Bayesian Machine Scientist to Compare Data Collapses for the Nikuradse Dataset

Ignasi Reichardt[1,*] Jordi Pallarès[2,†] Marta Sales-Pardo[1,‡] and Roger Guimerà[1,3,§]

[1]*Department of Chemical Engineering, Universitat Rovira i Virgili, Tarragona 43007, Catalonia, Spain*
[2]*Department of Mechanical Engineering, Universitat Rovira i Virgili, Tarragona 43007, Catalonia, Spain*
[3]*ICREA, Barcelona 08010, Catalonia, Spain*

Ever since Nikuradse's experiments on turbulent friction in 1933, there have been theoretical attempts to describe his measurements by collapsing the data into single-variable functions. However, this approach, which is common in other areas of physics and in other fields, is limited by the lack of rigorous quantitative methods to compare alternative data collapses. Here, we address this limitation by using an unsupervised method to find analytic functions that optimally describe each of the data collapses for the Nikuradse dataset. By descaling these analytic functions, we show that a low dispersion of the scaled data does not guarantee that a data collapse is a good description of the original data. In fact, we find that, out of all the proposed data collapses, the original one proposed by Prandtl and Nikuradse over 80 years ago provides the best description of the data so far, and that it also agrees well with recent experimental data, provided that some model parameters are allowed to vary across experiments.

In the early 1930s, Nikuradse conducted experiments to measure the friction sustained by a turbulent flow in a rough pipe [1]. With remarkable accuracy, he measured the dependency of the friction factor f on two dimensionless quantities, the Reynolds number Re and the relative roughness, that is, the ratio r/k between the size of the irregularities r and the radius of the pipe k (Fig. 1). Over eight decades later, and despite the fundamental and practical importance of the problem, the functional relationship $f = h(\text{Re}, r/k)$ remains unknown.

Numerous works [1–7] have attempted to solve the problem by collapsing Nikuradse's data into a function $\bar{f} = \bar{h}(x)$ that depends on a single-variable x combining both the Reynolds number (or the so-called turbulent Reynolds number, Re_τ) and the relative roughness. For example, Prandtl and Nikuradse [3] proposed the collapse $f^{-1/2} + 2\log(r/k) = \bar{h}(\text{Re}_\tau \cdot r/k)$, which suggests that a transformed turbulent friction factor only depends on the product of the turbulent Reynolds number and the relative roughness [1,2].

Data collapses such as this are common in many areas, and aim at establishing the combinations of independent variables that actually affect the dependent variable, disregarding the exact functional form of the dependency. These approaches have been particularly useful in critical phenomena, because the scaling behavior of critical systems implies that, close to the critical point, all relevant functions must be generalized homogeneous functions and therefore collapse under particularly simple scaling transformations [8,9]. This realization has led to important insights in physics and other disciplines, even for systems that are not at a critical point. For example, data collapses

have enabled the characterization of the statistical laws that govern the growth of human organizations [10,11].

Perhaps the major challenge of approaches based on scaling and data collapse is the fact that, in principle, many alternative collapses are possible, especially when the data are noisy. In the Nikuradse dataset, for example, different arguments may lead to very different collapses (Fig. 2, Table I). In such cases, the goodness of a collapse is typically evaluated qualitatively by how well the empirical data obtained under different conditions fall onto a single well-defined "scaling function" $\bar{h}$ [12]. In the few remarkable instances in which the goodness of the collapse is quantified, existing methods rely on interpolation of the datasets to approximate the scaling function, and on measuring deviations in the collapsed data [13]. Additionally, they allow only to compare a scaling function with different
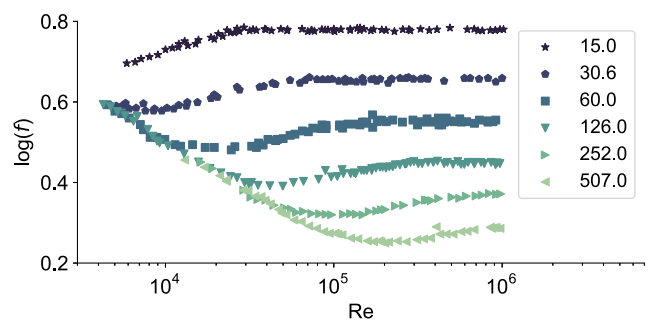

FIG. 1. The Nikuradse dataset, displaying the turbulent friction factor f as a function of the Reynolds number Re for different values of the inverse relative roughness k/r, which are indicated in the legend.
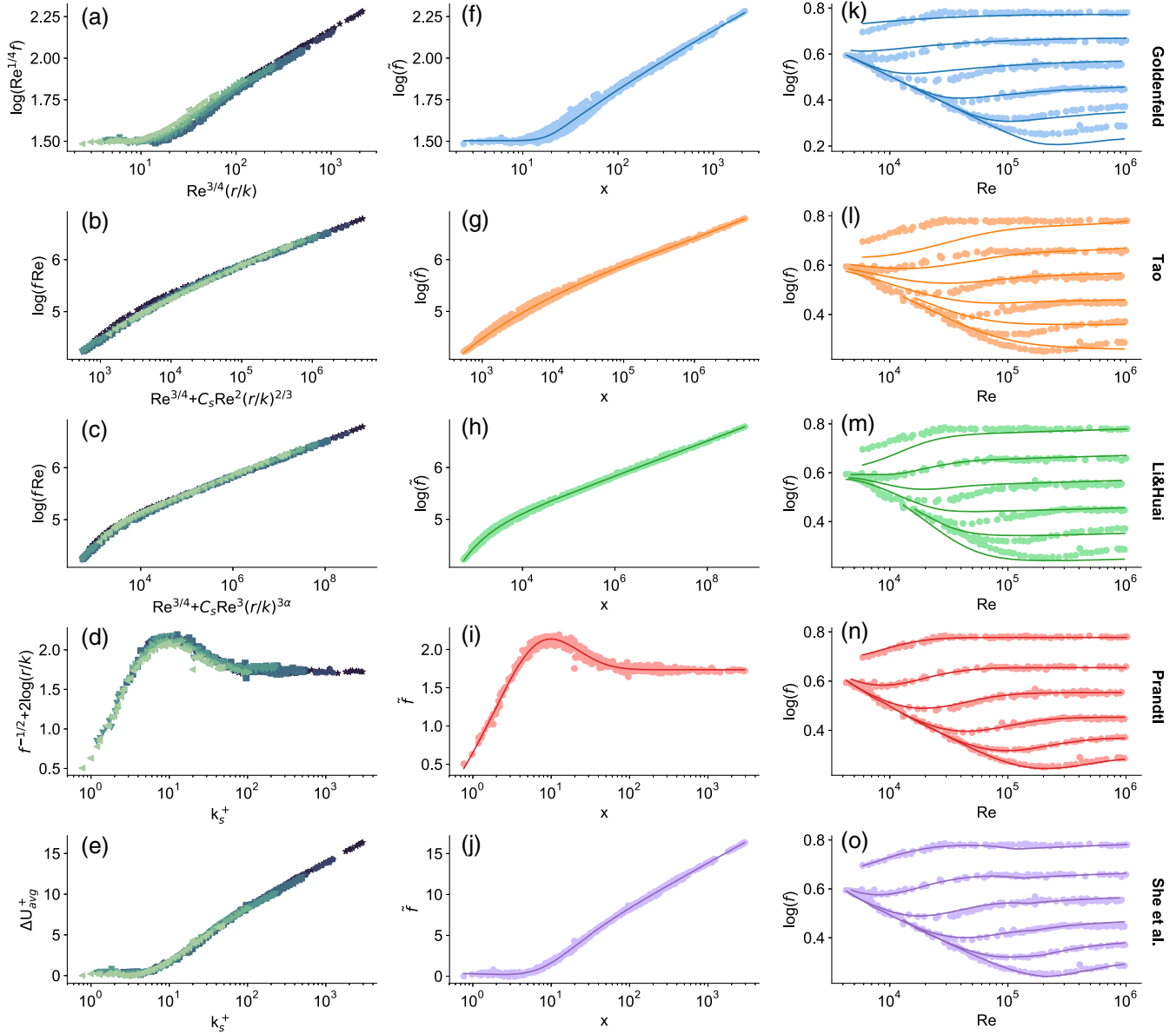
FIG. 2. (a)–(e) Nikuradse data collapsed according to each approach (Goldenfeld [5], Tao [6], Li and Huai [7], Prandtl [2], and She *et al.* [4]). (f)–(j) Models for the collapsed data. The solid line represents the most plausible closed-form mathematical model identified by the Bayesian machine scientist for each collapse. (k)–(o) Unscaled models. Same as the middle column but unscaled to the original variables.

parameter values, but not to compare scaling functions that are mathematically different.

Here, we argue that the goodness of a data collapse should be measured on the original unscaled data because (i) the scaling function distorts the data by stretching some regions and compressing others, thus potentially obscuring deviations in the collapse; (ii) the distortion can lead to collapse functions with very different ranges, thus making comparison between collapses meaningless. We propose a rigorous alternative to evaluate the quality of data collapses, and show that scaling functions that seem better at collapsing the data often do not describe the original

unscaled data well. Our approach proceeds by first obtaining the most plausible functional form for the scaling function $\bar{h}$, for which we use an unsupervised algorithm that explores systematically the space of possible functional forms for $\bar{h}$; following Refs. [14,15] we call this algorithm a machine scientist. Next, we unscale this function and quantify how it fits the original data, as opposed to the collapsed data. For the Nikuradse dataset, we find that the best collapse is the one originally proposed by Nikuradse and Prandtl [1,2]. We also study what is the applicability of the best models to recent datasets on turbulent friction in rough pipes.

TABLE I. Summary of the data collapses for the Nikuradse dataset.

Reference	Scaling variable x	Scaling function $\bar{h}$	Most plausible model $h(x)$
Goldenfeld [5]	$\mathrm{Re}^{3/4} r/k$	$f\,\mathrm{Re}^{1/4}$	$(1/c_3)\log\left[(c_1 x)^{c_2+x} + e^{c_1}\right]$
Tao[a] [6]	$\mathrm{Re}^{3/4} + C_s\mathrm{Re}^2(r/k)^{2/3}$	$f\,\mathrm{Re}$	$c_1(c_2 c_2^{-x} x + c_3) + x$
Li and Huai[b] [7]	$\mathrm{Re}^{3/4} + C_s\mathrm{Re}^3(r/k)^{3\alpha}$	$f\,\mathrm{Re}$	$c_1[c_2 + c_3 x(c_1 + c_4^x)]$
Prandtl[c] [2]	$\mathrm{Re}\sqrt{f/32}(r/k)$	$f^{-1/2} + 2\log(r/k)$	$c_1[c_2(c_1^x + c_3) + c_4^x]$
She et al.[c,d] [4]	$\mathrm{Re}\sqrt{f/32}(r/k)$	$(1/\kappa)\ln(\mathrm{Re}\sqrt{f/32}) + B - 1/(2\sqrt{f/32})$	$[-c_1 + \log(c_1^x c_2^2 + c_3 + x)]e^{c_4}$

[a]$C_s = 3 \times 10^{-5}$.
[b]$C_s = 1 \times 10^{-8}$ and $\alpha = 1/3 + \eta/2$, with $\eta = 0.02$.
[c]The independent variable defined by Prandtl and used by She et al. is denoted k_s^+ in Figs. 2(d) and 2(e).
[d]Following She et al., this function is denoted $\Delta U_{\mathrm{avg}}^+$ in Fig. 2(e). B is a Reynolds-dependent correction applied for $\mathrm{Re}\sqrt{f/32} < 5000$ (see definition in [4]).

Let us first formalize the problem. The data collapse hypothesis posits that, under the appropriate transformation, all the observed data D follow a single law $\bar{f} = \bar{h}(x)$ [16]. To identify the most plausible candidate for $\bar{h}$, we start by considering the probability $p(h_i|D)$ that an expression h_i is the correct one given the data, which can be written as [15]

$$p(h_i|D) = \frac{\exp\left[-\mathcal{L}(h_i)\right]}{Z}. \tag{1}$$

Here, $Z = \sum_i \exp\left[-\mathcal{L}(h_i)\right]$ is the partition function, and $\mathcal{L}(h_i) = -\log p(h_i, D)$ is the description length of the model [15,17]. The description length plays the role of a free energy in a physical system, and the model with minimum description length is the most plausible one. The description length cannot, in general, be calculated exactly; however, it can be approximated as $\mathcal{L}(h_i) = B(h_i)/2 - \log p(h_i)$, where $B(h_i)$ is the Bayesian information criterion of expression h_i [18,19] and the prior $p(h_i)$ is the probability assigned to h_i before any data are observed. This prior acts as an expression regularizer and requires certain hypotheses; following previous work [15] we use the maximum entropy $p(h_i)$ that is consistent with empirically observed frequencies of each operation [20].

We explore the space of possible mathematical expressions using the Metropolis algorithm, by means of what has been called a Bayesian machine scientist [15,20]. The Bayesian machine scientist draws upon concepts developed in symbolic regression [22], exponential random graphs [23], and Bayesian network sampling [15,24,25]; it is guaranteed to asymptotically sample from the stationary distribution $p(h_i|D)$ and is consistent, that is, given enough data it will assign the highest plausibility (the shortest description length) to the correct model with probability approaching one. Among all models explored by the machine scientist using the Metropolis algorithm, we select the model with the minimum description length as the most plausible one [26]. Specifically, we let the machine scientist sample expressions for each of the collapses $(\bar{f}, x)$ proposed for Nikuradse's data (Table I). From these samplings, we obtain the most plausible expression for each data collapse [Figs. 2(f)–2(j)].

Despite fitting the collapsed data tightly, most of the functions do not reproduce the features of the original data when unscaled, particularly in the low and intermediate Reynolds regimes [Figs. 2(k)–2(o)]. This is important because, as mentioned earlier, collapse and scaling theories are usually evaluated by how close the data appear to be in the collapse. Our results for the Nikuradse data show how this can be misleading; some of the data collapses that have been proposed are effectively a reduced-size view, in which the curves for each roughness are stretched until the separation in the vertical axis is no longer visible. This can be achieved by multiplying both the scaling variable and the scaling function by a quantity that spans a broad range, typically a power of the Reynolds number. As we show here, a function fitting such a stretched curve does not necessarily recover the correct $f = h(\mathrm{Re}, r/k)$ dependency when unscaled.

In Fig. 3, we show the mean absolute error for each of the scalings when the original variables are recovered. Unlike the deviations that one measures in the collapsed data, we argue that this quantity is a comparable and reliable measure of the true goodness of the collapse. From this, we conclude that only the data collapses proposed by Prandtl [2] and by She et al. [4] are accurate representations of the Nikuradse dataset in all regions. The remaining data collapses are only good representations of some regimes.

The collapses proposed by Prandtl [2] and by She et al. [4] both use as their scaling variable the roughness
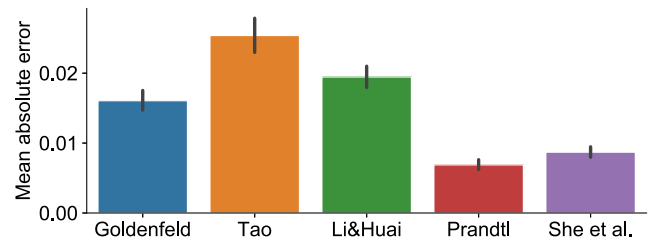

FIG. 3. Mean absolute error of the most plausible model for each data collapse, calculated on the original unscaled variables of the Nikuradse dataset [Figs. 2(k)–2(o)]. The error bars in indicate the 95% confidence interval for the mean.

Reynolds number k_s^+, and the scaling function depends, again on both cases, on the inverse square root of f. She *et al.* argued that the spread in the Prandtl-scaled data about $k_s^+ \sim 1$ is too large for it to be considered a good collapse. Therefore, they introduced an extra parameter p, which encodes the behavior at the transitionally rough regime, about $k_s^+ \sim 5$ in the scaled data (see Fig. 5 in [4]). Effectively, p is an exponent that tunes the sharpness of the transition between the different turbulent regimes with the roughness. Moreover, an *ad hoc* correction was added to improve the collapse (the so-called B term). This correction applies only to $\mathrm{Re}_\tau \equiv \mathrm{Re}\sqrt{f/32} < 5000$ and thus makes the scaling function defined piecewise, introducing a number of extra parameters. Despite these refinements, our approach suggests that the Prandtl collapse still provides smoother and more accurate unscaled curves.

To further understand the validity of the original scaling by Prandtl and the need for the corrections introduced by She *et al.*, we finally turn to the main criticism against Prandtl's scaling, namely that it does not hold for more recent empirical data from the Princeton experiments with honed [27] and commercial [28] pipes. These experiments used smoother pipes than those used by Nikuradse ($k/r = 8716$ and $k/r = 8065$, respectively), and explored flows with Re up to 2×10^7. To investigate the applicability of Prandtl's collapse to these data, we check whether the machine scientist is able to generate a new set of expressions fitting both Nikuradse's and the Princeton data (Fig. 4). Since the newer data seem to deviate from both Prandtl's and She *et al.*'s collapses (Fig. 4), we search for a single mathematical expression fitting the three datasets but with potentially different parameter values, accounting for different details arising, for example, from the geometry of the irregularities in each pipe. This is possible thanks to the fact that the machine scientist samples mathematical expressions and, once an expression is selected, it can be fit separately to each dataset. The plausibility of the expression is then given by the total description length

$$\mathcal{L}(h_i) = \frac{1}{2}\sum_d B^d(h_i) - \log p(h_i), \qquad (2)$$

where $B^d(h_i)$ is the Bayesian information criterion calculated on dataset d.

The most plausible model fitting both the Nikuradse and the Princeton datasets using Prandtl's collapse is

$$\bar{f}(x) = \sinh\left[c_1 + (c_2 + x)e^{-(c_3 x^2)^{c_4}}\right], \qquad (3)$$

whereas the most plausible model for She's collapse is

$$\bar{f}(x) = \left\{c_1 \log\left[(c_2 x^{c_3} + c_4)^2\right]\right\}^{c_1}. \qquad (4)$$

As above, the machine scientist uncovers, for each collapse, several models that are almost equally plausible. Therefore, the models in Eqs. (3) and (4) should be taken as just an instance of the models describing each of the data
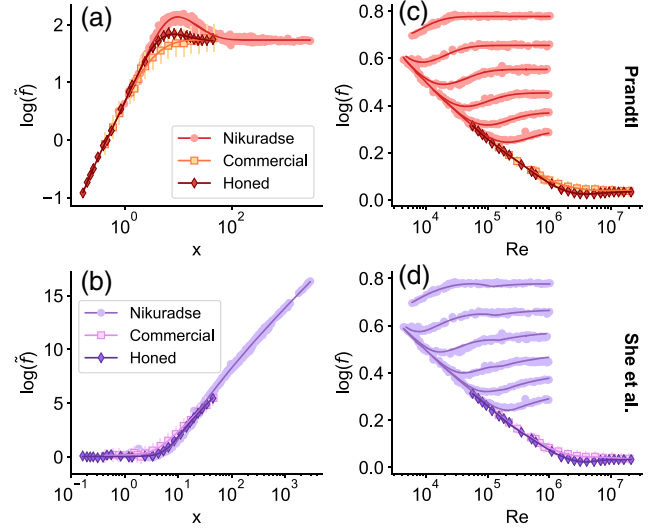

FIG. 4. (a),(b) Models for the Nikuradse and Princeton (honed pipe with $k/r = 8716$ and commercial pipe with $k/r = 8065$) datasets collapsed according to the approaches proposed by Prandtl [2] (top) and She *et al.* [4] (bottom). The solid lines represent the most plausible closed-form mathematical model identified by the Bayesian machine scientist for all datasets simultaneously. (c),(d) Unscaled models.

collapses, and it is beyond our aims to discuss their putative physical meaning here. However, in both scaling laws we get two parameters whose value is shared (or very similar) among pipe types, whereas two others take pipe-specific values over a broader range (Table II). Therefore, it seems plausible that two parameters are universal, whereas the other two may be dependent on properties of the material or the irregularities in it. In any case, the addition of the Princeton experiments does not seem to justify the corrections introduced by She *et al.* Indeed, our approach suggests that the commercial and the honed pipes are equally well described by both data collapses (Fig. 5).

Machine learning tools are increasingly applied to shed light into physics problems [29], from detecting phase transitions [30,31] to approximating the wave function of many-body quantum systems [32]. Here, we have used a Bayesian machine scientist [15], which automatically uncovers closed-form mathematical equations from data, to compare data collapses for the Nikuradse dataset. The

TABLE II. Parameter values for the expressions in Eqs. (3) and (4).

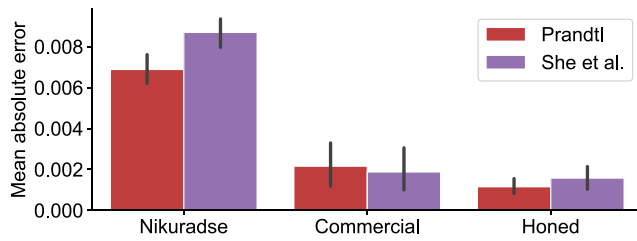| Scaling	Dataset	c_1	c_2	c_3	c_4
Prandtl	Nikuradse	1.3	-3.6	3.1	0.69
	Honed	1.3	-4.2	6.0	0.70
	Commercial	1.3	-14	1.5×10^3	0.62
She	Nikuradse	0.79	0.15	3.6	1.1
	Honed	0.71	0.14	4.1	1.0
	Commercial	0.77	0.34	2.7	1.1

FIG. 5. Mean absolute error of the most plausible model for Prandtl's and for She *et al.*'s data collapses, calculated on the original unscaled variables of the Nikuradse dataset and the two Princeton datasets (right column in Fig. 4). The error bars in indicate the 95% confidence interval for the mean.

machine scientist enables us to sample a wealth of analytic expressions that describe each data collapse, using only probabilistic model-selection arguments and without introducing any heuristics or biases. We then use the expressions obtained by the Bayesian machine scientist to evaluate the goodness of the models on the unscaled data; as we show, the common practice of evaluating goodness of fit on the collapsed data leads to misleading conclusions.

In the Nikuradse dataset, our approach favors the original data collapse proposed by Prandtl over 80 years ago and, conversely, disfavors the data collapses proposed more recently [4–7]. Our method suggests that these more recent approaches produce an apparent collapse of the data mainly by *compressing* the transition region between flow regimes; they only describe the experimental data in limiting regimes but not in these transition regions (although it is fair to note that their goal was, precisely, to describe some limiting behaviors and not all regimes). Our approach also shows that, contrary to what has been suggested, Prandtl's collapse is compatible with more recent experiments of turbulent friction, provided that some parameters in the scaling function are allowed to depend on the details of the pipe.

Although here we have focused on the Nikuradse dataset, we think that our approach can be applied to other systems on which data collapses are used to understand the underlying physics or mechanisms. Scalings and data collapses are, indeed, powerful tools; we argue that, combined with rigorous model selection and machine learning, they have the potential to become even more insightful.

*ignasi.reichardt@urv.cat
†jordi.pallares@urv.cat
‡marta.sales@urv.cat
§roger.guimera@urv.cat

[1] J. Nikuradse, https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19930093938.pdf (1933).
[2] L. Prandtl, https://ntrs.nasa.gov/search.jsp?R=19930094697 (1933).
[3] In his original paper, Nikuradse already presented his data scaled according to Prandtl's collapse. However, this collapse is mentioned in [4], where Bodenschatz attributes it to Prandtl, Nikuradse's supervisor.
[4] Z.-S. She, Y. Wu, X. Chen, and F. Hussain, New J. Phys. **14**, 093054 (2012).
[5] N. Goldenfeld, Phys. Rev. Lett. **96**, 044503 (2006).
[6] J. Tao, Phys. Rev. Lett. **103**, 264502 (2009).
[7] S. Li and W. Huai, PLoS One **11**, e0154408 (2016).
[8] H. E. Stanley, Rev. Mod. Phys. **71**, S358 (1999).
[9] G. I. Barenblatt, *Scaling*, Cambridge Texts in Applied Mathematics (Cambridge University Press, Cambridge, 2003).
[10] M. H. R. Stanley, L. A. N. Amaral, S. V. Buldyrev, S. Havlin, H. Leschhorn, P. Maass, M. A. Salinger, and H. E. Stanley, Nature (London) **379**, 804 (1996).
[11] Y. Lee, L. A. N. Amaral, M. Meyer, D. Canning, M. Meyer, and H. E. Stanley, Phys. Rev. Lett. **81**, 3275 (1998).
[12] We refer to $\bar{h}$ as the *scaling function* although, strictly speaking, this function is not always a generalized homogeneous function.
[13] S. M. Bhattacharjee and F. Seno, J. Phys. A **34**, 6375 (2001).
[14] J. Evans and A. Rzhetsky, Science **329**, 399 (2010).
[15] R. Guimerà, I. Reichardt, A. Aguilar-Mogas, F. A. Massucci, M. Miranda, J. Pallarès, and M. Sales-Pardo, Sci. Adv. **6**, eaav6971 (2020).
[16] In general, x could be a vector of independent variables instead of just one variable as in Nikuradse's data.
[17] P. D. Grünwald, *The Minimum Description Length Principle* (The MIT Press, Cambridge, Massachusetts, 2007).
[18] G. Schwarz, Ann. Stat. **6**, 461 (1978).
[19] T. Ando, *Bayesian Model Selection and Statistical Modeling* (CRC Press, Boca Raton, 2010).
[20] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.124.084503 for more details on the implementation of the Bayesian machine scientist, which includes Ref. [21].
[21] D. J. Earl and M. W. Deem, Phys. Chem. Chem. Phys. **7**, 3910 (2005).
[22] M. Schmidt and H. Lipson, Science **324**, 81 (2009).
[23] A. Caimo and N. Friel, Soc. Networks **33**, 41 (2011).
[24] S. Horvát, E. Czabarka, and Z. Toroczkai, Phys. Rev. Lett. **114**, 158701 (2015).
[25] R. Fischer, J. C. Leitão, T. P. Peixoto, and E. G. Altmann, Phys. Rev. Lett. **115**, 188701 (2015).
[26] In practice, all sampled expressions describe the data similarly well, and none of the results below depend on which expressions we choose.
[27] M. A. Shockling, J. J. Allen, and A. J. Smits, J. Fluid Mech. **564**, 267 (2006).
[28] L. I. Langelandsvik, G. J. Kunkel, and A. J. Smits, J. Fluid Mech. **595**, 323 (2008).
[29] L. Zdeborová, Nat. Phys. **13**, 420 (2017).
[30] J. Carrasquilla and R. G. Melko, Nat. Phys. **13**, 431 (2017).
[31] E. P. L. van Nieuwenburg, Y.-H. Liu, and S. D. Huber, Nat. Phys. **13**, 435 (2017).
[32] G. Carleo and M. Troyer, Science **355**, 602 (2017).