J. Fluid Mech. (2025), *vol.* 1020, A11, doi:10.1017/jfm.2025.10656

Extracting self-similarity from data

Nikos Bempedelis[1], **Luca Magri**[2,3,4] **and Konstantinos Steiros**[2]

[1]School of Engineering and Materials Science, Queen Mary University of London, London E1 4NS, UK

[2]Department of Aeronautics, Imperial College London, London SW7 2AZ, UK

[3]The Alan Turing Institute, London NW1 2DB, UK

[4]Politecnico di Torino, DIMEAS, Corso Duca degli Abruzzi 24, Torino 10129, Italy

Corresponding authors: Nikos Bempedelis, n.bempedelis@qmul.ac.uk; Konstantinos Steiros, k.steiros@imperial.ac.uk

(Received 4 February 2025; revised 30 June 2025; accepted 1 September 2025)

Identifying self-similarity is key to understanding and modelling a plethora of phenomena in fluid mechanics. Unfortunately, this is not always possible to perform formally in highly complex flows. We propose a methodology to extract the similarity variables of a self-similar physical process directly from data, without prior knowledge of the governing equations or boundary conditions, based on an optimisation problem and symbolic regression. We analyse the accuracy and robustness of our method in five problems which have been influential in fluid mechanics research: a laminar boundary layer, Burger's equation, a turbulent wake, a collapsing cavity and decaying turbulence. Our analysis considers datasets acquired via both numerical and wind tunnel experiments. The algorithm recovers the known self-similarity expressions in the first four problems and generates new insights into single length scale theories of homogeneous turbulence.

Key words: general fluid mechanics

1. Introduction

The various constraints that describe a physical phenomenon are often reflections of underlying symmetry principles, summarising regularities that exist independent of specific dynamics (Gross 1996). A notable example is the constraint of momentum conservation, which reflects a translational symmetry of the Euler–Lagrange equations (Landau & Lifshitz 1976). Dimensional analysis, historically linked to the discovery of scaling laws and non-dimensional numbers (Barenblatt 1996; Cantwell 2002), expresses the principle of covariance, i.e. the symmetry of any physical law under a dilational transformation of its units of measurement. More generally, when one considers a specific

problem, i.e. the governing physical laws and the corresponding boundary and/or initial conditions, multiple symmetries may be simultaneously present (e.g. spiral, reflectional, rotational) (Pakdemirli & Yurusoy 1998; Cantwell 2002).

A type of symmetry which is of fundamental importance across various branches of physics is that of self-similarity (Tennekes & Lumley 1972; Townsend 1976; Barenblatt 1996; Cantwell 2002). Following Townsend (1976), Barenblatt (1996), Pope (2000), we refer to self-similar (or self-preserving) phenomena as those whose evolution remains invariant under the group transformations of dilation (e.g. heat diffusion), translation (e.g. travelling waves) or a combination of both (e.g. turbulent wakes). Self-similarity allows for the reduction of an n-independent variable partial differential equation system into a system with $n-1$ similarity variables (Birkhoff 1960; Pakdemirli & Yurusoy 1998; Cantwell 2002). The transformation from original to similarity variables is known as a similarity transformation. Similarity transformations are particularly useful in problems involving $n=2$ independent variables because they transform a partial differential equation into an ordinary differential equation, which is more amenable to analytical solution (Pakdemirli & Yurusoy 1998; Cantwell 2002). Even if the governing equations are unknown, similarity considerations allow for the grouping of the problem parameters into similarity variables, reducing the effort required by the experimentalist when performing parametric characterisations of a problem by orders of magnitude (Barenblatt 1996). Additionally, the identification of similarity variables, subject to the constraints of the governing equations, can lead to the derivation of scaling laws, which describe the asymptotic evolution of the problem variables (Tennekes & Lumley 1972; Townsend 1976; Beaumard *et al.* 2024). The identification of similarity transformations and variables has played a prominent role in fluid mechanics research. Indeed, self-similarity is at the heart of theoretical efforts for the modelling of laminar and turbulent boundary layers (Prandtl 1904; Townsend 1976), free shear flows (Tennekes & Lumley 1972; Townsend 1976; George 1989; Pope 2000), cascade dynamics (Kolmogorov 1941*b*; Pope 2000; Vassilicos 2015; Steiros 2022*a*), linear and nonlinear waves (Taylor 1950*a*, *b*; Zel'Dovich & Raizer 1967; Whitham 2011), singularities (Eggers & Fontelos 2015) and high Mach number aerofoil design (Cantwell 2002), among many others.

There are three possible strategies to identify self-similarity. First, if the differential equations describing the problem are known and the boundary conditions are simple, one can formally extract the flow symmetries using the theory of Lie groups (Birkhoff 1960; Cantwell 1978; Cantwell 2002). Second, if the equations are unknown one may still invoke dimensional analysis to uncover dilational symmetries (Barenblatt 1996). Finally, on many occasions, intuition has allowed the uncovering of self-similarity directly, via visual inspection of the problem solution (as for example in the case of turbulent shear flows Townsend 1976). However, this is not always sufficient to reveal the underlying self-similarity of a phenomenon. Intuition depends on the skills of the practitioner and is generally confined to simple problems. Dimensional analysis can only treat problems with dimensional variables, while even then it may uncover only the 'tip of the iceberg' of possible self-similar solutions, i.e. some (but far from all) of the self-similarities connected to dilational transformations (Barenblatt 1996). Lie group theory may become unfeasible in cases where the boundary conditions are overly complicated. Even if this is not the case, it may reveal some but not all symmetries of the problem. A coordinate transformation may be necessary for additional, 'hidden' symmetries to be revealed (Cantwell 2002; Liu & Tegmark 2022). In other cases the governing equations may be inadequate. An example of the above occurs in turbulent flows, where it is customary to consider the ensemble-averaged flow equations of motion i.e. the Reynolds-averaged Navier–Stokes (RANS) equations. Given the chaotic turbulent dynamics and complex boundary conditions, certain

original symmetries of the Navier–Stokes equations are broken, but are still hypothesised to be recovered in a statistical sense (Frisch 1995). However, due to the underdetermined nature of the RANS equations (turbulence closure problem), extraction of symmetries is challenging (Oberlack 1999, 2001; Oberlack *et al.* 2006; Oberlack & Rosteck 2010; Oberlack *et al.* 2022).

The recent emergence of machine learning methodologies has provided a significant boost to our ability to uncover symmetries and self-similarities (Desai, Nachman & Thaler 2022; Yang *et al.* 2023; Otto *et al.* 2023). If self-similarity exists in a problem, it must appear in the observables (data), and can be, in principle, discovered by data-driven methods. For instance, a known challenge in conventional dimensional analysis (Buckingham Pi theorem) is that it yields a non-unique set of non-dimensional variables that governs the evolution of a physical problem (Barenblatt 1996; Cantwell 2002). Information from data can in that case provide a constraint to Buckingham Pi, and has been recently used to identify the most appropriate set of non-dimensional variables (Mendez & Ordóñez 2004; Constantine, del Rosario & Iaccarino 2017; Jofre, del Rosario & Iaccarino 2020; Saha *et al.* 2021; Xie *et al.* 2022; Bakarji *et al.* 2022; Yuan & Lozano-Durán 2025). Other examples are cases in which conservation laws and symmetries of ordinary differential equations (ODE) cannot be readily extracted because they are hidden, i.e. they require a coordinate transformation before they become manifest. Machine learning can be used to provide the necessary coordinate transformation for hidden symmetries to manifest (Liu & Tegmark 2022; Mototake 2023). Of high relevance are also efforts to leverage the information contained in data to reveal scale-invariant flow structures (Fukami, Goto & Taira 2024) or close underdetermined governing equations (Duraisamy, Iaccarino & Xiao 2019), which can support the training and testing of nonlinear machine-learning techniques (e.g. Fukami & Taira (2024), Duraisamy, Brunton & Taira (2025)).

In this work, we present a generalised methodology that can identify if similarity variables exist and, if they do, determine their mathematical expressions directly from data, without prior knowledge of the governing equations or boundary conditions. Our work differs from previous efforts (e.g. Bakarji *et al.* 2022; Xie *et al.* 2022; Yuan & Lozano-Durán 2025) in that it is not based on dimensional analysis (although it is dimensionally consistent) and can thus identify self-similarities beyond dilational transformations connected to dimensional problems. This is achieved by formulating a minimisation problem to identify the similarity variables, which are then interpreted analytically via symbolic regression. In particular, our method simultaneously searches for the optimal nonlinear transformations of both independent variables (spatio-temporal coordinates) and dependent variables (observables) that yield the similarity variables of the problem and is thus different to methods aiming to identify coordinate transformations that render hidden symmetries manifest (e.g. Liu & Tegmark 2022). The paper is structured as follows: in § 2 and Appendix A we describe the algorithm that extracts the similarity variables of a self-similar physical process from data. Section 3 demonstrates the potential utility of the proposed method by applying it to five problems which have been influential in fluid dynamics research and are exactly or approximately self-similar under different transformation types, with data derived from both laboratory and numerical experiments. Finally, § 4 discusses potential applications and limitations of the method and provides concluding remarks.

2. Data-driven identification of self-similarity

Let $q(s, t)$ be a quantity of interest governed by a set of nonlinear partial differential equations, with s and t being the independent variables, typically associated with a spatial

coordinate and time, respectively. In some problems (e.g. a laminar boundary layer), t may also correspond to a spatial coordinate. Consider the set of variables ξ and $\tilde{q}$, which lead to the transformation $q(s,t) \rightarrow \tilde{q}(\xi,t)$, with $\xi = \xi(s,t)$, where ξ may also depend on the parameters (constants) of the problem. If $\tilde{q}(\xi,t)$ is independent of t, i.e. there is a function $\hat{q}(\xi)$ such that $\tilde{q}(\xi,t) \equiv \hat{q}(\xi)$, then $q(s,t)$ is said to be self-similar (Pope 2000). In this case, the variables and the transformation are referred to as similarity variables and similarity transformation, respectively. Assuming that we have measurements of $q(s,t)$ at distinct instants (stations) t_i, where $i = 1, \ldots, n_t$ and n_t is the number of available instants (stations), we propose a two-stage workflow for extracting the similarity transformation directly from data, without previous knowledge of the governing equations. Because self-similarity can be inexact or the data may include errors, the equivalence $\tilde{q}(\xi,t) \equiv \hat{q}(\xi)$ can be at most expected to hold approximately. In higher dimensions, s and ξ are vectors.

2.1. *Step 1. Search for similarity variables*

We express the similarity variables as a superposition of elementary dilation and translation groups

$$\xi = \alpha(t)s + \beta(s,t), \quad \tilde{q} = \gamma(t)q + \delta(s,t). \tag{2.1}$$

Decomposing the similarity variables in the form of distinct transformations is important for facilitating the success of the ensuing optimisation and regression tasks, as well as for enhancing the interpretability of the method. The search for similarity variables is formulated as a minimisation problem

$$\arg\min_{\boldsymbol{w}} \frac{1}{2} \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} \left\| \tilde{q}(\xi,t_i) - \tilde{q}(\xi,t_j) \right\|_2^2, \tag{2.2}$$

where $\boldsymbol{w}$ is the design variable matrix containing the discrete values of the transformation functions α, β, γ, δ. The l_2-norms, $\|\cdot\|_2^2$, are computed following interpolation on the transformed coordinates (ξ) grid. The resolution of the ξ grid, i.e. the number of points where the l_2 norms are evaluated, can be freely chosen by the user and can be set equal to the number of points in the input (non-transformed) data (see also Appendix B). In some problems, solving (2.2) can lead to degenerate solutions (i.e. $\tilde{q} = 0$), in which case (2.2) is replaced by a mean-regularised cost functional, which prevents trivial optima

$$\arg\min_{\boldsymbol{w}} \frac{1}{2} \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} \left\| \frac{\tilde{q}(\xi,t_i) - \tilde{q}(\xi,t_j)}{\sqrt{|(\tilde{q}(\xi,t_i) + \tilde{q}(\xi,t_j))/2|}} \right\|_2^2. \tag{2.3}$$

This step provides the discrete values of the functions α, β, γ, δ and, by extension through (2.1), the discrete values of the similarity variables ξ and $\tilde{q}$, at all n_t instants. In practice, in order to further promote the success of the optimisation task whilst also preserving the automated character of the developed method, we propose the implementation of step 1 in successive iterations of increased decomposition complexity (i.e. building gradually from simple dilation ($\beta = \delta = 0$) to the general decomposition ($\xi = \alpha(t)s + \beta(s,t)$, $\tilde{q} = \gamma(t)q + \delta(s,t)$)). Depending on the problem (e.g. its boundary conditions), certain candidate transformations might be irrelevant or inadmissible (for example, translation of radial coordinates in an axisymmetric problem). Appendix A provides the pseudocode describing an example implementation of the method with clarifying remarks.

2.2. *Step 2. Analytic form of the transformations*

Given the knowledge of $\boldsymbol{w}$, i.e. the values of the functions α, β, γ, δ at all instants n_t, we employ symbolic regression to extract the analytic form of the transformations. Symbolic regression is a machine-learning technique that combines mathematical operators, functions, constants and state variables to construct a mathematical expression ψ that best represents a given dataset $\mathcal{D}$. In this work, $\mathcal{D}$ refers to the discrete values of each function (α, β, γ or δ) that were found in the first step of the workflow. Symbolic regression can identify arbitrary expressions as it does not make any assumptions about the underlying function. The set of variables (library) $\boldsymbol{A}$ which ψ can depend on, $\psi = \psi(\boldsymbol{A})$, is provided by the user, and is typically composed of the state variables and parameters of the problem.

To reduce the complexity of the regression problem, it is beneficial to take advantage of any prior knowledge regarding the characteristics of the dataset. We enforce two properties. The first property that is critical for discovering physical laws is that of dimensional homogeneity: the units of the identified expression $\psi(\boldsymbol{A})$ should match those of the given dataset $\mathcal{D}$. The second property that is relevant to this work is that scale invariance is related to power laws (Barenblatt 1996), allowing us in certain cases to restrict the search for expressions in the form of monomials.

This step of the algorithm can be carried out with any symbolic regression method, giving users the flexibility to choose their preferred tool. Here, we demonstrate the workflow with two different symbolic regression methods: the open-source library `PySR` (https://astroautomata.com/PySR/), which offers high performance, flexibility, configurability and generality (Cranmer 2023), and a custom, simple regression algorithm for the identification of monomials. In both cases, we use an l_2 loss function, with an optional penalisation term enforcing dimensional homogeneity

$$\arg\min_{\psi} \|\psi(\boldsymbol{A}) - \mathcal{D}\|_2^2 + w_D \| [\psi(\boldsymbol{A})] - [\mathcal{D}] \|, \tag{2.4}$$

where w_D is a non-negative regularisation factor and the brackets [ ] denote the units of a quantity in the form of a dimension vector (see Appendix A for an example). Equation (2.4) is individually applied to each transformation function (α, β, γ or δ). This step provides the analytic form of the similarity transformations α, β, γ, δ, and thereby of the similarity variables ξ and $\tilde{q}$, through (2.1).

3. Results

3.1. *Laminar boundary layer*

The> first validation example considers the well-known case of the two-dimensional, laminar and incompressible boundary layer, where Prandtl (1904) showed that some terms of the Navier–Stokes equations are negligible, and reduced the latter to the 'boundary layer equations' (BLE). By utilising the streamfunction Ψ, the BLE can be expressed in a single equation. Blasius (1908) noted that, for semi-infinite flat plate boundary conditions, the BLE is symmetrical under a dilational transformation, which reduces the problem to an ordinary differential equation for the non-dimensional streamfunction $f(\tilde{y}) = \Psi/\sqrt{\nu U_\infty x}$, with $\tilde{y} = y\sqrt{U_\infty/(\nu x)}$. In the above, U_∞ is the free-stream velocity away from the flat plate, x and y are the streamwise and normal-to-the-wall distances, respectively, and ν is the kinematic viscosity of the fluid. The BLE are thus reduced to the boundary value problem

$$2 f_{\tilde{y}\tilde{y}\tilde{y}} + f f_{\tilde{y}\tilde{y}} = 0, \tag{3.1a}$$

$$f(0) = 0,\ f_{\tilde{y}}(0) = 0,\ f_{\tilde{y}}(\infty) = 1. \tag{3.1b}$$

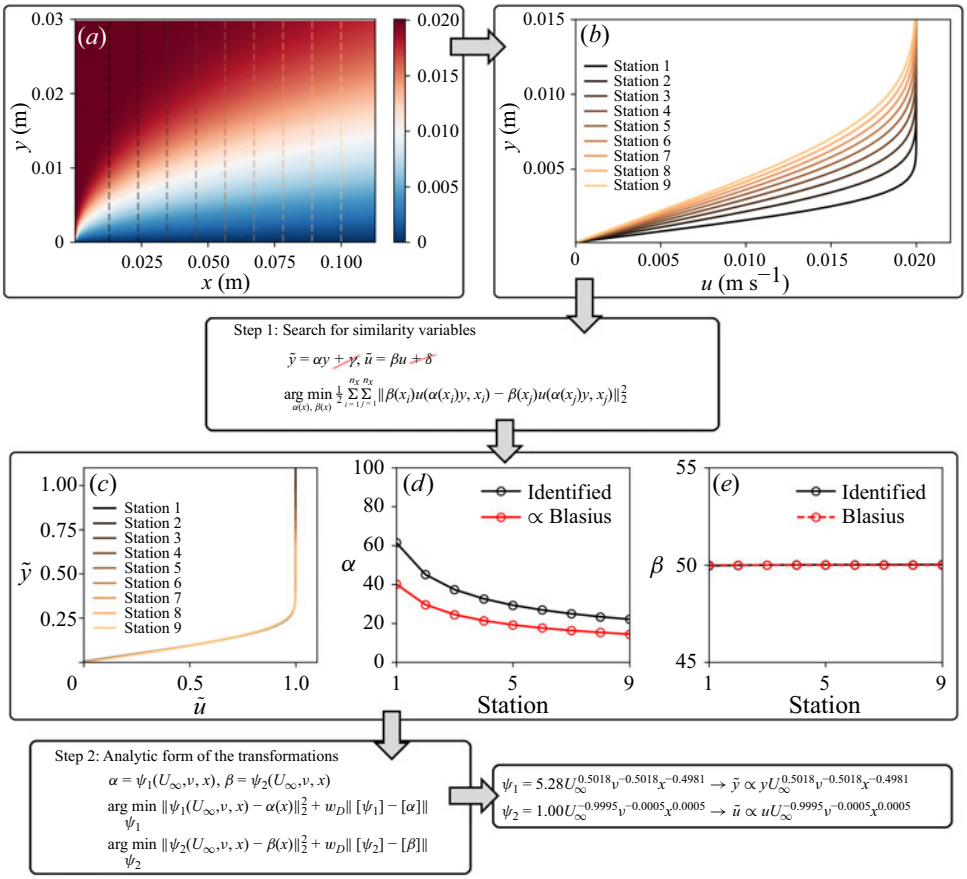

Figure 1. Data-driven identification of self-similarity in the Blasius boundary layer. (*a*) Streamwise velocity field (Blasius solution). The dashed vertical lines denote the nine velocity profile sampling stations. (*b*) Sampled velocity profiles. (*c*) Algorithmically collapsed velocity profiles. (*d*) Algorithmically identified scaling α of the wall-normal coordinate y. (*e*) Algorithmically identified scaling β of the streamwise velocity u.

Knowledge of $\tilde{u} = f_{\tilde{y}}$ (e.g. via numerical integration of the Blasius equation) allows for the calculation of the streamwise velocity distribution $u = U_\infty \tilde{u}$ at any location in the boundary layer. Here, we derive the Blasius similarity directly from data, without prior knowledge of the Navier–Stokes equations, Prandtl's scaling analysis or Blasius' similarity arguments. Our dataset consists of $u(x, y)$ data from a flat plate laminar boundary layer flow (figure 1*a*), acquired by solving the Blasius ODE with a shooting method (Bakarji *et al.* 2022). In particular, we extract velocity profiles $u(y)$ from nine stations at different streamwise coordinates x (figure 1*b*). The algorithm commences by seeking similarity under the dilational transformations $\tilde{y} = \alpha(x)y$, $\tilde{u} = \beta(x)u$. Figure 1(*c*) shows that the transformed profiles have collapsed onto a single curve. The identified transformations for the wall-normal coordinate y and the streamwise velocity u are shown in figures 1(*d*) and 1(*e*), respectively.

The second step consists of expressing the identified transformations as functions of the governing parameters and independent variables, i.e. a symbolic regression task. The library of candidate variables is composed of the free-stream velocity U_∞, the viscosity ν and the streamwise coordinate of the extracted profiles x, i.e. $\alpha = \alpha(U_\infty, \nu, x)$ and $\beta = \beta(U_\infty, \nu, x)$. Owing to the nature of the assumed dilational transformations, we look for

expressions in the form of monomials, i.e. $\alpha = c_1 U_\infty^{c_2} \nu^{c_3} x^{c_4}$ and $\beta = c_5 U_\infty^{c_6} \nu^{c_7} x^{c_8}$, where c_1 and c_5 are multiplicative constants that reflect the arbitrariness of the dilation factors between the original and similarity variables. Dimensional homogeneity is enforced by assuming $\tilde{y}$ and $\tilde{u}$ to be dimensionless and a non-zero weighting constant w_D in the objective function (2.4). The interpreted expressions are

$$\tilde{y} \propto y U_\infty^{0.5018} \nu^{-0.5018} x^{-0.4981}, \tag{3.2a}$$

$$\tilde{u} \propto u U_\infty^{-0.9995} \nu^{-0.0005} x^{0.0005}, \tag{3.2b}$$

which are in close agreement with the theoretically derived similarity variables of Blasius ($\tilde{y} = y\sqrt{U_\infty/(\nu x)}$, $\tilde{u} = u/U_\infty$). As previously noted, the identified and theoretical scalings can vary by an arbitrary multiplicative constant (i.e. the offset in figure 1*d*), depending as to whether and how the user normalises the collapsed profiles. Appendix B shows results from a sensitivity and robustness analysis with respect to the input data, parameters and noise.

3.2. *Burgers' equation*

The second validation example considers Burgers' equation, which is a partial differential equation that finds wide application in fluid dynamics, nonlinear acoustics and traffic flow, among others. It acts as a prototype for a variety of phenomena including shock wave formation, rarefaction waves and turbulence (Zel'Dovich & Raizer 1967; Whitham 2011). In the absence of diffusion, it is the simplest model for gas dynamics. Its mathematical expression in one dimension, along with its initial condition is

$$u_t + u u_x = 0, \quad u(x, 0) = f(x), \tag{3.3}$$

where $u(x, t)$ is the fluid velocity and x and t the spatial and temporal coordinates, respectively. The subscript denotes partial differentiation. Using the method of characteristics, the solution can be expressed as $u(x, t) = f(\xi) = f(x - ut) = f(x + \alpha)$, i.e. the problem described by (3.3) is invariant under a non-uniform translational transformation. In this work, we consider a sinusoidal initial condition, $u(x, 0) = \sin(2\pi x)$, with $x \in [0, 1]$. Figures 2(*a*) and 2(*b*) show the spatio-temporal evolution of velocity and profiles extracted at seven different time instants, respectively, illustrating the steepening of the solution as time evolves. The times shown are non-dimensionalised with the time of shock formation t_c, $t^* = t/t_c = 2\pi t$. The algorithm is provided with the data shown with markers in figure 2(*b*).

Figure 2(*c*) shows the algorithmically collapsed velocity profiles on the transformed coordinates $(\tilde{x}, t)$, i.e. the output of step 1 of the algorithm, following a transformation of the form $\tilde{x} = x + \alpha(x, t)$ (simpler transformations do not succeed at collapsing the data). Since this is an initial value problem, we only consider the distance of the profiles with respect to the initial condition, i.e. the objective function in (2.2) is simplified to $\sum_{i=1,\ldots,n_t} \|u(\tilde{x}, t_i) - u(x, 0)\|_2^2$.

We use `PySR` to interpret the algorithmically computed transformation α. The library of candidate variables is composed of the problem's independent and dependent variables, i.e. $\alpha = \alpha(u, x, t)$, and the library of operators includes the four basic mathematical operations $(+, -, \times, \div)$. `PySR` interprets the algorithmically computed transformation as $\alpha = -ut$, which is identical to the analytical similarity transformation. Close inspection of figure 2(*c*) shows that the profiles are not perfectly collapsed (for example, at $t^* = 0.936$). The range-normalised mean absolute error between the algorithmically computed discrete values of the transformation and the analytical ones is 2.2 %. However, in step 2 of the workflow (symbolic regression), the algorithm attempts to fit expressions that

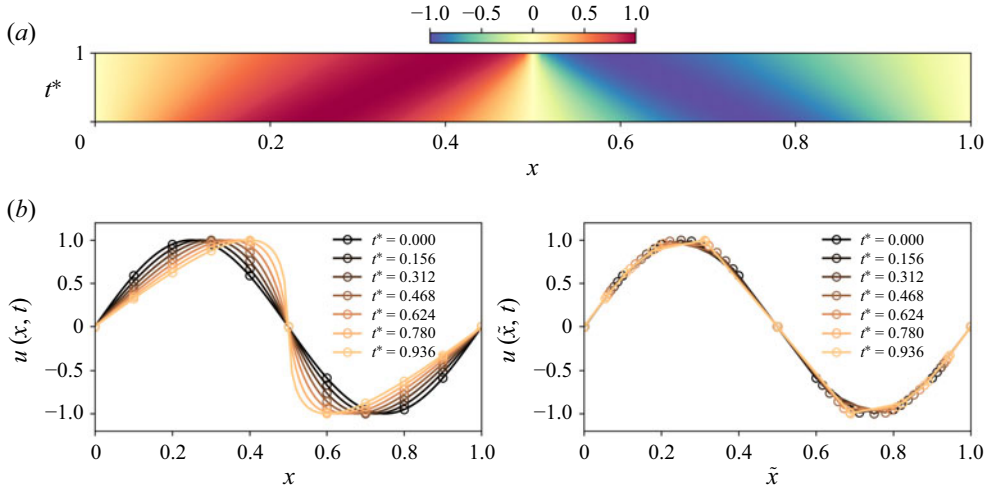

Figure 2. Data-driven identification of self-similarity in Burgers' equation. (*a*) Spatio-temporal evolution of the velocity. (*b*) Extracted profiles at different time instants. Markers show the data given to the algorithm. (*c*) Algorithmically collapsed profiles.

balance complexity and accuracy, allowing it to recover the exact analytical transformation $\alpha = -ut$.

3.3. *Free turbulent flow*

As a third example, we consider the statistically stationary flow past a slender bluff body at high Reynolds numbers, which is a characteristic example of a free (i.e. unconfined) turbulent shear flow. The term slender signifies an effectively infinite body length in one cross-flow direction, along which the flow statistics can be considered homogeneous. Such flows can be modelled using a boundary layer approximation, similar to the one employed by Prandtl for laminar boundary layers, applied to the time-averaged flow (Townsend 1976), i.e.

$$\bar{u}_x + \bar{v}_y = 0, \tag{3.4a}$$

$$\bar{u}\bar{u}_x + \bar{v}\bar{u}_y = \nu\bar{u}_{yy} - (\overline{u'v'})_y, \tag{3.4b}$$

where the overbar denotes time averaging and ν is the kinematic viscosity of the fluid. Here, u and v are the flow velocities along the streamwise x and cross-wise y directions, which are decomposed into time-averaged $(\bar{u}, \bar{v})$ and fluctuating (u', v') components. The formulation leading to (3.4) introduces the Reynolds stress term $\overline{u'v'}$ into the flow governing equations, i.e. an additional unknown which cannot be calculated implicitly (this is the well-known closure problem of turbulence). The identification of similarity in this case is, therefore, not possible from the equations themselves, unless an assumption is made regarding the relation of the unknown Reynolds stress terms to the mean velocity distribution (i.e. turbulence modelling). However, it is customary to hypothesise that the turbulent BLE accept self-similar solutions far from initial conditions, a fact also supported by experimental observations (Townsend 1976; Cantwell 2002). This approximate self-similarity is the origin behind various scaling laws for the evolution of free shear flows, widely used in many applications of the energy and transportation sectors, such as wind farm planning (Bempedelis & Steiros 2022) and jet engine noise prediction (Tam 2019). Returning to the particular example of the slender bluff body wake, we may

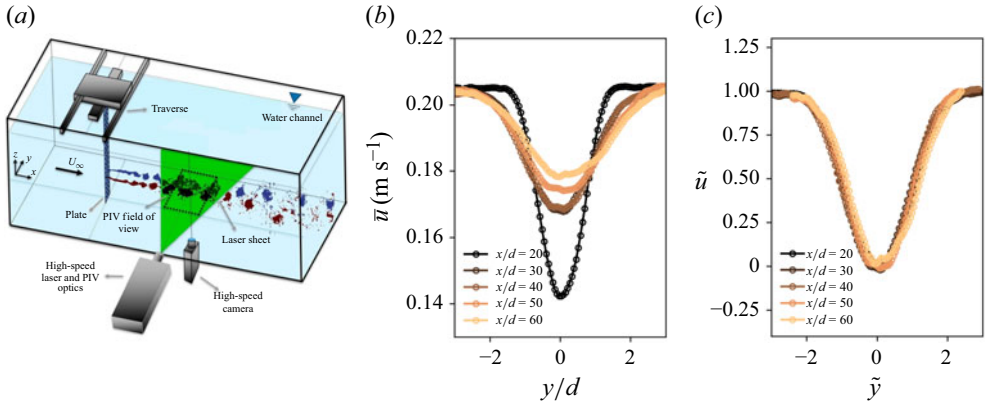

Figure 3. Data-driven identification of self-similarity in the wake of a porous plate. (*a*) Schematic of the experimental apparatus showing the flume, porous plate and PIV configuration. (*b*) Mean streamwise velocity profiles at different locations downstream of the plate. (*c*) Algorithmically collapsed velocity profiles.

consider the time-averaged velocity deficit $\zeta(x, y) = U_\infty - \bar{u}(x, y)$ that the body produces far (i.e. tens of characteristic body lengths) downstream. In the above, U_∞ is the constant free-stream velocity. The flow statistics are homogeneous along the spanwise direction z. Wake self-similarity is then known to assume the form (Townsend 1976; Pope 2000)

$$\zeta(x, y) = \tilde{\zeta}(\xi)(U_\infty - \bar{u}_{cntr}(x)), \quad \text{with } \xi = \frac{y}{y_{1/2}(x)}, \tag{3.5}$$

where $\bar{u}_{cntr} = \bar{u}(x, 0)$ denotes the centreline velocity and $y_{1/2}(x)$ is the wake half-width defined such that $\bar{u}(x, \pm y_{1/2}) = 0.5(U_\infty + u_{cntr})$. We attempt to extract the self-similar relation (3.5) directly from experimental data. To this end, we measured the turbulent wake of a plate of 53 % porosity immersed in a water flume normal to the flow (see figure 3*a*) at a Reynolds number based on the free-stream velocity and plate width $Re \approx 6000$. The velocity fields at various positions downstream of the plate were measured using Particle Image Velocimetry (PIV) (Phantom 4MP camera at 50 Hz acquisition frequency) as shown in figure 3(*a*). Each mean streamwise velocity profile was then calculated by averaging 3000 vector fields. More information regarding the experimental procedure can be found in Bekoglu *et al.* (2024, 2025). We consider the mean velocities at five stations in the wake of the plate (figure 3*b*). Figure 3(*c*) shows the collapse that is obtained via the proposed method, assuming transformations of the form $\tilde{y} = \alpha(x)y$ and $\tilde{u} = \beta(x)\bar{u} + \gamma(x)$ (simpler transformations cannot collapse the data). The identified transformations are regressed as functions of the characteristic scales and variables of the problem using `PySR`, with $\alpha = \alpha(x, y_w, y_{1/2})$ (with $y_w(x)$ defined as $\bar{u}(x, \pm y_w) = 0.99U_\infty$), $\beta = \beta(U_\infty, \bar{u}_{cntr}, x, \nu)$ and $\gamma = \gamma(U_\infty, \bar{u}_{cntr}, x, \nu)$. The library of operators consists of the four basic mathematical operations $(+, -, \times, \div)$. The interpreted similarity transformations are

$$\tilde{y} = \alpha y = \frac{1.0605}{y_{1/2}(x)} y, \tag{3.6a}$$

$$\tilde{u}(x, \tilde{y}) = \beta\bar{u} + \gamma = \frac{1}{U_\infty - \bar{u}_{cntr}}\bar{u} - \frac{\bar{u}_{cntr}}{U_\infty - \bar{u}_{cntr}} = 1 - \tilde{\zeta}, \tag{3.6b}$$

which match (ignoring the arbitrary multiplicative constant) the self-similar expressions for turbulent wakes found in the literature (Tennekes & Lumley 1972; Townsend 1976; Pope 2000). Besides our own experiments, we also test our method by using experimental

data available in the literature (Cimbala, Nagib & Roshko 1988), and again retrieve the self-similarity relations for turbulent wakes (see Appendix C).

3.4. *Cavity collapse*

The bursting of bubbles at the sea's surface plays a crucial role in the exchanges between oceans and the atmosphere, thereby being important for climate and weather (Deike 2022). When a bubble bursts at the sea's surface, aerosol is produced via two mechanisms: the rupture of the bubble's cap film, and the formation of a vertical jet that breaks into droplets following cavity collapse (Deike 2022). The jet and aerosol properties are a consequence of the nonlinear fluid dynamics near the points where the topology changes (Eggers 1997; Eggers & Fontelos 2015). In the case of a collapsing cavity, several studies have found that, near the collapse of the travelling capillary waves at the axis of symmetry, the liquid–gas interface evolves in a self-similar manner, enabling the derivation of scaling laws for the bubble and jet dynamics (Keller & Miksis 1983; Zeff *et al.* 2000; Duchemin *et al.* 2002; Ghabache *et al.* 2014; Gañán-Calvo 2017; Deike *et al.* 2018).

We simulate the collapse of a cavity ($Bo = \rho g R_0^2/\gamma = 10^{-3}$, $La = \rho\gamma R_0/\mu^2 = 2500$) by solving the two-phase incompressible axisymmetric Navier–Stokes equations using Basilisk (www.basilisk.fr), an open-source flow solver that has been previously used in computational studies of bursting bubbles (Lai, Eggers & Deike 2018; Berny *et al.* 2020; Sanjay, Lohse & Jalaal 2021). In the above, ρ and μ are the liquid density and viscosity, respectively, γ is the interfacial tension, R_0 is the initial bubble radius and g is the gravitational acceleration. The evolution of the liquid–gas interface is shown in figures 4(*a*)–4(*d*). We extract profiles of the interface $h(r, t)$ as it approaches the axis of symmetry, at $t_* = (t_0 - t)/t_c = [10, 9, 8, 7, 6]$, where t_c is the characteristic time of the horizontal capillary wave (Lai *et al.* 2018), and t_0 is the moment where the capillary waves meet at the axis of symmetry, $r = 0$. The extracted profiles are shown in figure 4(*e*). Duchemin *et al.* (2002) and Lai *et al.* (2018) observed that, in this time window, the interface approximately follows the $(t_0 - t)^{2/3}$ scaling of inviscid theory (Keller & Miksis 1983). The algorithm successfully collapses the extracted profiles into a single curve (figure 4*f*) for transformations of the form $\tilde{r} = \alpha(t)r$ for the radial coordinate and $\tilde{h} = \beta(t)h + \gamma(t)$ for the liquid–gas interface. By regressing the identified transformations α, β and γ as power-law functions of the non-dimensional time t_* (figure 4*g*–4*i*), we retrieve the expressions

$$\tilde{r} = c_1 r t_*^{-0.61}, \tag{3.7a}$$

$$\tilde{h} = c_2 h t_*^{-0.69} - c_3 h_0 t_*^{-0.66}, \tag{3.7b}$$

which are close (ignoring the arbitrary multiplicative constants c_1, c_2 and c_3) to the theoretically derived ones (Keller & Miksis 1983; Zeff *et al.* 2000; Duchemin *et al.* 2002; Ghabache *et al.* 2014; Lai *et al.* 2018).

3.5. *Decaying turbulence*

We consider the case of homogeneous decaying turbulence, experimentally realised by passing a stream of fluid through a uniformly spaced grid inside a wind tunnel (figure 5*a*). We analyse measurements of velocity time series, obtained via hot-wire anemometry at seven positions downstream of the grid (the details of the experimental campaign can be found in Steiros 2022*a*). The measured turbulence is fully developed, approximately homogeneous and yields the $-5/3$ law for the energy spectrum at intermediately sized eddies, as predicted by Kolmogorov's K41 framework (Kolmogorov 1941*a*,*b*,*c*).

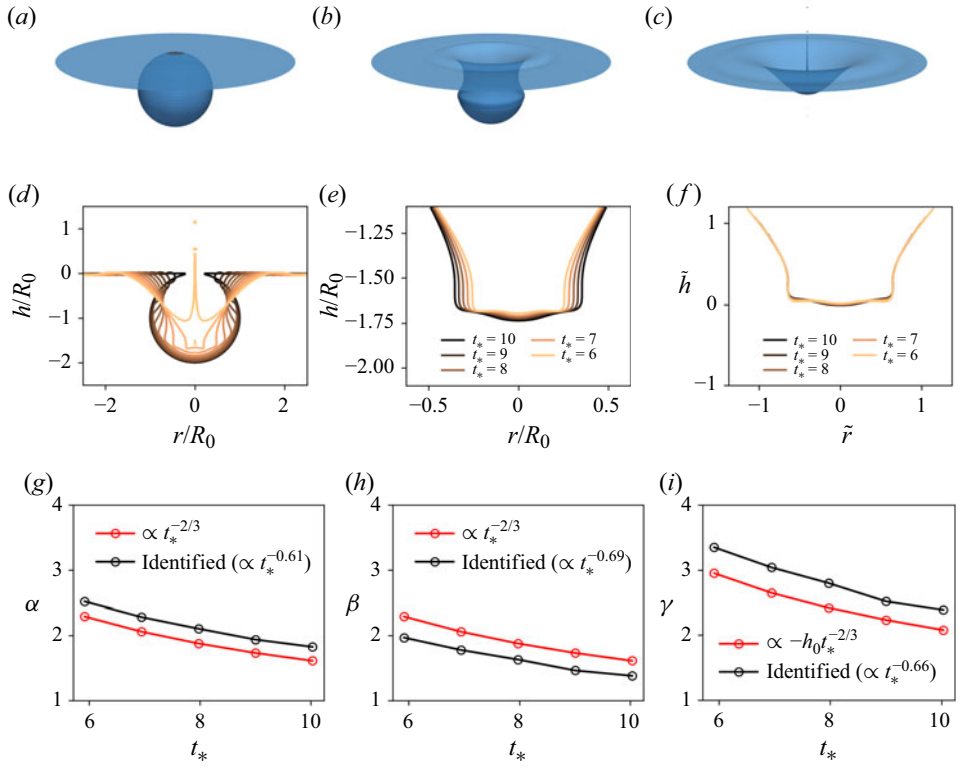

Figure 4. Data-driven identification of self-similarity in a collapsing cavity. Three-dimensional visualisation of the liquid–gas interface at (*a*) $t/\tau = 0$, (*b*) $t/\tau = 0.25$ and (*c*) $t/\tau = 0.5$, where $\tau = \sqrt{\rho R_0^3/\gamma}$ is the inertio-capillary time scale. (*d*) Liquid–gas interface time evolution, $t/\tau = (0, 0.05, \ldots, 0.5)$. (*e*) Interface profiles near cavity collapse. (*f*) Algorithmically collapsed interface profiles near cavity collapse. (*g–i*) Identified transformations α, β and γ. Comparison with theoretical scaling laws.

The theoretical derivation of the −5/3 law assumes that, at sufficiently high Reynolds numbers, an intermediate self-similar region forms in the cascade, which is independent of large- and small-scale effects (Batchelor 1953; Frisch 1995; Pope 2000). However, the above description is part of a more complex flow picture, as the widely studied paradigm of grid turbulence sufficiently far from initial conditions shows. The turbulence cascade in that case is generally accepted to be characterised by two self-similarities, both present at the same time and at different eddy sizes: One at large scales where viscosity is negligible (Steiros 2022*a*; Lundgren 2003), and one at small scales where non-equilibrium effects are negligible (Pope 2000; Lundgren 2003). The energy spectrum thus accepts the following general expression (Pope 2000):

$$E_{11}(\kappa, x) = \epsilon(x)^{2/3}\kappa^{-5/3} f(\kappa L) g(\kappa\eta)\,, \tag{3.8}$$

where x, κ, $E_{11}(\kappa, x)$ and $\epsilon(x)$ represent the distance from the grid, wavenumber, energy spectrum and dissipation rate, respectively. Here, L and η are the integral and Kolmogorov scales, characteristic of the large- and small-scale self-similarities, respectively. To retrieve the −5/3 law, $E_{11}(k, x) = \epsilon(x)^{2/3}\kappa^{-5/3}$, one needs to be asymptotically far from large scales (i.e. $\kappa L \to \infty$) and far from small scales ($\kappa\eta \to 0$) at the same time, as in that case K41 assures that both f and g tend to unity. It is of interest to see how our algorithm fares in this two-scale problem.

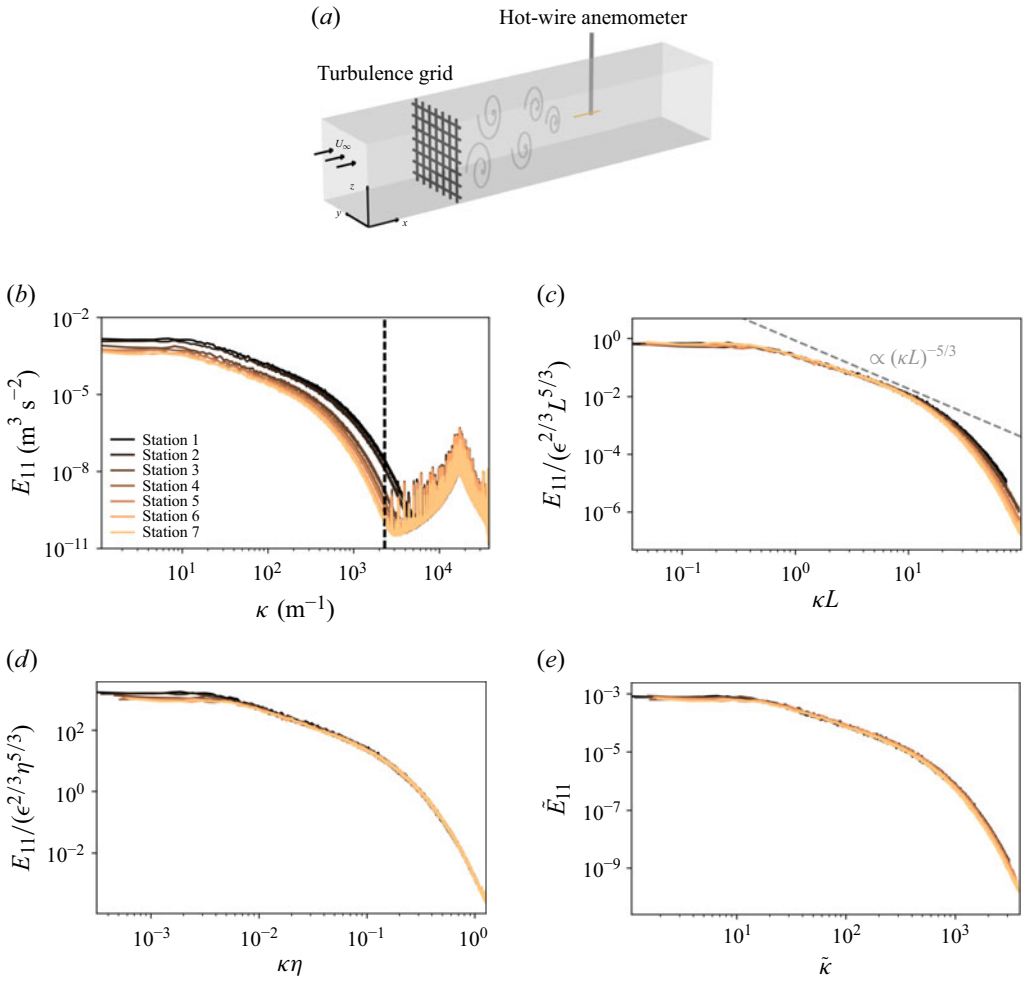

Figure 5. Data-driven identification of self-similarity in decaying turbulence. (*a*) Schematic of the experimental set-up. (*b*) Experimentally measured power spectral densities. The dashed vertical line delineates the range of the spectrum that is used. (*c*) Measured spectrum normalised by the inertial scales. (*d*) Measured spectrum normalised by the Kolmogorov scales. (*e*) Measured spectrum normalised by the algorithmically identified expression (3.9).

Figure 5(*b*) plots the measured power spectral densities versus the wavenumbers at different measurement stations. Figure 5(*c*,*d*) shows the collapse of the power spectral densities using large- and small-scale similarity variables, respectively. Using a library that consists of (L, η) for the x axis and (ϵ, L, η, k) for the y axis, symbolic regression of the algorithmically identified dilational transformations ($\tilde{\kappa} = \alpha(x)\kappa$, $\tilde{E_{11}} = \beta(x)E_{11}$) yields the following expressions, plotted in figure 5(*e*):

$$\tilde{\kappa} \propto \kappa L^{0.368}\eta^{0.632}, \qquad \tilde{E_{11}} \propto E_{11}\epsilon^{0.695}L^{0.709}\eta^{-1.015}k^{-2.042}. \tag{3.9}$$

As expected, the algorithm cannot derive the two-scale similarity expression (3.8), as the assumed dilational transformations effectively reduce it to a single-scale algorithm. Further insight may be obtained by using the dissipation scaling $\epsilon \propto k^{3/2}/L$, where k

is the turbulence kinetic energy, which is known to characterise homogeneous decaying turbulence far from initial conditions (Steiros 2022*a*,*b*; Vassilicos 2015). Expression (3.9) then becomes

$$\tilde{\kappa} \propto \kappa L^{0.368}\eta^{0.632}, \qquad \tilde{E_{11}} \propto E_{11}\epsilon^{-0.667}\left(L^{0.391}\eta^{0.609}\right)^{-5/3}. \tag{3.10}$$

Inspection of the above expression reveals that our algorithm has, seemingly, collapsed the totality of the cascade dynamics based on a single empirical length scale, $l \approx L^{0.38}\eta^{0.62}$, which is a combination of L and η. There have been several attempts to develop single length scale theories of turbulence, derived by assuming self-similarity of the turbulence statistics, i.e. see von Karman & Howarth (1938) and Sedov (2018), and the later theories of Barenblatt & Gavrilov (1974) and George (1992). The last two studies predict the Taylor microscale as the appropriate similarity variable, in close agreement with our data-driven methodology. In homogeneous decaying turbulence the Taylor microscale becomes $\lambda = L^{1/3}\eta^{2/3}$ (Pope 2000), which is close to our empirically derived length scale.

It might be tempting to interpret our empirical results as being in support of a single-scale picture for homogeneous turbulence, contrary to the commonly accepted view of turbulence as an intrinsically multi-length scale phenomenon (Batchelor 1953; Tennekes & Lumley 1972; Townsend 1976; Pope 2000). However, the results shown in figure 5 (and in most studies of homogeneous decaying turbulence) consider the far region of a turbulence grid where the integral length Reynolds number evolution is relatively slow, leading to a similarly slow evolution of L/η (Vassilicos 2015). As a result, our data show that all length scales in question (integral, Kolmogorov, Taylor) produce a (visually) 'good-enough' collapse, as shown in figure 5(*c*–*e*), and inspection of any of these graphs might give the impression of a single length scale self-similar evolution.

To better appreciate the single or multi-length scale nature of the turbulence cascade, we need data for which the integral length scale Reynolds number varies drastically. To achieve this, we analyse the periodic box Direct Numerical Simulation (DNS) data of decaying turbulence taken from Goto & Vassilicos (2016) (the details of the simulations can be found in the reference). Similar to the grid case, our goal is to collapse energy spectra taken at different decay times $\hat{t}$ (see figure 6*a*). Here, however, we consider datasets from two simulations which are identical in every aspect, apart from their size ($N = 1024^3$ and $N = 2048^3$) and the kinematic viscosity of the fluid, i.e. the Reynolds number of the flow varies drastically between the two cases, making the collapse of the curves less arbitrary.

Figure 6(*a*) presents the dimensional energy spectra from the two simulations, whereas figures 6(*b*) and 6(*c*) show the same data with the axes rescaled according to inertial and Kolmogorov similarity variables, respectively. The collapse of the large (small) scales is adequate only when the axes are normalised using the integral (Kolmogorov) scales, in agreement with a multi-scale view of the cascade. However, data from the same simulation size retain a 'good-enough' collapse independent of the choice of normalisation length, similar to our grid measurements, a fact which, if viewed in isolation, might give a false impression of a single length scale process. Figure 6(*d*) shows the algorithmically identified collapse of the data, that is, using the empirical transformations

$$\tilde{\kappa} \propto \kappa L^{0.342}\eta^{0.658}, \tag{3.11a}$$

$$\tilde{E_{11}} \propto E_{11}\epsilon^{-0.655}L^{-0.561}\eta^{-1.094}k^{-0.017} \approx E_{11}\epsilon^{-0.667}\left(L^{0.344}\eta^{0.656}\right)^{-5/3}. \tag{3.11b}$$

The algorithm again identifies, approximately, the Taylor microscale ($\lambda = L^{1/3}\eta^{2/3}$) as the appropriate length for the collapse of the turbulence dynamics, but this time it cannot

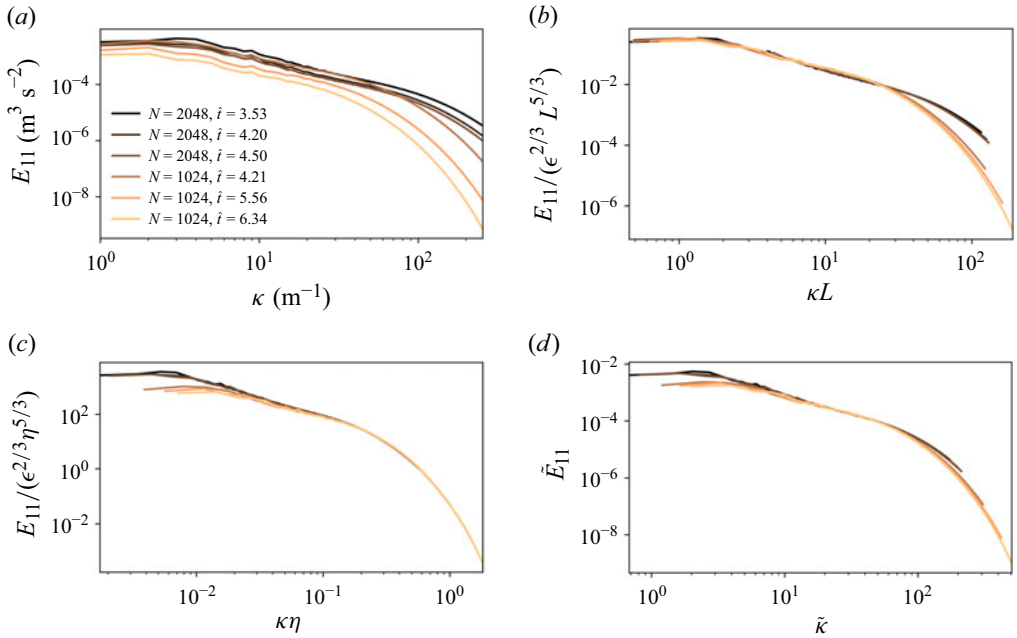

Figure 6. Data-driven identification of self-similarity in decaying turbulence, using the DNS data of Goto & Vassilicos (2016). (*a*) Power spectral densities. (*b*) Spectrum normalised by the inertial scales. (*c*) Spectrum normalised by the Kolmogorov scales. (*d*) Spectrum normalised by the algorithmically identified expression (3.11).

be claimed that the Taylor scale collapses both the large- and small-scale dynamics – only that it is an (algorithmically) optimal compromise when a single length scale collapse is enforced. We conjecture that the reason behind this could be explained from Lundgren's two-scale turbulence theory which is based on the technique of matched asymptotic expansions (Lundgren 2002, 2003). The repercussions of this theory are discussed in Obligado & Vassilicos (2019), Meldi & Vassilicos (2021): in particular, it is suggested that the Taylor microscale is the length scale at which the cascade dynamics is optimally distanced from both non-equilibrium effects (large scales) and dissipative effects (small scales) at the same time (in a matched asymptotic manner). The algorithmically identified collapse (figure 6*d*) produces best results in the intermediate, inertial range of the cascade where both above effects tend to become negligible. Based on this argument, the algorithm identifies the Taylor microscale scale because it is the optimal compromise of the two governing scales (inertial, Kolmogorov) of the cascade.

4. Discussion and conclusion

We have presented a methodology for identifying similarity variables from data, in the absence of governing equations. Given measurements of a quantity of interest, the method computes the discrete values of the similarity transformations that best collapse the data and expresses their analytic form via symbolic regression. The self-similarity that we identify is the optimal fit of the data, and is not derived from the governing partial differential equations and the boundary conditions of the specific problem.

The proposed method differs from dimensionless learning approaches that aim to reduce the dimensionality of parameter spaces via identification of dimensionless groups

(Xie *et al.* 2022; Bakarji *et al.* 2022; Yuan & Lozano-Durán 2025) because our focus is on discovering the similarity transformations that enable the collapse of entire solution profiles. Our approach is not restricted to dilational similarity transformations, and is capable of identifying more general, non-uniform relationships between variables, as demonstrated in the Burgers equation example. This flexibility broadens the range of problems for which similarity solutions can be uncovered. In practical terms, this approach also reduces the degree of prior knowledge required to reveal self-similar behaviour. For example, in the turbulent wake case, methods based on dimensional analysis would require the characteristic velocity difference to be provided as input in order to reveal the self-similarity. On the other hand, our method can uncover it autonomously. More generally, step 1 of the proposed framework can be used to detect self-similar behaviour whilst being effectively agnostic to the underlying physics, i.e. without requiring prior specification of the relevant variables or parameters. In the symbolic regression step (step 2), the user must supply the candidate parameters to obtain explicit analytic expressions, but the ability to first reveal the existence of self-similarity without specifying inputs may be particularly appealing for exploratory studies or complex systems. On the other hand, for cases of pure dilational self-similarity, methods based on dimensional analysis are expected to be more efficient. A note of caution: the computation of the l_2 norms in (2.2) is performed over the common support of the transformed profiles. If the supports differ significantly, the optimisation may become challenging. Although this issue did not arise in this work, we acknowledge it as a point that might need to be addressed in future work.

The capabilities of the method were demonstrated in five fluid mechanical problems, which are known or assumed to accept similarity solutions, based on data from both numerical and laboratory experiments. In four problems, the transformations identified by the method were in agreement to the analytically or empirically derived similarity transformations, whilst circumventing the need for scaling analysis or similarity arguments. In the fifth problem (turbulence cascade) our method identified an empirical length scale as the optimal similarity length of the cascade, which was close to the Taylor microscale. Further investigation revealed that this result cannot be treated as being unequivocally in support of single length scale theories of turbulence (von Karman & Howarth 1938; Barenblatt & Gavrilov 1974; George 1992; Sedov 2018), but rather expresses the role of the Taylor microscale as an intermediate scale between the integral and Kolmogorov scales, at which non-equilibrium and dissipative effects are optimally spaced (Meldi & Vassilicos 2021).

The results illustrate ways that the proposed algorithm can be used for the identification of similarity and scaling laws in situations where rigorous mathematical analysis is challenging. This includes a wide range of applications in fluid physics but also in other processes such as quasicrystal shape and growth (Kamiya *et al.* 2018), stellar collapse (Yahil 1983), single protein dynamics (Hu *et al.* 2016) and others. Other directions for future work include using the method to detect symmetry breaking, distinguish similarity across multiple scales and identify the onset or breakdown of self-similarity.

Acknowledgements. The authors would like to thank E. Bekoğlu for providing the flume data and the schematic of the experimental set-up. K.S. is grateful to Professor S. Goto for providing the periodic box DNS data.

Funding. Parts of this work were conducted while N.B. was at Imperial College London, supported by EPSRC, Grant No. EP/W026686/1. L.M. acknowledges support from the ERC Starting Grant PhyCo No. 949388 and the EU-PNRR YoungResearcher TWIN ERC-PI_0000005. K.S. acknowledges support from the ERC Starting Grant ONSET No. 101163321.

Declaration of interests. The authors report no conflict of interest.

Algorithm 1 Data-driven similarity inference: implementation example

Gather a set of n_t observations $q(s)$
$\boldsymbol{s} \in \mathbb{R}^{n_s \times n_t}$, $\boldsymbol{q} \in \mathbb{R}^{n_s \times n_t}$
Break the generalised similarity variable decomposition down to a set of admissible/relevant candidate transformations
e.g., pure dilation for s ($\beta = 0$), and pure ($\delta = 0$) or generalised dilation ($\delta = \delta(t)$) for q. The generalised decomposition is $\xi = \alpha(t)s + \beta(s, t)$ and $\tilde{q} = \gamma(t)q + \delta(s, t)$.
// Step 1: Search for similarity variables
for each combination $\in$ set of candidate transformations **do**
  Initialise design variable matrix $\boldsymbol{w}$
  Dimensions of $\boldsymbol{w}$ depend on the nature of the candidate transformations. In the case of pure dilation for both s and q, $\boldsymbol{w} \in \mathbb{R}^{n_t \times 2}$, with the first column corresponding to$s$ and the second column to q (i.e., containing the discrete values of α and γ, respectively).
  // Optimisation problem
  for iter $= 1$ **to** stopping criterion **do**
    Apply transformations to observations
    $\boldsymbol{\xi} = \boldsymbol{s} \odot (\mathrm{col}_1(\boldsymbol{w}))^{\mathsf{T}}, \tilde{\boldsymbol{q}} = \boldsymbol{q} \odot (\mathrm{col}_2(\boldsymbol{w}))^{\mathsf{T}}$
    Interpolate $\tilde{\boldsymbol{q}}$ on common $\boldsymbol{\xi}$
    Resolution of interpolation grid may vary.
    Compute objective function (OF) (2.2) or (2.3)
    Update $\boldsymbol{w}$ to minimise the OF
  end for
  if OF $\ll 1$ **then**
    $\boldsymbol{w}_{opt} \leftarrow \boldsymbol{w}$
    Exit step 1: similarity variables found. Potentially identified self-similarity.
  end if
end for
// Step 2: analytic form of the transformations
Normalise transformations and variables found in step 1 (optional)
for each transformation component (i.e. column j of $\boldsymbol{w}_{opt}$) **do**
  Define the n_v variables $\boldsymbol{A} \in \mathbb{R}^{n_v \times 1}$ that the component can be a function of
  Construct a matrix $\boldsymbol{A}_D$ containing the values of $\boldsymbol{A}$ at all locations
  In the example above, $\boldsymbol{A}_D \in \mathbb{R}^{n_v \times n_t}$
  // Symbolic regression task
  Find expression $\psi(\boldsymbol{A})$ that minimises
  $\|\psi(\boldsymbol{A}_D) - \mathrm{col}_j(\boldsymbol{w}_{opt})\|_2^2 + w_D \|[\psi(\boldsymbol{A})] - [\mathrm{col}_j(\boldsymbol{w}_{opt})]\|$
  The brackets [ ] denote the physical dimensions of a quantity in the form of a dimension vector, i.e., a vector containing the powers of its basic dimensions. For example, velocity has an associated dimension vector in the MLT (Mass, length, time) system $[u] = (0, 1, -1)$. Note that the term which enforces dimensional homogeneity is optional and only relevant when the quantity of interest is dimensional.
end for
return $\boldsymbol{w}$, $\psi(\boldsymbol{A})$

Data availability statement. The code and input data required to reproduce the work reported in the manuscript are available in the GitHub repository https://github.com/nbeb/extracting_self-similarity_from_data. The grid turbulence data are available upon request.

Author contributions. N.B., L.M. and K.S. designed research; N.B. and K.S. performed research; N.B. analysed data; and N.B. and K.S. wrote the paper.

Appendix A. Data-driven similarity inference: implementation example

A pseudocode exemplifying the implementation of the proposed method for the case of a quantity of interest $q(s, t)$ is provided in Algorithm 1. Clarifications or examples are shown in smaller font size below each pseudocode line. More implementation examples

can be found in the code and input data that reproduce the problems considered in the paper. These are available in the GitHub repository https://github.com/nbeb/extracting_self-similarity_from_data.

Appendix B. Sensitivity and robustness

In this section, we present results from a sensitivity analysis with respect to the (i) number of provided profiles (stations), (ii) number of points at each station, (iii) number of points where the l_2 norms of (2.2) are evaluated (i.e. discretisation of the transformed coordinates grid) and (iv) noise in the input data, for the Blasius boundary layer problem.

(i) Number of provided stations: in the example detailed in § 3, the algorithm is provided with velocity profiles at 9 different streamwise stations. Table 1 shows the identified scalings when fewer stations at regular or irregular spacings are considered. The algorithm shows robust performance across all tested cases.

(ii) Number of points at each station (N_{meas}): in the example detailed in § 3, the velocity profiles at each streamwise station consist of 100 points. Table 2 shows the identified scalings when fewer points are available (50 and 20, respectively). The algorithm shows robust performance across all tested cases. A slight increase in the y scaling error is observed as the number of available points decreases.

(iii) Number of points in the transformed coordinates grid ($N_{\xi-points}$): as discussed in § 2, the l_2-norms in (2.2) are computed following interpolation on the transformed coordinates (ξ) grid. In the example detailed in § 3, $N_{\xi-points}$ is set equal to the number of points in the input data N_{meas}. Table 3 shows the identified scalings when this ratio changes. The algorithm shows robust performance across all tested cases.

(iv) Lastly, we consider the robustness of the algorithm to noise. To this end, we add zero-mean Gaussian noise to the data, with standard deviation proportional to the local velocity magnitude, scaled by a relative noise level ϵ. The identified scalings for three different noise levels are given in table 4. The input and algorithmically scaled data are shown in figure 7. While the algorithm's accuracy decreases with increasing noise, the higher levels tested exceed typical experimental uncertainties, and the method can be expected to perform reliably under realistic noise conditions, as demonstrated in the examples using experimental datasets.

Appendix C. Application to data from literature: turbulent wake

We attempt to recover the self-similar expressions for turbulent wakes using our method and experimental data from the literature. We consider the mean velocities at five stations in the wake of a 47 % solid slender plate, as measured by Cimbala *et al.* (1988), in turbulent conditions ($Re = 5000$) (see figures 8*a* and 8*b*). Figure 8(*c*) shows the collapse that is obtained via the proposed method, assuming transformations of the form $\tilde{y} = \alpha(x)y$ and $\tilde{u} = \beta(x)\bar{u} + \gamma(x)$. The identified transformations are regressed using `PySR`, with $\alpha = \alpha(x, y_w, y_{1/2})$ (with $y_w(x)$ defined as $\bar{u}(x, \pm y_w) = 0.99U_\infty$), $\beta = \beta(U_\infty, \bar{u}_{cntr}, x, \nu)$ and $\gamma = \gamma(U_\infty, \bar{u}_{cntr}, x, \nu)$. The library of operators consists of the four basic mathematical operations ($+, -, \times, \div$). The interpreted similarity transformations are

$$\tilde{y} = \alpha y = \frac{1.2188}{y_{1/2}(x)} y, \tag{C1a}$$

$$\tilde{u}(x, \tilde{y}) = \beta\bar{u} + \gamma = \frac{1}{U_\infty - \bar{u}_{cntr}}\bar{u} - \frac{\bar{u}_{cntr}}{U_\infty - \bar{u}_{cntr}} = 1 - \tilde{\zeta}, \tag{C1b}$$

Stations	$\tilde{y}/y$	$\tilde{u}/u$
[1, 2, 3, 4, 5, 6, 7, 8, 9]	$U_\infty^{0.5018}\nu^{-0.5018}x^{-0.4981}$	$U_\infty^{-0.9995}\nu^{-0.0005}x^{0.0005}$
[1, 3, 5, 7, 9]	$U_\infty^{0.5010}\nu^{-0.5010}x^{-0.4990}$	$U_\infty^{-0.9980}\nu^{-0.0020}x^{0.0020}$
[1, 3, 9]	$U_\infty^{0.5008}\nu^{-0.5008}x^{-0.4992}$	$U_\infty^{-0.9998}\nu^{-0.0002}x^{0.0002}$
[2, 6, 8]	$U_\infty^{0.5022}\nu^{-0.5022}x^{-0.4978}$	$U_\infty^{-0.9990}\nu^{-0.0010}x^{0.0010}$

Table 1. Data-driven identification of self-similarity in the Blasius boundary layer. Identified transformations for different number of available profiles (stations).

N_{meas}	$\tilde{y}/y$	$\tilde{u}/u$
100	$U_\infty^{0.5018}\nu^{-0.5018}x^{-0.4981}$	$U_\infty^{-0.9995}\nu^{-0.0005}x^{0.0005}$
50	$U_\infty^{0.5033}\nu^{-0.5033}x^{-0.4967}$	$U_\infty^{-0.9997}\nu^{-0.0003}x^{0.0003}$
20	$U_\infty^{0.5115}\nu^{-0.5115}x^{-0.4885}$	$U_\infty^{-1.0001}\nu^{0.0001}x^{-0.0001}$

Table 2. Data-driven identification of self-similarity in the Blasius boundary layer. Identified transformations for different number of measurements available at each station.

$N_{\xi\text{-points}}/N_{meas}$	$\tilde{y}/y$	$\tilde{u}/u$
0.25	$U_\infty^{0.4993}\nu^{-0.4991}x^{-0.5013}$	$U_\infty^{-0.9997}\nu^{-0.0003}x^{0.0003}$
1.00	$U_\infty^{0.5018}\nu^{-0.5018}x^{-0.4981}$	$U_\infty^{-0.9995}\nu^{-0.0005}x^{0.0005}$
4.00	$U_\infty^{0.5016}\nu^{-0.5016}x^{-0.4984}$	$U_\infty^{-0.9992}\nu^{-0.0008}x^{0.0008}$

Table 3. Data-driven identification of self-similarity in the Blasius boundary layer. Identified transformations for different discretisations of the transformed coordinates grid.

ϵ	$\tilde{y}/y$	$\tilde{u}/u$
0	$U_\infty^{0.5018}\nu^{-0.5018}x^{-0.4981}$	$U_\infty^{-0.9995}\nu^{-0.0005}x^{0.0005}$
0.001	$U_\infty^{0.5023}\nu^{-0.5023}x^{-0.4976}$	$U_\infty^{-1.0000}\nu^{-0.0000}x^{0.0000}$
0.01	$U_\infty^{0.5366}\nu^{-0.5366}x^{-0.4634}$	$U_\infty^{-0.9950}\nu^{-0.0050}x^{0.0050}$
0.1	$U_\infty^{0.6689}\nu^{-0.6689}x^{-0.3311}$	$U_\infty^{-0.9820}\nu^{-0.0180}x^{0.0180}$

Table 4. Data-driven identification of self-similarity in the Blasius boundary layer. Identified transformations for different levels of added noise.

which match (ignoring the arbitrary multiplicative constant) the self-similar expressions for turbulent wakes found in the literature (Tennekes & Lumley 1972; Townsend 1976; Pope 2000).

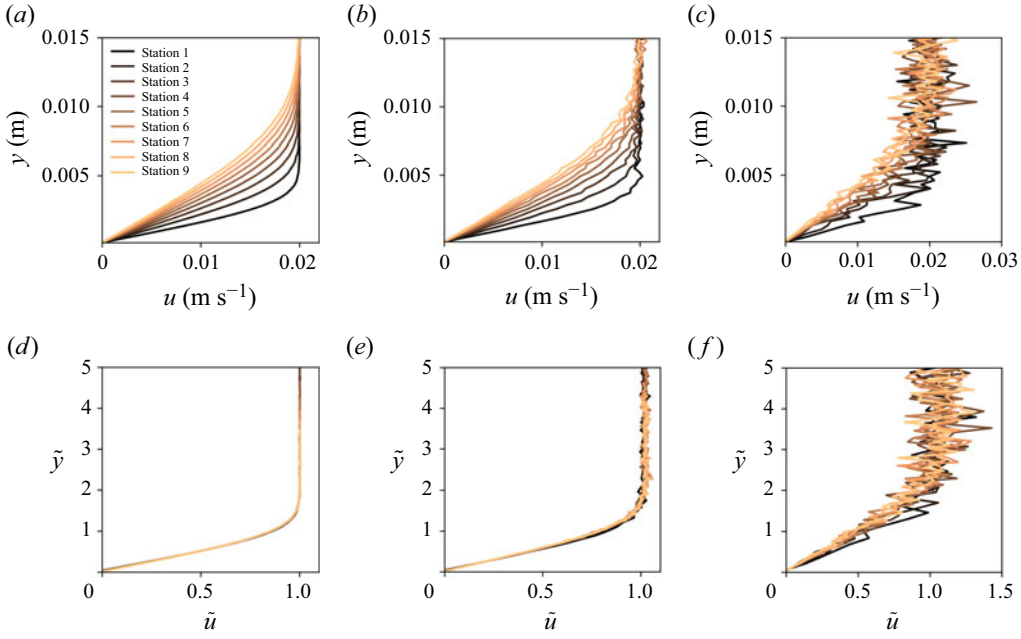

Figure 7. Data-driven identification of self-similarity in the Blasius boundary layer with added noise. Top row: input data, bottom row: input data collapsed with the algorithmically identified scalings (table 4). Panels show (*a*,*d*) $\epsilon = 0.001$, (*b*,*e*) $\epsilon = 0.01$, (*c*,*f*) $\epsilon = 0.1$.

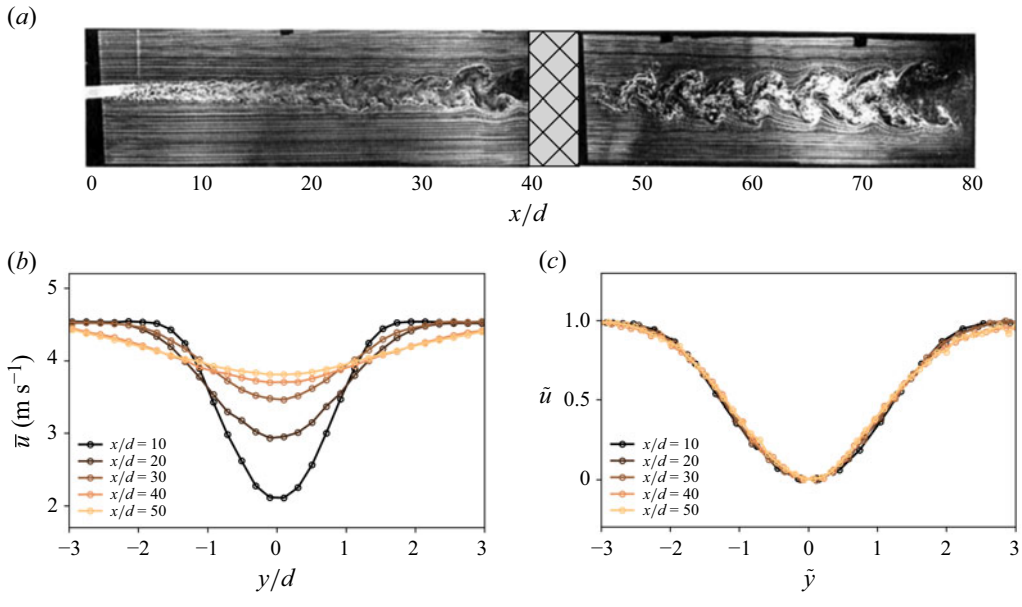

Figure 8. Data-driven identification of self-similarity in the wake of a porous plate. (*a*) Smoke-wire visualisation of a porous plate wake. Figure adapted from Cimbala *et al.* (1988). (*b*) Mean streamwise velocity profiles at different locations downstream of the plate. Experimental data extracted from Cimbala *et al.* (1988). (*c*) Algorithmically collapsed velocity profiles.

References

Bakarji, J., Callaham, J., Brunton, S.L. & Kutz, J.N. 2022 Dimensionally consistent learning with Buckingham Pi. *Nat. Comput. Sci.* **2** (12), 834–844.

Barenblatt, G.I. 1996 *Scaling, Self-Similarity, and Intermediate Asymptotics: Dimensional Analysis and Intermediate Asymptotics*. Cambridge University Press.

Barenblatt, G.I. & Gavrilov, A.A. 1974 On the theory of self-similar degeneracy of homogeneous isotropic turbulence. *Soviet Phys.-JETP* **38**, 399–402.

Batchelor, G.K. 1953 *the Theory of Homogeneous Turbulence*. Cambridge University Press.

Beaumard, P., Bragança, P., Cuvier, C., Steiros, K. & Vassilicos, J.C. 2024 Scale-by-scale non-equilibrium with Kolmogorov-like scalings in non-homogeneous stationary turbulence. *J. Fluid Mech.* **984**, A35.

Bekoglu, E., Bempedelis, N. & Steiros, K. 2024 On the formation of the primary and secondary vortex street instabilities. In *13th International Symposium on Turbulence and Shear Flow Phenomena (TSFP13)*.

Bekoglu, E., Bempedelis, N. & Steiros, K. 2025 *Formation of Turbulent Secondary Vortex Street in Absence of Vortex Shedding Instability*. Under review.

Bempedelis, N. & Steiros, K. 2022 Analytical all-induction state model for wind turbine wakes. *Phys. Rev. Fluids* **7** (3), 034605.

Berny, A., Deike, L., Séon, T. & Popinet, S. 2020 Role of all jet drops in mass transfer from bursting bubbles. *Phys. Rev. Fluids* **5** (3), 033605.

Birkhoff, G. 1960 *Hydrodynamics*, vol. 2234, Princeton University Press.

Blasius, H. 1908 Grenzschichten in flüssigkeiten mit kleiner reibung. *Zeitschrift für Mathematik Und Physik* **56**, 1–37.

Cantwell, B.J. 1978 Similarity transformations for the two-dimensional, unsteady, stream-function equation. *J. Fluid Mech.* **85** (2), 257–271.

Cantwell, B.J. 2002 *Introduction to Symmetry Analysis*. Cambridge University Press.

Cimbala, J.M., Nagib, H.M. & Roshko, A. 1988 Large structure in the far wakes of two-dimensional bluff bodies. *J. Fluid Mech.* **190**, 265–298.

Constantine, P.G., del Rosario, Z. & Iaccarino, G. 2017 Data-driven dimensional analysis: algorithms for unique and relevant dimensionless groups, arXiv preprint arXiv:1708.04303.

Cranmer, M. 2023 Interpretable machine learning for science with PySR and SymbolicRegression.jl, arXiv preprint arXiv:2305.01582.

Deike, L. 2022 Mass transfer at the ocean–atmosphere interface: the role of wave breaking, droplets, and bubbles. *Annu. Rev. Fluid Mech.* **54**, 191–224.

Deike, L., Ghabache, E., Liger-Belair, G., Das, A.K., Zaleski, S., Popinet, S. & Séon, T. 2018 Dynamics of jets produced by bursting bubbles. *Phys. Rev. Fluids* **3** (1), 013603.

Desai, K., Nachman, B. & Thaler, J. 2022 Symmetry discovery with deep learning. *Phys. Rev. D* **105** (9), 096031.

Duchemin, L., Popinet, S., Josserand, C. & Zaleski, Z. 2002 Jet formation in bubbles bursting at a free surface. *Phys. Fluids* **14** (9), 3000–3008.

Duraisamy, K., Brunton, S.L. & Taira, K. 2025 Introduction to turbulence & learning from data. In *Data Driven Analysis and Modeling of Turbulent Flows*, pp. 1–25. Elsevier.

Duraisamy, K., Iaccarino, G. & Xiao, H. 2019 Turbulence modeling in the age of data. *Annu. Rev. Fluid Mech.* **51**, 357–377.

Eggers, J. 1997 Nonlinear dynamics and breakup of free-surface flows. *Rev. Mod. Phys.* **69** (3), 865.

Eggers, J. & Fontelos, M.A. 2015 *Singularities: Formation, Structure, and Propagation*. vol. 53. Cambridge University Press.

Frisch, U. 1995 *Turbulence: the Legacy of AN Kolmogorov*. Cambridge University Press.

Fukami, K., Goto, S. & Taira, K. 2024 Data-driven nonlinear turbulent flow scaling with Buckingham Pi variables. *J. Fluid Mech.* **984**, R4.

Fukami, K. & Taira, K. 2024 Single-snapshot machine learning for super-resolution of turbulence. *J. Fluid Mech.* **1001**, A32.

Gañán-Calvo, A.M. 2017 Revision of bubble bursting: universal scaling laws of top jet drop size and speed. *Phys. Rev. Lett.* **119** (20), 204502.

George, W.K. 1989 *The Self-Preservation of Turbulent Flows and Its Relation to Initial Conditions and Coherent Structure*. Advances in Turbulence.

George, W.K. 1992 The decay of homogeneous isotropic turbulence. *Phys. Fluids A: Fluid Dyn.* **4** (7), 1492–1509.

Ghabache, E., Antkowiak, A., Josserand, C. & Séon, T. 2014 On the physics of fizziness: how bubble bursting controls droplets ejection. *Phys. Fluids* **26** (12), 121701.

GOTO, S. & VASSILICOS, J.C. 2016 Unsteady turbulence cascades. *Phys. Rev. E* **94** (5), 053108.
GROSS, D.J. 1996 The role of symmetry in fundamental physics. *Proc. Natl Acad. Sci.* **93** (25), 14256–14259.
HU, X., HONG, L., DEAN SMITH, M., NEUSIUS, T., CHENG, X. & SMITH, J.C. 2016 The dynamics of single protein molecules is non-equilibrium and self-similar over thirteen decades in time. *Nat. Phys.* **12** (2), 171–174.
JOFRE, L., DEL ROSARIO, Z.R. & IACCARINO, G. 2020 Data-driven dimensional analysis of heat transfer in irradiated particle-laden turbulent flow. *Intl J. Multiphase Flow* **125**, 103198.
KAMIYA, K., TAKEUCHI, T., KABEYA, N., WADA, N., ISHIMASA, T., OCHIAI, A., DEGUCHI, K., IMURA, K. & SATO, N.K. 2018 Discovery of superconductivity in quasicrystal. *Nat. Commun.* **9** (1), 154.
VON KARMAN, T. & HOWARTH, L. 1938 On the statistical theory of isotropic turbulence. Proceedings of the royal society of london. Series A. *Math. Phys. Sci.* **164** (917), 192–215.
KELLER, J.B. & MIKSIS, M.J. 1983 Surface tension driven flows. *SIAM J. Appl. Maths* **43** (2), 268–277.
KOLMOGOROV, A.N. 1941*a* Dissipation of energy in the locally isotropic turbulence. *Dokl. Akad. Nauk SSSR A* **32**, 16–18.
KOLMOGOROV, A.N. 1941*b* The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers. In *Dokl. Akad. Nauk SSSR A*, vol. 30, pp. 301–305.
KOLMOGOROV, A.N. 1941*c* On degeneration (decay) of isotropic turbulence in an incompressible viscous fluid. *Dokl. Akad. Nauk SSSR A* **31**, 538–540.
LAI, C.-Y., EGGERS, J. & DEIKE, L. 2018 Bubble bursting: universal cavity and jet profiles. *Phys. Rev. Lett.* **121** (14), 144501.
LANDAU, L.D. & LIFSHITZ, E.M. 1976 *Mechanics*. 3rd edn. Pergamon Press.
LIU, Z. & TEGMARK, M. 2022 Machine learning hidden symmetries. *Phys. Rev. Lett.* **128** (18), 180201.
LUNDGREN, T.S. 2002 Kolmogorov two-thirds law by matched asymptotic expansion. *Phys. Fluids* **14** (2), 638–642.
LUNDGREN, T.S. 2003 Kolmogorov turbulence by matched asymptotic expansions. *Phys. Fluids* **15** (4), 1074–1081.
MELDI, M. & VASSILICOS, J.C. 2021 Analysis of lundgren's matched asymptotic expansion approach to the Kármán–Howarth equation using the eddy damped quasinormal Markovian turbulence closure. *Phys. Rev. Fluids* **6** (6), 064602.
MENDEZ, P.F. & ORDÓÑEZ, F. 2004 Scaling laws from statistical data and dimensional analysis. *J. Appl. Mech.* **72** (5), 648–657.
MOTOTAKE, Y.-I. 2023 Extracting nonlinear symmetries from trained neural networks on dynamics data. In NeurIPS 2023 AI for Science Workshop.
OBERLACK, M. 1999 Similarity in non-rotating and rotating turbulent pipe flows. *J. Fluid Mech.* **379**, 1–22.
OBERLACK, M. 2001 A unified approach for symmetries in plane parallel turbulent shear flows. *J. Fluid Mech.* **427**, 299–328.
OBERLACK, M., CABOT, W., REIF, B.A.P. & WELLER, T. 2006 Group analysis, direct numerical simulation and modelling of a turbulent channel flow with streamwise rotation. *J. Fluid Mech.* **562**, 383–403.
OBERLACK, M., HOYAS, S., KRAHEBERGER, S.V., ALCÁNTARA-ÁVILA, F. & LAUX, J. 2022 Turbulence statistics of arbitrary moments of wall-bounded shear flows: a symmetry approach. *Phys. Rev. Lett.* **128** (2), 024502.
OBERLACK, M. & ROSTECK, A. 2010 New statistical symmetries of the multi-point equations and its importance for turbulent scaling laws. *Discrete Continuous Dyn. Syst.* **3**, 451–471.
OBLIGADO, M. & VASSILICOS, J.C. 2019 The non-equilibrium part of the inertial range in decaying homogeneous turbulence. *Europhys. Lett.* **127** (6), 64004.
OTTO, S.E., ZOLMAN, N., KUTZ, J.N. & BRUNTON, S.L. 2023 A unified framework to enforce, discover, and promote symmetry in machine learning, arXiv preprint arXiv:2311.00212.
PAKDEMIRLI, M. & YURUSOY, M. 1998 Similarity transformations for partial differential equations. *SIAM Rev.* **40** (1), 96–101.
POPE, S.B. 2000 *Turbulent Flows*. Cambridge University Press.
PRANDTL, L. 1904 *Proc. Third Int. Math. Cong. Heidelberg* 1904, pp. 484–491.
SAHA, S., GAN, Z., CHENG, L., GAO, J., KAFKA, O.L., XIE, X., LI, H., TAJDARI, M., KIM, H.A. & LIU, W.K. 2021 Hierarchical deep learning neural network (HiDeNN): an artificial intelligence (AI) framework for computational science and engineering. *Comput. Meth. Appl. Mech. Engng* **373**, 113452.
SANJAY, V., LOHSE, D. & JALAAL, M. 2021 Bursting bubble in a viscoplastic medium. *J. Fluid Mech.* **922**, A2.
SEDOV, L.I. 2018 Similarity and dimensional methods in mechanics. CRC press.
STEIROS, K. 2022*a* Balanced nonstationary turbulence. *Phys. Rev. E* **105** (3), 035109.
STEIROS, K. 2022*b* Turbulence near initial conditions. *Phys. Rev. Fluids* **7** (10), 104607.

TAM, C.K.W. 2019 A phenomenological approach to jet noise: the two-source model. *Phil. Trans. R. Soc. A* **377** (2159), 20190078.

TAYLOR, G.I. 1950*a* The formation of a blast wave by a very intense explosion i. theoretical discussion. *Proc. R. Soc. Lond. Ser. A Math. Phys. Sci.* **201** (1065), 159–174.

TAYLOR, G.I. 1950*b* The formation of a blast wave by a very intense explosion.-ii. the atomic explosion of 1945. *Proc. R. Soc. Lond. Ser. A. Math. Phys. Sci.* **201** (1065), 175–186.

TENNEKES, H. & LUMLEY, J.L. 1972 *A First Course in Turbulence*. MIT Press.

TOWNSEND, A.A. 1976 *the Structure of Turbulent Shear Flow*. Cambridge University Press.

VASSILICOS, J.C. 2015 Dissipation in turbulent flows. *Annu. Rev. Fluid Mech.* **47**, 95–114.

WHITHAM, G.B. 2011 *Linear and Nonlinear Waves*. John Wiley & Sons.

XIE, X., SAMAEI, A., GUO, J., LIU, W.K. & GAN, Z. 2022 Data-driven discovery of dimensionless numbers and governing laws from scarce measurements. *Nat. Commun.* **13** (1), 7562.

YAHIL, A. 1983 Self-similar stellar collapse. *Astrophys. J.* **265**, 1047–1055.

YANG, J., WALTERS, R., DEHMAMY, N. & YU, R. 2023 Generative adversarial symmetry discovery. In *International Conference on Machine Learning*, pp. 39488–39508. PMLR.

YUAN, Y. & LOZANO-DURÁN, A. 2025 Dimensionless learning based on information. arXiv preprint arXiv:2504.03927.

ZEFF, B.W., KLEBER, B., FINEBERG, J. & LATHROP, D.P. 2000 Singularity dynamics in curvature collapse and jet eruption on a fluid surface. *Nature* **403** (6768), 401–404.

ZEL'DOVICH, Y.B. & RAIZER, Y.P. 1967 *Physics of Shock Waves and High-Temperature Hydrodynamic Phenomena*. Academic Press.