Towards Multi-Fidelity Scaling Laws of Neural Surrogates in CFD

Paul Setinek [1] **Gianluca Galletti** [1] **Johannes Brandstetter**[1,2]

[1] ELLIS Unit, LIT AI Lab, Institute for Machine Learning, JKU Linz, Austria
[2] Emmi AI, Linz, Austria
setinek@ml.jku.at

Abstract

Scaling laws describe how model performance grows with data, parameters and compute. While large datasets can usually be collected at relatively low cost in domains such as language or vision, scientific machine learning is often limited by the high expense of generating training data through numerical simulations. However, by adjusting modeling assumptions and approximations, simulation fidelity can be traded for computational cost, an aspect absent in other domains. We investigate this trade-off between data fidelity and cost in neural surrogates using low- and high-fidelity Reynolds-Averaged Navier-Stokes (RANS) simulations. Reformulating classical scaling laws, we decompose the dataset axis into compute budget and dataset composition. Our experiments reveal compute-performance scaling behavior and exhibit budget-dependent optimal fidelity mixes for the given dataset configuration. These findings provide the first study of empirical scaling laws for multi-fidelity neural surrogate datasets and offer practical considerations for compute-efficient dataset generation in scientific machine learning.

1 Introduction

Machine learning has seen immense progress in recent years, which was not only driven by architectural or methodological innovations but also by the increasing availability of computational resources. This has enabled the scaling up of models, and as a result many SOTA models now contain hundreds of billions of parameters [38, 2]. Scaling laws, which originated in the domain of Large Language Models (LLMs) [20, 18], have expanded into various other areas like Computer Vision (CV) [46] or time series [35, 44]. These empirical studies describe how models improve as a function of *three axes*: (i) parameter count (N), (ii) dataset size (D), and (iii) compute (C).

In the meantime, scientific machine learning has similarly achieved remarkable success in modeling complex systems with neural surrogates. Notable examples include breakthroughs in weather and climate forecasting [21, 31, 29, 32, 8], material design [27, 45, 41] or protein folding [19, 1]. These advances have partly been enabled by large curated public datasets, such as ERA5 in weather modeling [17] or the Protein Data Bank (PDB) [6]. More recently, first large scale datasets have also been released in areas such as automotive aerodynamics [4, 14].

However, in many areas of science and engineering, such datasets are either not public or do not exist and therefore researchers are required to generate their own problem specific dataset prior to model training. Since the systems of interest are usually governed by Partial Differential Equations (PDEs) [15], generating training data requires solving these equations numerically, which is often coupled with significant computational costs [43]. This means that unlike other domains, where data can be sourced with little to no computation (*e.g.* text, real-world images and videos), the dataset size D in scientific machine learning is often no longer "free" to scale. While prior works study how

Accepted at the AI for Science workshop (NeurIPS 2025).

performance scales with dataset size, they do not take the computational cost associated with scaling the dataset axis into account [37, 16, 3, 30]. Moreover, many neural surrogates exist with the goal of "amortizing" the training and dataset cost with repeated, cheap evaluations down the line; if dataset generation becomes prohibitively expensive, it fundamentally defeats the purpose of such models.

The numerical solution of PDEs is a long-running research topic, which can be very nuanced: when designing numerical simulations, modeling assumptions and simplifications of the underlying physics need to be made, which trade fidelity for compute. Ordered by computational complexity, common approaches include Direct Numerical Simulation (DNS) which resolves all turbulent scales at prohibitive cost; Large Eddy Simulation (LES), which resolves only large scales while modeling subgrid-scale dynamics; and Reynolds-Averaged Navier-Stokes (RANS), where turbulence is entirely modeled through statistical averaging, making RANS the cheapest but also least accurate of the three. For example, a 3D LES over an airfoil can take several orders of magnitude longer than simulating the respective RANS equations. Even on simple scenarios, LES take an order of magnitude longer than RANS simulations [23], a factor which grows substantially for more chaotic systems.

This introduces a fundamental trade-off, which motivates our research question:

> *Under a fixed budget constraint, what is the optimal training set composition of low- and high-fidelity samples in order to maximize model performance?*

We move towards an answer to this essential question by proposing a reformulation of classical scaling laws, devised to account for this phenomenon. We argue for splitting the *dataset size axis* D into two components, namely *dataset compute budget* D_b (core hours allocated for data generation) and *dataset composition* D_c (controls the fidelity distribution of the training data).

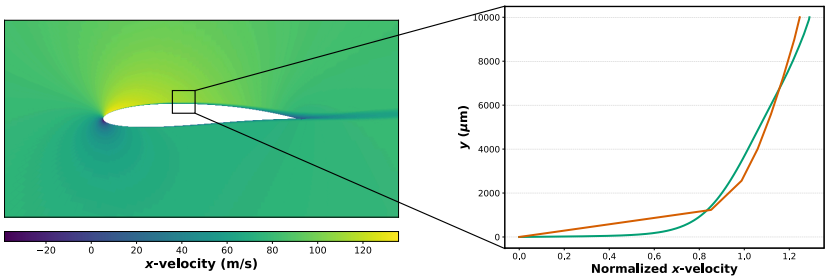
Since training data generation is typically the primary bottleneck in scientific machine learning, we focus on these two axes while assuming model size and compute for model training are non-limiting. Although restrictive, this assumption allows us to directly address our research question that has not been studied in prior work. Our contributions can be summarized as follows:

- We introduce a formulation of multi-fidelity scaling laws, extending classical scaling law analysis to settings where training data can be simulated at different fidelities.
- We design a multi-fidelity dataset of external aerodynamics around airfoils, incorporating different modeling assumptions across fidelity levels.
- We present the first empirical investigation of multi-fidelity scaling laws along the data axis on a Computational Fluid Dynamics (CFD) dataset, evaluating how model performance scales under varying dataset compositions and generation budgets.

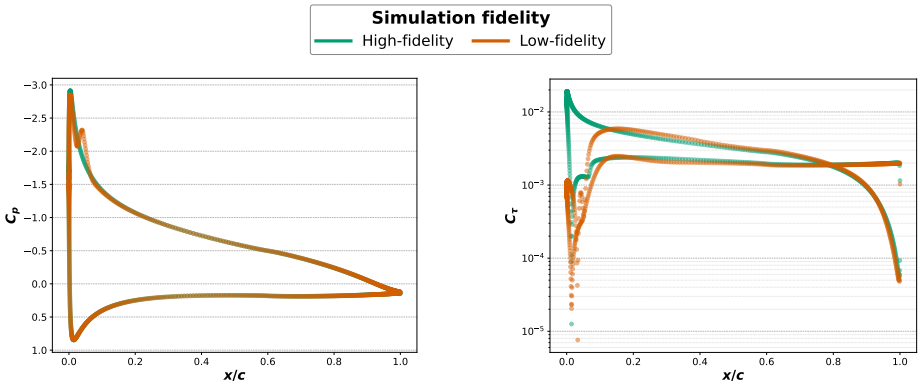
2 Related work

Learning from multi-fidelity data. Multi-fidelity data exists in different fields, and can come in different shapes and forms (e.g., varying realism, accuracy or resolution). In CV, *"resolution transfer"* [7, 34] or *"super-resolution"* [13, 28] are well researched directions, where models learn to predict fine-scale details from coarse observations. Similar ideas appear in scientific machine learning under *"discretization convergence"* in neural operators [22], where models are trained to generalize across mesh resolutions. While these methods may be invariant to changes in resolution (even though most of them have no theoretical guarantee [5]), they do not capture underlying physics and fidelity shifts.

Transfer learning for multi-fidelity data in scientific machine learning. Recent studies have explored transfer learning from low- to high-fidelity data [12, 36, 25]. However, in these works the distinction between fidelities is limited to changes in mesh resolution, while the underlying physical modeling assumptions remain the same. In contrast, a more recent study takes this further by transferring knowledge from low-fidelity RANS to high-fidelity LES simulations in the context of wind farm modeling, thereby altering the underlying modeling assumptions [47]. While these studies address an important aspect of scientific machine learning, our investigation pursues a complementary goal: we focus on identifying patterns suggesting the existence of optimal strategies for generating training data to maximize the performance of neural surrogates.



(a) Velocity field around the airfoil (left) and boundary layer detail (right).



(b) Pressure Distribution along the chord.

(c) Skin friction coefficient.

Figure 1: Visual comparison of a low- and high-fidelity simulation of a NACA4 airfoil with parameters (`M=2.408, P=5.987, XX=11.876`), at an Angle of Attack (AoA) of $7.57°$ with an inlet velocity of $81.645\,\mathrm{m\,s^{-1}}$. Figure 1a (left) shows the x-velocity of the high-fidelity simulation, and (right) zooms on the corresponding boundary layer profile at mid-chord ($0.5c$), highlighting the difference in modeling resolution between low- and high-fidelity. Bottom plots display the evolution of the pressure coefficient C_p (Figure 1b) and skin friction coefficient C_τ (Figure 1c), along the chord line x/c.

3 Dataset

The numerical solution of PDEs depends on assumptions and choices made while designing the simulation pipeline, aimed at balancing accuracy with computational feasibility. In CFD for external aerodynamics the goal is to solve the Navier-Stokes equations for the flow around rigid bodies. Given this problem setting, the following design choices can be made when simulating the system:

1. **Problem definition:** Define the high-level description of the system. For example, is the flow laminar or turbulent, compressible or incompressible, subsonic, transonic or supersonic. This also includes whether the problem should be solved in two or three dimensions and whether transient solutions are required or steady-state averages are sufficient.
2. **Physics modeling assumptions:** Choose the appropriate technique (e.g., LES, RANS, or hybrid methods). If turbulence is present, select a closure (e.g., one-equation models, two-equation models, etc.) and pick boundary layer treatment (fully resolved, wall functions).
3. **Initial and boundary conditions:** Set inflow, outlet and surface boundary conditions, as well as initial conditions of the system.
4. **Meshing:** Generate a mesh fine enough to support the modeling assumptions and simplifications defined in previous steps.
5. **Solver settings:** Select discretization schemes, time-stepping methods, relaxation factors and convergence criteria.

Table 1: Low- and high-fidelity modeling assumptions and resulting dataset characteristics.

Modeling Assumptions				Dataset Specifications		
Fidelity	Viscous sublayer	First cell height (μm)	First cell center (y^+)	Avg sim time (core hrs)	Avg number of nodes	Total size (GB)
High	Resolving	2	< 1	13.4	180K	18
Low	Modeling	1,200	30–300	4.8	96K	7.8

To study scaling, we select a dataset with the following criteria: (i) the problem setup should be realistic and not purely academic, (ii) the low- and high-fidelity datasets should differ in their physical modeling assumptions, not simply in mesh resolution, and (iii) the computational cost of the high-fidelity simulations should be noticeably larger than the low-fidelity simulations.

We identify aerodynamic airfoil simulations as an ideal testbed, since they are industrially relevant, well studied, and allow for different fidelity levels based on physical modeling assumptions. Due to the prohibitive cost of DNS or LES simulations for large dataset creation [10, 42], we base our study on simulating RANS equations.

In CFD and especially external aerodynamics, the boundary layer, i.e. the thin region of fluid close to the solid's surface, is of utmost importance. It is common to use a dimensionless wall distance y^+ (pronounced "y-plus") to describe the distance to the surface. The region where $y^+ < 5$ corresponds to the viscous sublayer. This layer is characterized by strong velocity gradients, and its accurate prediction is critical since key aerodynamic quantities, such as drag and lift, depend on these gradients. To create two distinct fidelity levels, we vary the boundary layer treatment. Our high-fidelity simulations *fully resolve* this region by ensuring that the first computational mesh cell center has $y^+ < 1$, leading to accurate but costly predictions. On the other hand, the low-fidelity setup uses a coarser mesh near the wall such that y^+ lies between 30 and 300. This allows for a wall function approach, in which the region close to the wall is not resolved directly but *modeled analytically* using the empirical law-of-the-wall derived from experimental data [33]. This modeling choice is widely used in RANS and wall-modeled LES (WMLES) applications.

We base our high-fidelity simulation setup on the AirfRANS dataset configuration [9]. AirfRANS models airfoils from the NASA's 4- and 5-digit series [11] in an incompressible regime (Mach number < 0.3), covering Reynolds numbers from 2×10^6 to 6×10^6, and AoAs between -5° and 15°. We run all simulations in OpenFOAM [39], using the `simpleFOAM` solver, with the k–ω SST turbulence model [26] as equation closure (a standard approach in airfoil aerodynamics). To highlight the important aspects of the resulting datasets, Table 1 summarizes the differences between high- and low-fidelity. All numerical simulations were run on an AMD EPYC9655P 96-Core CPU (192 threads, 4.5GHz and 2.2 TiB of RAM). We use OpenFOAMv2506 and Open MPI 4.1.1 for parallel execution.

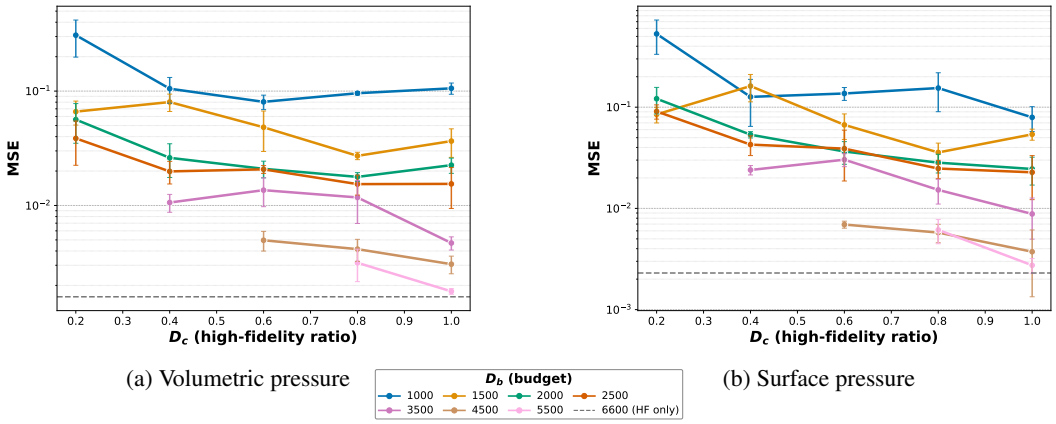
The final datasets consist of 611 matched pairs of low- and high-fidelity simulations (491 train/val, 120 test). The difference in dataset size compared to the original AirfRANS dataset is caused by the simplifications needed for the low-fidelity simulations, which can sometimes lead to poor convergence. Figure 1 illustrates the difference between the two fidelities in terms of boundary layer profile and the distribution of key physical quantities along the chord for a chosen dataset sample.

4 Experiments

Our experiments are designed to investigate *compute-optimal* model training in the setting where the available budget D_b for data generation is the primary constraint. Model size is fixed across all experiments, and our analysis focuses on the optimal composition of low- and high-fidelity data. For fixed D_bs (in core hours) we vary the ratio of low- to high-fidelity samples D_c in the training set. This is done by first estimating the number of datapoints based on the average cost and the desired fidelity distribution, then sampling random low-/high-fidelity simulations until D_b is matched, and finally applying an optional greedy repair step to ensure the final selection satisfies the budget constraint.

The task at hand is to predict the solution of the RANS simulation given the initial conditions and mesh node positions. For all nodes, we predict five quantities: velocity v in x- and y-direction, pressure p, and wall shear stress τ in x- and y-direction. We use Transolver [40], a SOTA transformer-based neural operator, with $\sim$4M parameters. Detailed model and training hyperparameters are provided in Appendix A.

The same architecture is trained on different datasets, generated at different budgets D_b and with varying ratios of low- and high-fidelity simulations D_c. Performance is evaluated on 120 unseen high-fidelity samples. Figure 2 and Figure 3 show the results of how model performance behaves with increasing dataset generation budgets and varying dataset compositions. We discuss our main findings below.



(a) Volumetric pressure

(b) Surface pressure

(c) Volumetric x-velocity

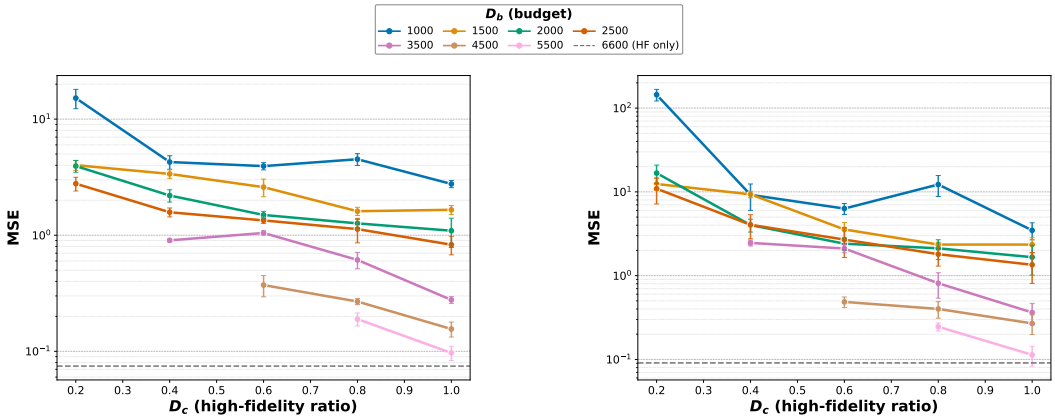
(d) Volumetric y-velocity

Figure 2: Scaling behavior for fields with positive transfer. We show the Mean Squared Error (MSE) of normalized fields averaged over four seeds with error bars indicating standard deviation. Lines of different colors show different dataset generation budgets in compute hours D_b, at growing percentage of high-fidelity ratio D_c. The dashed line indicates model performance when trained on the full high-fidelity dataset.

Compute Budget Scaling Law. Across all dataset compositions, we observe a clear trend that model error decreases with an increasing compute budget for dataset generation (Figures 2 and 3). This confirms that the used budget for training data simulations directly links to surrogate accuracy, analogous to scaling laws observed in model, data and compute size in other domains [20, 18].

Knowledge Transfer from Low- to High-Fidelity. For lower compute budgets, certain fields show signs of positive transfer from low- to high-fidelity samples. This behavior is visible for the pressure field in the volume (Figure 2a) and on the surface (Figure 2b) as well as the volumetric velocity field (Figures 2c and 2d). When the available dataset generation budget is limited, allocating all resources to high-fidelity samples does not lead to optimal test performance. Instead, models trained on a mixture of low- and high-fidelity data achieve better accuracy. This suggests that, *under tight compute constraints, the broader coverage of the data manifold offered by many low-fidelity samples*

outweighs the higher accuracy of a few high-fidelity ones. In general, the smaller the available budget, the more the optimal dataset composition shifts towards allocating more budget to lower fidelity samples. Above certain budgets, model performance continuously improves with more budget allocated towards high-fidelity samples, showing that beyond a certain budget threshold, model accuracy becomes primarily limited by the fidelity of the data rather than its quantity.



(a) Wall shear stress (x-component)

(b) Wall shear stress (y-component)

Figure 3: Scaling behavior for fields without positive transfer. We show the MSE of normalized fields averaged over four seeds with error bars indicating standard deviation. Lines of different colors show different training budgets in compute hours D_b, at growing percentage of high-fidelity composition D_c. The dashed line indicates model performance when trained on the full high-fidelity dataset.

Contrary to these trends, *we cannot observe any positive transfer from low- to high-fidelity samples for the wall shear stress.* Figures 3a and 3b show that for these quantities model performance consistently improves across all budgets when more dataset generation budget is allocated towards high-fidelity samples.

Physical Explanation. We hypothesize that the observed results can be linked to the discrepancies between the low- and high-fidelity simulations. The primary distinction lies in the treatment of the viscous sublayer at the airfoil surface: the low-fidelity setup models this region using relatively coarse meshing, whereas the high-fidelity simulation fully resolves it with a fine mesh. As a result, velocity and pressure fields remain largely consistent across fidelities, while the wall shear stress, which is highly sensitive to the boundary layer resolution, shows substantial deviations.

Table 2: nMAE (↓) per field between low- and high-fidelity simulations.

Field	Surface	Volume
(x-)Velocity	–	0.118
(y-)Velocity	–	0.303
Pressure	**0.043**	**0.040**
(x-)WSS	0.405	–
(y-)WSS	0.796	–

This explains why no positive transfer can be observed for this quantity even at small dataset generation budgets: the difference between the two simulations is simply too large. Table 2 quantifies these discrepancies per field by reporting the normalized Mean Absolute Error (nMAE) (Appendix B) of low-fidelity fields interpolated onto the corresponding high-fidelity mesh relative to their high-fidelity counterparts. It clearly shows a difference in nMAE for the pressure field compared to the two components of wall shear stress, aligning with the different multi-fidelity scaling behaviors shown in Figures 2a and 2b compared to Figures 3a and 3b. This also aligns with the visual comparison of the difference in pressure coefficients C_p and skin friction coefficients C_τ along the chord (see Figures 1b and 1c).

5 Conclusion and Future Work

Our work serves as an initial step towards understanding *scaling laws for neural surrogates trained on multi-fidelity data*, highlighting both the potential of optimal dataset budget allocation and the limitations arising when the fidelity gap between simulations becomes too large. Given our findings, we identify several promising directions.

Different simulation methods. Our experiments are currently limited to RANS simulations where fidelity is varied via boundary layer treatment. Exploring additional simulation methods as fidelities, such as time-averaged LES or hybrid RANS-LES approaches, could provide deeper insights into realistic multi-fidelity dataset design, albeit at increased computational cost.

Continuous fidelities. Our results support the development of continuous fidelity formulations rather than discrete fidelity levels. This is both more realistic, since every simulation inherently allows continuous mesh scaling, and potentially more effective, as it can mitigate situations where fidelity levels are too far apart for meaningful knowledge transfer.

Generalization of the framework. Extending our multi-fidelity scaling analysis to other scientific domains, such as thermomechanics, electromagnetics, or molecular dynamics, could reveal whether the identified scaling behaviors generalize across different physical systems. Additionally, while our study focuses on dataset generation cost and composition, future work should also explore scaling the remaining axes, namely model size and training compute in order to eventually establish a more complete formulation of multi-fidelity scientific scaling laws.

Acknowledgments

The authors thank Fabian Paischer for the insightful conversations and the constructive feedback, and Léo Cotteleer for the discussions about the numerical simulations.

The ELLIS Unit Linz, the LIT AI Lab, the Institute for Machine Learning, are supported by the Federal State Upper Austria. We thank the projects FWF AIRI FG 9- N (10.55776/FG9), AI4GreenHeatingGrids (FFG- 899943), Stars4Waters (HORIZON-CL6-2021- CLIMATE-01-01), FWF Bilateral Artificial Intelligence (10.55776/COE12). We acknowledge EuroHPC Joint Undertaking for awarding us access to Leonardo at CINECA, Italy, and MareNostrum5 at BSC, Spain.

References

[1] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

[2] Mistral AI. Mistral large 2 release announcement. `https://mistral.ai/news/mistral-large-2407`, 2025.

[3] Benedikt Alkin, Andreas Fürst, Simon Schmid, Lukas Gruber, Markus Holzleitner, and Johannes Brandstetter. Universal physics transformers: A framework for efficiently scaling neural operators. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL `http://papers.nips.cc/paper_files/paper/2024/hash/2cd36d327f33d47b372d4711edd08de0-Abstract-Conference.html`.

[4] Neil Ashton, Charles Mockett, Marian Fuchs, Louis Fliessbach, Hendrik Hetmann, Thilo Knacke, Norbert Schonwald, Vangelis Skaperdas, Grigoris Fotiadis, Astrid Walle, Burkhard Hupertz, and Danielle C. Maddix. Drivaerml: High-fidelity computational fluid dynamics dataset for road-car external aerodynamics. *CoRR*, abs/2408.11969, 2024. doi: 10.48550/ARXIV.2408.11969. URL `https://doi.org/10.48550/arXiv.2408.11969`.

[5] Francesca Bartolucci, Emmanuel de Bézenac, Bogdan Raoni, Roberto Molinaro, Siddhartha Mishra, and Rima Alaifari. Representation equivalent neural operators: a framework for alias-free operator learning, 2023. URL `https://arxiv.org/abs/2305.19913`.

[6] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235.

[7] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes, 2023. URL `https://arxiv.org/abs/2212.08013`.

[8] Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the earth system. *Nat.*, 641 (8065):1180–1187, 2025. doi: 10.1038/S41586-025-09005-Y. URL `https://doi.org/10.1038/s41586-025-09005-y`.

[9] Florent Bonnet, Jocelyn Ahmed Mazari, Paola Cinnella, and Patrick Gallinari. Airfrans: High fidelity computational fluid dynamics dataset for approximating reynolds-averaged navier-stokes solutions. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL `http://papers.nips.cc/paper_files/paper/2022/hash/94ab7b23a345f93333eac8748a66c763-Abstract-Datasets_and_Benchmarks.html`.

[10] Haecheon Choi and Parviz Moin. Grid-point requirements for large eddy simulation: Chapmans estimates revisited. *Physics of Fluids*, 24, 01 2012. doi: 10.1063/1.3676783.

[11] Russell M. Cummings, William H. Mason, Scott A. Morton, and David R. McDaniel. Applied computational aerodynamics: A modern engineering approach. In *Cambridge Aerospace Series*, pp. 731–765. Cambridge University Press, 2015. doi: 10.1017/CBO9781107284166.

[12] Subhayan De, Jolene Britton, Matthew J. Reynolds, Ryan Skinner, Kenneth E. Jansen, and Alireza Doostan. On transfer learning of neural networks using bi-fidelity data for uncertainty propagation. *CoRR*, abs/2002.04495, 2020. URL `https://arxiv.org/abs/2002.04495`.

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks, 2015. URL `https://arxiv.org/abs/1501.00092`.

[14] 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. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 499–536. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/013cf29a9e68e4411d0593040a8a1eb3-Paper-Datasets_and_Benchmarks_Track.pdf.

[15] Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 2nd edition, 2010.

[16] Maximilian Herde, Bogdan Raonic, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for pdes. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/84e1b1ec17bb11c57234e96433022a9a-Abstract-Conference.html.

[17] Hans Hersbach, Bill Bell, Paul Berrisford, Guido Biavati, András Horányi, Joaquín Muñoz Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Iryna Rozum, Dinand Schepers, Adrian Simmons, Ctlin Soci, Dick Dee, and Jean-Noël Thépaut. ERA5 hourly data on single levels from 1940 to present, 2023. URL https://doi.org/10.24381/cds.adbb2d47. Accessed on DD-MMM-YYYY.

[18] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022. doi: 10.48550/ARXIV.2203.15556. URL https://doi.org/10.48550/arXiv.2203.15556.

[19] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[20] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

[21] Ryan Keisler. Forecasting global weather with graph neural networks, 2022. URL https://arxiv.org/abs/2202.07575.

[22] Nikola B. Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces. *CoRR*, abs/2108.08481, 2021.

[23] D. Lopes, H. Puga, J. Teixeira, R. Lima, J. Grilo, J. Dueñas-Pamplona, and C. Ferrera. Comparison of rans and les turbulent flow models in a real stenosis. *International Journal of Heat and Fluid Flow*, 107:109340, 2024. ISSN 0142-727X. doi: https://doi.org/10.1016/j.ijheatfluidflow.2024.109340. URL https://www.sciencedirect.com/science/article/pii/S0142727X24000651.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[25] Lu Lu, Raphaël Pestourie, Steven G. Johnson, and Giuseppe Romano. Multifidelity deep neural operators for efficient learning of partial differential equations with application to fast inverse design of nanoscale heat transport. *CoRR*, abs/2204.06684, 2022. doi: 10.48550/ARXIV.2204.06684. URL https://doi.org/10.48550/arXiv.2204.06684.

[26] Florian Menter, M. Kuntz, and RB Langtry. Ten years of industrial experience with the sst turbulence model. *Heat and Mass Transfer*, 4, 01 2003.

[27] Amil Merchant, Simon L. Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nat.*, 624(7990):80–85, 2023. doi: 10.1038/S41586-023-06735-9.

[28] Brian B. Moser, Arundhati S. Shanbhag, Federico Raue, Stanislav Frolov, Sebastian Palacio, and Andreas Dengel. Diffusion models, image super-resolution, and everything: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7):1179311813, July 2025. ISSN 2162-2388. doi: 10.1109/tnnls.2024.3476671. URL http://dx.doi.org/10.1109/TNNLS.2024.3476671.

[29] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *CoRR*, abs/2301.10343, 2023. doi: 10.48550/ARXIV.2301.10343. URL https://doi.org/10.48550/arXiv.2301.10343.

[30] Fabian Paischer, Gianluca Galletti, William Hornsby, Paul Setinek, Lorenzo Zanisi, Naomi Carey, Stanislas Pamela, and Johannes Brandstetter. Gyroswin: 5d surrogates for gyrokinetic plasma turbulence simulations, 2025. URL https://arxiv.org/abs/2510.07314.

[31] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *CoRR*, abs/2202.11214, 2022. URL https://arxiv.org/abs/2202.11214.

[32] Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter W. Battaglia, Rémi R. Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nat.*, 637 (8044):84–90, 2025. doi: 10.1038/S41586-024-08252-9. URL https://doi.org/10.1038/s41586-024-08252-9.

[33] Hermann Schlichting and Klaus Gersten. *Boundary-Layer Theory*. Springer, Berlin, Heidelberg, 9 edition, 2016. doi: 10.1007/978-3-662-52919-5.

[34] Johannes Schusterbauer, Ming Gui, Pingchuan Ma, Nick Stracke, Stefan A. Baumann, Vincent Tao Hu, and Björn Ommer. Boosting latent diffusion with flow matching. In *ECCV*, 2024.

[35] Jingzhe Shi, Qinwei Ma, Huan Ma, and Lei Li. Scaling law for time series forecasting. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/97c2f0fac182353062d304d0322ae285-Abstract-Conference.html.

[36] Dong H. Song and Daniel M. Tartakovsky. Transfer learning on multi-fidelity data. *CoRR*, abs/2105.00856, 2021. URL https://arxiv.org/abs/2105.00856.

[37] Shashank Subramanian, Peter Harrington, Kurt Keutzer, Wahid Bhimji, Dmitriy Morozov, Michael W. Mahoney, and Amir Gholami. Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/e15790966a4a9d85d688635c88ee6d8a-Abstract-Conference.html.

[38] LLaMA-3 Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[39] H. G. Weller, G. Tabor, H. Jasak, and C. Fureby. A tensorial approach to computational continuum mechanics using object-oriented techniques. *Computers in Physics*, 12(6):620–631, 1998. doi: 10.1063/1.168744.

[40] Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for pdes on general geometries. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Ywl6pODXjB.

[41] Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, Matthew Horton, Robert Pinsler, Andrew Fowler, Daniel Zügner, Tian Xie, Jake Smith, Lixin Sun, Qian Wang, Lingyu Kong, Chang Liu, Hongxia Hao, and Ziheng Lu. Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024.

[42] Xiang I. A. Yang and Kevin P. Griffin. Grid-point and time-step requirements for direct numerical simulation and large-eddy simulation. *Physics of Fluids*, 33(1), January 2021. ISSN 1089-7666. doi: 10.1063/5.0036515. URL http://dx.doi.org/10.1063/5.0036515.

[43] Xin-She Yang, Slawomir Koziel, and Leifur Leifsson. Computational optimization, modelling and simulation: Recent trends and challenges. *Procedia Computer Science*, 18:855–860, 2013. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2013.05.250. URL https://www.sciencedirect.com/science/article/pii/S1877050913003931. 2013 International Conference on Computational Science.

[44] Qingren Yao, Chao-Han Huck Yang, Renhe Jiang, Yuxuan Liang, Ming Jin, and Shirui Pan. Towards neural scaling laws for time series foundation models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=uCqxDfLYrB.

[45] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, et al. A generative model for inorganic materials design. *Nature*, pp. 1–3, 2025.

[46] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 1204–1213. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01179. URL https://doi.org/10.1109/CVPR52688.2022.01179.

[47] Dichang Zhang, Zexia Zhang, Christian Santoni, Ali Khosronejad, and Dimitris Samaras. Transfer learning in multi-fidelity surrogate modeling: A wind farm case. In *ICML 2024 AI for Science Workshop*, 2024. URL https://openreview.net/forum?id=yBTDCqNcan.

Appendix

A Training details

We train our Transolver [40] models using AdamW [24] with a weight decay of 1×10^{-4}, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for 500 epochs, applying early stopping if there is no improvement in validation loss for 250 consecutive epochs. We employ a cosine decay learning rate scheduler with a 10 epoch linear warmup to an initial learning rate of 5×10^{-4}. We use gradient clipping and train in single precision float-32. We list the exact hyperparameter choices contributing to the total model size of $\sim$4M params in Table 3.

Table 3: Transolver hyperparameters used.

Hyperparameter	**Value**
Base dimension	256
# Attention heads	4
# Transformer layers	8
Slice base	128
MLP expansion ratio	2
Dropout (MLPs/projections)	0.1
Dropout (Attention)	0.1

B Normalized Mean Absolute Error

We define the nMAE as

$$\mathrm{nMAE} = \frac{\sum_{i=1}^{N} \left| \hat{y}_i^{\mathrm{LF}} - y_i^{\mathrm{HF}} \right|}{\sum_{i=1}^{N} \left| y_i^{\mathrm{HF}} \right|},$$

where $\hat{\boldsymbol{y}}_i^{\mathrm{LF}}$ are the fields from the low-fidelity simulation interpolated onto the corresponding high-fidelity mesh (nearest neighbor) and N is the number of mesh points of the high-fidelity sample. In our comparison, we report the average nMAE over all test samples.