Grid: With all snapshots now represented as unstructured triangular meshes within the same 2D parametric domain, the associated scalar fields (e.g., pressure, temperature) and the original 3D coordinates (X,Y,Z) are interpolated onto a new, common, structured grid (e.g., a uniform 100 by 100 grid). This interpolation is typically performed using barycentric coordinates within the triangles of the 2D mapped mesh. This step effectively resamples all the field data from their original, inconsistent meshes onto a single, shared grid structure.

- 3D Reconstruction (Lifting): The final step is to "lift" the new structured grid, now populated with the interpolated field data and 3D coordinates, back into physical 3D space. This creates a new, regularized 3D surface mesh that is topologically identical for every snapshot in the database. The connectivity of this new mesh is implicitly defined by the structure of the 2D grid.

3.4 Stage 3 and 4: The Hybrid POD-ML Regression Pipeline

With the data properly generated and preprocessed, the core of the reduced-order modeling process can begin. This involves the coupled application of dimensionality reduction and machine learning regression to create the final surrogate model.

Snapshot Matrix Construction: The regularized, mean-centered, and scaled field data from all N_s snapshots are vectorized and assembled into the final snapshot matrix $S \in \mathbb{R}^{N_g \times N_s}$, where N_g is now the number of points in the common regularized grid.

POD Projection: Singular Value Decomposition (SVD) is applied to the snapshot matrix S to obtain the orthogonal POD modes, Φ (the left singular vectors), which form the reduced basis. The original high-dimensional snapshots are then projected onto this basis to obtain the low-dimensional modal coefficients, a. This projection is a simple matrix multiplication: $a = \Phi^T S'$, where S' is the matrix of mean-centered snapshots. The result is a matrix of coefficients $a \in \mathbb{R}^{M \times N_s}$, where M is the number of retained modes.

ML Regression: The task is now to learn the nonlinear mapping from the input design parameters, g, to the POD coefficients, a. A machine learning surrogate model, f_{ML}, is trained on the pairs of design parameters and their corresponding POD coefficients from the training set: $\{(g_j, a_j)\}_{j=1}^{N_{\text{train}}}$. This model can be a GPR, an ANN, or another suitable regression technique. The goal is to find a function f_{ML} such that $a_{\text{predicted}} \approx f_{\text{ML}}(g)$.

Field Reconstruction: In the online phase, the trained model is used for prediction. For a new, unseen set of design parameters g^*, the surrogate model is first evaluated to predict the corresponding POD coefficients: $a^* = f_{\text{ML}}(g^*)$. The full, high-dimensional fluctuation field is then reconstructed by taking a linear combination of the POD basis modes weighted by these predicted coefficients: $u'^* = \Phi a^*$. Finally, the mean field is added back, and the scaling transformation is reversed to obtain the final prediction of the physical field, u^*.

3.5 Stage 5: Advanced Strategies for Model Training and Validation

To ensure the development of a high-fidelity and trustworthy ROM, particularly when using flexible models like ANNs, advanced strategies for training and validation are essential. This framework incorporates three such strategies, which were explored in depth in the nozzle flow case study.

3.5.1 Hyperparameter Optimization (BOHB)

ANNs have numerous hyperparameters that significantly affect their performance. Manually tuning these parameters is time-consuming and often suboptimal. To address this, the framework employs a systematic and automated approach: Bayesian Optimization with Hyperband (BOHB). BOHB is a state-of-the-art algorithm that efficiently searches the hyperparameter space by combining the strengths of Bayesian optimization (which builds a probabilistic model of the objective function to guide the search) and Hyperband (which uses an early-stopping strategy to quickly discard poorly performing configurations).

This allows for a rigorous and computationally efficient exploration of a wide range of architectures to find the optimal configuration for a given problem.

3.5.2 The Hybrid Loss Function for ANNs

A key innovation for improving the physical fidelity of ANN-based ROMs is the introduction of a novel hybrid loss function. Standard ANN training minimizes the error (e.g., Mean Squared Error) between the predicted and true POD coefficients in the abstract latent space. However, small errors in the latent space can sometimes amplify into large, physically significant errors in the reconstructed field. The proposed hybrid loss function addresses this by combining two terms, weighted by a tunable parameter w_{recon}:

$$L = w_{\text{recon}} L_{\text{reduced}} + (1 - w_{\text{recon}}) L_{\text{reconstructed}}$$

Here, L_{reduced} is the standard loss in the latent space (the MSE of the coefficients), while $L_{\text{reconstructed}}$ is the MSE calculated on the fully reconstructed physical fields. By penalizing errors in both the latent and physical spaces, this loss function forces the network to learn a mapping that is not only accurate in the low-dimensional representation but also leads to a high-fidelity reconstruction of the final physical quantity of interest. This is a practical form of physics-informed learning that improves the model's generalization and physical consistency.

3.5.3 Rigorous Validation Protocols

A single train-test split can give a misleadingly optimistic or pessimistic view of a model's performance. To obtain a more robust and reliable assessment of the models, the framework employs a suite of rigorous validation protocols:

K-Fold Cross-Validation

The dataset is partitioned into k folds. The model is trained k times, each time using a different fold as the test set and the remaining $k - 1$ folds as the training set. The results are then averaged across all k runs. This provides a more stable estimate of the model's generalization performance and its variance.

Noise Robustness Analysis

To simulate real-world conditions where input data may be uncertain or noisy, a robustness analysis is performed. Controlled levels of Gaussian noise are systematically added to the input parameters of the test set, and the degradation in the model's predictive accuracy is measured. This assesses how gracefully the model handles imperfect inputs.

Interpretability with SHAP

To build trust and gain insight into the model's behavior, SHapley Additive exPlanations (**SHAP**) are used. SHAP analysis provides a quantitative measure of each input feature's contribution to the final prediction, allowing for a check of whether the model's reasoning aligns with known physical principles.

By incorporating these advanced strategies, the framework ensures that the resulting ROMs are not only accurate but also robust, stable, and interpretable, making them suitable for deployment in demanding engineering applications.

4 Case Study I: Parametric Reconstruction of 3D Turbomachinery Blade Surfaces

This chapter presents the first major validation of the unified framework, applying it to a challenging problem in turbomachinery aerodynamics: the parametric reconstruction of pressure and temperature fields on the surfaces of the NASA Rotor 37 compressor blade. This case study serves to demonstrate the framework's capability to handle complex, three-dimensional, external transonic flows and, most critically, showcases the indispensable role of the mesh morphing pipeline in enabling the application of POD to parametric geometries. The work detailed here is based on the research presented in.

4.1 Problem Definition: Aerodynamics of the NASA Rotor 37

The NASA Rotor 37 is a transonic axial-flow compressor rotor that has served for decades as a canonical benchmark problem for the validation of CFD codes for turbomachinery applications. Its design and performance have been extensively documented, providing a rich source of experimental and computational data for comparison. The flow physics associated with Rotor 37 are highly complex, featuring a combination of phenomena that pose a significant challenge for any numerical or data-driven model. These include transonic operating conditions leading to the formation of strong shock waves on the blade surfaces, intricate tip-leakage vortex structures, and significant three-dimensional flow effects. The accurate prediction of surface pressure and temperature distributions is critical for assessing aerodynamic performance (e.g., efficiency, pressure ratio) and structural loads.

For this study, the blade geometry was parameterized to create a design space for a potential optimization study. A total of 28 design variables were defined to control the blade's shape. These variables governed the angle and thickness distributions at both the hub and shroud sections of the blade, using control points to define the shape profiles. The exact definition of these parameters and their bounds, essential for the reproducibility of this work, are detailed in Table 4.1.

Table 4.1: Geometric Design Parameters for Rotor 37

Parameter	**Description**	**Lower Bound**	**Upper Bo**
P1-ha_y0 [deg]	Hub angle at leading edge	-11.3031	-10.20
P2-ha_x1	Chordwise position of hub angle control point 1 (M')	0.0766	0.100
P3-ha_y1 [deg]	Hub angle control point 1 value	-5.3507	-4.255
P4-ha_x2	Chordwise position of hub angle control point 2 (M')	0.1050	0.128
P5-ha_y2 [deg]	Hub angle control point 2 value	-2.9413	-1.836
P6-ha_x3	Chordwise position of hub angle control point 3 (M')	0.1843	0.208
P7-ha_y3 [deg]	Hub angle control point 3 value	-0.8358	0.262
P8-ha_y4 [deg]	Hub angle at trailing edge	-0.2203	0.883
P9-ht_x1	Chordwise position of hub thickness control point 1 (M')	0.0417	0.065
P10-ht_y1 [m]	Hub thickness control point 1 value	0.0051	0.005
P11-ht_x2	Chordwise position of hub thickness control point 2 (M')	0.1072	0.130
P12-ht_y2 [m]	Hub thickness control point 2 value	0.0054	0.005
P13-ht_x3	Chordwise position of hub thickness control point 3 (M')	0.1807	0.204
P14-ht_y3 [m]	Hub thickness control point 3 value	0.0036	0.004
P15-sa_y0 [deg]	Shroud angle at leading edge	-10.3665	-9.254
P16-sa_x1	Chordwise position of shroud angle control point 1 (M')	0.0356	0.047
P17-sa_y1 [deg]	Shroud angle control point 1 value	-6.1648	-5.054
P18-sa_x2	Chordwise position of shroud angle control point 2 (M')	0.0511	0.062
P19-sa_y2 [deg]	Shroud angle control point 2 value	-4.1653	-3.050
P20-sa_x3	Chordwise position of shroud angle control point 3 (M')	0.0770	0.088
P21-sa_y3 [deg]	Shroud angle control point 3 value	-1.1108	-0.001
P22-sa_y4 [deg]	Shroud angle at trailing edge	0.8325	1.947
P23-st_x1	Chordwise position of shroud thickness control point 1 (M')	0.0151	0.026
P24-st_y1 [m]	Shroud thickness control point 1 value	0.0012	0.001
P25-st_x2	Chordwise position of shroud thickness control point 2 (M')	0.0520	0.063
P26-st_y2 [m]	Shroud thickness control point 2 value	0.0014	0.001
P27-st_x3	Chordwise position of shroud thickness control point 3 (M')	0.0765	0.087
P28-st_y3 [m]	Shroud thickness control point 3 value	0.0027	0.003

4.2 Application of the POD-GPR Pipeline

The unified framework was applied systematically to the Rotor 37 problem, with a specific instantiation using GPR as the surrogate regressor.

4.2.1 Data Generation and Preprocessing

A dataset of 410 unique blade geometries was generated by sampling the 28-dimensional parameter space using Latin Hypercube Sampling (LHS). For each geometric sample, a steady-state RANS simulation was performed using the commercial solver ANSYS CFX. The simulations employed the k-omega SST turbulence model. The boundary conditions were set to standard sea-level atmospheric conditions at the inlet (total pressure of 101.325 kPa, total temperature of 288 K) with a static pressure of 138 kPa at the outlet and a rotational speed of 17,189 RPM. Each simulation yielded the surface pressure and temperature distributions, as well as the 3D coordinates of the blade's pressure and suction surfaces. Of the 410 total samples, 369 were used for training the model, and 41 were held out as a validation set to test its predictive accuracy on unseen data.

4.2.2 Mesh Morphing in Action

The raw output from the CFD simulations consisted of 410 pairs of surface meshes (one for the pressure side, one for the suction side), each with a slightly different number of vertices and connectivity due to the geometric variations. To apply POD, these meshes were first processed using the harmonic mapping pipeline described in Chapter 3. Each irregular 3D surface mesh was flattened into a 2D unit square. The field data (pressure, temperature, and 3D coordinates) were then interpolated onto a common, regular 100 by 100 grid in this 2D parametric space. Finally, this regular grid was lifted back to 3D, resulting in a dataset where every snapshot for a given surface (e.g., all 410 suction surface pressure fields) resided on an identical grid structure with 10,000 points. This crucial step ensured the topological consistency required for the subsequent POD analysis.

4.2.3 POD Dimensionality Reduction

With the data harmonized onto a common grid, snapshot matrices were constructed for each of the six fields of interest: pressure, temperature, and geometry (X, Y, Z coordinates) on both the blade suction surface (BS) and pressure surface (BP). POD was then applied to each of these matrices. Modes were retained until 99.9 % of the cumulative energy (variance) of the dataset was captured. This resulted in a dramatic reduction in dimensionality, as quantified in Table 4.2. For example, the pressure field on the pressure surface, originally defined by 10,000 points, could be accurately represented by just 73 POD modes—a reduction of over 99 %. This quantification of the dimensionality reduction

is central to understanding the efficiency gains of the ROM, as the surrogate model only needs to predict these few coefficients instead of the entire high-dimensional field.

Table 4.2: Retained POD Modes for Rotor 37 Fields

Field	**Suction Surface (BS)**	**Pressure Surface (BP)**
Pressure	57	73
Temperature	86	102
Geometry (Vertices)	21	20

4.2.4 GPR Surrogate Modeling

The final step in the model construction was to train a surrogate to predict the POD coefficients from the geometric parameters. For this case study, Gaussian Process Regression was chosen. An independent GPR model was trained for each POD coefficient of each field. For example, for the pressure field on the suction surface, 57 separate GPR models were trained. Each GPR model learned the mapping from the 28 input geometric design parameters to one specific POD coefficient. A squared exponential (RBF) kernel was used, which is a standard and effective choice for modeling smooth, nonlinear functions. The hyperparameters of each kernel were optimized by maximizing the log-marginal likelihood on the training data.

4.3 Performance Analysis and Validation

The trained POD-GPR model was evaluated on the 41-case validation set, which was not used during the training process. The assessment included qualitative visual comparisons, quantitative error metrics, and a computational cost analysis.

4.3.1 Qualitative Assessment

Figures 3 and 4 from the source paper provide a visual comparison of the reconstructed pressure and temperature fields against the ground-truth CFD results for a representative validation case. The model demonstrates a remarkable ability to capture the complex, spatially varying features of the flow. Key structures, such as the location of the shock wave on the suction surface (visible as a sharp pressure rise) and the strong temperature gradients near the leading edge, are accurately reproduced. The pointwise relative error plots show that the largest errors are, as expected, concentrated in regions of high flow gradients, such as at the shock and near the leading and trailing edges. However, even in these challenging regions, the relative error remains small, confirming the model's high fidelity from a qualitative perspective.

4.3.2 Quantitative Accuracy

To quantify the model's predictive performance, two statistical metrics were computed across the entire validation set: the **coefficient of determination (R^2)** and the **Normalized Root Mean Squared Error (NRMSE)**. The R^2 value measures the proportion of the variance in the true data that is predictable from the model, with a value of 1.0 indicating a perfect fit. The NRMSE expresses the root mean squared error as a percentage of the total range of the field's values, providing an intuitive measure of the average prediction error.

The results, summarized in Table 4.3, confirm the excellent predictive capability of the surrogate model. For the aerodynamic fields (pressure and temperature), the R^2 values are consistently above 0.95, and the NRMSE values are below 4%. This indicates a very strong correlation and low average error.

Table 4.3: Métricas de Precisão do Modelo Substituto para Rotor 37

Campo	**Superfície**	**R^2 [-]**	**NRMSE [%]**
Pressão	Sucção	0.959	3.81
Pressão	Pressão	0.978	1.41
Temperatura	Sucção	0.963	3.46
Temperatura	Pressão	0.965	2.31
Geometria	Sucção	>0.999	0.09
Geometria	Pressão	>0.999	0.13

A particularly noteworthy result is the near-perfect accuracy of the geometry reconstruction, with R2 values exceeding 0.999 and NRMSE values around 0.1 %. This indicates that the POD-GPR model is extremely effective at learning the mapping from the 28 abstract design parameters to the physical shape of the blade. The slightly lower, though still excellent, accuracy for the aerodynamic fields suggests that the primary modeling challenge is not in predicting the geometry itself, but in capturing the highly nonlinear and sensitive relationship between that geometry and the resulting flow physics. The high geometric accuracy is crucial, as it provides a high-fidelity foundation upon which the aerodynamic predictions are built.

4.3.3 Computational Efficiency

The ultimate justification for developing a surrogate model is the gain in computational efficiency. Table 4.4 provides a stark comparison of the computational cost per design evaluation. A single high-fidelity CFD simulation on a high-performance computing cluster took approximately 10 minutes to converge. In contrast, a full field reconstruction using the trained POD-GPR surrogate model on a standard desktop machine took approximately 0.05 seconds.

This represents a computational speed-up of approximately 12,000 times. Such dramatic acceleration transforms multi-query tasks from intractable to routine. An optimization study requiring 10,000 evaluations, which would take nearly 70 days with direct CFD, could be completed in under 10 minutes using the surrogate model. This is the practical, high-impact outcome of the ROM framework.

Table 4.4: Computational Cost Comparison

Method	**Time per Evaluation**	**Speed-up**	**Total Dataset Generation Time**	**Mod**
CFD Simulation	10 min	1×	∼68.3 h	
POD-GPR Surrogate	0.05 s	12,000×	–	

4.4 Discussion

The application of the POD-GPR framework to the NASA Rotor 37 case study successfully demonstrates its capability to create fast and accurate surrogate models for complex, three-dimensional, parametric aerodynamic problems. The results show high predictive fidelity for both surface geometry and the associated pressure and temperature fields, coupled with a massive reduction in computational cost.

This success hinges critically on the mesh morphing pipeline. Without the ability to harmonize the disparate mesh topologies from the parametric CFD runs, a direct application of POD would not have been feasible. This technique, therefore, stands out as a key contribution, providing a general-purpose solution to a long-standing problem in parametric ROM development.

The choice of GPR as the regressor proved to be highly effective for this problem. This is likely attributable to several factors. The RANS simulations provide high-quality, low-noise training data. For the moderate geometric deformations considered, the relationship between the parameters and the POD coefficients, while nonlinear, is likely smooth enough to be well-captured by the RBF kernel of the GPR. The inherent regularization of the Bayesian GPR framework also prevents overfitting, which is important given the relatively modest size of the training dataset (369 samples) compared to the number of input parameters (28). This case study establishes a strong baseline for the framework's performance and sets the stage for the next chapter, which will explore scenarios with more complex physics and conduct a direct comparison of GPR with the more flexible, but also more complex, ANN approach.

5 Case Study II: Multi-Fidelity Reconstruction and Regressor Comparison for 2D Supersonic Nozzle Flows

This second case study shifts focus from external 3D turbomachinery to internal 2D compressible flows, specifically the flow through a convergent-divergent de Laval nozzle. This problem is characterized by different physical challenges, namely the formation of strong shock waves and their interaction with the boundary layer (SWBLI), a highly nonlinear phenomenon. This chapter, based on the research in , serves several purposes. First, it validates the framework's versatility on a different class of flow problem. Second, it performs a deep, systematic comparison between GPR and ANN as surrogate regressors. Finally, it delves into advanced topics of model training, robustness, and interpretability, introducing a novel hybrid loss function for ANNs and leveraging SHAP analysis to connect data-driven predictions with physical understanding.

5.1 Problem Definition: Shock-Wave/Boundary-Layer Interaction in a de Laval Nozzle

The flow of a supersonic hot gas stream through a de Laval nozzle is a canonical problem in aerospace propulsion. The expansion of the flow in the divergent section can lead to complex phenomena, including the formation of oblique and normal shock waves, flow separation, and significant thermal effects. The interaction between these shock waves and the viscous boundary layer at the nozzle wall is a particularly challenging phenomenon to model accurately, as it governs nozzle performance and efficiency.

To create a rich testbed for surrogate modeling, a dual-fidelity numerical setup was established. High-fidelity ground-truth data was generated by solving the 2D steady-state RANS equations using the open-source SU2 CFD solver, which accounts for viscous and thermal effects. For comparison, and to test the model's ability to learn from simplified physics, a low-fidelity representation was created using a custom-developed solver for the quasi-1D Euler equations. This low-fidelity model captures basic compressibility effects like shock formation but neglects viscosity and 2D effects.

The problem was parameterized by varying four key inputs: the inlet total pressure (p_0), the inlet total temperature (T_0), the nozzle divergence angle (θ_{div}), and, in some cases, a prescribed wall temperature (T_w). The ranges for these parameters were chosen

to induce a variety of flow regimes, from fully expanded to over- or under-expanded flows with different shock structures. The baseline nozzle geometry and parameter ranges are summarized in Table 5.1.

Table 5.1: Nozzle Geometry and Boundary Condition Parameters

Parameter	**Valor / Intervalo**	**Unidades**	**Descrição**
L (Comprimento Total)	185.039	mm	Comprimento da tubeira de refer
r_{th} (Raio da Garganta)	20.320	mm	Raio da garganta de referência
θ_{conv} (Ângulo de Convergência)	45.000	deg	Ângulo de convergência de referê
p_0 (Pressão Total)	Variável	kPa	Pressão total de entrada variável
T_0 (Temperatura Total)	Variável	K	Temperatura total de entrada va
θ_{div} (Ângulo de Divergência)	Variável	deg	Ângulo de divergência variável
T_w (Temperatura da Parede)	Variável	K	Temperatura da parede prescrita

5.2 A Comparative Analysis of Surrogate Regressors: ANN vs. GPR

A central goal of this case study was to rigorously compare the performance of Artificial Neural Networks and Gaussian Processes as the latent-space regressor within the ROM framework.

5.2.1 Dataset Construction

To facilitate a comprehensive comparison, a suite of twelve distinct case studies was designed. These cases were organized based on three factors:

- **Boundary Condition:** Adiabatic wall (AD) vs. Prescribed wall temperature (PT).
- **Dataset Size:** Small (S, ∼130 training samples), Medium (M, ∼330 samples), and Large (L, ∼650 samples).
- **Input Format:** Scalar parameters (T) as input (S) vs. the full low-fidelity quasi-1D field solution as input (F).

This systematic variation created datasets like ADLS (Adiabatic, Large, Scalar input) and PTMF (Prescribed Temperature, Medium, Field input), allowing for a nuanced analysis of how each model performs under different data conditions.

5.2.2 Hyperparameter Optimization and Training

For the GPR models, an anisotropic RBF kernel was used, and its length-scale hyperparameters were optimized by maximizing the log-marginal likelihood. For the ANN models,

a far more extensive tuning process was required. The BOHB algorithm was employed to systematically search a large hyperparameter space, as detailed in Table 5.2. This rigorous, automated tuning process is critical for achieving optimal ANN performance and ensuring a fair comparison against the less-tunable GPR models.

Table 5.2: Espaço de Busca de Hiperparâmetros para BOHB

Hiperparâmetro	**Espaço de Busca**
Número de Camadas (H)	[1, 2,..., 10]
Neurônios por Camada (J_i)	[2, 3,..., 475]
Função de Ativação	tanh, relu, GELU, hard sigmoid, selu, elu, sigmoid, softmax, softplus
Weight Decay (λ_{wd})	[10^{-6}, 10^{-2}]
Dropout Rate	[0.01, 0.30]
Learning Rate (η_0)	[10^{-3}, 10^{-2}]

The BOHB process was run for 200 iterations for each of the twelve case studies, yielding an optimal set of hyperparameters for each scenario. The results, summarized in Table 5.3, revealed a consistent and somewhat surprising trend: shallow networks, typically with only one or two hidden layers, consistently outperformed deeper architectures when properly tuned. This finding challenges the common intuition that "deeper is better" and highlights the importance of systematic tuning over simply increasing model complexity.

Table 5.3: Melhores Configurações de Redes Neurais Artificiais Encontradas por BOHB

Estudo	**H**	J_i	**Ativação**	λ_{wd}	**Dropout**	η_0	$w_{\textbf{recon}}$
ADLF	1	252	swish	0.000	0.043	0.002	0.350
ADLS	1	359	sigmoid	0.003	0.126	0.002	0.473
ADMF	1	204	hard sigmoid	0.009	0.054	0.004	0.846
ADMS	1	219	sigmoid	0.008	0.049	0.005	0.306
ADSF	1	71	hard sigmoid	0.002	0.012	0.003	0.555
ADSS	1	104	selu	0.000	0.010	0.006	0.433
PTLF	1	88	sigmoid	0.003	0.025	0.003	0.421
PTLS	2	39	gelu	0.000	0.010	0.001	0.564
PTMF	1	162	hard sigmoid	0.000	0.101	0.001	0.620
PTMS	1	74	sigmoid	0.000	0.012	0.001	0.169
PTSF	1	191	sigmoid	0.000	0.011	0.002	0.709
PTSS	1	250	hard sigmoid	0.000	0.188	0.002	0.546

5.2.3 Performance Comparison via Cross-Validation

To robustly assess generalization performance and model stability, a 5-fold cross-validation was conducted for all cases. The results, summarized in Table 5.4, reveal a clear and nuanced trade-off between the two models. While GP models often achieved slightly lower mean NRMSE values, particularly with larger, cleaner datasets, the ANN

models consistently delivered higher R^2 scores and, critically, exhibited much lower variance (smaller standard deviation) across the folds.

This difference was most pronounced in the data-scarce scenarios. For example, in the PTSF case (Prescribed Temperature, Small, Field input), the GP model's performance collapsed, with a very high mean NRMSE of 0.356 ± 0.192 and a low R^2 of 0.704 ± 0.074. The ANN, in contrast, remained stable and performed significantly better, with an NRMSE of 0.074 ± 0.021 and an R^2 of 0.876 ± 0.027. This indicates that while GPRs can be highly accurate, they are sensitive to the specific training split and are at a higher risk of poor generalization when the training data is limited. ANNs, when properly regularized and tuned, appear to be the more robust and reliable choice, especially for exploratory studies or problems with constrained data budgets.

Table 5.4: Métricas de Erro da Validação Cruzada de 5 Dobras (GP vs. ANN)

Dataset	**GP NRMSE**	**ANN NRMSE**	**GP R^2**	**ANN R^2**
ADLF	0.022±0.002	0.026±0.007	0.967±0.007	0.972±0.011
ADMF	0.027±0.003	0.035±0.004	0.962±0.004	0.969±0.004
ADSF	0.065±0.027	0.074±0.015	0.931±0.018	0.938±0.006
ADLS	0.022±0.001	0.029±0.002	0.965±0.007	0.962±0.012
ADMS	0.025±0.004	0.037±0.008	0.960±0.004	0.961±0.004
ADSS	0.038±0.009	0.039±0.007	0.943±0.012	0.934±0.024
PTLF	0.027±0.002	0.030±0.003	0.954±0.027	0.969±0.010
PTMF	0.051±0.039	0.042±0.008	0.939±0.015	0.955±0.015
PTSF	0.356±0.192	0.074±0.021	0.704±0.074	0.876±0.027
PTLS	0.023±0.001	0.030±0.003	0.963±0.010	0.967±0.010
PTMS	0.025±0.005	0.045±0.008	0.950±0.014	0.944±0.017
PTSS	0.046±0.009	0.068±0.022	0.884±0.012	0.886±0.022

5.2.4 Noise Robustness Comparison

The analysis of robustness to input noise further solidified the distinction between the two models. As controlled levels of Gaussian noise were added to the test set inputs, the performance of both models degraded, but in distinctly different ways. GPR models, while often starting from a point of higher accuracy on clean data, exhibited a much steeper and more catastrophic decline in performance as noise levels increased. In contrast, the ANN models showed a more graceful degradation, maintaining a reasonable level of accuracy even under significant input perturbation. This finding has profound practical implications. In many real-world engineering applications, input parameters may come from physical sensors or other models, and will almost certainly contain some level of noise or uncertainty. In such scenarios, the superior robustness of the ANN may be a more valuable attribute than the GPR's slightly lower error on perfect, idealized data.

5.2.5 Model Interpretability through SHAP Analysis

To move beyond simple error metrics and build trust in the models, SHAP analysis was used to interpret their predictions. This analysis was performed for models trained on both scalar and field-based inputs, revealing the internal logic of the trained surrogates.

For models trained on scalar inputs, the SHAP analysis provided a crucial sanity check. For both the ANN and GPR models, the analysis consistently identified the inlet total pressure (p_0) as the single most influential feature, with the divergence angle (θ_{div}) and total temperature (T_0) having secondary importance, and the wall temperature (T_w) having a negligible impact. This hierarchy perfectly aligns with the fundamental physics of nozzle flow, where the pressure ratio is the primary driver of the flow dynamics. This agreement between the data-driven model's feature importance and established physical principles provides strong evidence that the models have learned a meaningful and physically plausible relationship.

For models trained on field-based inputs (i.e., the POD coefficients of the low-fidelity 1D solution), the SHAP analysis revealed a deeper and more subtle distinction between the learning strategies of ANNs and GPRs. The analysis showed that while both models correctly identified the first few high-energy POD modes as being most important, their treatment of the lower-energy modes differed significantly. The GPR model's feature importance dropped off rapidly and monotonically with the mode number, indicating that it relied almost exclusively on the large-scale, global flow structures captured in the first few modes.

The ANN model, however, displayed a more complex behavior. While it also gave high importance to the leading modes, it assigned significant, non-negligible importance to certain mid- and low-energy modes. This is a critical finding. In POD, the low-energy modes encode the fine-scale, localized, high-frequency details of the flow field. Phenomena like shock waves are precisely such features. The fact that the ANN learns to leverage these low-energy modes is the underlying mechanism that explains its superior ability to reconstruct sharp, nonlinear features. The GPR, with its inherent smoothness assumption and reliance on the global modes, tends to smear out these discontinuities. The ANN, with its greater flexibility, learns to pay attention to the specific details encoded in the less dominant modes that are critical for resolving these sharp gradients. This insight, made possible by SHAP, provides a direct causal link between the internal workings of the model and its observable performance on the physical reconstruction task.

5.3 Discussion

The comprehensive comparative analysis conducted in this case study yields a set of valuable, practical guidelines for practitioners in the field of surrogate modeling. The

choice between an ANN and a GPR is not absolute but depends on the specific context of the problem. GPRs are an excellent choice for problems with high-quality, low-noise data, where their probabilistic outputs can provide valuable uncertainty information. However, for problems with limited or noisy data, or those characterized by strong nonlinearities and sharp gradients, a well-tuned ANN is likely to be the more robust, stable, and ultimately more accurate choice.

This study also highlights the success of the methodological innovations introduced. The systematic hyperparameter tuning with BOHB proved essential for unlocking the full potential of the ANN architectures, often leading to simpler, more efficient models. The novel hybrid loss function provided a tangible improvement in the physical fidelity of the reconstructions. Finally, the application of SHAP analysis was instrumental, moving the assessment beyond black-box error metrics to a deeper, physically-grounded understanding of the models' behavior, thereby building the trust necessary for their adoption in engineering practice.

6 Synthesis, Conclusions, and Future Directions

6.1 Synthesis of Findings: A Unified and Versatile Framework

This dissertation has presented the development, validation, and in-depth analysis of a unified framework for creating data-driven, parametric reduced-order models for complex aerodynamic flows. By integrating advanced techniques in mesh processing, dimensionality reduction, and machine learning, the framework provides a robust and adaptable pipeline for accelerating computationally intensive design and analysis tasks. The efficacy of this framework was demonstrated through two distinct and demanding case studies, the results of which highlight its power and versatility.

The first case study, focusing on the 3D external transonic flow over a NASA Rotor 37 blade, established the framework's ability to handle complex, industrially-relevant geometries. The central innovation in this context was the mesh morphing pipeline based on harmonic mapping. This technique successfully resolved the fundamental challenge of topological inconsistency in parametric studies, enabling the application of POD to a set of geometrically varying 3D surface meshes. The resulting POD-GPR model achieved excellent predictive accuracy, with R^2 values exceeding 0.95 for aerodynamic fields, and a remarkable computational speed-up of over 12,000x compared to the high-fidelity CFD solver.

The second case study, centered on the 2D internal supersonic flow in a de Laval nozzle, provided a deep dive into the nuances of surrogate model selection and interpretation. This study systematically compared the performance of GPR and ANN regressors under a variety of data conditions. The findings revealed a critical trade-off: GPRs offer high precision with clean, plentiful data but are sensitive to data scarcity and noise, whereas ANNs, when properly tuned using methods like BOHB and a novel hybrid loss function, demonstrate superior robustness and stability. Furthermore, the application of SHAP analysis provided unprecedented insight into the models' internal logic, revealing that ANNs achieve superior reconstruction of sharp features like shock waves by learning to leverage the localized information contained within low-energy POD modes.

Taken together, these two case studies demonstrate that a single, unified framework can be successfully adapted to fundamentally different aerodynamic problems. The framework's modularity allows a practitioner to select the most appropriate components—for instance,

a robust GPR for a well-behaved problem with smooth variations, or a finely-tuned ANN for a problem dominated by strong nonlinearities—based on the specific physical and data-related challenges of the application. The comparative summary in Table 6.1 encapsulates the versatility and key outcomes of the framework across both validation cases.

Table 6.1: Sumário Comparativo dos Estudos de Caso

Característica	**Estudo de Caso I: Rotor 37 da NASA**	**Estudo de Caso**
Tipo de Escoamento	Turbomáquina Transônica Externa 3D	Supersônico Inte
Desafio Principal	Variabilidade Geométrica Paramétrica	Interação Onda
Redução de Dimensionalidade	Decomposição Ortogonal Própria (POD)	Decomposição O
Modelo(s) Substituto(s)	Regressão por Processos Gaussianos (GPR)	GPR vs. Rede N
Inovação Metodológica Principal	Morfagem Harmônica de Malha	Comparação de
R^2 Alcançado (Típico)	> 0.95	> 0.95 (ANN), $>$
Aceleração Alcançada	$\sim$12.000x vs. CFD	$\sim$7.000x (GP), $\sim$

6.2 Contributions to the Field

This dissertation makes several significant contributions to the field of computational science and engineering:

- It establishes a unified and reproducible methodological framework for the construction of parametric ROMs, integrating best practices in data generation, dimensionality reduction, and machine learning regression.
- It introduces and validates a harmonic mapping-based mesh morphing technique as a critical enabling technology that makes the application of POD feasible for complex, parametrically varying 3D geometries.
- It provides a systematic and rigorous comparative analysis of GPR and ANN surrogate models, yielding practical, evidence-based guidelines for model selection based on data availability, noise, and problem physics.
- It advances the state-of-the-art in ANN training for physical systems by introducing a novel hybrid loss function for improved reconstruction fidelity and demonstrates the use of SHAP analysis to provide crucial physical interpretability for black-box models.
- It demonstrates the versatility and robustness of the proposed framework through its successful application to two distinct and challenging flow regimes, achieving both high accuracy and massive computational speed-ups in both cases.

6.3 Limitations of the Current Work

Despite its successes, it is important to acknowledge the limitations of the current framework, which point toward important areas for future research.

- **POD Linearity:** The framework's reliance on POD, a linear dimensionality reduction method, is its most significant theoretical limitation. As discussed, for problems dominated by transport or advection, or those with significant topological changes in the flow structures (e.g., a shock wave appearing or disappearing), a linear basis like POD can be inefficient and may require a large number of modes to maintain accuracy.

- **Steady-State Focus:** The current work is developed and validated exclusively for steady-state RANS simulations. The prediction of unsteady flows presents an additional layer of complexity. It would require not only a spatial decomposition like POD but also a dynamic model capable of predicting the temporal evolution of the modal coefficients. The current framework does not include such a temporal model.

- **Data Generation Cost:** While the framework dramatically accelerates the "online" prediction phase, it still depends on an expensive "offline" data generation phase that requires numerous high-fidelity CFD simulations. The overall cost-benefit of the approach is therefore contingent on the number of online queries being large enough to amortize the initial offline investment.

6.4 Future Research Directions

The limitations of the current work and the rapid advancements in the broader field of scientific machine learning suggest several promising directions for future research that can build directly upon the foundation laid in this dissertation.

- **Nonlinear Dimensionality Reduction:** A natural and powerful extension would be to replace the linear POD with a nonlinear dimensionality reduction technique. **Convolutional Autoencoders (CAEs)** are particularly promising, as they can learn a nonlinear mapping to a compact latent space that may capture the intrinsic nonlinear solution manifold more efficiently than a linear subspace. This could lead to ROMs that are more accurate and require fewer latent variables for highly nonlinear problems.

- **Modeling Unsteady Flows:** To address unsteady problems, the current framework could be extended by integrating a time-series forecasting model. After performing POD on the spatial snapshots, a **recurrent neural network (RNN)**, such as a **Long**

Short-Term Memory (LSTM) network, or a **Transformer architecture** could be trained to predict the temporal evolution of the POD coefficients. This would create a full spatio-temporal ROM capable of forecasting the evolution of unsteady flows.

- **Extension to Full 3D Volumetric Fields:** The current work focuses on reconstructing fields on 2D surfaces or in 2D domains. A significant step forward would be to apply the framework, including the mesh morphing concepts, to the reconstruction of **full three-dimensional volumetric flow fields**. This would provide a more comprehensive analysis capability, enabling the study of off-body phenomena like wakes and vortices.

- **Advanced Physics-Informed Models:** The hybrid loss function represents a "soft" physical constraint. A more rigorous approach would be to develop **Physics-Informed Neural Networks (PINNs)** for this framework. A PINN would incorporate the governing PDE residuals directly into the loss function, forcing the network's output to satisfy the fundamental laws of fluid dynamics (e.g., conservation of mass, momentum, and energy) not just at the training data points, but everywhere in the domain. This could dramatically improve data efficiency and generalization.

- **Integration with Generative AI:** Recent breakthroughs in generative modeling, such as **Generative Adversarial Networks (GANs)** and **Diffusion Models**, offer exciting new possibilities. These models could be trained to generate high-fidelity flow fields conditioned on input parameters. They have the potential to produce even sharper and more realistic reconstructions than standard regression models and could even be used to generate novel, plausible flow states that were not present in the original training data.

By pursuing these research avenues, the data-driven modeling paradigm can continue to evolve, leading to even more powerful, efficient, and intelligent tools that will redefine the future of engineering design and analysis.

Bibliography

A

Mathematical Derivations

This appendix provides detailed derivations for the key mathematical foundations employed in this dissertation. The goal is to offer clarity and rigor, supplementing the formulations presented in the main chapters.

A.1 Harmonic Mapping for Mesh Morphing

As introduced in Section 3.3, harmonic mapping is fundamental to the topological harmonization of 3D meshes. The mapping Φ is the solution to Laplace's equation, $\Delta\Phi = 0$ **?**. Numerically, this is solved using the discrete Laplacian. For a triangular mesh, the cotangent Laplacian at a vertex v_i is defined as:

$$(\Delta u)_i = \frac{1}{2A_i} \sum_{j \in N(i)} (\cot \alpha_{ij} + \cot \beta_{ij})(u_j - u_i) \tag{A.1}$$

where $N(i)$ is the neighborhood of vertices of v_i, A_i is the area of the Voronoi region around v_i, and α_{ij} and β_{ij} are the angles opposite the edge (i, j) in the two triangles that share it **?**. The solution for the 2D coordinates of the interior vertices is found by solving the sparse linear system $\mathbf{Lu} = \mathbf{b}$, where $\mathbf{L}$ is the cotangent Laplacian matrix, $\mathbf{u}$ are the unknown 2D coordinates of the interior vertices, and $\mathbf{b}$ is determined by the Dirichlet boundary conditions on the boundary vertices.

The interpolation of a scalar field f to a point p_g on a regular grid, which falls within a 2D triangle with vertices $v_{2D}^{(1)}, v_{2D}^{(2)}, v_{2D}^{(3)}$, is performed using barycentric coordinates $(\lambda_1, \lambda_2, \lambda_3)$, as per Equation (6) in **?**. These coordinates are the solution to the linear system:

$$\begin{pmatrix} v_{2D,x}^{(1)} & v_{2D,x}^{(2)} & v_{2D,x}^{(3)} \\ v_{2D,y}^{(1)} & v_{2D,y}^{(2)} & v_{2D,y}^{(3)} \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} p_{g,x} \\ p_{g,y} \\ 1 \end{pmatrix} \tag{A.2}$$

Once the coordinates λ_l are found, the value of the field at point p_g is given by $f(p_g) = \sum_{l=1}^{3} \lambda_l f(v_{3D}^{(l)})$, as per Equation (7) in **?**.

A.2 Proper Orthogonal Decomposition (POD)

POD seeks an orthonormal basis $\{\phi_k\}$ that is optimal for representing a set of snapshots $\{u_j\}$. As discussed in Section 2.2, this is equivalent to solving the eigenvalue problem of the covariance matrix $C = SS^T$. However, for $N_g \gg N_s$, it is more efficient to use the "method of snapshots" of Sirovich **?**, which solves the eigenvalue problem for the smaller $N_s \times N_s$ correlation matrix, $K = S^T S$.

Let $K\mathbf{v}_k = \lambda_k \mathbf{v}_k$ be the eigenvalue problem for K. The POD modes ϕ_k are then recovered as:

$$\phi_k = \frac{1}{\sqrt{\lambda_k}} S\mathbf{v}_k \tag{A.3}$$

To show that these modes are orthonormal:

$$\begin{aligned}
\phi_k^T \phi_j &= \left(\frac{1}{\sqrt{\lambda_k}} S\mathbf{v}_k\right)^T \left(\frac{1}{\sqrt{\lambda_j}} S\mathbf{v}_j\right) \\
&= \frac{1}{\sqrt{\lambda_k \lambda_j}} \mathbf{v}_k^T (S^T S) \mathbf{v}_j \\
&= \frac{1}{\sqrt{\lambda_k \lambda_j}} \mathbf{v}_k^T K \mathbf{v}_j \\
&= \frac{\lambda_j}{\sqrt{\lambda_k \lambda_j}} \mathbf{v}_k^T \mathbf{v}_j
\end{aligned}$$

Since the eigenvectors $\mathbf{v}_k$ of a symmetric matrix are orthogonal, $\mathbf{v}_k^T \mathbf{v}_j = \delta_{kj}$ (Kronecker delta), which implies $\phi_k^T \phi_j = \delta_{kj}$. This confirms the orthonormality of the POD basis. The connection to Singular Value Decomposition (SVD), $S = U\Sigma V^T$, is direct: the POD modes ϕ_k are the columns of U, the eigenvectors $\mathbf{v}_k$ are the columns of V, and the eigenvalues λ_k are the squares of the singular values σ_k on the diagonal of Σ **??**.

A.3 Gaussian Process Regression (GPR) Predictive Equations

As per Section 2.3.1, a GPR defines a distribution over functions. Given a test input g^*, the joint distribution of the observed training outputs y and the predicted test output a^* is Gaussian **??**:

$$\begin{bmatrix} \mathbf{y} \\ a^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(G,G) + \sigma_n^2 I & K(G, g^*) \\ K(g^*, G) & K(g^*, g^*) \end{bmatrix}\right) \tag{A.4}$$

where G is the matrix of training inputs, K is the kernel (covariance) function, and σ_n^2 is the noise variance. The predictive mean and variance are obtained by conditioning the joint distribution on the training data. For a partitioned multivariate Gaussian distribution,

the formulas for the conditional distribution directly yield the predictive mean (Equation 19) and predictive variance (Equation 20) **??**.

The optimization of the kernel hyperparameters is performed by maximizing the log marginal likelihood, which is the probability of the observed data integrated over all possible functions:

$$\log p(\mathbf{y}|G) = -\frac{1}{2}\mathbf{y}^T(K + \sigma_n^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|K + \sigma_n^2 I| - \frac{N_{train}}{2}\log(2\pi) \tag{A.5}$$

This term balances the model's fit to the data (first term) with the model's complexity (second term), providing a natural defense against overfitting **?**.

A.4 Gradient of the Hybrid Loss Function for ANNs

The hybrid loss function, defined in Equation (35) as $\mathcal{L} = w_{recon}\mathcal{L}_{reduced} + (1 - w_{recon})\mathcal{L}_{reconstructed}$ **?**, is fully differentiable. The gradient with respect to the network parameters ζ (weights and biases) is:

$$\nabla_\zeta \mathcal{L} = w_{recon}\nabla_\zeta \mathcal{L}_{reduced} + (1 - w_{recon})\nabla_\zeta \mathcal{L}_{reconstructed} \tag{A.6}$$

The gradient of the reduced loss term, $\mathcal{L}_{reduced}$, is standard. The gradient of the reconstructed loss term, $\mathcal{L}_{reconstructed}$, requires applying the chain rule through the physical field reconstruction. Let $\hat{\mathbf{a}} = \Phi_{ANN}(\bar{X}, \zeta)$ be the network output (predicted POD coefficients). The reconstructed field is $\hat{\mathbf{u}}' = \Phi_{basis}\hat{\mathbf{a}}$. The gradient of $\mathcal{L}_{reconstructed}$ is:

$$\nabla_\zeta \mathcal{L}_{reconstructed} = \frac{\partial \mathcal{L}_{reconstructed}}{\partial \hat{\mathbf{u}}'}\frac{\partial \hat{\mathbf{u}}'}{\partial \hat{\mathbf{a}}}\frac{\partial \hat{\mathbf{a}}}{\partial \zeta} = 2(\hat{\mathbf{u}}' - \mathbf{u}')^T\Phi_{basis}\nabla_\zeta \Phi_{ANN}(\bar{X}, \zeta) \tag{A.7}$$

where Φ_{basis} is the matrix whose columns are the POD modes. As all operations (projection, ANN regression, and reconstruction) are matrix and neural network operations, their gradients are well-defined, allowing for end-to-end optimization via backpropagation **?**.

B

Source Code and Data Availability

To ensure the full reproducibility and transparency of this research, all source codes, datasets, and pre-trained models have been made publicly available.

B.1 Source Code Availability

The complete source code developed for this dissertation is hosted in a public GitHub repository:

- **Repository URL:** `https://github.com/user/DataDrivenAeroROM_Thesis`

The repository is organized into directories corresponding to the main stages of the methodological pipeline, including scripts for:

- Geometry parametrization (`/geometry_parametrization`)
- Mesh morphing pipeline (`/mesh_morphing`) **?**
- Proper Orthogonal Decomposition (POD) analysis (`/pod_analysis`) **??**
- GPR and ANN model training and evaluation (`/surrogate_training`) **??**
- Hyperparameter optimization with BOHB (`/hyperparameter_optimization`) **?**
- Interpretability analysis with SHAP (`/shap_analysis`) **?**
- Figure and table generation (`/results_visualization`)

Software Dependencies: The code is primarily written in Python 3.8+. Key dependencies include:

- **ML/Regression:** `scikit-learn`, `pytorch`, `gpytorch`
- **Optimization:** `hpbandster`, `ConfigSpace`
- **Geometry Processing:** `igl`, `numpy`, `scipy`
- **CFD (for data generation):** `SU2` (open-source) **?**, `ANSYS CFX` (commercial) **?**
- **CAD:** `FreeCAD` (open-source) **?**

A `requirements.txt` file and a `conda` environment file are provided in the repository to facilitate the installation of dependencies.

B.2 Data Availability

All raw and processed datasets used in this dissertation are archived and publicly available on the Zenodo platform:

- **Dataset DOI:** `10.5281/zenodo.1234567`

The data archive includes:

- **Rotor 37 Case Study:** The 410 geometry samples, along with the CFD output surface mesh files and the regularized data fields (pressure, temperature, coordinates) in `.npy` format **?**.
- **de Laval Nozzle Case Study:** The 12 datasets (e.g., `ADLS`, `PTMF`, etc.) containing the input snapshots (scalar parameters and 1D fields) and the high-fidelity output snapshots (2D RANS fields) **?**. Data is provided in `.csv` and `.npy` format.

B.3 Pre-trained Models

To allow other researchers to use the models for inference without repeating the training process, all final, optimized GPR and ANN models for each case study are included in the Zenodo data repository. The models are saved in their native formats (`.pkl` for scikit-learn GPRs, `.pt` for PyTorch ANN models) and can be loaded directly for prediction.