



**Politecnico
di Torino**

ScuDo

Scuola di Dottorato - Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation

Doctoral Program in Energy Engineering (38th cycle)

Data-Efficient Surrogate Modeling of Diesel Engine Emissions via Prior-Informed Machine Learning Model

By

Filippo Aglietti

Supervisor(s):

Prof. Federico Millo, Supervisor

Prof. Andrea Piano, Co-Supervisor

Dr. Francesco Della Santa, Co-Supervisor

Doctoral Examination Committee:

Prof. Jakob Andert, RWTH Aachen University

Prof. Felix Leach, University of Oxford

Politecnico di Torino

2026

Abstract

The progressive tightening of emission regulations for internal combustion engines has made the accurate prediction of pollutant formation a central task in modern powertrain development. Data-driven surrogate models, most notably Multi-Layer Perceptrons (MLPs) and Gaussian Process Regression (GPR), have become established tools for this purpose, yet their accuracy ultimately depends on the size and representativeness of the experimental training set, which remains costly and time-consuming to acquire on real engine hardware. Prior-informed modeling strategies can alleviate this data dependency by incorporating domain knowledge into the training process: Physics-Informed Neural Networks (PINNs) embed closed-form governing equations as soft constraints on the loss function, while Sobolev training augments the data-fitting objective with quantitative gradient observations. However, in many engineering problems, including diesel emission modeling, neither ingredient is available. The governing physics is well established, but its equations act on in-cylinder thermochemical fields rather than on the control parameters that the surrogate takes as inputs, so no differentiable residual constrains the mapping to be learned; derivative observations, in turn, are not produced by experimental campaigns. What domain expertise does provide is a qualitative expectation on the regularity of the response, typically that it varies smoothly with respect to the control inputs. Such knowledge cannot be cast in the structured, quantitative form that the methods above require.

This thesis addresses the resulting gap through the development and assessment of *Gradient-Informed Neural Networks* (GradINNs), a novel deep learning framework that embeds qualitative prior beliefs about the gradient behavior of the target function directly into the training process. The framework couples a primary network, which approximates the target function, with an auxiliary network whose architecture and initialization encode the hypothesized qualitative gradient behavior. A customized loss function enforces consistency between the input-based Jacobian of the primary

network and the output of the auxiliary network at a set of unlabeled collocation points, while the joint optimization of both networks allows the prior constraint to be progressively relaxed as the auxiliary network adapts to the evolving data-driven gradients of the primary network.

After a comprehensive validation of the methodology across different classes of synthetic and real-world regression problems, GradINN has been applied to the prediction of NO_x , particulate matter, total unburned hydrocarbons, and carbon monoxide emissions from a production 2.0-liter, four-cylinder diesel engine equipped with variable-geometry turbocharging, dual-loop exhaust gas recirculation, and a common-rail multi-pulse injection system, representative of modern passenger-car powertrains. The model is trained on a dedicated Design of Experiments campaign conducted on a dynamometer test bench, and its predictive performance is assessed on both an in-sample test set and an independent engine map representative of the production calibration. A systematic comparison against rigorously tuned MLP and GPR baselines shows that GradINN consistently achieves the lowest prediction errors on all four emission targets, with RMSE reductions of 11.9% to 20.0% relative to the best MLP and of 12.3% to 52.1% relative to the best GPR on the experimental test set, uniform improvements across low, mid, and high emission intensities, and preserved accuracy under realistic production calibration conditions. A data-efficiency analysis further demonstrates that GradINN matches the accuracy of the full-data MLP baseline using approximately 25% fewer training samples, with direct implications for the reduction of experimental effort in engine calibration campaigns.

References