

Highlights

Physics-Constrained Neural Closure for Lattice Boltzmann Large-Eddy Simulation

Muhammad Idrees Khan, Sauro Succi, Hua-Dong Yao, Giacomo Falcucci

- Data-driven ML-LES closure with physics constraints and compact neural architecture
- 9 features: strain-rate (6) and vorticity (3)
- Physics-constrained training with Π -matching, rotational equivariance, and divergence-free SGS-force penalty
- Hybrid coupling: dissipative part via effective viscosity, anisotropic residual via forcing; captures backscatter
- ONNX Runtime deployment in Fortran/OpenACC with batched GPU inference
- A priori and a posteriori validation vs classical SGS models; preliminary transfer test in channel flow without retraining

Physics-Constrained Neural Closure for Lattice Boltzmann Large-Eddy Simulation

Muhammad Idrees Khan^{a,*}, Sauro Succi^{b,c}, Hua-Dong Yao^d, Giacomo Falcucci^{a,c}

^a*Department of Enterprise Engineering “Mario Lucertini”, University of Rome “Tor Vergata”, Via del Politecnico 1, Rome, 00133, Italy*

^b*Italian Institute of Technology, Piazzale Aldo Moro 1, Rome, 00185, Italy*

^c*Department of Physics, Harvard University, 33 Oxford Street, Cambridge, 02138, Massachusetts, USA*

^d*Department of Mechanics and Maritime Sciences, Chalmers University of Technology, , Gothenburg, 41296, Sweden*

Abstract

We present a physics-constrained, data-driven subgrid-scale (SGS) stress closure for large-eddy simulation (LES) in the lattice Boltzmann method (LBM). Trained on filtered-downsampled (FD) data from LBM direct numerical simulation (DNS) of forced homogeneous isotropic turbulence (FHIT) spanning multiple filter widths, a compact neural network maps nine macroscopic derivative inputs—six strain-rate and three vorticity components—to the six independent components of the SGS stress tensor; a deviatoric projection is applied post-inference to obtain the traceless stress used in the solver. Training combines a stress data loss with physics terms for SGS energy-transfer (II) matching, rotational equivariance under cube rotations, and compatibility of the implied SGS forcing with the divergence-based coupling.

The predicted stress is coupled to the solver through a split strategy: a dissipative, strain-aligned contribution is represented through an effective-viscosity projection, while the remaining anisotropic residual is applied through a forcing term. This construction is intended to retain both backscatter (via the effective viscosity) and non-dissipative anisotropic effects (via the resid-

*Corresponding author.

Email address: `muhammadidrees.khan@students.uniroma2.eu` (Muhammad Idrees Khan)

ual forcing), while remaining compatible with LBM deployment. In the cases considered here, a priori results show good agreement with FD references across stress components and SGS-transfer statistics, and a posteriori roll-outs improve several energetic and statistical measures relative to static and dynamic Smagorinsky baselines. A preliminary transfer test in turbulent channel flow is also reported without retraining. Finally, we demonstrate production deployment via ONNX Runtime, with throughput comparable to a dynamic Smagorinsky baseline in the tested configuration.

Keywords: Large-eddy simulation, subgrid-scale modeling, lattice Boltzmann method, machine learning, turbulence

1. Introduction

Large-eddy simulation (LES) is widely used to study turbulent flows at Reynolds numbers for which direct numerical simulation (DNS) remains prohibitively expensive. In LES, the governing equations are spatially filtered and the influence of unresolved scales must be represented through the subgrid-scale (SGS) stress. Classical SGS closures remain attractive because of their robustness and low cost, but they often trade structural fidelity for stability, which limits their ability to represent stress anisotropy and energy backscatter.

Recent data-driven turbulence modeling spans anisotropic Reynolds-stress modeling, supervised SGS closure learning, and reinforcement-learning-based closure discovery. Tensor-valued learning of Reynolds stresses has been explored, for example, in neural-network models for anisotropic Reynolds stresses in turbulent channel flow [1]. Supervised learning of LES closure terms and SGS stresses has been studied extensively in continuum Navier–Stokes LES [2], while reinforcement-learning approaches have been used to discover SGS closures in LES [3] and wall models in wall-modeled LES (WMLES) [4]. Related data-driven closure discovery has also recently been explored in coarse-grained lattice Boltzmann method (LBM) settings [5]. These developments motivate the present work, but most of them have been formulated in continuum Navier–Stokes solvers rather than in LBM frameworks.

Explicit SGS-stress learning has been studied mainly in continuum Navier–Stokes LES; its use in LBM LES remains much less explored. A key advantage of LBM for LES is that stress and pressure are locally available in space and time from the distribution functions, without requiring global pressure

solves; this local availability motivates strong interest in LES/LBM. Whether LBM confers additional computational advantages in training or extends to regimes where continuum solvers fail are open questions beyond the scope of this work. We focus instead on whether an explicit-stress ML closure can be coupled to LBM in a stable and practical way, and what accuracy and runtime it achieves relative to conventional viscosity-only baselines. In the cases examined here, the coupled closure improves stress and energy-transfer statistics and captures backscatter that Smagorinsky suppresses, while achieving throughput comparable to a dynamic Smagorinsky baseline (Appendix D, Table D.3).

These considerations lead to three questions that guide the paper:

1. Can a compact ML model predict the full SGS stress tensor accurately enough to be useful in closed-loop LBM-LES?
2. Can that predicted stress be coupled back to LBM in a stable and efficient way, despite the kinetic form of the solver?
3. In the test cases considered here, does such an explicit-stress coupling provide advantages over standard viscosity-only LBM closures, particularly for anisotropy and backscatter?

A broader motivation is whether ML can learn a closure more realistic, general, or accurate than the extended kinetic relaxation-time formulation of Chen *et al.* [6].

To address these questions, we develop an explicit-stress ML closure tailored to LBM. Training data are generated from LBM DNS of forced homogeneous isotropic turbulence (FHIT) and processed into filtered-downsampled (FD) fields at multiple filter widths to form paired datasets of resolved fields and exact SGS stresses. The neural network is trained on macroscopic derivative features (strain-rate and vorticity components) and predicts the six independent components of the SGS stress tensor; a deviatoric projection is applied post-inference before coupling to the solver. The predicted stress is then introduced through a hybrid scheme: the dissipative, strain-aligned part is represented through a modified relaxation rate, while the residual anisotropic part is applied through Guo forcing [7]. This split is designed to preserve the parts of the stress that are naturally compatible with viscosity-based LBM coupling while retaining non-dissipative contributions that would be lost in a purely eddy-viscosity formulation.

The contributions of this work are as follows. First, we study whether compact macroscopic derivative features provide a sufficient basis for pre-

dicting SGS stresses in an LBM setting. Second, we introduce a hybrid stress-splitting strategy that couples the dissipative part of the learned stress through relaxation and the residual part through forcing, enabling stable long-time LES in the cases examined here. Third, we assess the resulting model both a priori and a posteriori, by comparing LES with filtered DNS across stress and energy-transfer p.d.f.s, energy balance, and backscatter statistics. Finally, we show that the trained network can be exported to ONNX Runtime and integrated into a production Fortran/OpenACC LBM code with modest overhead in the tested deployment.

The remainder of the paper is organized as follows. Section 2 reviews classical SGS modeling approaches and recent ML-based closures, with emphasis on LBM implementations. Section 3 describes the governing equations, solver setup, feature design, and training methodology. Section 4 presents both a priori and a posteriori validation, including ablation studies. Section Appendix D summarizes deployment and performance. Section 5 summarizes the findings.

2. Related work

2.1. Classical SGS models

A wide range of subgrid-scale (SGS) closures has been proposed for LES. Functional (eddy-viscosity) models approximate the deviatoric SGS stress as $\tau'_{ij} = -2\nu_t S_{ij}$, with the eddy viscosity ν_t determined from resolved fields. Classical eddy-viscosity models such as the Smagorinsky model [8] and its dynamic variants [9, 10] remain attractive because of their robustness and low cost, but when constrained to nonnegative eddy viscosity they suppress backscatter and have limited ability to represent full stress anisotropy.

Structural models attempt to reconstruct tensorial features of the SGS stress from the resolved field. Representative examples include the scale-similarity (Bardina) model [11] and the gradient/Clark model [12]. These approaches can capture some non-dissipative behavior but often require additional stabilization (e.g. mixed formulations combining similarity/gradient terms with eddy viscosity [13]) to ensure numerical robustness and acceptable dissipation levels. Despite extensive refinements, classical closures still face challenges in universality across flow types, Reynolds numbers, filter widths, and grid resolutions.

2.2. ML-based SGS models

Data-driven turbulence modeling has developed rapidly across both Reynolds-averaged and LES settings. Representative tensor-valued learning studies showed that neural networks can be trained to predict anisotropic turbulence stresses in channel flow [1], illustrating the feasibility of learning structured stress tensors rather than only scalar closure coefficients. In LES, a substantial body of work in continuum Navier–Stokes solvers has focused on learning SGS stresses, SGS forces, or related closure terms from resolved flow features [2, 14–18], with a posteriori stability and generalization emerging as central concerns. These models can represent anisotropy and backscatter more flexibly than purely functional eddy-viscosity closures, but strong a priori agreement does not by itself guarantee stable closed-loop LES [15, 16].

The broader ML turbulence literature also includes reinforcement-learning approaches to closure discovery. For homogeneous isotropic turbulence (HIT), multi-agent reinforcement learning has been used to discover turbulence closures [3]; related work has extended this perspective to wall-model discovery in wall-bounded turbulence [4]. We cite these studies because they are important landmarks in data-driven turbulence modeling, even though their learning objectives, solver couplings, and deployment strategies differ from the supervised stress-regression approach adopted here.

Most explicit-SGS-stress ML studies have been developed in continuum Navier–Stokes LES settings rather than in LBM. Across that literature, common inputs include local velocity gradients and derived tensorial quantities such as strain rate, rotation/vorticity, and, in some cases, resolved-stress or Leonard/test-filtered terms; some studies also employ invariant-based or nonlocal feature representations [15, 16, 18].

A closely related example is Yang et al. [18], who predict the full SGS stress tensor in a Navier–Stokes LES framework using strain-rate and modified Leonard-stress inputs with data-only training. We cite that work as an important continuum-solver reference point; the distinction of the present work lies not in predicting a full stress tensor per se, but in doing so within LBM using derivative-only local features, physics-constrained training, and a dissipative/residual stress split for coupling.

2.3. LBM-specific LES and learned closures

A kinetic-theory perspective on turbulence modeling in LBM was introduced by Chen *et al.* in the extended Boltzmann kinetic equation for turbulent flows [6] and further developed in their expanded analogy between

Boltzmann kinetic theory and turbulence [19]. This kinetic-theory-based formulation is a natural target for comparison. Within the LBM, stress and pressure are locally available in space and time from the distribution functions, which makes LBM a natural setting for LES closures that depend on local strain-rate and related quantities. LES has most commonly been realized by modifying relaxation times to represent an eddy viscosity associated with a chosen SGS model [20–22]. This viscosity-based strategy fits naturally within LBM and has therefore remained the dominant LES pathway. More recent ML activity in LBM has explored data-driven tuning of relaxation parameters and learned collision operators [5, 23]. Other recent directions include physics-informed near-wall modeling [24] and learned kinetic operators [25–27].

Existing studies already show that machine learning is active in LBM, including turbulence-related closures, near-wall modeling, and learned kinetic operators [5, 23, 24, 26, 27]. Relative to the LBM studies cited above, the present work focuses on a different question: whether the full SGS stress tensor can be predicted from local derivative features and coupled to LBM through a hybrid dissipative/residual split. In the present formulation, the strain-aligned dissipative part is represented through an effective viscosity and the corresponding stress-mode relaxation-rate modification, while the remaining anisotropic residual is introduced through a forcing term (Guo forcing in our implementation) [7]. This construction is intended to retain SGS effects that are not naturally represented by a scalar eddy-viscosity closure alone.

Positioning. Classical SGS models emphasize robustness, but often at the cost of structural fidelity. The kinetic-theory formulation of Chen *et al.* [6] provides a principled baseline; a central question is whether ML can learn a closure that is more realistic, general, or accurate. ML closures in continuum solvers have shown that explicit stress prediction can recover richer SGS behavior, but their coupling strategies do not transfer directly to LBM. Existing LBM studies, meanwhile, have mostly remained within viscosity-based or kinetic-modification paradigms. The present study is positioned at the intersection of these lines: it examines whether a compact explicit-stress ML closure can be coupled stably and efficiently to LBM through a dissipative/residual split, and whether that strategy improves selected a priori and a posteriori behavior relative to standard viscosity-only baselines in the flows considered here.

3. Methods

This section describes the methodology for developing and deploying ML-based SGS closure models for LBM-based LES, including the LBM formulation, feature design, network architecture, training strategy, and integration with the production solver.

3.1. Filtered Navier–Stokes equations

For reference, we introduce the usual LES definition of the SGS stress by filtering the incompressible Navier–Stokes equations [28]. Applying a spatial filter $\overline{(\cdot)}$ and defining $\tau_{ij}^{\text{sgs}} = \overline{u_i u_j} - \bar{u}_i \bar{u}_j$, the filtered momentum equation reads

$$\frac{\partial \bar{u}_i}{\partial t} + \frac{\partial}{\partial x_j} (\bar{u}_i \bar{u}_j) = -\frac{1}{\rho} \frac{\partial \bar{p}}{\partial x_i} + \nu \nabla^2 \bar{u}_i - \frac{\partial \tau_{ij}^{\text{sgs}}}{\partial x_j} + \bar{f}_i. \quad (1)$$

We emphasize that our simulations are performed with MRT-LBM; the filtered Navier–Stokes form is introduced only to define τ_{ij}^{sgs} and its appearance through $-\nabla \cdot \boldsymbol{\tau}^{\text{sgs}}$.

3.2. MRT-LBM with Guo forcing

We use a D3Q19 multiple-relaxation-time (MRT) LBM formulation [29]. In compact form, the evolution with a body force \mathbf{F} (Guo forcing [7]) can be written as

$$f_\alpha(\mathbf{x} + \mathbf{c}_\alpha \Delta t, t + \Delta t) = f_\alpha(\mathbf{x}, t) - \left[\mathbf{M}^{-1} \boldsymbol{\Lambda} (\mathbf{m} - \mathbf{m}^{eq}) \right]_\alpha + \Delta t \left[\mathbf{M}^{-1} \left(\mathbf{I} - \frac{\boldsymbol{\Lambda}}{2} \right) \mathbf{m}^F \right]_\alpha. \quad (2)$$

where $\mathbf{m} = \mathbf{M}\mathbf{f}$ are moments, $\boldsymbol{\Lambda}$ is the diagonal matrix of MRT relaxation rates, and $\mathbf{m}^F = \mathbf{M}\boldsymbol{\Phi}(\mathbf{F})$ is the forcing term in moment space. The D3Q19 discrete-velocity set and the corresponding equilibrium distribution are summarized in Appendix A. Implementation and validation details of the MRT-LBM DNS/LES solver and forcing treatment used in this work can be found in [30]. All DNS, training-data generation, LES baselines (including dynamic Smagorinsky), and ONNX-coupled inference rollouts were performed on an NVIDIA A100 (80 GB) GPU. For deployment we use ONNX Runtime GPU (v1.16.3); the reported runs used CUDA v12.2. All quantities are reported in lattice units (lu).

3.3. Conventional LBM LES viscosity closure

In this conventional LBM LES approach, an eddy viscosity ν_t is introduced through an effective viscosity

$$\nu_e = \nu_0 + \nu_t, \quad (3)$$

$$\tau_e = \frac{1}{2} + 3\nu_e, \quad s_\nu = \frac{1}{\tau_e}, \quad (4)$$

where ν_0 is the molecular viscosity (lu), τ_e is the effective relaxation time for stress modes, and s_ν is the corresponding relaxation rate. For further discussion of the viscosity-based LBM LES formulation and its numerical implementation in FHIT, see [30]. For the static Smagorinsky baseline we use $C_s = 0.2$; in figure legends, ‘‘Smagorinsky’’ denotes this static model and ‘‘Dynamic’’ the dynamic procedure below.

3.3.1. Dynamic Smagorinsky

To benchmark the ML-LES model against a conventional dynamic eddy-viscosity closure in periodic FHIT, we implement a Germano-type dynamic Smagorinsky procedure in the LBM solver. The deviatoric SGS stress is modeled as

$$\tau'_{ij} = -2\nu_t \bar{S}_{ij}, \quad \nu_t = C \Delta^2 |\bar{S}|, \quad |\bar{S}| = \sqrt{2\bar{S}_{ij}\bar{S}_{ij}}, \quad (5)$$

where $C \equiv C_s^2$ is obtained dynamically and Δ is the grid filter width (taken as one lattice spacing in the baseline runs). Consistent with MRT-LBM practice, the strain-rate tensor is evaluated locally from non-equilibrium moments rather than finite differences. Denoting $\mathbf{m}^{neq} = \mathbf{m} - \mathbf{m}^{eq}$ and the current stress-mode relaxation rate by s_ν , the baseline uses an MRT moment-based relation $\bar{S}_{ij} = \mathcal{G}_{ij}(\mathbf{m}^{neq}; s_\nu)$.

We define a test filter $\widetilde{(\cdot)}$ with ratio $r \equiv \widetilde{\Delta}/\Delta$ (taken as $r = 2$ in our FHIT simulations). The Leonard stress is

$$L_{ij} = \widetilde{\widetilde{u_i u_j}} - \widetilde{u_i} \widetilde{u_j}, \quad (6)$$

and the model tensor is formed as

$$M_{ij} = \widetilde{\Delta}^2 |\widetilde{S}| \widetilde{S}_{ij} - \Delta^2 |\bar{S}| \bar{S}_{ij}. \quad (7)$$

The dynamic coefficient is computed using deviatoric contractions,

$$L'_{ij} = L_{ij} - \frac{1}{3}L_{kk}\delta_{ij}, \quad M'_{ij} = M_{ij} - \frac{1}{3}M_{kk}\delta_{ij}, \quad C = -\frac{1}{2}\frac{\langle L'_{ij}M'_{ij} \rangle}{\langle M'_{ij}M'_{ij} \rangle}, \quad (8)$$

where $\langle \cdot \rangle$ denotes a domain average appropriate for homogeneous turbulence. The resulting eddy viscosity is combined with the molecular value as $\nu_e = \nu_0 + \nu_t$ and mapped to the local stress relaxation rate through $\tau_e = \frac{1}{2} + 3\nu_e$ and $s_\nu = 1/\tau_e$. In the present baseline implementation, ν_t is clipped to be non-negative before updating s_ν [20].

3.4. ML-coupled LBM: effective viscosity and residual forcing

The ML model predicts the six independent components of the SGS stress tensor; a deviatoric projection (trace removal) is applied post-inference to obtain τ_{ij}^{ML} . We define the SGS energy transfer channel as

$$\Pi = -\tau_{ij}^{\text{ML}}S_{ij}, \quad (9)$$

where S_{ij} is the resolved strain-rate tensor and we use $|S| = \sqrt{2S_{ij}S_{ij}}$. We then extract an effective viscosity contribution by dissipation matching [11]:

$$\nu_t^{\text{eff}} = \frac{\Pi}{|S|^2}, \quad \nu_e = \nu_0 + \nu_t^{\text{eff}}. \quad (10)$$

In contrast to common eddy-viscosity baselines, we do not enforce non-negativity of Π or ν_t^{eff} ; negative values (backscatter) are therefore retained rather than suppressed by construction. We then compute the stress relaxation rate from the effective viscosity as $s_\nu = 1/(\frac{1}{2} + 3\nu_e)$. To retain non-dissipative SGS effects, we define the dissipative projection $\tau_{ij}^{\text{diss}} = -2\nu_t^{\text{eff}}S_{ij}$ and the residual stress

$$\tau_{ij}^{\text{res}} = \tau_{ij}^{\text{ML}} - \tau_{ij}^{\text{diss}}. \quad (11)$$

The corresponding residual force density is applied via Guo forcing [7] for the results reported herein; the exact-difference method [31] yields similar behavior:

$$\mathbf{F}^{\text{res}} = -\nabla \cdot \boldsymbol{\tau}^{\text{res}}, \quad \mathbf{F}_{\text{total}} = \mathbf{F}_{\text{ext}} + \mathbf{F}^{\text{res}}. \quad (12)$$

Table 1: Summary of main training cases used for the reported results. Here Δ denotes the filter width (in grid points of the DNS), and $\eta^+ = \eta/\Delta x$ is the dimensionless Kolmogorov length scale.

DNS	η^+	Δ	FD grid	Target ν
256^3	0.6	4	64^3	0.001
256^3	1.3	8	32^3	0.004
192^3	0.8, 1.5	6	32^3	0.002, 0.005

In practice we evaluate $\nabla \cdot \boldsymbol{\tau}^{\text{res}}$ using centered finite differences with periodic boundary conditions, consistent with the LBM discretization.

Complete ML-LES LBM update. Starting from the DNS MRT-LBM evolution with forcing (Eq. 2), the ML-coupled LES update is obtained by (i) replacing the stress-mode relaxation rate inside $\boldsymbol{\Lambda}$ by the ML-based value $s_\nu(\nu_e)$, and (ii) using the total forcing $\mathbf{F}_{\text{total}} = \mathbf{F}_{\text{ext}} + \mathbf{F}^{\text{res}}$:

$$f_\alpha(\mathbf{x} + \mathbf{c}_\alpha \Delta t, t + \Delta t) = f_\alpha(\mathbf{x}, t) - \left[\mathbf{M}^{-1} \boldsymbol{\Lambda}^{\text{ML}} (\mathbf{m} - \mathbf{m}^{\text{eq}}) \right]_\alpha + \Delta t \left[\mathbf{M}^{-1} \left(\mathbf{I} - \frac{\boldsymbol{\Lambda}^{\text{ML}}}{2} \right) \mathbf{m}^{\mathbf{F}_{\text{total}}} \right]_\alpha. \quad (13)$$

where $\boldsymbol{\Lambda}^{\text{ML}}$ equals the DNS relaxation matrix except for the stress relaxation rate s_ν , which is computed from $\nu_e = \nu_0 + \nu_t^{\text{eff}}$ with $\nu_t^{\text{eff}} = \Pi/|\mathbf{S}|^2$ and $\Pi = -\tau_{ij}^{\text{ML}} S_{ij}$.

Deployment details for the ONNX-coupled inference pathway are summarized in Appendix D; a step-by-step algorithm for reproducibility is given in Algorithm 1.

By construction, the residual stress $\boldsymbol{\tau}^{\text{res}}$ is orthogonal to the strain-rate tensor and therefore non-dissipative. Characterization of the stress split and numerical verification of this property during a posteriori rollouts are presented in Section 4.2.1 (Figure 4c,f).

3.5. Training data, features, and preprocessing

Table 1 summarizes the main training cases. The cases are pooled into a single dataset and split by files into train (80%), validation (10%), and test (10%); all DNS fields are generated with the MRT-LBM solver described in Section 3.

For the main ML results, the model employs 9 macroscopic derivative features: the six independent components of the symmetric strain-rate tensor

$$S_{ij} = \frac{1}{2} \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right), \quad (14)$$

namely S_{xx} , S_{yy} , S_{zz} , S_{xy} , S_{yz} , S_{xz} , together with the three components of the vorticity vector $\boldsymbol{\omega} = \nabla \times \mathbf{u}$, namely ω_x , ω_y , ω_z . For the ML closure, strain-rate components are evaluated locally using centered finite differences on the velocity field, both during data preparation and at inference. To ensure consistent input scaling across datasets generated at different filter widths, we apply a per-simulation Δ -normalization as part of the preprocessing described below.

Derivative-based features such as strain-rate and vorticity scale inversely with the filter width, i.e., S_{ij} , $\omega_i \sim U/\Delta$. As a consequence, their numerical magnitudes differ across datasets filtered at different resolutions, which would otherwise introduce a spurious dependency on Δ into the learning process. To make the derivative inputs comparable across filter widths, we apply a per-file (per-simulation) Δ -normalization to all derivative components:

$$S_{ij}^* = \Delta^{(s)} S_{ij}, \quad \omega_i^* = \Delta^{(s)} \omega_i, \quad (15)$$

where $\Delta^{(s)}$ is the filter width (equivalently, the downsampling factor) used to generate dataset s . In our pipeline, this multiplicative scaling is applied to the derivative features prior to z-score standardization during training, and the same convention is used at inference time.

During inference in LBM-LES, we run on the native lattice without downsampling, so $\Delta^{(s)} = 1$ and $S_{ij}^* = S_{ij}$ and $\omega_i^* = \omega_i$.

3.5.1. Network Architecture

The core network follows a hierarchical structure: an input linear layer mapping nine features to 64 hidden units, two hidden layers with $64 \rightarrow 64$ units, and an output linear layer mapping $64 \rightarrow 6$ units for the SGS stress components. We use GELU (Gaussian Error Linear Unit) [32] activations for all results reported here.

$$\text{GELU}(x) = x \cdot \Phi(x) = x \cdot \frac{1}{2} \left[1 + \text{erf} \left(\frac{x}{\sqrt{2}} \right) \right], \quad (16)$$

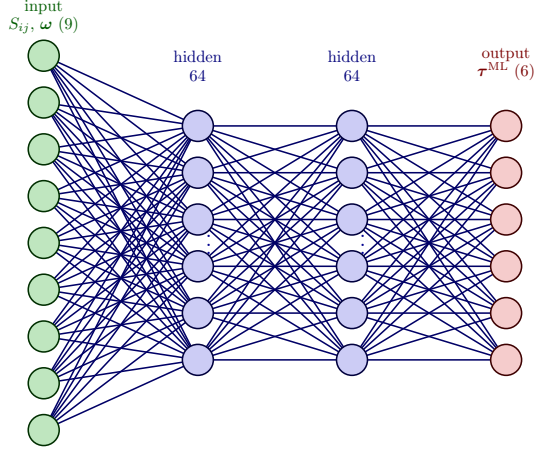


Figure 1: Neural-network architecture for SGS stress prediction (9→64→64→6). A deviatoric projection is applied post-inference.

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. The complete model contains approximately 5,190 trainable parameters: Layer 1 has 9×64 weights plus 64 bias parameters, Layer 2 has 64×64 weights plus 64 bias parameters, and the output head has 64×6 weights plus 6 bias parameters, for a total of $640 + 4,160 + 390 = 5,190$ trainable parameters.

3.6. Losses: data and physics

Training combines a data loss on the SGS stress with three physics-constrained losses used for the main results:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda_{\pi} \mathcal{L}_{\pi} + \lambda_{\text{eq}} \mathcal{L}_{\text{eq}} + \lambda_{\text{divF}} \mathcal{L}_{\text{divF}}. \quad (17)$$

Here \mathcal{L}_{MSE} is the mean-squared error on the six SGS stress components. The dissipation (energy transfer) channel is defined as

$$\Pi_{\text{pred}}^{(i)} = \mathbf{c}^{(i)T} \boldsymbol{\tau}_{\text{pred}}^{(i)}, \quad \mathbf{c}^{(i)} = [-S_{xx}^{(i)}, -S_{yy}^{(i)}, -S_{zz}^{(i)}, -2S_{xy}^{(i)}, -2S_{xz}^{(i)}, -2S_{yz}^{(i)}]^T. \quad (18)$$

The dissipation-matching loss is

$$\mathcal{L}_{\pi} = \text{MSE}(\Pi_{\text{pred}}, \Pi_{\text{true}}), \quad (19)$$

which preserves natural backscatter (no enforced non-negativity hinge). The equivariance loss \mathcal{L}_{eq} enforces rotational consistency under the discrete cube-rotation group, i.e. $\tau(Rx) \approx R\tau(x)R^T$. The divergence-free SGS-force loss $\mathcal{L}_{\text{divF}}$ penalizes $\nabla \cdot \mathbf{F}_{\text{sgs}}$ where $\mathbf{F}_{\text{sgs}} = -\nabla \cdot \boldsymbol{\tau}'$, encouraging numerical compatibility with the forcing formulation. Because $\mathcal{L}_{\text{divF}}$ requires spatial derivatives of the predicted stress, it is evaluated on full-grid snapshots rather than within the shuffled batch loop; we compute it once per epoch on a subset of training files and apply a separate gradient step, while the data loss, Π -matching, and equivariance loss are applied per batch. For the main-results model we use $\lambda_\pi = 0.4$, $\lambda_{\text{eq}} = 0.1$, and $\lambda_{\text{divF}} = 0.05$. No clipping or value-limiting is applied to SGS-stress targets or network outputs during training, nor to predicted stresses, Π , ν_t^{eff} , or the resulting SGS forcing during a posteriori rollouts.

4. Results

This section presents a priori validation, ablation studies, and a posteriori LES results across the parameter regimes and flow conditions considered. Unless otherwise stated, time is reported in normalized form as $t^* = t/t_0$ with $t_0 = L_0/U_0$ (L_0 domain length, $U_0 = 0.043$ reference velocity in lu; $t^* = \text{iter } U_0/N_x$ with $\Delta x = \Delta t = 1$). Statistics are computed from snapshots with $t^* \geq 20$ where the flow is statistically stationary (Fig. 9). All ML-based and Smagorinsky baseline results are for 128^3 at $\nu = 0.001$ (lu).

4.1. A priori validation

We assess model performance across the training parameter space using stress-correlation metrics and physical consistency checks.

As a primary a priori metric for stress prediction accuracy, we report the Pearson correlation coefficient between predicted and FD SGS stress components. For each independent component $k \in \{xx, yy, zz, xy, xz, yz\}$, we compute

$$\text{Corr}(\tau_{k,\text{pred}}, \tau_{k,\text{true}}) = \frac{\text{Cov}(\tau_{k,\text{pred}}, \tau_{k,\text{true}})}{\sigma_{\tau_{k,\text{pred}}} \sigma_{\tau_{k,\text{true}}}}. \quad (20)$$

We report correlations per component (and, when needed, aggregate across components).

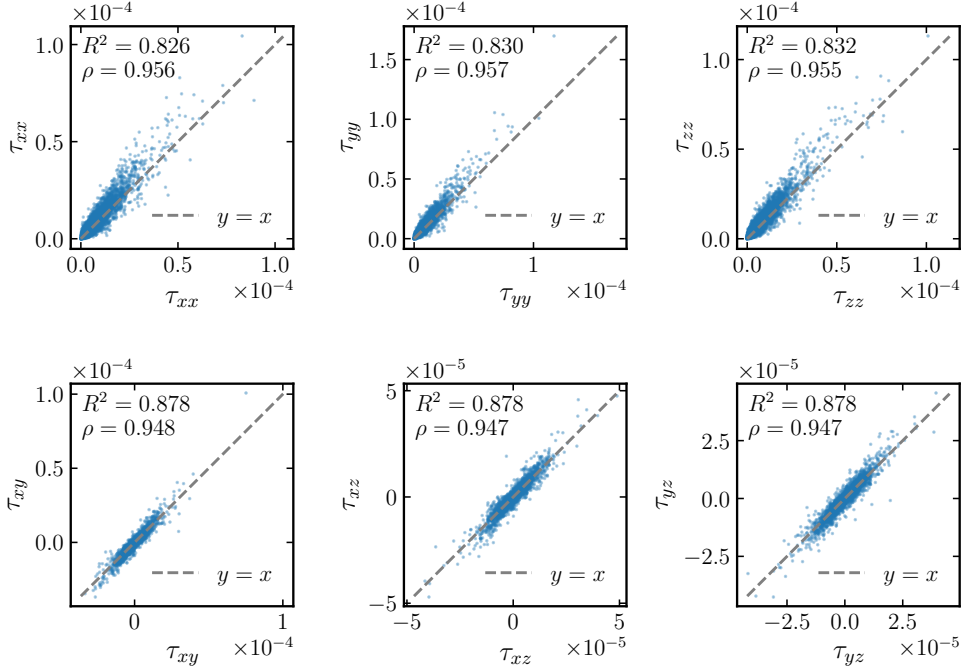


Figure 2: A priori validation scatter plots for SGS stress components. Each panel compares the predicted SGS stress component τ_{ij}^{pred} against the FD reference τ_{ij}^{true} over a validation dataset: FD grid 128^3 (256^3 DNS, $\Delta = 2$), $\nu = 0.001$. The dashed line indicates $y = x$. Reported values are the coefficient of determination R^2 and Pearson correlation ρ computed per component.

Figure 2 shows representative a priori scatter plots comparing predicted versus FD SGS stress components; annotations report R^2 and Pearson correlation ρ for each component.

To assess whether the model reproduces the distribution of SGS energy transfer and intermittent events [33], we also compare the probability density function (p.d.f.) of the SGS energy transfer proxy defined from the stress and strain-rate as

$$\Pi = -\tau_{ij}S_{ij}. \quad (21)$$

Figure 3 compares the p.d.f. of Π from the ML prediction and the FD reference. For completeness, we also include the a priori p.d.f. of the static Smagorinsky baseline (evaluated on the same FD fields). The negative tail corresponds to backscatter events (net transfer from subgrid to

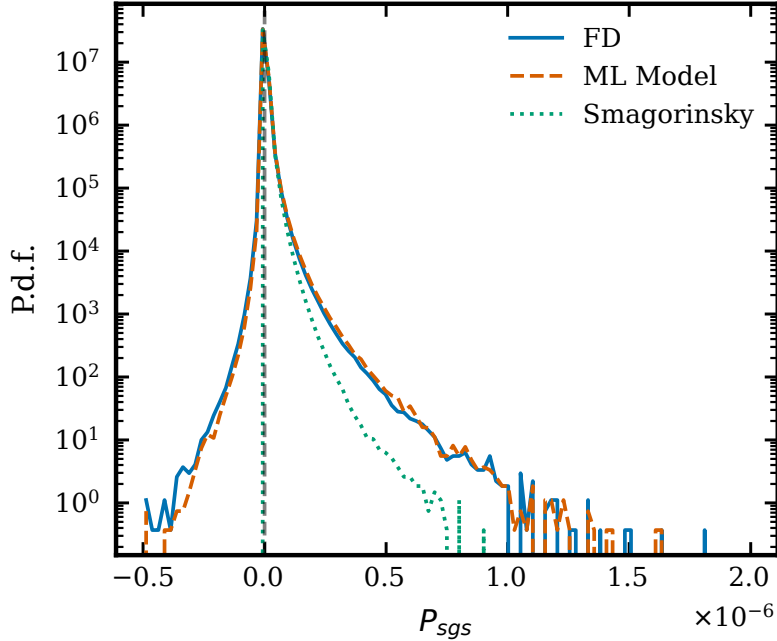


Figure 3: A priori p.d.f. of the SGS energy transfer Π comparing FD reference, ML prediction, and static Smagorinsky baseline (a priori). The ML model closely matches FD across the peak and tails, including backscatter (negative Π) and intermittent forward transfer (positive Π). For Smagorinsky, the non-negative eddy viscosity yields zero p.d.f. for $\Pi < 0$ and underpredicts the positive- Π tail.

resolved scales), while positive values correspond to forward transfer. The ML model closely matches the FD reference across the entire distribution, including the intermittent backscatter and forward-transfer tails. In contrast, the Smagorinsky model, by construction, enforces non-negative eddy viscosity and thus yields zero p.d.f. for $\Pi < 0$ and significantly underpredicts the probability of larger positive- Π events.

To quantify how stress-prediction accuracy varies across the training parameter space, we evaluate the model on each (Δ, ν) bucket of the full dataset from the training configurations (Table 1). Table 2 reports the coefficient of determination R^2 and macro-averaged Pearson correlation ρ per bucket. The Δ -normalization (Section 3.5) ensures comparable feature scaling across buckets. Performance is strongest for $(\Delta = 6, \nu = 0.005)$ and $(\Delta = 8, \nu = 0.004)$, with $R^2 \approx 0.80$ – 0.84 and $\rho \approx 0.90$ – 0.95 ; the $(\Delta = 4, \nu = 0.001)$ bucket, which contributes the largest sample count,

Table 2: R^2 and Pearson correlation ρ per (Δ, ν) bucket on the full dataset from the training configurations (Table 1). N is the number of samples per bucket.

Δ	ν	FD grid	N	R^2	ρ
4	0.001	64^3	13.4M	0.696	0.838
6	0.002	32^3	1.7M	0.710	0.844
6	0.005	32^3	1.7M	0.837	0.952
8	0.004	32^3	1.7M	0.804	0.904

yields $R^2 \approx 0.70$ and $\rho \approx 0.84$. This per-bucket breakdown confirms that the model achieves consistent accuracy across the full range of filter widths and viscosities used in training.

The results reveal a clear trend: accuracy improves with higher viscosity and coarser filter width. At fixed $\Delta = 6$, increasing ν from 0.002 to 0.005 raises R^2 from 0.71 to 0.84; similarly, $(\Delta = 8, \nu = 0.004)$ outperforms $(\Delta = 4, \nu = 0.001)$. The most challenging regime is the finest filter with lowest viscosity $(\Delta = 4, \nu = 0.001)$, which also contributes the largest sample count. This pattern is consistent with the expectation that more dissipative flows (higher ν) and coarser filters (larger Δ) yield more regular SGS structure and thus more predictable stress fields.

4.2. *A posteriori*

4.2.1. *Characterization and verification of the stress split*

Figure 4 summarizes both the structure and the numerical consistency of the proposed SGS closure during the coupled LES rollout. Panels (a) and (b) show that a substantial fraction of the learned SGS stress is not aligned with the strain-rate tensor and therefore cannot be represented by a purely eddy-viscosity-based model. Panel (e) confirms that the corresponding residual forcing is dynamically active. Panels (c) and (f) provide a discrete verification of the formulation: the residual stress is orthogonal to the strain rate and its explicit injection via forcing does not introduce net resolved-scale energy transfer, with both measures remaining at machine precision. Together, these results demonstrate a clean separation between dissipative and anisotropic SGS contributions.

4.2.2. *Robustness across collision operators*

To probe robustness with respect to the collision operator, we compare a posteriori statistics of Π obtained when the same trained network is coupled

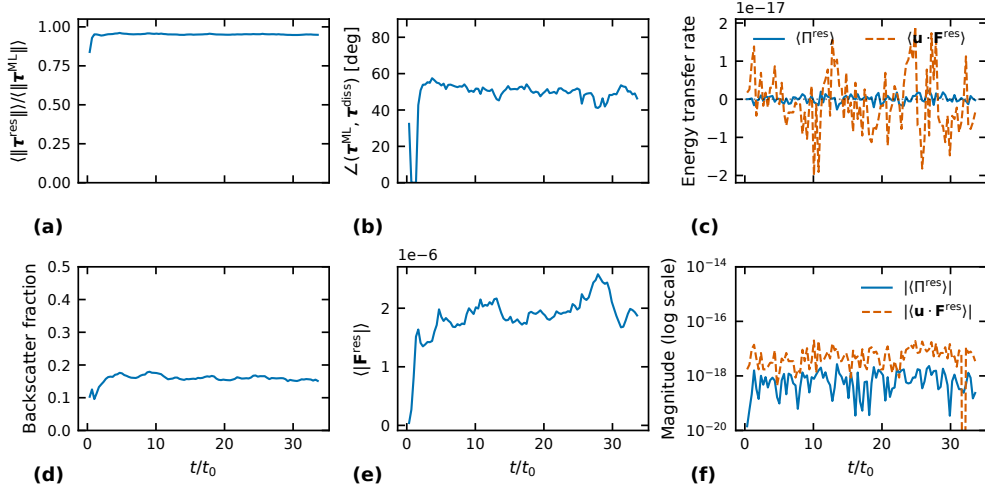


Figure 4: Characterization and verification of the hybrid SGS stress-splitting closure versus normalized time t/t_0 . (a) Residual fraction $\langle \|\boldsymbol{\tau}^{\text{res}}\| \rangle / \langle \|\boldsymbol{\tau}^{\text{ML}}\| \rangle$, quantifying the portion of the learned SGS stress not representable by an eddy-viscosity-aligned contribution. (b) Global stress-alignment proxy between the learned stress and its dissipative component, computed from domain-averaged inner products and magnitudes. (c) Energy consistency: residual dissipation $\Pi^{\text{res}} = -\tau_{ij}^{\text{res}} S_{ij}$ and resolved-scale work of the residual forcing $\langle \mathbf{u} \cdot \mathbf{F}^{\text{res}} \rangle$, with $\mathbf{F}^{\text{res}} = -\nabla \cdot \boldsymbol{\tau}^{\text{res}}$. (d) Backscatter fraction, defined as the fraction of grid cells where $\Pi = -\tau_{ij}^{\text{ML}} S_{ij} < 0$. (e) Mean magnitude of the residual forcing $\langle |\mathbf{F}^{\text{res}}| \rangle$, indicating the activity level of the explicit anisotropic SGS contribution. (f) Discrete orthogonality shown on a logarithmic scale; both $|\langle \Pi^{\text{res}} \rangle|$ and $|\langle \mathbf{u} \cdot \mathbf{F}^{\text{res}} \rangle|$ remain at numerical noise levels. Panels (c) and (f) demonstrate that the residual stress is non-dissipative and that its explicit injection via forcing is energy-consistent at the discrete level. Data from a representative FHIT rollout (128^3 , $\nu = 0.001$).

to LBM with either MRT or Bhatnagar–Gross–Krook (BGK) collision operators. Figure 5 shows that both configurations closely follow the FD reference, including the negative- Π tail (backscatter), even though the training dataset was generated from MRT-LBM DNS.

Figure 6 summarizes a posteriori probability density function (p.d.f.) comparisons computed from the coupled LES rollouts. Panel (a) compares the p.d.f. of a representative SGS stress component and shows that the ML-based closure closely follows the FD reference distribution, indicating that the rollout reproduces the typical stress magnitudes and intermittency level of the reference case. Panel (b) compares the p.d.f. of Π . The ML inference remains well aligned with the FD curve across the core of the distribution and into the tails. In contrast, the eddy-viscosity baselines (dynamic LES

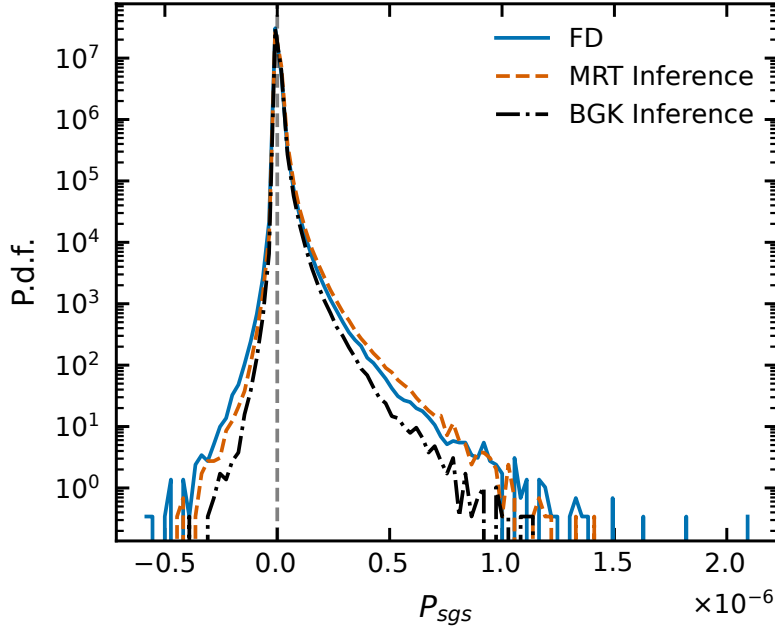


Figure 5: A posteriori p.d.f. of Π comparing FD reference with ML inference deployed in LBM with MRT and BGK collision operators. The BGK configuration remains statistically consistent with FD, indicating transferability beyond the collision model used to generate the training data.

and static Smagorinsky) deviate strongly from FD, with distributions biased toward large positive Π and an essentially absent negative- Π tail due to the strictly non-negative eddy viscosity ($\nu_t \geq 0$) used in the baseline implementations, which suppresses backscatter.

Figure 7 provides a spatial comparison of the SGS energy transfer Π at a representative z -slice (iteration 89,000). The top row compares ML-LES against the FD reference; the bottom row compares dynamic Smagorinsky against the same FD reference. The ML closure reproduces the spatial structure of Π (filamentary patterns, positive and negative regions) with smaller error than the dynamic baseline, whose error map shows larger and more widespread discrepancies.

Figure 8 compares a posteriori Π statistics for two training objectives. With the full objective (data loss plus physics losses; Eq. 17), the predicted Π p.d.f. aligns closely with FD. When all physics losses are disabled and training uses only the stress data loss \mathcal{L}_{MSE} , the model still reproduces the

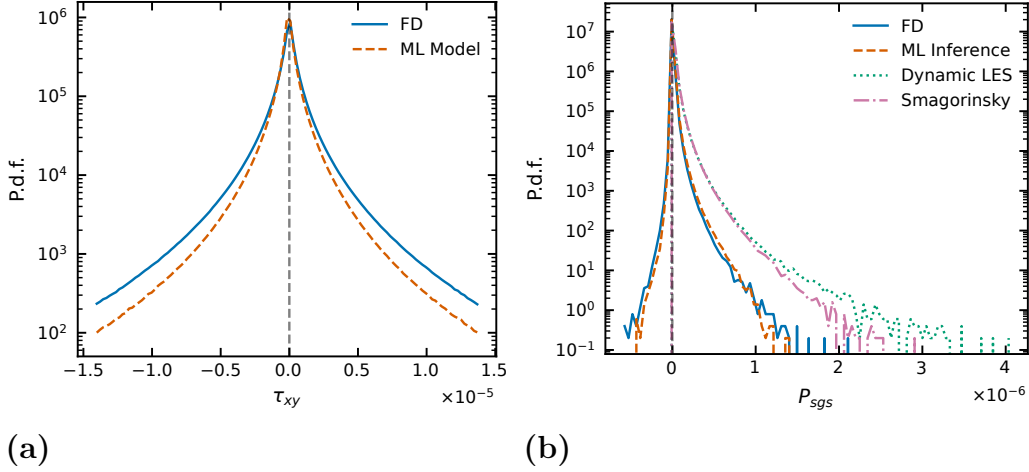


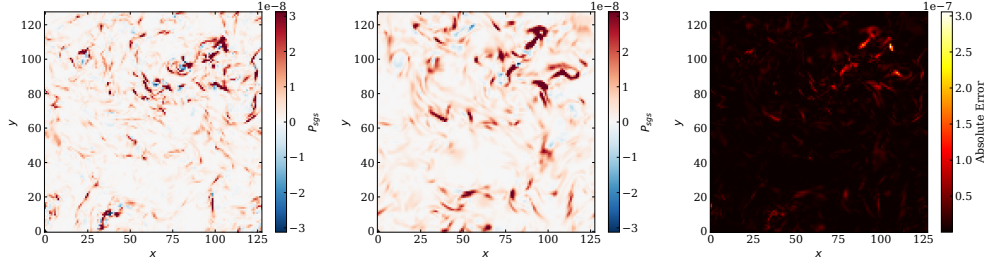
Figure 6: A posteriori p.d.f. comparisons: (a) representative SGS stress component distribution, and (b) p.d.f. of Π comparing FD, ML inference, and classical SGS baselines.

negative- Π tail (backscatter), but shows a noticeable deviation for positive Π , indicating overly strong forward-transfer events.

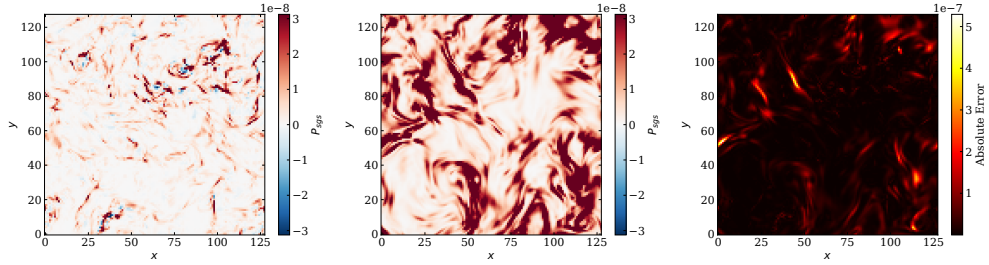
To assess whether the closures reproduce correct global energetics in closed-loop LES, we monitor the instantaneous energy-budget mismatch based on the volume-averaged kinetic-energy equation, i.e. the deviation of ε from the balance $-dK/dt + \langle \mathbf{f} \cdot \mathbf{u} \rangle$. Figure 9 shows that the ML-coupled LES remains significantly closer to the DNS reference (256^3) than the Smagorinsky baselines on the 128^3 grid. In particular, the dynamic Smagorinsky case exhibits the largest imbalance, indicating that its viscosity-based dissipation does not track the resolved energy evolution as accurately as the learned explicit-stress closure.

The 128^3 rollouts above extend beyond the exact training configuration: the training data use FD grids of size 64^3 and 32^3 (Table 1), whereas inference runs on the native LES grid at 128^3 without downsampling. To probe stability at even larger resolutions, we run ML inference on a 512^3 grid. After an initial transient, the energy-balance error stabilizes around 10–12% (see Appendix C), indicating that the block-based inference (Algorithm 1) remains numerically stable and maintains approximate energy consistency at production-scale resolutions.

As a preliminary a posteriori test of transfer beyond FHIT, we deploy the same trained model in turbulent channel flow; the results are reported in Appendix B.



(a) ML-LES vs FD



(b) Dynamic Smagorinsky vs FD

Figure 7: Spatial slice comparison of SGS energy transfer Π at z -slice 64/128, iteration 89,000 (time step index). In both panels, from left to right: FD (ground truth), model prediction, and absolute error map. (a) ML-LES vs FD. (b) Dynamic Smagorinsky vs FD. The ML closure reproduces the spatial structure of Π with smaller error than the dynamic baseline.

5. Conclusion

We presented a physics-constrained ML-based SGS closure for LBM LES that predicts the SGS stress tensor from local derivative features and couples it through a dissipative/residual split. The validation reported in Sections 4 and Appendix B suggests that the closure improves upon Smagorinsky baselines in FHIT (a priori and a posteriori) and remains consistent across MRT and BGK collision operators, transfers to channel flow without retraining, and achieves throughput comparable to a dynamic Smagorinsky baseline when deployed via ONNX.

The main contribution is to show that an explicit-stress ML closure can be integrated into LBM in a practical and stable way, with the split mechanism enabling non-dissipative SGS effects that viscosity-only closures suppress. Kang *et al.* [34] established that some cross-flow transfer from FHIT to channel flow is possible, though only partially in their formulation. Our

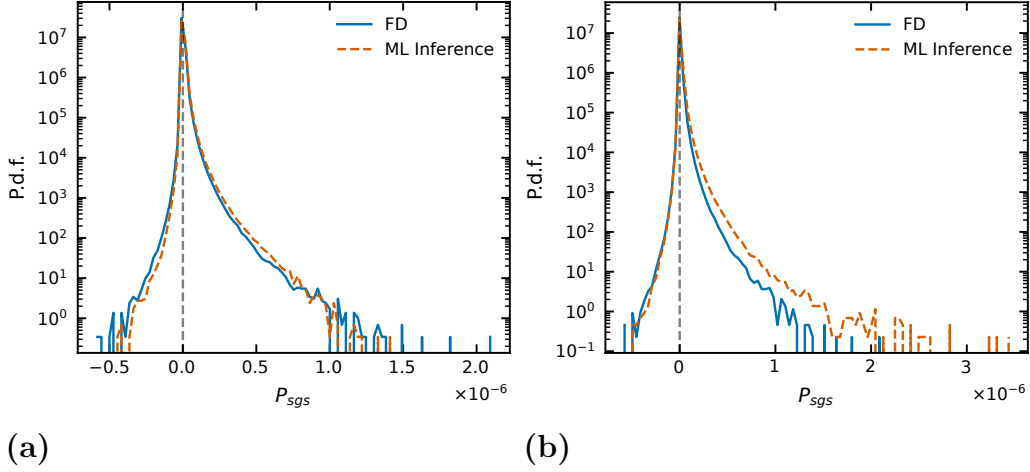


Figure 8: A posteriori Π p.d.f. from LES rollouts for two ML variants: (a) with physics loss, (b) without physics loss.

LBM results show transfer in a formulation that uses Δ -normalized features and a dissipative/residual stress split (Appendix B).

Validation is limited to FHIT (primary) and a preliminary channel-flow transfer test. Future work should clarify the mechanism behind the observed transfer through a rigorous analysis of feature-space overlap and prediction accuracy across flow regimes, extend the assessment to additional flow classes and Reynolds numbers, report skin-friction metrics for channel flow, and explore multi-GPU scaling for larger problem sizes.

Appendix A. D3Q19 velocity set and equilibrium distribution

For completeness, we summarize the discrete-velocity stencil and standard second-order equilibrium used by the D3Q19 LBM referenced in Section 3. The D3Q19 velocity set consists of 19 discrete velocity vectors

$$\mathbf{c}_\alpha = c (c_{\alpha x}, c_{\alpha y}, c_{\alpha z})^T, \quad (\text{A.1})$$

where $c = \Delta x / \Delta t$ is the lattice speed and $(c_{\alpha x}, c_{\alpha y}, c_{\alpha z})$ are the dimensionless direction vectors. The velocity vectors are

$$\mathbf{c}_\alpha = c \begin{cases} (0, 0, 0) & \alpha = 1, \\ (\pm 1, 0, 0), (0, \pm 1, 0), (0, 0, \pm 1) & \alpha = 2, \dots, 7, \\ (\pm 1, \pm 1, 0), (\pm 1, 0, \pm 1), (0, \pm 1, \pm 1) & \alpha = 8, \dots, 19. \end{cases} \quad (\text{A.2})$$

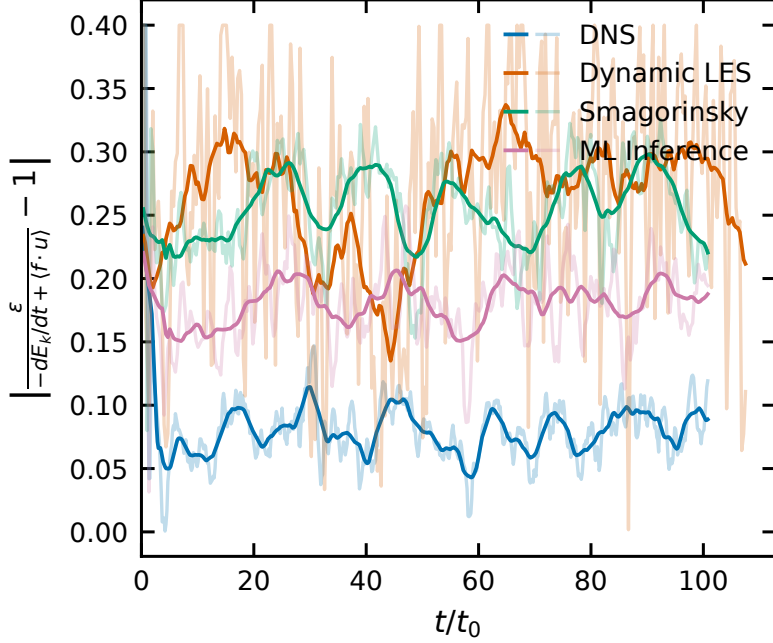


Figure 9: Energy-balance error versus normalized time t/t_0 : $|\varepsilon/(-dK/dt + (\mathbf{f} \cdot \mathbf{u})) - 1|$ for DNS (256^3) and LES rollouts (128^3). ML inference stays closest to DNS, while dynamic Smagorinsky shows the largest imbalance.

The corresponding lattice weights are $w_0 = 1/3$ (rest), $w_s = 1/18$ (axial), and $w_d = 1/36$ (diagonal), and the lattice sound speed satisfies $c_s^2 = c^2/3$.

The equilibrium distribution is taken in the standard second-order form

$$f_\alpha^{\text{eq}} = w_\alpha \rho \left[1 + \frac{\mathbf{c}_\alpha \cdot \mathbf{u}}{c_s^2} + \frac{(\mathbf{c}_\alpha \cdot \mathbf{u})^2}{2c_s^4} - \frac{\mathbf{u} \cdot \mathbf{u}}{2c_s^2} \right]. \quad (\text{A.3})$$

All simulations reported in this work are performed in a cubic domain with periodic boundary conditions in all three directions.

Appendix B. Transfer to channel flow

We perform an a posteriori test by deploying the same trained model (Section 3, Table 1) in turbulent channel flow at $\text{Re}_\tau \approx 160$. The model was trained exclusively on forced homogeneous isotropic turbulence (FHIT); applying it without retraining to a wall-bounded flow with strong mean shear tests transfer to flow configurations not seen during training—a recognized

challenge in ML-based LES [15]. Unlike prior physics-informed LBM work for wall-bounded flows, which required neural networks trained on near-wall data [24], our bulk-trained closure is deployed directly with wall boundary conditions. The simulation uses the hybrid ML–LES coupling on a $1536 \times 64 \times 128$ grid (streamwise \times spanwise \times wall-normal), viscosity $\nu = 0.001$, and uniform streamwise forcing; statistics are time-averaged after an initial transient.

Kang *et al.* [34] showed that transfer from FHIT-trained artificial neural network (ANN) closures to channel flow is possible but limited, and argued that this depends importantly on the model inputs, especially the inclusion of the Leonard stress L_{ij} . In contrast, our LBM closure transfers without retraining, suggesting that generalization depends not only on training data but also on representation, normalization, and solver coupling. Our ablation without Δ -normalization gave poorer inference results, suggesting that, in our formulation, Δ -normalization contributes to making inputs more comparable across filter widths and flow regimes (Section 3). One interpretation is that the closure depends only on local gradient information—the strain-rate and vorticity at each point—and does not require knowledge of the global flow geometry or wall position. It therefore approximates a local constitutive relation between filtered derivatives and SGS stress, rather than a flow-specific model. Transfer may then occur wherever the distribution of these local features in channel flow overlaps with that encountered in FHIT, despite the different large-scale flow structure. This interpretation remains speculative; rigorous support would require a quantitative analysis of feature-space overlap and its relation to prediction accuracy.

Figure B.10 compares the resulting mean velocity, RMS fluctuation, and Reynolds shear stress profiles against the DNS benchmark of Kim *et al.* [35] at $\text{Re}_\tau \approx 180$; our simulation is at $\text{Re}_\tau \approx 160$. Panel (a) shows the mean streamwise velocity u^+ versus y^+ ; the ML-coupled LES shows close agreement with the benchmark and captures both the viscous sublayer and the logarithmic region despite the slight Reynolds-number mismatch. Panel (b) shows the RMS fluctuations u_{rms}^+ , v_{rms}^+ , and w_{rms}^+ , with good agreement across all three components. Panel (c) shows the normalized Reynolds shear stress $-\langle u'w' \rangle^+$ versus y^+ (w wall-normal in our convention); the ML closure reproduces the benchmark profile closely across the near-wall, buffer, and logarithmic regions.

Appendix C. Stability at production-scale resolution

Figure C.11 shows the energy-balance error for ML inference at 512^3 , referenced in Section 4. The error stabilizes after the transient, supporting stable deployment at production-scale resolutions.

Appendix D. Implementation and computational performance

The ML model is deployed in a production Fortran/OpenACC MRT-LBM solver using ONNX Runtime [36]. We summarize the inference pathway at a high level for reproducibility. At each time step, the solver computes a compact feature vector at every lattice site and evaluates the trained network to obtain the SGS stress used by the ML-LES closure. The deployed model uses nine macroscopic derivative features: the six independent strain-rate components and three vorticity components, ordered as S_{xx} , S_{yy} , S_{zz} , S_{xy} , S_{xz} , S_{yz} , and ω_x , ω_y , ω_z . The derivatives are evaluated on the periodic lattice using centered finite differences, and the same preprocessing convention as training is applied through a filter-width scaling based on the downsampling factor Δ . In the coupled LES rollouts shown in this paper we run on the native lattice so $\Delta = 1$.

Inference is executed through an ONNX Runtime session created once at initialization and reused across time steps. The exported ONNX model embeds z-score normalization of features and denormalization of targets in the graph, so the solver passes raw feature values and receives physical stress values directly. The feature tensor is assembled in Fortran and passed to the inference engine; for GPU execution, the implementation avoids repeated allocations by reusing persistent buffers. For larger problem sizes, the inference call can be executed in batches to cap peak memory while maintaining throughput.

The network outputs the six independent components of the SGS stress tensor. Before coupling to the LBM update, we enforce a deviatoric projection by removing the trace of the predicted stress so that the injected stress is traceless by construction; the original (pre-projection) outputs can optionally be saved for post-processing.

Algorithm 1 summarizes the per-time-step procedure for reproducibility.

Table D.3: End-to-end performance comparison for a representative rollout to $t/t_0 = 40$.

Closure	Elapsed time (s)	MLUP/s	Relative diff. (MLUP/s)
Dynamic Smagorinsky LES	2929.319	71.592	–
ML-LES (ONNX inference)	2947.285	71.155	–0.61%

Algorithm 1 ML-LES inference and coupling (per time step)

Require: Resolved velocity \mathbf{u} , density ρ ; molecular viscosity ν_0 ; filter width Δ (unity on native lattice)

Ensure: Updated distribution functions f_α via MRT collision with ML-based closure

- 1: **1. Compute features** at all lattice sites:
 - 2: Using velocity $\tilde{\mathbf{u}} = \mathbf{u} + \frac{1}{2}\rho^{-1}\mathbf{F}_{\text{ext}}$ (Guo half-force correction [7]), evaluate strain-rate $S_{ij} = \frac{1}{2}(\partial_j \tilde{u}_i + \partial_i \tilde{u}_j)$ and vorticity $\boldsymbol{\omega} = \nabla \times \tilde{\mathbf{u}}$ using centered finite differences
 - 3: Apply Δ -normalization: $S_{ij}^* \leftarrow \Delta S_{ij}$, $\omega_i^* \leftarrow \Delta \omega_i$
 - 4: Assemble feature vector $\mathbf{x} = (S_{xx}, S_{yy}, S_{zz}, S_{xy}, S_{xz}, S_{yz}, \omega_x, \omega_y, \omega_z)$
 - 5: **2. Run network inference** to obtain raw stress $\boldsymbol{\tau}^{\text{raw}}$ ($N =$ number of lattice sites):
 - 6: **if** $N > 200,000$ **then**
 - 7: run in batches of size 200,000
 - 8: **else if** $N > 10,000$ **then**
 - 9: run in batches of size 50,000
 - 10: **else**
 - 11: run full-field inference
 - 12: **end if**
 - 13: **3. Deviatoric projection:** $\tau_{ij}^{\text{ML}} \leftarrow \tau_{ij}^{\text{raw}} - \frac{1}{3}\tau_{kk}^{\text{raw}}\delta_{ij}$
 - 14: **4. Stress split** (at each lattice site):
 - 15: $\Pi \leftarrow -\tau_{ij}^{\text{ML}}S_{ij}$; $|\mathcal{S}|^2 \leftarrow 2S_{ij}S_{ij}$
 - 16: $\nu_t^{\text{eff}} \leftarrow \Pi/|\mathcal{S}|^2$; $\nu_e \leftarrow \nu_0 + \nu_t^{\text{eff}}$
 - 17: $s_\nu \leftarrow 1/(\frac{1}{2} + 3\nu_e)$
 - 18: $\tau_{ij}^{\text{diss}} \leftarrow -2\nu_t^{\text{eff}}S_{ij}$; $\tau_{ij}^{\text{res}} \leftarrow \tau_{ij}^{\text{ML}} - \tau_{ij}^{\text{diss}}$
 - 19: **5. Coupling:**
 - 20: $\mathbf{F}^{\text{res}} \leftarrow -\nabla \cdot \boldsymbol{\tau}^{\text{res}}$ (centered differences, periodic BCs)
 - 21: $\mathbf{F}_{\text{total}} \leftarrow \mathbf{F}_{\text{ext}} + \mathbf{F}^{\text{res}}$
 - 22: Apply MRT collision (Eq. 13) with $\boldsymbol{\Lambda}^{\text{ML}}(s_\nu)$ and $\mathbf{m}^{\mathbf{F}_{\text{total}}}$
-

Table D.3 reports end-to-end timing for a representative rollout to $t/t_0 = 40$ for a dynamic Smagorinsky LES baseline and the ONNX-coupled ML-LES variant, using the same executable and hardware configuration (NVIDIA A100 80 GB GPU; ONNX Runtime GPU v1.16.3; CUDA v12.2). The measured throughput differs by $\approx 0.6\%$.

The wall-clock timing in Table D.3 includes both feature computation and inference. In our tested configuration, the ML-coupled rollout achieves

throughput comparable to the dynamic Smagorinsky baseline.

Declaration of competing interest

The authors declare no competing interests.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments

The authors wish to acknowledge the support of the National Center for HPC, Big Data and Quantum Computing, Project CN_00000013—CUP E83C22003230001, Mission 4 Component 2 Investment 1.4, funded by the European Union—NextGenerationEU; the support of Project PRIN 2022F422R2—CUP E53D23003210006, financed by the European Union—Next Generation EU; the support of Project PRIN PNRR P202298P25—CUP E53D23016990001, financed by the European Union—Next Generation EU; and the support of Project ECS 0000024 Rome Technopole—CUP B83C22002820006, NRP Mission 4 Component 2 Investment 1.5, funded by the European Union—NextGenerationEU.

We used KI-TURB 3D [37] for literature and code search during manuscript preparation.

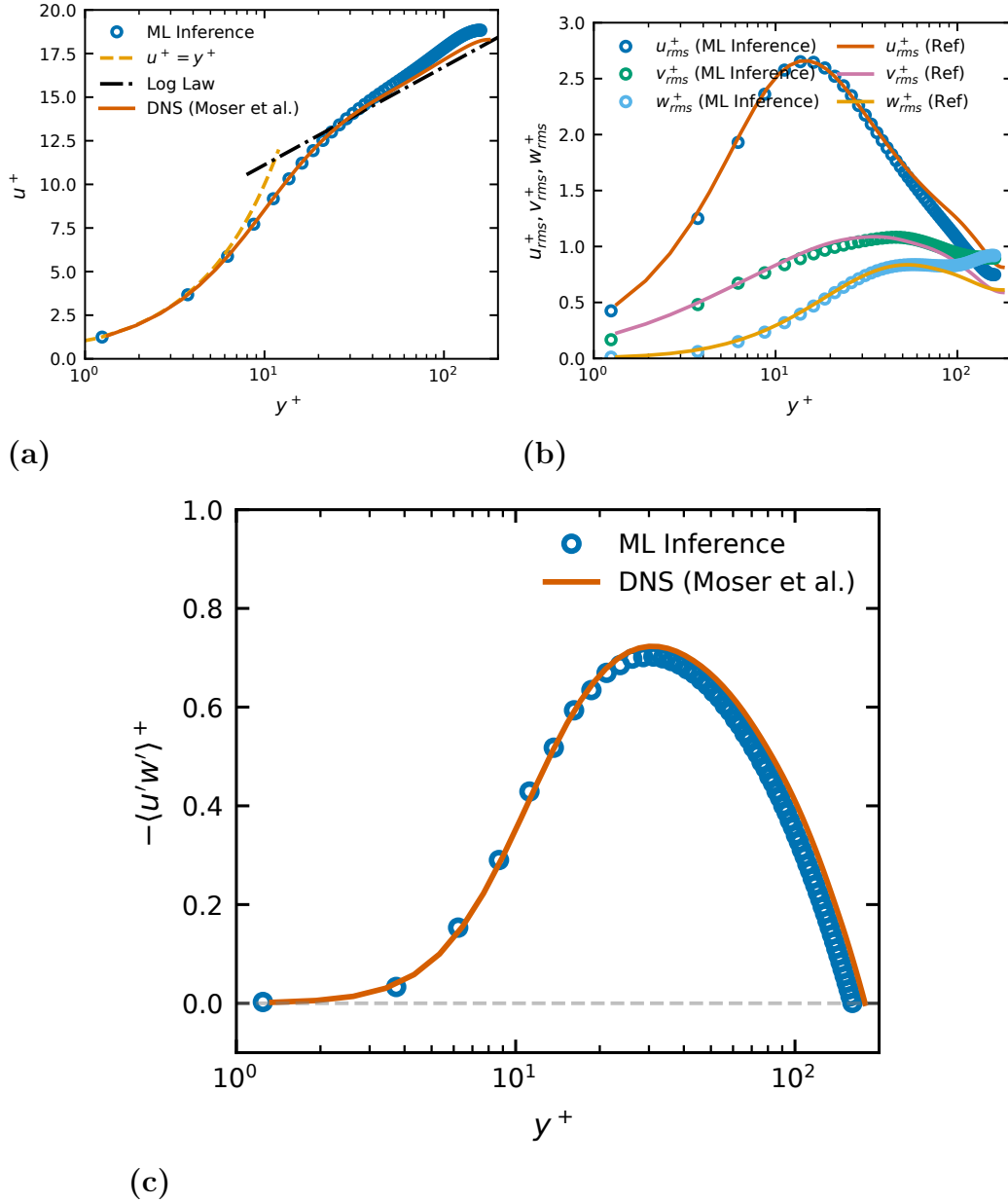


Figure B.10: A posteriori channel flow at $Re_\tau \approx 160$ (ML simulation) vs DNS benchmark at $Re_\tau \approx 180$ [35]: (a) mean streamwise velocity u^+ versus y^+ ; (b) RMS fluctuations u_{rms}^+ , v_{rms}^+ , w_{rms}^+ ; (c) Reynolds shear stress $-(u'w')^+$ versus y^+ (w wall-normal).

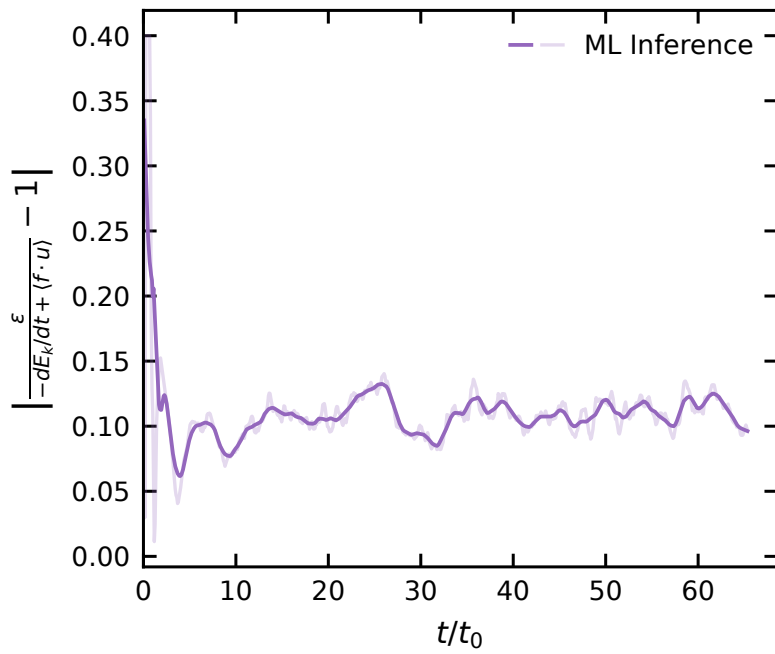


Figure C.11: Energy-balance error $|\varepsilon/(-dK/dt + \langle \mathbf{f} \cdot \mathbf{u} \rangle) - 1|$ for ML inference at 512^3 ($\nu = 0.001$, $t/t_0 \approx 60$). Error stabilizes after the transient.

References

- [1] R. Fang, D. Sondak, P. Protopapas, S. Succi, Neural network models for the anisotropic reynolds stress tensor in turbulent channel flow, *Journal of Turbulence* 21 (9-10) (2020) 525–543. doi:10.1080/14685248.2019.1706742.
- [2] A. Beck, D. Flad, C.-D. Munz, Deep neural networks for data-driven les closure models, *Journal of Computational Physics* 398 (2019) 108910.
- [3] G. Novati, H. L. De Laroussilhe, P. Koumoutsakos, Automating turbulence modelling by multi-agent reinforcement learning, *Nature Machine Intelligence* 3 (1) (2021) 87–96.
- [4] H. J. Bae, P. Koumoutsakos, Scientific multi-agent reinforcement learning for wall-models of turbulent flows, *Nature Communications* 13 (1) (2022) 1443.
- [5] P. Fischer, S. Kaltenbach, S. Litvinov, S. Succi, P. Koumoutsakos, Optimal lattice boltzmann closures through multi-agent reinforcement learning, arXiv preprint arXiv:2504.14422 (2025).
- [6] H. Chen, S. Kandasamy, S. Orszag, R. Shock, S. Succi, V. Yakhot, Extended boltzmann kinetic equation for turbulent flows, *Science* 301 (5633) (2003) 633–636. doi:10.1126/science.1085048.
- [7] Z. Guo, C. Zheng, B. Shi, Discrete lattice effects on the forcing term in the lattice boltzmann method, *Physical review E* 65 (4) (2002) 046308.
- [8] J. Smagorinsky, General circulation experiments with the primitive equations: I. the basic experiment, *Monthly Weather Review* 91 (3) (1963) 99–164.
- [9] M. Germano, U. Piomelli, P. Moin, W. H. Cabot, A dynamic subgrid-scale eddy viscosity model, *Physics of Fluids A: Fluid Dynamics* 3 (7) (1991) 1760–1765.
- [10] D. K. Lilly, A proposed modification of the germano subgrid-scale closure method, *Physics of Fluids A: Fluid Dynamics* 4 (3) (1992) 633–635.

- [11] J. Bardina, J. H. Ferziger, W. C. Reynolds, Improved subgrid scale models for large eddy simulation, in: Proceedings of the 13th Fluid and Plasma Dynamics Conference, AIAA, 1980.
- [12] R. A. Clark, J. H. Ferziger, W. C. Reynolds, Evaluation of subgrid-scale models using an accurately simulated turbulent flow, *Journal of Fluid Mechanics* 91 (1) (1979) 1–16.
- [13] C. Meneveau, J. Katz, Scale-invariance and turbulence models for large-eddy simulation, *Annual Review of Fluid Mechanics* 32 (1) (2000) 1–32.
- [14] R. Bose, A. M. Roy, Accurate deep learning sub-grid scale models for large eddy simulations, *arXiv preprint arXiv:2307.10060* (2023).
- [15] H. Choi, C. Cho, M. Kim, J. Park, Perspective on machine-learning-based large-eddy simulation, *Physical Review Fluids* 10 (11) (2025) 110701.
- [16] J. Park, H. Choi, Toward neural-network-based large eddy simulation: Application to turbulent channel flow, *Journal of Fluid Mechanics* 914 (2021) A16.
- [17] C. Xie, J. Wang, H. Li, M. Wan, S. Chen, Modeling subgrid-scale force and divergence of heat flux of compressible isotropic turbulence by artificial neural network, *Physical Review Fluids* 5 (5) (2020) 054605.
- [18] L. Yang, X. Zhang, S. Wang, Z. Lu, K. Zhang, D. Li, A neural-network-based mixed model of the subgrid-scale stress for large-eddy simulation of forced isotropic turbulence, *Physics of Fluids* 37 (2) (2025).
- [19] H. Chen, S. A. Orszag, I. Staroselsky, S. Succi, Expanded analogy between boltzmann kinetic theory of fluids and turbulence, *Journal of Fluid Mechanics* 519 (2004) 301–314. doi:10.1017/S0022112004001211.
- [20] K. N. Premnath, M. J. Pattison, S. Banerjee, Dynamic subgrid scale modeling for turbulent flows using lattice-boltzmann method, *Physica A: Statistical Mechanics and its Applications* 388 (13) (2009) 2640–2658.
- [21] Y. Zhang, H. Chen, Z. Guo, J. D. Sterling, Discrete lattice effects in large-eddy simulation, *Physical Review E* 77 (4) (2018) 046709.

- [22] H. Zhang, H. Hu, F. Zhang, X. Chen, Subgrid-scale model for large eddy simulations of incompressible turbulent flows within the lattice boltzmann framework, *Physical Review E* 110 (4) (2024) 045305.
- [23] M. C. Bedrunka, T. Horstmann, B. Picard, D. Reith, H. Foysi, Machine-learning-enhanced collision operator for the lattice boltzmann method based on invariant networks, *Physical Review E* 112 (5) (2025) 055308.
- [24] X. Xue, S. Wang, H.-D. Yao, L. Davidson, P. V. Coveney, Physics informed data-driven near-wall modelling for lattice boltzmann simulation of high reynolds number turbulent flows, *Communications Physics* 7 (1) (2024) 338.
- [25] X. Xue, M. F. P. ten Eikelder, M. Gao, X. Cheng, Y. Yang, Y. He, S. Wang, S. Cheng, Y. Hu, P. V. Coveney, Fast-forward lattice boltzmann: Learning kinetic behaviour with physics-informed neural operators, arXiv preprint arXiv:2509.22411 (2025).
- [26] A. Corbetta, A. Gabbana, V. Gyrya, D. Livescu, J. Prins, F. Toschi, Toward learning lattice boltzmann collision operators, *The European Physical Journal E* 46 (3) (2023) 10.
- [27] G. Ortali, A. Gabbana, I. Atmodimedjo, A. Corbetta, Enhancing lattice kinetic schemes for fluid dynamics with lattice-equivariant neural networks, *AIAA Journal* 63 (2) (2025) 716–731.
- [28] S. B. Pope, *Turbulent Flows*, Cambridge University Press, 2000.
- [29] D. d’Humières, Multiple-relaxation-time lattice boltzmann models in three dimensions, *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 360 (1792) (2002) 437–451.
- [30] M. I. Khan, S. Succi, G. Falcucci, Validating the boltzmann approach to the large-eddy simulations of forced homogeneous incompressible turbulence, *International Journal of Modern Physics C* (2025). doi: 10.1142/S0129183127500641.
URL <https://doi.org/10.1142/S0129183127500641>

- [31] A. L. Kupershtokh, D. Medvedev, D. Karpov, On equations of state in a lattice boltzmann method, *Computers & Mathematics with Applications* 58 (5) (2009) 965–974.
- [32] D. Hendrycks, Gaussian error linear units (gelus), arXiv preprint arXiv:1606.08415 (2016).
- [33] V. Borue, S. A. Orszag, Local energy flux and subgrid-scale statistics in three-dimensional turbulence, *Journal of Fluid Mechanics* 366 (1998) 1–31.
- [34] M. Kang, Y. Jeon, D. You, Neural-network-based mixed subgrid-scale model for turbulent flow, *Journal of Fluid Mechanics* 962 (2023) A38. doi:10.1017/jfm.2023.260.
- [35] J. Kim, P. Moin, R. Moser, Turbulence statistics in fully developed channel flow at low reynolds number, *Journal of Fluid Mechanics* 177 (1987) 133–166.
- [36] ONNX Runtime developers, Onnx runtime, <https://onnxruntime.ai/>, version: 1.16.3 (2021).
- [37] M. I. Khan, O. Contributors, Ki-turb 3d: Turbulence analysis and visualization suite, computer software (2025). URL <https://github.com/IdreesKhan3/ki-turb-3d>