Solving fluid flow problems using semi-supervised symbolic regression on sparse data

El Hasadi, Yousef M.F.; Padding, Johan T.

DOI
10.1063/1.5116183

Publication date
2019

Document Version
Final published version

Published in
AIP Advances

Solving fluid flow problems using semi-supervised symbolic regression on sparse data

Cite as: AIP Advances **9**, 115218 (2019); https://doi.org/10.1063/1.5116183
Submitted: 24 June 2019 . Accepted: 05 November 2019 . Published Online: 26 November 2019

Yousef M. F. El Hasadi, and Johan T. Padding

Solving fluid flow problems using semi-supervised symbolic regression on sparse data

Cite as: AIP Advances 9, 115218 (2019); doi: 10.1063/1.5116183
Submitted: 24 June 2019 • Accepted: 5 November 2019 •
Published Online: 26 November 2019

Yousef M. F. El Hasadi[a] and Johan T. Padding

AFFILIATIONS

Process and Energy Department, Delft University of Technology, Leeghwaterstraat 39, 2628 CB Delft, The Netherlands

[a]Author to whom correspondence should be addressed: G.DamianidisAlChasanti@tudelft.nl and yme0001@auburn.edu

ABSTRACT

The twenty first century is the century of data. Machine learning data and driven methods start to lead the way in many fields. In this contribution, we will show how symbolic regression machine learning methods, based on genetic programming, can be used to solve fluid flow problems. In particular, we will focus on the fluid drag experienced by ellipsoidal and spherocylinder particles of arbitrary aspect ratio. The machine learning algorithm is trained semisupervised by using a very limited amount of data for a specific single aspect ratio of 2.5 for ellipsoidal and 4 for spherocylindrical particles. The effect of the aspect ratio is informed to the algorithm through what we call previous knowledge, for example, known analytical solutions in certain limits, or through interbreeding of different flow solutions from the literature. Our results show good agreement with literature results, while they are obtained computationally faster and with less computing resources. Also, the machine learning algorithm discovered that for the case of prolate spheroids, the difference between the drag coefficients perpendicular and parallel to the flow in the high Reynolds number regime only depend on the aspect ratio of the geometry, even when the individual drag coefficients still decrease with increasing *Re*.

I. INTRODUCTION

Fluid mechanics is one of the most widely investigated topics in scientific literature due to its wide range of applications, ranging from spacecraft design to bacterial flows. Fluid flow is described by the Navier-Stokes equations. Due to the strong nonlinear nature of these equations, finding analytical or numerical solutions imposes a challenging task. The most popular numerical schemes used for solving the Navier-Stokes equations are finite difference,[1] finite volume,[2] finite element,[3] and lattice Boltzmann[4] schemes. For more accurate numerical schemes, finer grids and smaller time steps must be used, which leads to an increase in the computational time and usage of more computational resources. Even with the introduction of parallelization into computational fluid mechanics (CFD) algorithms, the algorithms start to reach their peak performance for two reasons. The first reason is that the algorithms themselves seem to have reached their peak, making it increasingly difficult to find higher order, more accurate discretization schemes.[5] The second reason is the limitation of the computing capacity due to the restrictions that physical laws enforce onto the computer hardware.[6]

In the last 50 years, an enormous amount of data has been generated related to fluid mechanics problems, either from experiments or from computer simulations. As a result of this substantial increase in the available data, our understanding of different fluid mechanics problems has improved significantly.[7] The data take different forms, such as simple datasets stored in computer hard drives or as analytical solutions[8] and mathematical correlations,[9] or as graphical images.[10] Most of these data are available on the World Wide Web, through scientific journals, or in scientific libraries and are readily accessible. With the increasing availability of data, data driven methods of machine learning start to gain substantial ground in the field of predicting the outcome of fluid dynamics problems. Machine learning methods used in fluid mechanics take different forms such as nonlinear regression,[11] neural networks,[12] combined smooth particle hydrodynamics (SPH) with regression,[13] symbolic regression,[14,15] and finding the governing differential equations by using sparse identification of nonlinear dynamics.[16,17] All these methods of machine learning use enormous amounts of data, which are sometimes very expensive to obtain computationally or experimentally.

However, machine learning can also be employed to find accurate correlations from relatively sparse datasets. The purpose of this paper is to show how accurate predictions in the field of fluid dynamics can be made by feeding machine learning methods a low amount of high quality measurements. As a leading example, we will focus on correlations for hydrodynamic forces on a single axisymmetric nonspherical particle, which is a classical fluid mechanics problem with a wide range of applications in different engineering disciplines. Before introducing our machine learning approach, it is insightful to first give an overview of the more classical theoretical and computational fluid dynamics approaches.

During the previous decades, lots of research efforts have focused on the flow behavior of particulate suspensions. Particulate suspensions are interesting because of their wide range of applications, ranging from blood to controlling the flow behavior of biomass, pastes, and ceramics. Researchers tend to investigate the flow behavior of a single particle first, before introducing the complicating factor of particle interactions. The simplest case is fluid flow around a single sphere in the Stokes regime, i.e., at a vanishingly small Reynolds number $Re \ll 1$. Stokes[18] found the analytical solution for the velocity field and the drag force. After this ground breaking work, Oseen[19] extended the analysis to include the effects of fluid inertia for small but finite Re numbers, and he obtained an analytical solution for the drag coefficient, which was found to be logarithmically dependent on Re. Using the method of asymptotic expansion, Proudman and Pearson[20] extended the work of Oseen to Reynolds numbers of order 0.1. For higher Reynolds numbers, analytical methods cease to exist, and the only tool available is direct numerical methods as in Ref. 21. There are additional hydrodynamic forces such as lift and torque, which are especially important for nonspherical particles. As for the case of spherical particles, there is a very limited number of analytical solutions for very small Reynolds number and mostly for highly symmetric and high aspect ratio particles. Examples include cylinders,[22,20] prolate ellipsoids,[23] and oblate ellipsoids.[24] For higher values of Re, as for the case of spherical particles, numerical methods are utilized to elucidate the flow around, and forces on, nonspherical particles. The first numerical results for spheroids were obtained by Pitter *et al.*[25] for oblate spheroids and by Masliyah and Epstein[26] for oblate and prolate spheroids. They solved the steady state form of the Navier- Stokes equations using a finite difference scheme. They obtained the drag coefficient for the case of a particle oriented parallel to the flow, for Re up to 100, and a very limited set of aspect ratios. For higher $Re > 100$, there is a limited number of investigations for certain particle shapes and aspect ratios. Vakil and Green[27] solved the flow around a circular cylinder for $1 < Re < 40$ and a range of aspect ratios between 1 and 20 using the commercial software Fluent, which uses the finite volume method. They obtained results for the drag and lift coefficients for different orientations of the particle relative to the flow. They summarized their results in two sets of correlations and showed that the maximum value for the lift to drag ratio occurs at an angle of attack between 40° and 50° for nearly all aspect ratios tested. Using a similar approach, Ouchene *et al.*[28] investigated prolate spheroids, over a wide range of aspect ratios from 1 to 32, at much higher Re (up to $Re = 240$). They presented correlations for the drag, lift, and torque, as a function of Re, angle of attack, and aspect ratio. Zastawny *et al.*[29] investigated the flow around four different particles, two prolate spheroids of aspect ratios 1.25 and 2.5, an oblate spheroid with aspect ratio 0.2, and a fiber particle with aspect ratio 5. They used the immersed boundary to solve the fluid flow around the particle for Re up to 300 and provided correlations for the drag, lift, and torque coefficients. These correlations contain multiple fitting parameters and are specific for each particle shape. Sanjeevi *et al.*[30] investigated flow around nonspherical particles for even higher Re up to 2000, well beyond the steady state regime, using a lattice Boltzmann scheme combined with a second order immersed boundary method. They investigated prolate spheroids with aspect ratio 2.5, oblate spheroids with aspect ratio 0.4, and fibers with aspect ratio 4. They were careful to make the correlations consistent with known physical limits for very low and very high Re. Their correlations are similar to those of Ref. 29. They found that in the Stokes regime, the lift and torque coefficients show a symmetric dependence on the angle of attack with a maximum at 45°, while at higher Re, this dependence becomes slightly skewed, related to the appearance of unsteady flow patterns.

In all the above works, the correlations were obtained by suggesting a functional (analytical) form with fit parameters that best describe the measured drag, lift, and torque coefficients. The optimization of the parameters is relatively easy, e.g., by application of a nonlinear optimization routine that minimizes the root mean square (rms) error between measurement and correlation prediction. However, *the choice of the functional form itself* is still largely a matter of trial-and-error. In this paper, we will use machine learning, specifically symbolic regression, to solve the issue of functional form selection. Symbolic regression uses concepts akin to genetic evolution by random mutations, interbreeding, and natural selection to find the best functional form. This is done without user interference, with one exception: the user needs to define the so-called function space of possible functions and operations on these functions. We will illustrate our approach by focusing on the drag experienced by fibers and prolate particles. We will use only a limited amount of data from Ref. 30 combined with some previous literature knowledge for the training purpose. To the best of our knowledge, this is the first time that generalized regression equations for the drag coefficients for a wide range of Re, aspect ratios and angle of attacks are obtained in this way. There are a number of novelties in this work. The first one is a new form of correlations for the drag coefficient of fibers and prolate spheroids for different Reynolds number, angles of attack, and aspect ratios. The second one is the way they are obtained from a very small set of data. We will show how symbolic regression is able to predict the role of particle aspect ratio, i.e., the role of a latent variable that was not explicitly varied in the original dataset, by choosing the function space based on functional forms suggested by previous (literature) knowledge. We believe that our method can be used more generally to solve a variety of fluid mechanics problems in a faster way compared to the trial-and-error method mentioned above. The final novelty of this work is that it shows that symbolic regression can find new hidden physics related to the evolution of the drag coefficient of nonspherical particles that will be reported for the first time.

II. METHODOLOGY

The main computational tool that we will use in the current investigation is symbolic regression suggested by Koza,[31] part of

the genetic programming ecosystem. The main purpose of symbolic regression is to find symbolically the mathematical relationship between a set of independent variables $x = \{x_1, x_2, x_3, \ldots\}$ and dependent variables $y = \{y_1, y_2, y_3, \ldots\}$. It achieves this goal by searching the function space, in a way similar to biological evolution, by means of mutation and crossover, as exemplified in Fig. 1. As an evolutionary algorithm, symbolic regression is based on a fitness function. Its primary purpose is to minimize the difference between the current values of the dependent variables and the predicted ones. The mathematical functions that will have a better fitness will survive the extreme process of evolution. In most applications that we are aware of,[32,33] symbolic regression learned about the problem with a supervised way of learning because each independent variable in the training set was varied to some extent with a corresponding value for the dependent variable. The philosophy of symbolic regression is different from that of other regression methods such as linear and nonlinear regression. Symbolic regression solves the problem by itself (it is sometimes said that it writes its own computer program) by finding the appropriate function that relates the independent variables to the dependent variables. In contrast, in traditional regression methods, the algorithm is finding the values of the coefficients of a predefined mathematical function that defines the relation between the input and output variables.

In this paper, we explore the potential of using symbolic regression with semisupervised learning. In our case, the independent variables are the Reynolds number Re, angle of attack ϕ, and aspect ratio p_a, while our only dependent variable is the drag coefficient C_D.

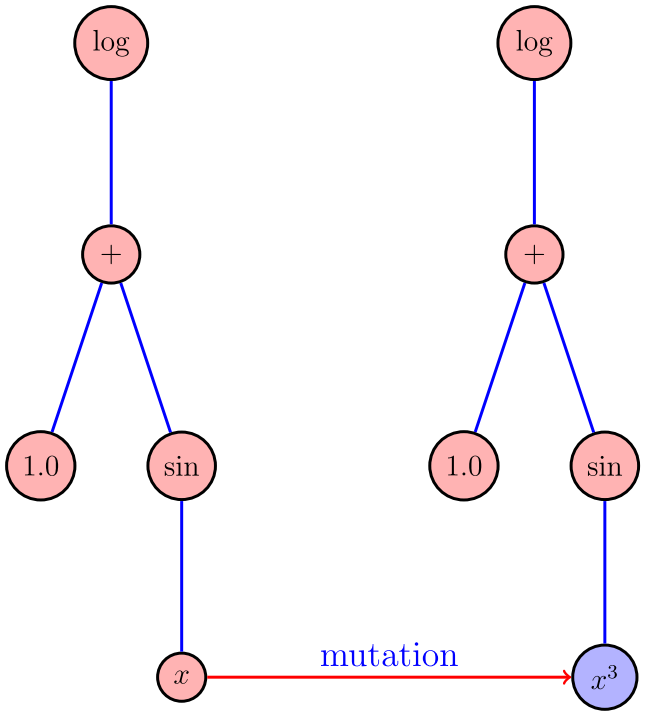

FIG. 1. Example of the mutation process of a mathematical function log(1 + sin(x)) –log(1 + sin(x^3)).

However, in our training dataset, we have only data for a single value of p_a for each geometry of the particle that we consider. This situation resembles the case of semisupervised learning because there are no corresponding values of C_D for multiple for values of p_a. The aspect ratio will act in our case as a latent variable, which the symbolic regression algorithm will learn about in a semisupervised way. We achieve this by providing the symbolic regression algorithm with functional forms that couple p_a to independent variables that *are* varied in the training dataset. In the current investigation, the training data are composed of the results reported by Sanjeevi *et al.*[30] In particular, the data consist of only 57 values of the drag coefficient C_D (for various values of Re and particle orientation) for a single aspect ratio [$p_a = 4.0$ for the spherocylinder (fiber) and $p_a = 2.5$ for the ellipsoid]. We feed the algorithm relations for the dependence on p_a based on (elements of) analytical solutions valid in certain limits of the independent parameters. Those relations will help the algorithm detect a generalized relation of how C_D is varying with all the independent variables, including p_a.

The applicability of the current algorithm depends on two main factors, the accuracy of the dataset that will be used for training and the availability of the functional forms. For example, this algorithm is an excellent candidate to be used for fluid mechanics problems that are associated with the creeping flow and laminar flow regimes. Those flow regimes are rich in terms of analytical solutions and high fidelity data. The predictability accuracy of the algorithm will depend on both the accuracy of the training dataset and the functional forms that are provided. The algorithm follows the following steps:

- Provide mathematical formulations related to the flow problem at hand to the symbolic regression algorithm. Functional forms are available as analytical solutions or information from the literature.
- The symbolic regression algorithm will generate a "soup" of mathematical relations through mutations and interbreeding.
- The mean square error of the generated mathematical forms is tested $\frac{1}{N}\sum_{i=1}^{N}(y - f(x_i))^2$, where N is the number of the data points in the training dataset, y is the dependent variable from the training dataset, and $f(x_i)$ is the predicted function. Also, the complexity fitness of each function will be checked. If the fitness conditions are met, the algorithm will supply the mathematical forms to the user; otherwise, it will go to the second step.

We use Eureqa software[15] as the symbolic regression platform. We use a randomly selected 70% of the dataset for training and the remaining 30% for validation, and we use the square error fitness function plus a fitness function that measures the complexity of the mathematical form. Here and in the following, log denotes the natural algorithm. Finally, we emphasize that the definition that we use for C_D is

$$C_D = \frac{|F_D|}{\frac{1}{2}\rho_f |u_\infty|^2 \frac{\pi}{4} d_{eq}^2}, \tag{1}$$

where F_D is the drag force exerted by the fluid on the particle, ρ_f is the fluid density, u_∞ is the uniform velocity of the fluid far away

from the particle, and d_{eq} is the diameter of the volume equivalent sphere.

III. RESULTS

In Subsections III A and III B, we will present the results for the two different types of particles that we consider in this paper, namely, spherocylinders (fibers) and prolate spheroids. The results will include correlations for the drag as a function of the Reynolds number, the angle of attack, and the aspect ratio of the geometry of the particles. In Subsection III C, we will explore what happens when we interbreed different solutions obtained from different flow solutions.

A. Spherocylinder (fiber) particles

We will give an extensive description of the way we arrived at the final correlation. As stated in the Introduction, our goal is to obtain a drag correlation that takes into account also the effect of particle aspect ratio, using data that were generated for a single aspect ratio. Because of the limited amount of data and more importantly because one of the parameters is not varied in the database, we have to explore the literature for appropriate relations, which will help the algorithm to find the best relation that can fit the data. First, we added an extra column for the value of the aspect ratio ($p_a = 4.0$) to the data given by Sanjeevi *et al.*[30] As a first trial, we will assume the following functional form for C_D:

$$C_D = f(\phi, \log(Re), Re, \sin^2(\phi), p_a). \tag{2}$$

We choose $\log(Re)$ and $\sin^2(\phi)$ in our initial guess for the functional form for the C_D because according to the literature,[20] for Oseen flows and beyond, $\log(Re)$ is an essential ingredient of the functional space that describes C_D and the same applies to $\sin^2(\phi)$.[34,30] However, if we use this form of the function as initial input to the genetic programming algorithm, we get a number of correlations, without the appearance of p_a, in their structure. This result is logical because p_a is not varied in the original data. To better guide the symbolic regression algorithm, we need to guess first about the shape of the functional dependence of C_D on the aspect ratio. A good start is to look at the Stokes flow regime. We will first assume that C_D is inversely proportional to $\log(p_a)$, as indicated by Cox.[35] The new initial functional form is

$$C_D = f\left(\phi, \log(Re), Re, \sin^2(\phi), \frac{1}{Re\,\log(p_a)}\right). \tag{3}$$

After the algorithm reaches a steady state, it proposes several equations, among them is the following:

$$C_D = 2.2 + 4.0 \times 10^{-4} Re + 1.7\,\sin(\phi)^2 + \frac{(35.2 + 9.7\,\sin(\phi)^2)}{(Re\log(p_a))} - 0.3\,\log(Re) - 0.1\,\sin(\phi)^2\,\log(Re). \tag{4}$$

We note that it is still important to use our physical knowledge to distill the appropriate relations because the algorithm provides us with many possible equations in each run, some of which do not agree with known physics in certain limits. For example, one solution that also describes the data well is

$$C_D = 0.2 + 1.7\,\sin(\phi)^2 + 2.0\left(\frac{1.0}{Re\,\log(p_a)}\right)^2 + \frac{(36.4 + 9.7\,\log(Re) + 3.1\pi\,\sin^2(\phi))}{(Re\,\log(p_a))} - 0.1\,\sin^2(\phi)\log(Re). \tag{5}$$

In Eq. (5), as $Re \to 0$, the leading term becomes proportional to Re^{-2}, which violates the linearity of the Stokes flow analytical solution, and thus, this solution is inadmissible. Equation (4) shows that the aspect ratio dependence appears only on the fourth term that represents the Stokes regime. However, this is may be due to the functional dependence that has been selected in Eq. (3), where the aspect ratio only appears in the term $\frac{1}{Re\,\log(p_a)}$. To further investigate if the aspect ratio effect will appear in another term in the fitting function, we conducted a new run with the following initial function:

$$C_D = f\left(\phi, \log(Re), \frac{1}{Re\log(p_a)}, \frac{Re}{\log(p_a)}, \sin(\phi)^2\right). \tag{6}$$

This time we obtained the following equation for the drag coefficient:

$$C_D = 3.7 + \sin(\phi)^2 + 0.07\log(Re)^2 + \frac{(34.3 + 9.9\sin^2(\phi) - 0.47\log(Re))}{(Re\log(p_a))} - \log(Re). \tag{7}$$

In Eq. (7), the dependency on the aspect ratio appears only in the Stokes flow term and is very similar to Eq. (4).

In Sec. II, we showed the steps by which we have obtained a fitting relation for the drag coefficient that takes into account the effect of the aspect ratio from a limited number of data points and from a single aspect ratio. However, to validate the derived dependence on the aspect ratio, we turn to the literature for numerical data for fibers or similar shapes such as cylinders. The only data that we found are for cylindrical particles of aspect ratios 2, 5, 10, and 20 by Vakil and Green.[27] We extracted the drag coefficient data from their Fig. 17 digitally and fed them to Eureqa with the following initial function:

$$C_D = f\left(\log(Re), \frac{1}{Re\,\log(p_a)}, \frac{Re}{\log(p_a)}, \sin^2(\phi), p_a\right). \tag{8}$$

We add a general dependency on p_a to ensure that we capture all the forms of the drag coefficient dependence on the aspect ratio. Nevertheless, the best function that fits the data from Ref. 27 takes the following form:

$$C_D = 1.4 + 7.3\,\sin^2(\phi) + 0.4\,\sin^2(\phi)\log(Re)^2 + \frac{(15.1 - 5.7\,\sin^2(\phi))}{(Re\log(p_a))} - 0.3\log(Re) - 3.4\sin(\phi)^2\log(Re). \tag{9}$$

The form of Eq. (9) has a lot of similarities with Eq. (7), especially with respect to the functional dependence on the aspect ratio, which is identical. The $\frac{1}{\log(p_a)}$ dependence of C_D in both cases shows that we captured correctly the dependence of C_D on p_a in Eq. (7), which

TABLE I. Coefficients for Eq. (10).

Coefficients	Equation (10)
a_0	3.7270
a_1	1.0
a_2	0.070
a_3	34.467
a_4	9.993
a_5	0.470

arose from a single aspect ratio dataset. The difference between Eqs. (7) and (9) could be due to several reasons. Among them is the slight difference in the geometry between the fiber and cylindrical particles and also the range of Re used in both cases.

Until now, we showed the process of deriving a generalized fitting equation from a low volume of data. We will choose Eq. (7) as a general equation that represents the variation of C_D with ϕ, Re, and p_a. All the coefficients that appeared in the previous equations were approximated to the first decimal for displaying purposes. To have a more accurate form of Eq. (7), we will rewrite it in the following form:

$$C_D = a_0 + a_1 \sin^2(\phi) + a_2 \log(Re)^2 + \frac{(a_3 + a_4 \sin(\phi)^2 - a_5 \log(Re))}{(Re \log(p_a))} - \log(Re), \quad (10)$$

where the values of the coefficients are listed in Table I. To prove the validity of Eq. (10), we will test its predictions against the available data from the literature for different Re, ϕ, and p_a.

The first test we did was to compare the results of Eq. (10) with those of Sanjeevi *et al.*,[30] the source of our training data. The comparison is shown in Fig. 2. For low Re, the agreement is quite close. For higher Reynolds numbers, the difference between the prediction of our correlation (10) and that of Ref. 30 is increasing. However, the overall relative error for the whole Re range investigated is about 3%, which is acceptable. For high (within our range of application) Re, the values of C_D get smaller, and thus, any small difference between the data of Ref. 30 and our data can amplify the relative

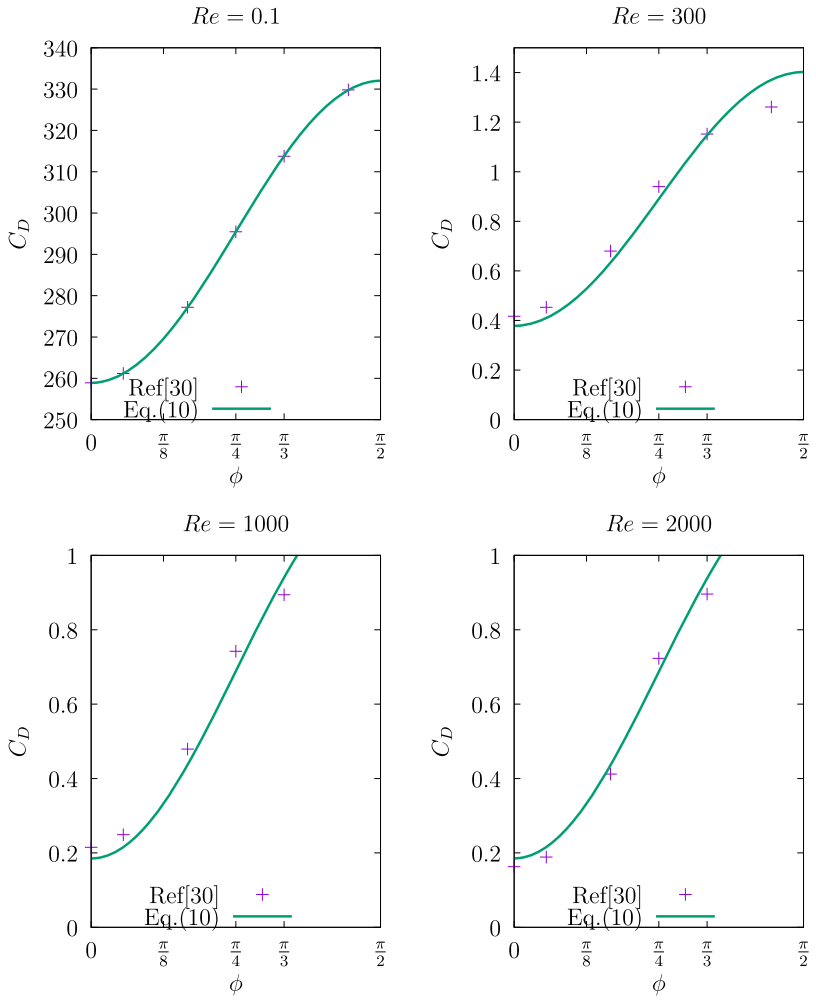

FIG. 2. Comparison between the results of Eq. (10) and those of Sanjeevi *et al.*[30] for different Re, ϕ, and p_a = 4.

error easily. To further evaluate the applicability of Eq. (10), we will compare it with the results of Zastawny *et al.*[29] for their fiber geometry with p_a of 5. In comparison, we also added the results of Ref. 30 for $p_a = 4$ to see if Eq. (10) is biased toward the data that had been used to obtain the correlation. What we see is the opposite: Eq. (10) captures very well the data of Ref. 29 for all *Re* tested, as shown in Fig. 3, and the average relative error between Eq. (10) and the results of Ref. 29 is about 6% for the whole *Re* used, while the relative difference between the data of Refs. 29 and 30 (caused by the difference in aspect ratio) is about 15%. For the case of $Re = 1.0$, the predictions of Eq. (10) and the data of Ref. 29 are close, while those of Ref. 30, corresponding to a lower aspect ratio, are higher than the former two. This interesting observation shows that Eq. (10) captures fairly well the effect of the variation of the aspect ratio. The overall results of Fig. 3 show that the geometry of the fiber plays a significant role mostly in the Stokes flow regime, which agrees with our previous statement. Furthermore, we believe that Eq. (10) applies to higher *Re* than the ones considered in Fig. 3. However, there are no data available for high *Re* for a fiber geometry other than Ref. 30. For a further exploration of the validity of Eq. (10), we will compare its predictions against those of Vakil and Green[27] for cylindrical particles. We selected two aspects ratios 10 and 20, which are 2.5 times and 5 times, respectively, longer than the original particle of Ref. 30, which the training data are based on. By the selection of the current aspect ratios, we have the opportunity to test Eq. (10) far from the original geometry that it is derived from, thus putting a solid ground for its generic form. We believe that the difference in the geometry of the fiber and cylindrical particle will impose some difference in the predicted C_D. Nevertheless, Fig. 4 shows fairly good agreement, especially for the case of $p_a = 10$ where the average relative error is 7%. Even though, for the case of $p_a = 20$, the deviations are slightly higher, especially in the low *Re* regime, the overall relative error is about 15%. These comparisons show that Eq. (10) is approximately applicable to a wide spectrum of *Re*, ϕ, and p_a. We will return to the C_D dependency on the aspect ratio p_a in Sec. IV.

B. Prolate spheroids particles

Prolate spheroids exhibit more complex behavior than fibers when it comes to the way that they interact with the

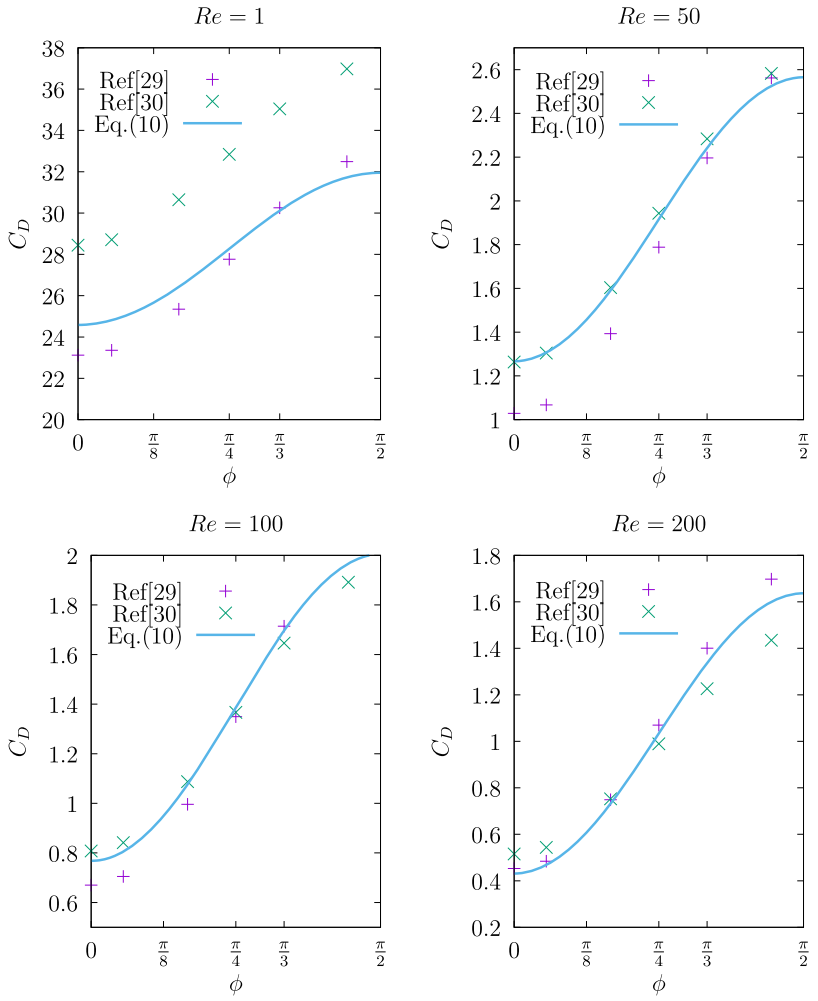

FIG. 3. Comparison between the results of Eq. (10) and those of Zastawny *et al.*,[29] $p_a = 5$, and Sanjeevi *et al.*,[30] $p_a = 4$, for different *Re* and ϕ.

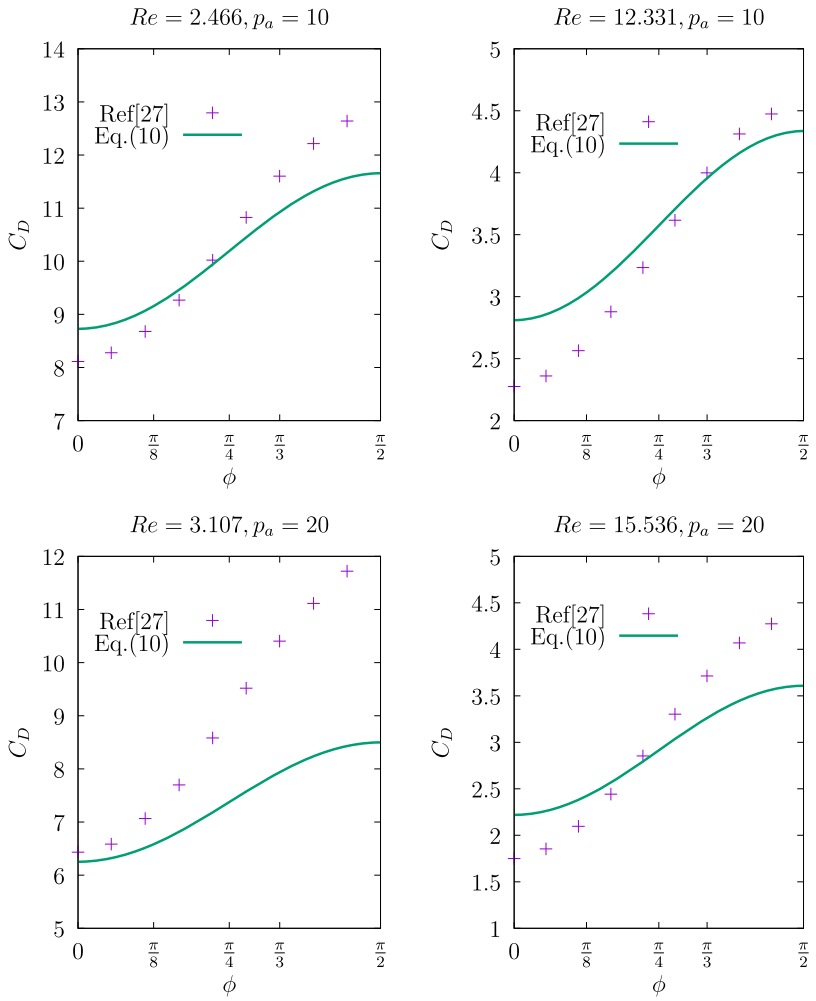

FIG. 4. Comparison between the results of Eq. (10) and those of Vakil and Green[27] for different *Re* and ϕ.

surrounding fluid. Their drag coefficient shows a complex functional dependency with the aspect ratio of their geometry.[34] We will first impose the C_D formula for the case of $\phi = 0$ derived analytically by Ref. 23 and for flow conditions that correspond to Oseen flow, which has the following form:

$$C_{D,\phi=0} = \frac{2\pi B}{Re(1-e_1^2)}\left(1 + \frac{BRe}{24} + \frac{B^2Re^2\log Re}{360}\right), \tag{11a}$$

$$e_1 = \sqrt{\left(1 - \frac{b^2}{a^2}\right)}, \tag{11b}$$

$$B = 24e_1^3\left((1+e_1^2)\log\frac{1+e_1}{1-e_1} - 2e_1\right)^{-1}. \tag{11c}$$

Here, e_1 is the eccentricity of the geometry of the particle, a is the semimajor axis, b is the semiminor axis, and B is a constant that depends on eccentricity. The shape of our first initial function is

$$C_D = f\left(\phi, \left(\frac{1}{Re}\right)\left(1 + \frac{BRe}{24} + \frac{B^2Re^2\log Re}{360}\right), Re, \log Re, \sin(\phi)^2\right), \tag{12}$$

where we used $B = 4.698$ for the dataset of Ref. 30. After feeding this dataset to the symbolic algorithm, we obtained several relations, but most of them were nonphysical because of the appearance of the Re^2 term in the denominator, so we excluded them. The one with higher accuracy and with physical significance is

$$C_D = 0.2 + 1.3e^{-0.01Re} + 0.7\sin(\phi)^2 + \frac{24.2 + 4.7\sin(\phi)^2}{Re} - 9.3\times10^{-5}Re. \tag{13}$$

The interesting observation from Eq. (13) is the absence of any dependence of C_D on the geometrical parameter B. It shows that the signature of geometrical parameters from Eq. (11) is very weak, and this is why it cannot be detected by the genetic algorithm. This weakness of detection may be a result of the fact that Eq. (11) is valid only for a single angle of attack ($\phi = 0.0$), while the dependency for other

angles of attack is not known. Another interesting observation from Eq. (13) is that the genetic algorithm will skip the initial functions if it finds that they are not relevant to the training data. This shows that the algorithm is not biased to the mathematical formulas that are given as an initial guess but only to the ones that are relevant to the training data.

After our first failed attempt to find the aspect ratio dependency of C_D, we turned to the Stokes solution for flow over prolate particles,[34] which has the following form:

$$C_{D,\phi=0} = \frac{24k_{a0}}{Re}, \tag{14a}$$

$$k_{a0} = \frac{8}{3}p_a^{-\frac{1}{3}}\left(\frac{-2p_a}{p_a+1} + \frac{2p_a^2-1}{p_a^2-1}\log\frac{p_a+\sqrt{p_a^2-1}}{\sqrt{p_a^2-1}}\right)^{-1}, \tag{14b}$$

$$C_{D,\phi=90} = \frac{24k_{a90}}{Re}, \tag{14c}$$

$$k_{a90} = \frac{8}{3}p_a^{-\frac{1}{3}}\left(\frac{p_a}{p_a^2-1} + \frac{2p_a^2-3}{p_a^2-1}\log(p_a+\sqrt{p_a^2-1})\right)^{-1}, \tag{14d}$$

$$C_D = C_{D,\phi=0} + (C_{D,\phi=90} - C_{D,\phi=0})\sin^2\phi. \tag{14e}$$

Based on this, we chose the following initial function:

$$C_D = f\Bigg(\phi, p_aRe, p_a\log(Re), \sin^2(\phi), \times\frac{(k_{a0}+(k_{a90}-k_{a0})\sin(\phi)^2)}{Re}, Re, \log(Re), \log(Re)^2\Bigg), \tag{15}$$

where k_{a0} and k_{a90} are the functions given by Eqs. (14b) and (14d), respectively. For the dataset of Ref. 30, there is only a single aspect ratio $p_a = 2.5$ for which $k_{a0} = 0.9615$ and $k_{a90} = 1.146$. By this way, we feed Eureqa the analytical Stokes solution for a single prolate particle and help the genetic algorithm to find accurate relations beyond its training data. We obtained the following relation for C_D:

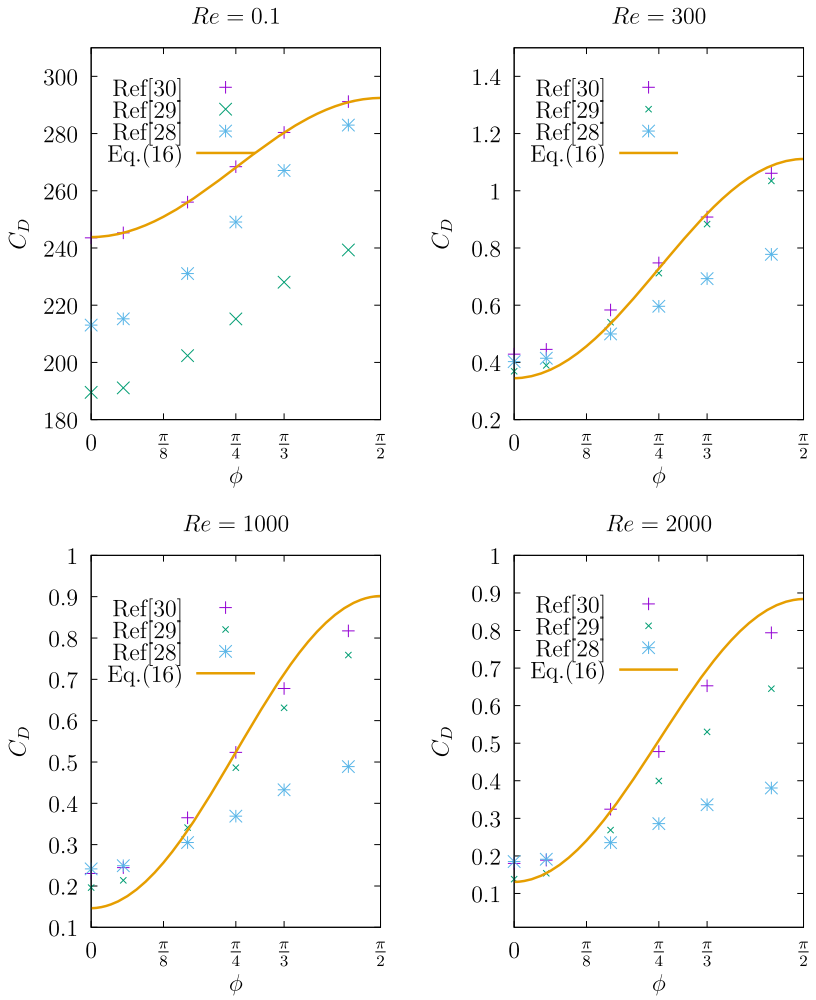

FIG. 5. Comparison between the results of Eq. (16) and those of Zastawny *et al.*,[29] Sanjeevi *et al.*,[30] and Ouchene *et al.*[28] for different Re, ϕ, and $p_a = 2.5$.

$$C_D = a_1 + a_2 \sin(\phi)^2 + a_3 \log(Re)^2 + \frac{\sum_{i=1}^{i=2} A_i}{Re} + a_7 \log(Re), \quad (16a)$$

$$A_1 = a_4(k_{a0} + (k_{a90} - k_{a0}) \sin(\phi)^2), \quad (16b)$$

$$A_2 = a_5 + a_6 \log(Re)^2. \quad (16c)$$

We test the validity of Eq. (16) by comparing its results with those of Refs. 30, 28, and 29, as shown in Fig. 5. For the case of $Re = 0.1$, Eq. (16) matches the results of Ref. 30, and the C_D values from Ref. 28 are in close proximity with our results, while those of Ref. 29 are significantly lower. This discrepancy in the results may be attributed to the grid size used by Zastawny *et al.*,[29] as discussed in Ref. 30. However, as Re is increasing, the C_D values of Ref. 29 are getting closer to the predictions by Eq. (16) and to those of Ref. 30, while those of Ref. 28 are lower than the rest, especially at high values of ϕ. Interestingly, the C_D values of Ref. 29 for $Re = 1000$ and 2000, even though they are coming from a correlation for Re up to 300, are in close agreement with our results and those of Ref. 30.

Prolate spheroids give a rare opportunity to explore the validity of a generalized C_D formula because of the existence of many test cases in the literature. The real test of Eq. (16) is its capability to predict the results of Ref. 28 for different aspect ratios, Re, and ϕ. We decided to choose two aspects ratios, 10 and 15, both of which are far larger than the aspect ratio of 2.5 of the initial training dataset, and two Reynolds numbers, 1 and 100, which cover flows with strong viscous effects and flows dominated by inertial fluid forces. The variation of C_D is shown in Fig. 6. Equation (16) predicts quite closely the results for the case of $Re = 1$ for both p_a selected. As Re increases to 100, our results start diverting from those of Ref. 28. However, the overall average relative error for the selected Re regime is 13.7% and 15.7% for $p_a = 10$ and 15, respectively, which is within an acceptable range, given the very large difference with the p_a of the training set. We believe that the good agreement of Eq. (16) with the data of Ref. 28 for low Re is mainly due to the initial function that we propose in Eq. (15), which contains the Stokes flow analytical solution for flow over a prolate spheroid. Thus, the symbolic regression algorithm had sufficient ingredients to uncover the general solution around and beyond the Stokes regime. We can explain the divergence of our data from those of Ref. 28 by the two following

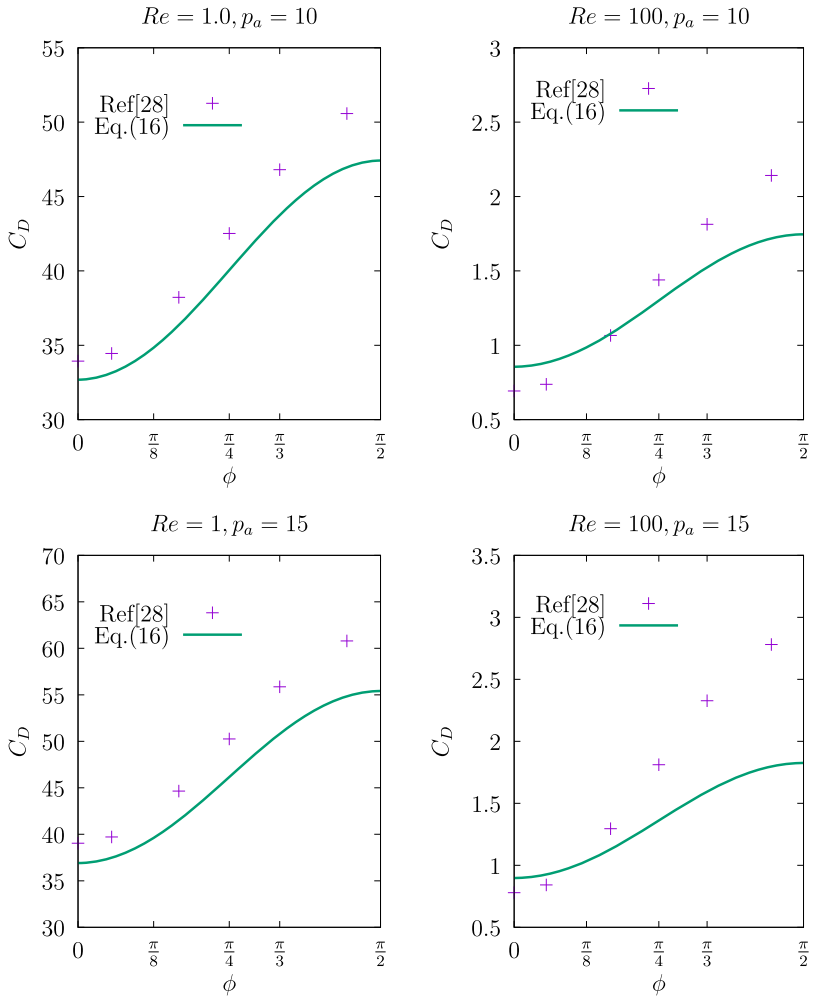

FIG. 6. Comparison between the results of Eq. (16) and those of Ouchene *et al.*[28] for different Re, ϕ, and $p_a = 10$ and 15.

reasons. First, Eq. (16) loses its accuracy as it departs from its original geometry. Second, the data of Ref. 30, used to obtain Eq. (16), were obtained from a different numerical scheme than that of Ref. 28. We believe that both reasons are essential, and we will explore them both, starting with the second one, while the first one will be explored in Sec. IV.

We believe that the second reason may play a significant role in the deviation. As shown in Fig. 5, the C_D values of Ref. 28 are deviating from our results and those of Ref. 30, especially as Re is increased. This is reflected in the results of Eq. (16) in Fig. 6. We believe that Eq. (16) has inherited the characteristics of C_D predicted by the numerical scheme used in Ref. 30, which is different from that used by Ref. 28. What we mean by the solver characteristics is how much error is propagated into the numerical scheme by a finite grid size and time step, for example. We come to this conclusion because Eq. (16) and the results of Ref. 28 have the same overall trend in which the C_D is increasing as the value of the p_a is increased. The last test for Eq. (16) is the examination of its behavior for the case of p_a = 1.25. This ellipsoid has many features similar to the spherical geometry and can be considered as another extreme case for an ellipsoidal geometry. It also gives us a rare opportunity to compare our results with those obtained from two different numerical schemes. The comparison of the results of Eq. (16) and those of Refs. 29 and 28 plus the Stokes flow analytical solution Eq. (14) is shown in Fig. 7. For the cases of Re = 0.1 and 1, which are adjacent to the Stokes flow regime, we make two observations. The first observation is that for Eq. (16) and the results from Ref. 28, the values of C_D are higher than the Stokes flow solution. The other observation is that for the results from Ref. 29, the C_D values are lower than those of the Stokes flow solution for almost the whole range of angles of attack.[29] We believe that for Re = 0.1 and 1.0, Eq. (16) and Ref. 28 are the two solutions that capture the variation of C_D most accurately because they produce values of C_D that are higher than the Stokes flow solution [Eq. (14)]. This is in agreement with the available theory, which states that for Oseen flows and beyond, the C_D is higher than for Stokes flow. The C_D predictions of Ref. 29 are significantly diverting from those of Eq. (14), so we conclude with a strong certainty that they do not represent physical reality. This shows that Eq. (16) significantly outperforms a direct numerical scheme, which was designed to solve the problem of flow over an ellipsoid

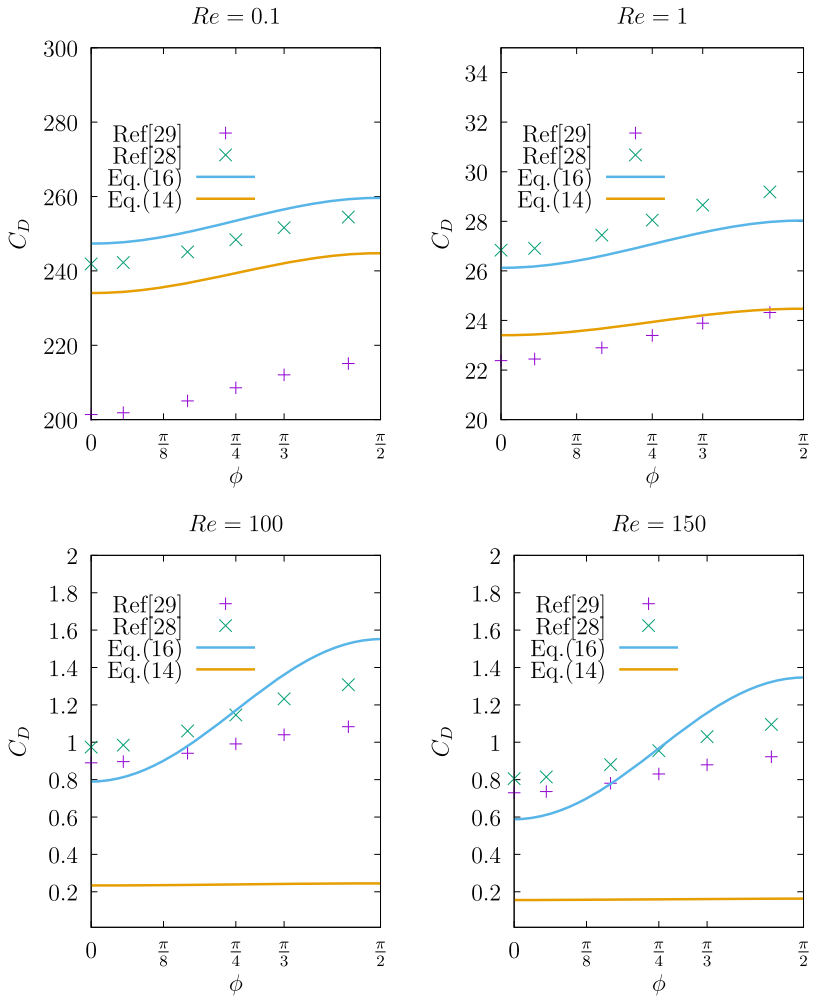

FIG. 7. Comparison between the results of Eq. (16) and those of Ouchene *et al.*,[28] Zastawny *et al.*,[29] and Stokes flow solution [Eq. (14)] for different Re, ϕ, and p_a = 1.25.

particle. This outperformance is not only with respect to accurate predictions of C_D but also with respect to the computational time and resources used to obtain the predictions. For our predictions, we just used a personal laptop computer and less than an hour of computational time. For simulations similar to those of Ref. 29, usually use a powerful workstation or a computer cluster with several cores must be used for several hours.

As we are moving toward the fluid inertial regime (higher Re), the results of the three numerical schemes are approaching each other, especially at low angles of attack. There is a divergence between our results and those of Refs. 28 and 29, especially at higher angles of attack. The average relative error for the whole Re regime between our results and those of Ref. 28 is 8.9%, while it is 21.5% compared to Ref. 29. We believe that this divergence is the result of the different numerical schemes used in the investigations, and it is difficult to assess which numerical scheme is more accurate. Nevertheless, our results are closely resembling those of Ref. 28. In Subsection III C, we will introduce a new method to interbreed different flow solutions together. This will help explore if the different flow numerical schemes are the reason for the deviation between the data of Refs. 28 and 29 and the results of our Eq. (16).

C. Discrete interbreeding of different flow solutions

We will first make the assumption that the characteristics of the numerical schemes are stored in the mathematical formulations that represent their results, such as mathematical correlations. Discrete interbreeding of different flow solutions (DIDFS) is the process of selecting specific mathematical terms (we call them genes or elements) from a specific numerical scheme solution and inserting them as initial functions in the symbolic regression algorithm to drive a regression equation for a different numerical scheme. We want to emphasize that DIDFS is different from the crossover process that was mentioned before, since the latter is part of the internal structure of the symbolic regression algorithm. We will use the interbreeding method only for ellipsoidal particles because of the existence of more information on how C_D varies with p_a in the literature compared to the spherocylinder case. Interbreeding will help to inherit some numerical scheme characteristics to different numerical schemes, without introducing any noise to the data that will be used for training of the symbolic regression algorithm. We want to point out that we will use the same data that have been utilized for deriving Eq. (16) to obtain a new ecosystem of equations for C_D.

In our first test of interbreeding, we will try to merge some of the flow numerical scheme properties of Refs. 28 and 30 by supplying mathematical functions that describe the behavior of the former numerical scheme as initial functions for the symbolic regression analysis of the latter numerical scheme. We start by creating a database of values of C_D from the correlation provided by Ref. 28 for Re ranging between 0.2 and 220 and p_a between 2 and 32. After this, the data will be used for regression analysis with an initial function of the following form:

$$C_D = f\left(\frac{1}{Re}, Re, \log(Re), p_a, p_a \log(Re), \log(Re)^2, \sqrt{p_a}, \sin(\phi)^2\right). \tag{17}$$

The symbolic regression algorithm obtains the following formula for C_D:

$$C_D = 0.489 + 0.137\,\sin(\phi)^2 + \frac{(3.296 + 6.28\sqrt{p_a} + 15.723\,\sin(\phi)^2 + 0.6160 p_a\,\sin(\phi)^2 \log(Re)^2 - 0.901 p_a \sin(\phi)^2)}{Re}. \tag{18}$$

Now we will inject some of the numerical scheme characteristics of Ref. 28 into the numerical scheme of Ref. 30. The interbreeding process will be done by selecting the mathematical functions $(6.28\sqrt{p_a}, (0.616 p_a - 0.901 p_a \log(Re)^2))$ (gene 1), $6.28\sqrt{p_a}$ (gene 2), $0.616 p_a - 0.901\,\log(Re)^2$(gene 3), $-0.901 p_a$ (gene 4), and $(\sqrt{p_a}, (p_a + p_a \log(Re)^2)$ (gene 5), which are part of Eq. (18) as initial functions for the symbolic regression algorithm that will use the dataset of Ref. 30 [i.e., similar to Eq. (16)] to search for a new functional form of C_D. We selected those different parts of Eq. (18) because they represent how the C_D varies with the aspect ratio. Those mathematical formulas will act as genes that will carry that information. We used a very generic way of selecting the genes, based on each functional element in the addition that forms Eq. (18), including their coefficients (except for gene 5, for reasons that we will discuss below). This is the most straightforward path that someone can take if the flow problem in hand is complex. However, a more in-depth investigation is needed to find the optimal way of selecting genes, which will maximize the amount of physics that will be learned by the symbolic regression algorithm and minimize the time needed to learn those physics. The reason that we divided the functional form dependence of Eq. (18) on p_a in different genes is because we wanted to explore their individual effect on the learning process of the symbolic regression algorithm. For gene 5, we took only the functional form of the aspect ratio dependence and we replaced all its coefficients by 1. In this way, we wanted to test the predictive ability of DIDFS for the case that we only supply the functional form. The symbolic regression algorithm will search the functional space for similar functions using the data of Ref. 30 [i.e., similar to Eq. (16)] for training. The mathematical formula of C_D that will contain similar functions as those that represent the genes of Ref. 28 will be selected. We will first inject gene 1, which is composed of the complete mathematical formulation that represents the variation of C_D with p_a, and the initial function is

$$C_D = f\left(\phi, Re, \frac{6.28\sqrt{p_a}}{Re}, \frac{(0.616 p_a \log(Re)^2 - 0.901 p_a)}{Re}, Re, \log(Re), p_a, \sin(\phi)^2\right). \tag{19}$$

We obtained the following equation for C_D:

$$C_D = a_1 + a_2 Re + a_3 \sin(\phi)^2 + \frac{(a_4\sqrt{p_a} + a_5 \sin(\phi)^2\sqrt{p_a})}{Re} + a_6 \log(Re)^2. \quad (20)$$

The coefficients of Eq. (20) are listed in Table II. In the following steps, we will inject genes 2, 3, and 4 individually. For the case of gene 2, the initial function is

$$C_D = f\left(\phi, Re, \frac{6.28\sqrt{p_a}}{Re}, Re, \log(Re), p_a, \sin(\phi)^2\right). \quad (21)$$

We obtained the following equation for the C_D:

$$C_D = a_1 + a_2 Re + a_3 \sin(\phi)^2 + \frac{(a_4\sqrt{p_a} + a_5 \sin(\phi)^2\sqrt{p_a})}{Re} + a_6 \log(Re). \quad (22)$$

For the case of gene 3, the initial function is

$$C_D = f\left(\phi, Re, \frac{0.616 p_a \log(Re)^2}{Re}, Re, \log(Re), p_a, \sin(\phi)^2\right). \quad (23)$$

The equation for the C_D is

$$C_D = a_1 + a_2 \sin(\phi)^2 + a_3 \log(Re)^2 + \frac{A_1}{Re} + B_1, \quad (24a)$$

$$A_1 = a_4 + a_5 \sin(\phi)^2 + a_6 p_a \log(Re)^2 \log(Re), \quad (24b)$$

$$B_1 = a_7 \log(Re) + a_8 \sin(\phi)^2 \log(Re). \quad (24c)$$

For case of gene 4, the initial function is

$$C_D = f\left(\phi, Re, \frac{-0.901 p_a}{Re}, Re, \log(Re), p_a, \sin(\phi)^2\right). \quad (25)$$

The formula for C_D is

$$C_D = a_1 + a_2 Re + a_3 \sin(\phi)^2 + \frac{(a_4 p_a + a_5 p_a \sin(\phi)^2)}{Re} + a_6 \log(Re)^2. \quad (26)$$

Finally, the initial function for the case of gene 5 is

TABLE II. Coefficients for Eqs. (16) and (20).

Coefficients	Equation (16)	Equation (20)
a_1	3.151	1.559
a_2	0.750	0.0005
a_3	0.0579	0.711
a_4	25.873	15.334
a_5	−2.258	3.034
a_6	0.2304	−0.044
a_7	−0.840	...

TABLE III. Coefficients for Eqs. (22) and (24).

Coefficients	Equation (22)	Equation (24)
a_1	1.832	3.541
a_2	0.0001	1.0
a_3	0.748	0.059
a_4	15.264	22.11
a_5	3.033	4.752
a_6	−0.261	−0.055
a_7	...	−0.889
a_8	...	0.051

$$C_D = f\left(\phi, Re, \frac{\sqrt{p_a}}{Re}, \frac{(p_a + p_a \log(Re)^2)}{Re}, Re, \log(Re), p_a, \sin(\phi)^2\right). \quad (27)$$

We can obtain the following equation for C_D:

$$C_D = a_1 \sin(\phi)^2 + a_2\sqrt{\sqrt{\frac{p_a}{Re}}} + \frac{(a_3 p_a + a_4\sqrt{p_a} + a_5 p_a \log(Re)^2 + a_6 \sin(\phi)^2\sqrt{p_a})}{Re}. \quad (28)$$

The values of the constants for Eqs. (22), (24), (26), and (28) are listed in Tables III and IV. Before proceeding to further analyze the newly derived ecosystem of equations for C_D, we will first compare their results with the data of Ref. 30 (training data) and Ref. 29, for the case of $p_a = 2.5$, and different Re, as shown in Fig. 8. For the case of $Re = 0.1$ and 300, the whole group of C_D equations from different interbreeding genes perfectly follow the data of Ref. 30, except for the case of Eq. (22), which at $Re = 300$ slightly overpredicts the results of Ref. 30. Moving to flows where the inertia of the fluid is dominant ($Re = 1000$), we observe that the different fitting equations for C_D are segregated into two groups, the first one follows accurately [Eqs. (22), (24), and (28)] the data of Ref. 30, while the second [Eqs. (20) and (26)] underpredicts the values of Ref. 30, especially at low values of ϕ, and closely follows those of Ref. 29 for higher values of ϕ. However, when Re is increased further to 2000, all the fitting equations for the C_D from the discrete interbreeding process are following in a close manner the results of Ref. 30. We believe the predictive discrepancies that we observe from Eqs. (20) and (26) could be due to the choice of the fitness function. It seems that the square error function did not converge appropriately in some cases when it has to deal with small values. Overall, the new ecosystem of

TABLE IV. Coefficients for Eqs. (26) and (28).

Coefficients	Equation (26)	Equation (28)
a_1	1.565	0.713
a_2	0.0005	4.154
a_3	0.734	0.161
a_4	9.698	12.762
a_5	1.919	0.161
a_6	−0.044	3.0341

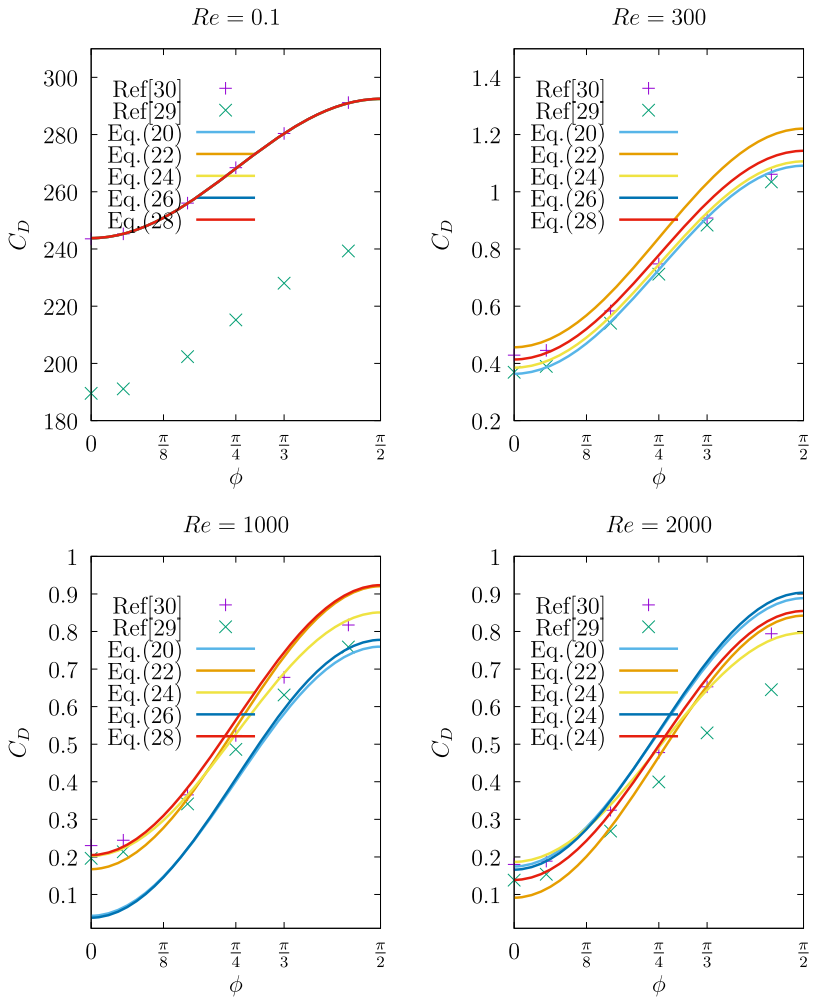

FIG. 8. Comparison between the results of different predictors of C_D, Zastawny *et al.*[29] and Sanjeevi,[30] for different Re, ϕ, and p_a = 2.5.

equations for C_D predicts accurately the data that are used for their training.

All the recent derived expressions for C_D from the interbreeding process are carrying elements of the numerical schemes used to obtain the data of Ref. 30 (Lattice Boltzmann) and Ref. 28 (finite volume), and also they do not contain any elements of the Stokes flow solution. We will first investigate the learning behavior of the new ecosystem of C_D equations for the case of p_a = 10 and different Re values. For Re = 0.1, where the viscous forces are dominant, the predicted correlations for C_D are grouped in three groups. The first group, which only includes Eq. (24), is in close proximity with the values of Ref. 28, as shown in Fig. 9. Moving to the remaining two groups of equations, we see that the predictions of Eqs. (20), (22), and (28) are close to each other, and they overpredict the results of Ref. 28 by an average of 40%, while the last group that only includes Eq. (26) overpredicts the results of Ref. 28 by an average of 160%. For the case of Re = 1.0, the trend is quite similar to that of Re = 0.1, except that the predictions of Eqs. (20), (22), and (28) are getting closer to those of Ref. 28. As we move to the inertial flow regime, Re = 100 and 150, we see that the C_D correlations that were significantly overpredicting the values of Ref. 28 are now significantly closing the gap, and Eq. (28), especially at high angles of attack. At the same time, the predictions of Eq. (24) are drifting away from those of Ref. 28. The mosaic of the different predictive behaviors that we see in Fig. 9 shows the importance of the interbreeding genes in the process of learning and that different genes enhance the learning process at different flow regimes. The dependency of Eq. (18) on p_a is divided into different genes in such a way that each gene will carry different amounts of information about the p_a dependency. We can see that role of gene 3 was essential for Eq. (24) to learn about the low Re regime, which is difficult because in this regime C_D varies significantly with the aspect ratio. This difficulty is reflected in the significant overprediction of the other set of equations. Surprisingly, Eq. (20), which has been derived from gene 1, containing all information of the p_a dependency, could not achieve reasonable accuracy in this regime. This shows that using a gene that contains the complete mathematical information about a latent variable such as p_a in our case does not guarantee that the symbolic regression algorithm will learn effectively about this variable in the whole spectrum of mathematical and physical regimes the training data may contain. Equations (20) and (22) have similar predictive behavior, which shows that equations that are

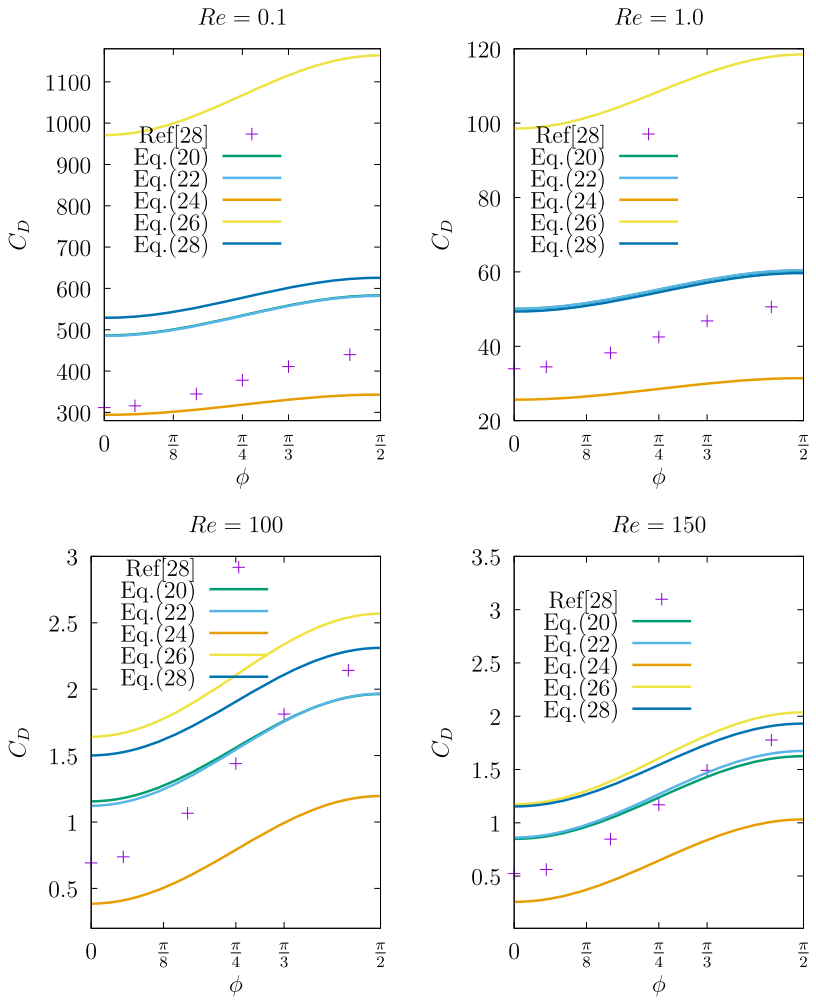

FIG. 9. Comparison between the results of different predictors of C_D and those of Ouchene *et al.*[28] for different *Re*, ϕ, and p_a = 10.

derived from genes that contain partial information about the latent variable can have similar performance with those that are derived using genes that contain the whole information.

In order to check the observations that we made from the previous test case, we will test the same set of equations for the case of p_a = 1.25 and different *Re*, as shown in Fig. 10. As we expected, Eq. (24) captures accurately the behavior of Ref. 28 for the low *Re* regime. However, the trio of equations [Eqs. (20), (22), and (28)] follows quite closely the approximate independence of ϕ observed in Ref. 29 at *Re* = 0.1 and 1.0. This is surprising because the trio of equations did not get any training by the data of Ref. 29 or got any genes neither from their numerical schemes nor from the Stokes solution. This agreement shows that the trio of the mentioned equations and the drag resulting from the numerical schemes of Ref. 29 share some common genes. However, there is a question that may be raised: what if the proposed genes are just over-fits? The answer to the this question can be inferred from the variation of the set of C_D equations derived from DIDFS with p_a. The general trend of Eq. (14), and that of Ref. 28, is that C_D is increasing as p_a is increasing in the low and high *Re* regimes. The same behavior has been captured by the new ecosystem of C_D equations. The interesting thing about the equations of C_D derived from DIDFS is that they only learned about the effect of the aspect ratio through the genes (mathematical formulations) that are given as initial functions to the symbolic regression algorithm. This demonstrates that the genes that were selected encode in their mathematical structure the evolution of real physical phenomena and cannot be considered as overfits. As for the inertial flow regime, we observe that the predictions of most of the equations are quite similar, except that of Eq. (28) whose prediction is close to that of Ref. 28 at very high angles of attack.

The structure of Eq. (20) carries elements of the numerical schemes used to obtain the data of Ref. 30 (lattice Boltzmann) and Ref. 28 (finite volume). In addition, it is not carrying any elements or genes (used here as a metaphor) from the Stokes flow analytical solution. We have to mention that neither the data of Ref. 28 nor the data of Ref. 30 contain any elements of a pure Stokes flow solution since the starting *Re* for both cases is 0.1. Our next quest is to obtain a regression equation for C_D that contains elements of the solutions of numerical schemes from Refs. 30 and 28, including the elements of the Stokes flow solution. We first impose the following initial function:

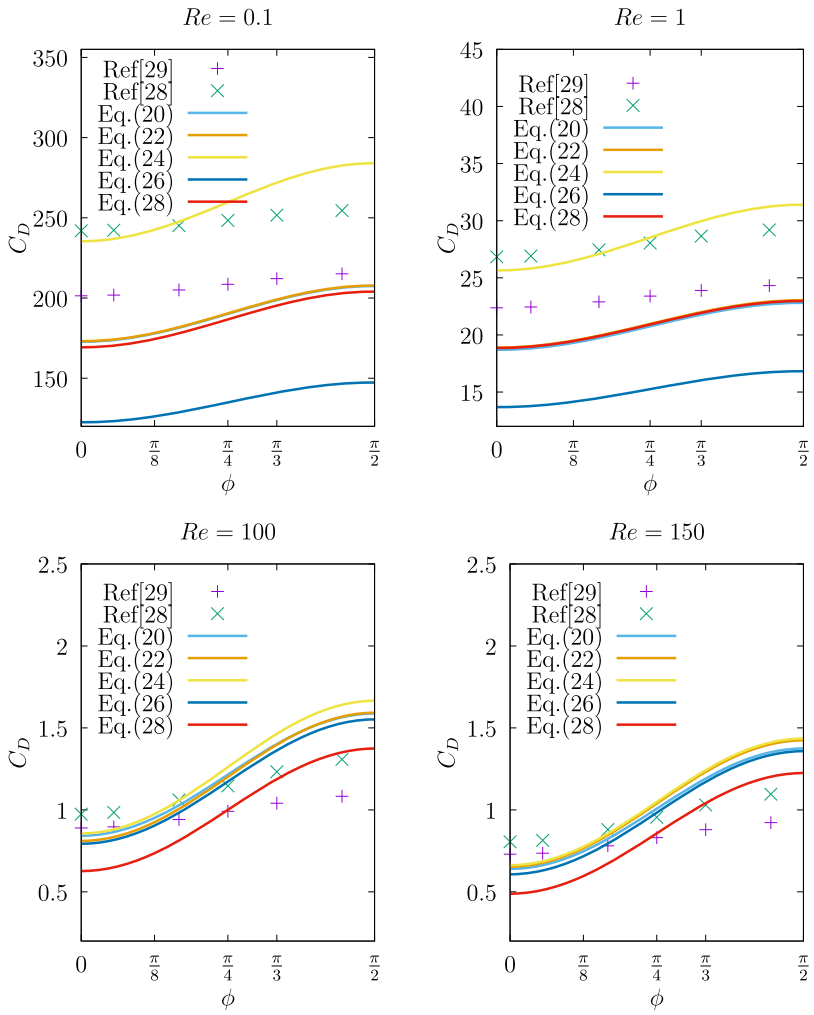

FIG. 10. Comparison between the results of different predictors of C_D and those of Ouchene *et al.*[28] and Zastawny *et al.*[29] for different *Re*, ϕ, and p_a = 1.25.

$$C_D = f\left(\phi, Re, \frac{6.28\sqrt{p_a}}{Re}, \frac{(0.616p_a - 0.91p_a\log(Re)^2)}{Re}, Re, \log(Re), p_a, \sin(\phi)^2, \frac{k_{a0} + (k_{a90} - k_{a0})\sin^2\phi}{Re}, k_{a0}, ka_{90}\right). \quad (29)$$

It was challenging to obtain a regression equation for C_D that contains the trio of the elements that we are looking for, and we obtained only one equation that satisfied the condition that we set, which is

$$C_D = a_1 + a_2\sin(\phi)^2 + \frac{\sum_{i=1}^{i=2} A_i}{Re} + a_5\log(Re), \quad (30a)$$

$$A_1 = a_3k_{a0} + a_3k_{a90}\sin(\phi)^2 + a_4\sqrt{pa}, \quad (30b)$$

$$A_2 = a_3\, k_{a90}\sin(\phi)^2. \quad (30c)$$

The coefficients of Eq. (30) are listed in Table V. Equation (30) shows that the gene numerical scheme of Ref. 28, namely, $\frac{6.28\sqrt{p_a}}{Re}$ and $\frac{(0.61p_a - 0.91p_a\log(Re)^2)}{Re}$, played a minor role in its fabric. On the contrary, the genes of the Stokes flow solution $\frac{k_{a0}+(k_{a90}-k_{a0})\sin^2\phi}{Re}$ had an overwhelming influence on the final C_D formulation. This may explain the troublesome finding of an appropriate expression for C_D. It seems that the genes for the analytical Stokes solution are in

TABLE V. Coefficients for Eq. (30).

Coefficients	Equation (30)
a_1	1.739
a_2	0.740
a_3	25.929
a_4	−0.493
a_5	−0.231

conflict with the genes of the numerical scheme.[28] We observed from the previous results that different equations can predict quite well the C_{D0} and C_{D90}. We will use this observation and we will also take advantage of the $\sin(\phi)^2$ dependency of C_D[36] for prolate spheroids to construct a new hybrid equation for C_D from Eqs. (16) and (28). In this newly constructed equation, the C_{D0} component comes from Eq. (16) and the C_{D90} component from Eq. (28) (since the latter gives good predictions at high ϕ),

$$C_D = C_{D0_{Eq.(16)}} + \left(C_{D90_{Eq.(28)}} - C_{D0_{Eq.(16)}}\right)\sin(\phi)^2. \quad (31)$$

For now on, we will focus on the high Re regime since in this regime we see the most significant deviation between our equations that we develop for C_D and those of Ref. 28. We selected two p_a values, 8 and 15, to test the predictive capabilities of Eqs. (30) and (31). For the case of $p_a = 8$, we see that Eq. (16) is predicting quite accurately the results of Ref. 28 for both *Re*, as shown in Fig. 11. We see also that Eqs. (20) and (30) predict with great precision the values of C_D at high angles of attack for both *Re* used. This shows that using discrete interbreeding with the Stokes solution leads to better predictions at high angles of attack compared to Eq. (16). However, Eqs. (20) and (30) overpredict the values of C_D at low values of ϕ for reasons that we cannot explain, while Eq. (31) slightly overpredicts the values of C_D at very high values of ϕ. If we increase the aspect ratio to 15, we see that the equations that result from DIDFS have better predictions for high angles of attack compared to those of Eq. (16). However, their predictions are less accurate compared to the case of $p_a = 8$. On the other hand, the C_D values from Eq. (31) almost match the values of C_D from Ouchene *et al.*,[28] for both *Re* considered, which shows that Eq. (31) can be used as a good predictor equation for C_D for aspect ratios ranging from 10 to 15 and for the whole range of *Re* considered. However, Eq. (31) loses its accuracy as p_a increases. We report errors for the proposed models for the following two cases of $Re = 200$: for $p_a = 8$ and 15. Furthermore, errors are reported with respect to the range of ϕ sampled. Table VI shows that Eq. (31) has the highest accuracy. However, surprisingly, Eq. (16), which is a result of a single aspect ratio, performs better than Eqs. (20) and (30), which carries in its structure genes from the C_D variation of Ref. 28. Generally, DIDFS helped to enhance the predictive capacity at certain regimes of the problem. This explains the predictive success of Eq. (31) as a mix between Eqs. (16) and (28), which is a product of the DIDFS process. Based on the work discussed, we find that the

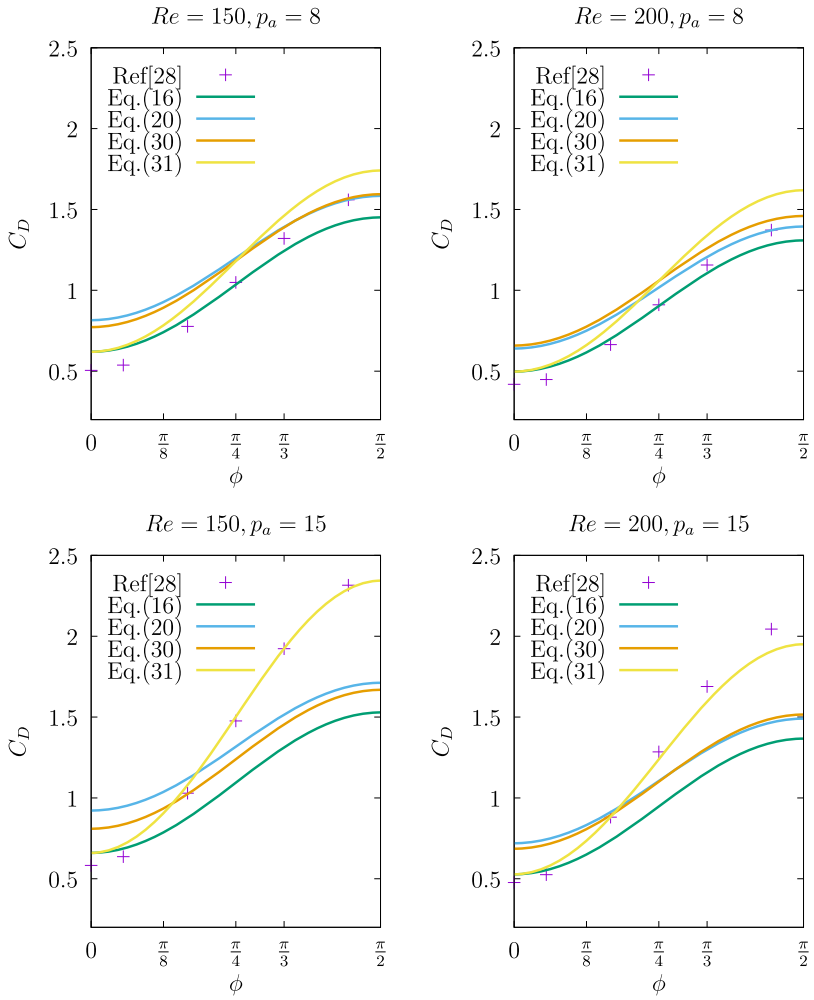

FIG. 11. Comparison between the results of different predictors of C_D and those of Ouchene *et al.*[28] for different *Re*, ϕ, and p_a = 8 and 15.

TABLE VI. The relative error between different C_D predictor equations and the data of Ouchene *et al.*[28] for prolate spheroids at $Re = 200$.

	$p_a = 8$ (%)	$p_a = 15$ (%)
Equation (16)	8	22
Equation (20)	20	27
Equation (30)	19	21
Equation (31)	16	6

previous knowledge in the form of genes accelerate the convergence of the symbolic regression algorithm. This observation is consistent with that of Schmidt and Lipson,[37] who recommended to use previous knowledge to accelerate the convergence of symbolic regression algorithms. Similarly, Loiseau and Brunton[38] concluded that for the nonlinear dynamics sparse identification to give physical results, it also has to be supplied with some physical insights.

In summary, we carefully examined the effect of the characteristics of the numerical scheme on the predictive performance of C_D correlations. We concluded that the numerical scheme characteristics played an essential role to enhance the predictive behavior of C_D predictors. However, we still do not have a correlation that can predict with great accuracy for the whole spectrum of Re and p_a the results of Ref. 28. We still believe that Eqs. (10) and (16) represent most of the physics related to fluid flow over spherocylindrical and prolate spheroids, respectively. However, there may be some missing physics that we could not discover due to the extreme sparsity of the data that we used. In Sec. IV, we will try to explore the physics that those equations describe and try to improve their predictive capability.

IV. THE DISCOVERY OF NEW PHYSICS

Up to this point, most of this paper was dedicated to obtaining predictive equations for C_D without paying significant attention to the physics that may evolve out of them. In this section, we will try to explore the physics that those equations are implying.

Equations (10) and (16) are the first correlations in the literature that show that C_D for nonspherical particles is a function of different $\log(Re)$ powers and $\frac{1}{Re}$. Another interesting observation is that the C_D for both shapes consists of similar functions for the terms

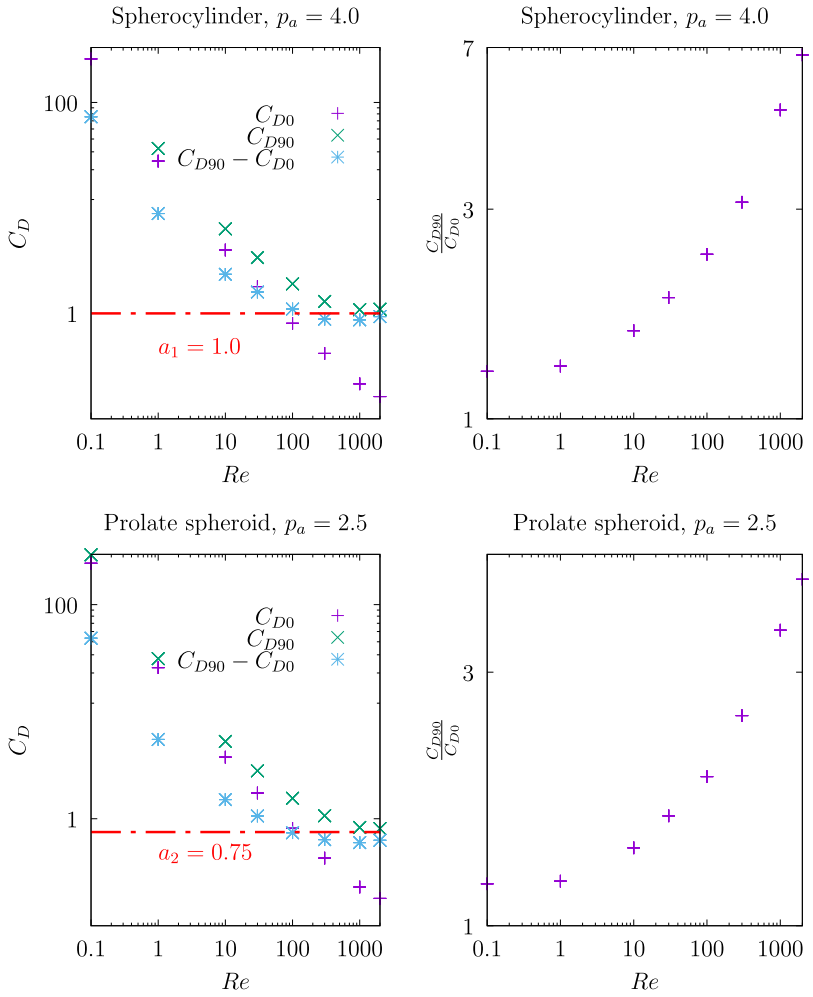

FIG. 12. The variation of C_{D0}, C_{D90}, and $\frac{C_{D90}}{C_{D0}}$ from the data of Sanjeevi *et al.*[30]

that do not involve $\frac{1}{Re}$. In addition, the values of the coefficients for the non-$\frac{1}{Re}$ terms are similar, for example, their average relative difference is about 14%. This perhaps surprising result may imply that both prolate spheroid and spherocylindrical particles could behave in a similar fashion at high Re. For Eqs. (10) and (16), if we take the asymptotic limit for $Re \to \infty$, the C_{D0} and C_{D90} depend only on Re but not on p_a. The most interesting finding of the C_D correlations is that $C_{D90} - C_{D0}$ at $Re \to \infty$ is constant, and it is dependent neither on Re nor on p_a, which is reflected in the constant values of the coefficients a_0 and a_2 for Eqs. (10) and (16), respectively. First, we will test these assumptions for the original data of Sanjeevi *et al.*,[30] which have been used to obtain Eqs. (10) and (16). What we observe for C_{D0} and C_{D90} for both geometries is that they are both decreasing with Re, as shown in Fig. (12), and their ratio is increasing with Re. The interesting behavior is that $C_{D90} - C_{D0}$ reaches an asymptotic behavior well before C_{D90} itself. Interestingly, both geometries share nearly the same Re at which the asymptotic region for $C_{D90} - C_{D0}$ starts. Equations (10) and (16) predicted the asymptotic value of $C_{D90} - C_{D0}$ for the data of Ref. 30 with great accuracy. To the best of our knowledge, we are the first to report such behavior, which is of great importance since it will help us to significantly reduce the number of expensive runs that are needed to explore the high Re regime.

The discovery of this new physics is attributed mainly to the symbolic regression machine learning algorithm, which searched for billions of mathematical formula combinations that helped significantly to find an optimum formula that describes the physics of the phenomena in a simple way. However, this discovery would never been achieved, if we did not select $\log(Re)$ and $\log(Re)^2$ as initial guess functions following the suggestion of Proudman and Pearson,[20] who stated that C_D will be a function of the powers of $\log(Re)$ multiplied by powers of Re. The failure of previous investigations[28–30] to report the asymptotic behavior of $C_{D90} - C_{D0}$ is mainly because of the complex nature of their correlations. Up to now, we tested the new physics that we extracted from Eqs. (10) and (16) on a single aspect ratio particle. Now we will examine the same physics for different aspect ratios.

From Figs. 3 and 6, we can see that C_{D0} from the results of other investigations[28,29] at different p_a matches our predictions. This is strong evidence that at high Re, the C_{D0} is solely dependent on Re but not p_a, and surprisingly, the values for C_{D0} are similar for both geometries that we used in the current investigation. However, if we just look at the same figures we observe that C_{D90} does depend on p_a and Re, which the current equations fail to predict. This leads one to assume that $C_{D90} - C_{D0}$ may also dependent on Re and p_a as well. What made us think in the beginning that the form of Eqs. (10) and (16) captures the overall physics of the problem was that Eqs. (9) and (18) show nearly the same physics as our derived equations in which C_{D0} and C_{D90} only depend on Re but not on p_a for a high inertial regime. It is true that in Eq. (9), the $C_{D90} - C_{D0}$ at high Re depends also on $\log(Re)^2$. However, this equation is derived for $1 < Re < 40$, which is in the lower Re end. As for Eq. (18), the symbolic algorithm failed to show any dependency of C_{90} on p_a or Re, even though the data that used for its derivation were from the higher Re(<200) end. Also, Eq. (18) showed that $C_{D90} - C_{D0}$ is constant and does not depend on p_a or Re.

We will take the case of the prolate spheroid to investigate further why we did not get any dependency of C_{D90} on p_a or Re because we have a suitable amount of data from different sources to compare with. We will assume that $C_{D90} - C_{D0}$ is just depending on p_a, not on Re. Then, we will create a mixed dataset that contains data of $C_{D90} - C_{D0}$ for different values of p_a from the data of Refs. 28–30 for $Re = 200$, which we believe is suitable for $C_{D90} - C_{D0}$ to reach an asymptotic value. We get the following equation from the symbolic regression algorithm:

$$C_{D90} - C_{D0} = 0.091 + 0.066p_a + 0.153\,\log(p_a) - \frac{0.0321\,\log(p_a)}{\log(\log(p_a))}. \quad (32)$$

We will modify Eq. (16) by equating Eq. (32) with a_2 resulting in the following equation:

$$C_D = a_1 + a_2\sin(\phi)^2 + a_3\log(Re)^2 + \frac{\sum_{i=1}^{i=2} A_i}{Re} + a_7\log(Re), \quad (33a)$$

$$A_1 = a_4(k_{a0} + (k_{a90} - k_{a0})\sin(\phi)^2), \quad (33b)$$

$$A_2 = a_5 + a_6\log(Re)^2, \quad (33c)$$

$$a_2 = 0.091 + 0.066p_a + 0.153\,\log(p_a) - \frac{0.0321\,\log(p_a)}{\log(\log(p_a))}. \quad (33d)$$

Equation (33) is a modified version of Eq. (16) with the same coefficients, except that of a_2. However, our efforts will not stop here, and we will use the DIDFS to obtain an equation for C_D through interbreeding the data of Ref. 30, with the Stokes flow solution, and Eq. (32). The initial function is

$$C_D = f\left(\phi, p_aRe, \frac{(ka0 + (k_{a90} - k_{a0})\sin(\phi)^2)}{Re}, Re, \log(Re), \log(Re)^2, \left(0.091 + 0.066p_a + 0.153\,\log(p_a) - \frac{0.032\,\log(p_a)}{\log(\log(p_a))}\right)\sin(\phi)^2\right). \quad (34)$$

We obtained the following equation for C_D:

$$C_D = a_1 + a_2Re + \left(a_3 + a_4p_a + a_5\log(p_a) + a_6\frac{\log(p_a)}{\log(\log(p_a))}\right) \times \sin(\phi)^2 + \frac{a_7(k_{a0} + k_{a90}\,\sin(\phi)^2 - k_{a0}\,\sin(\phi)^2)}{Re} + a_8\,\log(Re)^2. \quad (35)$$

The coefficients of Eq. (35) are listed in Table VII. Equation (35) is not similar to that of Eq. (16), and we may have introduced some noise in the data, since the gene for the $C_{D90} - C_{D0}$ came from three different sources.[20–30]

To test the validity of our assumption on the variation of $C_{D90} - C_{D0}$, Eqs. (33) and (35) will be tested extensively for different cases, especially for the high Re regime. We will compare our predictive C_D

TABLE VII. Coefficients for Eq. (35).

Coefficients	Equation (35)
a_1	1.539
a_2	0.005
a_3	0.091
a_4	0.006
a_5	0.153
a_6	−0.032
a_7	25.28
a_8	−0.0438

equations with three additional sources. The first one is Ke *et al.*,[39] who conducted lattice Boltzmann simulations for two prolate spheroids ($p_a = 2$ and 2.5) for $10 < Re < 200$ but also provided a correlation for C_D as a function of p_a that is built from data from spheres, oblate, and prolate spheroids. The second source is also from lattice Boltzmann simulations provided by Hölzer and Sommerfeld[40] for the case of a prolate spheroid with $p_a = 1.5$ for a range of Re between 0.3 and 240. The final source is a generalized equation derived also by Hölzer and Sommerfeld[41] by fitting experimental and theoretical data for different particle shapes.

For the case of $p_a = 15$, Eqs. (33) and (35) capture with great accuracy the results of Ref. 28 for the whole range of ϕ and the two Re considered, as shown in Fig. 13. The relative error between Eq. (33) and the results of Ref. 28 for $Re = 200$ is about 6.0% and that of Eq. (35) is about 10%. While the results of Ref. 39 underpredict the values of Ref. 28 at low angles of attack, they overpredict them at high angles of attack. Their relative difference with the results of Ref. 28 is about 27%. This is not surprising since the correlation of Ref. 39 is based on two aspect ratios 2 and 2.5, which may, therefore, be accurate only in that range. However, the surprising observation is that Eq. (16) has a better relative error of 22% (Table VI) than that of the correlation derived in Ref. 39, even though it is derived from a single aspect ratio ($p_a = 2.5$) and the Stokes flow solution. The predictions of Hölzer and Sommerfeld[41] gave an average relative error of 17%, which is good for an equation

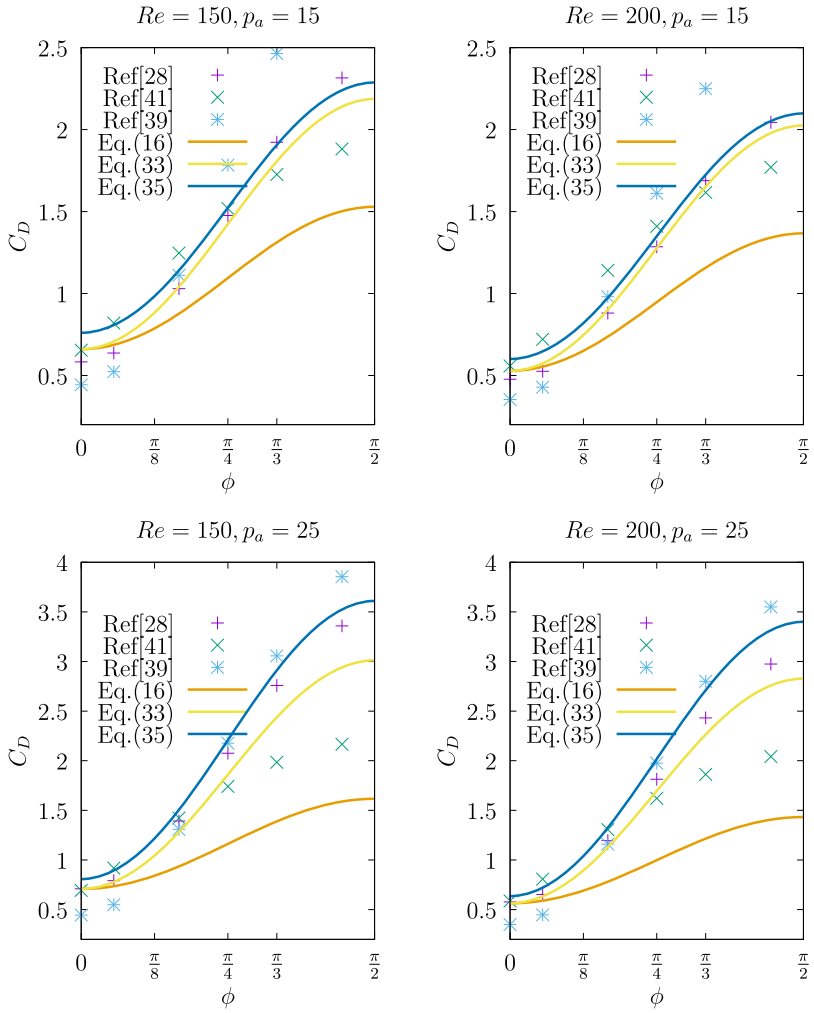

FIG. 13. Comparison between the results of different predictors of C_D and those of Ouchene *et al.*,[28] Hölzer and Sommerfeld,[41] and Ke *et al.*[39] for different Re, ϕ, and $p_a = 15$ and 25.

that is derived for general arbitrary nonspherical shape. However, its accuracy is deteriorating with an increase in Re and does not also reveal the physics that governs the evolution of the drag force for nonspherical particles. For the case of p_a = 25, Eqs. (33) and (35) gave the best predictions with just 5% relative error for each of them at Re = 200. On the other hand, the relative error for Eq. (16) is deteriorated to 35%, the relative error of Hölzer and Sommerfeld[41] remains nearly constant at 18%, similar to the case of p_a = 15, and the relative error of the Ke *et al.*[39] correlation is reduced to 19%.

We now move to geometries with lower aspect ratios as in Fig. 14. For those cases, we have the opportunity to compare our results with the results of additional numerical solvers. For the case of p_a = 1.25, we observe that the results of Refs. 28, 39, and 41 are close to each other, while the results of Eqs. (33) and (35) are in a very close proximity with those of Ref. 29. The reason why our data reassemble those of Ref. 29 and not Ref. 28 is because in the dataset that we used to derive Eq. (32) the values of $C_{D90} - C_{D0}$ for $p_a < 2.5$ are obtained from the data of Ref. 29. The reason for this selection is because we believed that the data of Ref. 29 are more accurate than Ref. 28 for small aspect ratios. However, we also wanted to test if the statement that $C_{D90} - C_{D0}$ is independent of Re for the inertial regime and still holds for the results of Ref. 29. The results for the case of p_a = 1.25 shows without doubt that $C_{D90} - C_{D0}$ depends only on p_a and not on Re for the data of Ref. 29. Similar behavior for the $C_{D90} - C_{D0}$ variation is observed also for the numerical results of Ref. 40. Furthermore, our results of Eq. (35) resemble better their numerical results than their own correlation.[41]

From Figs. 13 and 14, we proved that $C_{D90} - C_{D0}$ only depends on p_a but not on Re in the inertial regime for prolate spheroids. We have great confidence that the same law applies to other sufficiently smooth axisymmetric nonspherical particles such as spherocylinders or oblate spheroids. We believe that the $C_{D90} - C_{D0}$ asymptotic behavior could be independent of the $\sin(\phi)^2$ dependency of the drag, which has been observed for these types of particles.[36] The first evidence of this statement came from recent results of Pierson *et al.*[42] They conducted simulations for finite cylinders for an aspect ratio of 3, for 25 < Re < 250, and for different angles of attack. They reported that the C_D in the inertial regime does not vary according to $\sin(\phi)^2$ law. However, if we inspect their Fig. 22, we found that $C_{D90} - C_{D0}$

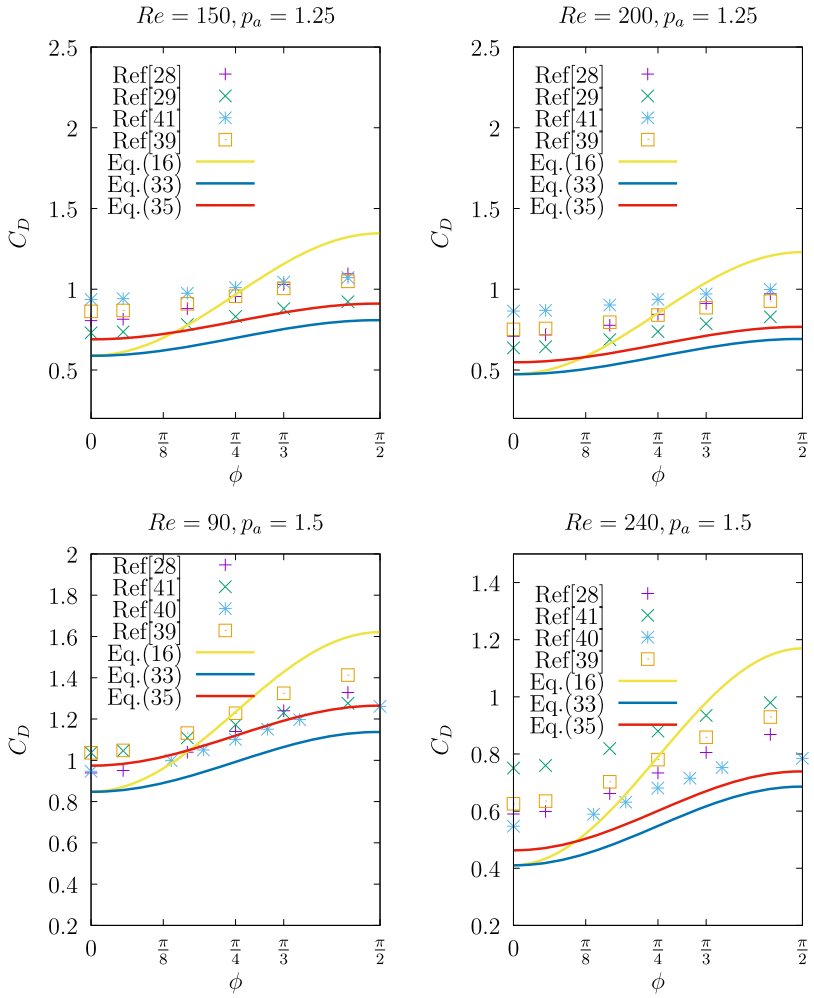

FIG. 14. Comparison between the results of different predictors of C_D and those of Ouchene *et al.*,[28] Zastawny *et al.*,[29] Hölzer and Sommerfeld,[41] Hölzer and Sommerfeld,[40] and Ke *et al.*[39] for different Re, ϕ, and p_a = 1.25 and 1.5.

TABLE VIII. Comparison between the results of different C_D predictors with the results of Ouchene *et al.*[28] for $p_a = 15$.

C_D[28]	ϕ	Re	Equation (33)	Relative error (%)	Equation (35)	Relative error (%)
0.582	0	150	0.660	13.22	0.760	30.42
1.47	$\frac{\pi}{4}$	150	1.42	3.47	1.52	3.22
2.36	$\frac{\pi}{2}$	150	2.18	7.58	2.28	3.46
0.47	0	200	0.52	10.61	0.60	25.86
1.28	$\frac{\pi}{4}$	200	1.27	0.56	1.34	5.02
2.09	$\frac{\pi}{2}$	200	2.02	3.10	2.09	0.27

for both $Re = 75$ and 100 is about 0.55. This shows that $C_{D90} - C_{D0}$ reaches an asymptotic value, which does not depend on Re similar to the prolate spheroids. However, in the current investigation, we could not shed light on the physical nature of Eq. (32). As for our second finding that C_{D0} is only dependent on Re, it is clear that it is held. For example, C_{D0} is 0.5 for $Re = 200$, regardless of p_a or the numerical solver. Table VIII contains more information about the accuracy of Eqs. (33) and (35).

After the extensive verification of our correlations in the high Re regime, we will compare our results with those of Andersson and Jiang[43] for the case of $p_a = 6.0$. Andersson and Jiang[43] used the immersed boundary method as their numerical method of choice, and in their investigation, they were only interested in the low Re regime. The comparison is listed in Table IX. In general, our results are within 10% from those of Ref. 43, except for the case of $\phi = \frac{\pi}{2}$ where the deviation jumps to 24.7%. This deviation is due to the different numerical schemes used to generate the data that helped in the derivation of Eqs. (16) and (35) and those of Ref. 43. In the literature, there is a great need for a bench mark investigation that will compare the results of C_D for different numerical schemes for different particle shapes and Re and set the standards for an accurate solution. This will help in training the machine learning algorithms, thus obtaining a more accurate prediction out of them. The interesting part of the results of Table IX is that Eq. (16) performed very well in regimes well above the Stokes flow, even though it was derived from data for a single $p_a = 2.5$. This shows that the symbolic regression algorithm learned about the flow regime beyond Stokes flow.

TABLE IX. Comparison between the results of different C_D predictors with the results of Andersson and Jiang[43] for $p_a = 6$.

C_D[43]	ϕ	Re	Equation (16)	Relative error (%)	Equation (33)	Relative error (%)
138.39	0	0.181	150.07	8.40	151.95	9.79
174.22	$\frac{\pi}{4}$	0.181	178.16	2.2	179.40	2.9
203.44	$\frac{\pi}{2}$	0.181	206.33	1.3	206.86	1.6
2.54	0	18.17	2.725	7.08	2.686	5.74
3.59	$\frac{\pi}{4}$	18.17	3.335	7.10	3.290	8.03
5.17	$\frac{\pi}{2}$	18.17	3.944	23.71	3.893	24.70

V. CONCLUSIONS

We demonstrated the feasibility of using a symbolic regression machine learning method for solving a very specific problem of predicting the fluid drag felt by ellipsoidal and spherocylinder particles. The way that we used the symbolic regression algorithm is far from being just as a fitting tool as in Ref. 32. On the contrary, the way that it learned about the data was semisupervised. For example, it found the dependence of C_D on the aspect ratio p_a even though this was a latent variable, i.e., a variable that was not varied in the initial dataset given to the algorithm for training. We presented a set of new drag correlations that we believe are valid in the high Re regime and for different p_a of the particle geometry, which is a substantial extension of the current correlations that exist in the literature. We also showed how the DIDFS method helps to insert the characteristics of one numerical scheme into another. We also presented new physics partially discovered by the machine learning algorithm. We found that $C_{D90} - C_{D0}$ in the high Re regime depends only on p_a, while C_{D0} depends only on Re at the same conditions. In our opinion, one of the main findings of the current research is that it is possible to construct generic drag correlations from the Stokes flow solutions, which were already known from the 1950s and a handful of data obtained in the current century. We end with the following recommendations:

- We believe that we can improve even further the accuracy of our correlations by finding new ways of training the algorithm. A way to speed up the learning by the symbolic regression algorithm is by changing the initial functional forms during the execution period of the algorithm.
- The genetic algorithm on which symbolic regression is based may be customized to suit fluid mechanics problems.
- More complex problems in fluid dynamics could be taken into account to understand the volume of the training data needed, the complexity of the initial functions, and the number of latent variables that we can solve at once.
- The machine learning method that we used can be an excellent candidate to find a generalized formula for the resistance tensor of arbitrary shape particles. The resistance tensor is the main component that controls the hydrodynamics of a particle in the Stokes flow regime. We know the resistance tensor for simplified shapes, while we use numerical techniques to calculate it for more complex geometries. We believe that using the results from both theoretical and numerical techniques as previous knowledge for symbolic regression will help to formulate an accurate mathematical picture of a general flow resistance tensor. This will help shed more light on the general physics and will also help in the application of Stokesian dynamics based methods[44] to more complex geometries.

ACKNOWLEDGMENTS

We thank Sathish K. P. Sanjeevi for providing the datasets for the drag coefficient used for obtaining the correlations. El Hasadi is thankful to Sathish K. P. Sanjeevi and Vinay Mahajan about the instrumental discussions about the physical nature of the drag force around nonspherical particles.

REFERENCES

[1] R. H. Pletcher, J. C. Tannehill, and D. Anderson, *Computational Fluid Mechanics and Heat Transfer* (CRC Press, 2012).
[2] S. Patankar, *Numerical Heat Transfer and Fluid Flow* (CRC Press, 1980).
[3] T. J. Hughes, L. P. Franca, and G. M. Hulbert, Comput. Methods Appl. Mech. Eng. **73**(2), 173–189 (1989).
[4] S. Chen and G. D. Doolen, Annu. Rev. Fluid Mech. **30**(1), 329–364 (1998)
[5] J. P. Boris, Annu. Rev. Fluid Mech. **21**(1), 345–385 (1989).
[6] S. Lloyd, Nature **406**(6799), 1047 (2000).
[7] G. K. Batchelor, *An Introduction to Fluid Dynamics* (Cambridge University Press, 2000).
[8] D. Jeffrey and Y. Onishi, J. Fluid Mech. **139**, 261–290 (1984).
[9] M. A. van der Hoef, R. Beetstra, and J. Kuipers, J. Fluid Mech. **528**, 233–254 (2005).
[10] X. Wu and P. Moin, J. Fluid Mech. **630**, 5–41 (2009).
[11] H. U. Voss, P. Kolodner, M. Abel, and J. Kurths, Phys. Rev. Lett. **83**(17), 3422 (1999).
[12] J. Tompson, K. Schlachter, P. Sprechmann, and K. Perlin, "Accelerating Eulerian fluid simulation with convolutional networks," in *ICML'17 Proceedings of the 34th International Conference on Machine Learning* (PMLR, 2017), Vol. 70, pp. 3424–3433.
[13] S. Jeong, B. Solenthaler, M. Pollefeys, M. Gross *et al.*, ACM Trans. Graphics **34**(6), 1 (2015).
[14] J. Bongard and H. Lipson, Proc. Natl. Acad. Sci. U. S. A. **104**(24), 9943–9948 (2007).
[15] M. Schmidt and H. Lipson, Science **324**(5923), 81–85 (2009).
[16] S. L. Brunton, J. L. Proctor, and J. N. Kutz, Proc. Natl. Acad. Sci. U. S. A. **113**(15), 3932–3937 (2016).
[17] N. M. Mangan, S. L. Brunton, J. L. Proctor, and J. N. Kutz, IEEE Trans. Mol., Biol. Multi-Scale Commun. **2**(1), 52–63 (2016).
[18] G. G. Stokes, *On the Effect of the Internal Friction of Fluids on the Motion of Pendulums* (Pitt Press, Cambridge, 1851), Vol. 9.
[19] C. W. Oseen Ark. Mat., Astron. Fys. **6**, 1 (1910).
[20] I. Proudman and J. Pearson, J. Fluid Mech. **2**(3), 237–262 (1957).
[21] E. J. Chang and M. R. Maxey, J. Fluid Mech. **277**, 347–379 (1994).
[22] R. Cox, J. Fluid Mech. **23**(4), 625–643 (1965).
[23] D. Breach, J. Fluid Mech. **10**(2), 306–314 (1961).
[24] T. Aoi, J. Phys. Soc. Jpn. **10**(2), 119–129 (1955).
[25] R. Pitter, H. Pruppacher, and A. Hamielec, J. Atmos. Sci. **30**(1), 125–134 (1973).
[26] J. H. Masliyah and N. Epstein, J. Fluid Mech. **44**(3), 493–512 (1970).
[27] A. Vakil and S. I. Green, Comput. Fluids **38**(9), 1771–1781 (2009).
[28] R. Ouchene, M. Khalij, B. Arcen, and A. Tanière, Powder Technol. **303**, 33–43 (2016).
[29] M. Zastawny, G. Mallouppas, F. Zhao, and B. Van Wachem, Int. J. Multiphase Flow **39**, 227–239 (2012).
[30] S. K. Sanjeevi, J. Kuipers, and J. T. Padding, Int. J. Multiphase Flow **106**, 325–337 (2018).
[31] J. Koza, *Genetic Programming* (MIT Press Cambridge, 1992).
[32] R. Barati, S. A. A. S. Neyshabouri, and G. Ahmadi, Powder Technol. **257**, 11–19 (2014).
[33] M. Quade, M. Abel, K. Shafi, R. K. Niven, and B. R. Noack, Phys. Rev. E **94**(1), 012214 (2016).
[34] J. Happel and H. Brenner, *Low Reynolds Number Hydrodynamics: With Special Applications to Particulate Media* (Springer Science & Business Media, 2012), Vol. 1.
[35] R. Cox, J. Fluid Mech. **44**(4), 791–810 (1970).
[36] S. K. Sanjeevi and J. T. Padding, J. Fluid Mech. **820**, R1 (2017).
[37] M. D. Schmidt and H. Lipson, in *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation* (ACM, 2009), pp. 1091–1098.
[38] J.-C. Loiseau and S. L. Brunton, J. Fluid Mech. **838**, 42–67 (2018).
[39] C. Ke, S. Shu, H. Zhang, H. Yuan, and D. Yang, Powder Technol. **325**, 134–144 (2018).
[40] A. Hölzer and M. Sommerfeld, Comput. Fluids **38**(3), 572–589 (2009).
[41] A. Hölzer and M. Sommerfeld, Powder Technol. **184**(3), 361–365 (2008).
[42] J.-L. Pierson, F. Auguste, A. Hammouti, and A. Wachs, Phys. Rev. Fluids **4**(4), 044802 (2019).
[43] H. I. Andersson and F. Jiang, Acta Mech. **230**(2), 431–447 (2019).
[44] J. F. Brady and G. Bossis, Annu. Rev. Fluid Mech. **20**(1), 111–157 (1988).