POLITECNICO
MILANO 1863

RE.PUBLIC@POLIMI
Research Publications at Politecnico di Milano

Post-Print

This is the accepted version of:

E. Monti, N. Singh, J. Sirignano, J.F. MacArt, M. Panesi, G. Gori
Physics-Constrained Deep Learning-Based Model for Non-Equilibrium Flows
in: AIAA Scitech 2024 Forum, AIAA, 2024, ISBN: 9781624107115, p. 1-21, AIAA 2024-0654
[AIAA Scitech 2024 Forum, Orlando, FL, USA, 8-12 Jan. 2024]
doi:10.2514/6.2024-0654

The final publication is available at https://doi.org/10.2514/6.2024-0654

Access to the published version may require subscription.

When citing this work, cite the original published paper.

Permanent link to this version
http://hdl.handle.net/11311/1258640

Physics-constrained deep learning-based model for non-equilibrium flows

Edoardo Monti *
Department of Aerospace Science and Technology, Politecnico di Milano, Via La Masa 34, 20156, Milano, Italy

Narendra Singh †
Center for Hypersonic and Entry Systems Studies (CHESS), University of Illinois, Urbana, Illinois, USA

Justin Sirignano ‡
University of Oxford, Oxford, UK

Jonathan F. MacArt §
University of Notre Dame, Notre Dame, Indiana, USA

Marco Panesi ¶
Center for Hypersonic and Entry Systems Studies (CHESS), University of Illinois, Urbana, Illinois, USA

Giulio Gori ‖
Department of Aerospace Science and Technology, Politecnico di Milano, Via La Masa 34, 20156, Milano, Italy

In hypersonic applications, the construction of accurate and computationally affordable models for simulating non-equilibrium fluid flows is a challenging task. In particular, design-relevant cases are complex and data availability is poor, de facto hampering the development of constitutive relations of general validity. To circumvent this issue, we propose a methodology for building physics-constrained neural networks providing a correction to the constitutive relation included in the Navier-Stokes model, with a specific focus on rarefied flows. The approach is based on the premise that physical laws should be inherently encoded in robust and accurate closures. By requiring the fulfillment of these laws i.e., by introducing specific constraints to the training process of the neural network, we obtain correction terms coherent with the physics, enhancing the modeling of both the viscous stress tensor and the heat flux vector. The goal is to demonstrate the feasibility of the proposed approach and its potential to benchmark the test case of the 1D shock.

I. Nomenclature

ρ	=	mass density
$\boldsymbol{u}$	=	bulk velocity
E	=	total energy per unit mass
p	=	pressure
a	=	speed of sound

*MSc, Department of Aerospace Science and Technology, edoardo2.monti@mail.polimi.it

†Postdoc, Department of Aerospace Engineering, University of Illinois at Urbana–Champaign; Center for Hypersonics and Entry Systems Studies (CHESS) narsingh@illinois.edu

‡Associate Professor, Mathematical Institute, University of Oxford, Andrew Wiles Building, Oxford OX2 6GG, UK,

§Assistant Professor, Department of Aerospace and Mechanical Engineering, University of Notre Dame, 369 Fitzpatrick Hall of Engineering, Notre Dame, IN 46556, USA; Member AIAA

¶Associate Professor, Department of Aerospace Engineering, University of Illinois at Urbana–Champaign; Director, Center for Hypersonics and Entry Systems Studies (CHESS); 306 Talbot Laboratory, 104 S. Wright St., Urbana, IL 61801, USA; Senior AIAA member,

‖Assistant Professor, Department of Aerospace Science and Technology, Politecnico di Milano, Via La Masa 34, 20156, Milano, Italy, giulio.gori@polimi.it

M = Mach number
ω = VHS parameter
$\boldsymbol{e}$ = symmetric part of the velocity gradient
$\boldsymbol{\Omega}$ = antisymmetric part of the velocity gradient
ϕ = Boltzmann velocity distribution function
Kn = Knudsen number
$\mathbb{I}$ = identity tensor
$\boldsymbol{\tau}$ = viscous stress tensor
$\boldsymbol{q}$ = heat flux vector
$\boldsymbol{\tau}_{\text{NS}}$ = viscous stress tensor for Navier-Stokes model
$\boldsymbol{q}_{\text{NS}}$ = heat flux vector for Navier-Stokes model
F = Augmented Navier-Stokes augmented equations
f, g = Deep Learning corrections to Navier-Stokes model
$\boldsymbol{\theta}$ = learning weights
$\boldsymbol{U}$ = conserved variables vector
$\boldsymbol{U}_A$ = adjoint variables vector
$\boldsymbol{Q}$ = thermodynamic state vector
J = objective function
Λ = molecular mean free path
ϵ_{rel} = relative error w.r.t. target data
ϵ_{T} = scheduler threshold value
α_{LR} = learning rate
α = regularization coefficient

II. Introduction

The construction of accurate and computationally affordable computer models for simulating non-equilibrium fluid flows is a challenging task, particularly when dealing with hypersonic applications. The popular Navier-Stokes (NS) model yields unreliable predictions whenever the continuum assumption fails (approx. for a Knudsen number $\text{Kn} > 0.1$). On the other hand, the large computational cost entailed in solving the more suited Boltzmann equation makes its numerical solution impossible for non-trivial cases. In such cases, even the most efficient Direct-Simulation Monte Carlo (DSMC) solution methods may result prohibitively expensive, in particular if approx. $\text{Kn} < 10$ and if thermochemical nonequilibrium is considered.

This paper focuses on improving the computational capabilities for predicting nonequilibrium hypersonic flows in the transitional regime $0.1 < \text{Kn} < 10$. Namely, we build physics-constrained Deep Learning (DL) augmentation terms that take advantage of the known and resolvable physics e.g., first principles, thermodynamics laws, and conservation laws, and exploit them to improve the accuracy of the NS solution. The aim is to achieve a prediction accuracy comparable to that of the DSMC solution while maintaining a computational cost comparable to solving the standard NS equations.

Previous studies employed Deep Neural Networks (DNNs) for enhancing the modeling of incompressible fluid flows, with particular reference to turbulent phenomena, but their potential has not been fully deployed for hypersonic applications, for which literature concerning DL-DNNs augmentation models is still scarce. Recent research by Sirignano *et al.* [1] proposed an innovative *online* training strategy that leverages the adjoint method to optimize the parameters of a DL model by solving a constrained PDE. This approach allows consistent mathematical-physical training and was successfully applied to hypersonic problems by Nair *et al.* [2] for a 1D shock flow in the transitional regime. However, this model does not embed the first principle of physics, making it case-dependent and preventing its extension to general 2D/3D applications.

The present work aims to overcome these issues and extend the state-of-the-art by proposing a flexible framework for building physics-constrained neural network closures of general application. To overcome these drawbacks, the pursued approach is based on the premise that physical laws and constraints are inherently embedded in real applications and they can be exploited to guide the development of more robust and accurate closures. By incorporating these constraints into the training process, we aim to develop a model that can capture both the universal physics, which is general, and the physics encoded in the data, which is case-dependent. It is important to note that the main advantage of the approach proposed in this work, compared to other alternative approaches, is its ability to generalize: the structure of the developed model is case-independent and permits the use of the same model on very different problems. In principle, the model

trained for a 1D problem can easily be used for 2D or 3D problems involving the same physics, since no case-specific parameter or geometry setup is involved, and its physics-informed core aims to reproduce the universal dependencies between input flow variables and output corrections. The prescriptions of physics are translated into mathematical constraints that are hard-imposed on the NN correction. As an example, the NN model input/output mathematical functional relation satisfies Noll's principles for macroscopic governing equations, guaranteeing its validity for different geometrical setups (e.g., altering the reference system would not compromise the model's accuracy). This result is particularly important for applying this methodology in design-relevant cases since it can enable the development of corrections that can be trained on simple scenarios while maintaining the ability to generalize to any case, easing the training process by partially overcoming the challenging task of gathering consistent amount of reliable data in the context of complex applications.

To indicate the feasibility of the proposed approach and its potential, a benchmark test case has been studied: the investigation of the inner structure of a 1D shock low-pressure Argon flow. The paper is organized as follows: Sec. III covers the theoretical background of the equations governing the fluid flow. A brief review of deep learning algorithms and their training strategy is also given. Section IV outlines the novel framework for building physics-constrained deep learning closures. In Section V we apply the developed methodology on the steady 1D shock case. Finally, Sec. VI presents a summary and discussion of future work.

III. Theoretical framework

In this Section, the theory underlying this work and the terminology used are briefly reviewed.

A. Governing equations

Consider the velocity distribution function $\phi(t, \boldsymbol{x}, \boldsymbol{v})$, which represents the solution to the Boltzmann equation, with $\boldsymbol{v}$ denoting the molecular velocity. The Boltzmann equation for a single component non-reactive monoatomic gas can be formulated as

$$\frac{\partial \phi}{\partial t} + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} \phi = \left(\frac{\partial \phi}{\partial t} \right)_{\text{collision}} \tag{1}$$

The mass density ρ, bulk momentum $\rho \boldsymbol{u}$ and total energy ρE can be determined as statistical moments of the velocity distribution function ϕ with respect to collision invariants $m, m\boldsymbol{v}, \frac{1}{2} m |\boldsymbol{v}|^2$, respectively. The balance equations for these quantities are derived directly from the moments of the Boltzmann equation [3].

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}) &= 0, \\ \frac{\partial \rho \boldsymbol{u}}{\partial t} + \nabla \cdot (\rho \boldsymbol{u} \otimes \boldsymbol{u} + p \mathbb{I} - \boldsymbol{\tau}) &= 0, \\ \frac{\partial \rho E}{\partial t} + \nabla \cdot ((\rho E + p) \boldsymbol{u} - \boldsymbol{\tau} \boldsymbol{u} + \boldsymbol{q}) &= 0, \end{aligned} \tag{2}$$

where $\boldsymbol{\tau}, \boldsymbol{q}$ are undefined terms which represent respectively the viscous stress tensor and the heat flux vector. They can be expressed as functions of ϕ as

$$\begin{aligned} \tau_{ij} &= - \int_{\boldsymbol{c} \in \mathbb{R}^3} (c_i c_j - |\boldsymbol{c}|^2) \phi d^3 \boldsymbol{c}, \\ q_i &= \int_{\boldsymbol{c} \in \mathbb{R}^3} c_i |\boldsymbol{c}|^2 \phi d^3 \boldsymbol{c}, \end{aligned} \tag{3}$$

where $\boldsymbol{c} = \boldsymbol{v} - \boldsymbol{u}$ is the peculiar velocity. The Chapman-Enskog expansion of ϕ truncated at the first order retrieves expressions for $\boldsymbol{\tau} = \boldsymbol{\tau}_{\text{NS}}$ and $\boldsymbol{q} = \boldsymbol{q}_{\text{NS}}$ in the limit of small perturbations from the local equilibrium condition, thereby providing balance equations in the form of the Navier Stokes equations. However, the Chapman-Enskog expansion entails a minimal deviation of the velocity distribution function ϕ from its equilibrium Maxwellian form. In transitional flows, the continuum assumption loses its validity and the local equilibrium hypothesis may not be always fulfilled since restoring equilibrium after a perturbation requires a non-negligible finite relaxation time. Thus, ϕ may significantly deviate from the Maxwellian equilibrium form and be instead bimodal [4, 5]. In such cases, the hydrodynamic problem is undetermined with respect to the conserved variables, as obtaining an expression for $\boldsymbol{\tau}$ and $\boldsymbol{q}$ requires the knowledge of ϕ, which in turn necessitates the resolution of the Boltzmann equation.

In this work, DL techniques are used to provide closures for $\boldsymbol{\tau}$ and $\boldsymbol{q}$ and solve the hydrodynamic problem, bypassing the solution of the Boltzmann equation. To achieve this purpose, we express $\boldsymbol{\tau}$ and $\boldsymbol{q}$, respectively, as closed and computable functions of conserved variables $\boldsymbol{U} = \rho, \rho\boldsymbol{u}, \rho E$ and the DL-based correction terms $\boldsymbol{f}$ and $\boldsymbol{g}$

$$\begin{aligned} \boldsymbol{\tau} &= \boldsymbol{\tau}(\boldsymbol{U}, \boldsymbol{f}), \\ \boldsymbol{q} &= \boldsymbol{q}(\boldsymbol{U}, \boldsymbol{g}). \end{aligned} \tag{4}$$

As shown in [2], there are two strategies for performing an augmentation of the model using a Deep Learning closure:

(a) Augmenting the Chapman-Enskog model of shear-stresses and heat-flux with corrective neural networks:

$$\begin{aligned} \boldsymbol{\tau} &= \boldsymbol{\tau}_{\mathrm{NS}}(\boldsymbol{U}) + \boldsymbol{\tau}_{\mathrm{C}}(\boldsymbol{U}, \boldsymbol{f}(\boldsymbol{U}; \boldsymbol{\theta})), \\ \boldsymbol{q} &= \boldsymbol{q}_{\mathrm{NS}}(\boldsymbol{U}) + \boldsymbol{q}_{\mathrm{C}}(\boldsymbol{U}\boldsymbol{g}(\boldsymbol{U}; \boldsymbol{\theta})), \end{aligned} \tag{5}$$

where $(...)_{\mathrm{C}}$ indicate corrective terms to the NS model, in the form of neural network closures.

(b) Directly replacing $\boldsymbol{\tau}$ and $\boldsymbol{q}$ with neural networks.

$$\begin{aligned} \boldsymbol{\tau} &= \boldsymbol{f}(\boldsymbol{U}; \boldsymbol{\theta}), \\ \boldsymbol{q} &= \boldsymbol{g}(\boldsymbol{U}; \boldsymbol{\theta}). \end{aligned} \tag{6}$$

Approach (b) has the potential to accurately represent the functional form for $\boldsymbol{\tau}$ and $\boldsymbol{q}$, leading to the corrected flow solution. However, experience showed that the online optimization process for the parameters in approach (b) is significantly more unstable due to the absence of physical constraints during the exploration phase. This instability can result in nonphysical and numerically unstable functional forms for stresses and heat flux, ultimately causing the failure of the entire online optimization method. For example, models that do not satisfy the Second Law of Thermodynamics can exhibit heat flux moving from lower to higher temperatures, potentially resulting in critical conditions such as $T < 0$. Additionally, the training process is slower due to the lack of proper scaling, as the network must spend optimization steps learning the factors of proportionality. This behavior could be mitigated by imposing constraints that enforce first-principles physics. However, this would reduce the function space that the deep learning operator could generate, thus losing the advantages of approach (b). To avoid the detrimental drawbacks of approach (b), this work adopts the (a) strategy. Physics-based constraints in approach (a) are more straightforward to implement, and its intrinsic scaling renders the training process faster and more efficient. Consequently, based on these considerations, the NN closure should function as a correction to the Chapman-Enskog shear stress tensor and the heat flux vector

$$\begin{aligned} \tau_{ij} &= \tau_{\mathrm{NS}ij}(1 + f_{ij}), \\ q_i &= q_{\mathrm{NS}_i}(1 + g_i). \end{aligned} \tag{7}$$

In this context, $\boldsymbol{f}$ and $\boldsymbol{g}$ represent the Navier-Stokes stress tensor and heat flux corrections, respectively. It is important to note that this formulation can be interpreted as a scaling of the neural network closure

$$\begin{aligned} \tau_{ij} &= \tau_{\mathrm{NS}ij} + \tau_{\mathrm{NS}ij} f_{ij}, \\ q_i &= q_{\mathrm{NS}_i} + q_{\mathrm{NS}_i} g_i, \end{aligned} \tag{8}$$

with

$$f_{ij} = \frac{\tau_{\mathrm{C}ij}}{\tau_{\mathrm{NS}ij}}, \quad \text{and} \quad g_i \;\; = \frac{q_{\mathrm{C}i}}{q_{\mathrm{NS}_i}}. \tag{9}$$

Due to this scaling, better training performance can be achieved under the reasonable assumption that the order of magnitude of $\boldsymbol{\tau}_{\mathrm{C}}$ and $\boldsymbol{q}_{\mathrm{C}}$ is comparable to that of $\boldsymbol{\tau}_{\mathrm{NS}}$ and $\boldsymbol{q}_{\mathrm{NS}}$. More importantly, this scaling makes enforcing physical constraints on the model significantly easier, such as the Second Law of Thermodynamics, as demonstrated in Sec. IV.

B. Deep Learning closure

Deep learning techniques rely on NN models which are fundamental nonlinear function approximators. The universal approximation theorem [6] states that any function can be approximated by a sufficiently large and deep network. This theoretical foundation allows for the development of NN models such as $\boldsymbol{f}(\boldsymbol{U};\boldsymbol{\theta})$ and $\boldsymbol{g}(\boldsymbol{U};\boldsymbol{\theta})$, that approximate the nonlinear function dependency of the undetermined terms from the physical variables $\boldsymbol{U}$. The NN takes the same inputs that characterize the physical model being emulated, but it has additional parameters i.e., the learning weights $\boldsymbol{\theta}$, that need to be calibrated using data. By training the NN on trusted data, we can obtain an approximation of the physical model's mapping between outputs and inputs (e.g., between viscous stresses and the flow solution).

There exist different architectures for building NN. These can be classified by type, structure, and learning algorithm, and perform best depending on the specific application. Therefore, the choice of the architecture is critical and the interested readers are encouraged to refer to the literature for a more in-depth understanding of the subject [7, 8]. In this work, we select the DeepONet feedforward deep architecture [9] which particularly suits the learning of continuous nonlinear operators.

C. Training strategies

There are two primary strategies for gradient-based training: *online* optimization and *offline* optimization. Figure 1 illustrates the main key aspects of the *online* training loop.

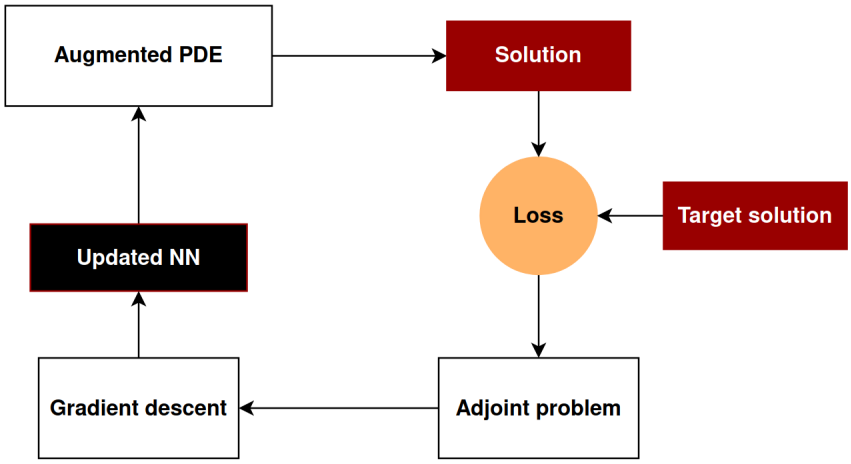


Fig. 1 Online training loop.

In CFD applications, the current state-of-the-art consists in the *offline* approach in which the models are trained using objective functions that depend explicitly on data and do not require solving a PDE. For instance, modeling the viscosity would require minimizing the objective function $J = \int_V (\mu(\boldsymbol{U}, \boldsymbol{f}(\boldsymbol{U};\boldsymbol{\theta})) - \mu_T)^2 dV$, where μ_T are reference data. However, this approach faces limitations when exact data for the higher-order terms being modeled is scarce or unavailable i.e., when the data source is experimental. Furthermore, since the closure is trained without considering its effects on the system dynamics, the *offline* optimization may be inconsistent with the governing equations [1]. Additionally, the optimization is not constrained to fulfilling a PDE. Therefore only the physics encoded in the reference data is possibly captured.

To address these limitations, this work employs the *online* optimization algorithm. By defining arbitrary objective functions, it is possible to target any quantity of interest and enable a training process focused on improving predictions in relevant flow regions such as boundaries and shocks. The training loop leverages an adjoint-based optimization algorithm including a constraint designed to enforce the fulfillment of the governing equations. Namely, the constraint requires that a PDE has be solved at every optimization step, resulting in a "PDE-aware" training process.

Summing up, the *online* neural network training strategy employed in this study facilitates the development of a physics-constrained model. The flexibility of the training strategy concerning data source availability is a crucial feature, mainly when higher-order data is limited, as is the case with complex hypersonic flows. This adaptable training strategy allows for integrating experimental and DSMC (Direct Simulation Monte Carlo) data sources, creating more

comprehensive and precise models better equipped to handle a diverse range of operating conditions.

D. Optimization problem

As reported in [2], the whole online training procedure can be mathematically translated in a constrained minimization problem as follows

$$\begin{aligned} &\boldsymbol{\theta} = \text{argmin}\{J(\boldsymbol{U}(\boldsymbol{\theta}^*))\}, \\ &\text{subject to} \quad \boldsymbol{F}(\boldsymbol{U};\boldsymbol{\theta}^*) = 0, \end{aligned} \tag{10}$$

where J is the objective function and $\boldsymbol{F}$ are the Navier-Stokes augmented equations shown in Section III. The objective function J is constructed using the relative discrepancy between the conserved variables of the augmented PDE and target data. This discrepancy is normalized with respect to the target data to homogenize the importance of the different physical quantities in the optimization process,

$$\begin{aligned} J(\boldsymbol{U}(\boldsymbol{\theta}),\boldsymbol{\theta}) =& \frac{1}{2}\int_V \left[\left(\frac{\rho(\boldsymbol{\theta}) - \rho_T}{\rho_T} \right)^2 + \right. \\ &+ \left(\frac{\rho\boldsymbol{u}(\boldsymbol{\theta}) - (\rho\boldsymbol{u})_T}{(\rho a)_T} \right) \cdot \left(\frac{\rho\boldsymbol{u}(\boldsymbol{\theta}) - (\rho\boldsymbol{u})_T}{(\rho a)_T} \right) + \\ &+ \left. \left(\frac{\rho E(\boldsymbol{\theta}) - (\rho E)_T}{(\rho E)_T} \right)^2 \right] dV + \alpha ||\boldsymbol{\theta}||^2, \end{aligned} \tag{11}$$

with $a = \sqrt{\gamma RT}$ as the speed of sound. An additional quadratic cost term in $\boldsymbol{\theta}$ is added to enhance robustness [10]. The regularization coefficient α has been set to 0.01 through a process of trial and error.

There are several strategies to solve this nonlinear minimization problem. Here, the adjoint method has been used to efficiently compute the gradients of the objective function w.r.t. the learning weights $\boldsymbol{\theta}$ and update them exploiting gradient descent iterations. For steady-state problems, the adjoint equations are

$$\frac{\partial \boldsymbol{F}^T}{\partial \boldsymbol{U}} \boldsymbol{U_A} + \frac{\partial J}{\partial \boldsymbol{U}} = 0, \tag{12}$$

where $\boldsymbol{U}_A$ are the adjoint variables. Thus, it is possible to calculate the sensitivity of the learning weights by

$$\nabla_{\boldsymbol{\theta}} J = \frac{\partial \boldsymbol{F}^T}{\partial \boldsymbol{\theta}} \boldsymbol{U_A} + \frac{\partial J}{\partial \boldsymbol{\theta}}. \tag{13}$$

The discrete adjoint equations and the forward equations are concurrently solved to the steady state during training. After the residuals have decreased to a specified threshold level, the solution is considered to be in its steady state. At this point, the gradient descent step is executed

$$\boldsymbol{\theta}^{\text{new}} = \boldsymbol{\theta}^{\text{old}} - \alpha_{\text{LR}} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^{\text{old}}). \tag{14}$$

The training procedure can be visualized through the following algorithm Algo. 1. An adaptive learning rate schedule is employed to enhance the robustness of the optimization process. A relative error parameter ϵ_{rel} is defined as the ratio of the loss of the trained model, J, to the loss of the unmodified Navier-Stokes solution, J_0, to measure the accuracy of the models. As can be seen in Algo. 1, whenever ϵ_{rel} drops below a designated threshold value, ϵ_T, the learning rate α_{LR} is modified by multiplying it with a scaling factor of $\beta_1 < 1$. The initial value of ϵ_T was established at 0.9, and after each adjustment of the learning rate, it was decreased by a factor of $\beta_2 < 1$ to further refine the training process.

Algorithm 1 Pseudo code describing the training process

```
Restart from unmodified Navier-Stokes solution
for Iteration do
    Time advancing towards steady state
    if Residuals < Threshold then
        Gradient descent step
    end if
    Compute ϵ_rel
    if ϵ_rel < ϵ_T then
        α_LR = β_1 α_LR
        ϵ_T = β_2 ϵ_T
    end if
end for
Save optimized learning weights θ
```

The final output of the optimization is the trained model $\boldsymbol{f} = \boldsymbol{f}(\boldsymbol{U};\boldsymbol{\theta})$, $\boldsymbol{g} = \boldsymbol{g}(\boldsymbol{U};\boldsymbol{\theta})$ that closes the balance equations (2).

IV. A novel framework for building physics-informed deep learning closures

In this section, we propose a novel framework to develop a universal and case-independent physics-based model that, once trained on physics-representative simple scenarios, aspires to be used to simulate more complex flows while maintaining its accuracy. To obtain this result, the proposed closure model embeds fundamental principles of physics, which are "hard-coded", and it avoids any dependency on problem-specific parameters.

A. Physical constraints

In the following, we delve into explaining how constraints enforcing the fulfillment of physics principles are incorporated into the proposed model. By "hard-coding" fundamental principles into the model, and by eliminating parameters dependent on specific scenarios, we aim to increase the robustness of the closure and the range of its applicability. Note that the closure is already constrained to conservation laws because of the *online* optimization approach exposed in Sec. III.C.

1. Noll's principles

Since the corrective terms $\boldsymbol{f}$, $\boldsymbol{g}$ apply to constitutive relations for macroscopic governing equations, they must be constrained by Noll's principles. Noll's principles are physical requirements that constitutive relations must fulfill. They can be expressed through three important axioms [11, 12]

I Principle of determinism.
Constitutive relations must depend solely on the flow's time history up to time t.

II Principle of local effect.
The motion of the fluid outside a small enough neighborhood of a material particle P, can be ignored in determining constitutive relations at that point.

III Principle of material frame-indifference.
Constitutive equations must be invariant for changes in the reference system and the observer.

These axioms can be employed to obtain constitutive relations for $\boldsymbol{f}$ and $\boldsymbol{g}$. Namely, the determinism principle can be satisfied by assuming that the tensors $\boldsymbol{f}$, $\boldsymbol{g}$ depend only on the fluid's current state of motion. The principle of local effect is fulfilled if one assumes that $\boldsymbol{f}$, $\boldsymbol{g}$ are local functions of the solution $\boldsymbol{U}$ and its gradient $\nabla\boldsymbol{U}$ so that $\boldsymbol{f}$, $\boldsymbol{g}$ are only implicit functions of the point's coordinates. This reduces the phenomenon of model overfitting, therefore

$$\begin{aligned} \boldsymbol{f} &= \boldsymbol{f}(\boldsymbol{U}, \nabla\boldsymbol{U};\boldsymbol{\theta}), \\ \boldsymbol{g} &= \boldsymbol{g}(\boldsymbol{U}, \nabla\boldsymbol{U};\boldsymbol{\theta}). \end{aligned} \tag{15}$$

The velocity gradient tensor can be expressed using the well-known relation

$$\frac{\partial u_i}{\partial x_j} = e_{ij} + \Omega_{ij}, \tag{16}$$

where Ω_{ij} is the antisymmetric part of the velocity gradient, representing the rigid body rotation, and e_{ij} is the symmetric part, representing the strain rate tensor, defined as

$$\begin{aligned} \Omega_{ij} &= \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} - \frac{\partial u_j}{\partial x_i}\right), \\ e_{ij} &= \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right). \end{aligned} \tag{17}$$

The principle of material frame-indifference imposes that $\boldsymbol{f}, \boldsymbol{g}$ is Galilean invariant, so $\boldsymbol{f}, \boldsymbol{g}$ must be independent of the velocity $\boldsymbol{u}$ and of rigid rotation $\boldsymbol{\Omega}$. Indeed, consider a first observer in a reference system fixed to the laboratory and a second observer in a reference system that translates and rotates with the fluid. With respect to this latter, the fluid has $\boldsymbol{u} = 0, \boldsymbol{\Omega} = 0$. Assuming $\boldsymbol{f} = \boldsymbol{f}(\boldsymbol{U}, \nabla\boldsymbol{U}, \boldsymbol{\theta})$, $\boldsymbol{g} = \boldsymbol{g}(\boldsymbol{U}, \nabla\boldsymbol{U}, \boldsymbol{\theta})$ to be valid, the first observer would find a dependence of $\boldsymbol{f}$ and $\boldsymbol{g}$ on $\boldsymbol{u}$ and $\boldsymbol{\Omega}$, whereas the second would not. Let's define the thermodynamic state vector $\boldsymbol{Q} = \{\rho, p, T\}$. Since the behavior of the fluid must be the same for both observers, we deduce that

$$\begin{aligned} \boldsymbol{f} &= \boldsymbol{f}(\boldsymbol{Q}, \nabla\boldsymbol{Q}, \boldsymbol{e}; \boldsymbol{\theta}), \\ \boldsymbol{g} &= \boldsymbol{g}(\boldsymbol{Q}, \nabla\boldsymbol{Q}, \boldsymbol{e}; \boldsymbol{\theta}). \end{aligned} \tag{18}$$

2. Viscous stress tensor symmetry

Given that the viscous stress tensor $\boldsymbol{\tau}$ is inherently symmetric, as shown in Eq. (3), it follows that the correction $\boldsymbol{f}$ must be also symmetric.

3. Indifference to equilibrium condition

To obtain a closure consistent with continuum theory, the augmented hydrodynamic model must reduce to the NS equations in the regions of the flow where continuum assumption is valid or, equivalently, where the velocity distribution function is well approximated by the first-order terms of the Chapman-Enskog expansion. In other words, the corrections $\boldsymbol{f}, \boldsymbol{g}$ must be negligible whenever the local Kn is sufficiently small. This constraint is imposed in two steps. First, input gradients are scaled using the hard-sphere mean free path Λ_{HS} as the reference length (21). In this way, input gradients serve as a measure of the rarefaction of the flow since their magnitude corresponds to the gradient length local Knudsen number [13]. For a thermodynamic quantity Q_i and for the strain rate e_{ij}, we define

$$\begin{aligned} \overline{\nabla Q_i} &= \frac{\Lambda_{\mathrm{HS}}}{Q_i}\nabla Q_i, \\ \overline{e_{ij}} &= \frac{\Lambda_{\mathrm{HS}}}{a} e_{ij}, \end{aligned} \tag{19}$$

where $\overline{\boldsymbol{e}}$, $\overline{\nabla\boldsymbol{Q}}$ are the scaled input gradients and a is the speed of sound. Then, the model must guarantee that, if the scaled input gradients go to zero, the corrections contribution to stresses and heat flux must be negligible

$$\begin{aligned} &1. \quad \lim_{\overline{\nabla\boldsymbol{Q}}, \overline{\nabla\boldsymbol{e}} \to \boldsymbol{0}} \boldsymbol{f} = \boldsymbol{0}, \\ &2. \quad \lim_{\overline{\nabla\boldsymbol{Q}}, \overline{\nabla\boldsymbol{e}} \to \boldsymbol{0}} \boldsymbol{g} = \boldsymbol{0}. \end{aligned} \tag{20}$$

If the model satisfies these conditions, then the indifference to equilibrium condition is satisfied. To maintain consistency with this non-dimensional formulation, and to make $\boldsymbol{f}$ and $\boldsymbol{g}$ independent from the adopted metric, Q must also be scaled

$$\overline{Q_i} = \frac{Q_i}{Q_{i\mathrm{ref}}}, \tag{21}$$

being $Q_{i\mathrm{ref}}$ a fixed reference quantity used to ensure that changes in the metric system are correctly taken into account (e.g., $p_{\mathrm{ref}} = 1\mathrm{Pa}$).

4. II Principle of Thermodynamics

As already observed in Sec. III, the macroscopic balance equations are derived directly from the Boltzmann equation, thus they are valid irrespective of Kn. From balance equations, it is possible to compute the entropy production term as

$$\begin{aligned}\mathcal{P}_S &= \frac{1}{T}(\boldsymbol{\tau} : \boldsymbol{e}) - \frac{1}{T^2}(\boldsymbol{q} \cdot \nabla T) = \\ &= \sum_{i,j} \frac{1}{T}(\tau_{ij} e_{ij}) - \frac{1}{T^2} \sum_i (q_i \frac{\partial T}{\partial x_i}).\end{aligned} \tag{22}$$

Therefore, the II Principle of Thermodynamics can be imposed by enforcing the Clausius-Duhem inequality $\mathcal{P}_S \geq 0$ [14]

$$\underbrace{\sum_{i,j} \frac{1}{T}(\tau_{\mathrm{NS}ij}(1 + f_{ij}) e_{ij})}_{\text{(A)}} + \underbrace{\sum_i \frac{1}{T^2}(-q_{\mathrm{NS}i}(1 + g_i) \frac{\partial T}{\partial x_i})}_{\text{(B)}} \geq 0. \tag{23}$$

It can be proven that enforcing the constraints reported in Eq. (24) over the corrections $\boldsymbol{f}$ and $\boldsymbol{g}$ guarantees that the II Law of Thermodynamics is always satisfied.

The (A) term in Eq. (23) is larger than zero if

$$\begin{aligned}&1. \quad f_{ii} = f \geq -1, \quad \forall i, \\ &2. \quad f_{ij} \geq -1, \quad \forall i, j \text{ with } i \neq j.\end{aligned} \tag{24}$$

On the other hand, if also

$$1. \quad g_i \geq -1, \quad \forall i, \tag{25}$$

then the (B) term is larger than zero too, and thus the entropy condition is satisfied. For a detailed mathematical proof, we kindly refer the interested reader to the appendix [.A].

B. General constrained model

By enforcing the conditions listed in Sec. IV.A, we can obtain full physics-constrained anisotropic augmentation closures (f_{ij}, g_i) for the hydrodynamic equations (7), with the following structure

$$\begin{cases} f_{ij} = f_{ij}(\overline{\boldsymbol{Q}}, \overline{\nabla \boldsymbol{Q}}, \overline{\boldsymbol{e}}; \boldsymbol{\theta}), \\ g_i = g_i(\overline{\boldsymbol{Q}}, \overline{\nabla \boldsymbol{Q}}, \overline{\boldsymbol{e}}; \boldsymbol{\theta}), \end{cases} \tag{26}$$

subject to the constraints

$$\begin{aligned}&1. \quad f_{ii} = f \geq -1, \quad \forall i, \\ &2. \quad f_{ij} \geq -1, \quad \forall i, j \text{ with } i \neq j, \\ &3. \quad g_i \geq -1, \quad \forall i, \\ &4. \quad \lim_{\overline{\nabla \boldsymbol{Q}}, \overline{\nabla \boldsymbol{e}} \to \boldsymbol{0}} \boldsymbol{f} = \boldsymbol{0}, \\ &5. \quad \lim_{\overline{\nabla \boldsymbol{Q}}, \overline{\nabla \boldsymbol{e}} \to \boldsymbol{0}} \boldsymbol{g} = \boldsymbol{0}.\end{aligned} \tag{27}$$

Note that the proposed methodology does not require any information concerning the dimensionality of the problem (or any other case-specific variable). Therefore, the model is general and usable in a general context. This result is particularly relevant for the training step, since it makes it possible to carry out the training considering simple problems that are representative of the physics of the rarefied transitional flows (e.g., 1D normal shock, 2D flat plate boundary layer) and then use the very same trained models in a broader context.

C. Simplified isotropic model

Given the limited literature concerning augmentation closures for rarefied flows, it was decided to begin with a simplified study. Consequently, in this work we implement a simplified isotropic scalar version of the full augmentation model presented in Sec. IV.B. Using an isotropic model implies that the correction to the stress tensor (or heat flux) is proportional to the stress tensor (or heat flux) itself. Therefore, this kind of correction is scalar and it can be also seen as an augmentation of the transport coefficients μ, k, allowing a straightforward physical interpretation and an easier implementation and numerical solution, since transport coefficients have stabilizing properties for their own nature. Nonetheless, the simplified model provides valuable insights concerning the behavior of the system and it leads to significant performance improvement. The simplified model reads

$$\begin{aligned} \tau_{ij} &= \tau_{\mathrm{NS}\,ij}(1+f), \\ q_i &= q_{\mathrm{NS}\,i}(1+g). \end{aligned}$$

In light of the research conducted by Sharma (2023) [15], it has been demonstrated that the impact of bulk viscosity for Argon monoatomic gas is negligible. Consequently, throughout this work, the value for bulk viscosity has been assumed to be zero. Therefore, the augmented terms can be expressed as

$$\begin{aligned} \tau_{ij} &= \mu_{\mathrm{NS}}(1+f)(2e_{ij} - \frac{2}{3}\mathrm{Tr}(\boldsymbol{e})\delta_{ij}), \\ q_i &= -k_{\mathrm{NS}}(1+g)\frac{\partial T}{\partial x_i}, \end{aligned} \tag{28}$$

being μ and k, respectively, the augmented viscosity and thermal conductivity

$$\begin{aligned} \mu &= \mu_{\mathrm{NS}}(1+f), \\ k &= k_{\mathrm{NS}}(1+g). \end{aligned} \tag{29}$$

Since the corrections f, g are scalars, it is necessary to impose an additional isotropic constraint. In fact, a scalar quantity must not depend on the observer orientation of the axis. This means that inputs of f, g must be purged by all those parameters that have a dependence on the orientation of the axis. In particular, f, g can no longer depend on the gradient of the thermodynamic state, but only on its magnitude. Additionally, the corrections can not depend on the e_{ij} tensor as a whole, but only on its eigenvalues $\mathrm{Eigs}(\boldsymbol{e})$, which are invariant to the reference system, resulting into the following functional form

$$\begin{aligned} f &= f(\overline{\boldsymbol{Q}}, |\overline{\nabla \boldsymbol{Q}}|, \mathrm{Eigs}(\overline{\boldsymbol{e}}); \boldsymbol{\theta}), \\ g &= g(\overline{\boldsymbol{Q}}, |\overline{\nabla \boldsymbol{Q}}|, \mathrm{Eigs}(\overline{\boldsymbol{e}}); \boldsymbol{\theta}), \end{aligned} \tag{30}$$

with the entropy constraint enforced by

1. $f \geq -1$,
2. $g \geq -1$.

D. Neural Network architecture

The neural network closure is based on a DeepONet architecture [9] which has a structure specialized for approximating non-linear operators, with further adaptation choices related to the specific application tackled in this work. Figure 2 shows the schematic representation of the DL closure.

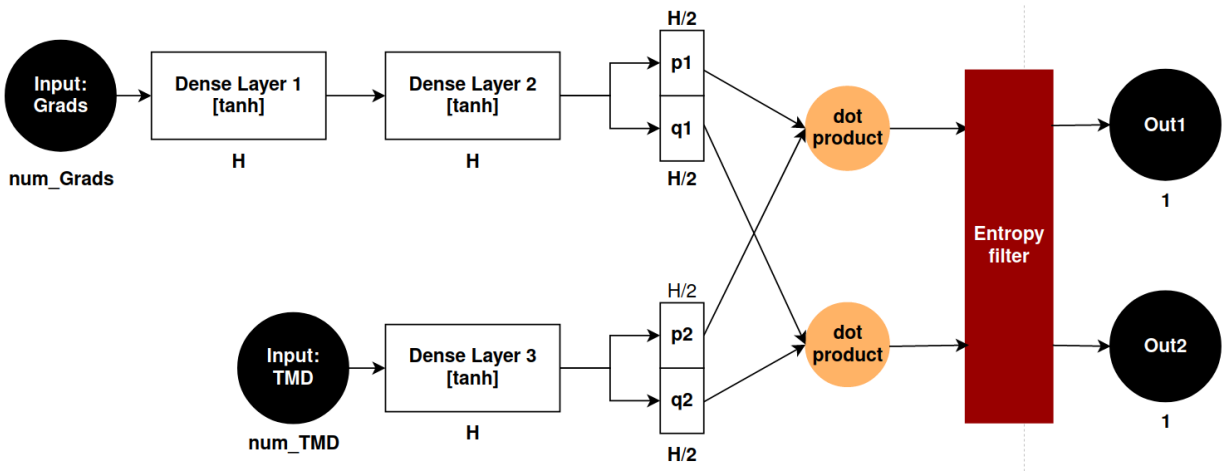


Fig. 2 Neural network architecture.

The network is characterized by multiple branches: the upper one processes the input gradients, whereas the lower one processes the thermodynamic state. The outputs of the branches are merged via a dot product, to provide the corrections with a complete non-linear basis that is a function of both the thermodynamic state and the input gradients, see Eq.(31). The dense layers 2 and 3 are split to separate the channels of information and make independent basis for $\boldsymbol{f}$ and $\boldsymbol{g}$.

$$\text{Out}_i = s_f\left(\sum_{p=1}^{H/2} \psi_{p_i}(\mathbf{Grads})\phi_{p_i}(\mathbf{TMD})\right), \tag{31}$$

with s_f the function which enforces the entropy constraint, $H = 24$ the dimension of the hidden layers, and Out_i the i-th correction for the viscous terms. Note that in this work $Out_1 = f$, $Out_2 = g$. The inputs are:

$$\begin{aligned} \mathbf{Grads} &= \{|\overline{\nabla\rho}|, |\overline{\nabla p}|, |\overline{\nabla T}|, \text{Eigs}(\overline{\boldsymbol{e}})\}, \\ \mathbf{TMD} &= \{\overline{\rho}, \overline{p}, \overline{T}\}. \end{aligned} \tag{32}$$

The neural network graph can be expressed as

$$\begin{aligned} \boldsymbol{D_1} &= \sigma(\boldsymbol{\mathcal{W}}_1\mathbf{Grads}), \\ \boldsymbol{D_2} &= \sigma(\boldsymbol{\mathcal{W}}_2\boldsymbol{D_1}), \\ \boldsymbol{D_3} &= \sigma(\boldsymbol{\mathcal{W}}_3\mathbf{TMD}), \end{aligned}$$

$$\begin{aligned} [\boldsymbol{p_1}, \boldsymbol{p_2}]^T &= \text{split}(\boldsymbol{D}_2), \\ [\boldsymbol{q_1}, \boldsymbol{q_2}]^T &= \text{split}(\boldsymbol{D}_3), \end{aligned} \tag{33}$$

$$\begin{aligned} f = \text{Out}_1 &= 0.9\ \text{s}_f(\boldsymbol{p}_1{}^T\boldsymbol{\mathcal{W}}_4\boldsymbol{p}_2), \\ g = \text{Out}_2 &= 0.9\ \text{s}_f(\boldsymbol{q}_1{}^T\boldsymbol{\mathcal{W}}_5\boldsymbol{q}_2), \end{aligned}$$

where $\sigma = \tanh()$ is the activation function of the dense layer and s_f is the exponential linear unit function (ELU) enforcing the entropy constraint (IV.C), since $\text{Im}(\text{s}_f) \in (-1, +\infty)$. The learnable parameters $\boldsymbol{\theta}$ are the weights of the dense layers

$$\boldsymbol{\theta} = \{(\boldsymbol{\mathcal{W}}_1)_{N_{\text{Grads}}\times H}, (\boldsymbol{\mathcal{W}}_2)_{H\times H}, (\boldsymbol{\mathcal{W}}_3)_{N_{\text{TMD}}\times H}, (\boldsymbol{\mathcal{W}}_4)_{H/2\times H/2}, (\boldsymbol{\mathcal{W}}_5)_{H/2\times H/2}\}.$$

Biases are not present to guarantee the equilibrium indifference condition, expressed in Eq. (20).

V. Application: 1D shock test case

The objective of this Section is to verify and assess the proposed methodology, augmenting the standard Navier Stokes equations with the ML corrective model in the 1D test case of a rarefied Argon flow across a normal shock in transitional regime. Highly resolved DSMC simulations have been used as target data for training the ML model. Our approach leverages an *online* adjoint optimization procedure to train closures. It is important to remark that the main advantage of the approach proposed in this work, compared to other alternative approaches [2], is its ability to generalize. The model designed for the 1D shock can, in theory, extend its applicability to 2D or 3D scenarios featuring normal shocks. This adaptability arises from the absence of case-specific parameters, and the model's coherence with fundamental physics. For instance, changing the dimensionality of the problem or altering the reference system would not jeopardize the model's validity (e.g., changing the direction of the x-axis would not cause all inputs to change their sign, preserving the model's effectiveness).

A. Training details

The results presented in this chapter were obtained using a modified version of the PyShock code, utilized in Ref. [16], which implements the model developed in Sec. IV.D. The initial learning rate $\alpha_{\text{LR}} = 10^{-6}$ is adjusted using the adaptive learning rate scheduler, specified as in Algo. 1 (lines 8-10). This scheduler reduces α_{LR} each time the relative error diminishes below a predetermined value. The coefficients of the scheduler are $\beta_1 = \beta_2 = 0.8$. Non-augmented Navier-Stokes solutions have been used to initialize the simulations. Navier-Stokes and adjoint solutions are solved concurrently until a steady state is reached, leveraging a pseudo-time advancing forward scheme. Optimization steps are executed only after the system attains a steady state, which is determined when residuals fall below a specific threshold.

To maximize the ability to learn the different behaviors of the non-equilibrium solution at different free stream Mach numbers M_∞ regimes, the model was trained by targeting simultaneously three free stream Mach number conditions ($M_{\text{train}} = 2, 5, 8$) and tested within a range of Mach numbers spanning $M_\infty = 2$ to $M_\infty = 10$, with $M_\infty = 3, 4, 6, 7, 9, 10$ considered out-of-sample. These tests were intended to evaluate the ability of the models to extrapolate accurately from the training dataset.

B. DSMC target data and Navier-Stokes accuracy

An in-house software has been used to generate the DSMC training and testing dataset. The free stream conditions at the inlet of the argon flow are assumed as follows

- $T_\infty = 300$K,
- $p_\infty = 6.667$Pa,

while the pre-shock Mach number M_∞ spans from $M_\infty = 2$ to $M_\infty = 10$. For collision modeling, the Variable Hard Sphere (VHS) model is employed to estimate the collision cross-section. In this context, the parameter ω was set to 0.75, where $\omega = 0.5$ corresponds to the Hard Sphere model and $\omega = 1$ corresponds to Maxwell molecules.

To be consistent with the DSMC setup, we used the Navier-Stokes equations with the VHS viscosity model (NS-VHS) as the reference non-augmented system of governing equations. VHS viscosity and thermal conductivity models can be derived in closed form in the context of the Chapman-Enskog theory and, hence, small deviance from equilibrium conditions is required to assume their validity. VHS viscosity can be expressed as:

$$\begin{aligned} \mu &= \mu_{\text{ref}} \left(\frac{T}{T_{\text{ref}}} \right)^{\omega}, \\ k &= k_{\text{ref}} \left(\frac{T}{T_{\text{ref}}} \right)^{\omega}. \end{aligned} \tag{34}$$

with μ_{ref} function of the mass-specific gas constant R, the molecular reference diameter d_{ref} and the ω VHS parameter:

$$\mu_{\text{ref}} = \frac{15\sqrt{2\pi R T_{\text{ref}}}}{2(5-2\omega)(7-2\omega)\pi d_{\text{ref}}^2} \tag{35}$$

To make the analysis consistent, the ω parameter and the free stream conditions in the non-augmented NS simulations align with those used for the DSMC target data.

The discrepancies between non-augmented NS-VHS and DSMC shock profiles are quantified by means of the loss function $J(\boldsymbol{U})$ eq. (11) evaluated with respect to the DSMC solution.

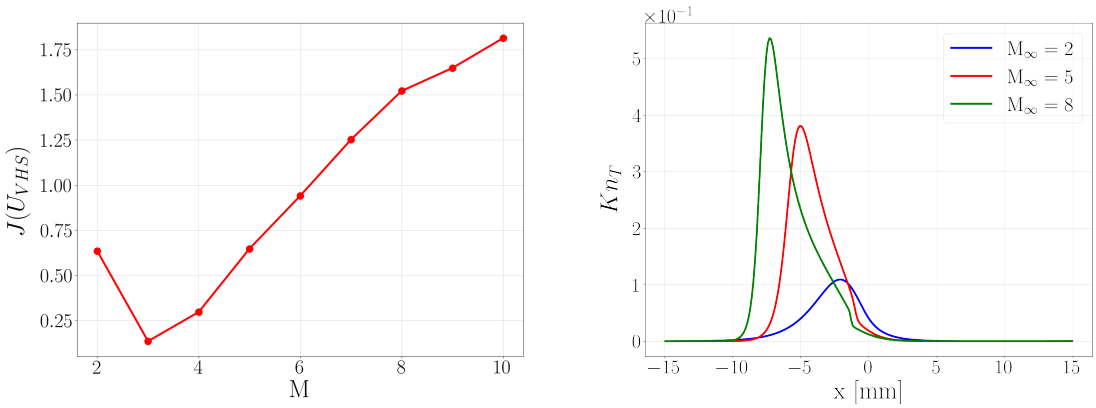


Fig. 3 (a) Error of NS-VHS with respect to DSMC. (b) Local Knudsen number across shock at various Mach numbers.

Figure 3a reveals that the VHS viscosity model is capable of well approximating the non-equilibrium solution only in the neighborhood of $M_\infty = 3$, while increasing or decreasing the Mach number leads to rapidly jeopardizing the accuracy of the simulation. Therefore, we expect a ML-corrective model capable of improving the performance of the NS-VHS equations at low and high Mach numbers while maintaining the accuracy at $M_\infty = 3$. For a more detailed analysis of NS-VHS performances, please refer to the Appendix where the complete results are presented.

C. Continuum breakdown analysis

The regions of the domain in which the continuum flow hypothesis fails have been identified through the use of a continuum breakdown parameter. Although the literature provides multiple definitions of such indicators, according to Lofthouse [13], the gradient-length (GL) local Knudsen number is the most suitable indicator for hypersonic compressible flows

$$\mathrm{Kn}_{Q_f}{}^{\mathrm{GL}} = \frac{\Lambda_{\mathrm{VHS}}}{Q_f}|\nabla Q_f|, \tag{36}$$

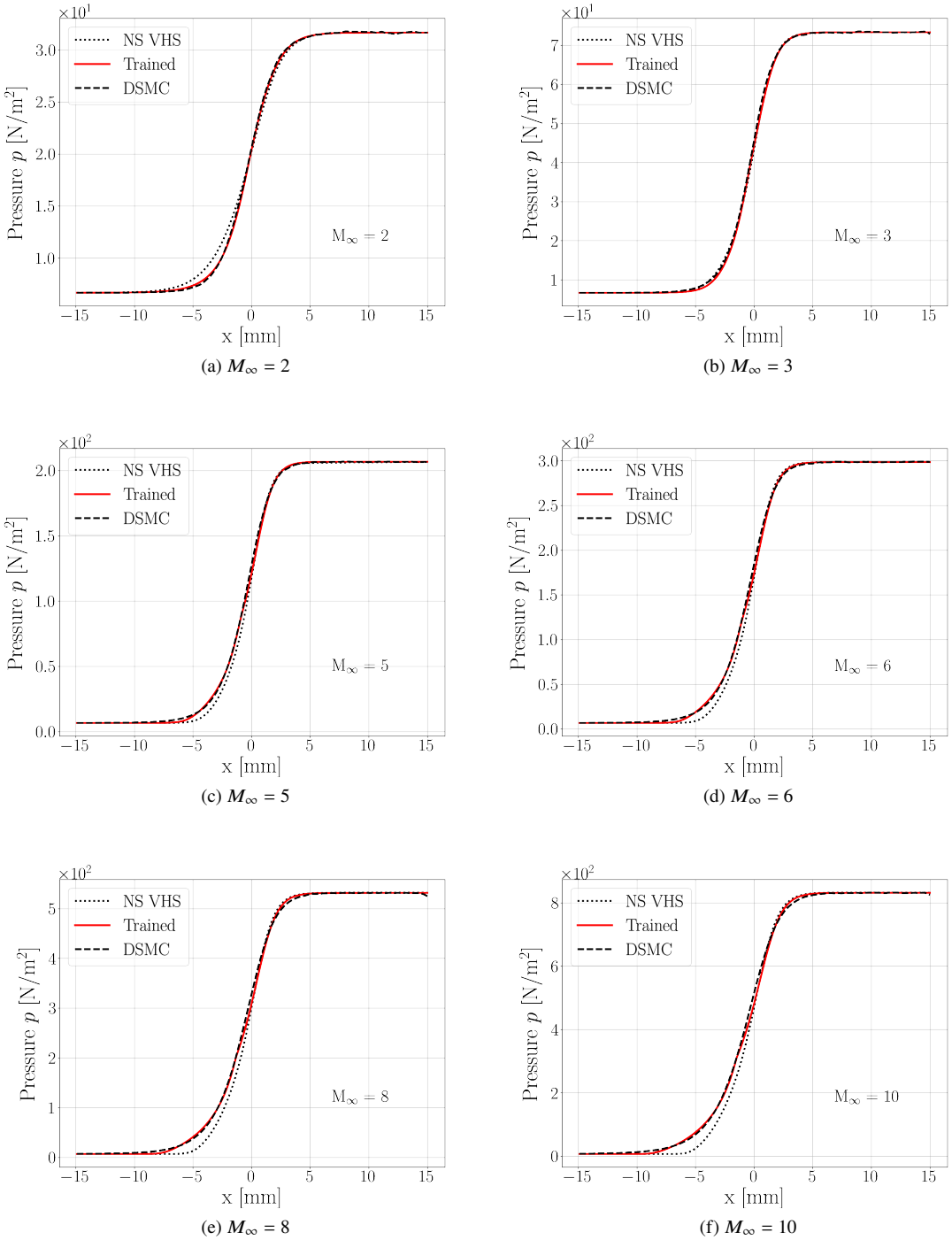
where Q_f is a flow variable set here as $Q_f = T$. Figure 3b reports the scalar field associated with $\mathrm{Kn}_T{}^{\mathrm{GL}}$. The $\mathrm{Kn}^{\mathrm{GL}}$ field shows that is not possible to describe properly large areas of the flow using the continuum assumption. Examining the shock profiles depicted in Fig. 4 and 5, we can establish a correlation between the areas of larger $\mathrm{Kn}^{\mathrm{GL}}$, and the across-the-shock region, characterized by strong gradients. Macroscopic quantities experience important variations across distances comparable to the local mean free path. Therefore, the number of collisions is small and the velocity distribution function differs consistently from the Maxwellian, resulting in a bimodal-like distribution [5]. This microscopic behavior causes the breaking of the hypotheses underlying the Chapman-Enskgog expansion across the shock.

D. Results and discussion

The solutions obtained with the trained models are shown in Figs 4 and 5 for the whole testing dataset. Simulation results (Trained) are set side by side with the unmodified Navier-Stokes VHS solution (NN VHS) and with the DSMC target data (DSMC) available for the respective setups. For brevity, only pressure 4 and temperature 5 profiles are displayed; however, similar trends were observed for the other primitive variables.

1. Pressure

The ML-corrected pressure profile, as illustrated in Fig. 4, demonstrates good agreement with the target non-equilibrium data, improving the performances of the standard Navier-Stokes predictions for all the M_∞ range.



(a) $M_\infty = 2$ (b) $M_\infty = 3$

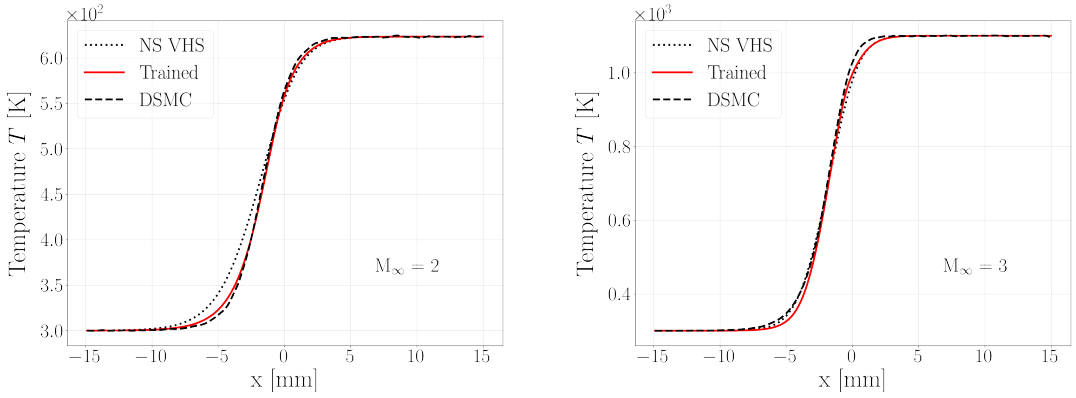
(c) $M_\infty = 5$ (d) $M_\infty = 6$

(e) $M_\infty = 8$ (f) $M_\infty = 10$

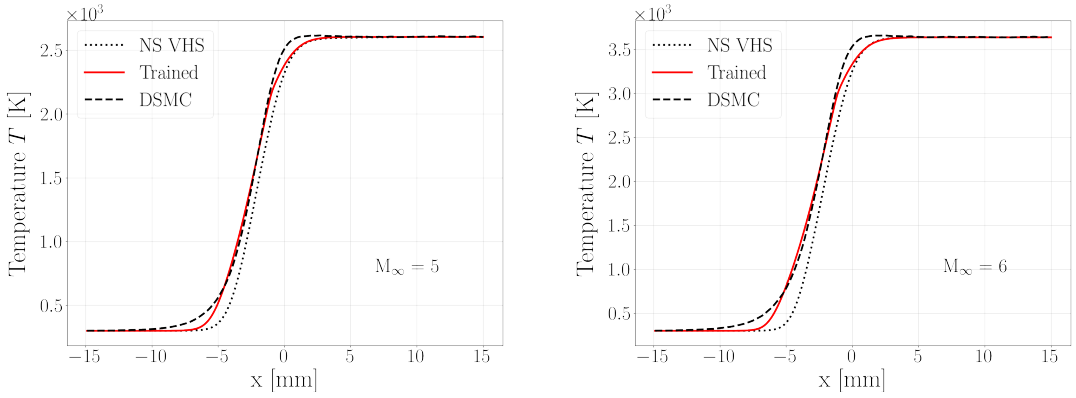
Fig. 4 Comparison of computed pressure profiles from the NN model, NS equations, and DSMC data.

2. Temperature

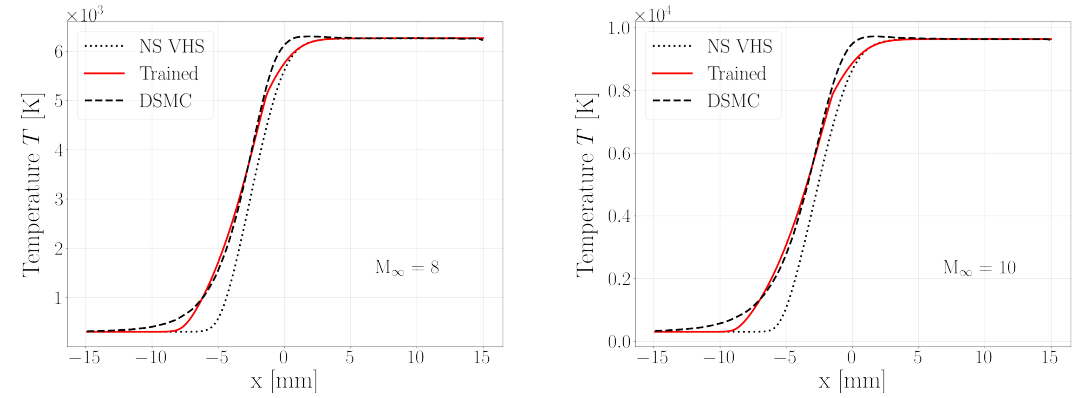
The temperature profiles, as reported in Fig 5, show good capabilities in replicating DSMC results. The NN-trained model has an effective thermal diffusivity with magnitude in between the NS equations and DSMC predictions. The discrepancies in the pre-shock region are consistent with the large gradient length based on the Knudsen number variation across the shock. However, a slight overshoot is observed in the shock tail, for all the cases. Mott-Smith



(a) $M_\infty = 2$ (b) $M_\infty = 3$



(c) $M_\infty = 5$ (d) $M_\infty = 6$



(e) $M_\infty = 8$ (f) $M_\infty = 10$

Fig. 5 Comparison of computed temperature profiles from the NN model, NS equations, and DSMC data.

analysis [5, 17] highlighted that temperature overshoot is due to the presence of shock areas in which the thermal conductivity coefficient k, considered as the proportionality factor between the heat flux and the temperature gradient, assumes negative values. The entropy condition we derived in Sec. section IV.A relies by construction on positive values of k, making it impossible for the model to reproduce this behavior.

3. Accuracy

The overall accuracy improvement is quantified by means of the relative error parameter $\epsilon_{\text{rel}} = J/J_0$, being J and J_0 the objective function evaluated respectively on the augmented NS and the unmodified NS-VHS solution.

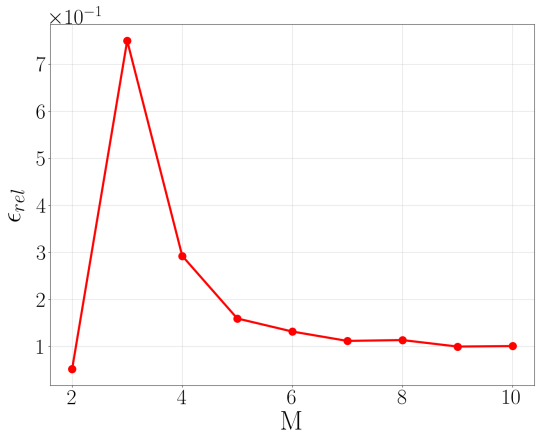


Fig. 6 Relative error.

Figure fig. 6 reports the ϵ_{rel} parameter evaluated for the trained model on the testing data set. As a reference, the non-augmented NS-VHS solution has $\epsilon_{\text{rel}} = 1$ and, therefore, a $\epsilon_{\text{rel}} < 1$ indicates an improvement to the non-augmented predictions. These findings reveal that the proposed neural network-enhanced CFD approach outperforms the standard one in terms of accuracy concerning the target data for all training Mach numbers on the whole testing data set. As expected, the model exhibits an improvement with respect to DSMC target data at low and high Mach numbers, where the NS-VHS error is greater, while maintaining the already good accuracy at $M_\infty = 3$, and displaying at the same time improved extrapolation capabilities to the out of samples testing cases. In particular, the trained model is able to reach a drop in the loss function of around 90% for a wide range of the testing data set ($M_\infty > 6$). This suggests the model effectively captures the underlying physical laws, hinting at its potential to accurately predict unseen data solutions.

4. Corrective DL model

Figures 7a and 7b depict the variation of the f and g corrections across the shock, providing a visual representation of the magnitude of their intervention. DL corrections significantly alter the μ and k profiles since non-equilibrium effects manifest in large gradient regions. Viscosity and thermal conductivity are continuum counterparts of sub-continuum processes, dependent on the microscopic velocity distribution ϕ. Thus, wherever ϕ markedly deviates from the quasi-equilibrium distribution, transport coefficients are no longer well-approximated by continuum-based models. It is worth emphasizing that, at locations where the corrections f and g are negligible, the values of μ and k revert to their non-corrected counterparts $\mu = \mu_{\text{NS}}$, $k = k_{\text{NS}}$, coherently to conditions expressed in Eq. 20. Consequently, the functions f and g provide a local quantification of the deviations between the quasi-equilibrium models, and the true physical behavior. The DL model gives an insight into the physics occurring in normal shocks, suggesting the behavior of μ and k, which represent respectively the momentum diffusion rate and internal energy diffusion rate, in non-equilibrium flows. Compared to continuum VHS predictions, the viscosity coefficient is overall decreased while thermal conductivity reduces its magnitude at lower Mach numbers and increases at high Mach numbers, where non-equilibrium effects are more influential and promote internal energy diffusion.

VI. Conclusions and future directions

In this work, we have extended the deep learning approach provided in Ref. [2] by designing a framework for building flexible and generalized NN closures aimed at improving predictions from the Navier-Stokes equations for rarefied flows in the translational regime. The main idea has been to devise a deep learning approach that allows the model to be effectively trained on problems governed by simple but representative physics, while simultaneously seeking general applicability. To accomplish this, the model must remain flexible and independent from the specific case setup,

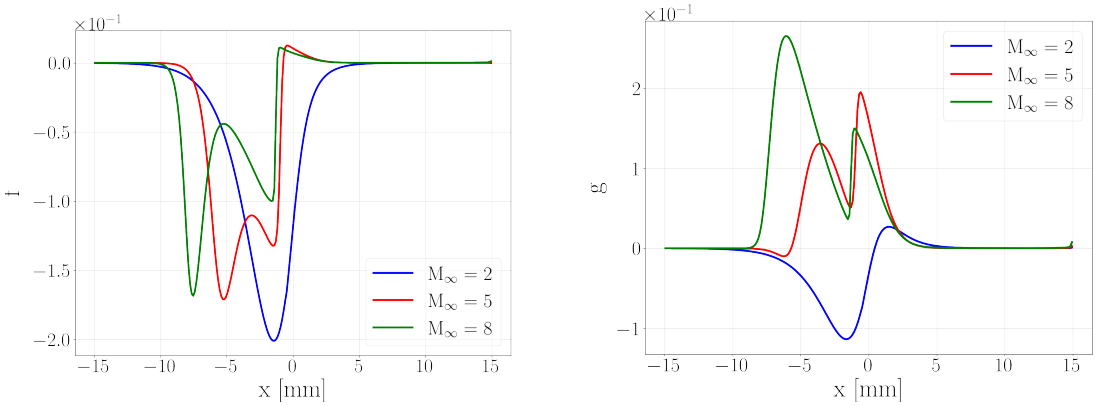


Fig. 7 (a) Viscosity correction f. (b) Thermal conductivity correction g.

such as geometry and free-stream conditions, enforcing the fundamental principles of physics over the NN corrections to theoretically build more robust and general closures. Specifically, the prescriptions of physics are translated into mathematical constraints that are imposed on the NN models. In this regard, a full mathematical proof that demonstrates a set of conditions under which the augmentation model always satisfies the second Law of thermodynamics, paramount to avoid non-physical and numerically unstable solutions, is also presented.

Feasibility and potential benefits of the proposed methodology are tested on the benchmark test case of the 1D shock problem and findings demonstrate the potential of deep learning approaches in enhancing the precision of Navier-Stokes predictions for rarefied gas flow simulations, providing accurate solutions at a comparable cost of solving the computational-affordable Navier-Stokes equations.

However, some critical aspect exists, as can be seen from the discrepancies between the augmented solution and the DSMC data in reproducing the temperature overshoot in the shock tail. The model is incapable of learning this behavior because of restrictive hypotheses enforced to satisfy the Second Law of Thermodynamics. A new set of conditions more flexible conditions is thus needed to improve the performance of the augmentation model. Additionally, further studies are required to evaluate the performance of the model on unseen setups e.g., testing the model at different Mach numbers, temperatures, or geometries, to ensure the accuracy of its predictions across a wide range of conditions. Further studies will also focus on testing the model on 2D physics-representative cases (e.g., flat plate), to assess the performance of the corrections on more complex scenarios.

Appendix

A. Proof for the entropy condition

Consider the augmentation terms $\varphi_{ij} = (1 + f_{ij})$ and $\gamma_i = (1 + g_i)$ applying, respectively, on each component of the stress tensor $\boldsymbol{\tau}_{\mathrm{NS}}$ and on the heat flux vector $\boldsymbol{q}_{\mathrm{NS}}$. In order to provide mathematical conditions on $\boldsymbol{f}$ and $\boldsymbol{g}$ for which the Second Law of Thermodynamics is always fulfilled, the Clausius-Duhem inequality for the entropy production term $\mathcal{P}_S \geq 0$ must be satisfied. The entropy production term can be expressed as

$$\mathcal{P}_S = \frac{1}{T}\left(\boldsymbol{\tau}_{\mathrm{NS}} \odot \boldsymbol{\varphi}\right) : \boldsymbol{e} - \frac{1}{T^2}\left(\boldsymbol{q}_{\mathrm{NS}} \odot \boldsymbol{\gamma}\right) \cdot \nabla T, \tag{37}$$

being $\odot$ the Hadamart product (the element-wise or entrywise product) and : the tensor dot product defined as the index contraction of two tensors $A : B = \sum_{i,j} a_{ij} b_{ij}$. Recalling the constitutive relations for the stress tensor we obtain

$$\begin{aligned}
\mathcal{P}_S &= \frac{1}{T} \left(\boldsymbol{\tau}_{\mathrm{NS}} \odot \boldsymbol{\varphi}\right) : \boldsymbol{e} - \frac{1}{T^2} \left(\boldsymbol{q}_{\mathrm{NS}} \odot \boldsymbol{\gamma}\right) \cdot \nabla T, \\
&= \frac{1}{T} \left[\left(2\mu \boldsymbol{e} + \lambda \mathrm{Tr}\left(\boldsymbol{e}\right) \mathbb{I}\right) \odot \boldsymbol{\varphi}\right] : \boldsymbol{e} + \frac{1}{T^2} \left(k \nabla T \odot \boldsymbol{\gamma}\right) \cdot \nabla T, \\
&= \frac{1}{T} \sum_{i,j} \left(2\mu e_{ij} + \lambda \mathrm{Tr}\left(\boldsymbol{e}\right) \delta_{ij}\right) \varphi_{ij} e_{ij} + \frac{1}{T^2} \sum_{i} \left(\frac{\partial T}{\partial x_i} \gamma_i\right) \frac{\partial T}{\partial x_i}, \\
&= \frac{1}{T} \sum_{i,j} \left(2\mu e_{ij} + \lambda \mathrm{Tr}\left(\boldsymbol{e}\right) \delta_{ij}\right) \varphi_{ij} e_{ij} + \frac{1}{T^2} \sum_{i,j} \frac{\partial T}{\partial x_j} \left(\delta_{ij} \gamma_j\right) \frac{\partial T}{\partial x_i},
\end{aligned} \tag{38}$$

Ultimately, the Clausius-Duhem inequality of the augmented system requires that

$$\mathcal{P}_S = \underbrace{\frac{1}{T} \sum_{i,j} \left(2\mu e_{ij} + \lambda \mathrm{Tr}\left(\boldsymbol{e}\right) \delta_{ij}\right) \varphi_{ij} e_{ij}}_{\text{(I)}} + \underbrace{\frac{1}{T^2} \sum_{i,j} \frac{\partial T}{\partial x_j} \left(\delta_{ij} \gamma_j\right) \frac{\partial T}{\partial x_i}}_{\text{(II)}} \geq 0, \tag{39}$$

where we identified the two separate contributions related to the stresses (I) and to the heat flux (II).

- (I) The contribution of the stress term to the entropy production can be written as

$$\underbrace{2\mu \sum_{i,j} e_{ij} \varphi_{ij} e_{ij}}_{(\mathrm{I}_\alpha)} + \underbrace{\lambda \sum_{i,j} \mathrm{Tr}(\boldsymbol{e}) \delta_{ij} \varphi_{ij} e_{ij}}_{(\mathrm{I}_\beta)}. \tag{40}$$

Let $\boldsymbol{v} = (e_{11}, e_{22}, e_{33})^T$, $\boldsymbol{w} = (e_{12}, e_{13}, e_{23})^T$, and the concatenation $\boldsymbol{V} = (\boldsymbol{v}, \boldsymbol{w}) = (e_{11}, e_{22}, e_{33}, e_{12}, e_{13}, e_{23})^T$. Let also define $\boldsymbol{A}_1 = \mathrm{diag}\left(\varphi_{11}, \varphi_{22}, \varphi_{33}\right)$ and $\boldsymbol{A}_2 = 2\mathrm{diag}\left(\varphi_{12}, \varphi_{13}, \varphi_{23}\right)$.
We can then write in a compact form

$$\begin{aligned}
(\mathrm{I}_\alpha) &= \boldsymbol{v}^T \boldsymbol{A}_1 \boldsymbol{v} + \boldsymbol{w}^T \boldsymbol{A}_2 \boldsymbol{w} = \boldsymbol{V}^T \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_2 \end{bmatrix} \boldsymbol{V} \\
&= \boldsymbol{V}^T \boldsymbol{A} \boldsymbol{V}
\end{aligned} \tag{41}$$

and

$$(\mathrm{I}_\beta) = \mathrm{Tr}(\boldsymbol{e}) \sum_{i,j} \delta_{ij} \varphi_{ij} e_{ij} = \sum_k e_{kk} \cdot \sum_j e_{jj} \varphi_{jj} = \sum_j \sum_k e_{jj} \varphi_{jj} e_{kk} = \sum_j \sum_k e_{jj} D_{jk} e_{kk}, \tag{42}$$

being $D_{jk} = \varphi_{jj}$. Let's decompose D_{jk} in a symmentric part $D^{(s)}_{jk} = \frac{1}{2}(D_{jk} + D_{kj}) = \frac{1}{2}(\varphi_{jj} + \varphi_{kk})$ and its antisymmetric part $D^{(a)}_{jk} = \frac{1}{2}(D_{jk} - D_{kj})$, obtaining

$$(\mathrm{I}_\beta) = \sum_j \sum_k e_{jj} D^{(s)}_{jk} e_{kk} + \sum_j \sum_k e_{jj} D^{(a)}_{jk} e_{kk} = \sum_j \sum_k e_{jj} D^{(s)}_{jk} e_{kk}, \tag{43}$$

because the anti-symmetric contribution is zero. Note that

$$\boldsymbol{D}^{(s)} = \begin{bmatrix} \varphi_{11} & \frac{1}{2}(\varphi_{11} + \varphi_{22}) & \frac{1}{2}(\varphi_{11} + \varphi_{33}) \\ \frac{1}{2}(\varphi_{11} + \varphi_{22}) & \varphi_{22} & \frac{1}{2}(\varphi_{22} + \varphi_{33}) \\ \frac{1}{2}(\varphi_{11} + \varphi_{33}) & \frac{1}{2}(\varphi_{22} + \varphi_{33}) & \varphi_{33} \end{bmatrix}. \tag{44}$$

Eventually, we obtain

$$(\mathrm{I}_\beta) = \boldsymbol{V}^T \begin{bmatrix} \boldsymbol{D}^{(s)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{V} = \boldsymbol{V}^T \boldsymbol{B} \boldsymbol{V}, \tag{45}$$

Ultimately, we can re-write the (I) as

$$\begin{aligned}(\mathrm{I}) &= \boldsymbol{V}^T\left(2\mu\boldsymbol{A} + \lambda\boldsymbol{B}\right)\boldsymbol{V} = \\ &= \boldsymbol{V}^T\left(\mu(\boldsymbol{B} + 2\boldsymbol{A} - \boldsymbol{B}) + \lambda\boldsymbol{B}\right)\boldsymbol{V} = \\ &= (\mu+\lambda)\boldsymbol{V}^T\boldsymbol{B}\boldsymbol{V} + \mu\boldsymbol{V}^T(2\boldsymbol{A} - \boldsymbol{B})\boldsymbol{V}^T\end{aligned} \tag{46}$$

Since $(\mu + \lambda) \geq 0$ and $\mu \geq 0$, the term (I) is larger, or at most equal, than zero $\forall \boldsymbol{V}$ if the bilinear forms $\boldsymbol{V}^T\boldsymbol{B}\boldsymbol{V}$ and $\boldsymbol{V}^T(2\boldsymbol{A} - \boldsymbol{B})\boldsymbol{V}$ are larger or equal than zero. In other words, $\boldsymbol{B}$ and $(2\boldsymbol{A} - \boldsymbol{B})$ must be positive semi-definite. Exploiting Sylvester's criterion, we found that the matrix $\boldsymbol{B}$ is positive semi-definite if and only if $\varphi_{ii} = \varphi \geq 0$, $\forall i$, whereas the matrix $2\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite if and only if $\varphi_{ij} \geq 0$, $\forall i, j$ with $i \neq j$.

- (II) manipulated to obtain

$$(\mathrm{II}) = k\sum_{i,j}\left(\frac{\partial T}{\partial x_i}\gamma_i\delta_{ij}\frac{\partial T}{\partial x_j}\right) = k\nabla T^T(\mathrm{diag}(\gamma_1,\gamma_2,\gamma_3))\nabla T, \tag{47}$$

therefore, (II) ≥ 0 if and only if $\mathrm{diag}(\gamma_1, \gamma_2)$ is positive semi-definite. According to Sylvester's criterion, this translates into the condition $\gamma_i \geq 0$

In conclusion, fulfilling the following relations on the correction term entries

1. $\varphi_{ii} = \varphi \geq 0, \quad \forall i,$
2. $\varphi_{ij} \geq 0, \forall i, j$ with $i \neq j,$
3. $\gamma_i \geq 0,$

is a sufficient condition for ensuring that Clausius-Duhem inequality is satisfied. With reference to f_{ij} and g_i, these conditions translate into

1. $f_{ii} = f \geq -1, \quad \forall i$
2. $f_{ij} \geq -1, \quad \forall i, j$ with $i \neq j$
3. $g_i \geq -1, \quad \forall i$

It's worth noting that, for the standard Navier-Stokes stress tensor and heat flux, we have $f_{ij} = 0$ and $g_i = 0$. As a result, we retrieve as expected that the Second Law of Thermodynamics is fulfilled for the standard Navier-Stokes equations.

B. Comparison between Navier-Stokes and DSMC solutions

The Variable Hard Sphere (VHS) model exhibits different behaviors at varying Mach numbers. Compared to DSMC data, the VHS model tends to be more diffusive at lower Mach numbers, while at higher Mach numbers it displays reduced diffusivity. Notably, at Mach number 3 (M=3), the VHS model demonstrates a commendable capability to replicate data obtained through DSMC methods. In contrast, the Sutherland model proves to be excessively steep for all Mach numbers, indicating limitations in its predictive accuracy for this test case.

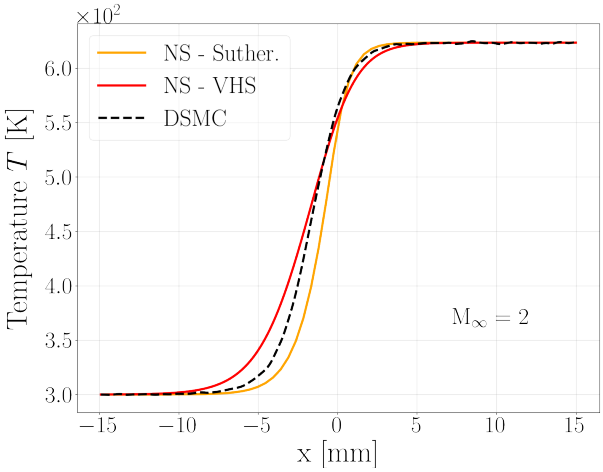


Fig. 8 x-Velocity

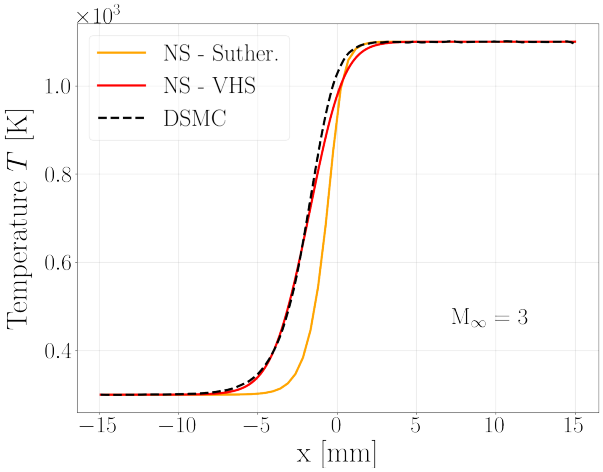


Fig. 9 y-Velocity

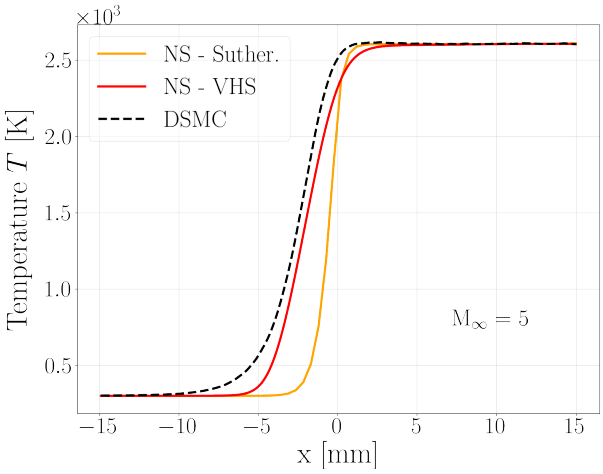


Fig. 10 Temperature

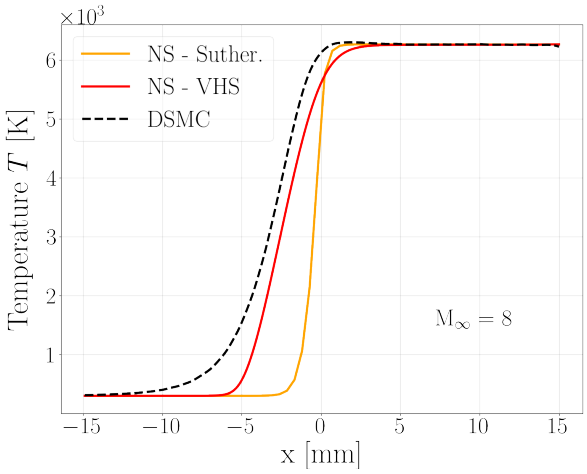


Fig. 11 Pressure

Acknowledgments

This material is based upon work supported by the Department of Defense, Office of Naval Research. The authors gratefully acknowledge computing resources provided by the CHESS department, University of Illinois at Urbana-Champaign. G. Gori would like to acknowledge that this work was partially funded by the European Union (Project 101059320 - UN-BIASED). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the European Research Executive Agency can be held responsible for them.

References

[1] Sirignano, J., MacArt, J., and Freund, J., "DPM: A deep learning PDE augmentation method with application to large-eddy simulation," *Journal of Computational Physics*, Vol. 423, 2020, p. 109811. https://doi.org/10.1016/j.jcp.2020.109811.

[2] Nair, A. S., Sirignano, J., Panesi, M., and MacArt, J. F., "Deep Learning Closure of the Navier–Stokes Equations for Transition-Continuum Flows," *AIAA Journal*, Vol. 61, No. 12, 2023, pp. 5484–5497. https://doi.org/10.2514/1.J062935.

[3] Vincenti, W. G., and Kruger, C. H., *Introduction to physical gas dynamics*, New York, Wiley, 1965.

[4] Chen, J., Ou, J., and Zhao, L., "Simulation of hypersonic flows in near-continuum regime using DSMC method and new extended continuum model," *AIP Conference Proceedings*, Vol. 2132, No. 1, 2019, p. 100007. https://doi.org/10.1063/1.5119602, URL https://aip.scitation.org/doi/abs/10.1063/1.5119602.

[5] Mott-Smith, H. M., "The Solution of the Boltzmann Equation for a Shock Wave," *Phys. Rev.*, Vol. 82, 1935, p. 885.

[6] Hornik, K., Stinchcombe, M., and White, H., "Multilayer feedforward networks are universal approximators," *Neural Networks*, Vol. 2, No. 5, 1989, pp. 359–366. https://doi.org/https://doi.org/10.1016/0893-6080(89)90020-8, URL https://www.sciencedirect.com/science/article/pii/0893608089900208.

[7] Brunton, S. L., and Kutz, J. N., *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*, Cambridge University Press, 2019. https://doi.org/10.1017/9781108380690.

[8] Brunton, S. L., Noack, B. R., and Koumoutsakos, P., "Machine Learning for Fluid Mechanics," *Annual Review of Fluid Mechanics*, Vol. 52, No. 1, 2020, pp. 477–508. https://doi.org/10.1146/annurev-fluid-010719-060214, URL https://doi.org/10.1146/annurev-fluid-010719-060214.

[9] Lu, L., Jin, P., Pang, G., Zhang, Z., and Karniadakis, G., "Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators," *Nature Machine Intelligence*, Vol. 3, 2021, pp. 218–229. https://doi.org/10.1038/s42256-021-00302-5.

[10] Bengio, Y., "Practical recommendations for gradient-based training of deep architectures," *Arxiv*, 2012.

[11] Noll, W., "The Foundations of Classical Mechanics in the Light of Recent Advances in Continuum Mechanics," *The Axiomatic Method*, Studies in Logic and the Foundations of Mathematics, Vol. 27, edited by L. Henkin, P. Suppes, and A. Tarski, Elsevier, 1959, pp. 266–281. https://doi.org/https://doi.org/10.1016/S0049-237X(09)70033-3, URL https://www.sciencedirect.com/science/article/pii/S0049237X09700333.

[12] Smith, D. R., *An Introduction to Continuum Mechanics - after Truesdell and Noll*, Springer Dordrecht, 1993.

[13] Lofthouse, A. J., Scalabrin, L. C., and Boyd, I. D., "Velocity Slip and Temperature Jump in Hypersonic Aerothermodynamics," *Journal of Thermophysics and Heat Transfer*, Vol. 22, No. 1, 2008, pp. 38–49. https://doi.org/10.2514/1.31280, URL https://doi.org/10.2514/1.31280.

[14] Frémond, M., "The Clausius-Duhem Inequality, an Interesting and Productive Inequality," *Nonsmooth Mechanics and Analysis*, edited by P. Alart, O. Maisonneuve, and R. T. Rockafellar, Springer US, Boston, MA, 2006, pp. 107–118.

[15] Sharma, B., Pareek, S., and Kumar, R., "Bulk viscosity of dilute monatomic gases revisited," *European Journal of Mechanics - B/Fluids*, Vol. 98, 2023, pp. 32–39. https://doi.org/https://doi.org/10.1016/j.euromechflu.2022.10.009, URL https://www.sciencedirect.com/science/article/pii/S0997754622001340.

[16] MacArt, J., "PyShock," , 2023.

[17] Solovchuk, M. A., and Sheu, T. W. H., "Prediction of strong-shock structure using the bimodal distribution function," *Phys. Rev. E*, Vol. 83, 2011, p. 026301.