5D NEURAL SURROGATES FOR NONLINEAR GYROKINETIC SIMULATIONS OF PLASMA TURBULENCE

Gianluca Galletti[1], **Fabian Paischer**[1], **Paul Setinek**[1], **William Hornsby**[2], **Lorenzo Zanisi**[2], **Naomi Carey**[2] **Stanislas Pamela**[2] **Johannes Brandstetter**[1,3]
[1] ELLIS Unit, LIT AI Lab, Institute for Machine Learning, JKU Linz, Austria
[2] UKAEA (United Kingdom Atomic Energy Authority), Culham Campus, Abingdon
[3] Emmi AI GmbH, Linz, Austria
paischer@ml.jku.at

ABSTRACT

Nuclear fusion plays a pivotal role in the quest for reliable and sustainable energy production. A major roadblock to achieving commercially viable fusion power is understanding plasma turbulence, which can significantly degrade plasma confinement. Modelling turbulence is crucial to design performing plasma scenarios for next-generation reactor-class devices and current experimental machines. The nonlinear gyrokinetic equation underpinning turbulence modelling evolves a 5D distribution function over time. Solving this equation numerically is extremely expensive, requiring up to weeks for a single run to converge, making it unfeasible for iterative optimisation and control studies. In this work, we propose a method for training neural surrogates for 5D gyrokinetic simulations. Our method extends a hierarchical vision transformer to five dimensions and is trained on the 5D distribution function for the adiabatic electron approximation. We demonstrate that our model can accurately infer downstream physical quantities such as heat flux time trace and electrostatic potentials for single-step predictions two orders of magnitude faster than numerical codes. Our work paves the way towards neural surrogates for plasma turbulence simulations to accelerate deployment of commercial energy production via nuclear fusion.

1 INTRODUCTION

Turbulence is a key driver of plasma confinement degradation, as it causes Plasma to diffuse towards the reactor wall, resulting in most of the heat and particle transport in magnetic fusion devices, such as Tokamaks. The growing turbulence is dampened by the zonal flow system, which regulates turbulence to reach a quasi-stationary saturated state (Itoh et al., 2006). The design and control of performing plasma scenarios strictly requires knowledge of the turbulent transport in the saturated state. This can be obtained via complex nonlinear gyrokinetic simulations, which involve evolution of a 5D distribution function over time.

Computationally affordable reduced-order quasilinear models of turbulent transport, such as QuaLiKiz (Bourdelle et al., 2015; Citrin et al., 2017) and TGLF (Staebler et al., 2007; Staebler & Kinsey, 2010), are routinely adopted. They make simplifying assumptions about the mechanism of turbulence saturation and modality expressed as so-called saturation rules. As different saturation rules are fitted to specific turbulence modalities, a general quasilinear model is not available to date, which limits their usefulness. A fast, general and reliable estimate of turbulent fluxes is only attainable via expensive nonlinear gyrokinetic simulations.

Neural surrogates can reduce the complexity of nonlinear gyrokinetic simulations. Current approaches operate on reduced input spaces (Narita et al., 2022; Honda et al., 2023), which do not model the 5D distribution function. This results in information loss, as they do not capture interactions across all dimensions. In this work, we attempt for the first time to train a neural surrogate for the evolution of the 5D distribution function. To this end, we first collect simulation data using a state-of-the-art nonlinear solver, namely GKW (Peeters et al., 2009), run with adiabatic electrons for different values of Ion Temperature Gradient (ITG). To compress the high-dimensional input

into a compact representation, we extend a hierarchical vision transformer to process the 5D input. We call our method 5D Swin-UNet and train it to evolve the 5D distribution function over time.

Our neural surrogate accurately models physical quantities for single-step predictions. To verify this, we integrate the predicted 5D distribution function to infer physical quantities such as heat flux time trace and electrostatic potentials for single-step predictions of 5D Swin-UNet. We find that the predicted quantities align well with the ground truth for a simulation with a value for the ion temperature gradient (ITG) which was never observed during training, except for an overestimation of the zonal flow. Furthermore, 5D Swin-UNet is approximately two orders of magnitude faster than GKW for the adiabatic electron approximation. In summary, our contributions are as follows.

- We provide a simple 2D visualization for the 5D distribution function.
- We introduce 5D Swin-UNet, a hierarchical vision Transformer to compress, process and reconstruct 5D data, without convolutions.
- We propose a physics evaluation to validate that our surrogate models accurately capture turbulent transport in Plasma.

2 Related Work

Machine Learning for Gyrokinetics. Most of the literature to date has focused on multilayer perceptrons as surrogate models of turbulence models that adopt the quasilinear approximation. Faster integrated models (Romanelli et al., 2014; Pereverzev & Yushmanov, 2002) of tokamak discharges were obtained for interpretative modelling in existing machines (Meneghini et al., 2017; van de Plassche et al., 2020) as well as predictive modelling (Citrin et al., 2023) and control studies (Mulders et al., 2021) of future reactor-class devices, clearly highlighting the benefits of surrogate modelling, albeit on reduced order models.

The literature on surrogates for higher-fidelity models is sparse. To model the linear spectra of micro-tearing modes, Hornsby et al. (2024) propose leveraging gaussian process regression (GPR). GPR also enables uncertainty quantification, but does not scale well to larger datasets and dimensionalities. Therefore, Narita et al. (2022) leverage convolutional neural networks based on the spectral images of the absolute values of the distribution function along with the electric potential at a fixed time slice and predict the corresponding heat flux and the time to saturation. Honda et al. (2023) extends this idea to a two-dimensional multimodal input space, including electrostatic potentials. Our method fundamentally differs in that it directly models the time evolution of the 5D distribution function, thus enabling the computation of turbulent fluxes at any time.

Neural Operators. Importantly, our 5D Swin-UNet does not fall into the category of neural operators, as it is not resolution invariant but operates on a fixed resolution. However, in future work we aim to extend our model to a neural operator, hence we elaborate on important related work in this regard. Neural operators (Kovachki et al., 2023; Alkin et al., 2024) are formulated with the objective of learning a mapping between function spaces, usually defined as Banach spaces U, V of functions defined on compact input and output domains X and Y, respectively. Neural operators enable continuous outputs that remain consistent across varying input sampling resolutions. A neural operator $\hat{\mathcal{G}} : \mathcal{U} \mapsto \mathcal{V}$ approximates the ground truth operator $\mathcal{G} : \mathrm{U} \mapsto \mathrm{V}$, and is usually composed of three maps $\hat{\mathcal{G}} := \mathrm{D} \circ \mathrm{A} \circ \mathrm{E}$ (Seidman et al., 2022), comprising the encoder E, the approximator A, and the decoder D. Training a neural operator involves constructing a dataset of input-output function pairs evaluated at discrete spatial locations. Training minimizes a mean squared error loss over the discretized space using gradient descent.

Over the recent years, driven by advances in neural operator learning (Lu et al., 2021; Li et al., 2020; Kovachki et al., 2023), deep neural network-based surrogates have emerged as a computationally efficient alternative in science and engineering (Thuerey et al., 2021; Zhang et al., 2023; Brunton et al., 2020), impacting e.g., weather forecasting (Kurth et al., 2023; Bi et al., 2023; Lam et al., 2023; Nguyen et al., 2023; Bodnar et al., 2024), protein folding (Jumper et al., 2021; Abramson et al., 2024), and material design (Merchant et al., 2023; Zeni et al., 2025; Yang et al., 2024).

Swin Transformers. Other works, such as the 4D fMRI Swin Transformer (Kim et al., 2023, SwiFT) extend Swin Transformers to the specific case of four dimensions (three spatial and time) to

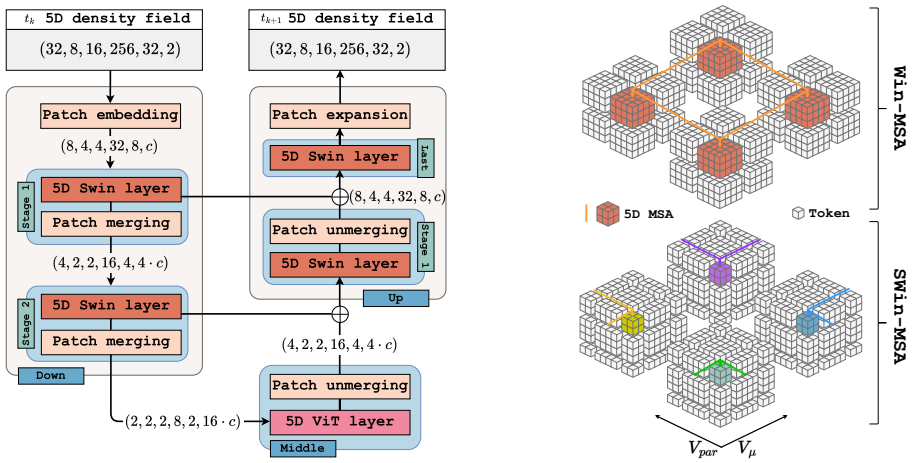


Figure 1: Overview of the model architecture and n-dimensional attention. **Left:** Our 5D Swin-UNet with two down/upsampling stages. We indicate shapes at each resolution change. **Right:** The locality of the n-dimensional shifted window attention for nDWin-MSA (top) and nDSWin-MSA (bottom), illustrated as a 2D plane of 3D window-partitioned tokens. Connected components are highlighted as colored tokens in the 3D blocks, and as lines connecting them across dimensions.

process fMRI scans. Since our nD layers do not rely on assumptions about the number of dimensions, we effectively automate the process used to derive SwiFT (4D) and Video-Swin (Liu et al., 2022) (3D) from standard 2D Swin layers.

3 METHODS

First, we elaborate on the data generation, preprocessing, and visualization techniques in Section 3.1. Next, we explain our proposed 5D Swin-UNet and elaborate on the architectural details in Section 3.2. Finally, in Section 3.3, we elaborate on our evaluation protocol.

3.1 DATA GENERATION

Gyrokinetic simulations are usually initialised from noise and evolved according to nonlinear gyrokinetic equations (Equation (1)). The simulations usually follow a certain pattern. In the linear phase the linear modes start growing, resulting in an initial increase in the heat flux time trace. Afterwards, the simulations enter the nonlinear or saturated regime where the linear modes start interacting, resulting in an oscillatory behaviour of heat flux. Turbulence is generally established in the nonlinear regime and complex integrals over long time intervals are required to obtain an estimate for the heat flux.

We rely on a nonlinear direct numerical gyrokinetic solver to produce samples of the 5D distribution function, namely GKW (Peeters et al., 2009). For background on the gyrokinetic framework, we refer the reader to Appendix B. We consider the adiabatic electron case, where the velocity distribution of electrons is assumed to be Boltzmann-like. Therefore, only the distribution function of the ions is considered. To obtain different trajectories, we vary the ITG and collect 5D fields of resolution $(32 \times 8 \times 16 \times 255 \times 32 \times 2)$ where the sixth dimension corresponds to the real and imaginary part of the ballooning transform, which is usually used to describe plasma coordinates. The five remaining spatial dimensions are denoted as $(V_{||} \times V_\mu \times s \times k_x \times k_y)$, where $V_{||}$ and V_μ represent velocities parallel and perpendicular to the field lines, s is the toroidal angle and k_x and k_y are spatial coordinates in the spectral space. We run GKW for different values of ITG and dump the distribution function every 20 steps to ensure sufficient change. The collected trajectories have 500 such steps that are additionally subsampled every third, resulting in an overall time-coarsening of 60 times and a total of 166 snapshots per trajectory. The dataset comprises five trajectories, resulting in 660 training samples and one unseen trajectory (165 samples). Appendix C provides a detailed discussion on the choice of ITGs used for data generation.

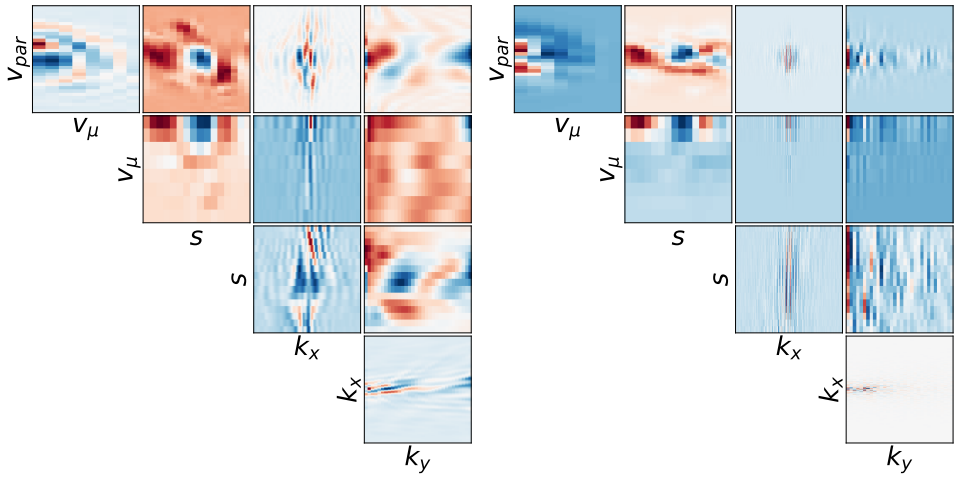


Figure 2: **Left:** 5D distribution function with an ITG of 7.9 in the linear phase ($t = 1.9$). **Right:** 5D distribution function with an ITG of 7.9 in the saturated phase ($t = 50.6$).

To visually inspect the data, we slice trajectories according to timesteps and visualize the 5D data as $\binom{5}{2} = 10$ images of non-repeating combinations of the five different axes. For each combination of two axes, we either reduce or slice the remaining three axes to project the field down to an image. An example visualization for reduction via averaging for time slices of the linear and nonlinear phase of a simulation can be observed in Figure 2. The visualizations for the sliced samples are presented in Figure 7 and Figure 8 in Appendix C.

3.2 5D Swin-UNet

To process the 5D input, we generalize the hierarchical Swin Transformer (Liu et al., 2021) to an arbitrary number of dimensions. The input to 5D Swin-UNet is a 5D field of shape $V_{par} \times V_\mu \times s \times x \times y \times 2$, which consists of the real and imaginary parts of a time snapshot of the nonlinear gyrokinetics simulation. Similarly to other conventional 2D Vision Transformers (Dosovitskiy et al., 2021; Liu et al., 2021), we first partition the input into non-overlapping patches with an n-dimensional patch embedding layer, mapping patch-local information into tokens. Patches are then processed with a UNet-style architecture (Ronneberger et al., 2015), using patch merging and unmerging to produce multiscale hierarchical representations. The original space is reconstructed from patch tokens using a patch expansion layer, which nonlinearly expands the patch space to the original resolution. The proposed 5D Swin-UNet architecture is shown in Figure 1, left.

The multi-head self-attention (MSA) used in Vision Transformers scales quadratically with the sequence length in the 2D case. This effect is exacerbated in 5D, making MSA prohibitively expensive. Therefore, we apply MSA only on local windows as in (Liu et al., 2021, SWin). This is agnostic to dimensionality, i.e., it can be expanded to any number of dimensions with varying window sizes per dimension (nDWin-MSA). To enable interaction between neighboring windows, we also extend cyclic shifts of SWin to 5D (nDSWin-MSA). Finally, we apply self-attention within each window in parallel. Figure 1 (right) visualizes the attention scope for one window in 5D for nDWin-MSA (top) and nDSWin-MSA (bottom).

3.3 Evaluation

To evaluate our method, we consider three approaches. As a first evaluation, we visualize the outputs of the surrogate model. This is done in the same manner as elaborated in Section 3.1 and we apply this scheme to the ground truth data and to the model prediction to obtain a side-by-side comparison. The second evaluation corresponds to visual inspection of the electrostatic potentials. Potentials are computed by an integral over the velocities $V_{||}$ and V_μ of the 5D distribution function. on the predictions of our surrogate model. Since electrostatic potentials are 3D fields, we slice them along the middle of the toroidal angle s and visualize them as images. Finally, we evaluate whether

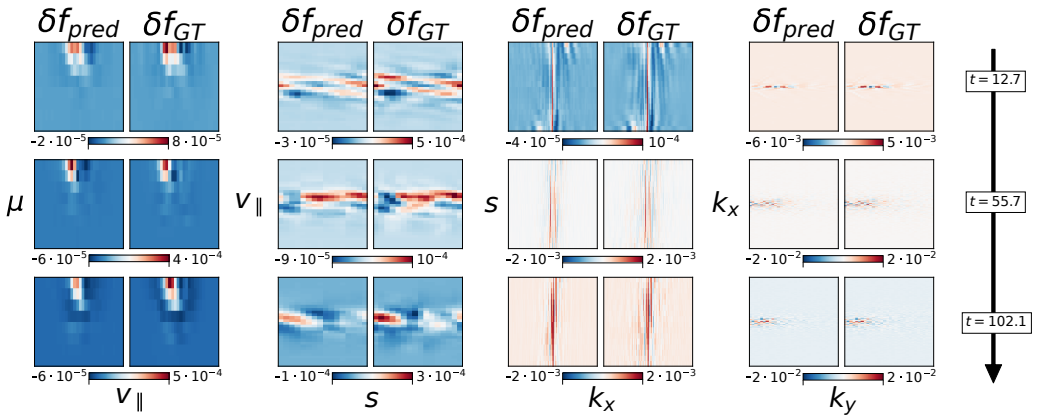


Figure 3: 5D Swin-UNet can accurately predict the distribution functon in five dimensions. We report one-step model predictions of the distribution function, δf_{pred}, versus ground truth, δf_{GT}, over time. For each axis pair, the corresponding 2D projections are obtained by averaging over the remaining dimensions.

our surrogate model accurately predicts the heat flux time trace. The heat flux is a scalar value that oscillates along an average in the saturated phase, and it is computed as a function of the 5D distribution function and the electrostatic potential. It is a key quantity that is used in downstream integrated modelling tools (Romanelli et al., 2014). This evaluation provides insight into whether the model accurately captures the underlying physics.

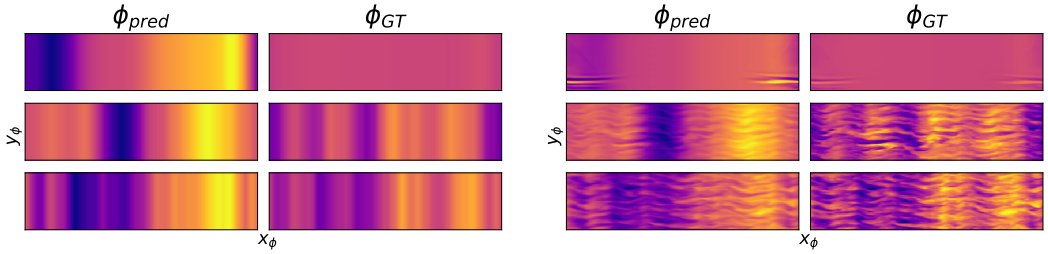
4 Experiments

In our experiments, we train 5D Swin-UNet for next-step prediction of the nonlinear gyrokinetic simulation. We split the data according to trajectories into training and validation. Specifically, we evaluate on one holdout trajectory and use the remaining ones for training. This way, we evaluate generalization to a holdout trajectory produced with a different ITG. Each sample of a trajectory is normalized on a per-sample basis. We train our model for 100 epochs and evaluate every 20 epochs on the holdout trajectory. For implementation details, we refer the reader to Appendix D.

Distribution function. We visualize the predicted 5D field of the distribution function over time. Figure 3 shows averages for four of the 10 possible combinations of the different axes for three different timesteps. The first timestep is from the linear regime, whereas the latter are from the saturated phase. The model prediction generally aligns very well with the ground truth, as both model prediction and ground truth are displayed with euqal colour scale.

Electrostatic Potentials. We obtain the electrostatic potentials as described in Section 3.3 and provide a visual comparison of the predicted and ground-truth electrostatic potential in Figure 4b. It shows that the the vertical wave vector is well reproduced, indicating that our surrogate model captures key aspects of the underlying physics. However, there is a misalignment with respect to the first mode (zonal mode) that represents the zonal flow. To further investigate this, we show the first mode of the electrostatic potential in Figure 4a. Interestingly, 5D Swin-UNet strongly overestimates the presence of zonal flow in the linear phase where it is usually not present (top row). This is likely caused by an imbalanced dataset, where the saturated phase – with a stronger zonal mode – is overrepresented. The zonal flow does not directly contribute to the resulting heat flux in the saturated phase; however, it dampens the development of turbulent transport. This is problematic, especially for autoregressive rollouts, where the predicted zonal flow in the linear phase results in a decaying heat flux. Therefore, we aim to explore different sampling strategies or tailored losses emphasizing the zonal mode in the future.

Heat flux time trace. We assess whether the model correctly infers the heat flux time trace of the holdout simulation (ITG=6.9). Figure 5 shows the predicted heat flux versus the ground truth. The model predictions align very well with the ground-truth heat flux time trace. This can be traced back to the vertical wave vector in the electrostatic potential (see Figure 3, right), which is successfully



(a) Zonal mode of the electrostatic potential.

(b) Electrostatic potential with zonal mode.

Figure 4: Visualization and comparison of the electrostatic potentials for model predictions (ϕ_{pred}) versus ground truth (ϕ_{GT}). In accordance with Figure 3, rows correspond to timesteps $[12.7, 55.7, 102.1]$. Figure 4a shows only the zonal component, while Figure 4b displays all modes of the field. While the vertical waves are well captured, the model overestimates the zonal flow, particularly at the first timestep, corresponding to the linear phase.

recovered by the surrogate model. This indicates that 5D Swin-UNet successfully learns to capture parts of the underlying physics. However, the zonal mode predicted by the surrogate is incorrect. As zonal flow does not contribute to heat flux, it is not visible in Figure 5.

Efficiency. One-step prediction only requires 360 ms on average for 5D Swin-UNet on a Nvidia A100 GPU, while the numerical solver GKW requires approximately 80s using 64 cpu cores. Therefore, 5D Swin-UNet is around two orders of magnitude faster than GKW.

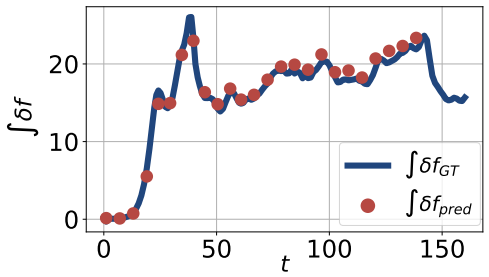


Figure 5: Heat flux time trace ($\int \delta f$) for ground-truth (GT) and single-step prediction of 5D Swin-UNet (pred) for the holdout trajectory (with ITG=6.9). 5D Swin-UNet accurately models the heat flux.

5 Conclusions and Future Work

We present a neural surrogate model for nonlinear gyrokinetic equations modelling turbulent transport in Plasmas. We first present a simple recipe to visually inspect the 5D distributon function. Then, we propose a neural surrogate that operates in 5D space, namely 5D Swin-UNet, and demonstrate that it accurately captures the underlying physics for single-step prediction. Our work paves the way towards neural surrogates for 5D turbulence modelling for nuclear fusion.

As a first step for future work we plan to incorporate additional diagnostics to verify that the surrogate model accurately captures physics beyond electrostatic potentials. Furthermore, a fruitful direction is to include a separate head for explicit prediction of electrostaic potentials. This alleviates the requirement for expensive GKW calls, needed for computing the potential field and the heat flux integrals. Eventually, our aim is to achieve stable long-term autoregressive rollouts with our surrogate model which is currently hindered by the overemphasis on the zonal flow.

REFERENCES

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.

Benedikt Alkin, Andreas Fürst, Simon Schmid, Lukas Gruber, Markus Holzleitner, and Johannes Brandstetter. Universal physics transformers. CoRR, abs/2402.12365, 2024. doi: 10.48550/ARXIV.2402.12365.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. Nat., 619(7970):533–538, 2023. doi: 10.1038/S41586-023-06185-3.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Anna Vaughan, Jayesh K. Gupta, Kit Thambiratnam, Alex Archibald, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. Aurora: A foundation model of the atmosphere. CoRR, abs/2405.13063, 2024. doi: 10.48550/ARXIV.2405.13063.

C Bourdelle, J Citrin, B Baiocchi, A Casati, P Cottier, X Garbet, and F Imbeaux and. Core turbulent transport in tokamak plasmas: bridging theory and experiment with QuaLiKiz. Plasma Physics and Controlled Fusion, 58(1):014036, December 2015. doi: 10.1088/0741-3335/58/1/014036.

Steven L. Brunton, Bernd R. Noack, and Petros Koumoutsakos. Machine learning for fluid mechanics. Annual Review of Fluid Mechanics, 52(Volume 52, 2020):477–508, 2020. ISSN 1545-4479. doi: https://doi.org/10.1146/annurev-fluid-010719-060214.

J Citrin, C Bourdelle, F J Casson, C Angioni, N Bonanomi, Y Camenen, X Garbet, L Garzotti, T Görler, O Gürcan, F Koechl, F Imbeaux, O Linder, K van de Plassche, P Strand, and G Szepesi and. Tractable flux-driven temperature, density, and rotation profile evolution with the quasilinear gyrokinetic transport model QuaLiKiz. Plasma Physics and Controlled Fusion, 59(12):124005, November 2017. doi: 10.1088/1361-6587/aa8aeb.

J. Citrin, P. Trochim, T. Goerler, D. Pfau, K. L. van de Plassche, and F. Jenko. Fast transport simulations with higher-fidelity surrogate models for ITER. Physics of Plasmas, 30(6), jun 2023. doi: 10.1063/5.0136752.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.

Mitsuru Honda, Emi Narita, Shinya Maeyama, and Tomo-Hiko Watanabe. Multimodal convolutional neural networks for predicting evolution of gyrokinetic simulations. Contributions to Plasma Physics, 63(5-6):e202200137, 2023. doi: https://doi.org/10.1002/ctpp.202200137.

W. A Hornsby, A. Gray, J. Buchanan, B. S. Patel, D. Kennedy, F. J. Casson, C. M. Roach, M. B. Lykkegaard, H. Nguyen, N. Papadimas, B. Fourcin, and J. Hart. Gaussian process regression models for the properties of micro-tearing modes in spherical tokamaks. Physics of Plasmas, 31 (1), jan 2024. ISSN 1089-7674. doi: 10.1063/5.0174478.

K. Itoh, S.-I. Itoh, P. H. Diamond, T. S. Hahm, A. Fujisawa, G. R. Tynan, M. Yagi, and Y. Nagashima. Physics of zonal flowsa). Physics of Plasmas, 13(5):055502, 05 2006. ISSN 1070-664X. doi: 10.1063/1.2178779.

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.

Peter Kim, Junbeom Kwon, Sunghwan Joo, Sangyoon Bae, Donggyu Lee, Yoonho Jung, Shinjae Yoo, Jiook Cha, and Taesup Moon. Swift: Swin 4d fmri transformer. Advances in Neural Information Processing Systems, 36:42015–42037, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

Nikola B. Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. J. Mach. Learn. Res., 24:89:1–89:97, 2023.

Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In Proceedings of the Platform for Advanced Scientific Computing Conference, PASC '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701900. doi: 10.1145/3592979.3593412.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. Science, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336.

Zongyi Li, Nikola B. Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Neural operator: Graph kernel network for partial differential equations. CoRR, abs/2003.03485, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pp. 9992–10002. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00986.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pp. 3192–3201. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00320.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. Nat. Mach. Intell., 3(3):218–229, 2021. doi: 10.1038/S42256-021-00302-5.

O. Meneghini, S.P. Smith, P.B. Snyder, G.M. Staebler, J. Candy, E. Belli, L. Lao, M. Kostuk, T. Luce, T. Luda, J.M. Park, and F. Poli. Self-consistent core-pedestal transport simulations with neural network accelerated models. Nuclear Fusion, 57(8):086034, July 2017. doi: 10.1088/1741-4326/aa7776.

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.

S. Van Mulders, F. Felici, O. Sauter, J. Citrin, A. Ho, M. Marin, and K.L. van de Plassche. Rapid optimization of stationary tokamak plasmas in RAPTOR: demonstration for the ITER hybrid scenario with neural network surrogate transport model QLKNN. Nuclear Fusion, 61(8):086019, July 2021. doi: 10.1088/1741-4326/ac0d12.

E. Narita, M. Honda, S. Maeyama, and T.-H. Watanabe. Toward efficient runs of nonlinear gyrokinetic simulations assisted by a convolutional neural network model recognizing wavenumber-space images. Nuclear Fusion, 62(8):086037, jun 2022. doi: 10.1088/1741-4326/ac70e8.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, volume 202 of Proceedings of Machine Learning Research, pp. 25904–25938. PMLR, 2023.

A.G. Peeters, Y. Camenen, F.J. Casson, W.A. Hornsby, A.P. Snodin, D. Strintzi, and G. Szepesi. The nonlinear gyro-kinetic flux tube code gkw. Computer Physics Communications, 180(12): 2650–2672, 2009. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2009.07.001. 40 YEARS OF CPC: A celebratory issue focused on quality software for high performance, grid and novel computing architectures.

G.V. Pereverzev and P.N. Yushmanov. Astra - automated system for transport analysis, 2002.

M Romanelli, G Corrigan, V Parail, Sven Wiesen, Roberto Ambrosino, P Da Silva Aresta Belo, Luca Garzotti, P Harting, F Köchl, Tuomas Koskela, L Lauro-Taroni, Chiara Marchetto, Massimiliano Mattei, E Militello-Asp, M Nave, Stanislas Pamela, A Salmi, P Strand, and G Szepesi. JINTRAC: A system of codes for integrated simulation of tokamak scenarios. Plasma and Fusion Research, 9, 01 2014. doi: 10.1585/pfr.9.3403023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi (eds.), Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III, volume 9351 of Lecture Notes in Computer Science, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Jacob H. Seidman, Georgios Kissas, Paris Perdikaris, and George J. Pappas. NOMAD: nonlinear manifold decoders for operator learning. 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.

G. M. Staebler and J. E. Kinsey. Electron collisions in the trapped gyro-landau fluid transport model. Physics of Plasmas, 17(12), dec 2010. ISSN 1089-7674. doi: 10.1063/1.3505308.

G. M. Staebler, J. E. Kinsey, and R. E. Waltz. A theory-based transport model with comprehensive physics. Physics of Plasmas, 14(5), may 2007. ISSN 1089-7674. doi: 10.1063/1.2436852.

Nils Thuerey, Philipp Holl, Maximilian Müller, Patrick Schnell, Felix Trost, and Kiwon Um. Physics-based deep learning. CoRR, abs/2109.05237, 2021.

K. L. van de Plassche, J. Citrin, C. Bourdelle, Y. Camenen, F. J. Casson, V. I. Dagnelie, F. Felici, A. Ho, S. Van Mulders, JET Contributors, and et al. Fast modeling of turbulent transport in fusion plasmas using neural networks. AIP Publishing, Feb 2020.

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.

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.

Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, Keir Adams, Maurice Weiler, Xiner Li, Tianfan Fu, Yucheng Wang, Haiyang Yu, Yuqing Xie, Xiang Fu, Alex Strasser, Shenglong Xu, Yi Liu, Yuanqi Du, Alexandra Saxton, Hongyi Ling, Hannah Lawrence, Hannes Stärk, Shurui Gui, Carl Edwards, Nicholas Gao, Adriana Ladera, Tailin Wu, Elyssa F. Hofgard, Aria Mansouri Tehrani, Rui Wang, Ameya Daigavane, Montgomery Bohde, Jerry Kurtin, Qian Huang, Tuong Phung, Minkai Xu, Chaitanya K. Joshi, Simon V. Mathis, Kamyar Azizzadenesheli, Ada Fang, Alán Aspuru-Guzik, Erik J. Bekkers, Michael M. Bronstein, Marinka Zitnik, Anima Anandkumar, Stefano Ermon, Pietro Liò, Rose Yu, Stephan Günnemann, Jure Leskovec, Heng Ji, Jimeng Sun, Regina Barzilay, Tommi S. Jaakkola, Connor W. Coley, Xiaoning Qian, Xiaofeng Qian, Tess E. Smidt, and Shuiwang Ji. Artificial intelligence for science in quantum, atomistic, and continuum systems. CoRR, abs/2307.08423, 2023. doi: 10.48550/ARXIV.2307.08423.

CONTENTS

A REPRODUCIBILITY STATEMENT

Along with our submission, we provide the code in a zipped archive. The code contains a readme file with all necessary instructions to run it. For data generation, we used the publicly available numerical code GKW, which can be found at https://bitbucket.org/gkw/workspace/projects/GKW. The `input.dat` file in the code archive specifies all the input parameters required to reproduce the training set.

B GYROKINETIC FRAMEWORK

Solved within GKW is the gyrokinetic set of equations. The full details can be found in Peeters et al. (2009) and references therein. The δf approximation is used, in which the distribution function is split into a background, Maxwellian distribution function F_M, and a perturbed distribution f which is a 5 dimensional function, $f = f(k_x, k_y, s, V_{||}, V_\mu)$. The final equation for the perturbed distribution function f, for each species can be written in the form

$$\frac{\partial f}{\partial t} + (v_\parallel \mathbf{b} + \mathbf{v}_D) \cdot \nabla f + \mathbf{v}_\chi \cdot \nabla f - \frac{\mu B}{m} \frac{\mathbf{B} \cdot \nabla B}{B^2} \frac{\partial f}{\partial v_\parallel} = S, \tag{1}$$

where S is the source term which is determined by the background distribution function, μ is the magnetic moment, $v_{||}$ is the velocity along the magnetic field, $\mathbf{v}_D$ denotes the sum of the drift velocities, $\mathbf{v}_\chi$ is the $E \times B$ velocity, B is the magnetic field strength, m and Z are the particle mass and charge number respectively. The background is assumed to be a shifted Maxwellian (F_M), whose terms are moved to the right-hand side as a source term S, which represent the drives of the underlying instabilities that are studied. The electrostatic potential, ϕ, is calculated from the gyrokinetic Poisson equation.

The thermal velocity $v_{\text{th}} \equiv \sqrt{2T/m}$, and the major radius ($R$) are used to normalise the length and time scales. Using standard gyrokinetic ordering, the length scale of perturbations along the field line ($R\nabla_\parallel \approx 1$) is significantly longer than those perpendicular to the field ($R\nabla_\perp \approx 1/\rho_*$). Here, $\rho_* = \rho_i/R$ is the normalised ion Larmor radius (where $\rho_i = m_i v_{th}/eB$ and $v_{th} = \sqrt{2T_i/m_i}$). To harness this, GKW uses straight field-line Hamada coordinates (s, ζ, ψ) where s is the coordinate along the magnetic field and ζ is the generalised toroidal angle. GKW uses a Fourier representation in the toroidal (y) and radial directions (x), perpendicular to the magnetic field.

C DATA GENERATION AND VISUALISATION

Data Generation.

The choice of ITG significantly impacts the behaviour of the simulations as it is a major driver of plasma turbulence. In general, higher values lead to the accumulation of particle density, resulting in drift effects that eventually turn into turbulence. This can be visualized by comparing flux traces across different ITGs (see Figure 6). Generally, as ITG increases, the system exhibits more nonlinear interactions, leading to higher heat flux values, $\int \delta f$. For values of $\{5.9, 7.9, 10, 15\}$, one can clearly separate between a linear and a saturated phase; this distinction is not evident for lower values. As values above 15 are usually not observed in Tokamaks and those below 3 do not cause any turbulent behaviour at all (see violet curve in Figure 6), we sample within the range $[4, 10]$ to collect the first

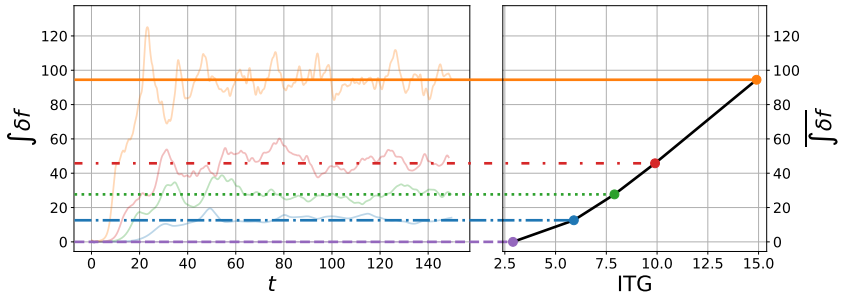
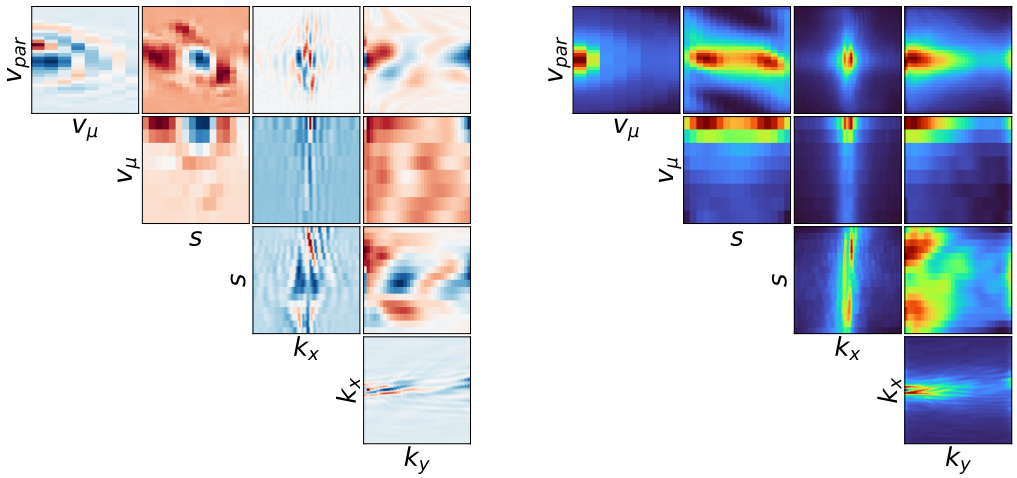


Figure 6: Flux trace over time (left) and averaged (right) for selected ITG values.

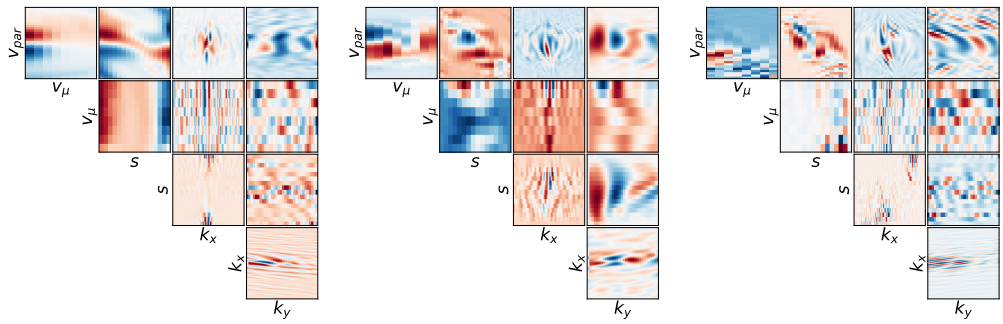
dataset. Specifically, we use ITG $\in [4.9, 5.9, 6.9, 7.9, 8.9]$, resulting in a training set of 4 trajectories with 660 samples and a validation set of 165 samples. Stored with single precision, this leads to a dataset size of 456 GB.

5D Data Visualization. We visualize an example of the generated data in Figure 7 and Figure 8. They show the 5D distribution function for an ITG of 7.9 (green curve in Figure 6) in the linear and saturated phases, respectively. Those figures show all 10 pairwise combinations of axes. To represent the remaining 3 axes in a 2D image, we compress them by taking the average (see Figures 7a and 8a), computing the standard deviation (see Figures 7b and 8b), or extracting slices at specific positions mainly at the start (see Figures 7c and 8c), in the middle (see Figures 7d and 8d), or at the end (see Figures 7e and 8e).



(a) Mean

(b) Standard Deviation

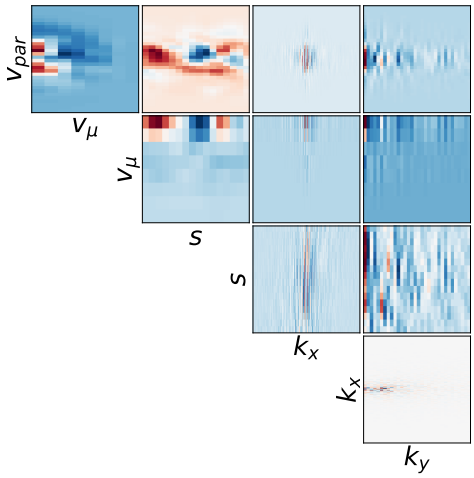


(c) Start slice

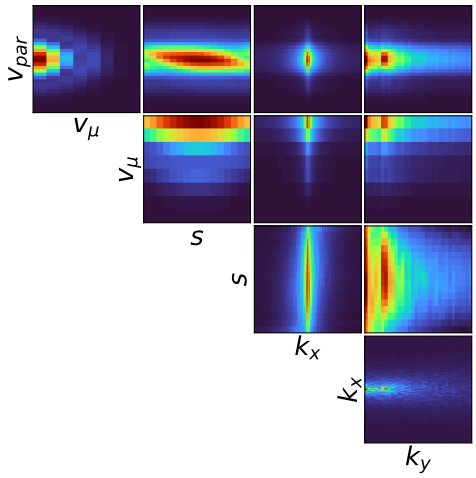
(d) Middle slice

(e) End slice

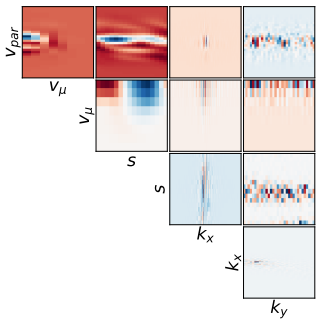
Figure 7: Distribution function δf with an ITG of 7.9 at time $t = 1.9$ (Linear Phase).



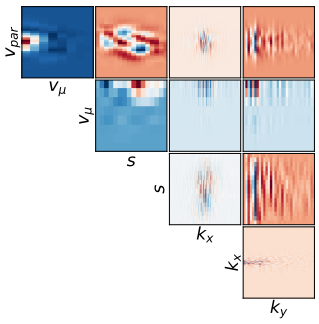
(a) Mean



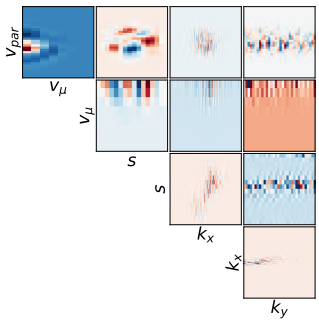
(b) Standard Deviation



(c) Start slice



(d) Middle slice



(e) End slice

Figure 8: Distribution function δf with an ITG of 7.9 at time $t = 50.6$ (Saturated Phase).

D IMPLEMENTATION DETAILS

While Vision Transformers usually employ convolutions for patch operations (Dosovitskiy et al., 2021; Liu et al., 2021; 2022), we use simple fully connected layers because the computational cost of generalized convolutions can become significant in higher dimensions without an optimized implementation. Patch embedding, merging, and expansions are implemented as linear layers or MLPs. The efficient application of nDSWin-MSA is also implemented similarly to Swin. The shift is performed as a cyclic roll along all the window-partitioned axes. Because such operations can produce batches of windows that are not adjacent, we apply a n-dimensional window mask.

We use the Adam optimizer (Kingma & Ba, 2015) with a weight decay of 1e-5 and a cosine learning rate scheduler with linear warmups with a peak at 1e-3, decayed to 0. During training we employ automatic mixed precision and gradient clipping to a magnitude of 1. We also transfer the spectral domains (k_x and k_y) to real space for training by applying inverse FFT. Since a single training sample comprises several GBs of data, we perform lazy dataloading. Model selection is performed every 20 training epochs based on the mean squared error on the holdout trajectory.