A Benchmarking Framework for AI models in Automotive Aerodynamics

Kaustubh Tangsali
NVIDIA
ktangsali@nvidia.com

Rishikesh Ranade
NVIDIA

Mohammad Amin Nabian
NVIDIA

Alexey Kamenev
NVIDIA

Peter Sharpe
NVIDIA

Neil Ashton
NVIDIA

Ram Cherukuri
NVIDIA

Sanjay Choudhry
NVIDIA

Abstract

In this paper, we introduce a benchmarking framework within the open-source PhysicsNeMo-CFD framework designed to systematically and consistently assess the accuracy, performance, scalability, and generalization capabilities of AI models for automotive aerodynamics predictions. The open extensible framework enables a diverse set of metrics relevant to the Computer-Aided Engineering (CAE) community. By providing a standardized methodology for comparing AI models, the framework enhances transparency and consistency in performance assessment, with the overarching goal of improving the understanding and development of these models to accelerate research and innovation in the field. To demonstrate its utility, the framework includes evaluation of both surface and volumetric flow field predictions on three AI surrogate models — DoMINO, X-MeshGraphNet, and FIGConvNet — using the DrivAerML dataset. It also includes guidelines for integrating additional models and datasets, making it extensible for physically consistent metrics. This benchmarking study aims to enable researchers and industry professionals in selecting, refining, and advancing AI-driven aerodynamic modeling approaches, ultimately fostering the development of more efficient, accurate, and interpretable solutions in automotive aerodynamics.

1 Introduction

Computational Fluid Dynamics (CFD) is essential in automotive aerodynamics, aiding in vehicle optimization alongside wind-tunnel testing. The introduction of the Worldwide Harmonised Light vehicles Test Procedure (WLTP)[1] and increased global competition, especially with the shift to electric vehicles, has heightened the need for CFD. The computational cost and accuracy of CFD simulations vary with turbulence modeling (e.g., RANS, WMLES, HRLES), car modeling choices, and available HPC resources. Wall-Modeled Large Eddy Simulations (WMLES) and hybrid RANS/LES methods (HRLES) [11] offer more accurate flow physics but are slower compared to RANS. GPU-based solvers have reduced the cost of HRLES/WMLES but still require substantial resources.

The interest in AI surrogate models in automotive aerodynamics has grown due to the need for faster design iterations. AI models provide quick feedback, significantly speeding up the design process. Final designs are still validated with traditional CFD solvers, but this approach minimizes their need during the initial development.

AI has been applied to various tasks, including direct prediction of drag coefficients [9, 16, 17], inferring drag from predicted surface fields [10, 2, 8, 7], and reconstructing volume flow fields

[1] https://www.wltpfacts.eu/what-is-wltp-how-will-it-work/

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

[14, 12, 6, 13]. These methods map geometric representations to aerodynamic quantities. Other strategies include generative adversarial networks for flow field super-resolution [18] and AI for shape optimization and design feedback [1]. These studies use datasets like DrivAerML [4], DrivAerNet [9], DrivAerNet++ [10], WindsorML [2], AhmedML [3], ShapeNet [5], and proprietary data, employing architectures from graph neural networks to generative models. High-quality datasets and physics-informed loss functions have been crucial for advancing this research area.

AI models and methods for automotive aerodynamics are often task-specific, with trade-offs in scalability, accuracy, and generalizability. Inconsistent error metrics make it challenging to compare models and hence, developing consistent performance metrics is crucial for creating scalable, accurate AI models with real-world utility.

We introduce a benchmarking framework for automotive aerodynamics in the open-source PhysicsNeMo-CFD framework, providing a thorough and consistent set of performance metrics to compare AI models from a CFD perspective. The framework reports L_2 errors, surface and volume contour comparisons, and other metrics such as aerodynamic forces regression and design trends to provide a thorough model comparison. Section 2 details the performance metrics. We conclude in Section 3 and suggest future work. The code is available at `https://github.com/NVIDIA/physicsnemo-cfd`.

2 Benchmarking in PhysicsNeMo-CFD

NVIDIA PhysicsNeMo-CFD is a sub-module of the open-source (Apache 2.0) NVIDIA PhysicsNeMo framework that provides the tools needed to integrate pretrained AI models into engineering and CFD workflows. The library is a collection of loosely-coupled workflows around the pretrained AI models for CFD, with abstractions and relevant data structures.

PhysicsNeMo-CFD includes benchmarking capabilities that enables evaluation of a wide variety of AI models developed for prediction of automotive aerodynamics. The objective of this work is:

- Equip researchers and engineers with tools to compare AI models using a consistent basis that is important to the CFD community.
- Provide a mechanism to understand the strengths and weaknesses of these models, thereby encouraging further research and development.

Existing literature predominantly assesses model performance using the coefficient of determination (R^2) for drag prediction and point-wise error metrics averaged over mesh elements within the computational domain. These metrics provide a basic sense of accuracy but can mask important deficiencies in aerodynamic predictions. A comprehensive evaluation should include metrics that capture both global accuracy and specific flow features, as well as quantify uncertainty.

As a result, traditional metrics alone offer limited insight into model interpretability and are insufficient for evaluating the practical applicability of these models as predictive tools in CAE. Furthermore, variations in dataset partitioning, evaluation criteria, and performance metrics introduce inconsistencies that preclude direct comparisons among different modeling approaches. Benchmarking tools in PhysicsNeMo-CFD addresses these limitations by providing a comprehensive suite of tools designed to enable robust, standardized, and reproducible model evaluation.

2.1 Evaluation Metrics

2.1.1 Aerodynamic forces

PhysicsNeMo-CFD computes aerodynamic forces such as drag and lift, along with their non-dimensional coefficients, for analyzing surface and volumetric flow predictions around vehicles. It also includes tools for regression plots and trend analysis to assess model performance.

2.1.2 Field comparisons

Flow field visualization is critical for aerodynamic evaluation, revealing key phenomena like pressure regions, flow separation, and vortical structures. The benchmarking framework supports comparisons

of flow fields in various formats (1D, 2D, 3D, 3D manifold) and allows for statistical averaging and resampling, facilitating analysis across different geometries.

2.1.3 Physics based metrics

AI models may not explicitly enforce physical constraints like mass and momentum conservation. PhysicsNeMo-CFD includes utilities to compute residuals, assessing how well AI models adhere to these conservation laws, which is vital for validating AI-driven aerodynamic predictions.

2.1.4 Confidence Scores and Reliability

AI-augmented CAE applications must offer appropriate reliability or confidence metrics when using AI surrogates. PhysicsNeMo-CFD workflows demonstrates several methods to estimate the confidence of AI surrogate model predictions such as sensitivity to input STL resolution, sensitivity to model bias, and sensitivity to ground truth data distribution.

2.2 Including new models and datasets

2.2.1 New models

PhysicsNeMo-CFD provides workflows for evaluating AI models on the DrivAerML [4][2] dataset for surface and volume predictions. These workflows standardize the evaluation and comparison process by using AI model predictions post-processed into .vtp (surface) and .vtu (volume) formats. Users are responsible for generating these files based on their model's architecture, but guidance is available in the GitHub repository. The .vtp and .vtu files must contain both the true and AI-predicted results on the same points for successful evaluation.

Once the .vtp and .vtu files are generated, the workflows can be run via the command line with the appropriate arguments. Detailed instructions are provided in the GitHub Documentation. A Jupyter notebook version is also available for better understanding the evaluation criteria and mechanics.

To evaluate a new model using PhysicsNeMo-CFD on the DrivAerML dataset, ensure the model is trained according to the train-validation split, evaluate the model on the validation samples, and save the results in .vtp and .vtu files. Then use the PhysicsNeMo-CFD workflows for comparisons.

2.2.2 New datasets

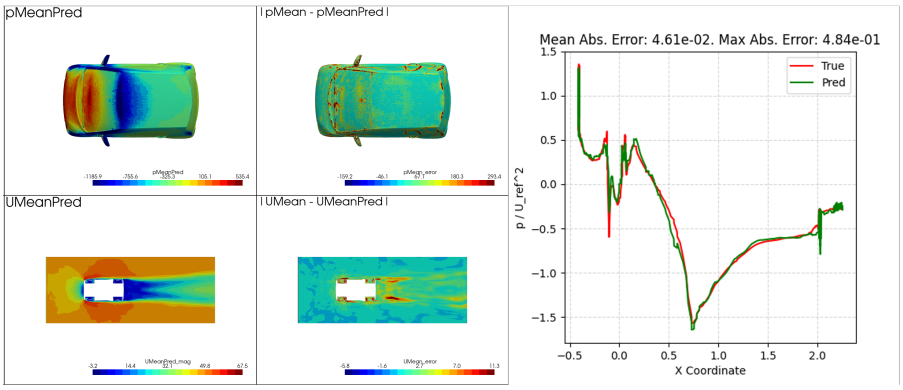
PhysicsNeMo-CFD workflows are customizable for different CFD datasets. The steps to test models on new datasets are similar to those for DrivAerML. Parameters such as variable names and data paths can be configured through a configuration file.

The benchmarking setup has been tested on several other datasets, including DriveSim (an internally generated RANS dataset for passenger cars using OpenFOAM), DriveSim+ (an internally generated RANS dataset for passenger cars with a commercial solver), and datasets involving aircraft aerodynamics (Figure 1). While designed for ML model benchmarking, PhysicsNeMo-CFD workflows can also be used for general CFD model verification and comparison.

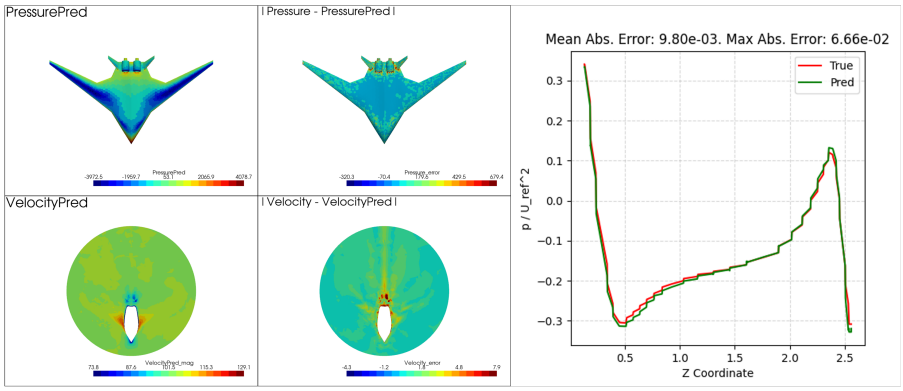
2.3 Experimental setup

In this section, we detail the training and evaluation of the three AI models: DoMINO, X-MeshGraphNet, and FIGConvNet. All models use the same data split for the DrivAerML dataset—436 samples for training and 48 for testing — with consistent IDs across models. X-MeshGraphNet and FIGConvNet are trained to evaluate only surface fields, while DoMINO predicts both surface and volume fields using a coupled approach. For X-MeshGraphNet and FIGConvNet, a uniform point cloud is sampled from the vehicle STL (4.5 million points for X-MeshGraphNet, 500K for FIGConvNet), and simulation results are linearly interpolated onto these points. During evaluation, surface fields predicted on the point clouds are interpolated back onto the simulation mesh for visualization and benchmarking. DoMINO is trained and evaluated directly on the simulation mesh. Model architectures, configurations, and pipelines are published in the PhysicsNeMo GitHub repo. At inference, predictions are written to raw simulation files and used for metric calculations.

[2] https://huggingface.co/datasets/neashton/drivaerml



(a) DoMINO model results trained on DriveSim dataset



(b) DoMINO model results trained on Aircraft dataset

Figure 1: Using PhysicsNeMo-CFD for evaluating AI models trained on new datasets. Figures show surface contours, volume contours and 1D line plots of normalized surface pressure (along vehicle centerline) of DoMINO model output. DoMINO model was trained on two datasets, one on a passenger car dataset (DriveSim) and one on an Aircraft dataset.

Table 1: Relative L2 errors on validation samples for surface predictions (Lower is better)

	X-MeshGraphNet	FIGConvNet	DoMINO
Pressure	0.14	0.21	0.10
Wall Shear Stress (x)	0.17	0.32	0.18
Wall Shear Stress (y)	0.22	0.62	0.26
Wall Shear Stress (z)	0.29	0.53	0.28

2.4 Results and discussions

In this section, we present selected results evaluated using the benchmarking framework for a few key metrics. Due to space constraints, a comprehensive evaluation of all metrics is not included here but will be available in the attached GitHub repository and the final preprint.

2.4.1 Surface results

Table 1 shows the L2 errors evaluated on the cell centers of the mesh while Table 2 and Table 3 provide a quantitative comparison between the design trends of the three models. Generally, the performance of DoMINO and FigConvNet is better than X-MeshGraphNet with DoMINO being slightly better. The reasons for the performance differences can be traced back to the data representation and choice of loss functions during training.

Table 2: Trend analysis (Spearman coefficient) on validation samples (Higher is better)

	X-MeshGraphNet	FIGConvNet	DoMINO
Spearman Coeff. Drag	0.96	0.99	0.99
Spearman Coeff. Lift	0.81	0.98	0.98

Table 3: Trend analysis (Errors) on validation samples (Lower is better)

	X-MeshGraphNet	FIGConvNet	DoMINO
Mean Abs. Error Drag (N)	15.23	8.86	6.64
Max Abs. Error Drag (N)	58.16	25.72	23.08
Mean Abs. Error Lift (N)	63.75	19.00	15.04
Max Abs. Error Lift (N)	187.42	56.90	79.71

2.4.2 Volume results

For volume comparisons, we only provide the results of relative L2 errors for DoMINO in Table 4 (evaluated on the nodes of the mesh). Development of FIGConvNet for volume predictions is currently ongoing. X-MeshGraphNet provides volume prediction capabilities, but due to resource and time constraints, we have not fine-tuned a volume model for X-MeshGraphNet. Hence, for volume comparisons, we only include the results from the DoMINO model. However, similar to the surface benchmarking, the workflows from PhysicsNeMo-CFD can be used for making inter-model comparisons.

3 Conclusion and Future work

In this work, we introduce a benchmarking framework in PhysicsNeMo-CFD that offers consistent CFD-specific metrics for the development and thorough analysis of AI models. This customizable framework is applicable to various external aerodynamics use cases beyond the datasets discussed here. We demonstrate its utility with three AI models: FIGConvNet, X-MeshGraphNet, and DoMINO. The framework validated choices in architecture, data representation, and loss functions, significantly aiding model development.

The landscape of AI Physics is becoming increasingly rich with the availability of standardized datasets and benchmarks, similar to the advancements seen in domains like NLP and vision. Open-source datasets like ERA5 for weather and DrivAerML for aerodynamics, along with benchmarks like WeatherBench [15] and Earth2Studio, have standardized their respective fields. However, CFD lacks such capabilities. PhysicsNeMo-CFD, through its benchmarking capabilities and downstream workflows aims to fill this gap, enabling the AI Physics community to develop accurate and generalizable models.

Future work will add more domain-specific metrics and improve usability for scientific exploration with trained models. This includes features like solver initialization, design sensitivity analysis, and enhancements in physics-based testing and uncertainty quantification. Additionally, we aim to extend PhysicsNeMo-CFD to provide benchmarks for other standardized AI Physics CFD use cases, such as turbulence modeling.

Table 4: Relative L2 errors on validation samples for volume predictions

	DoMINO
X-Velocity	0.1107
Y-Velocity	0.2139
Z-Velocity	0.2302
Pressure	0.1172

References

[1] Nikos Arechiga, Frank Permenter, Binyang Song, and Chenyang Yuan. Drag-guided diffusion models for vehicle image generation. *arXiv preprint arXiv:2306.09935*, 2023.

[2] Neil Ashton, Jordan B Angel, Aditya S Ghate, Gaetan K Kenway, Man L Wong, Cetin Kiris, Astrid Walle, Danielle C Maddix, and Gary Page. Windsorml: High-fidelity computational fluid dynamics dataset for automotive aerodynamics. *Advances in Neural Information Processing Systems*, 37:37823–37835, 2024.

[3] Neil Ashton, Danielle C Maddix, Samuel Gundry, and Parisa M Shabestari. Ahmedml: High-fidelity computational fluid dynamics dataset for incompressible, low-speed bluff body aerodynamics. *arXiv preprint arXiv:2407.20801*, 2024.

[4] Neil Ashton, Charles Mockett, Marian Fuchs, Louis Fliessbach, Hendrik Hetmann, Thilo Knacke, Norbert Schonwald, Vangelis Skaperdas, Grigoris Fotiadis, Astrid Walle, et al. Drivaerml: High-fidelity computational fluid dynamics dataset for road-car external aerodynamics. *arXiv preprint arXiv:2408.11969*, 2024.

[5] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository, 2015. URL `https://arxiv.org/abs/1512.03012`.

[6] Fangge Chen and Kei Akasaka. 3d flow field estimation around a vehicle using convolutional neural networks. In *BMVC*, page 396, 2021.

[7] Chris Choy, Alexey Kamenev, Jean Kossaifi, Max Rietmann, Jan Kautz, and Kamyar Azizzadenesheli. Factorized implicit global convolution for automotive computational fluid dynamics prediction. *arXiv preprint arXiv:2502.04317*, 2025.

[8] PhysicsNeMo Contributors. NVIDIA PhysicsNeMo: An open-source framework for physics-based deep learning in science and engineering, 2023. [Online]. Available: `https://github.com/NVIDIA/physicsnemo`.

[9] Mohamed Elrefaie, Faez Ahmed, and Angela Dai. Drivaernet: A parametric car dataset for data-driven aerodynamic design and graph-based drag prediction. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 88360, page V03AT03A019. American Society of Mechanical Engineers, 2024.

[10] Mohamed Elrefaie, Florin Morar, Angela Dai, and Faez Ahmed. Drivaernet++: A large-scale multimodal car dataset with computational fluid dynamics simulations and deep learning benchmarks. *Advances in Neural Information Processing Systems*, 37:499–536, 2024.

[11] Jochen Fröhlich and Dominic Von Terzi. Hybrid les/rans methods for the simulation of turbulent flows. *Progress in Aerospace Sciences*, 44(5):349–377, 2008.

[12] Sam Jacob Jacob, Markus Mrosek, Carsten Othmer, and Harald Köstler. Deep learning for real-time aerodynamic evaluations of arbitrary vehicle shapes. *arXiv preprint arXiv:2108.05798*, 2021.

[13] Mohammad Amin Nabian, Chang Liu, Rishikesh Ranade, and Sanjay Choudhry. X-meshgraphnet: Scalable multi-scale graph neural networks for physics simulation. *arXiv preprint arXiv:2411.17164*, 2024.

[14] Rishikesh Ranade, Mohammad Amin Nabian, Kaustubh Tangsali, Alexey Kamenev, Oliver Hennigh, Ram Cherukuri, and Sanjay Choudhry. Domino: A decomposable multi-scale iterative neural operator for modeling large scale engineering simulations. *arXiv preprint arXiv:2501.13350*, 2025.

[15] Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russel, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. Weatherbench 2: A benchmark for the next generation of data-driven global weather models, 2023.

[16] Binyang Song, Chenyang Yuan, Frank Permenter, Nikos Arechiga, and Faez Ahmed. Surrogate modeling of car drag coefficient with depth and normal renderings. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 87301, page V03AT03A029. American Society of Mechanical Engineers, 2023.

[17] Jonathan Tran, Kai Fukami, Kenta Inada, Daisuke Umehara, Yoshimichi Ono, Kenta Ogawa, and Kunihiko Taira. Aerodynamics-guided machine learning for design optimization of electric vehicles. *Communications Engineering*, 3(1):174, 2024.

[18] Thanh Luan Trinh, Fangge Chen, Takuya Nanri, and Kei Akasaka. 3d super-resolution model for vehicle flow field enrichment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5826–5835, 2024.