Prediction of thermal conductance and friction coefficients at solid-gas interface from statistical learning of collisions

Meng Liao, Quy-Dong To, Céline Léonard, Wenlu Yang

Prediction of thermal conductance and friction coefficients at solid-gas interface from statistical learning of collisions

Meng Liao, Quy-Dong To,* and Céline Léonard

Université Paris-Est, Laboratoire Modelisation et Simulation Multi Echelle,

UMR 8208 CNRS, UPEC, F-77454 Marne-la-Vallée, France

Wenlu Yang

Sorbonne University, LIP6 CNRS, Paris, France

(Dated: September 11, 2018)

Abstract

In this paper, we present the construction of statistical models of gas-wall collision based on data issued from Molecular Dynamics (MD) simulations and use them to predict the velocity slip and temperature jump coefficients at the gas-solid interface. The Gaussian Mixture (GM) model, an unsupervised learning technique, is chosen for this purpose. The model shares some similarities with the well-known Cercignani-Lampis model in kinetic theory but it is more robust due to the unlimited number of Gaussian functions used and the ability to deal with correlated data of high dimensions. Applications to real gas-wall systems (Argon-Gold and Helium-Gold) confirm the good performance of the model. The trained GM model predicts physical and statistical properties including accommodation, friction and thermal conductance coefficients in excellent agreement with the MD model.

PACS numbers:

---

* Corresponding author: quy-dong.to@u-pem.fr

I. INTRODUCTION

Velocity slip and temperature jump are important interface phenomena which are frequently encountered in micro-nanofluidic systems [1]. The common way to model the imperfect boundary conditions is to use the following phenomenological relations

$$\beta_1 v_k = -\sigma_{kz}, \quad \beta_2(T - T_w) = -q_z, \quad k = x, y, \tag{1}$$

Here the wall is assumed to be stationary at temperature T_w and normal to direction z. The slip velocity v_k is proportional to the (minus) shear stress σ_{kz} and the temperature jump $T - T_w$ is proportional to the heat flux q_z. The constants, β_1 and β_2, are respectively the Navier isotropic friction and Kapitza thermal conductance coefficients. For gas flow problems, it is well-known that the interface phenomena become significant when the mean free path λ is comparable to the channel height H, starting from $\lambda/H \geq 0.01$. The origin of the Knudsen layer effect comes from two sources: the finite actual velocity/temperature jump at the boundary and the deviation from the bulk solution within the distance λ from the wall. In this paper, we do not consider the variation of the velocity/temperature in the Knudsen layer and assume that the linear relations Eq. (1) are valid at the solid boundary [2]. Since both the bulk and interfacial transport properties of gases are governed by collisions at the atomic scale [3], they are expected to be proportional to the collision rate. It is thus natural to non-dimensionalize the boundary coefficients β_1 and β_2 as follows

$$\overline{\beta}_1 = \frac{\beta_1}{n}\sqrt{\frac{\pi}{2k_B T m}}, \quad \overline{\beta}_2 = \frac{\beta_2}{n}\sqrt{\frac{m\pi}{8k_B^3 T}}, \quad k = x, y, \tag{2}$$

In Eq. (2), n is the gas density at the wall, m the atomic mass of the monatomic gas and k_B the Boltzmann constant. The other reason to non-dimensionalize the interfacial coefficients is that $\overline{\beta}_1$ and $\overline{\beta}_2$ are connected to tangential momentum accommodation coefficient α_1 and energy accommodation coefficients α_2 [1]

$$\overline{\beta}_1 = \frac{\alpha_1}{2 - \alpha_1}, \quad \overline{\beta}_2 = \frac{\alpha_2}{2 - \alpha_2} \tag{3}$$

The origin of the accommodation coefficients comes from the collision models of gas atoms with the solid boundary [4–6]. These constants appear in the linear relation between the pre-collision and post-collision momenta and energy of gas atoms with respect to the wall.

However, MD simulations showed that the linearity postulate is not always valid. Errors from the linear regression procedure to obtain α_1 and especially α_2 are observed [7, 8]. It is suggested that the more effective way to model slip and jump phenomena is to directly use coefficients $\overline{\beta}_1$ and $\overline{\beta}_2$ to quantify them from realistic collision data whenever possible. On the other hand, the formulation based on coefficients $\bar{\beta}_1, \bar{\beta}_2$ can avoid the errors due to the use of (3) and the intermediate accommodation coefficients. It is noted that while Eqs. (3) are widely used, they are derived under restrictive conditions, e.g the Chapman-Enskog distribution [3], an approximate solution of the Boltzmann equation, and the Maxwell or Cercignani-Lampis wall model [9, 10]. As a result, if we want to extend or investigate the validity (3) in general situations, more accurate methods (see e.g Refs. [11, 12]) should be adopted.

In the framework of probabilistic modeling, there are numerous ways of reconstructing the (collision) behavior of the system based on the available (collision) data, which are realizations via Molecular Dynamics simulations. One of the method, proposed in our previous work is to use the nonparametric (NP) techniques to estimate the probability density functions (PDF) and generate new collision data [8]. To avoid the sparsity of the data in high dimension, the reflection process was assumed to be independent for each velocity components (c'_x, c_x), (c'_y, c_y) and (c'_z, c_z) and the model can be constructed with a good accuracy [13]. Although such modeling is relatively general, the uncorrelation hypothesis can be strong and oversimplify the true behavior. This motivates us to investigate the use of unsupervised learning techniques and their performance with high dimension data. Specifically, the present work considers the Gaussian Mixture (GM) model, which is a popular technique and has proved to be successful in many data science applications. The GM are probabilistic models using linear combination of multiple multidimensional gaussians to estimate the probability density of data. Like all collision models (or scattering kernels) in kinetic theory, the construction of GM are based on parameters which can be determined from collision data for each gas-solid couple under consideration. While the existing collision models are limited by a finite number of parameters (e.g accommodation coefficients), the GM models are not. On the other hand, the GM models require a special algorithm, i.e Expectation Maximization (EM) algorithm to tune the parameters.

The paper is organized as follows. In Sect. II, we present a literature review of statistical models of collisions, the general principles to construct them from MD data and their properties. GM model is then introduced with the EM algorithm to tune the model parameters. Criteria based on both statistical and physical parameters including accommodation coefficients, interfacial coefficients, reflection angles, and Kullback-Leibler divergence will be used to evaluate the performance of the model. Applications to real material systems (Ar and He gases vs Au surface) are presented in Section III, confirming the superiority of the GM model when compared with nonparametric and parametric models of literature. Finally, some remarks and conclusions are given.

II. DESCRIPTION OF COMPUTATION METHODS

A. Overview of statistical models of collision and their properties

The collision model, also called scattering kernel in kinetic theory, serves as boundary conditions for Boltzmann equations and aims at describing how gas atoms are reflected after collision with the wall at temperature T_w. It is defined by the conditional probability distribution, $P(\text{output} = \boldsymbol{c}|\text{input} = \boldsymbol{c}', \text{temperature} = T_w)$ or denoted shortly $P(\boldsymbol{c}|\boldsymbol{c}')$ of velocity $\boldsymbol{c}$ for a given incoming velocity $\boldsymbol{c}'$. In addition to the usual PDF properties like non-negativeness and normalization, $P(\boldsymbol{c}|\boldsymbol{c}')$ is also expected to satisfy the physical reciprocity conditions [14–16]:

$$P_{eq}(\boldsymbol{c}')P(\boldsymbol{c}|\boldsymbol{c}') = P_{eq}(-\boldsymbol{c})P(-\boldsymbol{c}'|-\boldsymbol{c}), \quad \boldsymbol{c} \in \mathbb{R}^2 \times \mathbb{R}^+, \quad \boldsymbol{c}' \in \mathbb{R}^2 \times \mathbb{R}^- \tag{4}$$

where, $P_{eq}(\boldsymbol{c}')$ is the distribution associated to the equilibrium gas at the same temperature as the wall $T = T_w$ and motionless with respect to the wall (zero mean velocity)

$$P_{eq}(\boldsymbol{c}') = \sqrt{\frac{2\pi}{\theta}}|c_z'|M_B(\boldsymbol{c}'), \quad M_B(\boldsymbol{c}') = \frac{e^{-\boldsymbol{c}'^2/2\theta}}{(2\pi\theta)^{3/2}}, \quad \theta = k_B T/m \tag{5}$$

This condition implies that the incidence-reflection is microscopically reversible when the wall is in equilibrium state.

It can be shown that all popular collision models in literature can be classified as parametric

4

models, in most cases, taking from one to three accommodation coefficients as parameters. For example, in the Maxwell model [4] where the gas atom can be reflected either specularly or diffusively with percentage $1 - \alpha$ and α, only one accommodation coefficient is used. In the anisotropic Cercignani-Lampis (ACL) model [5, 6], three different accommodation coefficients associated to tangential momentums along x, y directions and normal kinetic energy along z direction are used. While the Maxwell model exhibits discontinuities in the probability space due to the portion of the mirror reflections (Dirac distribution), the smooth PDF from ACL model agrees better with MD simulations. However, like all parametric models, they are subject to common limitations: the requirement of the existence of a limited number of constant accommodation coefficients and the lack of flexibility in the case of complex collision data. Since the flow solution depends both on the gas-gas and gas-wall collision, those limitations may have consequences on the determination of boundary conditions' coefficients β_1 and β_2 in Navier-Stokes-Fourier equations as mentioned earlier and can cause errors in other simulation methods, e.g direct simulation Monte Carlo (DSMC) [17], MD [18, 19], moment equations [10, 20], etc. based on the collision model. It suggests that to achieve better accuracy, we need to rely on the whole data range and to not be limited by the number of model parameters.

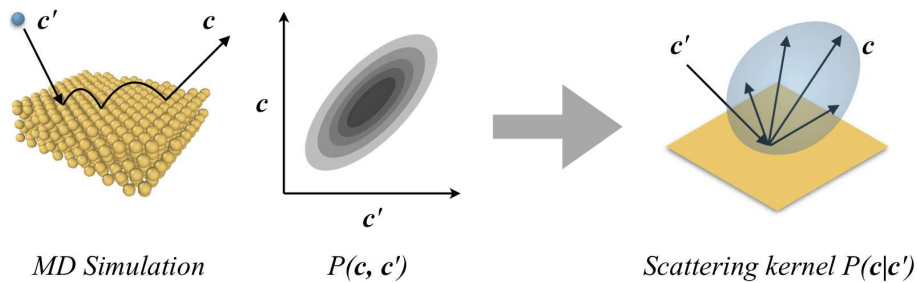*MD Simulation*   *P(c, c′)*   *Scattering kernel P(c|c′)*

FIG. 1. Procedure of reconstructing the scattering kernel from MD simulations of collisions.

The procedure of reconstructing the scattering kernel from MD simulations of collisions is shown in Fig. 1. The gas atoms are inserted with velocity $\boldsymbol{c}'$ at a plane beyond the cut-off distance and interact with the well thermostated solid atoms before escaping by crossing the same plane with velocity $\boldsymbol{c}$. Each couple of input and output velocities $(\boldsymbol{c}, \boldsymbol{c}')$ recorded is thus a realization of the joint PDF $P(\boldsymbol{c}, \boldsymbol{c}')$. In principle, if we can estimate accurately

$P(\boldsymbol{c}, \boldsymbol{c}')$, we can derive accurately the scattering kernel via the equation

$$P(\boldsymbol{c}|\boldsymbol{c}') = \frac{P(\boldsymbol{c}, \boldsymbol{c}')}{P(\boldsymbol{c}')}, \tag{6}$$

where $P(\boldsymbol{c}')$ is the distribution of input variable $\boldsymbol{c}'$. While $P(\boldsymbol{c}|\boldsymbol{c}')$ is a unique function of $\boldsymbol{c}$ and $\boldsymbol{c}'$, the joint PDF $P(\boldsymbol{c}, \boldsymbol{c}')$ depends on the input distribution $P(\boldsymbol{c}')$ via (6). In the special case where the input distribution is the equilibrium distribution, $P(\boldsymbol{c}') = P_{eq}(\boldsymbol{c}')$, the joint PDF becomes the equilibrium joint PDF $P(\boldsymbol{c}, \boldsymbol{c}') = P_{eq}(\boldsymbol{c}, \boldsymbol{c}')$. The reciprocity condition (4) is equivalent to the symmetry condition of $P_{eq}(\boldsymbol{c}, \boldsymbol{c}')$ in hyperspace

$$P_{eq}(\boldsymbol{c}, \boldsymbol{c}') = P_{eq}(-\boldsymbol{c}', -\boldsymbol{c}) \tag{7}$$

As a result, to model the scattering kernel via $P_{eq}(\boldsymbol{c}, \boldsymbol{c}')$ and the associated collision data, it is necessary that the above constraints are satisfied.

We remark that in many parametric scattering kernels [5, 6, 21], the reflection process occurs independently in x, y, and z directions. Under this assumption, both $P(\boldsymbol{c}|\boldsymbol{c}')$ and $P_{eq}(\boldsymbol{c}, \boldsymbol{c}')$ can be decomposed as:

$$P(\boldsymbol{c}|\boldsymbol{c}') = \prod_{k=1}^{3} P(c_k|c_k'), \quad P_{eq}(\boldsymbol{c}, \boldsymbol{c}') = \prod_{k=1}^{3} P_{eq}(c_k, c_k'), \quad k = x, y, z \tag{8}$$

A typical example is the well-known Cercignani-Lampis kernel and its associated joint PDF at equilibrium $P_{eq}^{CL}(\boldsymbol{c}, \boldsymbol{c}')$

$$P^{CL}(\boldsymbol{c}|\boldsymbol{c}') = \frac{e^{-\frac{[c_x - (1-\alpha_t)c_x']^2}{2\alpha_t(2-\alpha_t)\theta_w}}}{\sqrt{2\pi\theta_w\alpha_t(2-\alpha_t)}} \frac{e^{-\frac{[c_y - (1-\alpha_t)c_y']^2}{2\alpha_t(2-\alpha_t)\theta_w}}}{\sqrt{2\pi\theta_w\alpha_t(2-\alpha_t)}} \frac{c_z e^{-\frac{c_z^2 + (1-\alpha_e)c_z'^2}{2\alpha_e\theta_w}}}{\alpha_e\theta_w} \times$$

$$I_0\left(\frac{\sqrt{1-\alpha_e}c_z c_z'}{\alpha_e\theta_w}\right)$$

$$P_{eq}^{CL}(\boldsymbol{c}, \boldsymbol{c}') = \frac{e^{-\frac{(c_x^2 - 2(1-\alpha_t)c_x c_x' + c_x'^2)}{2\alpha_t(2-\alpha_t)\theta_w}}}{2\pi\theta_w\sqrt{\alpha_t(2-\alpha_t)}} \frac{e^{-\frac{(c_y^2 - 2(1-\alpha_t)c_y c_y' + c_y'^2)}{2\alpha_t(2-\alpha_t)\theta_w}}}{2\pi\theta_w\sqrt{\alpha_t(2-\alpha_t)}} \frac{|c_z c_z'| e^{-\frac{(c_z^2 + c_z'^2)}{2\alpha_e\theta_w}}}{\alpha_e\theta_w^2} \times$$

$$I_0\left(\frac{\sqrt{1-\alpha_e}c_z c_z'}{\alpha_e\theta_w}\right), \tag{9}$$

have this property. In particular, they are based on two parameters α_t, α_e, combination of multivariate Gaussian functions, and I_0 as modified Bessel function of first kind and zeroth order. Additionally, the joint PDF $P_{eq}^{CL}(\boldsymbol{c}, \boldsymbol{c}')$ also satisfies the symmetry condition (7).

It is interesting to note that a general class of 1D nonparametric collision model can also be constructed based on Eq. (8) and collision data. The uncorrelation hypothesis reduces the model from 3D (or 6D in data space $\boldsymbol{c}', \boldsymbol{c}$) to 1D (or 2D in each of 3 data spaces c'_x, c_x, c'_y, c_y, c'_z, c_z), which greatly simplifies the kernel construction. Using kernel density estimation (KDE) techniques, the constructed collision model has good performances when determining velocity slip and temperature jump coefficients like for the CH_4-graphite or Ar-Au systems [8, 22]. However, for nearly perfect elastic-reflection system like the He-Au system [22], the independent assumption of Eq. (8) can overestimate the reflective energy and can cause errors in the thermal conductance coefficient. In what follows, we shall explore the GM model, a popular and general purpose machine learning technique, which is not subject to above limitations. The model shares some similarities with the CL kernel which is among the best parametric models, by the presence of smooth multivariate Gaussian functions but it is much more powerful and flexible by means of superposition and unlimited number of the parameters used. Furthermore, the algorithms to identify those parameters from data are well developed in literature and in commercial software. Thus the efforts to construct an accurate collision model will be significantly minimized.

B. Preprocessing of training data and Gaussian mixture model (GM)

As mentioned earlier, the data are collected from MD simulations of independent gas-wall collisions at a given temperature T_w. The pre-collision velocity is drawn from equilibrium distribution, i.e $P(\boldsymbol{c}') = P_{eq}(\boldsymbol{c}')$ and $P(\boldsymbol{c}, \boldsymbol{c}') = P_{eq}(\boldsymbol{c}, \boldsymbol{c}')$ and we can fit the GM model with data, the realizations of $P_{eq}(\boldsymbol{c}, \boldsymbol{c}')$. To achieve the best performance of the method, we propose to precondition the data in the following way

- Symmetrization of the data: For each realization $(\boldsymbol{c}, \boldsymbol{c}')$, a virtual copy $(-\boldsymbol{c}', -\boldsymbol{c})$ is also added to the existing data. This step will double the size of the training data and ensure the symmetry of $P_{eq}(\boldsymbol{c}, \boldsymbol{c}')$ as shown in (7).

- Transformation of the random variables $\boldsymbol{c}$ and $\boldsymbol{c}'$: This is because of the support constraints

7

$c_z > 0$ and $c'_z < 0$ are not compatible with gaussian functions. Specifically, while c_x, c_y, c'_x, c'_y have Gaussian distributions, c_z, c'_z have Rayleighian distributions. In this case, we propose to adopt a transformation $\mathcal{T}$ that changes the distributions of c_z, c'_z to the same type as c_x, c_y, c'_x, c'_y. For example

$$\mathcal{T}(U) = \sqrt{2\theta}\,\text{erf}^{-1}\left[1 - 2\exp\left(-\frac{U^2}{2\theta}\right)\right], \quad \mathcal{T}^{-1}(U) = \sqrt{-2\theta\ln\left[\frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{U}{\sqrt{2\theta}}\right)\right]},$$

$$\boldsymbol{x}_R = \begin{bmatrix} c_x \\ c_y \\ \mathcal{T}(c_z) \end{bmatrix}, \quad \boldsymbol{x}_I = \begin{bmatrix} c'_x \\ c'_y \\ \mathcal{T}(c'_z) \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} x_{R1} \\ x_{R2} \\ \mathcal{T}^{-1}(x_{R3}) \end{bmatrix}, \quad \boldsymbol{c}' = \begin{bmatrix} x_{I1} \\ x_{I2} \\ -\mathcal{T}^{-1}(x_{I3}) \end{bmatrix}, \quad (10)$$

where $\boldsymbol{x}_R$ and $\boldsymbol{x}_I$ are two vector which are used to record the reflective value or incident value respectively. We note that to map the Gaussian to the Rayleigh distribution, we compute the CDFs of Gaussian distribution and of Rayleigh distribution

$$F^G(\mathcal{T}) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\mathcal{T}}{\sqrt{2\theta}}\right)\right], \quad F^R(U) = 1 - \exp\left[-U^2/(2\theta)\right]. \quad (11)$$

and use the relation $F^G(\mathcal{T}) = F^R(U)$ to recover the transformation (10) concerning c_z, c'_z.

The GM model is a probabilistic model that assumes all the data points are generated from a mixture of Gaussian distributions [23]. More importantly, it is general enough to estimate complex PDF function. Specifically, the GM estimator of $P(\boldsymbol{x}_R, \boldsymbol{x}_I)$ can be written as a superposition of M Gaussian functions with weights $\varphi_1, \varphi_2, .., \varphi_M$ as follows

$$P(\boldsymbol{x}_R, \boldsymbol{x}_I) \simeq P^{GM}(\boldsymbol{x}_R, \boldsymbol{x}_I) = \sum_{m=1}^{M} \varphi_m p_m(\boldsymbol{x}_R, \boldsymbol{x}_I), \quad p_m(\boldsymbol{x}_R, \boldsymbol{x}_I) = \mathcal{N}(\boldsymbol{X}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \quad (12)$$

Each Gaussian $p_m(\boldsymbol{x}_R, \boldsymbol{x}_I)$ has its own parameters $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ as mean vector and covariance matrix

$$\mathcal{N}(\boldsymbol{X}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\boldsymbol{\Sigma}_m|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_m)\right],$$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_R \\ \boldsymbol{x}_I \end{bmatrix}, \quad \boldsymbol{\mu}_m = \begin{bmatrix} \boldsymbol{\mu}_{mR} \\ \boldsymbol{\mu}_{mI} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m = \begin{bmatrix} \boldsymbol{\Sigma}_{mRR} & \boldsymbol{\Sigma}_{mRI} \\ \boldsymbol{\Sigma}_{mIR} & \boldsymbol{\Sigma}_{mII} \end{bmatrix}, \quad (13)$$

where d is the dimension of variables $\boldsymbol{X}$. Comparing with the parametric CL model (9), the bivariate gaussians also appear in the independent reflection process x and y where the covariance parameters $(1 - \alpha_t)$ are connected to the tangential momentum accommodation coefficients. In the case $M = 1$, the GM model also shares some strong similarities with Nocilla [24] model based on a drifted Maxwellian for reflected velocities. However, the application of the GM model on the preprocessed collision data rather than the original data, as mentioned previously has fully eliminated the negative reflective velocity c_z that the drifted Maxwellian can not. Furthermore, as seen in (12,13), the GM model is much more general. Constituted of multiple multivariate Gaussian functions, it does not rely on independent reflection hypothesis and contains numerous parameters to fit with data. Those are φ_m, $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ with $m = 1, 2, .., M$ which can be effectively determined using the EM algorithm. (see Appendix 1) [25–27].

The scattering kernel is the conditional probability function of reflective velocity $P(\boldsymbol{x}_R | \boldsymbol{x}_I)$ given incident velocity and this function can also be written as a mixture form[23, 28]:

$$P^{GM}(\boldsymbol{x}_R | \boldsymbol{x}_I) = \sum_{m=1}^{M} \tilde{\varphi}_m(\boldsymbol{x}_I) p_m(\boldsymbol{x}_R | \boldsymbol{x}_I). \tag{14}$$

with the following weights

$$\tilde{\varphi}_m(\boldsymbol{x}_I) = \frac{\varphi_m p_m(\boldsymbol{x}_I)}{\sum_{m=1}^{M} \varphi_m p_m(\boldsymbol{x}_I)}. \tag{15}$$

The marginal and the conditional distributions appearing in equations (14) and (15) are calculated with the formula

$$p_m(\boldsymbol{x}_I) = \int p_m(\boldsymbol{x}_R, \boldsymbol{x}_I) d\boldsymbol{x}_R = \mathcal{N}(\boldsymbol{x}_I | \boldsymbol{\mu}_{mI}, \boldsymbol{\Sigma}_{mII}), \tag{16}$$

and

$$\begin{aligned} p_m(\boldsymbol{x}_R | \boldsymbol{x}_I) &= \frac{p_m(\boldsymbol{x}_R, \boldsymbol{x}_I)}{p_m(\boldsymbol{x}_I)} = \mathcal{N}(\boldsymbol{x}_R | \boldsymbol{\mu}_{mR|I}, \boldsymbol{\Sigma}_{mR|I}), \\ \boldsymbol{\mu}_{mR|I} &= \boldsymbol{\mu}_{mR} + \boldsymbol{\Sigma}_{mRI} \boldsymbol{\Sigma}_{mII}^{-1} (\boldsymbol{x}_I - \boldsymbol{\mu}_{mI}), \\ \boldsymbol{\Sigma}_{mR|I} &= \boldsymbol{\Sigma}_{mRR} - \boldsymbol{\Sigma}_{mRI} \boldsymbol{\Sigma}_{mII}^{-1} \boldsymbol{\Sigma}_{mIR}. \end{aligned} \tag{17}$$

In the next section, we will study the performance of the GM model in comparison with the NP and CL kernels.

C. Evaluation of model performance and comparisons

Generally, the criteria to judge the quality of a statistical model is the accuracy of fitting with the given data and more importantly, the ability to generate new data predicting quantities consistent with the realistic system. For evaluation, we propose to compute physical based parameters including accommodation coefficients, the reflective deviation angle and boundary conditions coefficients $\bar{\beta}_1$ and $\bar{\beta}_2$ at the interface using the GM model and compare with the same parameters issued from NP, CL kernels and the realistic MD simulations.

As mentioned earlier, accommodation coefficients appear in parametric collision models as constants. However, for realistic atomistic systems, these constants may not be properly defined. In literature, there exist several methods to identify those coefficients but they cannot guarantee the uniqueness of the obtained results (see examples in Appendix 2). Among those works, we consider the method based on correlation analysis [8, 29]. In this case, the accommodation coefficients are defined via the correlation parameters of the data

$$\alpha_1 = 1 - \frac{\langle c'_x c_x \rangle - \langle c'_x \rangle \langle c_x \rangle}{\langle c'^2_x \rangle - \langle c'_x \rangle^2}, \quad \alpha_2 = 1 - \frac{\langle c'^2 c^2 \rangle - \langle c'^2 \rangle \langle c^2 \rangle}{\langle c'^4 \rangle - \langle c'^2 \rangle^2}$$

$$\alpha_n = 1 - \frac{\langle |c'_z| c_z \rangle - \langle |c'_z| \rangle \langle c_z \rangle}{\langle c'^2_z \rangle - \langle c'_z \rangle^2}, \quad \alpha_e = 1 - \frac{\langle c'^2_z c^2_z \rangle - \langle c'^2_z \rangle \langle c^2_z \rangle}{\langle c'^4_z \rangle - \langle c'^2_z \rangle^2} \tag{18}$$

While α_1 and α_2 are the tangential momentum accommodation coefficient (TMAC) and the energy accommodation coefficient (EAC), respectively (see equation (3)), α_e is the normal momentum accommodation coefficients (NMAC) and α_e is the normal energy accommodation coefficients (NEAC). Those expressions are derived from the linear regression of collision data (c'_x, c_x), (c'^2, c^2), $(|c'_z|, c_z)$ and (c'^2_z, c^2_z), respectively. Being special composite parameters depending on the moments of $P(\boldsymbol{c}, \boldsymbol{c}')$, they can be sensitive to the choice of $P(\boldsymbol{c}, \boldsymbol{c}')$ and $P(\boldsymbol{c}')$. In what follows, we assume that those parameters are obtained from the equilibrium distribution $P_{eq}(\boldsymbol{c}, \boldsymbol{c}')$ which is also the basis of our model. For the sake of simplicity, we continue to call them accommodation coefficients despite knowing that from definition (18), they are more statistical than physical quantities.

In addition to the correlation analysis of velocities in association with the accommodation coefficients described above, we also examine in details the statistical relations between the incident γ', reflection γ and deviation η angles as defined in Fig. 2.
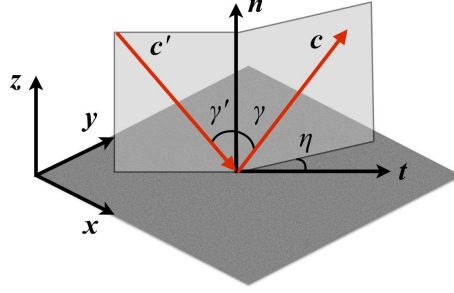

FIG. 2. A simple scheme of the incident-reflective velocities, c'-c, the incident-reflective vertical angles, γ'-γ, and the deviation angle η.

Other physical parameters to be checked are the interfacial coefficients $\bar{\beta}_1$ and $\bar{\beta}_2$. From the kinetic consideration of the terms involved in equations (1) and (2), $\bar{\beta}_1$ and $\bar{\beta}_2$ can be computed by the expressions

$$\bar{\beta}_1 = \frac{\sqrt{\pi/2\theta}\langle c'_x - c_x\rangle}{\langle c'_x/|c'_z| + c_x/|c_z|\rangle}, \quad \bar{\beta}_2 = \frac{\sqrt{9\pi/32\theta}\langle c'^2 - c^2\rangle}{\langle c'^2/|c'_z| + c^2/|c_z|\rangle - 3\theta_w\langle 1/|c'_z| + 1/|c_z|\rangle} \tag{19}$$

Because of the non-equilibrium origin of the expressions, those coefficients cannot be determined using the equilibrium distribution for the gas at the same temperature and the same velocity as the wall, i.e $P(c') \neq P_{eq}(c'|\theta = \theta_w, \langle c'_x\rangle = 0)$, otherwise we shall encounter numerical issues. Instead, for the determination of $\bar{\beta}_1$, the equilibrium distribution with a non zero mean velocity $\langle c'_x\rangle \neq 0$ and the temperature of the wall $\theta = \theta_w$ is used as the input gas distribution $P(c') = P_{eq}(c'|\theta = \theta_w, \langle c'_x\rangle \neq 0)$ and for $\bar{\beta}_2$, we used the equilibrium distribution with zero mean velocity $\langle c'_x\rangle = 0$ and a temperature different from the wall temperature $\theta \neq \theta_w$ or $P(c') = P_{eq}(c'|\theta \neq \theta_w, \langle c'_x\rangle = 0)$. We also highlight the importance of transforming the variables c_z and c'_z in the data preconditioning step before training the GM model. This step guarantees the vanishing probability at $c'_z = 0$ and $c_z = 0$ like the Rayleigh distribution and the convergence of the rational moments containing $1/c'_z$ and $1/c_z$. Unlike the previous analysis of accommodation coefficients and reflective angle based on the existing collision data at equilibrium, this method requires new realizations,

i.e using the new input $P(\boldsymbol{c}')$ to generate new data $P(\boldsymbol{c}', \boldsymbol{c})$ with the wall model $P(\boldsymbol{c}|\boldsymbol{c}')$ (parametric/nonparametric kernels, GM model or atomic wall).

Another criteria to evaluate the statistical performance of different scattering kernels is based on the relative entropy, the Kullback-Leibler divergence (D_{KL}) [30]. This quantity is a measure of the differences between the probabilities P and Q. Defined as the expectation of the logarithmic difference, its discrete form is written as:

$$D_{\mathrm{KL}}(P\|Q) = \sum_j P(\boldsymbol{x}_j) \log \frac{P(\boldsymbol{x}_j)}{Q(\boldsymbol{x}_j)}, \tag{20}$$

To compute $P(\boldsymbol{x}_j)$ and $Q(\boldsymbol{x}_i)$, we divide the data space into n bins and choose a Parzen-window size h, then the probability density of data generated by MD simulation or a scattering kernel is estimated. The density in each bin can be written as:

$$P(\boldsymbol{x}_j) = \frac{N_j}{\sum_j N_j}, \tag{21}$$

where, $\boldsymbol{x}_j$ represents the center of bin j, N_j is the number of data in the window of length h centered on $\boldsymbol{x}_j$. We note that the above formula can be applied to any distribution of random variables $\boldsymbol{x}$, for example velocity, kinetic energy, collision angles etc..

III.   APPLICATIONS TO REAL GAS-WALL SYSTEMS

A.   Construction of scattering kernels from Molecular Dynamics simulation data

Based on the same systems as Refs. [22, 31], we shall carry out more detailed analysis of the MD simulation results, in order to construct and evaluate statistical models of collision. We simulate the process of He or Ar gas atoms impacting smooth/rough Au atomic surface and investigated the gas-solid friction and thermal conductance. The rough surface is generated by depositing randomly atoms on the smooth surface. The He or Ar molecules are inserted one by one in the collision zone, only after the molecule interacts with the wall and goes out of the collision zone, another molecule is inserted in the zone from a random posi-

Pair	A[eV]	B[eVÅ^{-1}]	C[eVÅ]	α[Å^{-1}]	γ[Å^{-2}]	C_6[eVÅ^6]	C_8[eVÅ^8]	β[Å^{-1}]
He-Au	-2.9755	0.47351	4.8980	-2.576	0.8204	9.5073	131.18	3.618
Ar-Au	1.623×10^4	-	-	3.356	-	76.785	1066.9	3.051

TABLE I. The parameters of the He-Au and Ar-Au pairwise interaction potentials [22, 31].

tion with a random velocity. The MD simulations are performed by LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator)[32]. Each collision simulation was run with one gas atom at a time. The simulation parameters will be described briefly as follows as follows:

- temperature of the Au surface: 300 K, controlled by the Nosé-Hoover thermostat

- to compute β_1 we apply a average velocity to the incident gas atom: ranging from -0.5 Å/ps to 0.5 Å/ps (with a step of 0.25 Å/ps)

- to compute β_2 we apply a initial temperature to the incident gas: ranging from 100 K to 400 K (with a step of 100 K)

The highly optimized version of multibody EAM potential [33] is employed for the description of the interaction between the Au atoms. The Au-Ar and Au-He pairwise potentials, which are recently determined from ab-initio studies [22, 31], are implemented in LAMMPS and adopted in this work. The functional form of the pair potentials $V_{\mathrm{g-Au}}(R)$ is given by

$$V_{\mathrm{g-Au}}(R) = \left(A + BR + \frac{C}{R}\right) e^{(-\alpha R - \gamma R^2)} + \sum_{n=6,8} \frac{C_n}{R^n} f_n\left(\beta R\right). \tag{22}$$

where R is the distance between the gas atom ($'$g$'$ standing for He or Ar) and a gold atom (Au). The parameters A, B, C, α, and γ are associated with the repulsive part and the remaining parameters correspond to the attractive part of the potential. The function f_n is the damping function of Tang and Toennies [34],

$$f_n(x) = 1 - e^{-x}\left(1 + x + ... + \frac{x^n}{n!}\right). \tag{23}$$

The numerical values of the pairwise potential parameters are given in Tab.I

In this paper, to construct the scattering kernel, we use the velocity data of the MD simulations under the equilibrium state, in which the temperature of gas flow is equal to the

temperature of the atomic wall (300 K) and the mean velocity of the flow equals to 0. We can calculate the accommodation coefficients for the CL kernel using Eq. (18). For the NP kernel, we followed the model of Ref. [22], the He velocities range from -40 to 40 Å/ps, and the Ar velocities range from -20 to 20 Å/ps. The incident and reflective velocity spaces were discretized into 100 and 1000 intervals, respectively, the Parzen-window size $h = 1$ Å/ps. The probability that a reflective velocity does not belong to a category is determined by interpolation. In the GM model, the number of mixtures M is a free parameter which needs to be chosen before the learning. To avoid the risk of overfitting or underfitting, there is an optimal value of M for a given MD data size. We tested the number of Gaussian functions from $M = 1$ to 512 with different initial covariance matrices and mean vectors. To find the optimal M we calculate the energy accommodation coefficient α_2 using the reflective velocity generated by GM model and Eq. (18). In Fig.3, we found that when M is equal to 64 (2^6), the performance of the GM model is the best and much better than CL and NP models. So the GM model with $M = 64$ has a good accuracy and a reasonable computation time and is used in the present study.
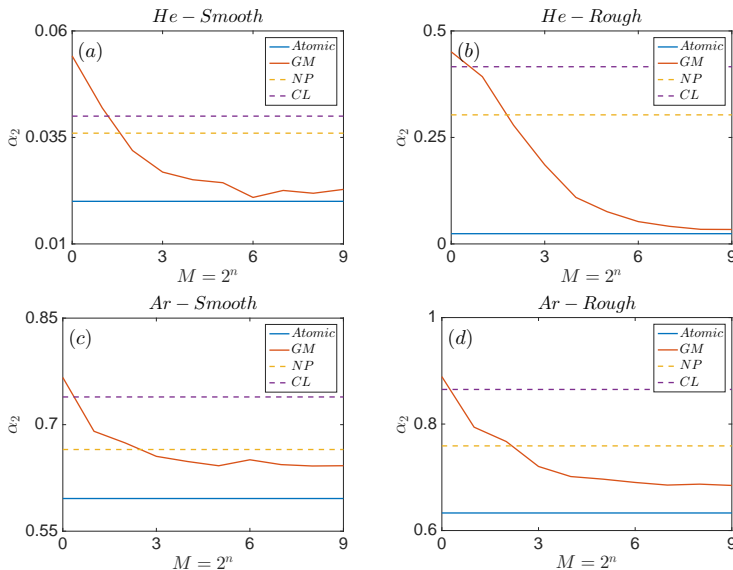

FIG. 3. Number of Gaussian vs. α_2 determined by statistical models and MD

We then used each kernel, CL, NP and GM, to generate new incident-reflective velocities (about 10^5 data points) based on incident velocities under equilibrium or non-equilibrium states. We investigated the characteristics of the scattering kernels according to the following

14

aspects:

- the accommodation coefficients and the momentum/energy distribution

- the incident, reflective, and deflection angles

- the surface friction and the thermal conductance

B. Accommodation coefficients

We decomposed the velocity into the normal and the tangential components and then studied the correlations between the incident-reflective velocities and the kinetic energy. Using Eq. (18), we can calculate TMAC (α_1), NMAC (α_n), and EAC (α_2) from incident-reflective velocity data. Tab. II presents TMAC, NMAC and EAC of He/Ar gas on the smooth/rough Au surface using the atomic wall and the different statistical models of collision (GM, NP, CL). From this table, we find that TMAC and NMAC obtained by the different kernels are very close to the atomic simulation results. However, significant differences are observed for the coefficient EAC. The result of the GM kernel is consistent with the atomic model, but the NP and CL kernels considerably overestimate EAC, especially for the He-Au rough surface system. We note that, unlike Ar, He is a special gas which has a very small atomic mass and its interaction with the gold surface is very weak (see ref. [22]). These points can explain the nearly perfect elastic collision of an helium atom on a Au wall, even for the rough atomic surface. In other words, while the atomic motion can change its direction after collisions due to the surface corrugation (e.g via the influences of the atomic roughness, the thermal motion, and the energy landscape), the atomic kinetic energy is almost conserved.

Using the density estimator presented in Eq. (21), we can estimate the joint probability of the incident-reflective velocities, and of the kinetic energy. The kinetic energy distributions determined by the atomic wall use and by different scattering kernels are plotted in Fig. 4. The energy clouds of the atomic and GM models are visibly narrower than those of the CL and NP kernels. From the column figures of both atomic and GM surface, we can see that roughness does not significantly change EAC (16.7% for He, 5.9% for Ar). However, this trend cannot be captured by both NP and CL kernels.

Gas	Surface model	Smooth			Rough		
		α_1	α_n	α_2	α_1	α_n	α_2
He	Atomic	0.012	0.052	0.020	0.326	0.293	0.024
	GM	0.012	0.052	0.021	0.327	0.297	0.031
	NP	0.015	0.058	0.040	0.323	0.295	0.303
	CL	0.012	0.055	0.036	0.325	0.317	0.416
Ar	Atomic	0.425	0.803	0.596	0.688	0.831	0.633
	GM	0.425	0.798	0.597	0.686	0.835	0.643
	NP	0.450	0.823	0.665	0.701	0.844	0.759
	CL	0.424	0.814	0.739	0.689	0.841	0.865

TABLE II. The tangential momentum accommodation coefficient (TMAC) α_1, the normal momentum accommodation coefficient (NMAC) α_n, and the energy accommodation coefficient (EAC) α_2, computed using different statistical models: GM, NP, CL, and the atomic (MD) model. The construction of the statistical models NP, GM and CL from MD data is discussed in sec. IIIA. The accommodation coefficients associated to the statistical models are computed based on generation of new data and Eq. (18). It is noted that the CL model uses tangential momentum accommodation coefficient (TMAC) and normal energy accommodation coefficient (NEAC) computed by MD as parameters. Consequently, numerical TMAC values of CL model are extremely close to MD results.

These phenomenas are further illustrated in Fig. 5. The KL divergence D_{KL} between different statistical models and the atomic model was calculated. We can notice that for tangential and normal momentums, the D_{KL} of the GM and NP kernels is at the same level. For the CL kernel, although the D_{KL} for c_x and c_z are larger than those of the other two kernels, its good performance in determining TMAC and NMAC (Tab. II) indicates that CL kernel is still a good kernel to generate a single velocity component. Meanwhile, Fig. 5b presents anomalies for the kinetic energy. The energy D_{KL} of NP and CL kernels are close to 1 for the He/rough Au case, meaning that the incident-reflective kinetic energy distributions determined by the NP and CL kernels differ greatly from the results of the atomic simulations. Considering the good performance of CL and NP kernels for the momentum reflection, and less for the kinetic energy, it is suggested that the independent assumption of Eq. (8) may oversimplify the true behavior especially for nearly perfect elastic systems like He/Au. This deficiency has been properly covered by the GM model. We note that the EAC value difference between the statistical models is due to the correlation parameter

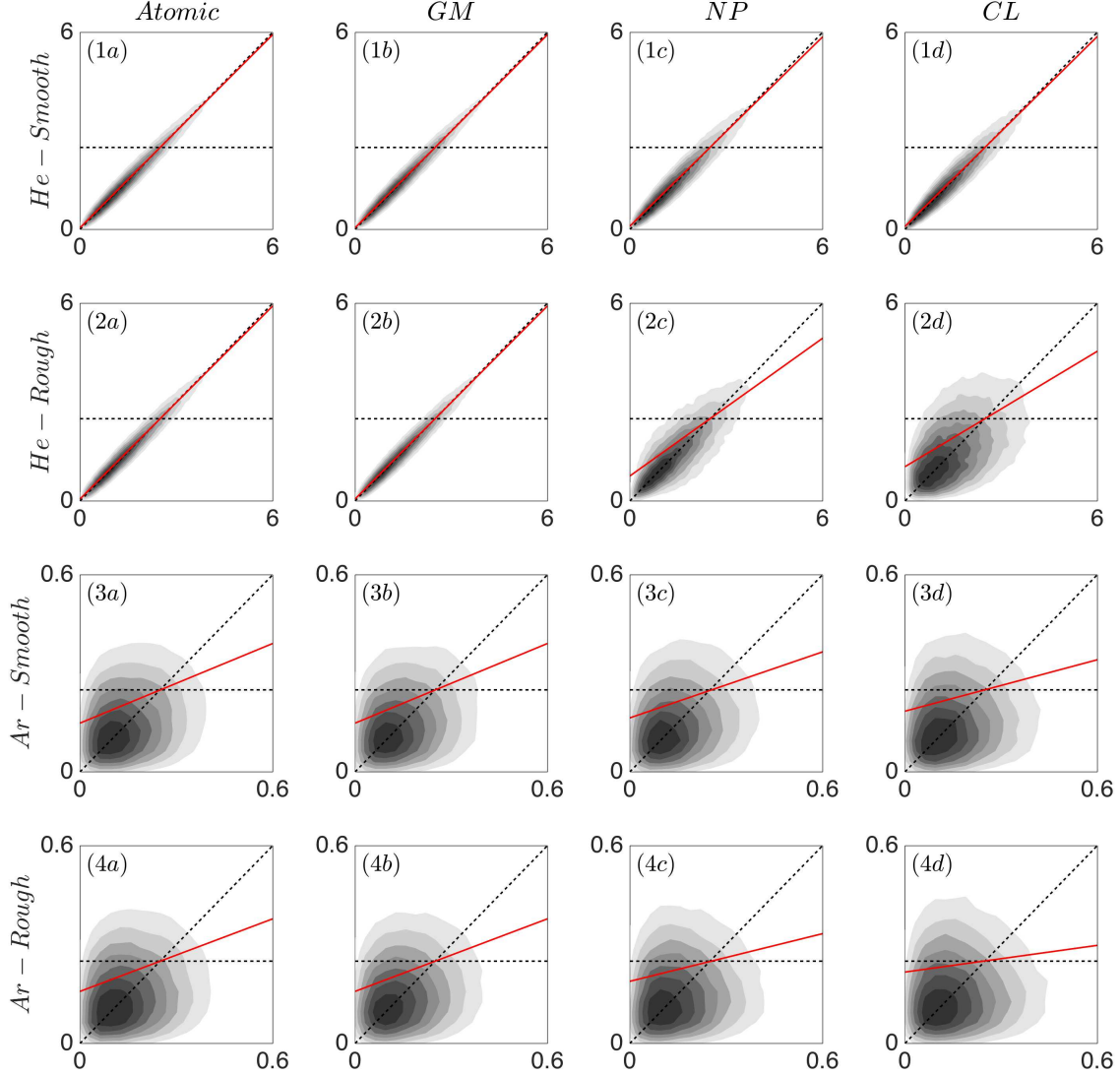$$\langle c'^2 c^2 \rangle = \int c'^2 c^2 P_{eq}(\boldsymbol{c}, \boldsymbol{c}') d\boldsymbol{c} d\boldsymbol{c}'$$

FIG. 4. The joint probability density of incident-reflective kinetic energy with unite $\text{Å}^2/ps^2$ (the prefactor m/2 is removed for simplicity) determined by the atomic simulations, the Gaussian mixture (GM), the nonparametric (NP), and the Cercignani-Lampis (CL) kernels. The horizontal axis indicates the incident energy $\boldsymbol{c'}^2$ and the vertical axis indicates the reflective energy $\boldsymbol{c}^2$. The red line is the least-square linear regression of kinetic data, its slope is equal to 1-α. When the red line is close to the diagonal dashed line, the reflection is close to be specular (elastic collision). When the red line is close to the horizontal dashed line, the reflection is diffusive.

As a result, the accuracy of $P_{eq}(\boldsymbol{c}, \boldsymbol{c'})$ governs the accuracy of the correlation parameter and EAC. While it is not trivial to connect the elastic collision phenomena to the property of the probability function $P_{eq}(\boldsymbol{c}, \boldsymbol{c'})$ or $P(\boldsymbol{c}|\boldsymbol{c'})$, we can conclude mathematically that using the decomposition postulate (8) like CL and NP models oversimplifies the real behavior of $P_{eq}(\boldsymbol{c}, \boldsymbol{c'})$. In other words, due to the exchange between the velocity components in order to conserve the energy, each of post-collision velocity components is expected to depend on all
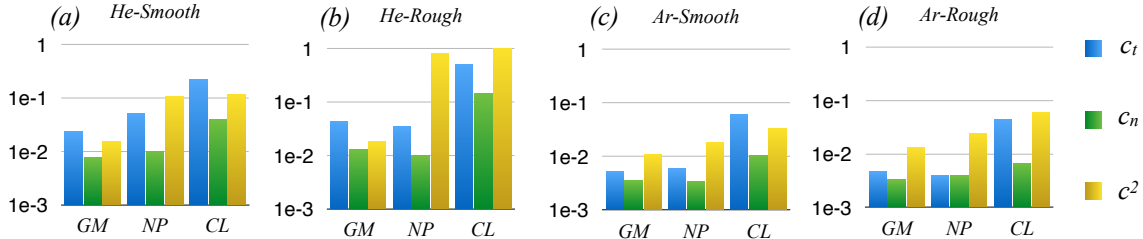
the pre-collision components.

FIG. 5. The Kullback Leibler divergence $D_{KL}(Atomic||Kernel)$ of momentum and kinetic energy determined by the atomic simulations, the Gaussian mixture (GM) kernel, the nonparametric (NP) kernel, and the Cercignani-Lampis (CL) kernel.

C. Collision angles

To further study the performance of each scattering kernel, the probability densities of deviation angle η and incident-reflective vertical angles γ'-γ determined by atomic simulations and scattering kernels were analyzed. The geometric definitions of η, γ', γ are illustrated in Fig. 2.

Angle η can be calculated from velocity components c'_x, c'_y, c_x, c_y and measures the deviation of the gas atom reflection from the incident direction in xy-plane. The incident angle γ' and reflective angle γ depend on the incident, $\boldsymbol{c}'$, and reflective, $\boldsymbol{c}$, velocities, respectively. By calculating the correlation between γ' and γ, we can understand the aberration of the reflection in the normal direction.

Figure 6 shows the probability of the reflective direction on the tangential plane obtained from atomic simulation data, GM, NP and CL kernels under equilibrium state. Numerical results show that the rough surface can increase the deviation of the reflective direction with respect the incident direction in xy-plane, especially the percentage of backscattering ($\pi/2 < \eta < 3\pi/2$). As the result, the overall scattering is more diffusive and the probability contour is shifted towards and envelope the origin. In contrast, the contour of reflection on smooth surfaces is narrower. From analysis of gas atoms, the adhesion between Au wall and Ar gas atoms is stronger than for the He-Au system, so that the reflection direction has a greater probability of deviating from the specular direction. From Fig. 6, we can find that
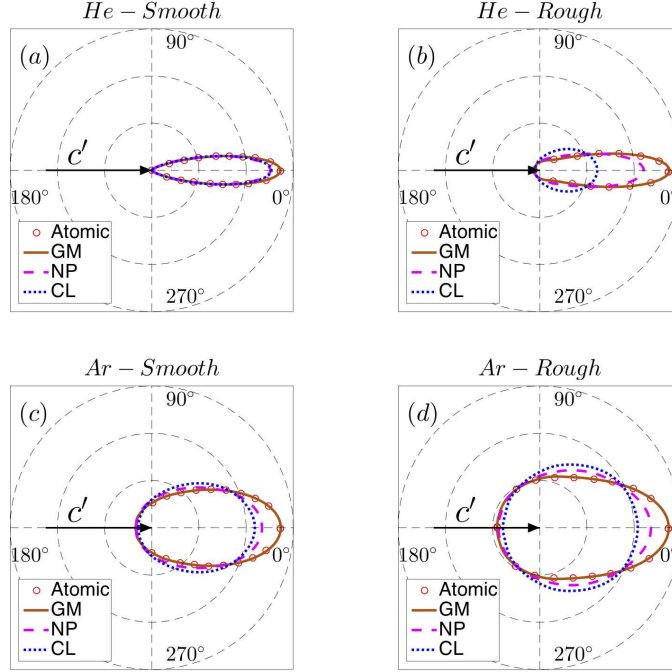
18

FIG. 6. The probability density of the reflective deviation angle η determined by the atomic simulations, the Gaussian mixture (GM), the nonparametric (NP), and the Cercignani-Lampis (CL) kernels in polar coordinate system. The distance from the curve to the origin, the radius, indicates the probability of associated to angle η, which coincides with the angle coordinate of the system.

both CL and NP kernels overestimate the diffusivity of the reflective direction. Only the probability contour of the GM kernel is identical with the one of the the atomic simulations.

Figure 7 shows the joint probability density of γ' and γ. When the deviation angle is small, the data population becomes narrow and shrinks toward the diagonal line. Through these cloud figures, we can also confirm that the roughness of the wall surface and the bonding properties between the gas and the wall affect the reflective direction of the gas atom. To quantitatively compare the differences between the three types of kernels and the atomic simulations, we calculate the KL divergence between the scattering kernels and the atomic wall results. The KL divergence of the reflective angle probability is presented in Fig. 8. The results of the figure confirm the differences observed in Figures 6 and 7. The KL divergence of the CL kernel is over 3 times larger than that of the GM kernel, especially for He-Au rough surface system. Basically, the CL kernel incorrectly estimates the distribution of the reflection direction. The results of the NP kernel for the Ar-Au system are better
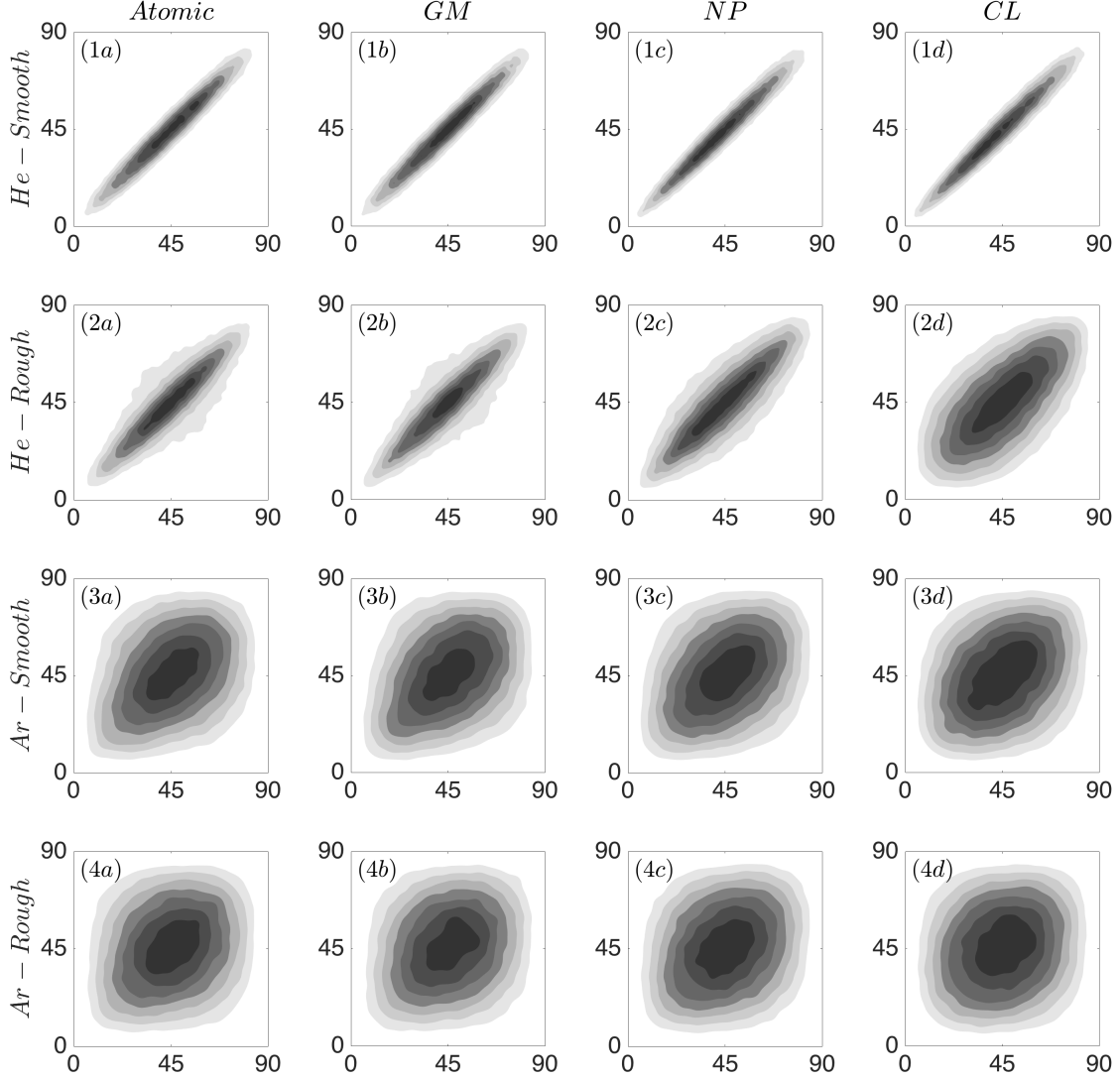
FIG. 7. The probability density of the incident-reflective vertical angles determined by the atomic simulations, the Gaussian mixture (GM), the nonparametric (NP), and the Cercignani-Lampis (CL) kernels. The horizontal axis indicates the incident angle γ' and the vertical axis indicates the reflective angle γ. The incident and reflective angles range from 0 to 90°

than the CL kernel ones, the corresponding KL divergence is under 2 times larger than the GM kernel ones in terms of tangential deviation angle η. We can clearly see that the GM kernel has a great advantage in predicting the distribution of the reflection directions.
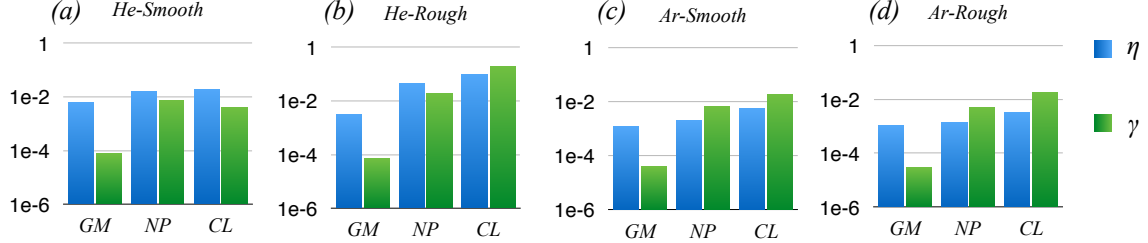
20

FIG. 8. The Kullback-Leibler divergence $D_{KL}(Atomic||Kernel)$ derived form Fig. 6 and 7.

D.   Interfacial friction and thermal conductance

The surface friction and thermal conductance properties can be characterized by the velocity slip coefficient, $\bar{\beta}_1$, and the temperature jump coefficient, $\bar{\beta}_2$, respectively. These two coefficients can be determined by the accommodation coefficients determined previously and relation (3) or by a direct method based on equation (19) and the non-equilibrium collision simulation data. This step permits to test the ability of the GM model in terms of generating new data and reproducing these important physical properties for real systems. We note that if the final objective is to construct kernels usable in kinetic simulation methods, this step is crucial and guarantees the reliability of the results.

The results for two coefficients are reported in Tab. III. The coefficient $\bar{\beta}_1$ depends on the accommodation momentums and is sensitive to the surface roughness. It means that the friction significantly increases $\bar{\beta}_1$. These effects are observed for both Ar/Au et He/Au systems. From smooth surface to rough surface, the $\bar{\beta}_1$ value increases about 34 times for He-Au system and about 2 times for Ar-Au system. In contrast, the contribution of the surface roughness to the increase of the thermal conductance is relative small, about 20% for He and 9% for Ar.

The numerical tests confirm again the superior performance of the GM model. This is the statistical model that produces the best results when compared with the atomic model. It is surprising to find that the simple analytical relation between β and α (Eq. (3)) works relatively well and yields values of the same order as MD and GM models. It is less accurate than the GM model but better than other models in terms of evaluating the thermal coefficient $\bar{\beta}_2$. In Appendix 2, we also carry out theoretical investigation of CL models.

21

Gas	Scheme	Smooth		Rough	
		$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_1$	$\bar{\beta}_2$
He	Atomic	0.0069	0.0123	0.236	0.015
	GM	0.0064	0.0120	0.240	0.020
	NP	0.0087	0.0194	0.194	0.176
	CL	0.0060	0.0186	0.205	0.252
	$\alpha/(2-\alpha)$	0.0060	0.0100	0.195	0.012
Ar	Atomic	0.276	0.445	0.571	0.485
	GM	0.283	0.442	0.555	0.484
	NP	0.301	0.459	0.559	0.548
	CL	0.273	0.518	0.530	0.644
	$\alpha/(2-\alpha)$	0.270	0.424	0.524	0.463

TABLE III. The friction coefficient ($\bar{\beta}_1$) and the thermal conductance ($\bar{\beta}_2$) determined by different schemes.

Specifically, we compute analytically the slip and jump coefficients $\bar{\beta}_1$ and $\bar{\beta}_2$ for the CL model with parameters α_t, α_e using the same approaches as the MD model. Results show that the obtained expressions for $\bar{\beta}_1$ and $\bar{\beta}_2$ in terms of α_t, α_e are different but the numerical differences are very small for the whole admissible range $(0, 1)$ of α_t and α_e. However as discussed previously, the coefficients α_1 and α_2 defined by Eq. (18) are more of statistical nature and furthermore α_2 is not associated to any parametric statistical models in literature (except for the Maxwell model). This fact will limit the application of those coefficients in particle and/or statistical based simulation methods like MD or DSMC. In contrary, the GM model is not subject to this limitation.
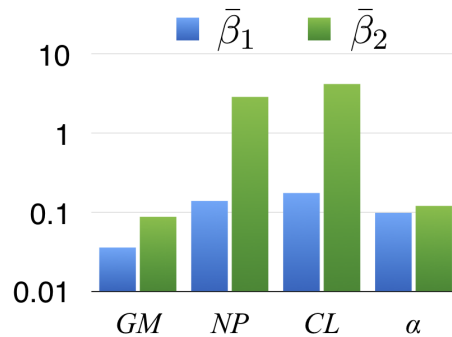

FIG. 9. The average relative error of the different schemes, GM, NP, CL, and α, comparing with the atomic simulations.

IV. CONCLUSION

In this paper, we present the application of unsupervised learning techniques in the construction of wall-gas models and the computation of interfacial properties like friction and thermal conductance. The data obtained from the Molecular Dynamics simulations of gas-wall collision is used to train the scattering kernels, i.e statistical collision models that predict the post-collision velocity from a given pre-collision velocity. Due to the common similarities with the well-known Cercignani-Lampis (CL) model in kinetic theory and the good overall performance in general data treatment, the GM model is chosen for this purpose.

Unlike CL and other parametric scattering kernels in literature, the GM model is flexible with unlimited number parameters and supported by EM algorithm, a robust technique to identify those parameters from data. We also propose to use physical and statistical based criteria e.g accommodation coefficients, friction and thermal conductance, collision angles to evaluate the performance of the GM model. Applications to real gas-wall systems including He-Au and Ar-Au confirm the accuracy of the model. All the coefficients obtained/predicted by the GM model are close to MD results. The overall accuracy of the trained GM model assessed by KL divergence is also superior to other models. The constructed model is useful both for theoretical studies of boundary conditions and for the gas simulation methods like MD and DSMC by saving considerable computation cost devoted to the atomic wall.

APPENDIX

1. Expectation maximization algorithm

Expectation maximization (EM) algorithm is an iterative method to find the maximum likelihood of parameters in statistical models, which is introduced by Dempster *et. al.* [26]. The parameter set of the GM model is noted as $\Theta = (\theta_1, \theta_2, ...\theta_M)$ and the mth parameter is $\theta_m = (\varphi_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$. Noting samples as $\boldsymbol{X}$, i.e $\boldsymbol{X}_i$ corresponds to the ith training data of total N training data, the likelihood can be written as

$$\mathcal{L}(\boldsymbol{X}|\Theta) = \prod_{i=1}^{N} p(\boldsymbol{X}_i|\Theta),$$

$$p(\boldsymbol{X}_i|\Theta) = \sum_{m=1}^{M} \varphi_m \mathcal{N}(\boldsymbol{X}_i|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{24}$$

To determine the parameter Θ, we can maximize the log-likelihood function:

$$\Theta = \arg \max_{\Theta} \log \mathcal{L}(\boldsymbol{X}|\Theta)$$

$$= \arg \max_{\Theta} \sum_{i=1}^{N} \log p(\boldsymbol{X}_i|\Theta), \tag{25}$$

and the sun of log-likelihood function in Eq. (25) can be replaced by the expectation of log-likelihood function:

$$\Theta = \arg \max_{\Theta} \mathbb{E}\left(\log p(\boldsymbol{X}_i|\Theta)\right). \tag{26}$$

Then, we can use an iterative algorithm to determine the parameter Θ of the GM model. The EM algorithm can be concluded in four steps: initial state, E-step, M-step and convergence test:

1. Initial state: take the initial value $\theta_m^{(0)} = (\varphi_m^{(0)}, \boldsymbol{\mu}_m^{(0)}, \boldsymbol{\Sigma}_m^{(0)})$ for each Gaussian distribution in the GM model.

2. E-step: use the parameters of the $(r-1)^{th}$ iteration to calculate the mixture weights w_{im}

of each sample $\boldsymbol{X}_i$ for the $(r)^{th}$ iteration:

$$w_{im} = \frac{\varphi_m^{(r-1)} \mathcal{N}(\boldsymbol{X}_i | \boldsymbol{\mu}_m^{(r-1)}, \boldsymbol{\Sigma}_m^{(r-1)})}{\sum_m \varphi_m^{(r-1)} \mathcal{N}(\boldsymbol{X}_i | \boldsymbol{\mu}_m^{(r-1)}, \boldsymbol{\Sigma}_m^{(r-1)})}. \tag{27}$$

3. M-step: use the w_{im} calculated by E-step to determine the new parameters $\boldsymbol{\mu}_m^{(r)}$, $\boldsymbol{\Sigma}_m^{(r)}$ and $\varphi_m^{(r)}$:

$$\mu_m^{(r)} = \frac{\sum_i w_{im} \boldsymbol{X}_i}{\sum_i w_{im}}$$

$$\Sigma_m^{(r)} = \frac{\sum_i w_{im}(\boldsymbol{X}_i - \mu_m^{(r)})(\boldsymbol{X}_i - \mu_m^{(r)})^T}{\sum_i w_{im}}$$

$$\varphi_m^{(r)} = \frac{1}{n} \sum_{i=1}^{n} w_{im}. \tag{28}$$

4. convergence test: conclude the iterative process if the following conditions are met:

$$\left| 1 - \frac{\mathbb{E}\left(\log p(\boldsymbol{X}_i | \Theta^{(r)})\right)}{\mathbb{E}\left(\log p(\boldsymbol{X}_i | \Theta^{(r-1)})\right)} \right| < \varepsilon, \tag{29}$$

where, ε is a threshold. Repeat the steps 2 to 4 until the expectation of the log-likelihood function converge.

The advantage of the EM algorithm is that it guarantees local convergence, but the drawback is that the algorithm is sensitive to the initial state. In practice, the k-means algorithm [25] and the iterative pairwise replacement algorithm (IPRA) [23] have a good performance in giving an initial state for the GM model. In addition, we can also use multiple random initial states and then compare the distributions of GMs obtained with different initial states.

2. Comparisons of interfacial coefficients associated to the CL model by different methods

We consider the CL kernel whose analytical expression is given by (9) and examine the coefficients α_1, α_2, $\bar{\beta}_1$, $\bar{\beta}_2$ computed from the correlation analysis and by the non-equilibrium method.

Using the equilibrium distribution $P_{eq}^{CL}(\boldsymbol{c}, \boldsymbol{c}')$, the moments $f(\boldsymbol{c}, \boldsymbol{c}')$ in correlation analysis can be calculated by

$$\langle f(\boldsymbol{c}, \boldsymbol{c}')\rangle = \int f(\boldsymbol{c}, \boldsymbol{c}')P_{eq}^{CL}(\boldsymbol{c}, \boldsymbol{c}')d\boldsymbol{c}d\boldsymbol{c}' \tag{30}$$

which yields the explicit expressions

$$\langle c_x\rangle = \langle c_x'\rangle = 0, \quad \langle c_x c_x'\rangle = (1 - \alpha_1)\theta_w, \quad \langle c_x^2\rangle = \langle c_x'^2\rangle = \theta_w, \quad \langle c_z^2\rangle = \langle|c_z'^2|\rangle = 2\theta_w,$$

$$\langle c^2\rangle = \langle c'^2\rangle = 4\theta_w, \quad \langle c^2 c'^2\rangle = 4\theta_w{}^2\left(\alpha_t{}^2 - 2\alpha_t - \alpha_e + 6\right), \quad \langle c^4\rangle = \langle c'^4\rangle = 24\theta_w^2. \tag{31}$$

Substituting the above results in (18), we obtain coefficients α_1, α_2

$$\alpha_1 = \alpha_t, \quad \alpha_2 = \frac{1}{2}[\alpha_e + \alpha_t(2 - \alpha_t)] \tag{32}$$

and the associated interfacial coefficients, $\bar{\beta}_1$, $\bar{\beta}_2$

$$\bar{\beta}_1 = \frac{\alpha_t}{2 - \alpha_t}, \quad \bar{\beta}_2 = \frac{\alpha_e + \alpha_t(2 - \alpha_t)}{2 + (1 - \alpha_e) + (1 - \alpha_t)^2}, \tag{33}$$

under assumption of the validity of the relation $\bar{\beta}_i = \alpha_i/(2 - \alpha_i)$ with $i = 1, 2$.

Regarding the non-equilibrium method, we impose the input distribution $P(\boldsymbol{c}')$ at a different temperature or velocity from the wall and use the scattering kernel to compute $P(\boldsymbol{c})$

$$P(\boldsymbol{c}) = \int P^{CL}(\boldsymbol{c}|\boldsymbol{c}')P(\boldsymbol{c}')d\boldsymbol{c}' \tag{34}$$

To determine $\bar{\beta}_1$ and $\bar{\beta}_2$ with (19), we need to compute the moments

$$\langle f(\boldsymbol{c}')\rangle = \int f(\boldsymbol{c}')P(\boldsymbol{c}')d\boldsymbol{c}', \quad \langle f(\boldsymbol{c})\rangle = \int f(\boldsymbol{c})P(\boldsymbol{c})d\boldsymbol{c} \tag{35}$$

yielding the final results (after considering $\theta \simeq \theta_w$)

$$\bar{\beta}_1 = \frac{\alpha_t}{2 - \alpha_t}, \quad \bar{\beta}_2 = \frac{3}{2}\left[\frac{\alpha_e + \alpha_t(2 - \alpha_t)}{(1 - \alpha_e) + 2(1 - \alpha_t)^2 + 3}\right]. \tag{36}$$

We also noted that in literature, there also exists different expressions of coefficients $\bar{\beta}_1, \bar{\beta}_2$ (see Ref. [5, 16])

$$\bar{\beta}_1 = \frac{\alpha_t}{2 - \alpha_t}, \quad \bar{\beta}_2 = \frac{\alpha_e + \alpha_t(2 - \alpha_t)}{4 - \frac{2}{5}[3\alpha_e + 2\alpha_t(2 - \alpha_t)]}. \tag{37}$$

Despite using non equilibrium gas distribution as input distribution, the derivation of these equations is different from the non equilibrium method used in the present work. It is based on the moment balance at the wall and the temperature, velocity and stress and heat flux are parameters of input distribution (Chapman-Enskog) and are not computed from both input and output distributions as in (19).

A general remark can be made here that while all the methods agree in terms of slip effects (coefficient $\bar{\beta}_1$), analytical expressions for the interfacial conductance are different (coefficient $\bar{\beta}_2$) . It is suggested that these discrepancies are due to the fact that the EAC coefficient is undefined for the CL wall kernel [8], except for the diffusive wall $\alpha_e = \alpha_t = 1$. Nevertheless, the discrepancies between the methods are not significant, maximum 6% at the extreme contrast cases $\alpha_t = 1, \alpha_e = 0$ or $\alpha_t = 0, \alpha_e = 1$ (see Fig. 10).
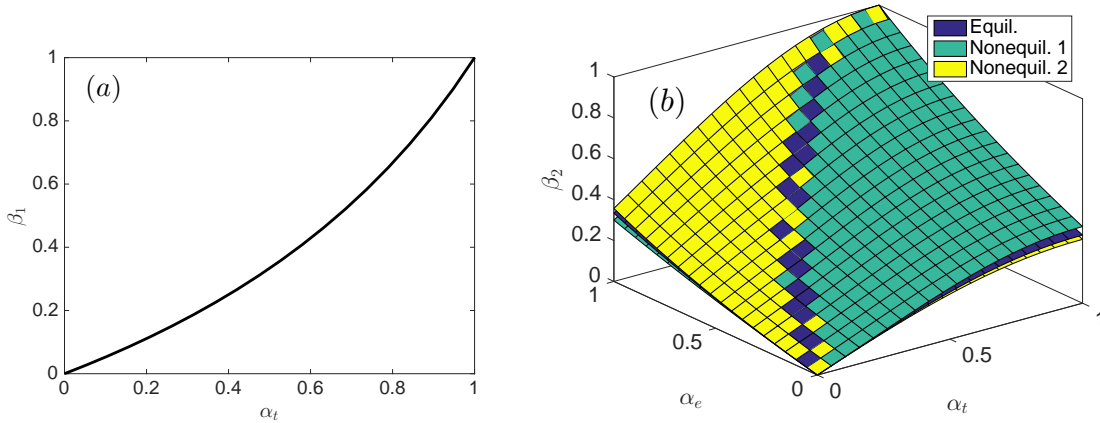

FIG. 10. (a) β_1 as a function of α_t. (b) Comparisons between different methods of computing β_2 with the same CL models: the equilibrium method (33), the nonequilibirum method 1 in this work (36), the nonequilibirum method 2 in literature (37).

ACKNOWLEDGMENTS

[1] G. Karniadakis, A. Beskok, and N. Aluru, *Microflows and nanoflows: fundamentals and simulation*, Vol. 29 (Springer Science & Business Media, 2006).

[2] Q. D. To, C. Léonard, and G. Lauriat, Physical Review E **91**, 023015 (2015).

[3] S. Chapman and T. Cowling, *The mathematical theory of non-uniform gases: an account of the kinetic theory of viscosity, thermal conduction, and diffusion in gases* (Cambridge University Press, 1970).

[4] J. Maxwell, Philos. T R. Soc. A **170**, 231 (1879).

[5] Q.-D. To, V.-H. Vu, G. Lauriat, and C. Léonard, J. Math. Phys. **56**, 103101 (2015).

[6] C. Cercignani and M. Lampis, Transport Theory Stat. Phys. **1**, 101 (1971).

[7] T. T. Pham, Q. D. To, G. Lauriat, C. Léonard, and V. Van Hoang, Phys. Rev. E **86**, 051201 (2012).

[8] M. Liao, Q.-D. To, C. Léonard, and V. Monchiet, Phys. Fluids **30**, 032008 (2018).

[9] H. Struchtrup, Continuum Mech. Therm. **17**, 43 (2005).

[10] H. Grad, in *Thermodynamik der Gase/Thermodynamics of Gases* (Springer, 1958) pp. 205–294.

[11] F. Sharipov, Eur. J. Mech. B Fluids **22**, 133 (2003).

[12] F. Sharipov, Eur. J. Mech. B Fluids **22**, 145 (2003).

[13] M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric statistical methods*, Vol. 751 (John Wiley & Sons, 2013).

[14] C. Cercignani, *Theory and application of the Boltzmann equation* (Scottish Academic Press, 1975).

[15] S. Brull, P. Charrier, and L. Mieussens, Kin. Rel. Mod **38**, 219 (2014).

[16] H. Struchtrup, Phys. Fluids **25**, 112001 (2013).

[17] G. A. Bird, NASA STI/Recon Technical Report A **76** (1976).

[18] D. Rapaport, *The Art of Molecular Dynamics Simulation* (Cambridge University Press, 2004).

[19] M. Allen and D. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, 1989).

[20] H. Struchtrup and M. Torrilhon, Phys. Fluids **15**, 2668 (2003).

[21] S. K. Dadzie and J. G. Méolans, J. Math. Phys. **45**, 1804 (2004).

[22] M. Liao, R. Grenier, Q.-D. To, M. P. de Lara-Castells, and C. Léonard, J. Phys. Chem. C, year=2018.

[23] H. G. Sung, *Gaussian mixture regression and classification*, Ph.D. thesis, Rice University (2004).

[24] F. G. Collins and E. Knox, AIAA J. **32**, 765 (1994).

[25] S. Calinon, *Robot Programming by Demonstration: A Probabilistic Approach* (EPFL/CRC Press, 2009) ePFL Press ISBN 978-2-940222-31-5, CRC Press ISBN 978-1-4398-0867-2.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, J. Royal Stat. Soc. Series B , 1 (1977).

[27] T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning , 236 (2001).

[28] C. E. Rasmussen and C. K. Williams, The MIT Press, Cambridge, MA, USA **38**, 715 (2006).

[29] P. Spijker, A. J. Markvoort, S. V. Nedea, and P. A. J. Hilbers, Phys. Rev. E **81**, 011203 (2010).

[30] S. Kullback and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1951).

[31] R. Grenier, Q.-D. To, M. P. de Lara Castells, and C. Léonard, J. Phys. Chem. A **119**, 6897 (2015).

[32] S. Plimpton, P. Crozier, and A. Thompson, "Large-scale atomic/molecular massively parallel simulator, http://lammps.sandia.gov (accessed may 20, 2018),".

[33] H. W. Sheng, M. J. Kramer, A. Cadien, T. Fujita, and M. W. Chen, Phys. Rev. B **83**, 134118 (2011).

[34] K. T. Tang and J. P. Toennies, J. Chem. Phys. **80**, 3726 (1984).