Do logarithmic terms exist in the drag coefficient of a single sphere at high Reynolds numbers?

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

DOI
10.1016/j.ces.2022.118195

Publication date
2023

Document Version
Final published version

Published in
Chemical Engineering Science

Citation (APA)
El Hasadi, Y. M. F., & Padding, J. T. (2023). Do logarithmic terms exist in the drag coefficient of a single sphere at high Reynolds numbers? *Chemical Engineering Science*, *265*, Article 118195. https://doi.org/10.1016/j.ces.2022.118195

Chemical Engineering Science 265 (2023) 118195

Contents lists available at ScienceDirect

Chemical Engineering Science

journal homepage: www.elsevier.com/locate/ces

Do logarithmic terms exist in the drag coefficient of a single sphere at high Reynolds numbers?

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

[a] Process and Energy Department, Delft University of Technology, Leeghwaterstraat 39, 2628 CB Delft, the Netherlands
[b] Civil Engineering and Geosciences Department, Delft University of Technology, Stevinweg 1, 2628 CN Delft, the Netherlands

HIGHLIGHTS

- Obtain predictive models for the drag coefficient of a sphere using symbolic regression.
- The drag coefficient of the sphere depends on logarithmic terms of the Reynolds number.
- The logarithmic drag models have a higher extrapolation range than the power-based models.
- The logarithmic drag models can predict the drag crisis at high Reynolds numbers.
- The logarithmic drag models predict the proper behaviour at a low Reynolds number regime.

ARTICLE INFO

Article history:
Received 12 July 2022
Received in revised form 22 September 2022
Accepted 5 October 2022
Available online 13 October 2022

Keywords:
sphere
Drag coefficient
Machine learning
Multi-phase flows
Matched asymptotic expansions

ABSTRACT

At the beginning of the second half of the twentieth century, Proudman and Pearson (J. Fluid. Mech.,2(3), 1956, pp.237–262) suggested that the functional form of the drag coefficient (C_D) of a single sphere subjected to uniform fluid flow consists of a series of logarithmic and power terms of the Reynolds number (Re). In this paper, we will explore the validity of the above statement for Reynolds numbers up to 10^6 by using a symbolic regression machine learning method. The algorithm is trained by available experimental data and data from well-known correlations from the literature for Re ranging from 0.1 to 2×10^5. Our results show that the functional form of C_D contains powers of $\log(Re)$, plus the Stokes term. The logarithmic C_D expressions can generalize (extrapolate) better beyond the training data than pure power series of Re and are the first in the literature to predict with acceptable accuracythe onset of the rapid decrease (drag crisis) of C_D at high Re, but also to follow the right behaviour towards zero Re. We also find a connection between the root of the Re-dependent terms in the C_D expression and the first point of laminar separation. The generalization behaviour of power-based drag coefficient equations is worse than logarithmic-based ones, especially towards the zero Re regime in which they give non-physical results. The logarithmic based C_D correctly describes the physics from the low Re regime to the onset of the drag crisis. Also, by applying a minor modification in the logarithmic based equations, we can predict the drag coefficient of an oblate spheroid in the high Re regime.

1. Introduction

Predicting the drag force on an object fixed in a planar flow has been the subject of extensive investigation from the early days of fluid mechanics when it emerged as an independent discipline. The analytical solution for the drag force experienced by a rigid sphere for creeping flow conditions, found by Stokes (1851) in 1851, is one of the first known analytical expression in the fluid mechanic's community. Stokes assumed in his solution that inertial effects of the fluid could be neglected throughout the solution domain. However, Oseen (1910) found an inconsistency in the Stokes solution. Specifically, he found that inertial fluid effects cannot be neglected far away from the sphere. He derived a new form of equations, known as Oseen equations Oseen (1910), that can handle this inconsistency, and he came up with an improved approximation for the drag coefficient, defined as $C_D = F_D / \left(\frac{1}{2}\rho v_\infty^2 \frac{\pi}{4} d^2\right)$, where F_D is the drag force, ρ the fluid density, v_∞ the fluid flow velocity far away from the sphere, and d the sphere diameter Oseen (1913). There are additional solutions to the Oseen equations, such as those of Goldstein (1929) and Faxen (1923). Proudman and Pearson (1957) and Kaplun and

E-mail addresses: yme0001@auburn.edu (Y.M.F. El Hasadi), J.T.Padding@tudelft.nl (J.T. Padding)

https://doi.org/10.1016/j.ces.2022.118195
0009-2509/

Lagerstrom (1957) used the matched asymptotic method to solve the Navier–Stokes equations to resolve the fluid flow around different blunt bodies. Proudman and Pearson (1957) divided the flow field around the sphere into two stream function expansions. The first one, which they called the Stokes expansion, controls the flow near the surface of the sphere. The second expansion, which they called the Oseen expansion, controls the flow far from the surface of the sphere. Both expansions are based on the Navier–Stokes equations, and the two expansions are matched at a certain distance from the sphere using the method of matched asymptotics. Evaluating stresses from the Stokes expansion they arrived at the following expression for the C_D of a sphere:

$$C_D = \frac{24}{Re}\left(1 + \frac{3}{16}Re + \frac{9}{160}Re^2 \log\left(\frac{Re}{2}\right)\right) \tag{1}$$

Here $Re = \rho v_\infty d/\mu$ is the Reynolds number. They made the following statement (conjecture) about the expansions that govern the flow field Proudman and Pearson (1957): "*The non-linearity of the Navier–Stokes equation then shows that both expansions must involve powers of* $\log(Re)$*, and it seems reasonable to suppose that both expansions are in powers of Re, each term of which is multiplied by a polynomial in* $\log(Re)$". This statement also reflects on the functional form of the drag coefficient. However, the authors did not mention the *Re* range for which the statement is valid. From now on, we will call this conjecture **P&P**. Graebel (2007) supported the **P&P** statement by mentioning that the C_D functional form that will result from asymptotic expansions of the Navier–Stokes equations will always be a function of $\log(Re)$. A few years later, Chester et al. (1969) added an extra term to Eq. (1), which was the last addition that came from the expansion of the Navier–Stokes equations. The logarithmic terms are not confined to solutions of flows of low *Re*, but also they appear in asymptotic solutions of high *Re* regime. For example, they appear in the asymptotic solution of the local coefficient of skin friction for the case of a semi-infinite plate Van Dyke (1975):

$$C_f = \frac{0.664}{\sqrt{Re_x}} - 0.551\frac{\log(Re_x)}{Re_x^{3/2}} + \frac{C_1}{Re_x^{3/2}} \tag{2}$$

where Re_x is the local Reynolds number, and C_1 is a constant. Recently, Khair and Chisholm (2018) obtained an expression for the drag force, on an elongated particle using matched asymptotic expansion by solving the Navier–Stokes equations for *Re* up to a value of 200. The derived drag force shows a logarithmic dependence on *Re* as follows:

$$\frac{\overline{F}_D}{4\pi} \sim \frac{1}{\log\left(\frac{1}{\epsilon}\right)} + \frac{\log(Re)}{\log^2\left(\frac{1}{\epsilon}\right)} \tag{3}$$

where $\overline{F}_D$ is the non-dimensional drag force and ϵ is the ratio of the characteristic width to the length of the particle. Also for a shear free cylinder, the drag coefficient shows a logarithmic dependence on *Re*, and it holds for *Re* as high as 10^5 as described by Kumar et al. (2021). They described the appearance of logarithmic terms as "*atypical* ". They used the matched asymptotic method to solve the inviscid and boundary layer equations, and they came up with the following drag coefficient:

$$C_D = \frac{16\pi}{Re}\left(1 + \frac{0.24\log(Re) - 1.74}{\sqrt{Re}}\right) \tag{4}$$

The appearance of logarithmic terms (alternatively known as logarithmic switchback terms Lagerstrom and Reinelt (1984)) in the asymptotic expansions have intrigued the scientific community, because in some instances they were not forced by the governing equations Popović (2005). Van Dyke (1975) dedicated a section in his book describing the proliferation of logarithmic terms in different fluid mechanics problems, and he made the following comment: "*one can philosophize that description by fractional powers fails to exhaust the myriad phenomena in the universe, and logarithms are the next simplest function*". Initially, the logarithms were tied with paradoxes in fluid mechanics, or to the singular perturbation techniques themselves. However, Lagerstrom and Reinelt (1984) showed that logarithmic terms are part of the solution of the governing equations, and the asymptotic expansion method is just one way to reach to the solution. This view is supported by other investigations using different mathematical methods Holzer and Kaper (2014), Popović and Szmolyan (2004).

There are analytical solutions for the Stokes and Oseen regimes for some non-spherical particles such as oblate or prolate spheroids, circular cylinders and few other particle geometries (Happel and Brenner, 2012; Cox, 1965; Breach, 1961; Aoi, 1955). Eq. (1) and all other analytical solutions, regardless of the shape of the particles, are valid up to $Re \approx 1.0$.

For higher *Re*, analytical solutions for the Navier–Stokes equations cease to exist due to its non-linearity. The flow around a sphere at high *Re* consists of a mosaic of different flow morphologies, depending on *Re* as described by Achenbach (1972), Kamble and Girimaji (2020). High Reynolds number flows ($Re \geqslant 10^4$) are usually classified into four flow regimes. In the subcritical flow regime, the C_D value is independent of *Re*. In contrast, in the critical flow regime, C_D starts to decrease rapidly as *Re* increases until a minimum is reached at a critical Reynolds number. For a smooth sphere $Re_{cr} \approx 3.7 \times 10^5$. This critical flow regime sometimes is referred to as the drag crisis. Beyond the critical *Re*, in the so-called supercritical regime, the drag coefficient slowly increases with increasing *Re* until it reaches a maximum value. Further increasing *Re*, the drag coefficient stays constant and this regime is called transcritical. For the prediction of C_D at high *Re* one usually resorts to numerical simulations (Dennis and Walker, 1971; Jenson, 1959; Nakhostin and Giljarhus, 2019; Constantinescu et al., 2002) or experiments (Achenbach, 1972; Deshpande et al., 2017; Maxworthy, 1969). The results of these numerical simulations and experiments are translated into fitting correlations, with a range of applicability limited to the range of the data that is used in the fitting process. This has resulted in a zoo of correlations that take different mathematical forms (Rouse, 1961; Engelund and Hansen, 1967; Clift, 1970; Morsi and Alexander, 1972; Graf, 1984; Flemmer and Banks, 1986; Khan and Richardson, 1987; Swamee and Ojha, 1991), as shown in the extensive list published in the recent review by Goossens (2019). The majority of correlations focus on the subcritical regime and take the following functional form:

$$C_D = \underbrace{\underbrace{\frac{24}{Re}\left(C_1 + C_2 Re^a\right)}_{\text{Schiller and Naumann}} + \frac{C_3}{1 + \frac{C_4}{Re}}}_{\text{Brown and Lawler}} \tag{5}$$

The second term of Eq. (5) arises from boundary layer theory (Schlichting and Gersten, 2016), which accounts for the inertial effects of the fluid. The value of the exponent *a* ranges from 0.5 to 0.68. These type of correlations are suitable for *Re* up to 2×10^5, right before drag-crisis. There is a similarity between the structure of Eq. (5) and the analytical solution of the Oseen equations for high *Re* obtained by Weisenborn and Ten Bosch (1995), who provided the following equation for C_D:

$$C_D = 1.058 + \frac{4.58}{Re^{2/3}} + \frac{53.67}{Re^2} \tag{6}$$

Oseen equations are used as a gauging tool for understanding the more complex behaviour of the Navier–Stokes equations. However, we know that there is significant difference between the solutions

of the Navier–Stokes equations, and those of the Oseen equations, and this difference already starts at very low *Re* (Van Dyke, 1970).

In summary, almost all correlations for drag found in literature are expressed as power law expansions, similar to Eqs. (5) and (6). Correlations with logarithmic terms, such as Eqs. (1)–(4), are very rare and seem to have been largely overlooked.

The improvement of high-performance computer architectures, plus the availability of data from numerical simulations and experiments, sparked an increase in interest to use machine learning methods to solve problems in many scientific disciplines. This has led to label machine learning as the fourth paradigm in science, next to experimentation, theory and simulation (Butler et al., 2018). When it comes to fluid mechanics, applying machine learning methods constitutes a challenge for several reasons, such as the transient nature of most fluid mechanics problems, the heterogeneity of most available data, the extensive non-linearities that govern fluid mechanics, and the multi-scale nature of most problems at hand (Brunton et al., 2020). To deal with these challenges, an ideal machine learning algorithm for fluid mechanics, should possess features such as interpretability, explainability, generalisability, and convergence (Brunton et al., 2020). One of the most popular machine learning frameworks that are used extensively in different fluid mechanics problems, from solving partial differential equations (Dissanayake and Phan-Thien, 1994; Raissi and Karniadakis, 2018), discovering physics (Iten et al., 2020), learning active-nematic hydrodynamics (Colen et al., 2016), to predicting physical properties (Kushvaha et al., 2020) are artificial neural networks (ANN). Other machine learning methods that are used for scientific discovery are sparse identification of nonlinear dynamical systems for discovering differential equations from sparse data (Brunton et al., 2016), and symbolic regression that is used for discovering laws of nature (Schmidt and Lipson, 2009), discovering new materials (Weng et al., 2020), and solving fluid flow problems (El Hasadi and Padding, 2019).

The main purpose of this paper is to explore existence of the logarithmic terms in the mathematical functional form that describes the variation of the drag coefficient with *Re*. For this purpose, we need a predictive method, whose output is a mathematical functional form, so we can check its nature consistently. The only method available in the literature is symbolic regression, which is a tool for unbiased determination of correlations (Koza, 1992). Symbolic regression provides mathematical functions that may describe the training data as its output. However, we did not only select symbolic regression because of its output, but also because it generalizes and needs less computational time to obtain suitable results compared to other machine learning methods (Thompson et al., 2020).

In this paper we will use symbolic regression to re-investigate known data on drag. We will show that symbolic regression actually rediscovers the logarithmic terms, suggesting that logarithmic expansions may represent the physics better than power law expansions. As a side result, we will show that there is an intriguing connection between the found logarithmic terms and the point of first boundary layer separation. In Appendix A we will show that similar logarithmic terms can be discovered in correlations for heat transfer coefficients.

2. Methodology

The principle that we will use in this investigation to drive our predictive equations will be based on symbolic regression, as will be explained in the next paragraph, using information about the possible mathematical formulation of the solution of the Navier - Stokes equations for the case of uniform flow over a sphere. The sources that we are basing our solution space on is the **P&P** conjecture (Proudman and Pearson, 1957), Lagerstrom and Reinelt (1984),Van Dyke (1975) who all indicated that logarithmic terms could constitute the solution. We will not use a physics basis for the choice of our guess functions, beyond the fact that the drag force will be expressed in terms of relevant physical dimensionless quantities. Because we could not find any physics-based theory that can merge the low, medium, and high Reynolds number regimes, the challenge we gave ourselves is to find mathematical relations that cover all three regimes. As we will show in the coming section, the symbolic regression algorithm finds a very similar form as the drag coefficient equation obtained by Abraham (1970), the only drag coefficient correlation that has been derived from physics principles. We will use three distinctive rules to distill which equations are best describing the physical phenomena at hand, and those rules are the following:

1. The equations must fit the training data. Equations with high correlation coefficient and low values of absolute errors will be selected. This rule is common for selecting fitting equations.
2. The equations must generalize (extrapolate beyond the training data). The equation that will generalize for more than one known flow regime will be selected. This rule will distill the equations that follow the anticipated physics. Most equations in the literature do not follow the second rule, which is why we label them as fitting equations.
3. The asymptotic behaviour of certain parts of the selected equation should comply with specific predefined physical laws in flow regimes where we have a detailed physical understanding. This rule is supplementary to the second rule and will be used only when we have sufficient information.

In this paper, we will use the symbolic regression machine learning method proposed by Koza (1992). Symbolic regression is a powerful tool for searching the mathematical space for an approximate functional relation between a certain number of input and output variables, and it is based on genetic programming proposed by Holland (1992). The framework of genetic programming is probabilistic, and is not based on mathematical principles, such as correctness, consistency, justifiability, certainty, orderliness, and decisiveness as outlined by Koza (1992), but solely on the principles of Darwinian evolution (Darwin, 1859). The idea of the genetic programming is simple, and it is based on transforming an initial population (in our case a population of mathematical functions) to a new population that survived a particular fitness constraint. The main operators that are used to create the new population are similar to those found in nature, namely that of reproduction and crossover (Koza, 1992).

The algorithm first generates a random pool of functions, that undergo genetic operations such as crossover, which corresponds to the combination of two functions to give a new offspring function. Another operation is a mutation in which a certain part of the mathematical function is changed randomly. Two indices measure the fitness of the newly obtained functions. The first index is minimizing the mean square difference between the training and predicted dependent values. The second index is to check the mathematical complexity of functions, and select the simplest ones, to prevent over-fitting. The guessed functional forms will constrain the search space for the symbolic regression algorithm, and helps to improve the generalizability of the predicted equations. The equations that will show significant generalization behaviour beyond the training data will be considered to have a physical significance and may represent the physical reality. We used the Eureqa software (Schmidt and Lipson, 2009) as symbolic regression platform. A rigorous description of the symbolic regression algorithm in use in the current investigation is given in El Hasadi and Padding (2019).

In Appendix B we illustrate that the machine learning algorithm we use can capture a known function's series expansion. It shows the ability of symbolic regression to find expansions of functions, that are valid beyond the training data used to obtain them, which gives symbolic regression an advantage compared to, artificial neural networks.

3. Results

3.1. Symbolic regression of C_D Data

We will start by exploring the C_D dependency on Re for the case of a sphere. We will create three data sets for the regression process. The first one will be generated from the correlation of Brown and Lawler (2003) which has the functional shape of Eq. (5). This data set contains about 8500 points in the range $0.1 < Re < 1.9 \times 10^5$, which is enough to capture the smallest details in the C_D variation. The second data set that we will use is the exact experimental data that Brown and Lawler (2003) used themselves to derive their correlation. It contains about 450 points in the range $0.1 < Re < 1.975 \times 10^5$. The final data set is based on the Schiller and Naumman (1933) correlation, and contains of 5020 points in the range $0.1 < Re < 700$.

We will start by examining the first data set, and we will let the symbolic regression algorithm guess about the functional form of the C_D dependence on Re. We can do this by specifying the most general initial functional form:

$$C_D = f(Re) \tag{7}$$

The algorithm derived several regression equations, but here we will show two, one because it accurately fits the results, and the other because it is simple. Both equations follow the first rule (fitting the data with reasonable accuracy). The equations are the following:

$$C_D = a_1 + \frac{a_2}{Re} + a_3\sqrt{Re} + \frac{a_4}{\sqrt{Re}} + \frac{a_5}{(a_6 + Re)} + a_7 Re \tag{8}$$

$$C_D = a_1 + \frac{a_2}{Re} + \frac{a_3}{\sqrt{Re}} \tag{9}$$

The coefficients of Eq. (8), and (9) are listed in Table 1. Eq. (9) contains the Stokes $\frac{1}{Re}$ term, and the first-order term from boundary layer theory $\frac{1}{\sqrt{Re}}$. The first known dependency of C_D on $\frac{1}{\sqrt{Re}}$ came from the Blasius solution (Blasius, 1908) of the boundary layer equations proposed by Prandtl (1904) for the case of a flat plate. The C_D for blunt bodies, like a sphere, has a similar dependency on Re (Leal, 2007; Abraham, 1970). A similar form as Eq. (9) was obtained previously by fitting experimental data (Brauer and Mewes, 1972; Hölzer and Sommerfeld, 2008), and also by using concepts of boundary layer theory (Abraham, 1970). Brauer and Mewes (1972), Holzer and Sommerfeld(2008) used non-linear fitting tools to obtain their correlations, which require *a priori* knowledge of the functional structure. A comparison between the coefficients of Eq. (9), and those of Brauer and Mewes (1972),Hölzer and Sommerfeld (2008), Abraham (1970) is given in Table 2. The coefficients of Eq. (9) have similar values to those of Brauer and Mewes (1972). Compared to those of Hölzer and Sommerfeld (2008) thereis only significant difference in the value of a_3. There is also a significant difference between the coefficients of Eq. (9) and those of Abraham (1970). This may be due to the pure theoretical nature of the equation proposed by Abraham.

It is important to note that both the Stokes term and the boundary layer term have been found without using any sophisticated mathematical approach. On the contrary, they have been found by a probabilistic genetic algorithm. The emergence of the boundary layer term in Eqs. (8) and (9) without human intervention can be added to the experimental and numerical results that support boundary layer theory, even though there is no general mathematical proof of its existence, as mentioned by Batchelor (2000).

We will now try to explore the existence of logarithmic switchback terms for the drag on a sphere for the higher Re regime. We will use for this the first data-set (i.e. data from the Brown and Lawler (2003) correlation). We will start by imposing the following initial functional form:

$$C_D = f\left(\frac{24}{Re}, \log(Re), Re\log(Re), \log^2(Re)\right) \tag{10}$$

We choose this form of the initial function because we want to ensure that logarithmic switchback terms similar to Eq. (1) will be part of the initial soup of functions that the symbolic algorithm will further evolve. The symbolic regression algorithm converged to the following equation:

$$C_D = a_1 + \frac{a_2}{Re} + a_3\log(Re) + a_4\log^2(Re) + a_5\log^4(Re) \tag{11}$$

The values of the coefficients of Eq. (11) are listed in Table 3. Eq. (11) depends on powers of $\log(Re)$ and also contains the Stokes law term, Interestingly the value of a_1 coefficient for Eq. (11) matches that of Hollandbatt (1972) drag coefficient correlation (Hollandbatt, 1972). The form of Eq. (11) is partially fulfilling the **P&P** conjecture (Proudman and Pearson, 1957) for Re as high as 2×10^5. Overall, Proudman and Pearson (1957) made a profound statement more than 64 years ago, using only mathematical intuition, and they may have been right when they suspected that logarithmic switchback terms are part of the solution. It may be difficult for the current form of the genetic algorithm to spot the entire logarithmic switchback series, because reducing the complexity of the equations is part of its optimization process. Therefore, terms that do not play a significant role in the variation of the dependent variable (C_D) will die out during the evolution process. The failure of detection of $Re^n\log^n(Re)$ terms, where n is an integer, after a significant number of mathematical formula evaluations exceeding 10^{11}, suggests that their signal is weak (a metaphor for their insignificant role in the dependence of C_D on Re). If we read more carefully the conjecture, we find that Proudman and Pearson (1957) used the following wording :*"It seems reasonable to suppose that both expansions are in powers of Re*". They used the word *'reasonable to suppose'*, expressing doubt, while for the $\log(Re)$ terms they used the word 'must' which reflects that the authors were sure

Table 1
Coefficients for Eq. (8), Eq. (9), and (12).

Coefficients	Eq. (8)	Eq. (9)	Eq. (12)
a_1	0.251	0.412	0.505
a_2	23.620	23.311	23.224
a_3	0.001	4.119	2.762
a_4	3.255	-	-
a_5	49.291	-	-
a_6	97.537	-	-
a_7	-2.709×10^{-6}	-	-

Table 2
Relative difference in the values of coefficients of Eq. (9) to that of Brauer and Mewes (1972),Hölzer and Sommerfeld (2008), and Abraham (1970).

Coefficients	Brauer and Mewes (1972)	Hölzer and Sommerfeld (2008)	Abraham (1970)
a_1	2.9%	−1.94%	29.01%%
a_2	−2.95%	−2.95%	−2.87%
a_3	2.88%	27.16%	−28.40%

Table 3
Coefficients for Eq. (11) and (13).

Coefficients	Eq. (11)	Eq. (13)
a_1	3.286	3.272
a_2	24.205	23.26
a_3	−0.818	0.112
a_4	0.064	−0.652
a_5	−0.000107	0.035

about their appearance in the two expansions. Adding to that, Chester et al. (1969) was frustrated about the poor convergence of his equation, mainly because it is only valid for extremely low values of *Re*. He suggested that the expansion in powers of Re may be a poor idea (Chester et al., 1969; Hunter et al., 1990).

To further validate the ecosystem of equations we obtained, we will compare their predictions with various sources in the literature, as shown in Fig. 1. The first insight from Fig. 1 is that Eq. (1) is valid only at low *Re*, and this was one of the main reasons we believe that the scientific community did not further explore the use of logarithmic terms, even as fitting functions. Eqs. (8) and (11) follow closely the correlation of Brown and Lawler (2003), and also the experimental data used to obtain their correlation. The average relative errors between the predictions of Eqs. (8) and (11) with respect to the experimental results of Brown and Lawler (2003) are 3.87% and 3.39%, respectively. We see that Eq. (9) follows closely the results of Hölzer and Sommerfeld (2008),Brauer and Mewes (1972), while it deviates from the predictions of Abraham (1970) especially for values of *Re* above 10^3. This is expected because the equation provided by Abraham (1970) is valid for *Re* up to 10^3. Also, Eq. (9) and those of references (Brauer and Mewes, 1972; Hölzer and Sommerfeld, 2008; Abraham, 1970) cannot capture the local minimum for Re between 10^3 and 10^4 that the experimental results of Brown and Lawler (2003) show.

Comparing Eqs. (8) and (11), we find that their complexity index is 34 and 19, respectively. The complexity index shows that the logarithmic series representation of C_D is mathematically simpler compared to the power series representation, making Eq. (11) more favourite to represent the physical phenomena of the C_D variation according to Occam's razor statements (Domingos, 1999). One of these statements is: *"Given two models with the same generalization error, the simpler one should be preferred because simplicity is desirable in itself."*. However, when it comes to the accuracy of fitting the data, both equations show a similar level of accuracy. For example, the power-based Eq. (8), has a mean square error of 5.9×10^{-5}, while the same error metric for the logarithmic based Eq. (11) is 9.74×10^{-5}. Even though both equations have similar fitting behaviour, their extrapolation behaviour is very different, as we will show in the coming subsection.

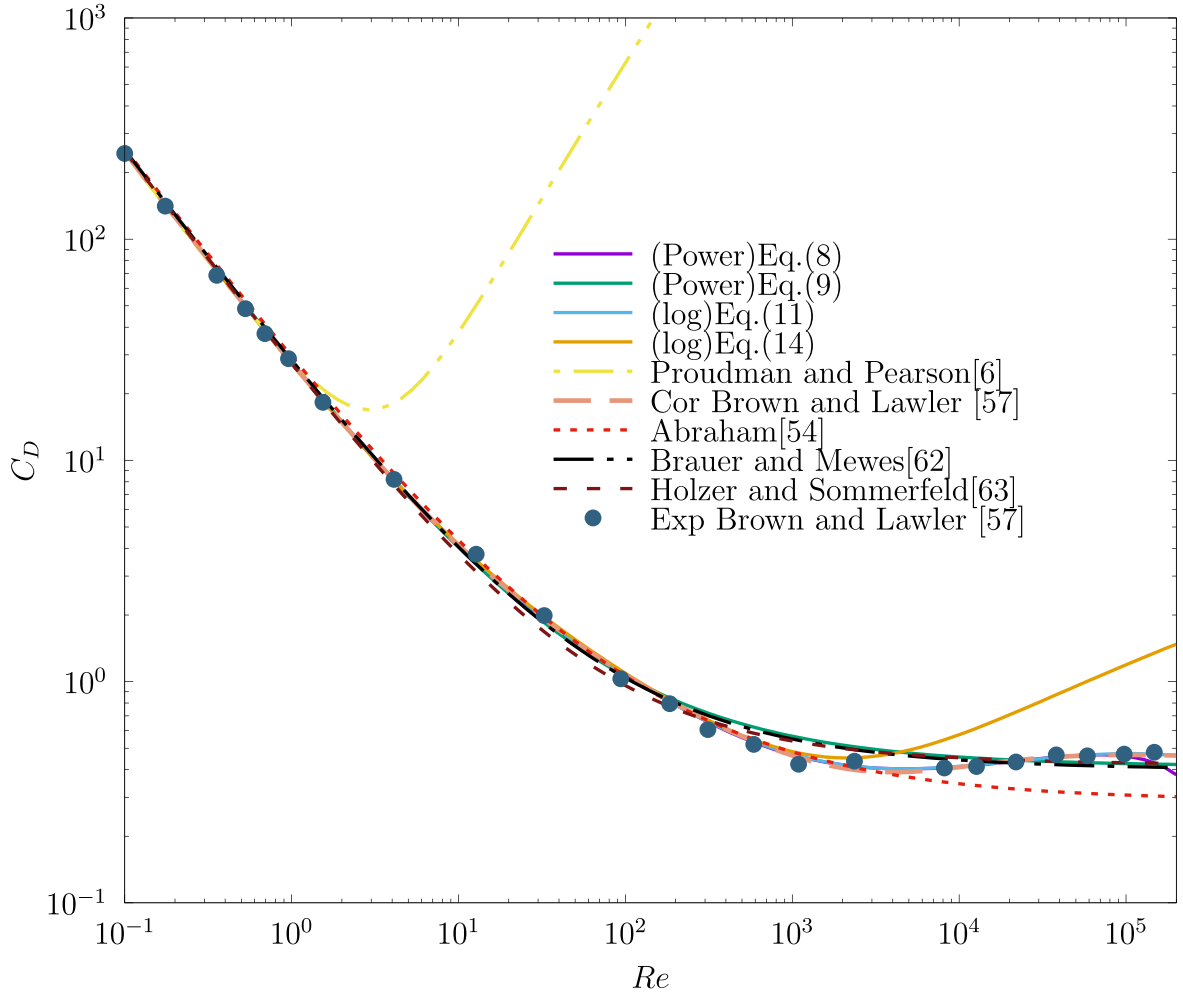

Fig. 1. Comparison between the drag coefficient C_D predicted by Eqs. (8), (9), (11) and (16)different sources from the literature. Dashed lines indicate literature correlations. Symbols indicate experimental values.

Now we will use the second (experimental) data set, to explore the feasibility of getting predictive equations for C_D from a limited amount of noisy experimental data. We will start by letting the algorithm guess the C_D dependence,by using the same initial functional form as that of Eq. (7).

The symbolic regression algorithm found the following equation:

$$C_D = a_1 + \frac{a_2}{Re} + \frac{a_3}{\sqrt{Re}} \tag{12}$$

The coefficients of Eq. (12) are listed in Table 1. Using the second data set we next explore if the data show any logarithmic dependence by imposing the following initial set of functions similar to Eq. (10).

We got the following equation for C_D:

$$C_D = a_1 + \frac{a_2}{Re} + \frac{a_3 \log^2(Re)}{Re} + a_4 \log(Re) + a_5 \log^2(Re) \tag{13}$$

The values of the coefficients are listed in Table 3. Eq. (12) is of the same functional form as Eq. (9), but the coefficients are not identical, because the second data set contains far less data, and also contains some noise. The derivation of Eq. (12) from pure experimental data, without imposing knowledge of any physics, except the definition of *Re,* shows that the symbolic regression algorithm discovered the Stokes limit and the term attributed to boundary layer theory without any external help. However, the human factor is still required since we have to select the equations that we think represent physical reality from the population of equations that the algorithm suggests. Eq. (13) shows that we can get the logarithmic dependence from a pure experimental data set, and it partially fulfils the **P&P** conjecture. Eqs. (13) and (11) are quite similar. We believe that Eq. (13) failed to capture the $\log^4(Re)$ term because this term influences C_D in the high *Re* regime where there are significant fluctuations in the experimental data set. Probably if there was a higher volume of data, especially at higher *Re,* the $\log^4(Re)$ term could also be captured from pure experimental results. A comparison of the performance of the power expansion Eq. (12) and the logarithmic expansion Eq. (13) against existing data in the literature is shown in Fig. 2. The average relative error for Eqs. (12) and (13) is 13.7% and 12.0%, respectively, against the experimental results of Brown and Lawler (2003). Eq. (13) shows a local minimum in the range of the Re close to that of the experimental results of Brown and Lawler (2003), while Eq. (12) fails to show any local minimum.

We will use the third and final data set from the Schiller and Naumman (1933) correlation which contains information about the variation of C_D for *Re* ranging from 0.1 to 700. We will use the following general initial functional form as that of Eq. (7). The symbolic regression algorithm found the following equation for C_D:

$$C_D = a_1 + \frac{a_2}{Re} + a_3 \log(Re) + a_4 \log^2(Re) \tag{14}$$

The coefficients of Eq. (14) are listed in *Table 4*. The genetic algorithm came up with the logarithmic dependence of C_D on *Re* without any external help, and it discovered the **P& P** conjecture partially. The value of a_1 = 3.1406, differs from the value of π by only about 0.03%. It will be very interesting in the future to investigate the value of a_1 by fitting to very accurate numerical or experimental data. Eq. (14) follows the Brown and Lawler correlation (Brown and Lawler, 2003) up to *Re* of 10^3, as shown in Fig. 1. This behaviour is expected because higher power logarithmic terms are missing from Eq. (14), since the training data was limited to *Re* up to 700.

Up to this point we have discussed the drag without referring to the flow around the sphere. The flow around a sphere is a rich mosaic of phenomena, and usually drag correlations fail to predict them. Among these phenomena is the emergence of a laminar separation point, which is well known to occur for sufficiently blunt objects, including a sphere. The point of laminar separation is identified by the formation of a closed recirculating ring eddy at the rear of the sphere. The first emergence of separation is difficult to detect either experimentally or theoretically. For this reason, there is some discrepancy in the literature on the value of the reported critical Re_s, and corresponding drag C_{Ds}, at first separation. The first experimental observations by Nisi and Porter (1923) suggested that Re_s = 10. This was confirmed by numerical simulations of Rimon and Cheng (1969). On the other hand, Proudman and Pearson (1957),Van Dyke (1975), by using the Stokes second expansion, estimated that Re_s = 16, close to the numerical results of Bourot (1969) and Jenson (1959) of 15.2 and 17, respectively, and the experiments of Payard and Coutanceau (1974) indicating Re_s = 17. Other simulation results (Dennis and Walker, 1971; Chang and Maxey, 1994) show that Re_s is equal to approximately 20, and the experiments of Taneda (1956) predict that Re_s = 24.

If we inspect a_1 of the logarithmic expansion Eq. (11) in Table 3 we see that its value is 3.286, which is quite similar to the value of the drag coefficient C_{Ds} at the initial laminar separation reported by Payard and Coutanceau (1974), which is 3.306. If the constant a_1 is the drag coefficient at initial laminar separation, then the following transcendental equation must have a positive root at the corresponding Reynolds number Re_s:

$$\frac{a_2}{Re} + a_3 \log(Re) + a_4 \log^2(Re) + a_5 \log^4(Re) = 0 \tag{15}$$

By solving Eq. (15) we find that Re_{rt} = 14.06 is its only root. That makes Re_{rt} the only *Re* value that zeroes off all terms beyond the constant a_1. This Re_{rt} is close to values of Re_s reported in literature. For example, the relative error with respect to the results of Bourot (1969) and Chang and Maxey (1994) is 8% and 30%, respectively. We conjecture that Re_{rt} is representing Re_s, even though we do not have any proof for this. We believe we are witnessing an instance where the machine learning algorithm found a mathematical description of a physical phenomenon, which needs human abilities to be interpreted in terms of physical laws. Otherwise, it will be a good approximation, that can describe some of the physics involved in the process of flow separation. As far as the authors are aware, there is only one analytical prediction for the point of first flow separation, from slow motion viscous theory (Proudman and Pearson, 1957; Van Dyke, 1970). However, that result was disputed by the authors of (Proudman and Pearson, 1957; Van Dyke, 1970), as we will show later. In practice, we depend on numerical simulations to find the point of zero local shear stress, as described by boundary layer theory (Schlichting and Gersten, 2016). However, Batchelor (2000) raised serious doubts about estimating the onset of separation by this method.

Beyond this point, we will assume that (the smallest, real) root Re_{rt} is equal to Re_s. Using the same procedure to calculate Re_s, from the logarithmic Eq. (13) by solving the following transcendental equation:

$$\frac{a_2}{Re} + \frac{a_3 \log^2(Re)}{Re} + a_4 \log(Re) + a_5 \log^2(Re) = 0 \tag{16}$$

we found the two following roots: Re_s = 15.76, *and* 9.52×10^7. The large root value of 9.52×10^7, is a non-physical result, which we believe is caused by the missing higher power $\log(Re)$ term from Eq. (13). However, Re_s = 15.76 compares very well with the results of Bourot (1969) and Chang and Maxey (1994), with a relative difference of 3.68% and 21.2%, respectively. If we do the same analysis for the logarithmic Eq. (14), we will find that Re_s = 15.19, and 3.518×10^6. For the smallest root, the relative difference with the

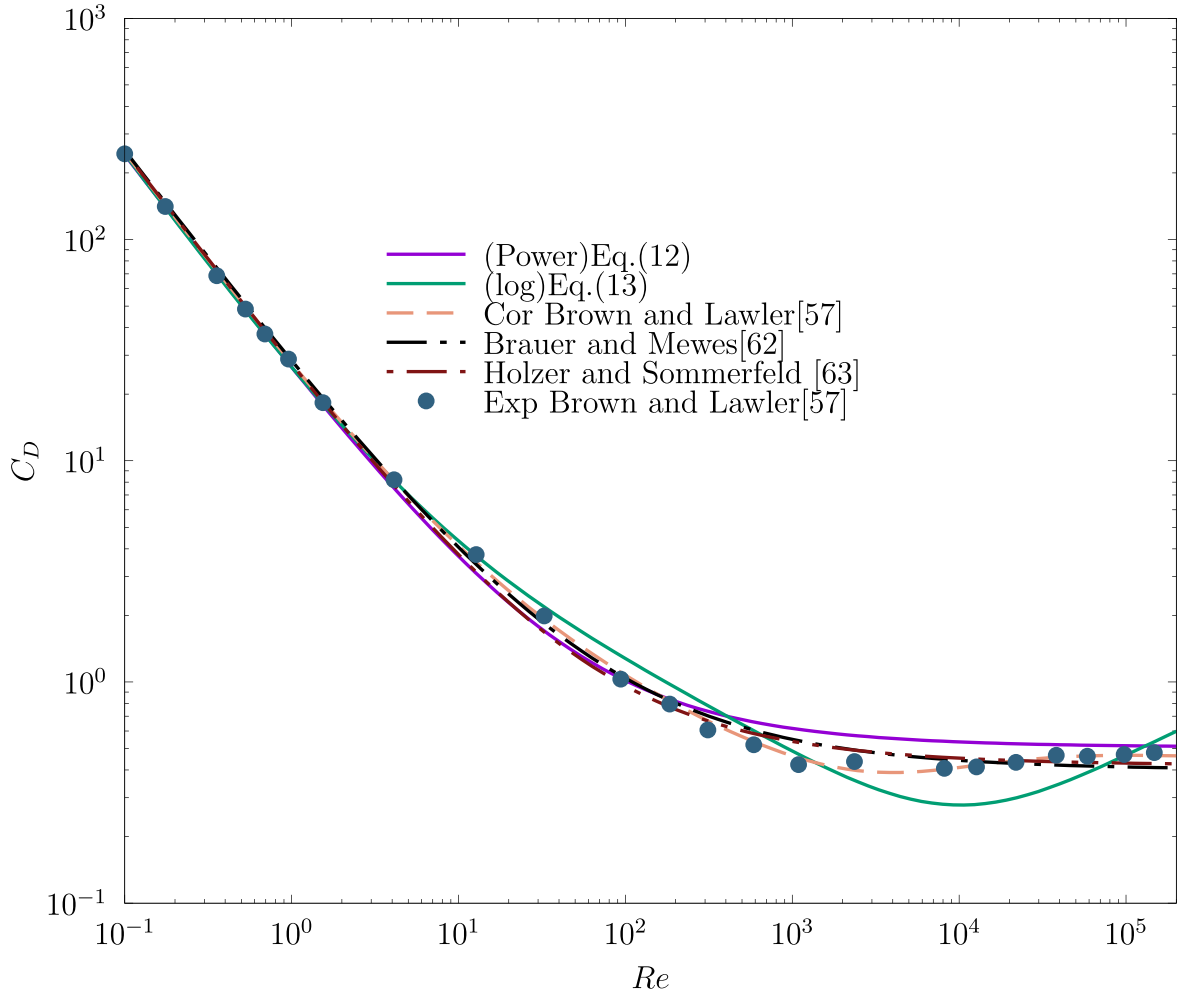

Fig. 2. Comparison between drag coefficient C_D predicted by Eq. (12) and (13), and different sources from the literature. Dashed lines indicate literature correlations. Symbols indicate experimental values.

Table 4
Coefficients for Eq. (14).

Coefficients	Eq. (14)
a_1	3.140
a_2	24.270
a_3	−0.716
a_4	0.047

results of Bourot (1969) *and* Chang and Maxey (1994) is 0.13%, and 24.0%, respectively.

We will next calculate Re_{rt} from the more popular power-law expressions Eq. (8) and (9) in the same way. For Eq. (8) $Re_{rt} = 3 \times 10^5$. This root closely approximates the critical Reynolds number ($Re_{cr} \approx 3.7 \times 10^5$) for the critical flow regime (drag crisis) as reported by Achenbach (1972). We will further discuss the physical significance of Re_{rt} in the generalization subsection since the value *of* Re_{rt3} is outside the training data range. As for power-law Eq. (9), it does not have any roots.

Returning to the logarithmic ecosystem of equations, in their seminal works, Proudman and Pearson (1957) *and* Van Dyke (1975) calculated the Re_s value to be 16 analytically from the first and second terms in the Stokes expansion. Proudman and Pearson (1957) made the following comment: *"This Reynolds number is far too large to make estimates based on only two terms of the Stokes expansion at all reliable. In fact, it cannot seriously be claimed that slow-motion theory gives even a qualitative expansion of the phenomena."* However, Van Dyke (1975) *and* Ranger (1972) tried to confirm the result of Proudman and Pearson (1957), by using extra terms in the Stokes expansion that contain the logarithmic terms from the results of Proudman and Pearson (1957) and those of Chester et al. (1969). They failed because the Stokes expansion equation that includes the logarithmic terms has only complex roots. Van Dyke (1975) commented on this issue saying that*"the logarithm needs reinterpretation."* In our work we now see that the values of Re_s from Eqs. (11), (13), and (14) are converging with different degree of accuracy toward a value of approximately 16.

3.2. Generalization beyond the training data

In this subsection, we will test our newly derived equations generalisation behaviour, for flow regimes that were not included in the training data. Specifically, we will test their behaviour for the low Reynolds number regime for Re down to 10^{-4}, and for the critical flow regime for Re up to 10^6.

3.2.1. Low Re flow regime

In the low Re regime, $\frac{24}{Re}$ is the dominant term for the drag coefficient, which will make it difficult to assess the performance of our equations, against the existing correlations, analytical solutions, experimental and numerical results. For this reason, we will use

the way Maxworthy (1965) plotted his drag coefficient data. He plotted the quantity $\frac{C_D}{C_{Ds}} - 1$ against Re, where $C_{Ds} = \frac{24}{Re}$ is the Stokes drag. This way, we eliminate the divergence of the Stokes term, which makes the comparison with different sources from the literature more precise. From low Reynolds number theory we know that $\frac{C_D}{C_{Ds}} - 1$ converges to $\frac{3}{16}Re$ (Oseen term).

The predictions for the variation of $\frac{C_D}{C_{Ds}} - 1$ against Re from our models and numerous sources from literature are shown in Fig. 3. In the range of Re 10^{-1} to 10, which is within the range of the training data, all our derived equations, plus the Brown and Lawler (2003) correlation, follow with reasonable accuracy the experimental results of Maxworthy (1965),John Veysey and Goldenfeld (2007), in addition to the numerical results of Jenson (1959),Dennis and Walker (1971). In the same Re range, the analytically derived equations of Proudman and Pearson (1957), Goldstein (1965), and Oseen (1910) deviate from experimental, and numerical results, because of their limited applicability range.

Next we turn to the Re range between 10^{-4} to 10^{-1}, which is beyond the training data range. In this flow regime, the logarithm-based Eqs. (11) and (14) follow closely the analytical results of Proudman and Pearson (1957),Goldstein (1965),Oseen (1910), and the semi-empirical and empirical correlations of Lewis and Carrier (1949),Beard and Pruppacher (1969), and the numerical simulations of Le Clair et al. (1970). On the contrary, the power-based Eqs. (9) and (8), as well the Brown and Lawler (2003) correlation, divert significantly from the analytical, experimental, and numerical data. For example the relative difference

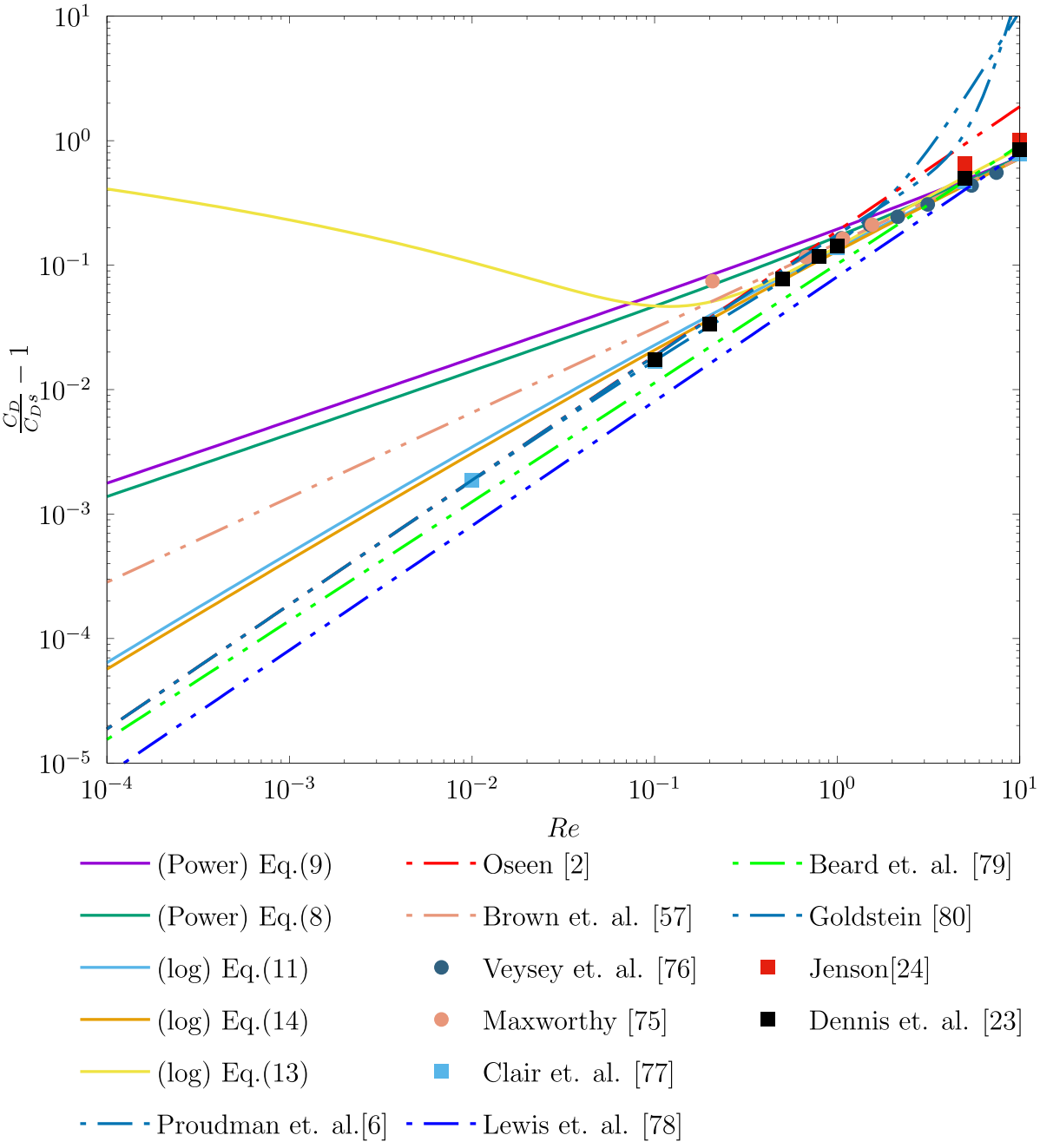

Fig. 3. Comparison between the $\frac{C_D}{C_{Ds}} - 1$ predictions in the low Re limit by Eqs. (8), (9), (11), (13), and (14), and different sources from the literature for low Re regime. Circles represents experiments, and squares represents numerical simulations.

for the prediction of $\frac{C_D}{C_{Ds}} - 1$ between Eq. (11) and the analytical solution of Proudman and Pearson (1957) is 240% at $Re = 10^{-4}$. At the same conditions, the relative difference between the Brown and Lawler (Brown and Lawler, 2003) correlation and Proudman and Pearson (1957) is 1410%, which is significantly higher than the error generated by both logarithmic equations. The five times increase in the accuracy of the logarithmic based Eqs. (11) and (14) compared to the power-based Eqs. (8) and (9) suggests that the logarithmic equations contain terms that describe the physical reality better. Another interesting aspect of the results of Fig. 3 is that it shows that we can improve the accuracy of machine learning models for the same training data set, by using previous physical knowledge about the problem at hand. The observation from Fig. 3 is similar to our observations for the Maclaurin expansion of the $\sin(x)$ function in the Appendix B. In both cases, only equations that have similar terms to the actual representation of a function, or the physical law that they are approximating, generalize well beyond their training data. The results from Fig. 3, show that the popular power-based representations of C_D fail to extrapolate beyond the range of Re that is used for their training, which indicates that power-based representations may have only been a convenient mathematical fit, rather than having physical significance. Finally, we want to explain why Eq. (13) diverges even though it consists of logarithmic terms similar to the previous two. The reason for the divergence is the $\frac{a_3 \log^2(Re)}{Re}$ term which increases its value as the value of Re decreases. This term can be considered an overfitting parameter, which is easy to spot, due to the interpretable nature of the results of symbolic regression. From the short analysis above, we conclude that the power-based equations violate the second rule because they erroneously extrapolate beyond their training data. We will further analyze the predicted equations' behaviour in the low Re regime to further assess their behaviour.

If we rearrange the Proudman and Pearson (1957) drag coefficient equation to take the following form:

$$C_D - C_{Ds} = 4.5 + 1.35 Re^2 \log\left(\frac{Re}{2}\right) \tag{17}$$

then, in the limit $Re \to 0$, the value of $C_D - C_{Ds}$ converges to 4.5. This shows that all terms that depend on Re except the Stokes term vanish. This behaviour conserves the correct physics that any drag coefficient predictive equation must follow. Also, it paves the way to test the third rule that we suggested for the predictive equation selection. The evolution of $C_D - C_{Ds}$ for the different predictive equations and correlations from the literature are shown in Fig. 4 for Re values ranging from 10^{-12} to 10^{-3}. The value of $C_D - C_{Ds}$ for Proudman and Pearson (1957) is constant, and it is 4.5 as expected for the range of Re tested. On the other hand, for the case of the power-based Eq. (8) and (9), Brown and Lawler (2003),Abraham (1970) correlations the value of $C_D - C_{Ds}$ is increasing steeply with decreasing Re which is non-physical. On the contrary, the values of $C_D - C_{Ds}$ from the logarithmic equations Eqs. (11) and (14) follow closely that of Proudman and Pearson (1957), and thus describe the correct physics more precisely. Specifically, Eq. (14) shows a very slow increase of the $C_D - C_{Ds}$ value as Re is deceasing. By increasing the number of logarithmic terms as in Eq. (11), the value of $C_D - C_{Ds}$ attains a constant value of ≈ 25 as shown in Fig. 4. At extremely low Re, the $C_D - C_{Ds}$ predicted by Eq. (11) drops to a value very close to that of Proudman and Pearson (1957), which shows that Eq. (11) follows the correct physics at extremely low Re.

Due to the violation of the second and third rules by the power-based equations in the low Re regime, we conclude that those equations can not describe the drag coefficient in all Re regimes, making them unsuitable to describe the physical evolution of the drag coefficient. In the following subsection, we will add more proof by comparing power-based equations with logarithmic based equations also in the large Re limit.

3.2.2. Critical flow regime

The critical flow regime is less well investigated, either experimentally or numerically, compared to the subcritical or lower Re regimes. There are no any analytical approximations for C_D in the critical flow regime. Even direct numerical simulations (DNS) are limited to the onset of the subcritical flow regime at $Re = 10^4$ (Beratlis et al., 2019). Current computational fluid dynamics (CFD) simulations that deal with the critical flow regime use different approximations to deal with turbulence. Constantinescu et al. (2002) use Detached-Eddy-Simulations (DES), which is a hybrid method that combines Reynolds-Averaged Navier–Stokes (RANS) and Large Eddy Simulations (LES). Nakhostin and Giljarhus (2019) used RANS turbulence models for their simulations, and Muto et al. (Muto et al., 2012) used Large Eddy Simulations coupled with the a subgrid-scale turbulence model. The most extensive numerical simulations in the critical and supercritical regime have been conducted by Geier et al. (2017) using a Cumulant Lattice Boltzmann method, and they do not use any turbulence models. Their high fidelity model uses a fourth-order accurate diffusion approach, suitable for low viscosity high Re flows. The accuracy of the Cumulant Lattice Boltzmann depends on the optimization of its parameters. The authors used a spectrum of three different mesh grid schemes, namely a course one with 40×10^6 nodes, a medium one with 75×10^6 nodes, and a fine grid mesh with 133×10^6 nodes.

Fig. 5 explores the performance of the power-based Eq. (8) and logarithm-based Eq. (11) in the subcritical, critical, and supercritical flow regimes, and compares their performance against experimental and numerical results. The training data for Eq. (8) and (11) was limited to Re up to 2×10^5. There is a significant discrepancy between the different experimental results, for different reasons, such as the turbulence intensity the positions of the sensors around the sphere (Batchelor, 2000). Eq. (8) follows the anticipated trend in the critical flow regime in which the C_D is decreasing with increasing Re. Note that on the contrary, the value of C_D from the correlation of Brown and Lawler (2003) stays constant for Re values higher than 10^4. The onset of the critical flow regime for the power-based equation Eq. (8) starts at approximately $Re \approx 10^5$, earlier than most experimental and numerical results, except the experimental data of Maxworthy (1969), in which the critical flow regime starts at much lower Re. At approximately $Re = 3 \times 10^5$ Eq. (8) drops to zero, and its values resemble the experimental values of Achenbach (1972). The drop of Eq. (8) to zero at $Re = 3 \times 10^5$ was already predicted algebraically in the previous section, and (8) is the first in literature that predicts with good accuracy the value of Re_{cr} reported by the experiments of Achenbach (1972). From Fig. 5, we can see that even the high fidelity simulations of Geier et al. (2017) with fine grid failed to predict Re_{cr} since they failed to resolve the Kolmogorov length scale at such high Re. The numerical results for the medium grid scheme of Geier et al. (2017) are close to the predictions of Eq. (8) for Re until the critical Reynolds number. The logarithmic based Eq. (11) predicts the onset of the critical flow regime with great accuracy since it follows the C_D values from the experiments of Suryanarayana et al. (1993), Achenbach (1972) from $Re = 5 \times 10^4$ to about 3×10^5. Eq. (11) does not drop to zero at Re_{cr} as Eq. (8), however it follows very closely the high fidelity numerical results of Gerier et al. (Geier et al., 2017) for the coarse grid case for Re up to 10^6. This shows that Eq. (11) follows an approximately physical reality for Re up to 10^6, since the results of Gerier et al. (Geier et al., 2017) are generated by solving an approximate form of the Navier–Stokes equa-

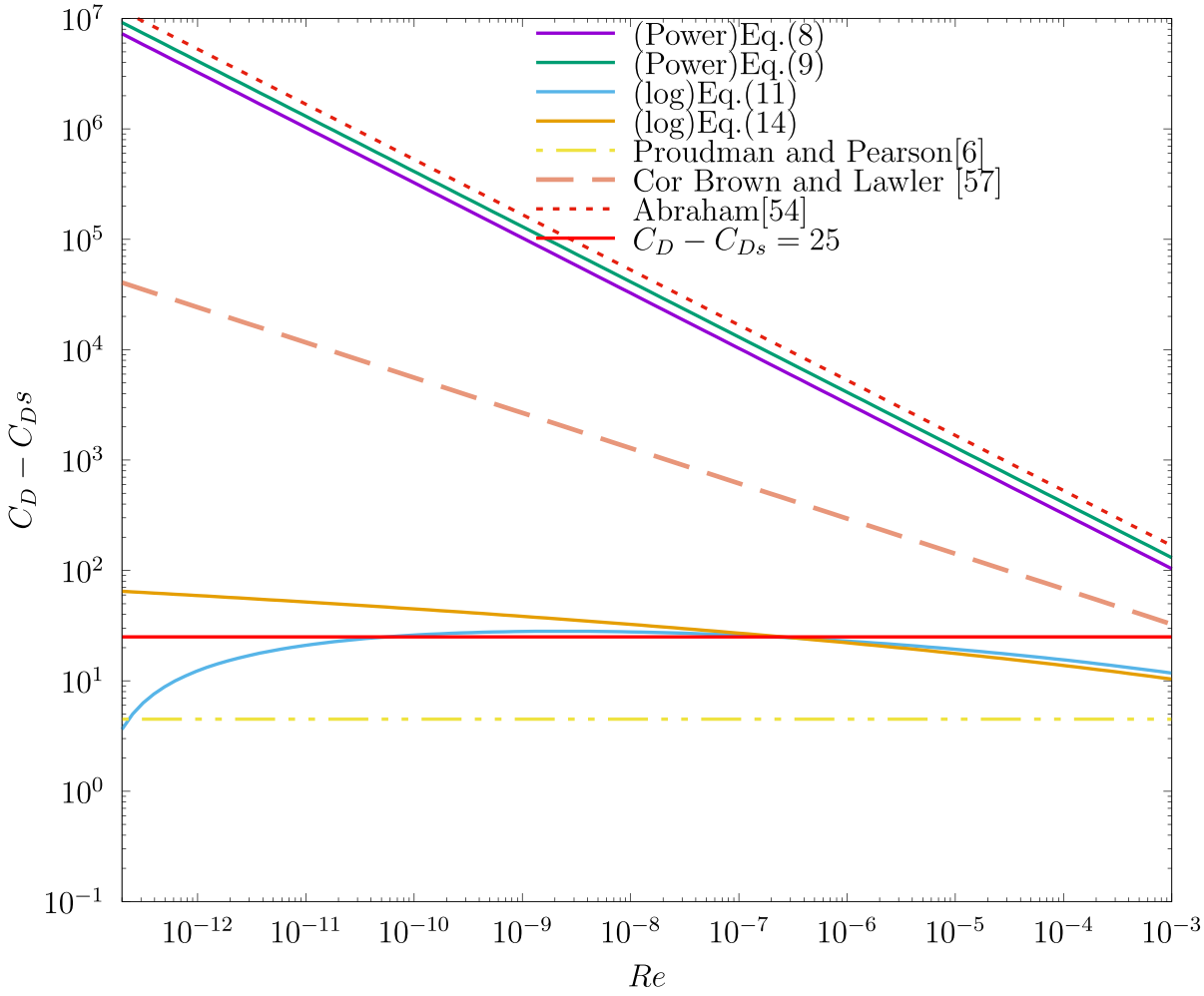

Fig. 4. Comparison between the $C_D - C_{Ds}$ predictions in the low *Re* limit by Eqs. (8), (9), (11), (13), and (14), and different sources from the literature for low *Re* regime.

tions. Both Eq. (8) and (11) fail to predict the increase of C_D after the end of the critical flow regime, and the start of the supercritical flow regime at which the boundary layer attached at the surface of the sphere changes from being partly laminar to being fully turbulent. This failure is attributed to the fact the training data used to obtain Eqs. (8) and (11) are far from the critical flow regime. Predicting C_D for the critical flow regime is difficult even for high fidelity solvers. For example, the non-optimized (Nonopt) solver of Geier et al. (2017) failed to predict the drag crisis. Instead, it predicts that C_D does not change with *Re*, similar to what the correlation of Brown and Lawler (2003) predicts. Eq. (8) and (11) performs better in the critical regime than the fitting correlation of Morrison (2013) which is a result of fitting experimental data from the literature. Another interesting observation is that the rate of change of C_D with *Re* in the critical flow regime, for both Eq. (8) and (11), follows the smooth trend similar to the experiments of Maxworthy (1969) and the high fidelity simulations of Geier et al. (2017), rather than the sharp nearly discontinuous change of C_D observed in the experiments of Achenbach (1972),Suryanarayana et al. (1993), Wieselsberger (1922).

To illustrate further the performance of the power-based Eq. (8), and that of the logarithmic Eq. (11) we will compare their predictions with the experimental results of Suryanarayana et al. (1993), and those of Achenbach (1972) as shown in Tables 5 and 6, respectively. The comparison shows that the logarithmic based Eq. (11) performance is superior to that of the power-based Eq. (8) since the relative error metric is always greater than that of the logarithmic-based Eq. (11), and always in the two-digit range. The power-based equation Eq. (8) only outperforms that of the logarithmic equation Eq. (11) in the thin diverging region. However, the overall performance of the logarithmic based Eq. (11) is better than the power-based equation Eq. (8) in the critical flow regime if we compare all their predictions with the available numerical and experimental data, as shown in Fig. 5. The power-based equation Eq. (8) perform better compared to its performance at the low Reynolds number regime. However, it still can not generalize well enough compared to logarithmic based equation Eq. (11). For this reason, we concluded that power-based equations cannot describe the evolution of the physics of the drag coefficient in different flow regimes.

Both Eqs. (8) and (11) predict different stages of the critical flow regime with surprising accuracy. They are the first in literature to make such predictions without being exposed to the critical flow regime, but only by using a limited amount of physics stored in the training data and the imposed functional forms. The question may arise whether these predictions are just a product of chance? Our short answer is no, for several reasons. The first reason is that the *Re* number changes by orders of magnitude in the critical flow regime, which gives many possibilities for the output of the predictive function, but Eq. (11) predicts with small error the experimental results of Suryanarayana et al. (1993) concerning the onset of the critical flow regime. The same applies to the Re_{cr} predicted by Eq. (8) compared to the experimental results of Achenbach (1972). The second and more supportive reason is that symbolic

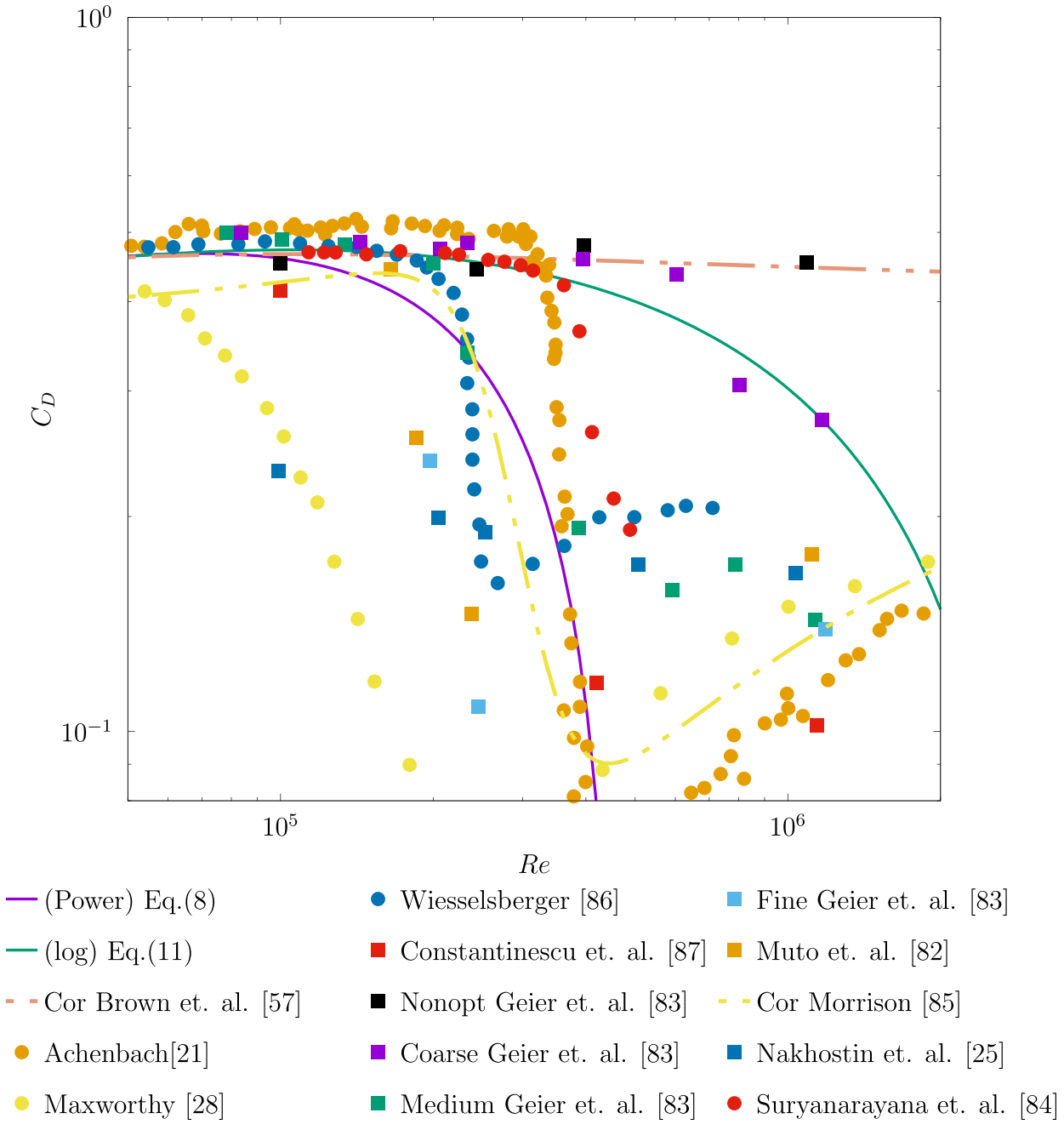

Fig. 5. Comparison between the C_D predictions by Eq. (8) and (11), and different sources in the high Re regime where the drag crisis occurs.

Table 5
Comparison between the values of the drag coefficient C_D from our predictive equations Eq. (8) and (11) with the experimental values of Suryanarayana et al. (1993) for different Re values.

Re	Suryanarayana et al. (1993)	Eq. (8)	Error %	Eq. (11)	Error %
2.1×10^5	0.468	0.367	21.5	0.447	4.4
2.5×10^5	0.457	0.313	31.5	0.451	1.3
2.7×10^5	0.454	0.287	36.7	0.447	1.5
2.9×10^5	0.449	0.259	42.3	0.443	1.3
3.1×10^5	0.441	0.240	45.4	0.440	0.2

regression can generalize and predict the approximated function's unexpected behaviour, similar to the example shown in Appendix B about the $\sin(x)$ approximation. The algorithm was trained to predict the peaks; however, it also accurately predicts the existence of valleys. We strongly believe that Eq. (11) contains terms that approximate the fundamental physical law that C_D is following, which is why it managed to generalize both the Stokes and critical flow regimes. This makes the logarithmic representation

Table 6
Comparison between the values of the drag coefficient C_D from our predictive equations Eq. (8) and (11) with the experimental values of Achenbach (1972) for different *Re* values.

Re	Achenbach (1972)	Eq. (8)	Erorr %	Eq. (11)	Error %
2.0×10^5	0.502	0.373	25.6	0.461	8.16
2.3×10^5	0.489	0.339	31.0	0.453	7.27
3.3×10^5	0.450	0.202	55.1	0.434	3.4
3.4×10^5	0.388	0.196	49.38	0.433	-11.8

of C_D a serious candidate of an analytical mathematical formulation that governs the variation of C_D with the *Re*.

In summary, we showed that the functional form of C_D could be represented by both powers and logarithmic functions of *Re*. However, the logarithmic representation conveys the physics in a different way than the power representation, and illuminates new physical phenomena, which are beyond the reach of current analytical or empirical C_D formulas. Because of the logarithmic equations' good generalization behaviour, especially Eq. (11), such equations should not be considered as merely fitting equations, but rather as semi-analytical equations. When appealing to mathematical aesthetics, our results suggest that the drag coefficient of a sphere might be well described by the form $C_D = \pi + 24/Re + f(\log Re)$, with $C_D = \pi$ at the first point of separation, occurring at a Reynolds number Re_s given by the transcendental equation $24/Re_s + f(\log Re_s) = 0$. Van Dyke (1975) described the appearance of logarithms in the asymptotic expansions as obscure, but it appears that these obscure entities can speak the language of fluid dynamics much better than powers. A similar situation exists in the field of turbulence, especially regarding channel flow, where there is an open debate in the scientific community whether power or logarithmic expansions best describe the velocity at the wall in certain flow regimes (Schultz and Flack, 2013). Note that the logarithmic dependence of the drag coefficient C_D also exists for geometries different than a sphere such as spherocylinders and prolate spheroids, as shown in our previous work (El Hasadi and Padding, 2019).

3.3. Drag coefficient for an oblate spheroid

This subsection is a continuation of testing the generalization behaviour of the derived equations. This time we will test their ability to predict the drag coefficient of a different geometry than a sphere. We selected the single geometry of an oblate spheroid with an aspect ratio p_a equal to 0.25 for the case that the flow is parallel to the equatorial diameter of the particle. We selected this specific geometry because the oblate spheroid and sphere are both parts of the geometrical space of spheroids, which makes their drag coefficient formulas share many similarities between them. In addition, we can compare our predictions with the direct numerical simulation results of Sanjeevi et al. (Sanjeevi et al., 2018) for a wide range of Reynolds numbers.

In line with the compression of the oblate geometry, we will modify Brown and Lawler (Brown and Lawler, 2003) correlation, power-based Eq. (8), and logarithmic based Eq. (11) equations. We will follow the work of Livi et al. (Livi et al., 2022), and we will multiply Brown and Lawler (Brown and Lawler, 2003) correlation, and Eq. (8) with the correction factor *K* for the Stokes drag derived by Happel and Brenner (2012), we leading to the following modified equations:

$$C_D = K\left(\frac{24}{Re}\left(1.0 + 0.15Re^{0.681}\right) + \frac{0.407}{1 + \frac{8710}{Re}}\right) \tag{18}$$

$$C_D = K\left(a_1 + \frac{a_2}{Re} + a_3\sqrt{Re} + \frac{a_4}{\sqrt{Re}} + \frac{a_5}{(a_6 + Re)} + a_7 Re\right) \tag{19}$$

The *Re* here after in this section is based on the diameter of equivalent sphere, the value of *K* for the specific particle geometry selected is 1.083. For the logarithmic based Eq. (11) we will multiply the geometry correction factor only with the Stokes term following our findings in our previous work (El Hasadi and Padding, 2019) for non-spherical particles. The new modified logarithmic equation for the oblate geometry has the following form:

$$C_D = a_1 + \frac{Ka_2}{Re} + a_3\log(Re) + a_4\log^2(Re) + a_5\log^4(Re) \tag{20}$$

The performance of the specially modified equations (Eqs. (18)–(20)) are tested against the numerical simulations of Sanjeevi et al. (Sanjeevi et al., 2018) as shown in Fig. 6. The predictions of the logarithmic modified equation (Eq. (20)) is closer to the numerical results of Sanjeevi et al. (2018) compared to the modified Brown and Lawler Eq. (18), and the modified power-based equation Eq. (19), at the moderate Reynolds numbers, while at high *Re*, the predictions of the three equations are similar. However, if we slightly modify further Eq. (20) by changing the value of the coefficient a_1 by 4.5% from 3.28 to π we will get the following equation:

$$C_D = \pi + \frac{Ka_2}{Re} + a_3\log(Re) + a_4\log^2(Re) + a_5\log^4(Re) \tag{21}$$

Eq. (21) predicts the numerical results of Sanjeevi et al. (2018) closely for the whole range of the *Re* and confirms that the logarithmic based equation Eq. (11) that we found for a sphere can be easily extended to predict the drag coefficient for a non-spherical geometry. We assigned π for the value of coefficient a_1 because, as mentioned in the previous section, we suspect that the value of a_1 converges toward π. This generalization behaviour shows that logarithms can represent the evolution of the drag coefficient beyond the spherical geometry. Eq. (21) is the first in literature that is capable to predict the drag coefficient for particle geometry without using experimental or numerical data from that specific geometry. We are also preparing a new study for the applicability of using Eq. (11) to predict the drag coefficient for non-spherical particles.

The worst performing predictive equation is the general-purpose correlation of Hölzer and Sommerfeld (2008). It predicts the results of Sanjeevi et al. (2018) for *Re* up to 100. After that, its predictions diverge significantly from the numerical results of Sanjeevi et al. (2018). The Hölzer and Sommerfeld (2008) correlation is derived to predict the drag coefficient of arbitrary particle geometry by fitting the drag coefficient data from different particle geometries, including disks and spheres. In principle the Hölzer and Sommerfeld (2008) correlation was exposed to a more extensive data set of non-spherical geometries than Eqs. 20,21.

We illustrate the accuracy of the predictive equations for the oblate particle geometry in Table 7. The worst predictive logarithmic-based equation Eq. (20) is about 10% more accurate compared to that of the power-based Eq. (19), and the modified Brown and Lawler (2003). Imposing a change of only 4.5% in the

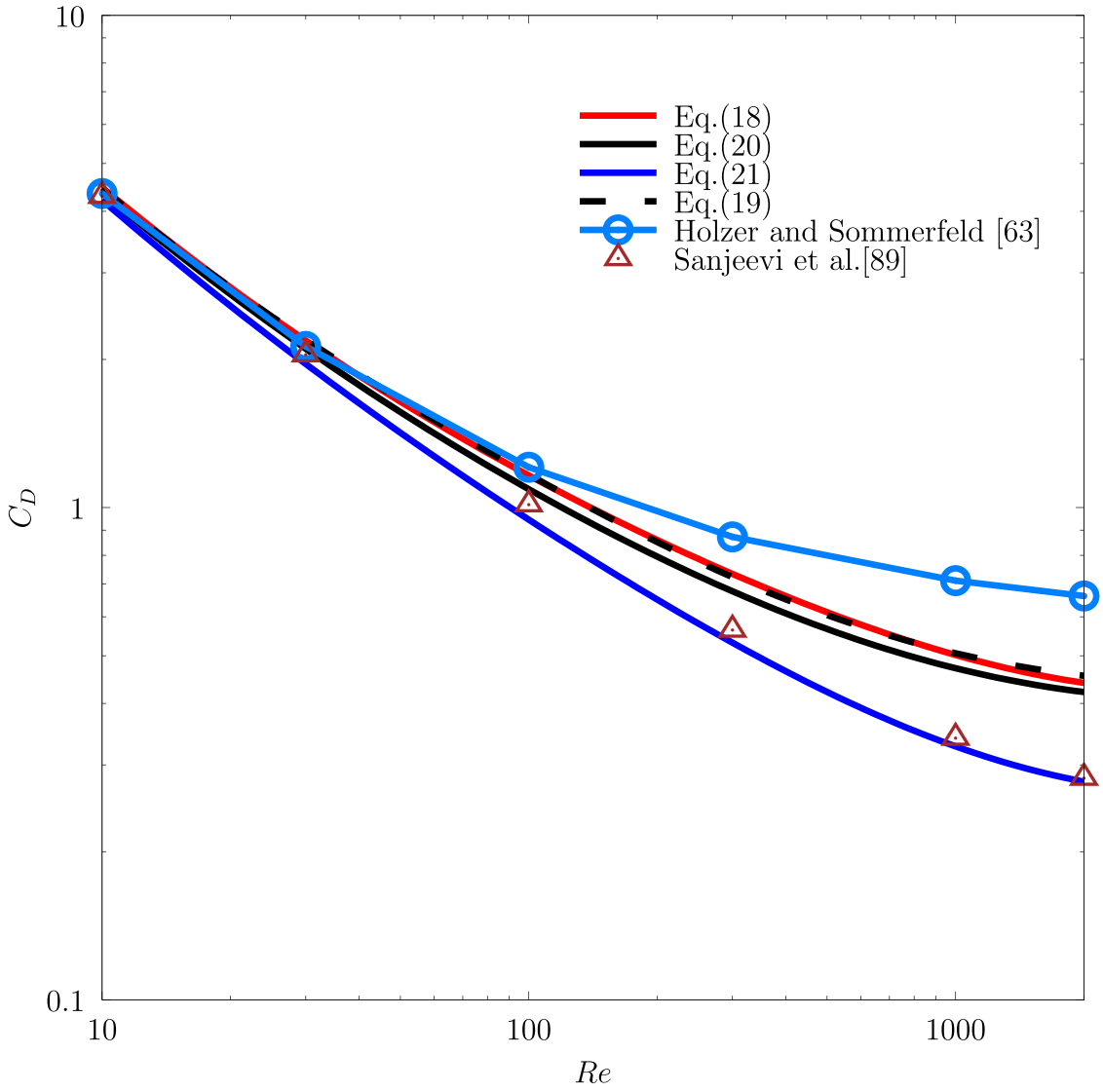

Fig. 6. Comparison between the C_D predictions by Eqs. (18), (20), (21) and (19), Holzer and Sommerfeld (Hölzer and Sommerfeld, 2008) correlation, and the numerical results of Sanjeevi et al. (Sanjeevi et al., 2018) for the case of an oblate spheroid with aspect ratio(pa) of 0.25, and a wide range of Reynolds numbers.

Table 7
Comparison between the drag coefficient values C_D for the oblate particle geometry p_a =0.25 from the numerical simulations of Sanjeevi et al. (2018) with those from derived predictive equations for different *Re* values.

Re	Sanjeevi et al. (2018)	Eq. (18)(%)	Eq (19)(%)	Eq. (20)(%)	Eq. (21)(%)
1000	0.340	0.501 (47%)	0.505(49%)	0.471 (39%)	0.327(4%)
2000	0.282	0.440 (57%)	0.455(61%)	0.421 (49%)	0.277(2%)

value of the a_1 coefficient in the logarithmic based equation Eq. (20) enhances the accuracy substantially by about 46%. The accuracy enhancement is far greater than the change in the coefficient's value. It supports our argument that the logarithmic-based equations far better interpret the physics of the drag coefficient around particles of different geometries compared to power-based equations. We also changed the values of the a_1 coefficient in Eq. (19), and the value of the constant of 0.407 in Eq. (18) by 4.5% but their predictive behaviour did not change compared to the original equations. In order for Eqs. (18) and (19) to have similar predictive accuracy as the logarithmic Eq. (21), the constant 0.407 and the coefficient a_1 their values must change by 75%, and 60%, respectively. For both equations, the change in the coefficients is substantially bigger than their predictive accuracy in their unchanged form, which could indicate that their predictions could be a result of over-fitting.

4. Conclusions

In this investigation, we explored the possibility of a logarithmic dependence of the drag coefficient C_D on the Reynolds number *Re* inspired by asymptotic solutions for creeping flow conditions. We used a symbolic regression machine learning algorithm, and our training data are based on experiments and data from well-known empirical correlations available in the literature. We can make the following conclusions:

- The drag coefficient C_D can be expressed as a function of powers in $\log(Re)$, partially fulfilling the Proudman and Pearson (1957) conjecture **P&P**.
- If an expansion in terms of $\log(Re)$ is made for the drag coefficient C_D, the value of the Re at which all the Re dependent terms go to zero is closely resembling the Re at the first emergence of laminar separation, as predicted analytically by Proudman and Pearson (1957).
- The logarithmic dependence of C_D on Re is found independently, without any prior knowledge, by the symbolic regression algorithm.
- The logarithmic based Eq. (11) can generalize in both low and high Re regimes. In the high Re regime Eq. (11) can predict the drag crisis, its results closely following experimental and numerical predictions from literature.
- Since Eq. (13) is derived from the experimental data of Brown and Lawler (2003), the appearance of the logarithmic terms in C_D equations is independent of the correlation that is used as a source of the training data.
- Eq. (11) is the only equation in literature that correctly describes the physics of the drag coefficient from the zero Re regime to the onset of the drag crisis, even though it is derived from data that does not include any information about those two regimes.
- Power-based equations, such as Eq. (8) and (9), do not describe the drag coefficient evolution for the entire Re range. For this reason, we believe they only represent a reasonable approximation of the real solution.
- We recommend using Eq. (11) for engineering applications, due to its generalization behaviour and following the appropriate physics in the low and high Re regimes. Also, we believe that the logarithms reassemble a functional form for the drag coefficient that can be applied to other particle shapes.

The bigger picture of our results is that, although our method cannot give answers as rigid mathematical proofs, it is highly probable that if one day we manage to solve in a closed form the Navier–Stokes equations, around a sphere, this solution will be expressed in terms of logarithms rather than powers. The logarithmic terms that symbolic regression found are related to the velocity and pressure fields around the sphere. Symbolic regression is an excellent candidate to further investigate the functional form of these fields, and we intend to conduct a future study toward this goal. Finally, we note that the machine learning framework that we developed is general and can be used in different scientific disciplines with the condition that experimental and numerical data exists, plus the availability of some limited analytical solutions.

Data Availability

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

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The first author thanks Dimitra Damianidou for the enlightening discussions about the subject. The authors thank Lorenzo Botto for the discussion about matched asymptotic methods. Finally, the authors thank the European Research Council for its financial support under its consolidator grant scheme, contract No. 615096 (NonSphereFlow).

Appendix A

Nusselt number Nu

In this Appendix will explore the possibility of a logarithmic dependence of the Nusselt number Nu on the Peclet number Pe and Reynolds number Re. First, we will briefly describe the available literature, and after that, we will explain our results thoroughly.

Concerning the heat transfer rate from a particle fixed in a fluid, most investigations available in the literature are related to the case of forced convection. In this type of flow, the velocity profile is decoupled from that of the temperature. For further simplification, there is also no variation in the transport properties of the fluid with temperature. These simplifications pave the way of obtaining several analytical solutions for a single sphere (Acrivos and Taylor, 1962) for limited cases of low Re and Peclet number $Pe = v_\infty d/\alpha$, where α is the thermal diffusivity of the fluid. Acrivos and Taylor (1962) used asymptotic expansions and the velocity profile of the Stokes solution to find the following relation for the Nusselt number $Nu = hd/k$, where h is the (convective and surface mean) heat transfer coefficient and k is the thermal conductivity of the fluid (linked to the thermal diffusivity through $k = \alpha\rho c_p$, with c_p the specific heat capacity of the fluid), for the case of $Pe \to 0$ and $Re \to 0$:

$$Nu = 2 + \frac{1}{2}Pe + \frac{1}{4}Pe^2\log(Pe) + 0.034Pe^2 + \frac{1}{16}Pe^2\log(Pe) \tag{A.1}$$

In practice, this solution is limited to $Re \lesssim 0.03$. Rimmer (1968) added an extra term to Eq. (A.1) from asymptotic expansions, and as far as we know this is the last term that evolved from the matched asymptomatic expansions in the low Pe and $Re \to 0$ regime. Conversely, for $Pe \to \infty$ and $Re \to 0$, Acrivos and Goddard (1965) used the matched asymptotic expansions to arrive at the following relation for Nu:

$$Nu = 0.922 + 1.249Pe^{\frac{1}{3}} \tag{A.2}$$

As for the case of drag, for higher Re we need to rely on semi-empirical relations to express the variation of Nu with the flow field parameters. Whitaker (1972) provided a correlation, which is still considered one of the most accurate available in literature (Sparrow et al., 2004):

$$Nu = 2 + \left(C_4Re^{a_1} + C_5Re^{a_2}\right)Pr^{a3} \tag{A.3}$$

where $Pr = c_p\mu/k$ is the Prandtl number (note that $Pe = RePr$). The values of a_1, a_2, and a_3 are $\frac{1}{2}, \frac{2}{3}$, and 0.4, respectively. The Whitaker correlation is valid for $1 \leqslant Re \leqslant 10^5$ and a wide range of Pr. The second, and third terms represent inertial fluid effects, and their functional form is inspired by boundary layer theory. Although the first term comes from the analytical solution for pure conduction from a sphere, all exponents in Eq. (A.3) are obtained from empirical fitting.

For the purpose of evaluating the existence of logarithmic terms for the problem of convective heat transfer over a sphere. We will create a data set of 26,796 points from the Whitaker (1972) correlation Eq. (A.3) for Pr in the ranging from 0.74 to 7.0, and Re in the range of 10^{-1} to 10^4. We will start with the simplest assumption by allowing the symbolic regression algorithm to guess about the dependency of Nu on Re, Pr and, or Pe, through the following initial function:

$$Nu = f(Re, Pr, Pe) \tag{A.4}$$

The resulting Nu correlation is the following:

$$Nu = a_1 + a_2\sqrt{Pe} + a_3\sqrt{Re}\sqrt{a_4 + a_5\sqrt{Pe}} + a_6 Pe + a_7 Re \tag{A.5}$$

The coefficients are listed in Table A.1. Most equations that the algorithm produces show that Nu is a function of Re and Pe, and excludes the explicit dependence on Pr. This is different from the source of our data (the Whitaker correlation Eq. (A.3)), which explicitly depends on Pr and Re. Even when we used a substantial amount of data, the algorithm failed to predict the exact structure of the Whitaker correlation (Whitaker, 1972). The recent investigation of Udrescu and Tegmark (2020) showed, consistent with our results, that Eureqa failed to predict the exact functional structure of many functions included in the Feynman lectures (Feynman et al., 1965). They attributed this failure due to the complexity of those functions, and the number of variables that they contain.

Examining the properties of Eq. (A.5), we find that as $Re \to 0$, Eq. (A.5) reduces to $a_1 + a_2\sqrt{Pe}$, which bears similarities with Eq. (A.2) for the Pe dependency, because for both cases the power of Pe is less than one, and both equations show that even at very low Re convection affects the heat transfer rate. This type of dependency did not exist in the Whitaker correlation Eq. (A.3), where for $Re \to 0$ (outside the range of validity of the Whitaker correlation) Nu converges to a value of 2.0, corresponding to pure conduction from a single sphere.

We will now examine the full dependence of Nu on logarithms of Pe, Re, and Pr. This structure of dependency is based on our previous knowledge of the physics of the problem of forced convection over a sphere. We know that for $Re \to 0$ and $Pe < 1, Nu$ depends on $\log(Pe)$ (Acrivos and Taylor, 1962) (Eq. A.1), so there may exist an intermediate Pe regime where logarithms will play a role as well, until we reach a high Pe regime where Eq. (A.2) is dominant. For the high Re regime we already showed that the drag coefficient $C_D$is a function of logarithms of Re, so because of the tight relation between flow and heat transfer (Duan et al., 2015) we expect that logarithms of Re will play a role in the convective heat transfer process as well. The initial function has the following form:

$$Nu = f\left(\log(Pe), Pe\log(Pe), \log^2(Pe), \log(Re), Re\log(Re), \log^2(Re), \log(Pr), Pr\log(Pr), \log^2(Pr)\right) \tag{A.6}$$

As initial guess we gave equal weight to all functional forms, to avoid any bias, toward any of the independent variables. The symbolic regression algorithm found the following two correlations:

$$Nu = a_1 + a_2\log^2(Re)\log(Pe)Pe^{a_3} + a_4 Pe^{a_5} \tag{A.7}$$

$$Nu = a_1 + a_2\log^2(Re) + a_3 Pe^{a_4} + a_5\log^2(Re)\log(Pe)Pe^{a_6} + a_7\log(Pe) \tag{A.8}$$

The second equation is more complex than the first. The coefficients of both Eqs. (A.7), and (A.8) are listed in Table A.1. Both equations posess very interesting features. We will start with Eq. (A.8), where the term $a_1 + a_3 Pe^{a_4}$ resembles closely the approximation of Eq. (A.2). The relative difference of the a_1, a_3 coefficients and those of Eq. (A.2) is 15%, and 8%, respectively. The relative error is remarkably small, if we take into account that the source of the data set is coming from an empirical correlation that has an average predictive error of 30%. The logarithmic Eqs. (A.7), and (A.8) follows the second and the third rules, similar to their drag coefficient counterparts. On the other hand, the power-based based Eq. (A.5) fails to extrapolate, especially in the low Re regime.

Table A.1
Coefficients for Eqs. (A.5), (A.7), and (A.8).

Coefficients	Eq. (A.5)	Eq. (A.7)	Eq. (A.8)
a_1	2.0	1.582	1.063
a_2	0.343	0.003	0.0067
a_3	0.0454	0.326	1.351
a_4	9.341	1.0	0.299
a_5	1.0	0.322	0.0028
a_6	-7.0×10^{-5}	-	0.332
a_7	−0.00131	-	−0.128

We believe that the combination of the logarithmic dependence of Pe and Re plays an essential role in the emergence of an asymptotic solution. It seems there are very few possible ways to represent the data of Whitaker (1972) using logarithms of Pe and Re and one of those few is using terms similar to Eq. (A.2). Our findings show the essential role played by previous physical knowledge of the problem in specific regimes, to help the machine learning algorithm to reach a physically meaningful result.

The genetic algorithm predicted the asymptotic solution for the high Pe (Eq. A.2) case, rather than for low Pe (Eq. A.1), probably because our training data is more biased toward the high Pe regime. Since the lowest Re and Pr used are 0.1 and 0.7 respectively, the lowest Pe we used is 0.07, which lies at the boundary of the high Pe regime. We could not use lower Pe because the Whitaker correlation (Whitaker, 1972) is based on Re ranging between 3.5 and 7.6×10^4, and Pr ranging between 0.7 and 380. Note that we did use the Whitaker correlation (Whitaker, 1972) also for lower Re, $0.1 < Re < 3.5$, to generate our training data. We test its validity against the experimental data of Will et al. (2017) for the lowest Prandtl number that we used, Pr = 0.7, and for Re as low as 0.1, and we found that the Whitaker correlation (Whitaker, 1972) follows closely the results of (Will et al., 2017), as shown in Fig. A.1. An indication that the hydrodynamics in the highly inertial regime may be governed by logarithmic terms of Re, is the the appearance of $\log^2(Re)$ terms both in Eqs. (A.7) and (A.8), similar to the case of C_D (see Eqs. (11), (13) and (14)). Also, the $\log^2(Re)$ terms for both Nu and C_D share the same sign, and their pre-factors are of the same order of magnitude.

We compare the performance of our predictor equations for different Pr, and Re numbers, in Fig. A.1. We select four cases, two of them lie within the training data set (Pr = 0.7 and 7.0) that we supplied to the algorithm. The other two test cases (Pr =50 and 300) lie outside the training data set to test the extrapolation capabilities of our predictor equations. For Pr = 0.7, Eqs. (A.5), (A.7) and (A.8) perfectly follow the Whitaker (1972) correlation and the experimental results of Will et al. (2017). At high Re they also follow the numerical results of Feng and Michaelides (2000). As expected, our ecosystem of equations do not follow the asymptotic solution of Acrivos and Goddard (1965) since their solution is only valid in the low Re and high Pe regime. For the case of Pr = 7.0, our ecosystem of equations predicts the evolution of Nu with great accuracy. For the cases of Pr = 50 and 300, Eqs. (A.7) and (A.8) predict with great accuracy the results of the Whitaker (1972) correlation, except in a very narrow region at low Re. The conditions in this low Re - high Pr regime are applicable to the asymptotic solution of Acrivos and Goddard (1965). This is why the whole ecosystem of our equations deviate from the results of the Whitaker (1972) correlation, and follow by different degrees of accuracy the asymptotic solution of Acrivos and Goddard (1965), Eq. (A.2). All of our equations are functions of Pe and Re. However, for low Re the Nu correlations switch to a dependency on Pe only, which is consistent with the physics of Eqs. (A.1) and (A.2).

If we perform a simple mathematical analysis on the Whitaker (1972) correlation such as taking the limit $Re \to 0$, and $Pr \to \infty$, which is the applicable range for the Acrivos and Goddard (1965) analytical solution, the result is 2.0. This shows that the Whitaker (1972) correlation does not contain any elements of the asymptotic solution of Acrivos and Goddard (1965). To further investigate the behaviour of our predictive equations in the range

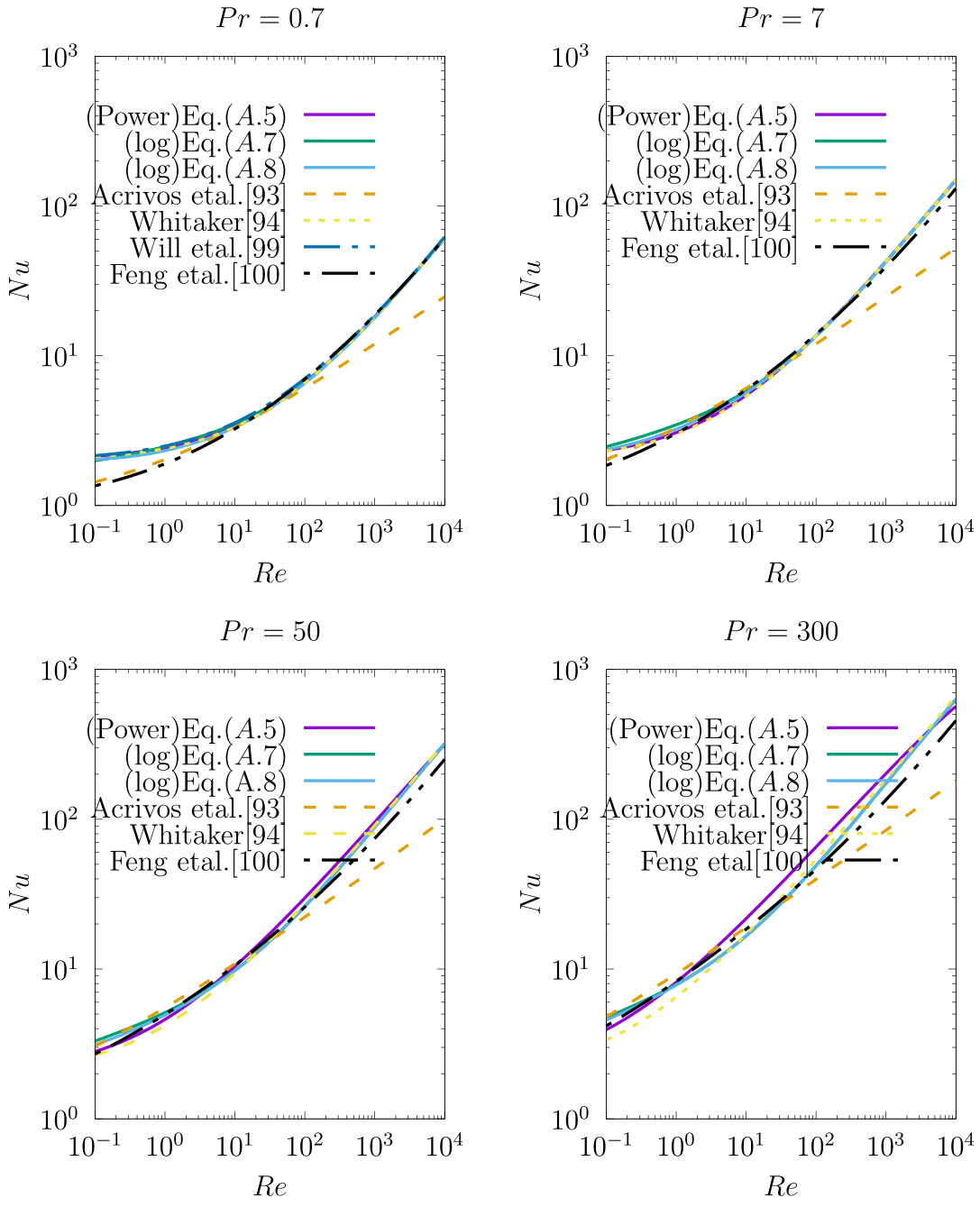

Fig. A.1. Comparison between the results of different predictor equations for the Nusselt number *Nu* with those from literature for four different Prandtl numbers *Pr*.

of applicability of the Acrivos and Goddard (1965) asymptotic solutions, we selected two cases with *Pr* equal to 1000, and 3000, and *Re* ranging between 10^{-2} to 10^{-1}, the results are shown in Fig. A.2. It is clear that the logarithmic based predictive equations are superior in their predictions compared to the power-based equations since they follow the solution of Acrivos and Goddard (1965) closely. Specifically, for the case of *Re* = 0.05, the *Nu* values from different sources are illustrated in Table A.2. For example the Whitaker (1972), correlation predictions differ by 35% to 41% with the respect to the Acrivos and Goddard (1965) predictions. The logarithmic based equations Eqs. (A.7), and (A.8) their values only differ only 0.2% to 4.0% from those of Acrivos and Goddard (1965). As for the power-based equation Eq. (A.5), the relative error is 19.14%, and 17% with respect to Acrivos and Goddard (1965) predictions. While, the predictions of Feng and Michaelides (2000) differ from those of Acrivos and Goddard (1965), by 14.09%, and 14.85%.

The high relative error of the Whitaker (1972) correlation indicates that it is not applicable in the range of *Re* and *Pr* where the asymptotic solution of Acrivos and Goddard (1965) is valid. This proves the correctness of the simple mathematical analysis that we did above. Thus, it is clear proof that in the training data that we use, there will be no trace of the Acrivos and Goddard (1965) asymptotic solution. Thus, the only reason for the high accuracy of the logarithmic based equations Eqs. (A.7) and (A.8) is that they describe the governing physics of the problem. This is similar to our conclusion on the drag coefficient. The inclusion of logarithmic

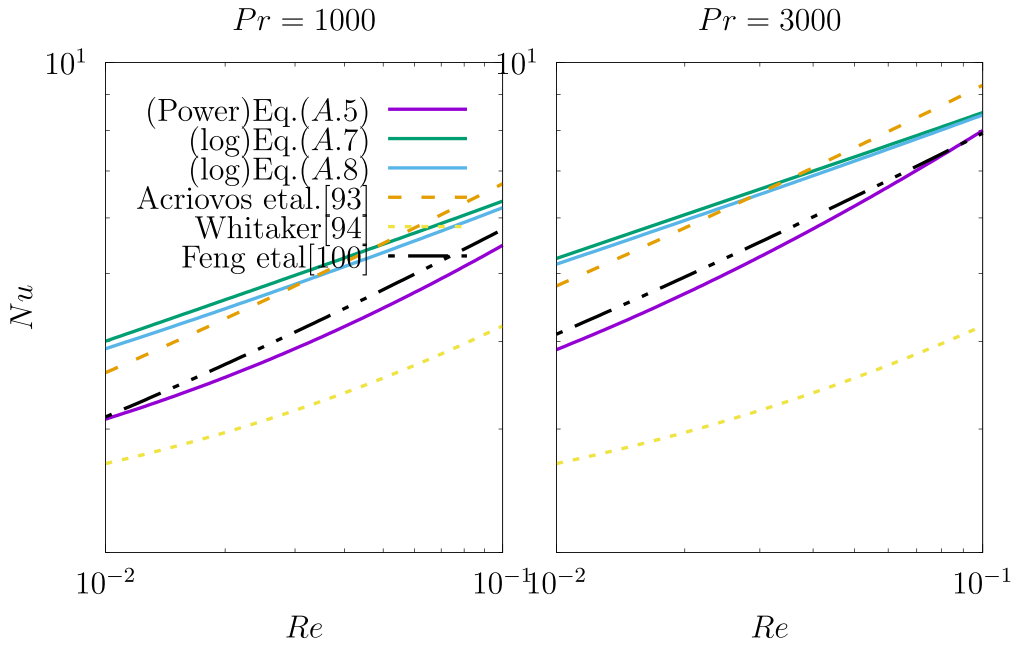

Fig. A.2. Comparison between the results of different predictor equations for the Nusselt number *Nu* with those from literature for four different at Prandtl *Pr*, numbers, and low Reynolds numbers *Re*.

Table A.2
Comparison between our predictive equations, and correlations from the literature with the results of Acrivos and Goddard (1965) for the cases of *Pr* = 1000, 3000, and *Re* =0.05.

	Pr =1000	Relative Error (%)	*Pr* = 3000	Relative Error (%)
Eq. (A.5)	4.46	-19.14	6.23	-17.41
Eq. (A.7)	5.05	−0.2	7.33	-2.93
Eq. (A.8)	5.35	-2.95	7.23	-4.25
Whitaker (1972)	3.546	-35.74	4.4	-41.74
Feng and Michaelides (2000)	4.79	14.09	6.43	-14.85

terms leads to a generalization of derived functions, leading to a description of physics not included in the training data. This leads us to conclude that the logarithmic terms could be part of the analytical solution of the two problems described in this investigation. A surprising observation is that the correlation of Feng and Michaelides (2000), which includes an approximate form of Acrivos and Goddard (1965) asymptotic solution, failed to make accurate predictions at high *Pe* numbers and shows how complex is the process of getting close to the predictions of the analytical solution of Acrivos and Goddard (1965). This shows the difficulty of capturing the asymptotic solution of Acrivos and Goddard (1965) even by using numerical data and parts of the asymptotic solution itself and supports our argument indirectly that the logarithmic terms represent part of the solution of the Navier–Stokes and energy equations.

The above shows that symbolic regression can find an asymptotic solution by using previous physical knowledge, rather than depending completely on the training data set. Feeding machine learning algorithms previous physical knowledge for the problem that they try to optimize, increases substantially the probability of better extrapolation predictions. For further discussion on how to implement previous knowledge into symbolic regression, the readers is referred to our recent publication (El Hasadi and Padding, 2019).

From this appendix we can make the following conclusions:

- The Nusselt number of a single sphere depends on logarithms of Re, Pe, as well as powers of *Pe*.
- If logarithmic functions of *Re* and *Pe* are used as initial functions for the symbolic regression algorithm, the algorithm produces with high accuracy the asymptotic solution derived by Acrivos and Goddard (1965) from the matched asymptotic method, in the low *Re* and high *Pe* regime. Interestingly, the training data that we used does not follow the asymptotic solution of Acrivos and Goddard (1965).
- There is a connection between the appearance of the logarithmic terms in both C_D and *Nu* expressions, and the ability of those expressions to generalize outside the training data range. This connection makes the logarithmic representation a strong candidate for the functional form of C_D and *Nu* that could result from solving the Navier–Stokes equations analytically for the problem of flow over a single sphere at high *Re*, and be a result of a generalized fluid mechanics theory that applies to both low and high *Re* regimes.

Appendix B

Maclaurin expansion of Sin function

A well-known result of applied mathematics is the representation of continuous functions by the Taylor expansion(Taylor, 1717):

$$f(x) = \sum_{n=0}^{\infty} \frac{f^n(a)(x-a)^n}{n!} \tag{B.1}$$

Table B.1
Coefficients for Eqs. (B.4) and (B.5).

Coefficients	Eq. (B.4)	Eq. (B.5)
a_1	0.9999	1.0001
a_2	0.1665	0.1682
a_3	0.00826	0.0031
a_4	0.000173	0.0065

When a = 0, the Taylor series reduces to the Maclaurin series. The following expansion gives the Maclaurin series for $\sin(x)$:

$$\sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots \tag{B.2}$$

One of the reasons we choose the $\sin(x)$ function as our test case for the symbolic regression algorithm is its non-monotonic nature, specifically its transition from an increasing to a decreasing function. This feature will help us assess the generalization behaviour of the algorithm. We generated 5000 uniform training points in the range $[0,\frac{\pi}{2}]$. We selected this specific range because we wanted to feed the algorithm only the monotonically increasing part of the $\sin(x)$ function, and see if it can generalize, and predict the decreasing part of the function between $[\frac{\pi}{2}, \pi]$. The algorithm does not possess any prior knowledge of the $\sin(x)$ function and starts by assuming the most primitive initial function for the symbolic regression algorithm:

$$y = f(x) \tag{B.3}$$

The symbolic regression algorithm suggested many equations, including the following two:

$$y(x) = a_1 x - a_2 x^3 + a_3 x^5 - a_4 x^7 \tag{B.4}$$

$$y(x) = a_1 x - a_2 x^3 + a_4 x^4 + a_5 x^5 \tag{B.5}$$

The values of the coefficients of Eq. (B.4) and (B.5) are listed in Table B.1. Eq. (B.4) contains the first four terms of the Maclaurin series for the $\sin(x)$ function. Although this may seem to be trivial, to the best of our knowledge this is the first time that a machine-learning algorithm managed to derive a Taylor or a Maclaurin series out of pure data. For the derivation of any Taylor series of a function we need to use the calculus invented simultaneously by Newton (1833) and Leibniz (1682).

First, we want to illustrate the effect of the different terms of Eq. (B.4) on its accuracy and generalization, as shown in Fig. B.1. For the $[0,\frac{\pi}{2}]$ domain, except for the first linear term, regardless of the number of terms we add, the decreasing nature of $\sin(x)$ for $x > \frac{\pi}{2}$ is predicted. Adding more terms increases the accuracy. While the first three terms are enough to predict with great accuracy the training data, the fourth term plays a significant role for values of $x > \frac{\pi}{2}$ which is beyond the range of the training data. We chose Eq. (B.4) not only because of its accuracy but due to its resemblance of the Maclaurin series, thus our selection is based on our own previous knowledge. What is missing is a generalization theorem which can tell us about the generalization behaviour of a specific machine learning algorithm, trained at a specific range of data. Without this theorem, we

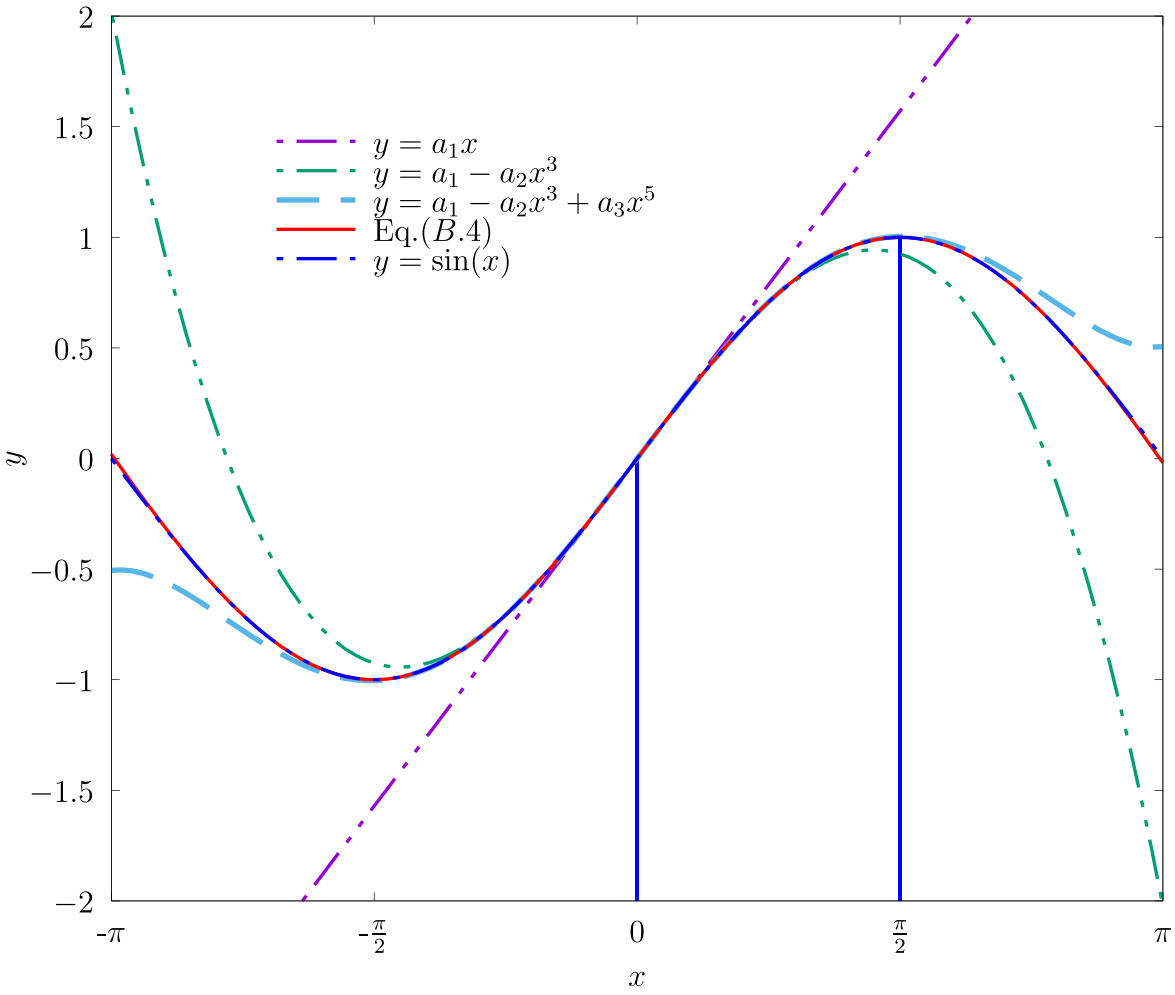

Fig. B.1. The influence of different terms of Eq. (B.4) on its variation with x. Blue bars indicate the range of training data.

Table B.2
Coefficients of polynomials of degree n = 3, and 7.

Coefficients	n = 3	n = 7
a_0	−0.002	-4.70×10^{-8}
a_1	1.027	1.0
a_2	−0.069	-2.339×10^{-5}
a_3	−0.138	−0.166
a_4	-	-2.45×10^{-4}
a_5	-	0.008
a_6	-	-2.046×10^{-4}
a_7	-	-1.377×10^{-4}

will always be hesitant to use machine learning predictions beyond their training range, specifically when dealing with problems for which we have minimal knowledge about the behaviour outside the training range. Finally, we want to compare the performance of the symbolic regression algorithm with other popular machine algorithms in literature, such as polynomial regression and artificial neural networks (ANN) for the same $\sin(x)$ case. Polynomial regression may be considered as one of the oldest machine learning algorithms (Brunton and Kutz, 2019), inspired by Legendre and Gauss's works, and implemented in a robust algorithm by Gregonne in 1815 (Stigler, 1974). Polynomial regression is the most appropriate "traditional" regression method to arrive at polynomials such as the Maclaurin series. In polynomial regression, the structure of the fitting equation and the degree of the polynomial are predefined. For our case we will use two different polynomials one with a degree of n = 3, and other one with n = 7. We use the same training data set that we used for the symbolic regression, and for implementation, we will use the Polyfit function from the open-source Numpy library written in python (www.numpy.org/doc/stable/user/, xxxx). The main output of the algorithm is the coefficients of the following equation:

$$y(x) = a_0 + a_1 x + \ldots a_n x^n \tag{B.6}$$

The coefficients for the two polynomials that we used are listed in Table B.2.

We selected the artificial neural network because it is considered as a universal function approximators (Cybenko, 1989; Hornik, 1991), but also because it does not need any prior knowledge about the structure of the equation to best fit the training data, similar to the symbolic regression algorithm. Contrary to symbolic regression, the product of a neural network approach is not a function but the trained neural network itself. We will use a feed-forward deep neural network, with eight hidden layers. The first hidden layer consists of 64 neurons, while, the remaining hidden layers contain 32 neurons, and finally an output layer containing a single neuron (Brunton and Kutz, 2019). In each hidden layer we use the Relu activation function, and also we apply L2 regularization to avoid overfitting. The algorithm minimizes the mean square difference between the predicted and training data, using a gradient descent algorithm. We use the open-source library TensorFlow (Abadi et al., 2016) to implement the artificial neural network framework. For training, we use 40,000 training points,

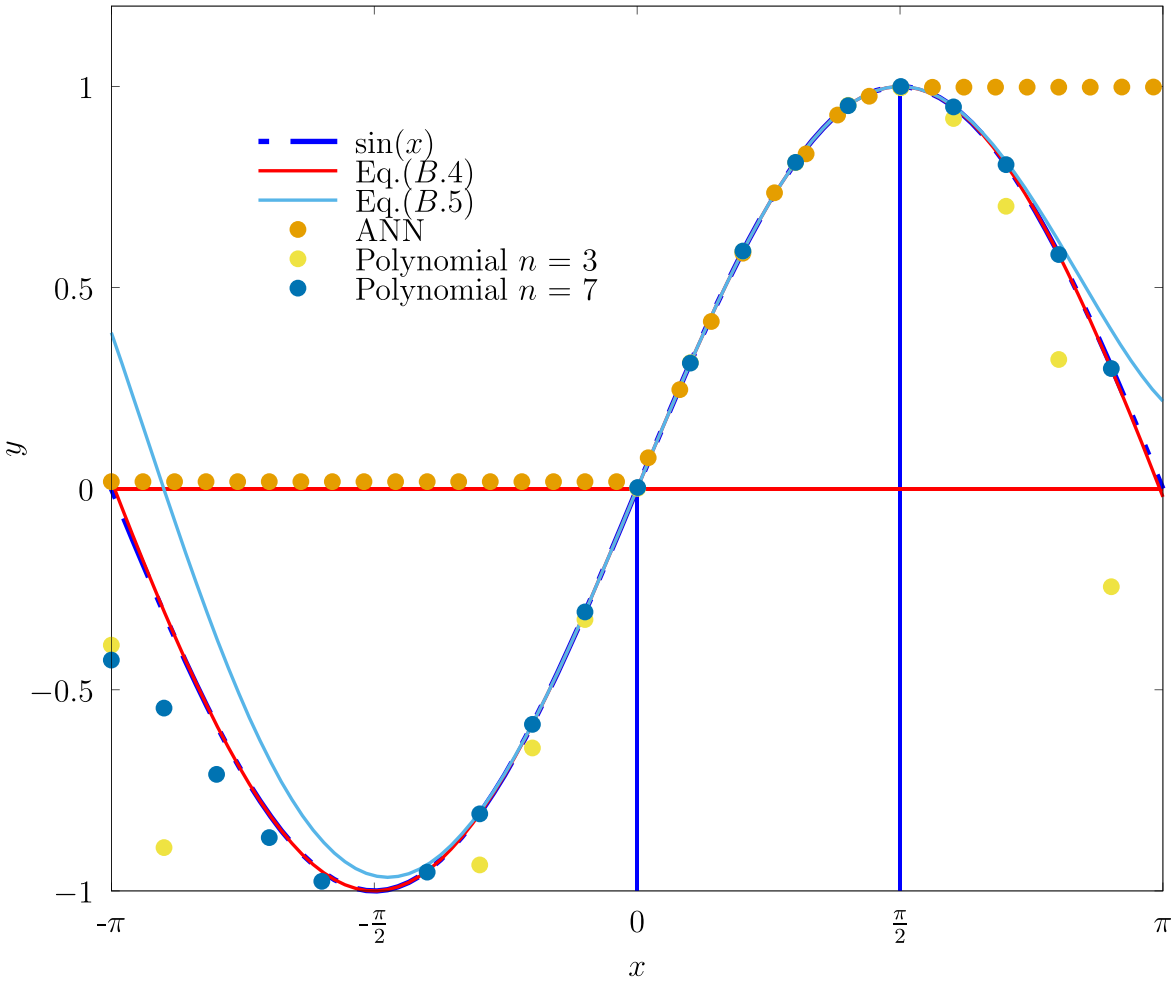

Fig. B.2. Comparison between different machine learning methods for the $\sin(x)$ example. Blue bars indicate the range of training data.

which is a much higher volume compared to the other two algorithms, because deep neural networks require a large amount of data to be trained appropriately(Brunton et al., 2020).

A comparison between the performance of the three algorithms is shown in Fig. B.2. Symbolic regression and polynomial regression were the only algorithms that predict the peaks and valleys of the $\sin(x)$ function within the range of $[-\pi, \pi]$. This success can be attributed to the fact that both algorithms represent the $\sin(x)$ function as a polynomial. For the case of the symbolic regression, it discovered the polynomial representation by itself. On the contrary, the ANN failed to generalize beyond the training data. We hoped that by making the network deeper, we could help the network extract sufficient features from the training data, and generalize. However, what we observe is that the ANN memorizes the training data instead of generalizing it. For example for $x > \frac{\pi}{2}$ the output of the ANN is always a constant value of one, which is the value of $\sin(\frac{\pi}{2})$, and for $x < 0$ the output of the ANN is always a constant value of zero, which is the value of $\sin(0)$. This type of memorization by an ANN is also observed in several other studies such as Zhang et al. (2016). Also, the work of Kim et al. (2020) showed that if feed-forward ANN is integrated with symbolic regression, one obtains a better generalization behaviour compared to pure ANN. Another interesting observation is that despite the fact that both symbolic regression and ANN optimize the mean square difference, they come up with totally different generalization behaviour.

This Appendix showed that symbolic regression can generalize beyond the training data, and can predict a change in the original function occurring beyond the training range. This shows the usefulness of using interpretable machine learning results, as recommended by Rudin (2019), and it helps us understand the output function behaviour within and beyond the training range.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al., 2016. arXiv preprint arXiv:1603.04467.
Abraham, F.F., 1970. Phys. Fluids 13 (8), 2194–2195.
Achenbach, E., 1972. J. Fluid Mech. 54 (3), 565–575.
Acrivos, A., Goddard, J., 1965. J. Fluid Mech. 23 (2), 273–291.
Acrivos, A., Taylor, T.D., 1962. Phys. Fluids 5 (4), 387–394.
Aoi, T., 1955. J. Phys. Soc. Jpn. 10 (2), 119–129.
Batchelor, G., 2000. An introduction to fluid dynamics. Cambridge University Press, Cambridge, UK.
Beard, K., Pruppacher, H., 1969. J. Atmos. Sci. 26 (5), 1066–1072.
Beratlis, N., Balaras, E., Squires, K., 2019. J. Fluid Mech. 879, 147–167.
Blasius, H., 1908. Grenzschichten in Flüssigkeiten mit kleiner Reibung. University of Gottingen, Germany.
Bourot, J., 1969. A 269, 1017–1020.
Brauer, H., Mewes, D., 1972. Chem. Ing. Tech. 44 (13), 865–868.
Breach, D., 1961. J. Fluid Mech. 10 (2), 306–314.
Brown, P.P., Lawler, D.F., 2003. J. Environ. Eng. 129 (3), 222–231.
Brunton, S.L., Kutz, J.N., 2019. Data-driven science and engineering: Machine learning, dynamical systems, and control. Cambridge University Press.
Brunton, S.L., Proctor, J.L., Kutz, J.N., 2016. Proc. Natl. Acad. Sci. 113 (15), 3932–3937.
Brunton, S.L., Noack, B.R., Koumoutsakos, P., 2020. Annu. Rev. Fluid Mech. 52, 477–508.
Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., Walsh, A., 2018. Nature 559 (7715), 547–555.
Chang, E.J., Maxey, M.R., 1994. J. Fluid Mech. 277, 347–379.
Chester, W., Breach, D., Proudman, I., 1969. J. Fluid Mech. 37 (4), 751–760.
Clift, R., 1970. Proc. Chemeca' 701, 14.
Colen, J., Han, M., Zhang, R., Redford, S.A., Lemma, L.M., Morgan, L., Ruijgrok, P.V., Adkins, R., Bryant, Z., Dogic, Z., et al., 2016. Proc. Natl. Acad. Sci. 118(10), 3932–3937.
Constantinescu, G., Pacheco, R., Squires, K., 2002. In: 40th AIAA Aerospace Sciences Meeting & Exhibit, 425.
Cox, R., 1965. J. Fluid Mech. 23 (4), 625–643.
Cybenko, G., 1989. Math. Control, Signals Syst. 2 (4), 303–314.
Darwin, C., 1859. On the origin of species. Johon Murray, London, UK.
Dennis, S., Walker, J., 1971. J. Fluid Mech. 48 (4), 771–789.
Deshpande, R., Kanti, V., Desai, A., Mittal, S., 2017. J. Fluid Mech. 812, 815–840.
Dissanayake, M., Phan-Thien, N., 1994. Commun. Numer. Methods Eng., 10(3), 195–201.
Domingos, P., 1999. Data Min. Knowl. Discov. 3 (4), 409–425.
Duan, Z., He, B., Duan, Y., 2015. Scient. Rep. 5, 12304.
El Hasadi, Y.M., Padding, J.T., 2019. AIP Adv. 9 (11), 115218.
Engelund, F., Hansen, E.A., 1967. monograph on sediment transport in alluvial streams. TEKNISKFORLAG Skelbrekgade 4 Copenhagen V, Denmark..
Faxen, H., 1923. Arkiv for Matemetik Astronomi och Fysik 17, 1–28.
Feng, Z.-G., Michaelides, E.E., 2000. Int. J. Heat Mass Transf. 43 (2), 219–229.
Feynman, R.P., Leighton, R.B., Sands, M., 1965. Am. J. Phys. 33 (9), 750–752.
Flemmer, R.L., Banks, C., 1986. Powder Technol. 48 (3), 217–221.
Geier, M., Pasquali, A., Schönherr, M., 2017. J. Comput. Phys. 348, 889–898.
Goldstein, S., 1929. Proc. Roy. Soc. London. Ser. A, Containing Papers Math. Phys. Character 123 (791), 225–235.
Goldstein, S., 1965. Aeronautical Research Council (Great Britain).
Goossens, W.R., 2019. Powder Technol. 352, 350–359.
Graebel, W., 2007. Advanced fluid mechanics. Academic Press, Burlington, MA, USA.
Graf, W.H., 1984. Hydraulics of sediment transport. Water Resources Publication, Highlands Ranch, CO USA.
Happel, J., Brenner, H., 2012. Low Reynolds number hydrodynamics: with special applications to particulate media, volume 1. Springer Science & Business Media.
Holland, J.H. et al., 1992. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press, Cambridge, MA, USA.
Hollandbatt, A., 1972. Transactions Of The InstItution Of Chemical. Engineers 50 (1). 12–+.
Holzer, M., Kaper, T.J., et al., 2014. Adv. Diff. Eqs. 19 (3/4), 245–282.
Hölzer, A., Sommerfeld, M., 2008. Powder Technol. 184 (3), 361–365.
Hornik, K., 1991. Neural networks 4 (2), 251–257.
Hunter, C., Tajdari, M., Boyer, S., 1990. SIAM J. Appl. Math. 50 (1), 48–63.
Iten, R., Metger, T., Wilming, H., Del Rio, L., Renner, R., 2020. Phys. Rev. Lett. 124 (1), 010508.
Jenson, V., 1959. Proc. Roy. Soc. London. Ser. A. Math. Phys. Sci. 249 (1258), 346–366.
John Veysey, I., Goldenfeld, N., 2007. Rev. Mod. Phys. 79 (3), 883.
Kamble, C., Girimaji, S., 2020. Phys. Fluids 32 (10), 105110.
Kaplun, S., Lagerstrom, P., 1957. J. Math. Mech. 6 (5), 585–593.
Khair, A.S., Chisholm, N.G., 2018. J. Fluid Mech. 855, 421–444.
Khan, A., Richardson, J., 1987. Chem. Eng. Commun. 62 (1–6), 135–150.
Kim, S., Lu, P.Y., Mukherjee, S., Gilbert, M., Jing, L., Čeperić, V., Soljačić, M., 2020. IEEE Trans. Neural Networks and Learn. Syst., 2020.
Koza, J.R., 1992. Genetic programming: on the programming of computers by means of natural selection, volume 1. MIT press, Cambridge, MA, USA.
Kumar, A., Rehman, N.M., Giri, P., Shukla, R.K., 2021. J. Fluid Mech. 920.
Kushvaha, V., Kumar, S.A., Madhushri, P., Sharma, A., 2020. J. Compos. Mater. 54 (22), 3099–3108.
Lagerstrom, P., Reinelt, D., 1984. SIAM J. Appl. Math. 44 (3), 451–462.
Leal, L.G., 2007. Advanced transport phenomena: fluid mechanics and convective transport processes, volume 7. Cambridge University Press, Cambridge, UK.
Le Clair, B., Hamielec, A., Pruppacher, H., 1970. J. Atmos. Sci. 27 (2), 308–315.
Leibniz, G.W. Acta eruditorum, 467–473.
Lewis, J., Carrier, G., 1949. Q. Appl. Math. 7 (2), 228–234.
Livi, C., Di Staso, G., Clercx, H.J., Toschi, F., 2022. Phys. Rev. E 105 (1), 015306.
Maxworthy, T., 1965. J. Fluid Mech. 23 (2), 369–372.
Maxworthy, T., 1969. J. Appl. Mech. 36, 598–607.
Morrison, F.A., 2013. An introduction to fluid mechanics. Cambridge University Press.
Morsi, S., Alexander, A., 1972. J. Fluid Mech. 55 (2), 193–208.
Muto, M., Tsubokura, M., Oshima, N., 2012. Phys. Fluids 24 (1), 014102.
Nakhostin, S., Giljarhus, K., 2019. IOP Conference Series: Materials Science and Engineering, volume 700. IOP Publishing, p. 012007.
Newton, I., 1833. Philosophiae naturalis principia mathematica, volume 1. G. Brookman.
Nisi, H., Porter, A.W., 1923. The London, Edinburgh, and Dublin Philos. Magaz. J. Sci. 46 (275), 754–768.
Oseen, C.W., 1910. Arkiv Mat., Astron. och Fysik 6, 1.
Oseen, C.W., 1913. Ueber den gueltigkeitsbereich der stokesschen widerstandsformel. Friedländer.
Payard, M., Coutanceau, M., 1974. CR Academie des Sci., Paris B 278, 369–372.
Popović, N., 2005. J. Phys: Conf. Ser., volume 22. IOP Publishing, p. 164.
Popović, N., Szmolyan, P., 2004. Nonlinear Analysis: Theory. Methods & Applications 59 (4), 531–565.
Prandtl, L., 1904. In Verhandlg. III. Intern. Math. Kongr. Heidelberg, 484–491. Math Congress.
Proudman, I., Pearson, J., 1957. J. Fluid Mech. 2 (3), 237–262.
Raissi, M., Karniadakis, G.E., 2018. J. Comput. Phys. 357, 125–141.
Ranger, K., 1972. SIAM J. Appl. Math. 23 (3), 325–333.
Rimmer, P.L., 1968. J. Fluid Mech. 32 (1), 1–7.
Rimon, Y., Cheng, S., 1969. Phys. Fluids 12 (5), 949–959.
Rouse, H., 1961. Technical report. Dover Publications Inc, New York, USA.
Rudin, C., 2019. Nat. Mach. Intell. 1 (5), 206–215.
Sanjeevi, S.K., Kuipers, J., Padding, J.T., 2018. Int. J. Multiphase Flow.
Schiller, L., Naumman, A.Z., 1933. Vereines Deutscher Inge. 77, 318–321.
Schlichting, H., Gersten, K., 2016. Boundary-layer theory. Springer, Berlin Heidelberg, Germany.
Schmidt, M., Lipson, H., 2009. Science 324(5923), 81–85.
Schultz, M.P., Flack, K.A., 2013. Phys. Fluids 25 (2), 025104.
Sparrow, E.M., Abraham, J.P., Tong, J.C., 2004. Int. J. Heat Mass Transf. 47 (24), 5285–5296.

Stigler, S.M., 1974. Historia Math. 1 (4), 431–439.
Stokes, G.G., 1851. On the effect of the internal friction of fluids on the motion of pendulums, volume 9. Pitt Press Cambridge, Cambridge, UK.
Suryanarayana, G., Pauer, H., Meier, G., 1993. Exp. Fluids 16 (2), 73–81.
Swamee, P.K., Ojha, C.S.P., 1991. J. Hydraul. Eng. 117 (5), 660–667.
Taneda, S., 1956. J. Phys. Soc. Jpn. 11 (10), 1104–1108.
Taylor, B., 1717. Methodus incrementorum directa & inversa. Inny.
Thompson, N.C., Greenewald, K., Lee, K., Manso, G.F., 2020. arXiv preprint arXiv:2007.05558.
Udrescu, S.-M., Tegmark, M., 2020. Sci. Adv. 6 (16), eaay2631.
Van Dyke, M., 1970. J. Fluid Mech. 44 (2), 365–372.
Van Dyke, M., 1975. NASA STI/Recon Technical Report A 75.
Weisenborn, A., Ten Bosch, B., 1995. SIAM J. Appl. Math. 55 (3), 577–592.
Weng, B., Song, Z., Zhu, R., Yan, Q., Sun, Q., Grice, C.G., Yan, Y., Yin, W.-J., 2020. Nat. Commun. 11 (1), 1–8.
Whitaker, S., 1972. AIChE J. 18 (2), 361–371.
Wieselsberger, C.v., 1922. Phys. J. 23, 219–224.
Will, J.B., Kruyt, N.P., Venner, C.H., 2017. Int. J. Heat Mass Transf. 109, 1059–1067.
www.numpy.org/doc/stable/user/.
Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2016. arXiv preprint arXiv:1611.03530.