



Research Article

TransportBench: A comprehensive benchmark for non-equilibrium flow transport

Xu Wang^{a,b,#}, Minghao Li^{a,#}, Qizhen Hong^a, Yang Liu^a, Chen-an Zhang^a, Shuai Zhang^a,
Wenhao Li^{a,c}, Yonghao Zhang^{a,c}, Tianbai Xiao^{a,c,*}

^a Institute of Mechanics, Chinese Academy of Sciences, Beijing 100190, China

^b School of Future Technology, University of Chinese Academy of Sciences, Beijing 101408, China

^c School of Engineering Science, University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Keywords:

Computational fluid dynamics
Rarefied gas dynamics
High-temperature gas dynamics
Non-equilibrium transport
Scientific machine learning

ABSTRACT

Scientific machine learning models, as versatile tools for numerical simulation and analysis, are increasingly transforming the landscape of fluid mechanics research. However, existing datasets and benchmarks are primarily limited to continuum fluids and provide limited support for non-equilibrium transport phenomena. To address this gap, we present TransportBench, a high-fidelity dataset and standardized benchmark for non-equilibrium flow transport, designed to reveal the strengths and limitations of neural network models across diverse flow regimes.

Specifically, the dataset encompasses a broad physical spectrum, covering continuum and rarefied regimes, low-speed and hypersonic flows, inert and chemically reactive gases, and both translational and internal-energy non-equilibrium effects. Built upon this dataset, we systematically benchmark representative neural architectures using unified evaluation protocols to probe key challenges in learning non-equilibrium flows, including robustness to shock-dominated discontinuities and multi-scale effects, as well as generalization across geometry and physical parameters.

Numerical results demonstrate that model performance exhibits a pronounced dependence upon the specific flow characteristics. No single architecture consistently performs best for all the tasks. Instead, different models exhibit distinct performance patterns, which may be related to their architectural priors when capturing smooth flow fields, shock-induced discontinuities, and high-order non-equilibrium statistics. By jointly providing the non-equilibrium flow dataset and model benchmark, TransportBench offers a new testbed for the development, evaluation, and diagnosis of scientific machine learning methods for fluid transport beyond the Navier-Stokes hydrodynamics. The benchmark datasets² and implementation codes³ are available under the MIT license.

1. Introduction

Accurate simulation of non-equilibrium flow transport has become increasingly important as fluid systems are pushed toward more extreme spatial and thermodynamic regimes. In hypersonic flight and atmospheric re-entry [1], gases may experience strong compression, shock heating, rarefaction, and internal-energy excitation; in micro/nano-electro-mechanical systems (MEMS/NEMS) [2], the characteristic length can become comparable to the molecular mean free path. These conditions give rise to pronounced non-equilibrium effects, under which the continuum assumption may break down, and the classical Navier-Stokes equations become insufficient for accurately describing the underlying transport phenomena. A more faithful description

requires statistical mechanics and gas-kinetic formulations such as the Boltzmann equation [3], or direct-modeling particle methods such as the direct simulation Monte Carlo (DSMC) method [4]. Although these approaches provide reliable descriptions of non-equilibrium fluid dynamics, their high computational cost, slow statistical convergence, and inherent stochastic noise make them difficult to use for rapid design, uncertainty analysis, and repeated parametric studies. This creates a strong demand for efficient surrogate modeling and simulation acceleration methods tailored to non-equilibrium flow transport.^{1,2}

Scientific machine learning is emerging as a promising paradigm for addressing these computational challenges by combining data-driven approximation with physical modeling [5,6]. However, the systematic evaluation of such methods remains largely constrained by benchmarks designed for continuum and near-equilibrium flow problems. Widely

* Corresponding author.

E-mail address: txiao@imech.ac.cn (T. Xiao).

These authors contributed equally to this work.

¹ <https://huggingface.co/datasets/CFDML/TransportBench>.

² <https://github.com/CFDML/TransportBench>.

Nomenclature

$\mathbf{x}$	Space variable (x, y, z)
$\mathbf{v}$	Particle velocity variable (u, v, w)
t	Time variable
f	Particle distribution function
s	Species index
l	Internal state index
Q	Collision operator in kinetic equation
ρ	Macroscopic density
$\mathbf{u}$	Macroscopic velocity
E	Macroscopic energy
$\mathbf{U}$	Conservative variables of flow field
$\mathbf{P}$	Pressure tensor
$\mathbf{q}$	Heat flux
$\mathbf{v}_{\text{pec}}$	Peculiar velocity
p	Macroscopic pressure
φ	Mass fraction
h	Enthalpy
S	Source term
N_s	Total number of chemical species
V_s	Number of vibrational states associated with the s th species
D_s	Diffusion coefficient of the s th species
η	Thermal conductivity associated with translational-rotational energy
μ	Shear viscosity
$\mathcal{A}$	Operating conditions
$\mathcal{U}$	Solution field
$\mathcal{G}$	Machine learning model
$\mathcal{L}$	Loss function
$\mathbf{b}$	Class-shape transformation (CST) coefficient
$\mathcal{U}$	Uniform distribution
r	Random variable sampled from $\mathcal{U}$
c	Airfoil chord
Ma	Mach number
Kn	Knudsen number
T	Macroscopic temperature
X	Horizontal or streamwise spatial coordinate
Y	Vertical or transverse spatial coordinate
Re	Reynolds number
ω	Viscosity index
d_{ref}	Diameter at reference temperature
$\mathcal{T}$	Non-linear logarithmic transformation
$\mathbf{M}$	Binary geometry mask
a	Scalar field
N	Number of spatial grid points
MSE	Mean squared error
MAE	Mean absolute error

used datasets, such as PDEBench [7] and FlowBench [8], primarily consider the Navier-Stokes-level dynamics close to thermodynamic equilibrium, in which the flow fields are often smooth, the output variables are mostly low-order macroscopic quantities, and the computational domains are regular or only mildly complex. They do not fully reflect the defining challenges of non-equilibrium transport, including rarefaction effects, Knudsen layers, high-order moment quantities, strong shock discontinuities, and multi-scale kinetic-to-continuum behavior. Consequently, strong performance on existing continuum-flow benchmarks does not necessarily indicate that a model can robustly predict rarefied or hypersonic non-equilibrium flows.

Beyond the issue of physical coverage, existing evaluation methods also leave room for more systematic comparison in non-equilibrium flow learning. In many studies, neural architectures are assessed un-

der different parameter budgets, grid resolutions, preprocessing strategies, and training protocols. Such differences can make it difficult to distinguish the effect of architectural inductive bias from that of model capacity or implementation details [9]. Moreover, different neural architectures encode distinct assumptions about locality, global coupling, mesh structure, and coordinate representation. These assumptions can strongly affect their ability to resolve sharp gradients, high-order non-equilibrium statistics, and localized flow structures. In particular, coordinate-based and attention-based models may exhibit spectral bias and tend to learn smoother, low-frequency mappings, while spectral or convolutional models may respond differently to high-frequency feature injection [10]. Therefore, a benchmark for non-equilibrium flow transport should not only report prediction errors, but also provide controlled settings for diagnosing how different architectures behave under specific physical challenges [11,12].

To address these issues, we introduce TransportBench, a benchmark framework with two closely connected objectives. First, it provides a high-fidelity dataset for non-equilibrium flow transport, covering rarefied and continuum regimes, low-speed and hypersonic flows, inert and reacting gases, and both translational and internal-energy non-equilibrium. The dataset is organized into four representative flow scenarios, including airfoil, cylinder, cavity, and double-cone configurations, which are designed to examine geometric generalization, physical-parameter generalization, high-order moment prediction, and shock-dominated hypersonic transport. Second, TransportBench establishes a standardized model benchmark for evaluating representative scientific machine learning architectures under controlled settings. By comparing operator, spectral, convolutional, attention-based, geometric, and latent models with unified preprocessing, training protocols, parameter budgets, and evaluation metrics, the benchmark enables a more systematic assessment of how different architectural inductive biases perform across distinct flow regimes. Through this dual design, TransportBench aims to serve not only as a reusable dataset for non-equilibrium fluid dynamics, but also as a diagnostic testbed for developing and evaluating learning-based methods for flow transport beyond the classical Navier-Stokes level.

2. Related work

2.1. Neural architectures for learning fluid dynamics

Scientific machine learning has introduced a broad range of neural architectures for approximating fluid dynamics across different spatial representations and physical scenarios. These models differ not only in the network design, but also in the inductive biases they impose on the learned solution fields. Understanding these biases is essential for evaluating their suitability for non-equilibrium transport problems, where the target fields may contain smooth regions, sharp discontinuities, multi-scale structures, and high-order kinetic statistics.

Convolutional and latent-representation models, such as convolutional neural networks, U-Net [13], and autoencoders [14], have been widely used for learning solution fields represented on structured grids. Convolutional models exploit local receptive fields, hierarchical feature extraction, and translational equivariance, making them effective for capturing spatial gradients and multi-scale coherent structures. U-Net, in particular, has become a common baseline for image-like flow-field prediction because the encoder-decoder structure and skip connections combine local details with broader contextual information. Convolutional autoencoders (AE) further introduce a latent bottleneck for compact flow-field representation and are often used for reduced-order modeling and manifold learning.

Operator-learning models, such as the deep operator network (DeepONet) [15] and Fourier neural operator (FNO) [16], provide another important route for data-driven fluid simulation. Rather than learning only a finite-dimensional mapping between fixed discretized inputs and outputs, these models aim to approximate mappings between func-

tion spaces. DeepONet represents nonlinear operators through a branch-trunk architecture and supports continuous spatial queries, while FNO parameterizes integral kernels in the Fourier domain and efficiently captures long-range correlations on structured grids. These properties make operator-learning models attractive for learning input-output mappings between operating conditions, geometries, and flow fields, regardless of whether the underlying data are generated by parametric PDE solvers, particle simulations, or experiments. Their performance, however, may depend strongly on whether the target flow structures are smooth, globally correlated, or dominated by localized discontinuities.

Attention-based and geometric learning models provide flexible ways to model spatial interactions beyond purely local convolutional kernels. Vision Transformers (ViT) [17] use patch embeddings and self-attention to capture long-range dependencies on structured grids, making them suitable for flow fields where global interactions are important. In contrast, point-based and graph-based models are designed for irregular spatial representations, such as point clouds or unstructured meshes. Point Transformer (PT) and related geometric architectures can directly aggregate information from neighboring points, thereby providing a natural representation of complex geometries without relying entirely on Cartesian grid interpolation [18]. Although these models offer greater flexibility in spatial representation and interaction modeling, their effectiveness in sparse, shock-dominated, or highly non-equilibrium regimes remains problem-dependent.

Overall, existing neural architectures provide complementary strengths for learning fluid dynamics. Convolutional models rely on local receptive fields and hierarchical feature extraction to capture spatial gradients and multi-scale structures. Operator-learning models are designed to learn mappings between functions or fields, with Fourier-based variants further exploiting spectral representations to capture long-range correlations. Attention-based models use token-level interactions to model global dependencies, while geometric models operate on point clouds, graphs, or meshes to accommodate flexible spatial representations. These strengths, however, can also become limitations when the target flow fields contain strong discontinuities or high-order non-equilibrium structures. Coordinate-based models, and in some cases attention-based models, may favor smoother or lower-frequency representations, whereas Fourier-based spectral representations may be prone to oscillatory artifacts near sharp discontinuities. Therefore, a systematic comparison of these architectures under non-equilibrium flow conditions is necessary for understanding which model families are suitable for different regimes.

2.2. Benchmarks for scientific machine learning and flow simulation

Alongside the rapid development of neural architectures, standardized benchmarks have become essential for evaluating scientific machine learning methods. General-purpose PDE benchmarks, such as PDEBench [7] and PINNacle [19], provide valuable testbeds for assessing neural solvers across a range of partial differential equations. Fluid-oriented benchmarks, including CFDBench [20], FlowBench [8], and FD-Bench [9], further extend the evaluation landscape toward flow simulation over complex geometries and varying physical parameters. These efforts have played an important role in promoting reproducible comparison and accelerating progress in data-driven fluid dynamics.

Nevertheless, most existing benchmarks remain primarily focused on continuum-level flow problems. Their governing equations are usually based on the Navier-Stokes equations or other hydrodynamic PDEs, and the corresponding datasets often involve moderate flow conditions, smooth field structures, and standard macroscopic variables such as density, velocity, pressure, and temperature. While such settings are valuable for evaluating neural representations in conventional CFD tasks, they do not fully cover the physical complexity of non-equilibrium flow transport. In rarefied and hypersonic regimes, models must deal with Knudsen layers, velocity-distribution distortion, high-order stress and

heat-flux moments, strong shock discontinuities, and multi-scale kinetic-to-continuum behavior.

Another challenge concerns the comparability of model evaluations. Existing studies often differ in parameter budgets, grid resolutions, preprocessing pipelines, train-test splits, optimization strategies, and levels of hyperparameter tuning. As a result, reported performance differences may be conflated with increased model capacity, higher data resolution, longer training, or more aggressive implementation-specific optimization. In such cases, it becomes difficult to determine whether a model performs better because of a more suitable architectural inductive bias or simply because it benefits from larger parameter counts and greater computational effort. This issue becomes particularly important in non-equilibrium flows, where model behavior is highly sensitive to spatial representation, normalization strategy, and the treatment of high-frequency features.

TransportBench is designed to complement existing benchmarks by focusing specifically on non-equilibrium flow transport. As summarized in Table 1, it differs from prior datasets in two main aspects. First, it includes high-fidelity non-equilibrium flow data generated from gas-kinetic simulations, enabling evaluation beyond standard Navier-Stokes-level benchmarks. Second, it incorporates high-order non-equilibrium quantities, such as stress tensor, heat flux, and particle distribution function, rather than only macroscopic conservative variables. In addition, it adopts unified preprocessing, training protocols, parameter budgets, and evaluation metrics to enable fair and manageable comparison across representative scientific machine learning architectures. Through this design, TransportBench serves both as a reusable dataset for non-equilibrium flow transport and as a standardized benchmark for assessing model performance across distinct flow regimes.

3. Dataset and benchmark design

3.1. Design principles

TransportBench is designed with two closely connected objectives: to provide physically representative datasets for non-equilibrium flow transport and to establish controlled settings for benchmarking scientific machine learning models. The first objective concerns dataset construction. To this end, it covers a broad range of flow regimes, including rarefied and continuum fluids, low-speed and hypersonic conditions, inert and reacting gases, and both translational and internal-energy non-equilibrium effects. These regimes are selected to expose learning models to physical phenomena that are usually absent or underrepresented in standard continuum-flow benchmarks, such as Knudsen layers, velocity-distribution distortion, stress and heat-flux moments, thermochemical non-equilibrium, and shock-shock and shock-boundary interactions.

The second objective concerns model evaluation. TransportBench includes tasks that probe distinct learning challenges, including geometric generalization, physical-parameter generalization, high-order moment prediction, and shock-dominated field reconstruction. It also adopts unified preprocessing, training protocols, parameter budgets, and evaluation metrics, so that model comparisons are less affected by differences in data handling, model size, or implementation-specific training choices.

A particular design consideration is the representation of sharp physical structures. Numerical analyses show that neural networks tend to learn smooth, low-frequency components more easily than localized high-frequency variations [10,21]. In shock-dominated and multi-scale flows, this tendency may lead to smeared discontinuities or inaccurate Rankine-Hugoniot values. A further question concerns the role of explicit high-frequency feature injection [22]. Such techniques are often introduced to mitigate spectral bias and improve the representation of sharp spatial variations. However, their effect may depend strongly on the underlying architecture. For models with different coordinate, con-

Table 1

Comparison of TransportBench with existing scientific machine learning and fluid dynamics benchmarks in terms of physical fidelity, generalization capability, and evaluation difficulty. “Non-equilibrium effects” indicates whether the benchmark includes transport phenomena beyond the standard continuum-flow level. “High-order statistics” denotes prediction targets such as the stress tensor and heat flux. “Kinetic” denotes the Boltzmann equation and related physics.

Benchmark	Physical fidelity			Generalization capability		Evaluation difficulty	
	Governing equations	Non-equilibrium effects	High-order statistics	Geometries	Physics	Discontinuities & shocks	Sparse grids & few-shot
PDEBench [7]	Continuum (NS)	×	×	×	✓	×	×
PINNacle [19]	Continuum (mixed)	×	×	×	×	×	×
CFDBench [20]	Continuum (NS)	×	×	✓	✓	×	×
FlowBench [8]	Continuum (NS)	×	×	✓	✓	×	×
FD-Bench [9]	Continuum (NS)	×	×	✓	✓	×	×
TransportBench	Kinetic	✓	✓	✓	✓	✓	✓

volutional, spectral, or integral representations, high-frequency injection may either improve shock resolution or introduce oscillatory artifacts, noisy background structures, or frequency-aliasing effects. TransportBench therefore includes shock-dominated configurations not only as difficult prediction tasks, but also as diagnostic settings for examining when high-frequency feature injection helps, when it fails, and how its effect interacts with architectural inductive bias.

Following these principles, the benchmark is constructed from four representative flow scenarios, which are introduced in detail below. These scenarios connect the three components of TransportBench: dataset construction defines the physical scenarios and target quantities (Sections 3 and 4), the unified benchmarking protocol standardizes preprocessing, training, and evaluation (Section 5), and the diagnostic analysis interprets how different architectures behave across distinct non-equilibrium regimes (Section 6).

3.2. Unified physical and learning formulation

Non-equilibrium fluid transport can be described from a unified statistical mechanics perspective, where the gas state is characterized by distribution functions over particle velocity and internal states. For a gas-dynamic system that may contain multiple species and internal energy modes, a generalized Boltzmann equation can be written as

$$\frac{\partial f_{s,l}}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{x}} f_{s,l} = \mathcal{Q}_{s,l}^{\text{el}} + \mathcal{Q}_{s,l}^{\text{inel}} + \mathcal{Q}_{s,l}^{\text{chem}}, \quad (1)$$

where $f_{s,l}(\mathbf{x}, \mathbf{v}, t)$ denotes the mass density distribution function at the s th species in the l th internal state. The terms on the right-hand side represent elastic collisions, inelastic internal-energy transitions, and chemical production or loss processes, respectively. This formulation provides a common physical basis for non-equilibrium transport, including rarefied and high-temperature gas dynamics.

For a monatomic gas without chemical reactions, Eq. (1) reduces to the classical Boltzmann equation,

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{x}} f = \mathcal{Q}^{\text{el}}(f). \quad (2)$$

Based on the distribution function, macroscopic flow variables can be obtained by taking moments over the velocity space, i.e.,

$$\rho = \int_{(\mathbb{R})^3} f d\mathbf{v}, \quad \rho \mathbf{u} = \int_{(\mathbb{R})^3} \mathbf{v} f d\mathbf{v}, \quad \rho E = \int_{(\mathbb{R})^3} \frac{1}{2} |\mathbf{v}|^2 f d\mathbf{v}. \quad (3)$$

We use $\mathbf{U}(\mathbf{x}, t) = [\rho, \rho \mathbf{u}, \rho E]^T$ to denote the conservative variables. Taking moments of the Boltzmann equation yields macroscopic transport equations,

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \nabla_{\mathbf{x}} \cdot (\rho \mathbf{u}) &= 0, \\ \frac{\partial \rho \mathbf{u}}{\partial t} + \nabla_{\mathbf{x}} \cdot (\rho \mathbf{u} \otimes \mathbf{u} + \mathbf{P}) &= 0, \\ \frac{\partial \rho E}{\partial t} + \nabla_{\mathbf{x}} \cdot (\rho E \mathbf{u} + \mathbf{P} \cdot \mathbf{u} + \mathbf{q}) &= 0, \end{aligned} \quad (4)$$

where the total pressure tensor and heat flux are given by

$$\mathbf{P} = \int_{(\mathbb{R})^3} \mathbf{v}_{\text{pec}} \otimes \mathbf{v}_{\text{pec}} f d\mathbf{v}, \quad \mathbf{q} = \int_{(\mathbb{R})^3} \frac{1}{2} |\mathbf{v}_{\text{pec}}|^2 \mathbf{v}_{\text{pec}} f d\mathbf{v}, \quad (5)$$

where $\mathbf{v}_{\text{pec}} = \mathbf{v} - \mathbf{u}$ is the peculiar velocity. These moment relations define the macroscopic and high-order quantities used as learning targets in TransportBench. In particular, Tasks I and II use DSMC data to provide macroscopic rarefied-flow fields and near-wall non-equilibrium transport information, while Task III uses deterministic DVM solutions of the Boltzmann-BGK equation to provide particle distribution functions and derived high-order moments such as the stress tensor and heat flux. Therefore, Eqs. (1)–(5) directly specify the physical quantities that the benchmark models are expected to reconstruct in the kinetic and rarefied-flow tasks.

For shock-dominated high-enthalpy flows, directly resolving the full kinetic equation with internal-state transitions is often prohibitively expensive. TransportBench therefore also includes a macroscopic non-equilibrium modeling level, which can be derived as an asymptotic limit of Eq. (1), and is used in Task IV to generate the high-enthalpy double-cone dataset. The vibrational state-specific conservation laws can be written as

$$\begin{aligned} \frac{\partial \rho_{s,l}}{\partial t} + \nabla_{\mathbf{x}} \cdot (\rho_{s,l} \mathbf{u}) &= \nabla_{\mathbf{x}} \cdot (\rho D_s \nabla_{\mathbf{x}} \varphi_{s,l}) + S_{s,l}, \\ \frac{\partial \rho \mathbf{u}}{\partial t} + \nabla_{\mathbf{x}} \cdot (\rho \mathbf{u} \otimes \mathbf{u} + \mathbf{P}) &= 0, \\ \frac{\partial \rho E}{\partial t} + \nabla_{\mathbf{x}} \cdot (\rho E \mathbf{u} + \mathbf{P} \cdot \mathbf{u} + \mathbf{q}) &= \nabla_{\mathbf{x}} \cdot \left(\rho \sum_{s=1}^{N_s} D_s \sum_{l=1}^{V_s} h_{s,l} \nabla_{\mathbf{x}} \varphi_{s,l} \right), \end{aligned} \quad (6)$$

where the total pressure tensor and heat flux are given by

$$\mathbf{P} = p \mathbf{I} - \mu (\nabla_{\mathbf{x}} \mathbf{u} + (\nabla_{\mathbf{x}} \mathbf{u})^T) + \frac{2}{3} \mu (\nabla_{\mathbf{x}} \cdot \mathbf{u}) \mathbf{I}, \quad \mathbf{q} = -\eta \nabla_{\mathbf{x}} T, \quad (7)$$

where $\rho_{s,l}$, $\varphi_{s,l}$, $h_{s,l}$, and $S_{s,l}$ denote the density, mass fraction, enthalpy, and source term of the s th species in the l th internal state, respectively. Besides, N_s represents the total number of chemical species, and V_s is the number of vibrational states associated with the s th species, D_s is the diffusion coefficient of the s th species, η is the thermal conductivity associated with translational-rotational energy, and μ is the shear viscosity. In this benchmark, Eqs. (6) and (7) do not apply to all tasks. Instead, they specify the modeling level and target variables of the high-enthalpy double-cone task (Task IV). This task focuses on the reconstruction of shock-dominated thermochemical flow fields, where strong gradients, localized extrema, and internal-energy non-equilibrium pose learning challenges different from those in the rarefied-flow and kinetic-moment tasks.

Based on the above physical formulations, the four tasks are cast into a unified supervised learning form. Let $\mathcal{A}$ denote the input operating condition, which may include geometry descriptions and freestream parameters. Let $\mathcal{U}$ denote the target solution field, which may include macroscopic quantities such as density, velocity, pressure, and temperature, as well as non-equilibrium quantities such as stress tensor, heat flux, and particle distribution function. The learning problem is written as

$$\mathcal{G} : \mathcal{A} \mapsto \mathcal{U}. \quad (8)$$

A scientific machine learning model $\mathcal{G}_\theta$ is trained to approximate this mapping using a reconstruction loss,

$$\mathcal{L}_{\text{rec}} = \|\mathcal{G}_\theta(\mathcal{A}) - \mathcal{U}\|_2^2. \quad (9)$$

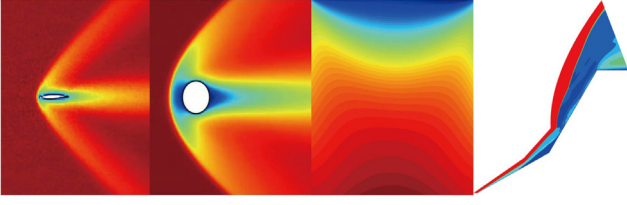


Fig. 1. Representative ground-truth flow fields of the four physical scenarios in TransportBench. From left to right: (a) rarefied airfoil flows with varying geometries for geometry-dependent predictions; (b) rarefied cylinder flow across a range of Mach and Knudsen numbers for operating-condition-dependent predictions; (c) lid-driven cavity flows for distribution-function and high-order moment predictions; and (d) high-enthalpy double-cone flows for shock-dominated hypersonic field reconstructions under thermochemical non-equilibrium. Together, these scenarios cover geometry variation, operating-condition variation, kinetic-moment prediction, and sharp-discontinuity representation in non-equilibrium fluid transport.

In the present work, this formulation is mainly used to standardize the task definitions and model evaluation, rather than to impose a single governing equation on all datasets. The purely data-driven baselines are trained with the reconstruction loss in Eq. (9), while the physical formulations above define the regimes, target quantities, and diagnostic objectives of the four benchmark tasks. This connection links the theoretical description in this subsection to the dataset construction in Section 4 and the model evaluation in Sections 5 and 6.

3.3. Physical scenarios

Based on the above design principles and unified formulation, TransportBench contains four representative flow scenarios. They are selected to cover different levels of non-equilibrium gas transport and to expose neural network models to complementary learning challenges, including geometry-dependent prediction, parameter-dependent prediction, high-order kinetic-moment prediction, and shock-dominated field reconstruction. Figure 1 shows the ground-truth flow fields for four representative tasks. Task I: airfoil flow with geometry-dependent prediction

Flow over an airfoil is a classical problem in aerodynamics. In rarefied gas dynamics, the interaction between the incoming flow and the airfoil surface can generate complex structures, including bow shocks near the leading edge, boundary-layer development, and wake regions downstream of the profile. When the airfoil geometry changes, the shock standoff distance, surface interaction, and wake morphology also vary. This makes the airfoil case suitable for evaluating whether neural models can learn geometry-dependent flow structures and predict flow fields for unseen airfoil shapes.

The baseline geometry is the RAE2822 airfoil, parameterized using the Class-Shape Transformation (CST) method. The upper and lower surfaces are represented by sixth-order Bernstein polynomials. The corresponding CST coefficients of the baseline airfoil are

$$\begin{aligned} \mathbf{b}_U &= [0.1266, 0.1379, 0.1617, 0.1642, 0.2205, 0.1719, 0.2132], \\ \mathbf{b}_L &= [-0.1289, -0.1426, -0.1311, -0.2615, -0.0253, -0.1145, 0.0727], \end{aligned} \quad (10)$$

where $\mathbf{b}_U$ and $\mathbf{b}_L$ denote the CST coefficients of the upper and lower surfaces, respectively. To generate geometric variations, each coefficient is independently perturbed within $\pm 30\%$,

$$b'_i = b_i \cdot (1 + r_i), \quad r_i \sim \mathcal{U}(-0.3, 0.3), \quad (11)$$

where $\mathcal{U}$ denotes the uniform distribution. A total of 100 shape-varied airfoil geometries are produced, as illustrated in Fig. 2.

The incoming flow consists of nitrogen and is set to the Mach number $Ma = 2.0$ and the Knudsen number $Kn = 0.026$. The Knudsen number is defined based on the freestream mean free path and the airfoil chord, and this flow condition corresponds to the slip-flow regime. The freestream temperature is 161 K, the freestream velocity is 517.26 m/s, and the wall temperature is fixed at 290 K. A fully accommodating diffuse reflection model is imposed at the airfoil surface interaction. These physical parameters are kept fixed for all 100 samples so that the task isolates the effect of geometric variation. The resulting flow fields are mapped onto a Cartesian grid of 128×128 . In the benchmark evaluation, the train-test split is performed over the set of airfoil geometries, so that the task assesses both interpolation within the geometry distribution and prediction for unseen shapes. Task II: cylinder flow with parameter-dependent prediction

Flow around a circular cylinder is a canonical configuration for studying flow separation and wake dynamics. In high-speed rarefied gas flow, the bow shock and downstream wake are strongly affected by compressibility and rarefaction effects. The cylinder case is therefore designed to evaluate whether neural models can learn parameter-dependent changes in flow topology over a wide range of operating conditions, especially Mach and Knudsen numbers.

In this task, the cylinder geometry is fixed, while the operating conditions vary over a broad parameter space. A total of 2400 flow cases are generated. The Knudsen number, defined using the cylinder diameter as the characteristic length, ranges from $Kn = 0.06$ to $Kn = 1.0$ with increments of 0.02, and the Mach number ranges from $Ma = 2.0$ to $Ma = 6.9$ with increments of 0.1. Argon is used as the working gas. The freestream temperature is set to $T_\infty = 273$ K, and the wall temperature is fixed at $T_w = 273$ K. The gas-surface interaction at the cylinder wall is modeled using fully diffuse reflection.

Across this parameter space, the flow topology changes significantly. At lower Knudsen numbers, the flow exhibits clearer bow-shock and wake structures. At higher Knudsen numbers, the frequency of inter-molecular collisions decreases, the rarefaction effect becomes more pronounced, and the scale of macroscopic flow structures increases accordingly, as shown in Fig. 3. The computational domain extends 11 cylinder radii in the streamwise direction and 10 cylinder radii in the transverse direction. A multi-level grid refinement strategy is used to improve near-wall resolution. Depending on the boundary-layer thickness and rarefaction level, the number of unstructured cells varies across cases. The data are then mapped onto an elongated Cartesian grid of 128×192 to capture the extended wake region and non-equilibrium transport.

This task focuses on predictions across varying operating conditions. It tests whether a neural architecture can learn how global flow structures vary with rarefaction and compressibility, rather than only interpolate between visually similar flow fields. The test set contains unseen combinations of Mach and Knudsen numbers, enabling evaluation of parameter-space prediction and generalization under controlled conditions. Task III: lid-driven cavity flow with high-order kinetic prediction

While Tasks I and II focus mainly on macroscopic flow-field prediction under geometry or parameter variation, Task III is designed to probe the micro-macro connection in non-equilibrium fluid transport. The lid-driven cavity case is selected as its simple geometry and confined domain help isolate the gas-kinetic challenge associated with the Boltzmann equation. In the rarefied flow regime, the particle distribution function can deviate substantially from the local Maxwellian, and non-equilibrium transport phenomena emerge. Therefore, predicting only low-order macroscopic variables is insufficient for fully characterizing the flow.

The physical domain is a two-dimensional unit square, $x, y \in [0, 1]$. A total of 99 simulation cases are generated, with the Knudsen number varying from $Kn = 0.02$ to $Kn = 1.0$ in increments of 0.01. The freestream Knudsen number is defined based on the cavity side length, and its range covers multiple flow regimes and captures the evolution from near-equilibrium to strongly non-equilibrium states. Argon is used as the working gas. The boundary conditions consist of isothermal sta-

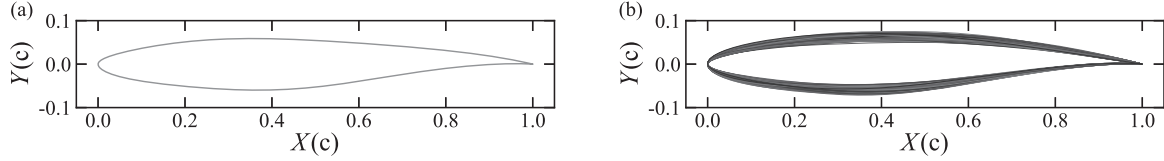


Fig. 2. Baseline RAE2822 airfoil and CST-generated geometric variations.

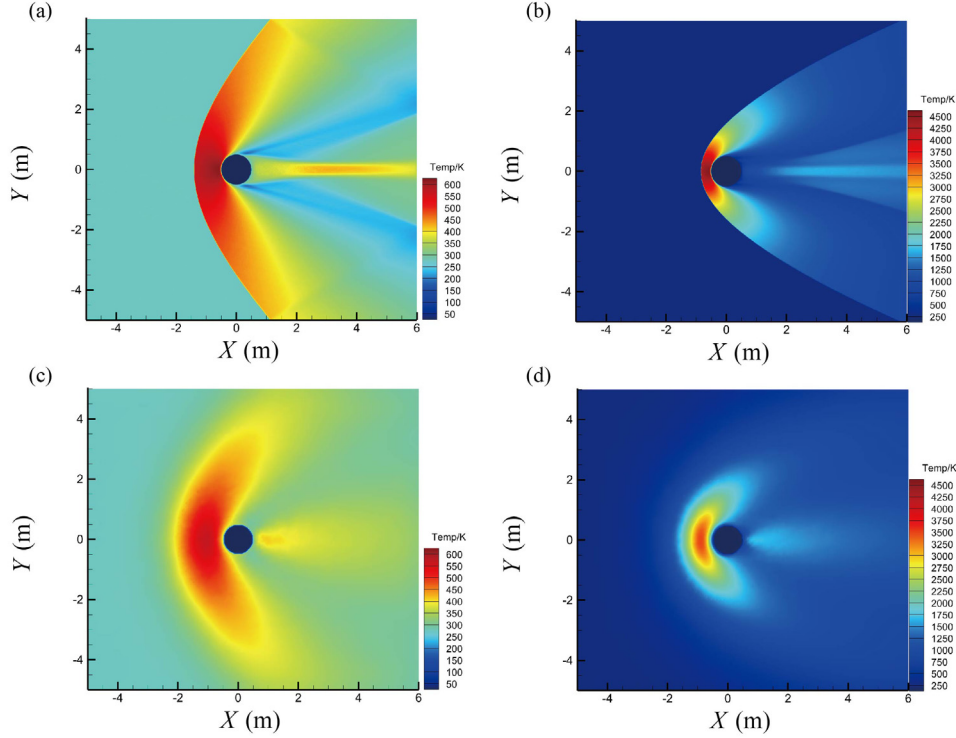


Fig. 3. Representative temperature fields in the cylinder flow. The four cases correspond to different combinations of Mach and Knudsen numbers. At lower Knudsen numbers, clearer bow-shock and wake structures are observed, whereas at higher Knudsen numbers the reduced collision frequency leads to stronger rarefaction effects and more diffuse macroscopic flow structures. This comparison illustrates the freestream-condition-dependent variation of shock and wake topology.

tionary walls at $T_w = 273$ K and a top lid moving at a constant velocity $u_w = 50.55$ m/s.

The output of this task includes both macroscopic flow quantities and particle distribution functions. The particle distribution function is resolved on a discrete velocity grid, and the macroscopic quantities are obtained by taking moments over velocity space. Higher-order non-equilibrium statistics, including the stress tensor $\mathbf{P}$ and heat flux $\mathbf{q}$, are of particular importance because they encode the distortion of the velocity distribution function away from local equilibrium at a macroscopic level. Compared with low-order macroscopic variables, these quantities provide a more sensitive measure of non-equilibrium transport.

Figure 4 shows representative stress and heat-flux fields at different Knudsen numbers. In the near-equilibrium regime at $Kn = 0.02$, these fields remain relatively smooth, whereas at higher Knudsen numbers they exhibit stronger non-equilibrium structures and pronounced gradients. This task therefore evaluates whether neural network models can serve as a micro-macro bridge. An effective model should not only reproduce macroscopic fields, but also capture deviations of the particle distribution function from equilibrium. Since the Knudsen number is varied across the dataset, this task also provides a controlled setting for testing prediction across different degrees of rarefaction. Task IV: double-cone flow with shock-dominated field reconstruction

This task considers high-enthalpy nitrogen flow over a $25^\circ - 55^\circ$ double-cone configuration based on the CUBRC LENS-I experiments [23]. Unlike Tasks I-III, which are dominated by translational non-

equilibrium effects, this case is strongly influenced by thermochemical non-equilibrium. Its physical complexity arises from detailed state-to-state (StS) kinetics for internal-energy relaxation and finite-rate processes. In the StS model, N_2 is represented using 55 discrete vibrational energy levels. The state-specific mass production rates are evaluated by accounting for vibrational-translational (V-T) energy exchange and dissociation-recombination (D-R) processes associated with $N_2 - N_2$ and $N_2 - N$ collisions [24]. At the same time, the flow contains strong shock discontinuities and shock-boundary-layer interactions, making it a challenging case for neural field representation.

An axisymmetric computational grid of 1024×512 nodes is employed in the axial and radial directions, respectively. The first near-wall grid spacing is set to 1×10^{-7} m so that the cell Reynolds number is on the order of unity. This grid resolution has been verified to be grid-independent for simulations of high-enthalpy double cone flows [24]. The freestream Mach number ranges from $Ma = 6.0$ to $Ma = 17.6$, and the Reynolds number ranges from 1.4×10^5 to 2.3×10^6 . As the Mach number increases, thermochemical non-equilibrium effects become more pronounced, and the vibrational energy distribution deviates more strongly from the equilibrium Boltzmann distribution. A representative Mach-number field is shown in Fig. 5.

For learning-based evaluation, the data are mapped onto a sparse and highly anisotropic representation of 17×384 . This setting is intentionally challenging. The dataset contains 51 high-fidelity snapshots, with 45 samples used for training and 6 samples used for testing. Meanwhile,

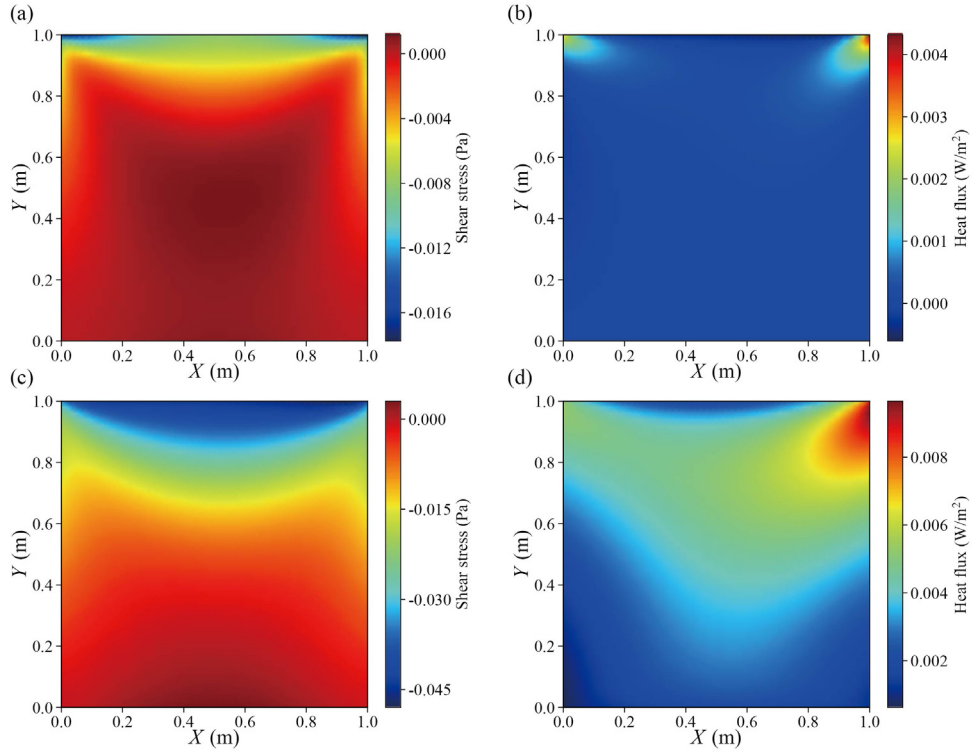


Fig. 4. Representative non-equilibrium moment fields in the lid-driven cavity flow. The figure shows the shear-stress component P_{xy} and heat-flux component q_x . The top row corresponds to the near-equilibrium regime at $Kn = 0.02$, where the fields are relatively smooth; the bottom row corresponds to $Kn = 1.0$, where stronger rarefaction effects lead to more pronounced gradients and non-equilibrium moment structures.

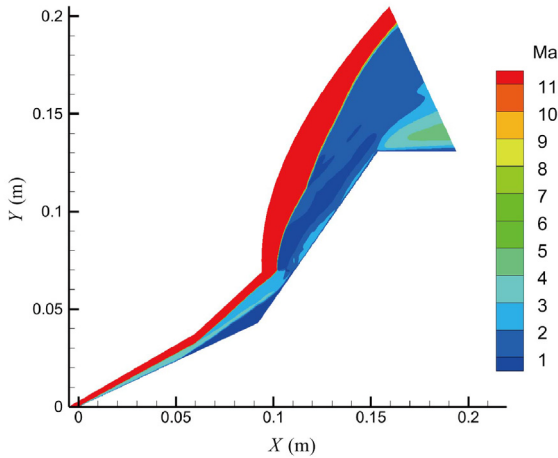


Fig. 5. Representative Mach-number contour for the high-enthalpy double-cone flow at $Ma = 11.628$ and $Re = 341500$.

the target fields contain localized shock peaks, steep gradients, and large property variations across narrow spatial regions. Therefore, this task is not merely a high-enthalpy flow prediction problem, but also a diagnostic test for the representational behavior of neural architectures under sparse and shock-dominated conditions.

This design directly reflects the frequency-representation issue discussed in the benchmark principles. Neural models that favor smooth, low-frequency mappings may capture the global flow topology but smear shock peaks and underestimate localized extrema. In contrast, models equipped with explicit high-frequency feature injection may better represent sharp structures in some cases, but may also introduce os-

cillatory artifacts, nonphysical noise, or frequency-aliasing effects. By including the double-cone configuration, this task provides a controlled setting for examining how architectural inductive bias, spectral bias, and high-frequency feature representations affect shock-dominated flow prediction.

Together, the four tasks form a benchmark suite that spans geometry-dependent prediction, parameter-dependent prediction, high-order kinetic-statistics prediction, and shock-dominated hypersonic field reconstruction. This organization allows TransportBench to evaluate both the predictive accuracy and the physical robustness of scientific machine learning models across distinct non-equilibrium flow regimes. To present this information in one place, Table 2 summarizes the data sources, input-output settings, grid representations, sample sizes, data splits, and evaluation objectives of the four tasks.

4. Dataset generation and preprocessing

4.1. High-fidelity data generation

Generating reliable reference data for non-equilibrium flow transport requires numerical methods that are consistent with the physical regime. In TransportBench, the four tasks are therefore generated using different high-fidelity solvers and modeling levels. The rarefied external flows are simulated using particle-based kinetic methods, the cavity flow is generated using a deterministic discrete velocity method for the Boltzmann equation, and the high-enthalpy double-cone task is computed using a continuum solver coupled with state-to-state thermochemical non-equilibrium modeling. The solver selection is consistent with the unified physical formulation described in Section 3.2.

For Tasks I and II, the reference data are generated using the direct simulation Monte Carlo (DSMC) method [4]. DSMC is a well-established gas-kinetic particle method for rarefied gas dynamics and is particularly suitable for high-speed flows with strong non-equilibrium effects [25].

Table 2

Summary of the four TransportBench tasks and their roles in the benchmark.

Task	Physical scenario	Input–output setting	Data format and split	Evaluation objective
I	Rarefied flow over CST-varied RAE2822 airfoils (DSMC).	<i>In</i> : CST-varied airfoil geometry. <i>Out</i> : macroscopic flow field.	128 × 128 grid; 100 samples, 80%/20% split over geometries.	Geometry-dependent prediction; generalization to unseen shapes.
II	Rarefied flow around a fixed circular cylinder (DSMC).	<i>In</i> : Ma , Kn . <i>Out</i> : flow field with shock wave and wake structures.	128 × 192 grid; 2400 samples, 80%/20% split over parameters.	Parameter-dependent prediction across rarefaction and compressibility.
III	Lid-driven cavity flow (DVM for Boltzmann-BGK).	<i>In</i> : Kn . <i>Out</i> : distribution functions, macroscopic fields, and high-order moments.	50 × 50 spatial grid, multi-channel moment/distribution fields; 99 samples, 80%/20% split over Kn .	High-order kinetic prediction; micro-macro consistency.
IV	High-enthalpy double-cone flow with thermochemical non-equilibrium (state-to-state).	<i>In</i> : Ma , Re . <i>Out</i> : shock-dominated thermochemical fields with localized peaks.	17 × 384 sparse anisotropic grid; 51 snapshots (45 train / 6 test).	Data-limited shock-dominated reconstruction; frequency-representation diagnosis.

Table 3VHS model parameters for Argon and Nitrogen ($T_{\text{ref}} = 273$ K).

Species	ω	d_{ref} (Å)
Ar	0.81	4.17
N ₂	0.74	4.17

A typical DSMC simulation alternates between particle transport and stochastic intermolecular collisions within computational cells. After the flow reaches a statistically steady state, macroscopic quantities such as density, velocity, temperature, stress, and heat flux can be obtained by sampling and averaging particle information over sufficiently long time intervals.

The open-source SPARTA solver [26] is used for the DSMC simulations. The variable hard sphere (VHS) molecule model is adopted to describe intermolecular collisions, and the no-time-counter (NTC) algorithm is used for efficient collision computations [4]. The VHS parameters for argon and nitrogen are listed in Table 3. For nitrogen internal-energy relaxation, a constant-collision-number model is used, with the rotational and vibrational collision numbers set to 5 and 50, respectively. Gas-surface interaction is modeled using fully diffuse reflection with complete thermal accommodation to the wall temperature. No additional surface reaction or recombination process is considered. Hierarchical Cartesian grids with adaptive refinement near solid boundaries are used to resolve near-wall rarefaction effects and Knudsen-layer structures. To reduce the stochastic noise inherent in particle methods, time-averaged sampling is performed over sufficiently long steady-state intervals.

For Task III, the lid-driven cavity dataset is generated by solving the Boltzmann-BGK equation using a discrete velocity method (DVM) [27,28]. Unlike DSMC, which relies on stochastic particle sampling, the DVM resolves the distribution function on a discrete velocity grid and evolves it deterministically. Macroscopic variables and high-order non-equilibrium moments are then obtained by numerical integration over the velocity space. In particular, the stress tensor and heat flux are computed directly from velocity moments of the distribution function. This setting provides a clean kinetic dataset for evaluating whether neural models can predict high-order non-equilibrium statistics associated with microscopic velocity-distribution distortion.

For Task IV, the high-enthalpy double-cone flow is computed using the in-house ARTIST-CFD solver [29]. The governing equations are the compressible Navier-Stokes equations coupled with a vibrational state-to-state model for thermochemical non-equilibrium. Inviscid fluxes are computed using a modified Steger-Warming upwind scheme combined with second-order MUSCL reconstruction and a minmod limiter, while viscous fluxes are discretized using a second-order central scheme. Time advancement is performed using a data-parallel line relaxation method. Species viscosities and thermal conductivities are evaluated using Blotner’s curve fitting and Eucken’s relation, respectively, and mixture transport coefficients are obtained using Wilke’s mixing rule. The species

mass diffusion coefficient is determined by assuming a constant Lewis number of 1.4.

4.2. Data preprocessing pipeline

The raw outputs from DSMC, DVM, and continuum CFD solvers differ in spatial format, resolution, variable range, and noise characteristics. A unified preprocessing pipeline is therefore applied to convert these heterogeneous simulation results into standardized learning datasets. This step is essential for fair comparison across neural architectures, since differences in data representation or normalization can strongly affect model performance.

First, all flow fields are mapped onto standardized spatial grids. For the airfoil task, the flow variables are interpolated onto a 128 × 128 Cartesian grid. For the cylinder task, an elongated 128 × 192 grid is used to capture the extended wake region. For the cavity task, the structured physical domain is retained as 50 × 50 while the kinetic moments are organized as multi-channel spatial fields. For the double-cone task, the high-resolution axisymmetric simulation results are mapped onto a sparse and highly anisotropic representation of 17 × 384, which is used to evaluate model behavior under shock-dominated and data-limited conditions.

Second, geometry masks are introduced for configurations containing solid bodies. The binary mask $M(\mathbf{x})$ distinguishes fluid regions from the solid or nonphysical regions. During training and evaluation, loss functions and error metrics are computed only over the fluid domain. This avoids forcing models to learn meaningless values inside solid obstacles and prevents invalid interior regions from biasing the comparison among architectures.

Third, variable normalization is applied to reduce scale imbalance across different physical quantities and flow regimes. For variables with moderate dynamic ranges, standard min-max normalization is used. For quantities with large variations, especially in the double-cone task, where pressure and thermodynamic variables may change sharply across shocks, logarithmic dynamic-range compression is applied before normalization. For a positive scalar field $a(\mathbf{x})$, the transformed variable can be written as

$$\tilde{a}(\mathbf{x}) = \frac{\mathcal{T}(a(\mathbf{x})) - \min(\mathcal{T}(a))}{\max(\mathcal{T}(a)) - \min(\mathcal{T}(a))} \in [0, 1], \quad (12)$$

where

$$\mathcal{T}(a) = \log_{10}(a + \epsilon), \quad (13)$$

and ϵ is a small positive constant introduced for numerical stability. The logarithmic compression helps prevent extreme shock-induced jumps from dominating the optimization process.

Finally, all datasets are organized into consistent input-output pairs according to the learning formulation in Section 3.2. The input may include geometric descriptors, physical parameters, or operating conditions, while the output consists of task-dependent flow variables. By applying consistent grid mapping, masking, and normalization procedures, TransportBench ensures that model comparisons primarily reflect architectural behavior rather than differences in data preparation.

5. Benchmarking protocols and baseline models

5.1. Baseline models

We evaluate six representative neural architectures that cover different spatial representations and architectural inductive biases commonly used in scientific machine learning for fluid dynamics. Convolutional and latent-representation models

U-Net is selected as a structured-grid convolutional baseline. Its encoder-decoder architecture, local receptive fields, and skip connections make it effective for capturing multiscale spatial structures and localized gradients. A convolutional autoencoder is also included to evaluate whether non-equilibrium flow fields can be represented through a compact latent bottleneck. This model provides a useful baseline for assessing the compressibility of high-dimensional flow fields and the potential loss of sharp or high-frequency information during latent-space projection. Operator-learning models

The deep operator network (DeepONet) and Fourier Neural Operator (FNO) are included as representative operator-learning architectures. DeepONet uses a branch-trunk structure to learn nonlinear input-output mappings and supports continuous spatial queries. FNO parameterizes integral kernels in the Fourier domain and captures long-range correlations on structured grids. These two models provide complementary examples of operator learning: DeepONet emphasizes coordinate-based field evaluation, while FNO combines operator learning with a global spectral representation. Attention-based and geometric learning models

Vision Transformer is included to assess the effectiveness of token-based self-attention for flow-field prediction on structured spatial representations. Point Transformer is selected as a geometric learning baseline that operates on spatial point sets and aggregates information through pointwise attention. These models allow us to evaluate whether global token interactions or flexible point-based representations are beneficial for non-equilibrium flow prediction, especially in settings involving complex geometries, parameter-induced flow variation, or sparse spatial data.

Together, these six architectures span convolutional, latent, operator-learning, spectral, attention-based, and geometric modeling strategies. This selection enables TransportBench to examine how different inductive biases perform under geometry and parameter variations, high-order kinetic-moment prediction, and shock-dominated field reconstruction.

5.2. Evaluation settings and training protocols

To make model comparisons more controlled, TransportBench adopts standardized training and evaluation settings. For Tasks I-III, all baseline models are constrained to approximately one million trainable parameters. This setting is intended to evaluate architectural efficiency under moderate model capacity and to reduce the possibility that performance differences are dominated by model size. For Task IV, which involves sparse data, strong shock discontinuities, and highly anisotropic spatial representation, the parameter budget is increased to approximately 33 million. This setting gives each model sufficient capacity to represent the difficult target fields and helps reveal whether remaining errors are associated with architectural limitations rather than severe under-parameterization.

For Tasks I-III, the datasets are split into training and testing subsets using an 80%/20% ratio. The test sets contain unseen samples from the corresponding geometry or parameter distributions. For the airfoil task, this means unseen airfoil geometries generated by CST perturbations. For the cylinder and cavity tasks, the test set contains unseen physical-parameter values or combinations. For the double-cone task, the dataset contains 51 high-fidelity snapshots in total, of which 45 samples are used for training and 6 samples are reserved for testing. The limited number of high-fidelity snapshots makes the task closer to a data-limited shock-dominated prediction setting. All experiments are conducted using the

Adam optimizer for 2500 epochs, and the baseline settings are kept consistent across model families unless otherwise specified.

The shock-dominated double-cone task requires additional training stabilization because of the large dynamic range and localized discontinuities in the target fields. For this task, we use a long learning-rate warmup with the OneCycleLR scheduler [30] to avoid unstable early-stage optimization. A spatial curriculum strategy is also adopted: the model first learns the smoother background field, and the contribution of localized shock regions is increased during later training stages. These strategies are applied consistently to all baseline models in Task IV, so that the comparison remains focused on architectural behavior rather than model-specific tuning.

In addition to the baseline training protocol, Fourier feature injection is evaluated as an ablation in the shock-dominated task. This ablation is designed to test whether explicit high-frequency coordinate embeddings help models resolve sharp shock structures, or whether they introduce oscillatory artifacts, nonphysical noise, or frequency-aliasing effects. The ablation therefore directly connects the model benchmark to the frequency-representation issue discussed in Section 3.1.

5.3. Masked optimization and evaluation metrics

For configurations containing solid bodies, the computational domain includes both fluid regions and nonphysical interior regions. To prevent these invalid regions from affecting training and evaluation, we use a binary geometry mask $M(\mathbf{x})$, where $M(\mathbf{x}) = 1$ denotes fluid points and $M(\mathbf{x}) = 0$ denotes solid or excluded regions. During training, models are optimized using a masked mean squared error loss,

$$\mathcal{L}_{\text{MSE}} = \frac{1}{\sum_{i=1}^N M(\mathbf{x}_i)} \sum_{i=1}^N M(\mathbf{x}_i) \|\mathcal{U}_\theta(\mathbf{x}_i) - \mathcal{U}(\mathbf{x}_i)\|_2^2, \quad (14)$$

where $\mathcal{U}_\theta$ is the model prediction, $\mathcal{U}$ is the reference solution, and N is the number of spatial grid points. The MSE loss serves as the optimization objective and penalizes large local deviations more strongly, given its squared-error form. The same mask is used during evaluation so that errors are computed only over the physically valid fluid region.

For quantitative evaluation, we report the masked MSE together with two complementary metrics. The reported MSE is consistent with the training objective and emphasizes larger pointwise deviations because of its squared-error form. The first complementary metric is the masked mean absolute error (MAE),

$$MAE = \frac{1}{\sum_{i=1}^N M(\mathbf{x}_i)} \sum_{i=1}^N M(\mathbf{x}_i) \|\mathcal{U}_\theta(\mathbf{x}_i) - \mathcal{U}(\mathbf{x}_i)\|_1. \quad (15)$$

MAE measures the average pointwise deviation between prediction and reference data. Compared with the squared training loss, it provides a more direct and interpretable measure of local reconstruction accuracy and is less dominated by a small number of large-error points. Besides, the masked relative L_2 error is defined as

$$\text{Relative } L_2 = \frac{\sqrt{\sum_{i=1}^N M(\mathbf{x}_i) \|\mathcal{U}_\theta(\mathbf{x}_i) - \mathcal{U}(\mathbf{x}_i)\|_2^2}}{\sqrt{\sum_{i=1}^N M(\mathbf{x}_i) \|\mathcal{U}(\mathbf{x}_i)\|_2^2}}. \quad (16)$$

The relative L_2 error measures the normalized global discrepancy of the predicted field. By scaling the error with the magnitude of the reference solution, it enables more meaningful comparisons across variables and flow regimes with different physical scales. For the double-cone task, the relative L_2 error is computed after mapping the normalized predictions and reference fields back to the physical-variable space. This inverse reconstruction includes the inverse logarithmic transformation introduced during preprocessing. The reason is that the logarithmic compression is mainly used to stabilize training and prevent large-magnitude shock-related variables, such as pressure, from dominating the optimization. If the relative L_2 error were evaluated only in the log-normalized space, it would measure the discrepancy of the

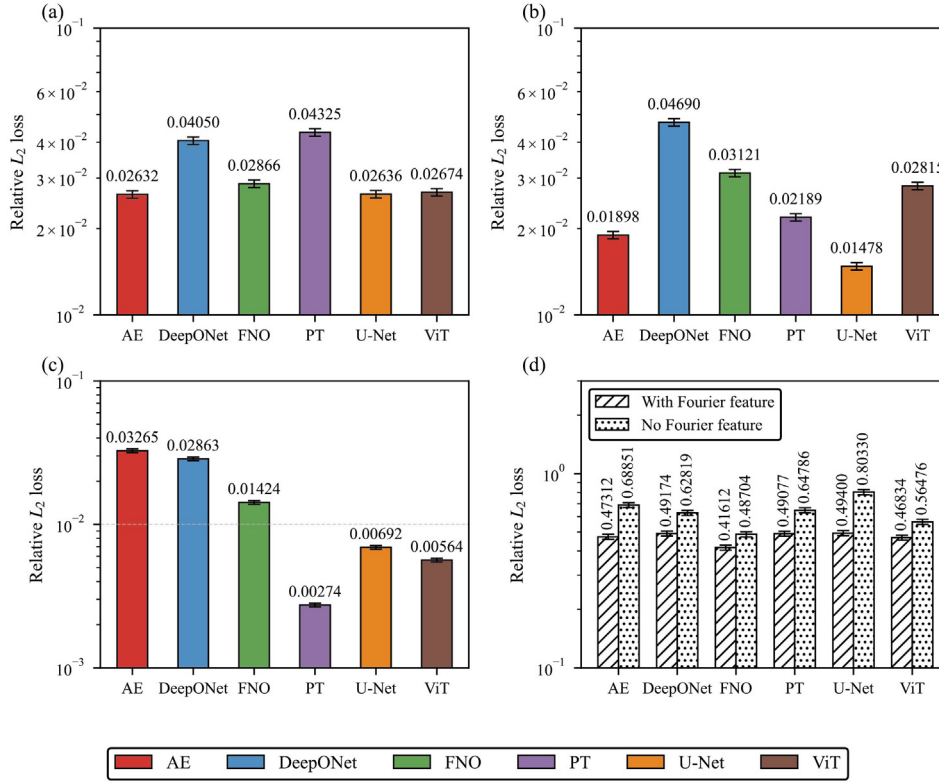


Fig. 6. Quantitative performance overview across the four TransportBench tasks. Subfigures (a)-(c) report the relative L_2 error of six baseline architectures on Tasks I-III. The results show regime-dependent model behavior: AE and U-Net achieve low errors in the airfoil task, U-Net performs best in the cylinder task, and PT achieves the lowest error in the cavity task. Subfigure (d) reports the Fourier-feature ablation on the double-cone task, showing that high-frequency feature injection reduces relative L_2 error for all tested architectures, with architecture-dependent improvement magnitudes.

compressed variables and could underestimate physically significant errors near shock-induced peaks. Therefore, evaluating relative L_2 in the reconstructed physical space provides a more faithful measure of the relative error associated with strong shocks and large property variations.

In summary, the masked MSE is used as the training objective and is also reported as an evaluation quantity, while MAE and relative L_2 provide complementary views of prediction accuracy. MAE characterizes average local accuracy, MSE emphasizes larger pointwise deviations, and relative L_2 measures field-level agreement after scaling by the magnitude of the reference solution. For the double-cone task, relative L_2 is further evaluated after inverse reconstruction to the physical-variable space so that shock-induced peak errors are not underestimated by the log-normalized representation.

6. Numerical experiments and analysis

In this section, we evaluate the six baseline architectures introduced in Section 5 on the four TransportBench tasks. The goal is not only to compare prediction errors, but also to examine how different architectural inductive biases behave under distinct non-equilibrium flow conditions.

Figure 6 provides an overview of the quantitative results. A clear regime dependence can be observed: no single architecture consistently performs best across all tasks. Convolutional and latent-representation models perform favorably in the geometry-varying airfoil task; U-Net, autoencoder, and Point Transformer show strong performance in the parameter-varying cylinder task; and Point Transformer achieves the lowest error in the cavity task, followed by Vision Transformer and U-Net. The double-cone task further highlights the difficulty of resolving localized shock discontinuities under sparse and highly anisotropic rep-

Table 4

Quantitative performance on Task I: airfoil flow with geometry-dependent prediction. MAE, MSE, and relative L_2 error are reported for the six baseline architectures on the test set.

Model	MAE	MSE (L2 Loss)	Relative L2
AE	0.0120010	0.0002772	0.02632
DeepONet	0.0136960	0.0006563	0.04050
FNO	0.0124002	0.0003291	0.02866
PT	0.0142392	0.0007484	0.04325
U-Net	0.0120156	0.0002781	0.02636
ViT	0.0121577	0.0002869	0.02674

resentations, and shows that Fourier feature injection has architecture-dependent effects.

6.1. Task I: airfoil flow with geometry-dependent prediction

Task I evaluates the ability of models to predict rarefied flow fields around airfoils with geometry variations. The training and test samples share the same physical operating conditions, while the airfoil shapes vary according to the CST perturbations described in Section 3.3. Therefore, this task focuses on whether a model can learn the relation between solid-boundary geometry and the corresponding non-equilibrium flow structures.

The quantitative results are summarized in Table 4, which corresponds to Fig. 6(a). The autoencoder (AE) and U-Net achieve the lowest relative L_2 errors among the tested baselines (0.02632 and 0.02636, respectively). Vision Transformer (ViT) performs closely, with a relative L_2 of 0.02674, while FNO remains competitive at 0.02866. In contrast, DeepONet and Point Transformer yield larger errors, suggesting that

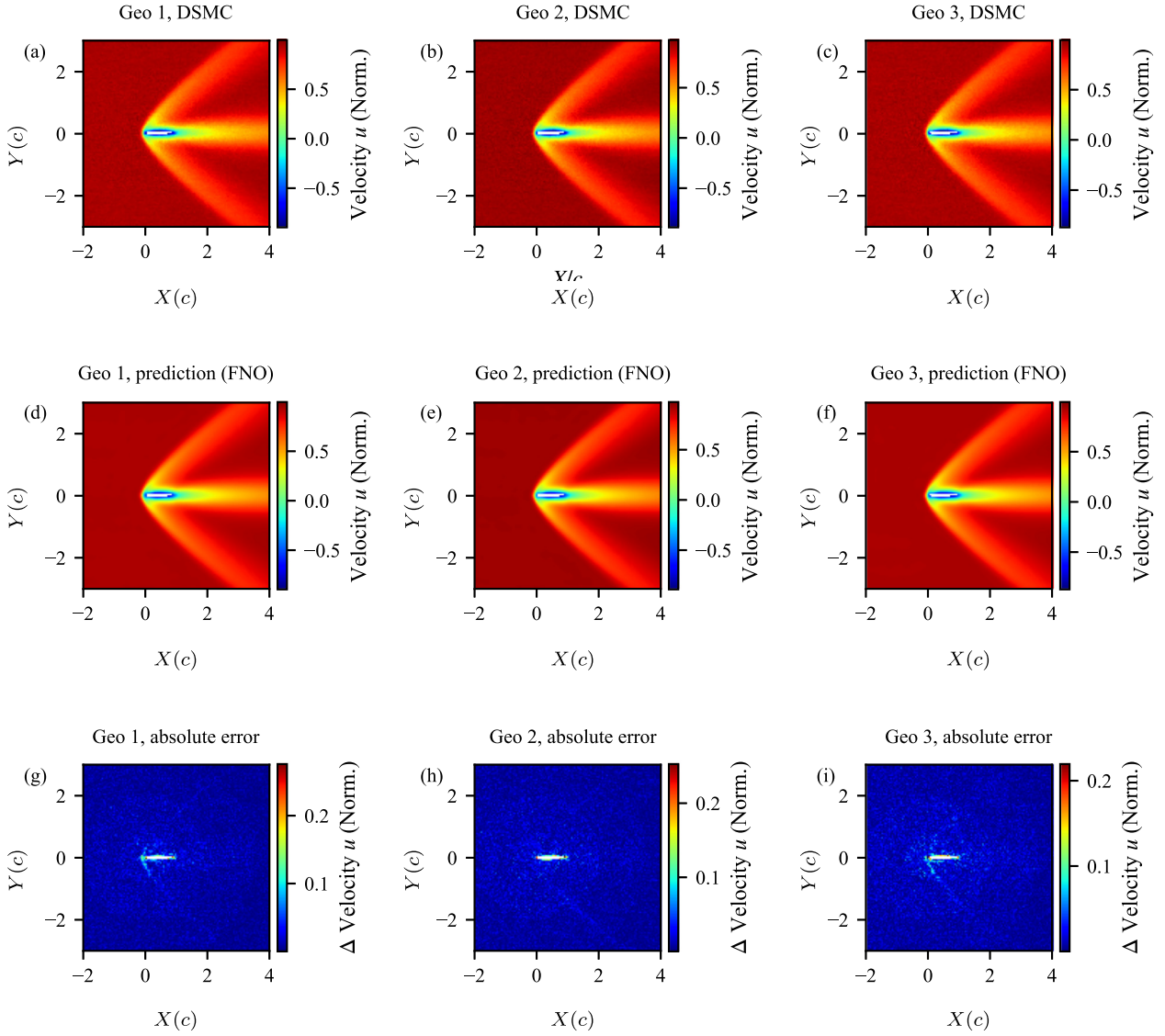


Fig. 7. Representative FNO predictions on Task I: rarefied airfoil flow. The rows show the DSMC reference solution, FNO prediction, and absolute error, respectively, for three representative airfoil geometries. FNO reconstructs the geometry-dependent shock and wake structures with relatively low localized errors.

coordinate-based field reconstruction and point-based aggregation can be less effective for this particular geometry-dependent prediction setting.

These results suggest that structured-grid spatial representations are well-suited for the airfoil task. The strong performance of AE and U-Net suggests that local spatial correlations and compact field representations are effective for mapping airfoil-shape variations to the corresponding shock and wake structures. The competitive performance of FNO further indicates that global spectral coupling can also be beneficial, especially for capturing coherent flow-field patterns on structured grids. By contrast, DeepONet lacks explicit local spatial priors, which may limit its ability to resolve geometry-induced shock structures near the leading edge.

Figures 7 and 8 compare predictions from coordinate-based (DeepONet) and Fourier-based (FNO) representations. Both models recover the major macroscopic flow features, including the bow shock and downstream wake. Because the reference data are generated by DSMC and therefore contain statistical fluctuations, the smoother model predictions also indicate an implicit denoising effect. However, clear differences appear near the leading edge and along the shock front. FNO

(Fig. 7) produces a more accurate reconstruction with relatively uniform error distribution, whereas DeepONet (Fig. 8) shows larger discrepancies concentrated around the leading-edge region and the V-shaped shock structure. Overall, Task I indicates that models with structured-grid spatial priors are well-suited for predicting geometry-dependent flow fields when the main challenge lies in mapping boundary-shape variation to shock and wake structures.

To quantitatively evaluate the predictive quality of localized structures under geometry variations, we further extract the 1D temperature profile along the stagnation line upstream of the airfoil leading edge. As shown in Fig. 9, the interaction between the hypersonic freestream and the airfoil geometry generates a distinct bow shock, characterized by a sharp temperature jump. Consistent with the 2D contour observations, the 1D profile explicitly reveals that coordinate-based architectures (e.g., DeepONet) suffer from spectral bias, heavily smearing the shock front and underpredicting the peak temperature. Conversely, U-Net accurately captures the steep gradient of the shock and closely aligns with the DSMC ground truth. This confirms that preserving local spatial priors is vital for resolving geometry-induced non-equilibrium shock structures.

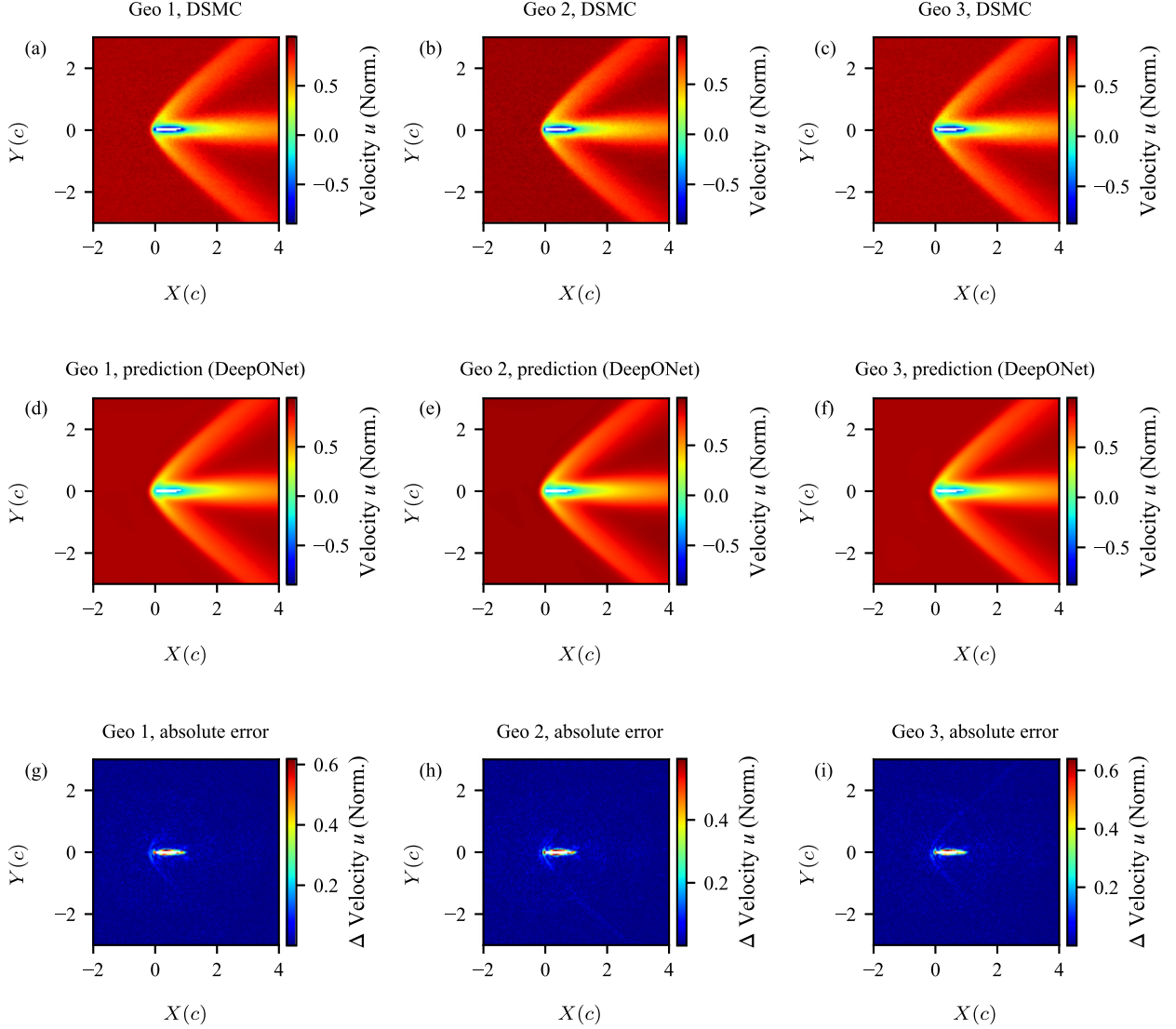


Fig. 8. Representative DeepONet predictions on Task I: rarefied airfoil flow. The rows show the DSMC reference solution, DeepONet prediction, and absolute error, respectively, for three representative airfoil geometries. DeepONet exhibits larger discrepancies near the leading edge and along the V-shaped shock front.

6.2. Task II: cylinder flow with parameter-dependent prediction

Task II examines prediction across a broad range of Mach and Knudsen numbers for multi-scale flow around a circular cylinder. Unlike Task I, where the geometry changes but the operating conditions remain fixed, this task requires models to learn how flow topology varies with compressibility and rarefaction. Across the parameter space, the bow shock, wake length, and degree of diffusion change substantially.

The quantitative results are summarized in Table 5, which corresponds to Fig. 6(b). U-Net achieves the lowest errors across all metrics (relative L_2 of 0.01478). AE also performs strongly, reaching a relative L_2 of 0.01898, followed by PT at 0.02189. FNO and ViT show intermediate performance, whereas DeepONet yields the largest error among the tested baselines. These results suggest that parameter-dependent cylinder flow benefits from architectures that can preserve local spatial structures while adapting to large-scale changes in shock and wake topology.

The strong performance of U-Net suggests that local convolutional priors may be beneficial in this task. Although the wake location and shock structure vary with Ma and Kn , the underlying flow fields remain organized on a structured spatial representation, where local gradients, boundary-layer features, and wake transitions can be captured

Table 5

Quantitative performance on Task II: cylinder flow with parameter-dependent prediction. MAE, MSE, and relative L_2 error are reported for the six baseline architectures across unseen combinations of Mach and Knudsen numbers.

Model	MAE	MSE (L2 Loss)	Relative L2
AE	0.0086424	0.0002037	0.01898
DeepONet	0.0161770	0.0010977	0.04690
FNO	0.0093879	0.0005211	0.03121
PT	0.0084134	0.0002568	0.02189
U-Net	0.0065303	0.0001158	0.01478
ViT	0.0103630	0.0004536	0.02815

efficiently by convolutional kernels. The competitive performance of AE further suggests that the dominant flow variations in this dataset are sufficiently coherent to be represented through a compact latent space. Point Transformer also performs well, indicating that pointwise aggregation can adapt to parameter-induced spatial deformation, particularly changes in wake extent and shock position.

Figures 10 and 11 compare representative predictions from DeepONet and U-Net under different parameter combinations. DeepONet

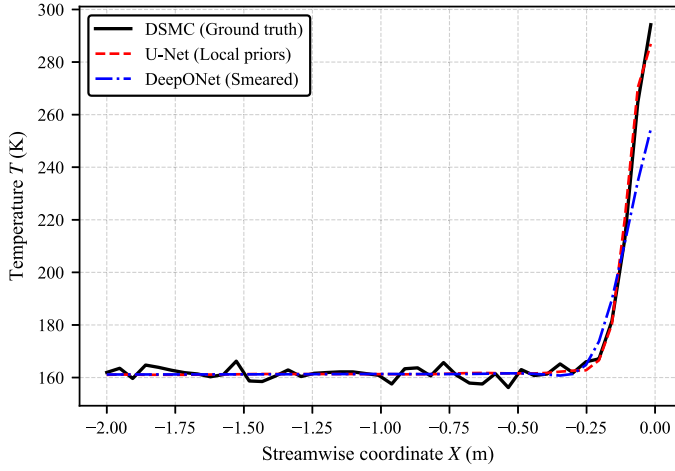


Fig. 9. Comparison of the 1D temperature profile along the stagnation line ($y = 0$) upstream of the airfoil. DSMC reference (black solid line), U-Net (red dashed line), and DeepONet (blue dash-dotted line) are compared. U-Net effectively captures the steep shock gradient, whereas DeepONet exhibits a smeared shock profile.

produces smooth flow fields but tends to misalign shock fronts and wake structures when the operating conditions vary substantially. This suggests that, in the present architecture, scalar inputs such as Ma and Kn , combined with pointwise coordinate queries, are less effective for reconstructing globally shifting flow topology. U-Net captures the main shock

and local flow structures more accurately. Nevertheless, localized error accumulation can still be observed in the extended wake region under more challenging parameter combinations, indicating that local convolutional priors do not completely remove the difficulty of parameter-dependent topological variation.

These results show that parameter-dependent rarefied-flow prediction requires models to capture both local gradients and global topological changes. Architectures with strong local spatial priors (like U-Net) or flexible spatial point aggregation (like Point Transformer) are therefore associated with favorable performance in this task. To explicitly demonstrate the architectures' capabilities in resolving sharp discontinuities and near-wall non-equilibrium effects—which are often obscured by global metrics—we also extract the 1D temperature profile along the stagnation line in front of the cylinder. As shown in Fig. 12, the flow experiences drastic shock heating at $x \approx -1.1$ and a rapid temperature drop within the Knudsen layer near the cold wall ($x = -0.5$). DeepONet completely smears the sharp bow shock, producing a non-physical early temperature rise and failing to capture the distinct boundaries of the shock wave. In contrast, U-Net accurately captures both the steep gradient of the shock front and the precise temperature jump near the wall. This diagnostic comparison strongly supports the conclusion that local convolutional priors are highly advantageous for reconstructing shock-dominated regions across varying operating conditions.

6.3. Task III: lid-driven cavity flow with high-order kinetic prediction

Task III evaluates the ability of neural models to predict both the particle distribution functions and the associated macroscopic moment fields. Compared with Tasks I and II, this task places less emphasis

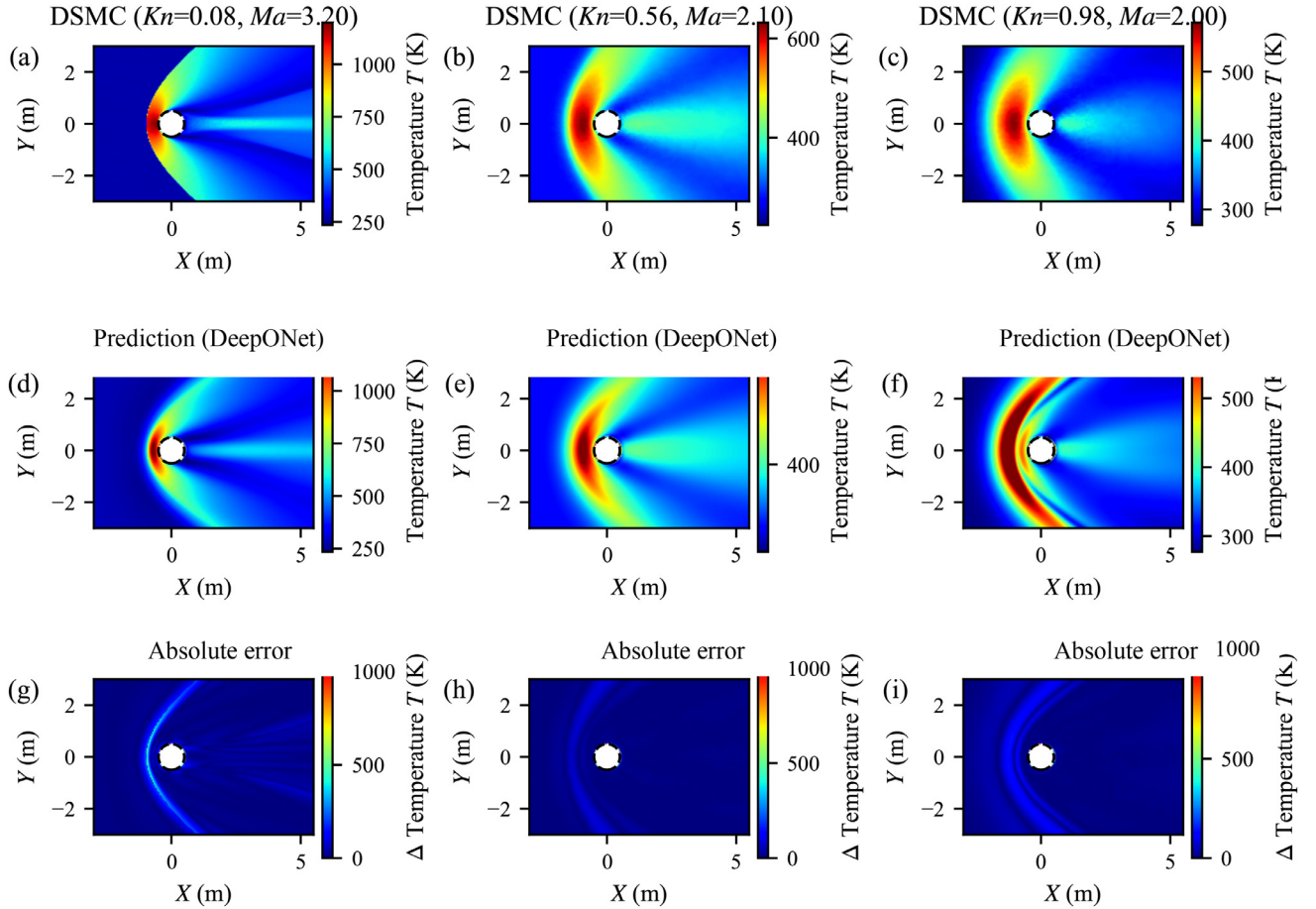


Fig. 10. Representative DeepONet predictions on Task II: rarefied cylinder flow. The rows show the DSMC reference solution, DeepONet prediction, and absolute error under three representative combinations of Kn and Ma . DeepONet produces smooth flow fields but tends to misalign shock and wake structures under large parameter variations.

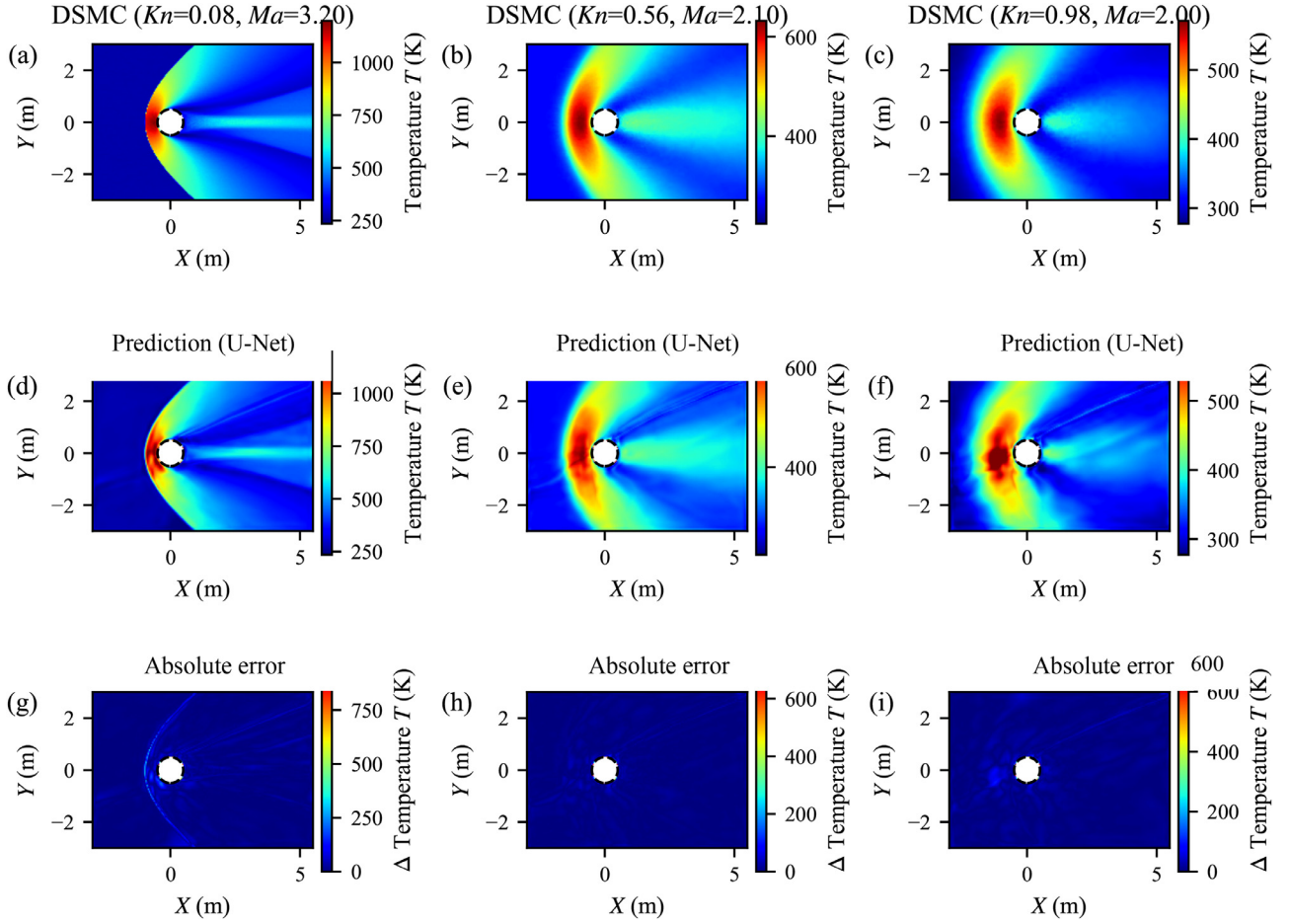


Fig. 11. Representative U-Net predictions on Task II: rarefied cylinder flow. The rows show the DSMC reference solution, U-Net prediction, and absolute error under three representative combinations of Kn and Ma . U-Net captures local shock and near-wall features more accurately, while localized errors remain visible in the extended wake region under challenging parameter combinations.

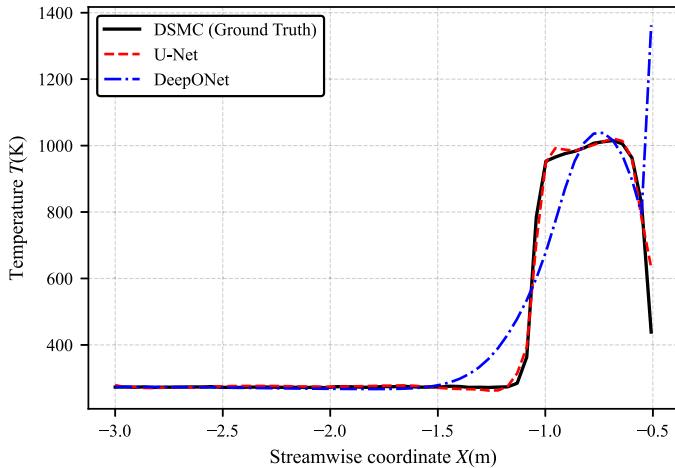


Fig. 12. Comparison of the 1D temperature profile along the stagnation line in front of the cylinder. The extraction highlights the capability of different architectures to resolve the sharp bow shock and the near-wall temperature jump within the Knudsen layer.

on complex geometry or strong shock structures and instead focuses on whether a model can capture the micro-macro structure of non-equilibrium flow transport. The distribution function contains the microscopic velocity-space information, while quantities such as density,

Table 6

Quantitative performance on Task III: lid-driven cavity flow with high-order kinetic prediction. MAE, MSE, and relative L_2 error are reported for prediction of particle distribution functions and associated macroscopic or high-order moment fields.

Model	MAE	MSE (L2 Loss)	Relative L2
AE	0.0164398	0.0011465	0.03265
DeepONet	0.0087809	0.0008484	0.02863
FNO	0.0026017	0.0002289	0.01424
PT	0.0011773	0.0000093	0.00274
U-Net	0.0042013	0.0000511	0.00692
ViT	0.0028403	0.0000370	0.00564

velocity, stress tensor, and heat flux are obtained from its velocity moments. Therefore, this task provides a direct test of whether neural models can learn not only macroscopic flow patterns, but also the mesoscopic gas-kinetic information.

The quantitative results are summarized in Table 6, which corresponds to Fig. 6(c). PT achieves the lowest errors across all metrics with a relative L_2 of 0.00274. ViT and U-Net also perform well, with relative L_2 errors of 0.00564 and 0.00692, respectively. These results indicate that the cavity task does not simply favor a broad model category such as operator learning or grid-based learning. Instead, performance depends on how effectively each architecture represents spatial correlations, boundary-induced non-equilibrium structures, and the coupling between distribution functions and their moment fields.

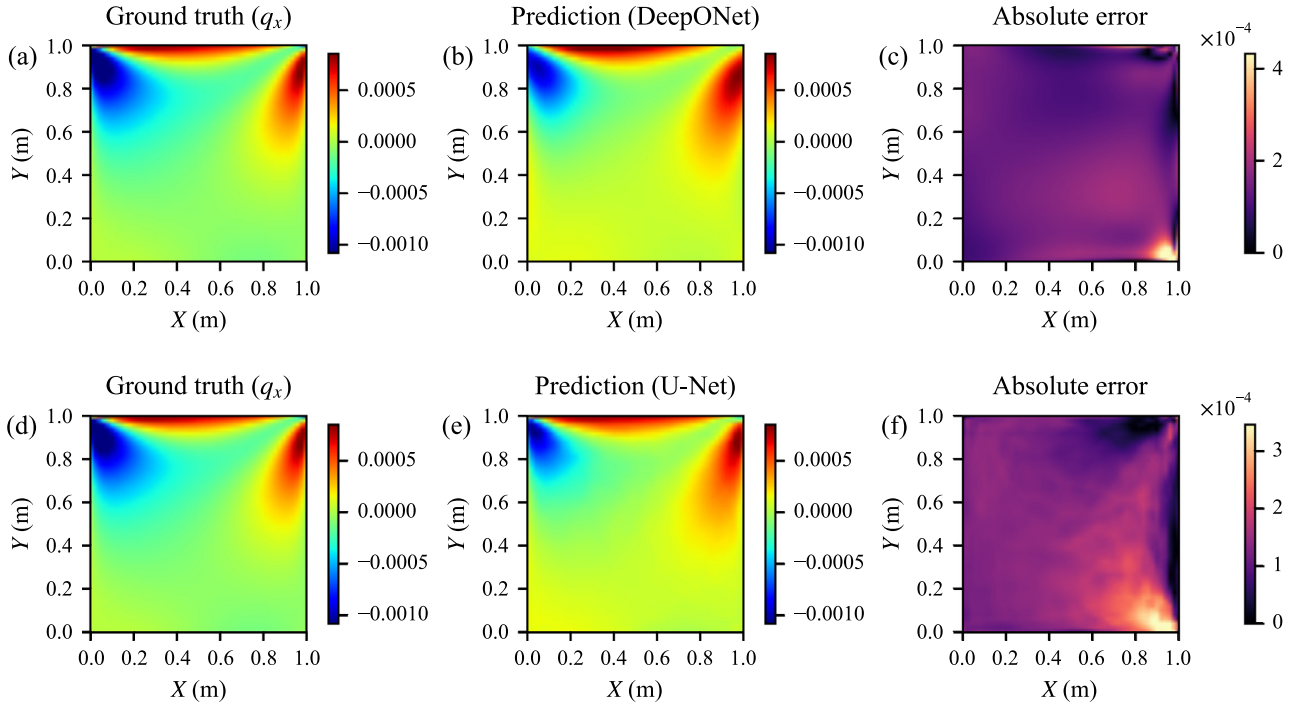


Fig. 13. Comparison on Task III: lid-driven cavity flow. The figure compares reference, predicted, and absolute-error fields for the heat-flux component q_x . DeepONet produces visually smooth predictions but oversmooths some non-equilibrium variations, whereas U-Net achieves much lower overall errors while exhibiting block-like artifacts near the cavity corners. This comparison illustrates two representative error patterns in high-order kinetic-moment prediction.

The strong performance of Point Transformer suggests that point-wise aggregation can be effective for this smooth but physically coupled kinetic problem. Although the cavity geometry is simple, the distribution function and high-order moments are influenced by wall-driven shear, corner effects, and rarefaction. These effects generate spatially distributed correlations that may be captured efficiently by flexible point-based aggregation. The competitive performance of ViT also suggests that token-level global interactions are useful for representing non-local correlations in the confined cavity domain. U-Net performs well overall, benefiting from local convolutional priors, but its structured-grid representation might introduce localized artifacts near the corners.

Figure 13 provides further insight into the representational biases of these models by comparing the heat-flux predictions of DeepONet and U-Net. DeepONet produces visually smooth predictions due to its continuous coordinate-based representation, but it over-smooths some non-equilibrium variations and yields relatively large quantitative errors. U-Net achieves a much lower error but exhibits block-like artifacts near the cavity corners. This comparison illustrates that visual smoothness alone is not sufficient for accurate high-order kinetic prediction, while grid-based local reconstruction may introduce discretization-related artifacts in regions affected by boundary singularities.

Figure 14 further examines the prediction results in the velocity space. It compares the reference and predicted distribution functions at representative spatial locations, i.e., the cavity center and the geometric singularity point at the top-right corner. At the cavity center (near-equilibrium region), the distribution function remains close to a Maxwellian shape, whereas near boundary-driven shear regions it becomes visibly distorted and anisotropic. Accurately reproducing these velocity-space structures is important because the high-order moments, such as the stress tensor and heat flux, are direct integrals of the same distribution function. Therefore, agreement at the distribution-function level provides stronger evidence that the model captures the underlying kinetic non-equilibrium, rather than merely fitting smooth macroscopic contours.

Overall, Task III shows that predicting non-equilibrium cavity flow involves a coupled micro-macro learning problem. A successful model

should reproduce the distribution function in velocity space and maintain consistency with the macroscopic and higher-order moments derived from it. This makes this task a useful benchmark for evaluating neural solvers on kinetic quantities beyond standard continuum-flow variables.

6.4. Task IV: double-cone flow with shock-dominated field reconstruction

Task IV is a demanding task in TransportBench. It involves high-enthalpy nitrogen flow over a double-cone configuration, with thermochemical non-equilibrium, strong shock discontinuities, shock-boundary-layer interaction, limited training samples, and a sparse anisotropic representation. The task presents a severe representational challenge for neural models because accurate prediction requires resolving localized shock peaks and steep gradients. Besides, the state-to-state model introduces a large number of coupled state-specific variables and stiff source terms associated with vibrational relaxation and dissociation-recombination processes. Therefore, the target fields are compressed manifestations of a high-dimensional thermochemical relaxation process, rather than merely shock contours to be interpolated.

We evaluate the six baseline architectures with and without Fourier feature (FF) high-frequency injection. This ablation is designed to examine whether explicit high-frequency coordinate embeddings improve shock resolution and how their effect depends on the underlying model architecture.

6.4.1. Local convolutional priors in shock-field reconstruction

Among the tested baselines, U-Net without Fourier feature injection achieves the lowest MAE and MSE on the double-cone task, indicating strong local reconstruction capability. However, it does not achieve the lowest relative L2 error. Instead, FNO with Fourier feature injection obtains the best relative L2 performance. The near-wall pressure profile in Fig. 15 further shows that U-Net without Fourier features captures the main shock-induced pressure peaks and the overall shock-boundary-layer interaction with relatively low error.

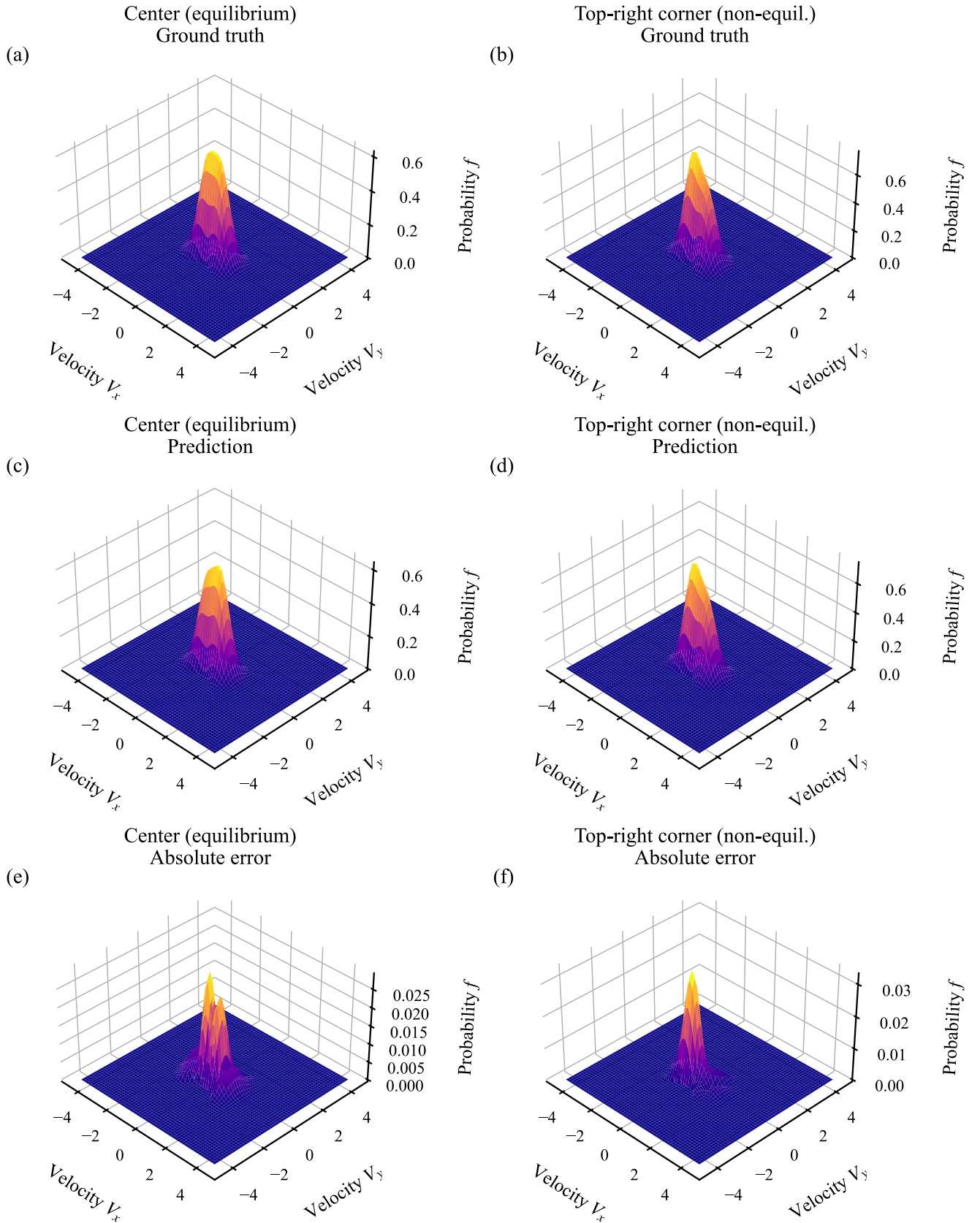


Fig. 14. Particle distribution function prediction in the lid-driven cavity flow. The figure compares the reference distribution function, FNO prediction, and absolute error at two representative spatial locations. The left column corresponds to the cavity center, where the distribution remains close to a near-Maxwellian shape. The right column corresponds to the top-right corner, where boundary-driven shear and rarefaction effects lead to an anisotropic and distorted distribution. Since macroscopic quantities and high-order moments are associated with velocity-space integration of the distribution function, this comparison directly evaluates whether the model captures the kinetic structures underlying the predicted moment fields.

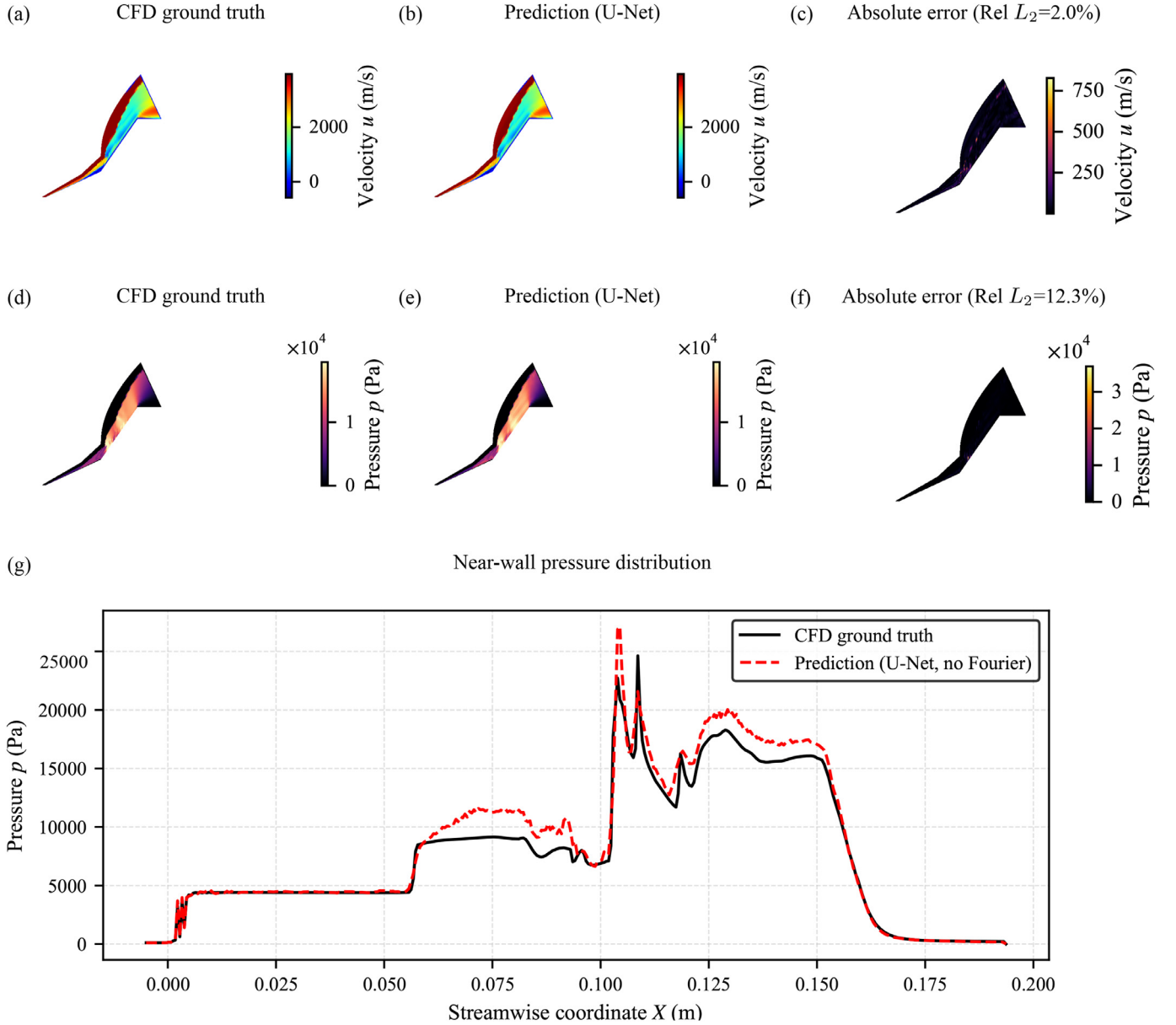


Fig. 15. U-Net prediction on the double-cone task without Fourier feature injection. The figure compares reference and predicted velocity and pressure fields, together with the absolute-error fields and near-wall pressure distribution. U-Net captures the main shock-induced pressure peaks and the overall shock-boundary-layer interaction with relatively low error in this setting, illustrating the advantage of local convolutional priors for shock-dominated field reconstruction.

This result indicates that local convolutional priors may be useful for shock-dominated field reconstruction. The local receptive fields of U-Net help preserve sharp spatial variations, while its encoder-decoder structure and skip connections combine local detail with broader contextual information. In the double-cone case, this locality is important not only for resolving shock fronts, but also for representing the rapid downstream variations induced by shock heating and state-specific relaxation.

This finding also highlights an important distinction between smooth-field prediction and shock-dominated prediction. Models that perform well on smooth or globally correlated fields do not necessarily resolve narrow shock structures accurately, especially when training data are scarce and the spatial representation is highly anisotropic.

6.4.2. Spectral bias and oscillatory artifacts

Coordinate-based models show limitations in resolving sharp shock peaks. DeepONet without Fourier feature injection captures the over-

all flow topology but tends to smooth localized pressure maxima near shock-interaction regions, as shown in Fig. 16. This behavior is consistent with the spectral-bias issue discussed in Section 3.1: coordinate-based models may favor smoother, low-frequency mappings and therefore underestimate high-frequency discontinuities.

Fourier-based models face a different challenge. FNO can capture long-range correlations and performs well in several tasks, but its global spectral representation may produce oscillatory artifacts near sharp discontinuities. When Fourier feature injection is applied, FNO improves the reconstruction of some high-frequency structures, but the pressure field may exhibit nonphysical oscillations, as illustrated in Fig. 17. This indicates that a lower global error does not necessarily eliminate local oscillatory artifacts near sharp discontinuities.

Together, Figs. 16 and 17 show two different challenges in shock-dominated prediction: low-frequency smoothing and oscillatory errors near sharp discontinuities. These observations suggest that resolving shock-dominated fields requires more than increasing model capacity

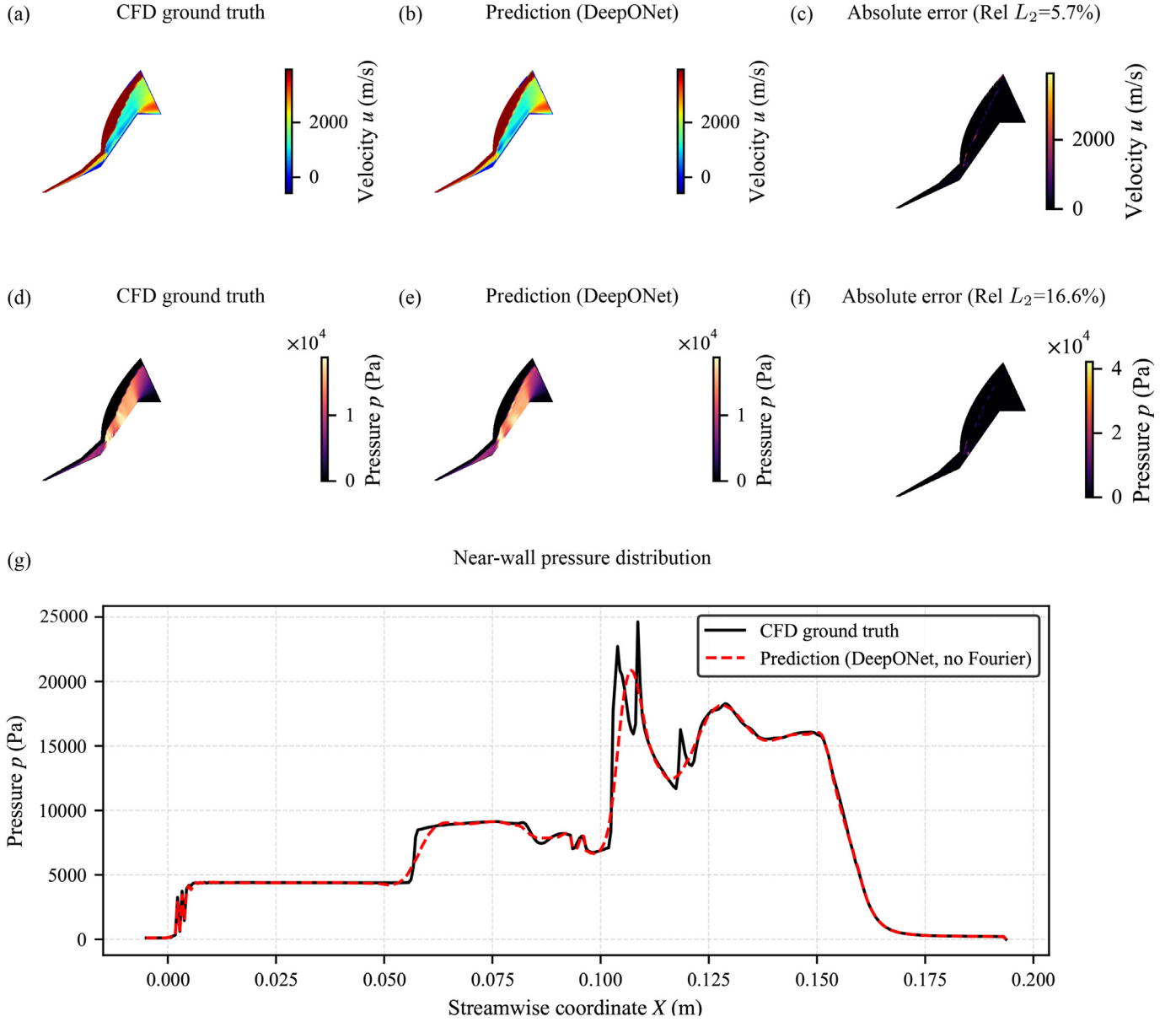


Fig. 16. Representative DeepONet prediction on the double-cone task without Fourier feature injection. The figure compares CFD reference fields, DeepONet predictions, absolute-error fields, and the near-wall pressure distribution. DeepONet captures the overall flow topology but smooths localized shock-induced pressure peaks, illustrating the spectral-bias limitation of coordinate-based field reconstruction in shock-dominated flows.

or adding high-frequency input features. The observed performance of a model may depend on the interaction between its architectural priors, data representation, optimization behavior, and the localized and discontinuous nature of the target flow field.

6.4.3. Architecture-dependent effect of Fourier feature injection

Table 7 shows that the effect of Fourier feature injection is metric-dependent. In terms of relative L_2 error, Fourier features improve all tested architectures. The improvement is most pronounced for U-Net, AE, PT, and DeepONet, indicating that high-frequency coordinate embeddings can enhance physical-space relative field agreement in shock-dominated reconstruction.

However, the same conclusion does not hold uniformly for MAE and MSE. DeepONet, PT, and ViT improve across all three metrics after Fourier feature injection, suggesting that these architectures benefit directly from explicit high-frequency information. FNO obtains the best relative L_2 error with Fourier features, although its MSE changes only

slightly. In contrast, AE and U-Net exhibit a trade-off: Fourier features reduce relative L_2 but increase MAE and MSE. For U-Net in particular, the no-Fourier version remains the best in absolute-error metrics, while the Fourier version substantially improves relative L_2 .

This metric-dependent discrepancy arises from the different evaluation spaces used by the absolute-error metrics and the relative L_2 metric. For stable learning and balanced multi-variable assessment, MAE and MSE are computed in the normalized learning space. This is especially important for the double-cone task, where shock-induced pressure and thermodynamic variables can span several orders of magnitude. Without logarithmic compression and normalization, the training loss and absolute-error metrics would be dominated by a few large-magnitude variables and localized peak regions. Therefore, normalized MAE and MSE mainly reflect the average reconstruction quality over the full masked field, including the broad smooth background.

In contrast, relative L_2 is computed after inverse reconstruction to the physical-variable space. For the double-cone task, this inverse recon-

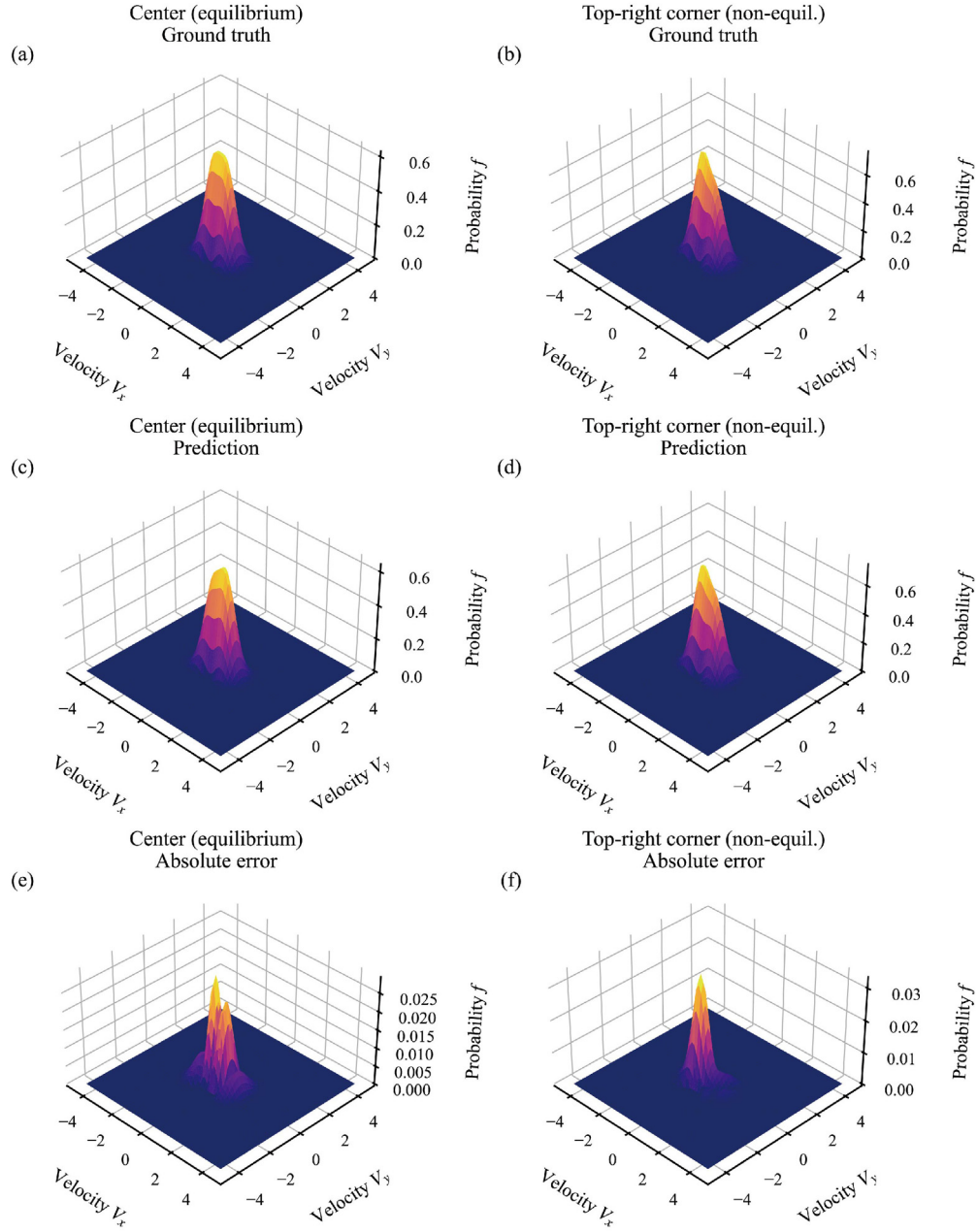


Fig. 17. Representative FNO prediction on the double-cone task with Fourier feature injection. The figure compares CFD reference fields, FNO predictions, absolute-error fields, and the near-wall pressure distribution. FNO captures some sharp structures and achieves strong global relative-error performance, but the pressure profile exhibits oscillatory discrepancies near shock-induced peaks, illustrating a local artifact associated with spectral representations in discontinuity-dominated fields.

struction includes the exponential inverse of the logarithmic transformation. As a result, small errors in the log-normalized representation near shock-induced peaks can correspond to large absolute discrepancies in physical space. Fourier feature injection helps models represent these sharp high-frequency structures more accurately and therefore reduces the physical-space relative L_2 error for all tested architectures. However, for architectures with strong local spatial priors, such as AE and U-Net, the same high-frequency embeddings may also introduce small oscillatory errors or background noise in smoother regions, increasing normalized MAE and MSE. Therefore, the results in Table 7 should be interpreted as a trade-off between full-field normalized reconstruction accuracy and physical-space shock-sensitive relative accuracy.

These results indicate that Fourier feature injection should not be treated as a universally positive or negative modification. Its benefit

depends on the interaction between model topology, normalization, and evaluation metric. In shock-dominated thermochemical flows, relative L2, MAE, MSE, and qualitative near-wall behavior should be considered together.

7. Discussion

The experiments on TransportBench show that learning non-equilibrium flow transport is strongly shaped by both the underlying physical regime and the model's architectural prior. Beyond reporting prediction errors, the benchmark provides diagnostic evidence on when different model families succeed or fail. We summarize the main implications from the following perspectives. Regime-dependent inductive bias

Table 7

Quantitative performance on Task IV: double-cone flow. MAE and MSE are computed in the normalized learning space, while relative L2 is computed after inverse reconstruction to the physical-variable space. The ablation study highlights the architecture- and metric-dependent impact of explicit Fourier feature injection in shock-dominated fields.

Model	Feature	MAE	MSE (L2 loss)	Relative L2
AE	With Fourier	0.0325	0.0058	0.47312
	No Fourier	0.0313	0.0052	0.68851
DeepONet	With Fourier	0.0397	0.0058	0.49174
	No Fourier	0.0565	0.0121	0.62819
FNO	With Fourier	0.0312	0.0060	0.41612
	No Fourier	0.0331	0.0059	0.48704
PT	With Fourier	0.0373	0.0074	0.49077
	No Fourier	0.0736	0.0210	0.64786
U-Net	With Fourier	0.0296	0.0055	0.49400
	No Fourier	0.0259	0.0046	0.80330
ViT	With Fourier	0.0335	0.0057	0.46834
	No Fourier	0.0378	0.0063	0.56476

A central observation from TransportBench is that model performance is strongly regime-dependent. No single architecture consistently achieves the best performance across all tasks. In the airfoil task, AE and U-Net achieve the lowest errors, while FNO remains competitive, suggesting that structured-grid spatial priors, local convolutional reconstruction, compact latent representation, and global spectral coupling can all be useful for geometry-dependent rarefied-flow prediction. In the cylinder task, U-Net performs best, followed by AE and Point Transformer, indicating that parameter-dependent flow prediction benefits from local spatial priors while also requiring adaptability to large-scale shock and wake deformation. In the cavity task, Point Transformer achieves the lowest error, followed by ViT and U-Net, showing that pointwise aggregation and token-level interactions can be effective for smooth but physically coupled kinetic and moment fields. In the double-cone task, U-Net without Fourier feature injection achieves the lowest MAE and MSE, highlighting the value of local convolutional priors for absolute shock-field reconstruction, whereas FNO with Fourier feature injection achieves the lowest relative L2 error. This result shows that shock-dominated reconstruction requires metric-aware assessment rather than a single overall ranking.

These results indicate that model selection should be guided by the dominant structure of the target problem rather than by a presumed universal architecture. Geometry-dependent prediction, parameter-induced topology variation, micromacro kinetic prediction, and shock-dominated hypersonic reconstruction impose different requirements on neural architectures. TransportBench therefore serves not only as a leaderboard-style benchmark, but also as a diagnostic tool for examining which model behaviors and architectural priors are associated with favorable performance in different non-equilibrium transport regimes. Beyond parameter scaling and black-box benchmarking

The benchmark also shows that increasing model capacity alone is not sufficient to overcome the difficulty of non-equilibrium flow prediction. This is especially clear in the double-cone task, where the parameter budget is enlarged to reduce severe under-parameterization, yet the observed errors remain strongly architecture-dependent. A model with large capacity can still fail when its spatial representation, frequency bias, or aggregation mechanism is poorly aligned with localized shock structures and thermochemical relaxation effects. Without standardized parameter budgets, grid resolutions, preprocessing strategies, training protocols, and evaluation metrics, performance gains may be difficult to attribute, since they may result from larger models, more aggressive tuning, or higher computational effort rather than from a more suitable architectural inductive bias.

The results further suggest that model assessment should not rely on a single aggregate number. In the double-cone Fourier-feature ablation, for example, high-frequency injection may improve one metric for a

specific model while degrading another, indicating that local accuracy, global field agreement, and physical behavior near shocks or boundaries should be considered together. TransportBench addresses these issues by evaluating representative architectures under unified settings, making it easier to separate architecture-related behavior from capacity-driven improvements and to interpret model performance in relation to physical flow structures. As noted in Section 6, the architecture-related observations in this benchmark are diagnostic rather than strictly causal, since the compared models differ not only in architecture but also in input representation, optimization dynamics, and implementation details. Establishing stricter causal relationships would require more controlled ablation studies, in which a single architectural factor is varied while the remaining components are held fixed, which we leave for future work. High-order variables and frequency-aware representations

TransportBench extends conventional fluid benchmarks by including kinetic-level quantities and high-order non-equilibrium statistics. In the cavity task, models are evaluated on both particle distribution functions and the macroscopic or high-order moment fields derived from them. These quantities are directly connected to the velocity-space structure of the Boltzmann solution and are more sensitive to rarefaction and non-equilibrium effects than standard macroscopic variables. Accurate prediction therefore requires more than reproducing density, velocity, or temperature fields; it requires preserving the micromacro consistency between the distribution function and its moments.

At the same time, the double-cone task shows that shock-dominated hypersonic flows pose a different but equally important representational challenge. Models that favor smooth or low-frequency mappings may capture the overall flow topology but under-resolve localized shock peaks. Explicit Fourier feature injection can mitigate this tendency in some cases, such as reducing the relative L_2 error of DeepONet, but it is not universally beneficial. In the double-cone results, explicit Fourier feature injection reduces relative L_2 errors for all tested architectures and improves all three metrics for DeepONet, Point Transformer, and ViT. However, it also produces metric-dependent trade-offs for AE and U-Net, where relative L2 improves while MAE and MSE deteriorate. These findings indicate that future models for non-equilibrium flow transport should be both kinetic-aware and frequency-aware, while also being evaluated with multiple complementary metrics.

8. Conclusion

In this work, we introduced TransportBench, a benchmark framework for non-equilibrium flow transport that integrates high-fidelity datasets with standardized model evaluation protocols. The benchmark is designed to address two related needs in scientific machine learning for fluid dynamics: the lack of representative datasets for non-equilibrium gas transport and the need for controlled comparison of neural architectures across different physical regimes.

TransportBench covers a broad range of transport phenomena, including rarefied and continuum flows, low-speed and hypersonic conditions, inert and reacting gases, translational and internal-energy non-equilibrium, and both smooth and shock-dominated field structures. The four benchmark tasks are designed to assess geometry-dependent prediction, parameter-dependent prediction, distribution-function and high-order kinetic prediction, and shock-dominated hypersonic reconstruction. By including non-equilibrium quantities such as distribution functions, stress, and heat flux, the current work extends existing fluid benchmarks beyond standard macroscopic variables and provides a more stringent test for learning non-equilibrium transport.

Through systematic evaluation of representative convolutional, latent-representation, operator-learning, attention-based, and geometric models, we showed that model performance is strongly regime-dependent. No single architecture performs uniformly best across all tasks. Instead, different inductive biases become advantageous under different physical and representational conditions. The results also show that explicit high-frequency feature injection has a non-uniform effect

in shock-dominated flows. In the double-cone benchmark, Fourier features reduce relative L2 errors across the tested architectures, but their influence on MAE and MSE remains architecture-dependent. Thus, high-frequency augmentation should be evaluated jointly with global normalized errors, absolute pointwise errors, and qualitative physical behavior, since improved aggregate accuracy may still coexist with local oscillatory artifacts or nonphysical noise.

Overall, TransportBench provides not only a reusable dataset for non-equilibrium gas dynamics, but also a diagnostic benchmark for understanding how scientific machine learning models behave across multiple levels of gas-transport physics. We expect it to support the development of more robust, physics-aware, and regime-adaptive neural solvers for rarefied flows, high-enthalpy hypersonic flows, and broader non-equilibrium transport problems.

CRedit authorship contribution statement

Xu Wang: Writing – original draft, Software, Methodology, Investigation, Data curation. **Minghao Li:** Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation. **Qizhen Hong:** Writing – review & editing, Software, Methodology, Investigation, Data curation. **Yang Liu:** Software, Methodology, Data curation. **Chen-an Zhang:** Supervision, Resources, Funding acquisition. **Shuai Zhang:** Resources, Funding acquisition. **Wenhao Li:** Writing – review & editing, Resources, Funding acquisition. **Yonghao Zhang:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Tianbai Xiao:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for this journal and was not involved in the editorial review or the decision to publish this article.

Acknowledgments

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDB0620403), the Chinese Academy of Sciences Project for Young Scientists in Basic Research (Grant No. YSBR107), the National Nature Science Foundation of China (Grant Nos. 12302381 and 12572340), and the Beijing Natural Science Foundation (Grant No. L252039).

Appendix A. Numerical implementation and hyperparameters

To ensure complete reproducibility of the TransportBench evaluation, we detail the exact hyperparameter configurations and architectural specifications for all baseline models evaluated in this study.

To decouple intrinsic architectural advantages from brute-force parameter scaling, all models in Tasks I-III are bottlenecked to approximately 1×10^6 (1M) parameters. For Task IV, the budget is scaled to approximately 33×10^6 (33M) parameters to combat severe physical discontinuities. The detailed architectural hyperparameters (e.g., hidden dimensions, number of layers/modes) configured to meet these budgets, alongside their specific optimization settings, are summarized in Table A.1.

All models are trained for 2500 epochs using the Adam optimizer. For Task IV, the OneCycleLR scheduler dictates the learning rate warm-up and annealing phases.

References

- [1] D. Szirczak, H. Smith, A review of design issues specific to hypersonic flight vehicles, *Prog. Aerosp. Sci.* 84 (2016) 1–28.
- [2] M. Gad-el Hak, *The MEMS Handbook*, CRC Press, 2001.
- [3] C. Cercignani, *The Boltzmann Equation and Its Applications*, Springer, 1988.
- [4] G.A. Bird, *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*, Oxford University Press, 1994.
- [5] S.L. Brunton, B.R. Noack, P. Koumoutsakos, Machine learning for fluid mechanics, *Annu. Rev. Fluid Mech.* 52 (2020) 477–508.
- [6] G.E. Karniadakis, I.G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, L. Yang, Physics-informed machine learning, *Nat. Rev. Phys.* 3 (6) (2021) 422–440.
- [7] M. Takamoto, T. Praditia, R. Leiteritz, D. MacKinlay, F. Alesiani, D. Pflüger, M. Niepert, PDEBench: an extensive benchmark for scientific machine learning, in: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 1596–1611.
- [8] R. Tali, A. Rabeh, C.-H. Yang, M. Shadkhah, S. Karki, A. Upadhyaya, et al., FlowBench: a Large Scale Benchmark for Flow Simulation over Complex Geometries, *arXiv preprint arXiv:2409.18285* (2024).
- [9] H. Wang, R. Li, F. Xu, F. Sun, et al., FD-Bench: a modular and fair benchmark for data-driven fluid simulation, *arXiv preprint* (2025).
- [10] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, A. Courville, On the spectral bias of neural networks, in: *International Conference on Machine Learning (ICML)*, PMLR, 2019, pp. 5301–5310.
- [11] T. Xiao, M. Frank, Using neural networks to accelerate the solution of the Boltzmann equation, *J. Comput. Phys.* 443 (2021) 110521.
- [12] T. Xiao, M. Frank, RelaxNet: a structure-preserving neural network to approximate the Boltzmann collision operator, *J. Comput. Phys.* 490 (2023) 112317.
- [13] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: *International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, 2015, pp. 234–241.
- [14] J. Zhai, S. Zhang, J. Chen, Q. He, Autoencoder and its various variants, in: *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2018, pp. 415–419.
- [15] L. Lu, P. Jin, G. Pang, Z. Zhang, G.E. Karniadakis, Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators, *Nat. Mach. Intell.* 3 (3) (2021) 218–229.
- [16] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, A. Anandkumar, Fourier neural operator for parametric partial differential equations, *arXiv preprint arXiv:2010.08895* (2020).
- [17] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al., A survey on vision transformer, *IEEE Trans. Pattern Anal. Mach. Intell.* 45 (1) (2022) 87–110.
- [18] H. Zhao, L. Jiang, J. Jia, P.H.S. Torr, V. Koltun, Point transformer, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16259–16268.
- [19] Z. Hao, et al., PINNacle: A comprehensive benchmark of physics-informed neural networks for solving PDEs, *arXiv preprint arXiv:2306.08827* (2023).
- [20] Y. Luo, Y. Chen, Z. Zhang, CFD-Bench: A comprehensive benchmark for machine learning methods in fluid dynamics, *arXiv preprint arXiv:2310.02672* (2023).

Table A1

Hyperparameter configurations of the six baseline architectures used in TransportBench. The table lists the parameter budget, learning rate, batch size, and key architectural settings for Tasks I-III. For Task IV, the parameter budget is increased as described in Section 5.2, while the same architecture families and training framework are retained.

Model	Params (standard)	Learning rate	Batch size	Key architectural configurations
DeepONet	~1.05M	1×10^{-3}	32	Branch layers: 4, Trunk layers: 4, Hidden dim: 128
FNO	~0.98M	1×10^{-3}	32	Fourier modes: 12, Width: 32, FFT layers: 4
U-Net	~1.10M	5×10^{-4}	16	Encoder/Decoder blocks: 4, Base channels: 32
ViT	~1.02M	1×10^{-4}	16	Patch size: 8×8 , Transformer layers: 6, Heads: 8
AutoEncoder	~1.00M	1×10^{-3}	32	Conv layers: 5, Latent dim: 128, Activation: GELU
Point trans.	~1.15M	5×10^{-4}	8	k -nearest neighbors: 16, Attention layers: 4

- [21] A. Jacot, F. Gabriel, C. Hongler, Neural tangent kernel: convergence and generalization in neural networks, *Advances in Neural Information Processing Systems* (NeurIPS), vol. 31, 2018.
- [22] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, R. Ng, Fourier features let networks learn high frequency functions in low dimensional domains, in: *Advances in Neural Information Processing Systems* (NeurIPS), volume 33, 2020, pp. 7537–7547.
- [23] I. Nompelis, G. Candler, M. Holden, Computational investigation of hypersonic viscous/inviscid interactions in high enthalpy flows, in: *36th AIAA Thermophysics Conference*, 2003, p. 3642.
- [24] X. Wang, J. Guo, Q. Hong, S. Li, High-fidelity state-to-state modeling of hypersonic flow over a double cone, *Phys. Fluids* 35 (11) (2023) 116101.
- [25] I.D. Boyd, T.E. Schwartzentruber, *Nonequilibrium Gas Dynamics and Molecular Simulation*, Cambridge Aerospace Series, Cambridge University Press, 2017.
- [26] S.J. Plimpton, S.G. Moore, A. Borner, A.K. Stagg, T.P. Koehler, M.A. Gallis, Direct simulation Monte Carlo on petaflop supercomputers and beyond, *Phys. Fluids* 31 (8) (2019) 086101.
- [27] T. Xiao, C. Liu, K. Xu, Q. Cai, A velocity-space adaptive unified gas kinetic scheme for continuum and rarefied flows, *J. Comput. Phys.* 415 (2020) 109535.
- [28] T. Xiao, *Kinetic.jl: a portable finite volume toolbox for scientific and neural computing*, *J. Open Source Softw.* 6 (62) (2021) 3060.
- [29] X. Wang, Q. Hong, Y. Hu, Q. Sun, On the accuracy of two-temperature models for hypersonic nonequilibrium flow, *Acta Mech. Sin.* 39 (2) (2023) 122193.
- [30] L.N. Smith, A disciplined approach to neural network hyper-parameters: Part 1—learning rate, batch size, momentum, and weight decay, *arXiv preprint arXiv:1803.09820*(2018).