Learning Effective Good Variables from Physical Data

machine learning & knowledge extraction

MDPI

Article

Learning *Effective Good* Variables from Physical Data

Giulio Barletta, Giovanni Trezza and Eliodoro Chiavazzo *

Department of Energy, Politecnico di Torino, C.so Duca degli Abruzzi 24, 10129 Torino, Italy; giulio.barletta@polito.it (G.B.); giovanni.trezza@polito.it (G.T.)
* Correspondence: eliodoro.chiavazzo@polito.it

Abstract: We assume that a sufficiently large database is available, where a physical property of interest and a number of associated ruling primitive variables or observables are stored. We introduce and test two machine learning approaches to discover possible groups or combinations of primitive variables, regardless of data origin, being it numerical or experimental: the first approach is based on regression models, whereas the second on classification models. The variable group (here referred to as the new effective good variable) can be considered as successfully found when the physical property of interest is characterized by the following effective invariant behavior: in the first method, invariance of the group implies invariance of the property up to a given accuracy; in the other method, upon partition of the physical property values into two or more classes, invariance of the group implies invariance of the class. For the sake of illustration, the two methods are successfully applied to two popular empirical correlations describing the convective heat transfer phenomenon and to the Newton's law of universal gravitation.

Keywords: machine learning in physics; primitive variable analysis; physical property invariance; feature grouping

check for updates

Citation: Barletta, G.; Trezza, G.; Chiavazzo, E. Learning *Effective Good* Variables from Physical Data. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1597–1618. https://doi.org/10.3390/make6030077

Academic Editor: Andreas Holzinger

Received: 6 May 2024
Revised: 1 July 2024
Accepted: 9 July 2024
Published: 12 July 2024

1. Introduction

Theoretical modeling and numerical simulations have become invaluable tools for the current scientific and technological advancement across various fields [1]. In essence, models make use of a mathematical description of the physical laws by establishing the relationships between physical variables. The latter variables have the key role of describing, in a complete and non-redundant manner, a system of interest. As such, their correct identification is never a trivial task, especially in systems with little prior knowledge [2–4].

Often, in order to effectively model complex systems, it is favorable to search for a more convenient description with a reduced number of effective variables [5,6]. In other words, the system behavior is not described by the directly accessible physical variables, but rather by some groups or combinations. In this respect, back in the late 19th century, the Buckingham theorem was introduced [7–9]. In the latter approach, based on dimensional analysis, it is possible to mix the primitive physical variables, thus creating fewer dimensionless numbers which are effectively relevant.

Over the years, sophisticated methodologies have also been developed based on machine learning, data mining and other data-driven approaches [10–12]. Some other approaches aim to unlock the discovery of symbolic expressions accurately matching data derived from an unknown function. This problem has been tackled with a number of methods [13,14], including sparse regression [15–18], genetic algorithms [19–21] and physics-inspired algorithms like AI Feynman, introduced by Udrescu and Tegmark [22]. Inspired by the latter, some authors of this work presented a multi-objective optimization procedure for reducing the set of composition-based material descriptors by optimally mixing them in power combination form. This resulted in improved classification performances, as demonstrated in a case study focused on superconductors [23]. Moreover, the same procedure was applied by Bonke et al. [24] to identify an effective reduced set of

variables in a micelle-based photocatalytic system for solar fuels production. Specifically, the algorithm allowed for analytically mixing five physical primitive variables into two synthetic features for the optimal binary separation of the experimental samples according to their performances. The practical benefit of such an approach lies in its capability to replace an experimental sample achieving high performance with an alternative combination of its elemental components. This gives the possibility of reducing the most expensive ones, and re-balancing the others, at no cost on the overall performance. To our knowledge, developing a robust method for identifying symmetries in descriptors, and thereby determining analytically mixed features that govern a specific phenomenon, remains an unresolved issue [2,22]. Within this framework, the scientific question we aim to address is the same of interpretability algorithms like SHAP [25], i.e., identifying relevant features over a trained ML model. However, SHAP works in the original features space, even when aggregations of such features more effectively would explain the phenomenon of interest. Other recent works deal with methods for feature selection/removal by means of genetic algorithms (but not mixing) [26], for identification of the minimal intrinsic variables (but not analytically) [3]. Also, Udrescu and Tegmark [22] consider only a few possible symmetries in the data.

In this work, we present a general and automated methodology, suitable for regression tasks, able to identify groups and/or group sets of variables in the power form $x_i^{\alpha_1} x_j^{\alpha_2} \cdots x_m^{\alpha_p}$. We show its effectiveness on popular thermo-fluid dynamics correlations (i.e., Dittus–Boelter and Gnielinski). Furthermore, we also demonstrate that the approach can be easily extended to a more general functional form, proving its ability on Newton's law of universal gravitation. Also, we employ the optimal feature mixing procedure in ref. [23] for classification tasks, showing its successful application on the former two problems of this study. We refer to the resulting variable groups/sets as effective good variables, since either the examined property of interest or the class effectively depend on groups or sets of features rather than on individual features considered separately. Thus, such effective good variables comply with the definition adopted by Chen et al. [2], as a "complete and non-redundant description of the relevant system". An overview of the two proposed methodologies is depicted in Figure 1.

The paper is organized in sections as follows. Section 2 presents the methodology for searching good variables in regression models and describes the implementation of classification models for optimal variable mixing towards class separation. This section is further subdivided into specific techniques, including the consideration of single and multiple invariant groups in power form, as well as a broader exploration into non-power forms. Section 3 presents numerical examples and discussions related to our study. We detail the procedure for generating the three datasets employed in this study, based on the three functional forms analyzed here (i.e., Dittus–Boelter correlation, Gnielinski correlation, and Newton's law of universal gravitation). Here, we also provide a comprehensive evaluation of our approach. Finally, in Section 4, we draw conclusions providing an overview of the main contributions. We point out that, since this work is methodological, it deals only with numerically generated data. However, those methodologies are blessed by generality and, as such, are agnostic to the data origin.

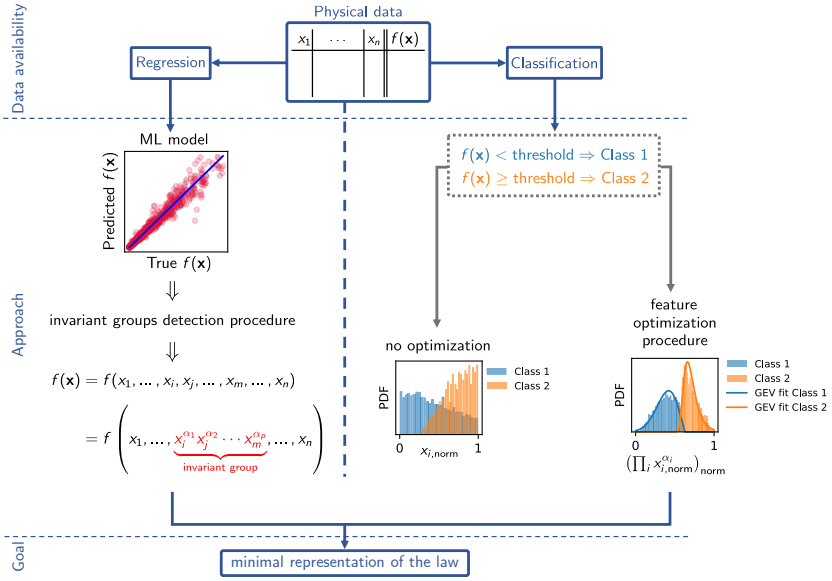

Figure 1. Overview of the protocol used to detect possible symmetries of a target property of interest with respect to its input variables, utilizing only data and ignoring the analytical functional dependence. Two distinct methodologies are presented: the former for identifying, in regression tasks, invariant groups in the form $x_i^{\alpha_1} x_j^{\alpha_2} \cdots x_m^{\alpha_p}$, among others; the latter for identifying, in classification tasks, one or several mixed features as power combinations of the input variables to achieve an optimal class separation.

2. Methods

2.1. Problem Statement

Within the context of analyzing complex physical phenomena, discovering effective combinations or groups of primitive variables that characterize a particular physical property of interest is highly desirable. Indeed, this deeper insight not only enhances understanding per se, but also offers practical benefits for optimization purposes, as already pointed out above. Herein, we introduce a methodology to detect invariant groups/sets of variables for regression tasks. In some cases, invariant groups may not be identifiable for regression tasks. Therefore, we also outline a method for identifying similar invariant groups for classification tasks, including the formulation of an analytical classifier.

2.2. Searching for Good Variables by Regression Models

2.2.1. Single Invariant Group in Power Form

Let $f(\mathbf{x}) = f(x_1, \dots, x_n)$ denote a function depending on n physical variables $x_1, \dots, x_n$. If $f(\mathbf{x})$ is invariant with respect to a group of p variables in the form $(x_i^{\alpha_1} x_j^{\alpha_2} \cdots x_m^{\alpha_p})$, $\alpha_1, \alpha_2, \dots, \alpha_p \in \mathbb{R}$, when this group is kept at a constant value $\bar{c}$, regardless of the value of each component $x_i, x_j, \dots, x_m$, $f(\mathbf{x})$ does not change. Stated differently, the above invariance condition requires:

$$\alpha_1 \ln(x_i) + \alpha_2 \ln(x_j) + \cdots + \alpha_p \ln(x_m) = c \tag{1}$$

($c = \ln(\bar{c})$), which upon differentiation yields:

$$\alpha_1 \frac{\mathrm{d}x_i}{x_i} + \alpha_2 \frac{\mathrm{d}x_j}{x_j} + \cdots + \alpha_p \frac{\mathrm{d}x_m}{x_m} = 0. \tag{2}$$

We can thus construct the $(1 \times p)$ matrix $\mathbf{B}$ at a generic point $\overline{\mathbf{x}_0} = (x_{i,0}, x_{j,0}, \dots, x_{m,0})$, as

$$\mathbf{B} = \left[\frac{\alpha_1}{x_{i,0}}, \frac{\alpha_2}{x_{j,0}}, \cdots, \frac{\alpha_p}{x_{m,0}} \right]. \tag{3}$$

Let $\mathbf{K}$ be a matrix whose columns form an ortho-normal basis in the null space (or kernel) of $\mathbf{B}$. Hence, the condition of invariance of $f(\mathbf{x})$ with respect to the group $(x_i^{\alpha_1} x_j^{\alpha_2} \cdots x_m^{\alpha_p})$ can be recast into an orthogonality condition in the configuration space between the normalized gradient of $f(\mathbf{x})$ and each column of $\mathbf{K}$, namely

$$(\nabla \tilde{f})_{\overline{\mathbf{x}_0}} \cdot \mathbf{K} = 0, \tag{4}$$

where

$$\nabla \tilde{f} = \frac{\nabla f}{\text{norm}(\nabla f)}. \tag{5}$$

Coefficients $\alpha_1, \alpha_2, \ldots, \alpha_p$ can be conveniently normalized, thus yielding the following algebraic system:

$$\begin{cases} (\nabla \tilde{f})_{\overline{\mathbf{x}_0}} \cdot \mathbf{K} = 0 \\ \alpha_1^2 + \alpha_2^2 + \cdots + \alpha_p^2 = 1 \end{cases}. \tag{6}$$

If the non-linear system in Equation (6) is satisfied for the same exponents ($\overline{\alpha_1}, \overline{\alpha_2}, \ldots, \overline{\alpha_p}$) over all the domains of the features ($x_i, x_j, \ldots, x_m$), the group $x_i^{\overline{\alpha_1}} x_j^{\overline{\alpha_2}} \cdots x_m^{\overline{\alpha_p}}$ represents an intrinsic variable and $f(\mathbf{x})$ is invariant with respect to that group.

2.2.2. Multiple Concurrent Invariant Groups in Power Form

Clearly, the above procedure can be extended to a function $f(\mathbf{x})$ being invariant with respect to a higher number of feature groups, with each group even sharing some of the primitive physical variables. Without losing generality, and for the sake of simplicity, we limit this generalized description to sets consisting of two concurrent groups.

Let $f(\mathbf{x}) = f(x_1, \ldots, x_n) = f(x_1, \ldots, x_i^{\alpha_1} x_j^{\alpha_2} \cdots x_m^{\alpha_p}, x_j^{\beta_1} x_k^{\beta_2} \cdots x_r^{\beta_q}, \ldots, x_n)$ denote a function where two groups share a generic primitive variable x_j. In this case, the above procedure applied to individual groups $x_i^{\alpha_1} x_j^{\alpha_2} \cdots x_m^{\alpha_p}$ and $x_j^{\beta_1} x_k^{\beta_2} \cdots x_r^{\beta_q}$ independently is not suitable any longer and requires a generalization as discussed below.

To investigate the invariance with respect to both groups, we now construct the $(2 \times l)$ matrix $\mathbf{B}$:

$$\mathbf{B} = \begin{bmatrix} \frac{\alpha_1}{x_{i,0}} & \frac{\alpha_2}{x_{j,0}} & 0 & \cdots & \frac{\alpha_p}{x_{m,0}} & 0 \\ 0 & \frac{\beta_1}{x_{j,0}} & \frac{\beta_2}{x_{k,0}} & \cdots & 0 & \frac{\beta_q}{x_{r,0}} \end{bmatrix}, \tag{7}$$

with $\max(p, q) \leq l \leq p + q$, $\mathbf{K}$ being a matrix whose columns represent an ortho-normal basis of the null space of $\mathbf{B}$.

As performed above, the condition of invariance requires that the normalized gradient of $f(\mathbf{x})$ is orthogonal to each column of the kernel matrix $\mathbf{K}$ (see also Equation (4)).

Adding the normalization condition of the coefficients $\alpha_1, \alpha_2, \ldots, \alpha_p$ and $\beta_1, \beta_2, \ldots, \beta_q$, we obtain:

$$\begin{cases} (\nabla \tilde{f})_{\overline{\mathbf{x}_0}} \cdot \mathbf{K} = 0 \\ \alpha_1^2 + \alpha_2^2 + \cdots + \alpha_p^2 = 1 \\ \beta_1^2 + \beta_2^2 + \cdots + \beta_q^2 = 1 \end{cases} \tag{8}$$

with the partial derivatives being evaluated in $\overline{\mathbf{x}_0}$. If the non-linear system in Equation (8) is satisfied for the same exponents ($\overline{\alpha_1}, \ldots, \overline{\alpha_p}, \overline{\beta_1} \ldots, \overline{\beta_q}$) in the entire feature domain ($x_i, x_j, \ldots, x_r$), the two groups $x_i^{\overline{\alpha_1}} x_j^{\overline{\alpha_2}} \cdots x_m^{\overline{\alpha_p}}$ and $x_j^{\overline{\beta_1}} x_k^{\overline{\beta_2}} \cdots x_r^{\overline{\beta_q}}$ are intrinsic variables as the function $f(\mathbf{x})$ is invariant with respect to both groups.

It is worth stressing that more general cases with three or more concurrent invariant groups will imply additional rows for the matrix $\mathbf{B}$. Furthermore, in general, the non-linear system (8) is not necessarily closed (see Section 2.2.4).

2.2.3. Further Generalization to Non Power Forms

The invariant group/set identification procedure, introduced for groups in the power form, can be generalized to other functional relationships. For the sake of illustration, in the following, we restrict to a single invariant group.

Let $f(\mathbf{x}) = f(x_1, \ldots, x_n)$ denote a function invariant with respect to a group involving p variables according to a generic functional dependence $g(x_i, x_j, \ldots, x_m)$, such that $f(\mathbf{x}) = f(x_1, \ldots, g(x_i, x_j, \ldots, x_m) \ldots, x_n)$. When such group is a constant $\overline{c}$, even varying its components $x_i, x_j, \ldots, x_m$ separately, $f(\mathbf{x})$ does not change. This yields:

$$g(\underbrace{x_i, x_j, \ldots, x_m}_{p}) = \overline{c} \tag{9}$$

which translates into:

$$\frac{\partial g}{\partial x_i}\mathrm{d}x_i + \frac{\partial g}{\partial x_j}\mathrm{d}x_j + \cdots + \frac{\partial g}{\partial x_m}\mathrm{d}x_m = 0. \tag{10}$$

We can thus construct the $(1 \times p)$ matrix $\mathbf{B}$ at a generic point $\overline{\mathbf{x}_0} = (x_{i,0}, x_{j,0}, \ldots, x_{m,0})$ as

$$\mathbf{B} = \left[\frac{\partial g}{\partial x_i}, \frac{\partial g}{\partial x_j}, \ldots, \frac{\partial g}{\partial x_m}\right]. \tag{11}$$

Let $\mathbf{K}$ be a matrix whose columns represent an orthonormal basis of the null space of $\mathbf{B}$. Applying the invariance condition of Equation (4) leads to a non-linear system analogous to Equation (6). Notably, the procedure illustrated here can be further extended to sets of groups in any functional form, following a reasoning similar to that applied in Section 2.2.2.

2.2.4. Regression Model and Procedure Implementation

In this study, we assume that a database of physical data is available. Each sample in our database consists of n features $(x_1, \ldots, x_n)$ corresponding to a target quantity $f(\mathbf{x})$. We further assume that the target quantity possibly depends on some invariant groups expressed for instance in the power form. We thus try to detect such invariance following the above procedure.

As already described, this requires the evaluation of the gradient $\nabla f(\mathbf{x})$. To this end, $f(\mathbf{x})$ can be conveniently approximated with a Deep Neural Network (DNN), allowing the computation of that gradient by means of automatic differentiation. The choice of a DNN (instead of a simpler approximator) is also beneficial to the method's versatility, making it suitable for the generalized cases of non power forms. All of the DNNs of this study are trained and validated over the 85% of the corresponding databases—of which the 85% is used for the training and the remaining 15% for the validation—and tested over the remaining 15%. The DNN structure, as reported in Supplementary Note S2, is used for all the examples of this study. Upon network training and validation, automatic differentiation is utilized to calculate the gradient of the function at a specific point $\mathbf{x}_0$ within its domain. Here, the variables in the examined group are randomly chosen within their domains, while the values of the remaining variables are held constant at their averages in the original database.

The function's gradient is computed and applied in Equations (6) and (8), depending on the specific case. Subsequently, the system is solved using a least-squares optimization method that utilizes the Trust Region Reflective [27] algorithm, which is suitable also for under-determined (non closed) systems that may arise during the process of identifying sets of groups.

This strategy is repeated with 20 different values of $\mathbf{x}_0$ and iterated multiple times, each time updating the initial guess to the mean value of the exponents $(\overline{\alpha_1}, \overline{\alpha_2}, \ldots, \overline{\alpha_p})$ averaged over the 20 solutions found in the previous iteration.

An overview of the methodology for identifying invariant groups/sets is depicted in Figure 2.

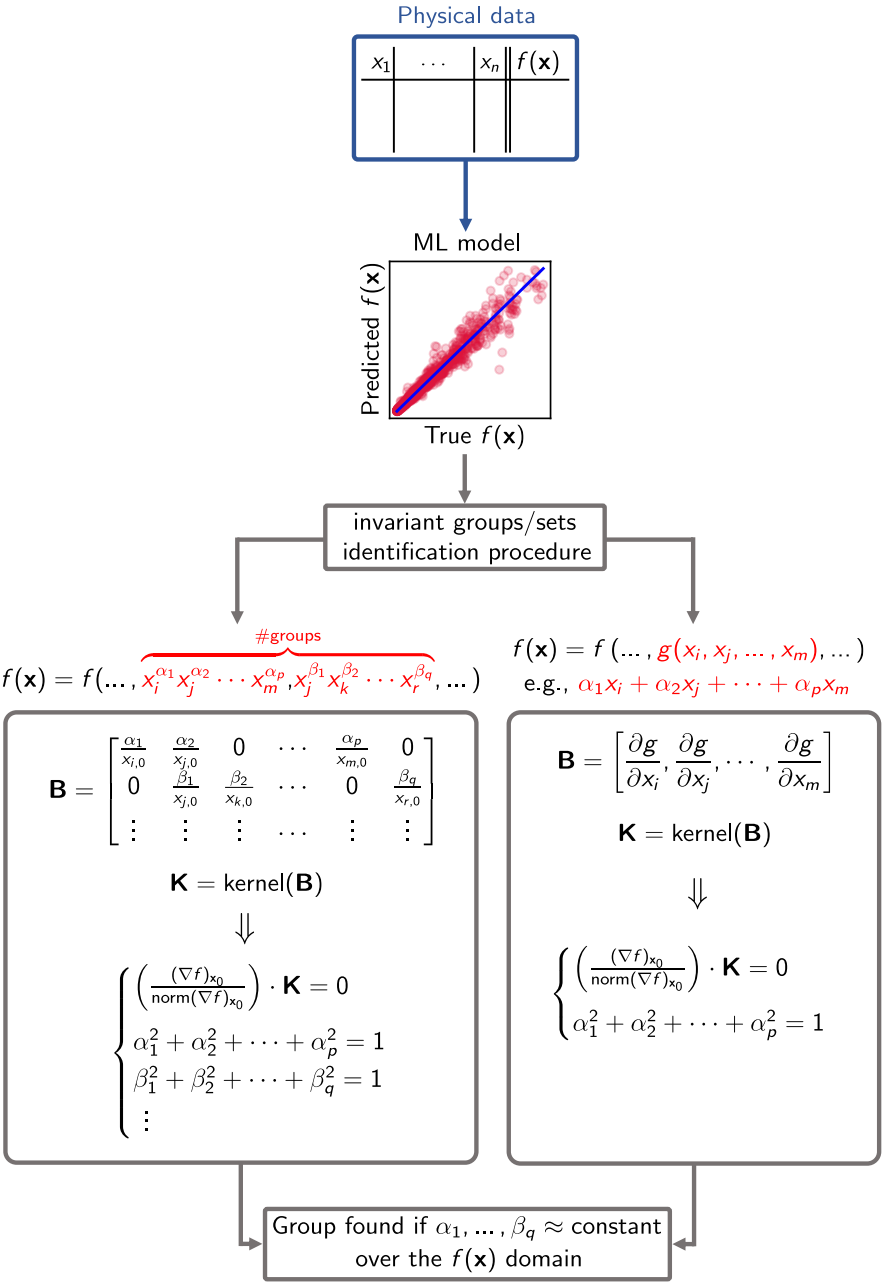

Figure 2. Overview of the procedure for identifying invariant groups/sets. A regression model is trained on the physical data and used to compute the gradient of the objective function in a point $\mathbf{x}_0$. The matrix $\mathbf{B}$ is constructed according to the functional structure of the investigated group/set, and its kernel $\mathbf{K}$ is computed. Finally, the condition of invariance between gradient and kernel is coupled to the normalization conditions of the coefficients. If the resulting non-linear system is satisfied for the same coefficients over the $f(\mathbf{x})$ domain, the group/set is an intrinsic variable and $f(\mathbf{x})$ is invariant with respect to it.

2.3. Searching for Good Variables by Classification Models

The methodology introduced above, based on regression models, is able to identify groups and/or group sets of physical variables upon which a generic function $f(\mathbf{x})$ depends. If successful, this approach efficiently reduces the number of such primitive variables by optimizing their combination.

However, a similar goal can be achieved by means of an entirely different approach based on classification models, and it has been initially proposed by Trezza and Chiavazzo in [23]. The latter methodology, briefly reviewed in the following, enables the optimal mixing of primitive features by performing classification and subsequent optimal class separation of the available data samples.

Let $x_1, \ldots, x_n$ denote n features. Let $(\tilde{x}_1, \ldots, \tilde{x}_n)$ be the corresponding dimensionless quantities

$$\tilde{x}_i = \frac{x_i - x_{i,\min}}{x_{i,\max} - x_{i,\min}} + 1 \tag{12}$$

normalized by construction within the range $[1, 2]$ to avoid singularities in the procedure below, where $x_{i,\min}$ and $x_{i,\max}$ represent the minimum and the maximum observed values for the ith feature over the training set, respectively. It is possible to synthetically create a set of $m \ll n$ mixed features $(y_1, \ldots, y_m)$ as

$$y_j = \prod_{i=1}^{n} \tilde{x}_i^{\alpha_{ij}} \tag{13}$$

with $\{\alpha_{ij}\}$ being an $(n \times m)$ matrix estimated by means of a multi-objective optimization algorithm, as described below. Finally, the new variables y_j can be normalized within the interval $[0, 1]$ according to

$$\tilde{y}_j = \frac{y_j - y_{j,\min}}{y_{j,\max} - y_{j,\min}}. \tag{14}$$

The main idea is that the matrix α_{ij} shall be selected following an optimization criterion attempting the largest possible separation between two (or more) different classes partitioning the values of a physical property of interest. Clearly, this can be accomplished by maximizing a certain distance between the classes. However, a multi-objective optimization procedure could also be pursued.

For instance, in a binary classification, the matrix α_{ij} in Equation (13) can be chosen to lie on the Pareto front while simultaneously pursuing: (i) maximization of a carefully selected distance between the two classes; (ii) minimization of a norm of the covariance matrix of the first class distribution; and (iii) minimization of a norm of the covariance matrix of the second class distribution.

The main rationale behind the minimization of a norm of the covariance matrix for class distribution (along with a distance between classes) is the aim of possibly obtaining smooth distribution functions that can be analytically bet fitted. Following the latter idea, below, we will pursue multi-objective optimization for all the examined cases and provide some optimal solutions from Pareto fronts.

In this study, we adopt genetic algorithms for optimization whereas the Bhattacharyya distance between the histograms of the two equally binned classes [28,29] is evaluated to be maximized during the multi-objective optimization. However, as shown in ref. [23], other options for class distance may be considered, such as the Wasserstein distance [30] or the averaged number within a fixed radius of nearest neighbors of one class to each samples of the other class. Furthermore, herein we utilize variable power-form mixing outlined in Equation (13). Nonetheless, alternative grouping options (e.g., linear mixing as employed in ref. [23]) are also possible with the overall methodology remaining unchanged. Also, the procedure can be further generalized to a number of classes > 2.

In this case, the genetic algorithm aims at simultaneously maximizing the pairwise distances between the classes [23]. For the remaining two objectives, in one-dimensional cases, a numerical estimate of the standard deviation of the binned data in the two classes is computed. In two-dimensional (or higher) cases, the determinant of the covariance matrix can be utilized. From the practical standpoint, considering a dataset of physical data samples characterized by features $x_1, \ldots, x_n$ and their corresponding response $f(\mathbf{x})$, such a dataset is partitioned into two classes based on a carefully chosen threshold for $f(\mathbf{x})$. After the aforementioned pre-processing steps, a genetic algorithm optimization is employed to

identify the Pareto front. This optimization concurrently seeks to minimize the variances of the two classes (in one dimension) or the determinants of the covariances of the two classes (in two or more dimensions), while also maximizing the Bhattacharyya distance between the classes. A summary of this procedure is illustrated in Figure 3.

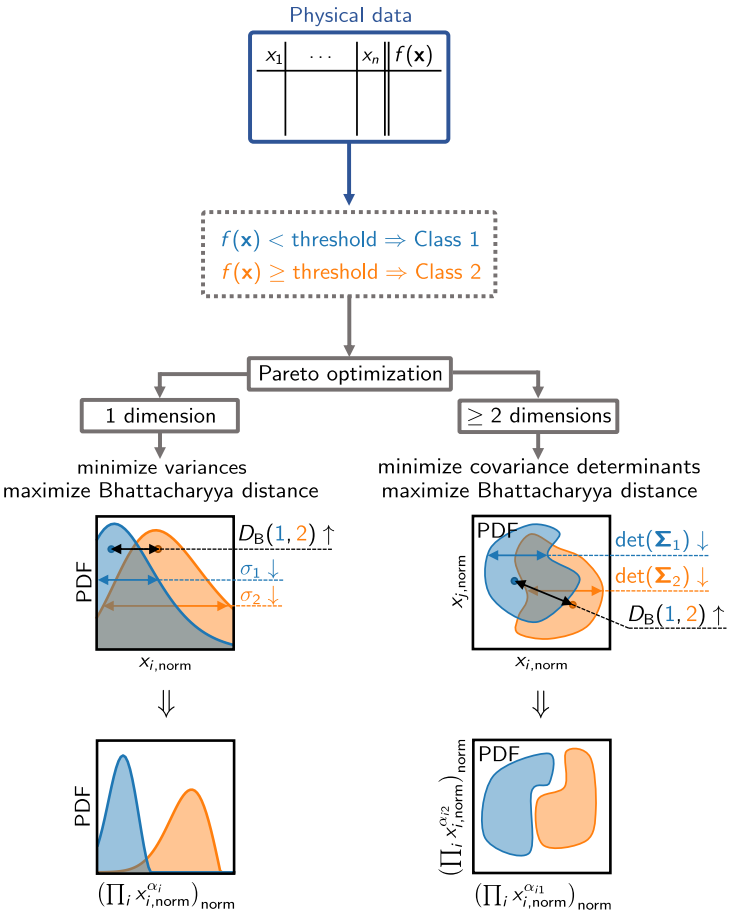

Figure 3. Overview of the procedure to identify optimal mixed variables for class separation. Threshold values are chosen to divide the physical data in classes. A Pareto optimization is performed to construct a reduced set of synthetic features that simultaneously maximizes the Bhattacharyya distance between the classes and (i) minimizes the variances of the class distributions, in the one dimensional case, or (ii) minimizes the determinants of the covariance matrix of the class distributions, in the multi dimensional case.

3. Numerical Examples and Discussion

3.1. Datasets Creation

We first create three datasets from physical models to be used to train and test Machine Learning (ML) tools, aiming to predict physical properties of interest. More specifically, the first two datasets are generated using popular thermo-fluid dynamic correlations (i.e., Dittus–Boelter and Gnielinski correlations) with the target quantity being the Nusselt number. Those datasets are created starting from the physical properties of 16 real liquids at ambient temperature and pressure [31], and evaluating the kinematic viscosity ν and the thermal diffusivity κ of each liquid from tabulated data (see Supplementary Note S1). Each fluid accounts for 500 value combinations of the flow speed u, the hydraulic diameter D, and the friction factor f (the latter is only requested for the Gnielinski correlation). The values of these three variables are randomly chosen in the ranges $[0.1, 1]$, $[0.01, 0.1]$, and $[0.02, 0.09]$, respectively. This leads to a total of 8000 samples for each dataset. The Nusselt number is thus calculated according to the corresponding equations. Finally,

emulating what is typically experienced in experimental measurements, noise is added on top of the correct target values, sampling 8000 points η_i from a Gaussian distribution, with mean $\mu = 0$ and standard deviation $\sigma = 0.1$; the target value of each entry Nu_i is then multiplied by $(1 + \eta_i)$. The third dataset is created on the basis of Newton's law of universal gravitation. It is constructed in a similar fashion, by generating random values within the range $[10^{16}, 10^{18}]$ for the masses m_1 and m_2, and within the ranges $[-10^{12}, -10^{10}]$ and $[10^{10}, 10^{12}]$ for the spatial coordinates $x_1, y_1, z_1, x_2, y_2, z_2$. Here, the target value is the gravitational force F_g, which is computed with some noise added following the same procedure as above.

3.2. Dittus–Boelter Equation

The Dittus–Boelter correlation for a fluid undergoing heating is expressed by the following equation:

$$\mathrm{Nu} = 0.023 \mathrm{Re}_D^{4/5} \mathrm{Pr}^{0.4} \tag{15}$$

where $\mathrm{Re}_D = \frac{ud}{\nu}$ represents the Reynolds number and $\mathrm{Pr} = \frac{\nu}{\kappa}$ denotes the Prandtl number. The equation can be rewritten by substituting the dimensionless quantities Re and Pr, thus obtaining

$$\mathrm{Nu} = 0.023 \left(\frac{ud}{\nu} \right)^{4/5} \left(\frac{\nu}{\kappa} \right)^{0.4} = 0.023 \frac{u^{0.8} d^{0.8}}{\nu^{0.4} \kappa^{0.4}}. \tag{16}$$

It is easy to verify that the objective value is invariant with respect to all the possible combinations of input features, u, d, ν, κ, either couples or triplets.

3.2.1. Use of Regression Models

We attempt to recover the symmetries of the Nusselt number with respect to binary and ternary groups in the form $x_i^{\alpha_1} x_j^{\alpha_2}$ and $x_i^{\alpha_1} x_j^{\alpha_2} x_k^{\alpha_3}$, only relying on noised data and by adopting the methodology illustrated in Section 2.2. As outlined above, this requires the evaluation of the gradient of the noised Nusselt number with respect to the input features, namely $\nabla \overline{\mathrm{Nu}}(u, d, \nu, \kappa)$, where for each sample i, $\overline{\mathrm{Nu}_i} = \mathrm{Nu}_i(1 + \eta_i)$ (see Methods for details). This can be conveniently computed by automatic differentiation over a DNN model approximating $\overline{\mathrm{Nu}}(u, d, \nu, \kappa)$. As input features of the DNN, the four variables u, d, ν, κ are used. The dataset is thus divided into three parts: (i) a training set, (ii) a validation set used to detect potential overfitting, and (iii) a testing set for the comprehensive evaluation of model performance. Figure 4a showcases the model predictions over the testing set, while in Figure 4b, the corresponding loss across epochs is depicted.

Notably, the model is highly predictive, with coefficient of determination $R^2 = 0.963$ over the testing set, with no evidence of overfitting observed. The model is thus fed to the optimization algorithm. For each of the ten expected invariant groups (six binary and four ternary), the algorithm estimates the normalized exponents $\alpha_1, \alpha_2, \alpha_3$ 20 times. Table 1 shows such true normalized exponents $\alpha_1, \alpha_2, \alpha_3$, together with the means $\mu_{\alpha_1}, \mu_{\alpha_2}, \mu_{\alpha_3}$ and the standard deviations $\sigma_{\alpha_1}, \sigma_{\alpha_2}, \sigma_{\alpha_3}$ over those 20 evaluations. Our findings indicate that the method correctly identifies invariant groups in both couple and triplet forms. Furthermore, it demonstrates the ability to estimate normalized exponents $\alpha_1, \alpha_2, \alpha_3$ for individual primitive variables within each group, with relative percent errors not exceeding 11.0% for couples and 16.6% for triplets. Clearly, we notice that normalized exponents are obtained up to the sign: the exponents determined for the pair (u, d) are negative instead of the expected positive values according to the Dittus–Boelter equation.

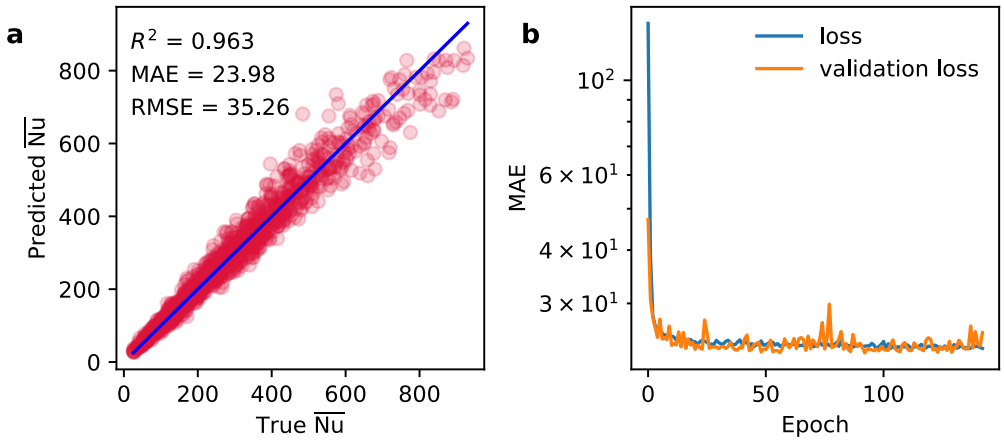

Figure 4. Results of the DNN regression model for the noised Nusselt number $\overline{\text{Nu}}$ in the Dittus–Boelter correlation. (**a**) Predictions over the testing set, and (**b**) corresponding loss curves for the DNN model. Model performances are shown in terms of coefficient of determination R^2, mean absolute error (MAE), and root mean squared error (RMSE).

In Table 1, we label as found, all the groups with standard deviations $\sigma_{\alpha_1}, \sigma_{\alpha_2}, \sigma_{\alpha_3} \leq 0.2$, since low values of σ imply that the results for the 20 evaluations are consistent. We also label as reliable all the groups where none of the evaluated exponents approach 0 or 1, since such cases would correspond to trivial solutions. Remarkably, all six binary groups and all four ternary groups of input features are found to comply with the condition of invariance, and the results for all of them are reliable.

Table 1. Normalized exponents $\alpha_1, \alpha_2, \alpha_3$ for the Dittus–Boelter correlation together with their average estimates over 20 evaluations $\mu_{\alpha_1}, \mu_{\alpha_2}, \mu_{\alpha_3}$ and the corresponding standard deviations $\sigma_{\alpha_1}, \sigma_{\alpha_2}, \sigma_{\alpha_3}$. Found groups refer to low standard deviation, reliable groups refer to average far from 1 and 0.

Group	α_1	α_2	α_3	μ_{α_1}	σ_{α_1}	μ_{α_2}	σ_{α_2}	μ_{α_3}	σ_{α_3}	Found	Reliable
(u,d)	0.71	0.71	-	−0.71	0.05	−0.70	0.05	-	-	yes	yes
(u,ν)	0.89	−0.45	-	0.91	0.04	−0.42	0.07	-	-	yes	yes
(u,κ)	0.89	−0.45	-	0.89	0.01	−0.45	0.03	-	-	yes	yes
(d,ν)	0.89	−0.45	-	0.90	0.04	−0.42	0.11	-	-	yes	yes
(d,κ)	0.89	−0.45	-	0.87	0.02	−0.50	0.04	-	-	yes	yes
(ν,κ)	−0.71	−0.71	-	−0.67	0.10	−0.73	0.10	-	-	yes	yes
(u,d,ν)	0.67	0.67	−0.33	0.68	0.03	0.65	0.04	−0.33	0.07	yes	yes
(u,d,κ)	0.67	0.67	−0.33	0.66	0.03	0.64	0.03	−0.37	0.03	yes	yes
(u,ν,κ)	0.82	−0.41	−0.41	0.87	0.04	−0.37	0.06	−0.34	0.10	yes	yes
(d,ν,κ)	0.82	−0.41	−0.41	0.83	0.06	−0.35	0.11	−0.41	0.05	yes	yes

3.2.2. Use of Classification Models

In a second attempt to reduce the number of input variables, we investigate the possible existence of symmetries for classification, aiming at the construction of m mixed features in the form $y_j = \prod_{i=1}^{n} \tilde{x}_i^{\alpha_{ij}}$, where $(\tilde{x}_1, \ldots, \tilde{x}_n)$ are the primitive variables properly normalized within the interval $[1, 2]$ and $\{\alpha_{ij}\}$ denotes an $n \times m$ matrix optimally estimated, as detailed in Section 2.3. Such features are finally properly normalized as $\tilde{y}_j$ in the interval $[0, 1]$. In the case of the Dittus-Bolter, we set class 1 for samples with $\overline{\text{Nu}} < 395$ and class 2 for samples with $\overline{\text{Nu}} \geq 395$. We thus create a single mixed feature ($m = 1$) according to the above procedure. Figure 5a reports the Probability Density Function (PDF) binning of the training set data over the two classes (a higher value of the PDF means a higher number of items in the corresponding bin), against the normalized flow velocity u_{norm}, while Figure 5b shows the same PDFs (i.e., class distributions) against the normalized mixed feature $\tilde{y}_1$.

It is noteworthy to observe that when represented against the primitive feature, the two classes exhibit considerable overlap, whereas the same two classes appear well separated when plotted against the mixed feature. As a result, it is particularly convenient to attempt an analytical bet-fitting of the two distributions depicted in Figure 5b approximated by a Generalized Extreme Value (GEV) distribution, with density

$$g(\tilde{y}_1) = \frac{1}{\sigma}\left(1+\omega\frac{\tilde{y}_1-\mu}{\sigma}\right)^{-\frac{\omega+1}{\omega}} \exp\left(-\left(1+\omega\frac{\tilde{y}_1-\mu}{\sigma}\right)^{-1/\omega}\right). \tag{17}$$

The fitting is performed by means of the SciPy Python package [32]. The specific computed GEV distribution for samples with $\overline{\mathrm{Nu}} < 395$ has factors $\mu = 0.338$, $\sigma = 0.137$, and $\omega = 0.342$, whereas the GEV distribution for samples with $\overline{\mathrm{Nu}} \geq 395$ has factors $\mu = 0.675$, $\sigma = 0.077$, and $\omega = 0.074$. Figure 5c shows the PDFs over the binned data of the testing set reported against the same mixed feature $\tilde{y}_1$, along with the GEV fittings computed on the training set. Notably, the classes are still well separated, with a good agreement between the GEV distributions and the testing set densities.

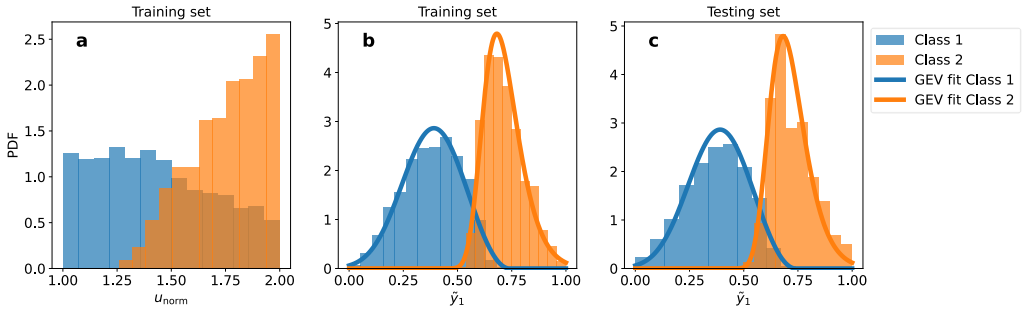

Figure 5. One dimensional example for classification on the Dittus–Boelter correlation. (**a**) PDFs over binned data of the training set for the two classes ($\overline{\mathrm{Nu}} < 395$ and $\overline{\mathrm{Nu}} \geq 395$) reported against the normalized flow velocity. (**b**) PDFs over binned data of the training set for the two classes reported against the mixed feature y_1, constructed according to Equation (13) and choosing the point of the Pareto front with the least overlapping of the two classes according to the Bhattacharyya distance, along with a GEV analytical fitting of the two binnings. (**c**) PDFs over binned data of the testing set for the two classes reported against the same mixed feature y_1 together with the same GEV fittings of the (**b**) subfigure. The mixed variable $\tilde{y}_1$ shown here is referred exclusively to this Dittus–Boelter one dimensional optimization.

Furthermore, we make an attempt at the creation of two mixed normalized features $\tilde{y}_1, \tilde{y}_2$ ($m = 2$) with the same classes. Specifically, Figure 6a shows the PDF two dimensional binning of the training set data over the two classes, against the normalized flow velocity u_{norm} and the hydraulic diameter d_{norm}. Figure 6b shows the same PDFs against the normalized mixed features $\tilde{y}_1, \tilde{y}_2$ constructed according to Equation (13) by power combination of the four primitive variables and choosing the point of the Pareto front with the least overlapping of the two distributions. Similarly to the one dimensional case, the classes appear well separated when plotted against the mixed features, whereas there is a wide overlap when plotted against the primitive variables. Figure 6c shows the PDFs over the binned data of the testing set reported against the same mixed features. The two classes still appear well separated.

Finally, we aim at the construction of $m = 1$ mixed feature for separating the data samples into three classes: class 0 for $\overline{\mathrm{Nu}} < 197.5$, class 1 for $197.5 \leq \overline{\mathrm{Nu}} < 395$, class 2 for $\overline{\mathrm{Nu}} \geq 395$. The multi-objective optimization is performed aiming at concurrently maximizing the pairwise distances between the classes. Figure 7 shows the separation in classes reported against the normalized flow velocity u_{norm} and the new mixed feature $\tilde{y}_1$: in line with previous cases, the substantial overlap observed when representing classes against the primitive feature is significantly reduced when using the optimized mixed

feature. The bet-fitting GEV distributions for the three classes are computed according to Equation (17) over the training set samples. Specifically the GEV associated with Class 0 has factors $\mu = 0.284$, $\sigma = 0.103$, $\omega = 0.494$, the GEV for Class 1 has factors $\mu = 0.475$, $\sigma = 0.067$, $\omega = 0.193$, and the GEV for Class 2 has factors $\mu = 0.684$, $\sigma = 0.080$, $\omega = 0.103$. When plotting the PDFs over the binned data of the testing set against the mixed feature, the classes appear well separated, and there is good agreement between the GEV distributions and the testing set densities.

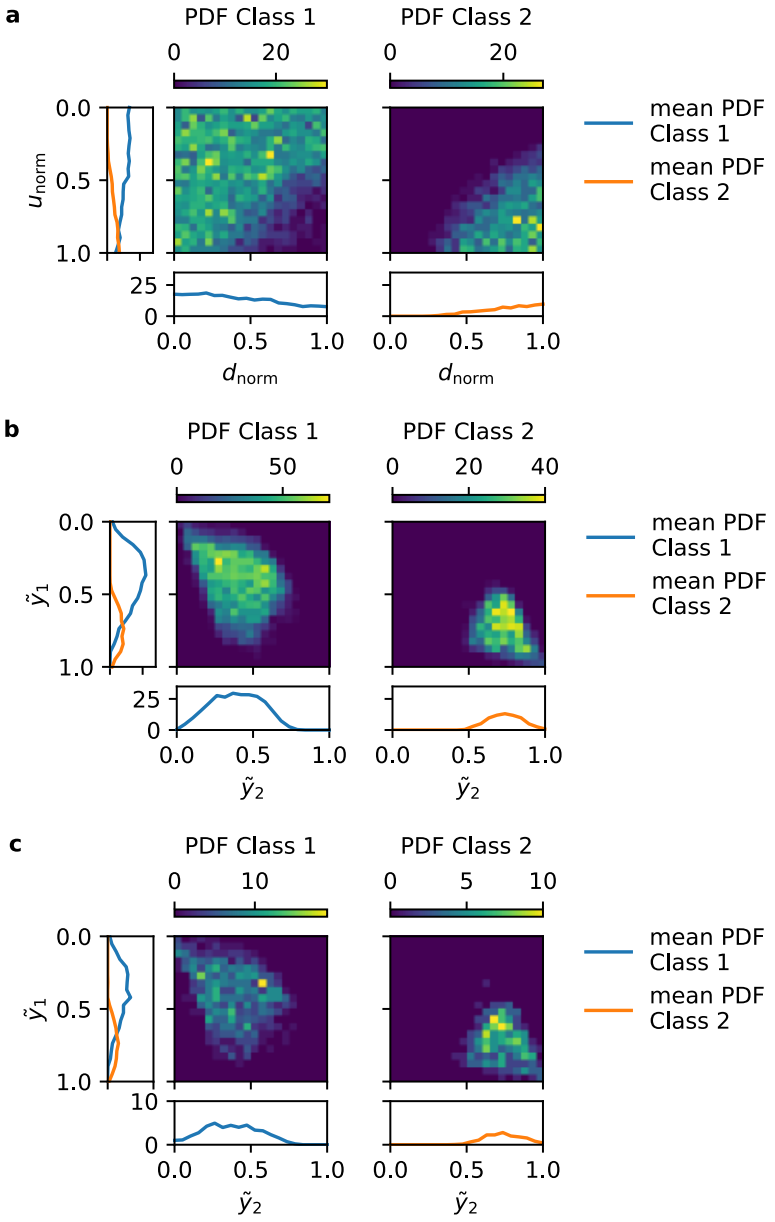

Figure 6. Two dimensional example for the classification on the Dittus–Boelter correlation. (**a**) PDFs over binned data of the training set for the two classes ($\overline{\text{Nu}} < 395$ and $\overline{\text{Nu}} \geq 395$) reported against the normalized flow velocity u_{norm} and the normalized hydraulic diameter d_{norm}. (**b**) PDFs over binned data of the training set for the two classes reported against the mixed features $\tilde{y}_1, \tilde{y}_2$, constructed according to Equation (13) and choosing the point of the Pareto front with the least overlapping of the two classes according to the Bhattacharyya distance. (**c**) PDFs over binned data of the testing set for the two classes reported against the same mixed features $\tilde{y}_1, \tilde{y}_2$. The mixed variables $\tilde{y}_1, \tilde{y}_2$ shown here are referred exclusively to this Dittus–Boelter two dimensional optimization.

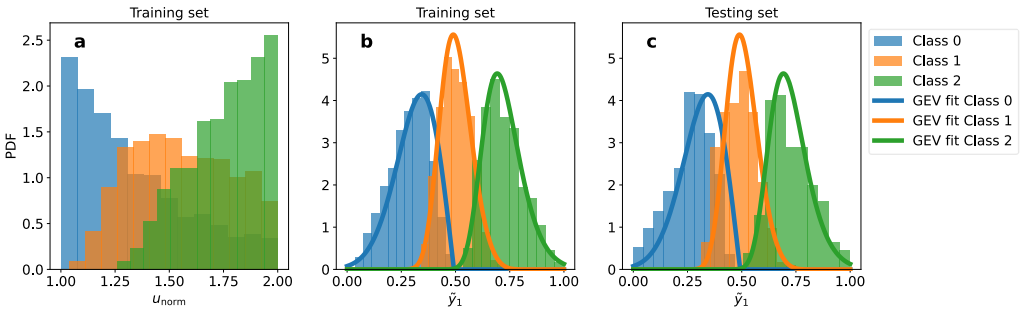

Figure 7. One dimensional example for the ternary classification on the Dittus–Boelter. (**a**) PDFs over binned data of the training set for the three classes ($\overline{\mathrm{Nu}} < 197.5$, $197.5 \leq \overline{\mathrm{Nu}} < 395$, and $\overline{\mathrm{Nu}} \geq 395$) reported against the normalized flow velocity u_{norm}. (**b**) PDFs over binned data of the training set for the three classes reported against the mixed feature $\tilde{y}_1$, constructed according to Equation (13) and choosing the point of the Pareto front with the least overlapping of the three classes according to the Bhattacharyya distance, along with a GEV analytical fitting of the three binnings. (**c**) PDFs over binned data of the testing set for the three classes reported against the same mixed feature $\tilde{y}_1$ together with the same GEV fittings of the (**b**) subfigure. The mixed variable $\tilde{y}_1$ shown here is referred exclusively to this Dittus–Boelter one-dimensional optimization.

3.3. Gnielinski Correlation

The Gnielinski correlation for turbulent flow in tubes is expressed by the equation

$$\mathrm{Nu} = \frac{(f/8)(\mathrm{Re}_D - 1000)\mathrm{Pr}}{1 + 12.7(f/8)^{1/2}(\mathrm{Pr}^{2/3} - 1)} \tag{18}$$

where $\mathrm{Re}_D = \frac{ud}{\nu}$ represents the Reynolds number, $\mathrm{Pr} = \frac{\nu}{\kappa}$ is the Prandtl number and f denotes the friction factor. The equation can be reformulated by substituting the dimensionless quantities Re and Pr, resulting in:

$$\mathrm{Nu} = \frac{(f/8)\left(\frac{ud}{\nu} - 1000\right)\left(\frac{\nu}{\kappa}\right)}{1 + 12.7(f/8)^{1/2}\left(\left(\frac{\nu}{\kappa}\right)^{2/3} - 1\right)}. \tag{19}$$

The Nusselt number is invariant only with respect to the combination of flow velocity and hydraulic diameter (u, d) with real exponents $(1, 1)$, normalized exponents $(0.707, 0.707)$.

3.3.1. Use of Regression Models

Here, we first attempt the detection of possible symmetries of the noised Nusselt number $\overline{\mathrm{Nu}}$ adopting the methodology proposed in Section 2.2.2. We obtain access to the gradient $\nabla\overline{\mathrm{Nu}}(u, d, \nu, \kappa, f)$ by means of automatic differentiation over a DNN approximating $\overline{\mathrm{Nu}}(u, d, \nu, \kappa, f)$. Figure 8a shows the model predictions over the testing set, while in Figure 8b the corresponding loss across epochs is depicted. The model is highly predictive, with coefficient of determination $R^2 = 0.978$ over the testing set, and no evidence of overfitting is observed. The trained model is thus fed to the optimization algorithm, looking for binary groups in the form $x_i^{\alpha_1} x_j^{\alpha_2}$. The average $\mu_{\alpha_1}, \mu_{\alpha_2}$ estimates of the exponents α_1, α_2, together with their standard deviations $\sigma_{\alpha_1}, \sigma_{\alpha_2}$, are reported in Table 2.

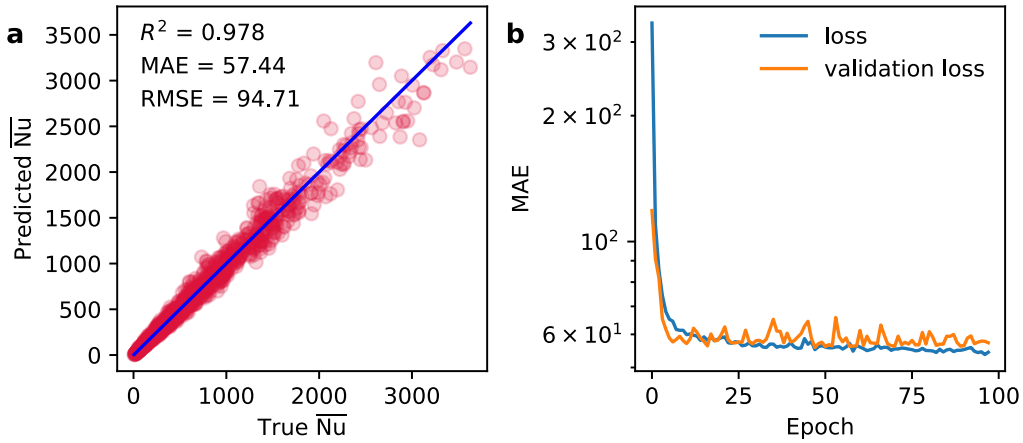

Figure 8. Results of the DNN regression model for the noised Nusselt number $\overline{\text{Nu}}$ in the Gnielinski correlation. (**a**) Predictions over the testing set, and (**b**) corresponding loss curves for the DNN model. Model performances are shown in terms of coefficient of determination R^2, mean absolute error (MAE), and root mean squared error (RMSE).

Specifically, the procedure correctly identifies the expected group (u, d), whereas for most other groups, the results correspond to either the trivial solution (i.e., the absolute value of one of the exponents is close to 1 and the other is close to 0), or at least one of the evaluated values presents too high variance (i.e., $\sigma_{\alpha_i} > 0.2$). Remarkably, for the remaining cases (namely (u, ν), (u, f), (d, ν), (d, f), (ν, κ), (ν, f)), the procedure identifies groups complying with the invariance condition, albeit not explicitly expected in the Gnielinski correlation. We thus concluded that the suggested procedure has detected a *local* invariance and this can be numerically verified as follows (here we limit to groups of two variables only). First, the features not appearing in the group are kept constant. Second, a random sample i in the dataset is considered for the evaluation of the quantity $\tilde{c} = x_{1,i}^{\alpha_1} x_{2,i}^{\alpha_2}$. Third, a vector $\overline{\mathbf{x}_1}$ of evenly spaced points in the domain of the first feature is created. As such, the array corresponding to the second feature is obtained as $\overline{\mathbf{x}_2} = (\tilde{c}\overline{\mathbf{x}_1}^{-\alpha_1})^{1/\alpha_2}$. A new dataset is thus constructed, with the variables out of the group being constant and with the variables in the group being replaced by $\overline{\mathbf{x}_1}$, $\overline{\mathbf{x}_2}$. For all those samples, the response value $f(\mathbf{x})$ is computed. The local invariance is demonstrated when $f(\mathbf{x})$ is approximately constant over all the new constructed dataset. Further details can be found in Supplementary Note S3.

Furthermore, armed with the same DNN regression model, we try to detect invariance of the Nusselt number with respect to sets groups in the form $x_i^{\alpha_1} x_j^{\alpha_2} x_k^{\alpha_3}$ and $x_k^{\beta_1} x_l^{\beta_2}$ employing the procedure presented in Section 2.2.2. Table 2 also reports the results for selected sets of two feature couples, including sets appearing explicitly in the Gnielinski correlation—i.e., $[(u, \nu), (\nu, \kappa)]$ and $[(d, \nu), (\nu, \kappa)]$—and for selected sets of features comprising a triplet and a couple—e.g., $[(u, d, \nu), (\nu, \kappa)]$, which is extremely relevant as it comprises the Reynolds and the Prandtl numbers. For all of the mentioned sets, the procedure correctly identifies the exponents that make the groups compliant with the condition of invariance. Indeed, the normalized exponents obtained with the procedure for the sets $[(u, \nu), (\nu, \kappa)]$ and $[(d, \nu), (\nu, \kappa)]$ are $[(0.730, -0.680), (-0.656, 0.755)]$ and $[(0.695, -0.718), (-0.687, 0.727)]$, respectively, whereas the analytical solution is $[(0.707, -0.707), (0.707, -0.707)]$ for both the cases. The normalized exponents obtained for the set $[(u, d, \nu), (\nu, \kappa)]$ turn out to be $[(0.602, 0.552, -0.570), (0.698, -0.716)]$, whereas the analytical solution corresponds to $[(0.577, 0.577, -0.577), (0.707, -0.707)]$.

Table 2. Normalized exponents $\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2$ for the Gnielinski correlation, together with their average estimates over 20 evaluations $\mu_{\alpha_1}, \mu_{\alpha_2}, \mu_{\alpha_3}, \mu_{\beta_1}, \mu_{\beta_2}$ and the corresponding standard deviations $\sigma_{\alpha_1}, \sigma_{\alpha_2}, \sigma_{\alpha_3}, \sigma_{\beta_1}, \sigma_{\beta_2}$. Found groups/sets refer to low standard deviation, reliable groups refer to average far from 1 and 0.

Group/Set	α_1	α_2	α_3	β_1	β_2	μ_{α_1}	σ_{α_1}	μ_{α_2}	σ_{α_2}	μ_{α_3}	σ_{α_3}	μ_{β_1}	σ_{β_1}	μ_{β_2}	σ_{β_2}	Found	Reliable
(u,d)	0.707	0.707	-	-	-	−0.673	0.067	−0.734	0.056	-	-	-	-	-	-	yes	yes
(u,ν)	-	-	-	-	-	0.801	0.022	−0.597	0.030	-	-	-	-	-	-	yes	yes
(u,κ)	-	-	-	-	-	0.949	0.010	−0.313	0.032	-	-	-	-	-	-	yes	no
(u,f)	-	-	-	-	-	−0.867	0.065	−0.468	0.161	-	-	-	-	-	-	yes	yes
(d,ν)	-	-	-	-	-	0.807	0.042	−0.587	0.055	-	-	-	-	-	-	yes	yes
(d,κ)	-	-	-	-	-	0.948	0.015	−0.314	0.049	-	-	-	-	-	-	yes	no
(d,f)	-	-	-	-	-	0.868	0.068	0.479	0.111	-	-	-	-	-	-	yes	yes
(ν,κ)	-	-	-	-	-	−0.918	0.035	−0.389	0.076	-	-	-	-	-	-	yes	no
(ν,f)	-	-	-	-	-	0.756	0.052	−0.650	0.056	-	-	-	-	-	-	yes	yes
(κ,f)	-	-	-	-	-	0.467	0.033	−0.883	0.017	-	-	-	-	-	-	yes	yes
$[(u,\nu),(\nu,\kappa)]$	0.707	−0.707	-	0.707	−0.707	0.730	0.047	−0.680	0.054	-	-	−0.656	0.011	0.755	0.010	yes	yes
$[(d,\nu),(\nu,\kappa)]$	0.707	−0.707	-	0.707	−0.707	0.695	0.028	−0.718	0.027	-	-	−0.687	0.008	0.727	0.007	yes	yes
$[(u,d),(u,\nu)]$	-	-	-	-	-	0.313	0.000	−0.950	0.000	-	-	−0.950	0.001	0.313	0.002	yes	no
$[(u,d),(d,f)]$	-	-	-	-	-	−0.576	0.056	−0.814	0.043	-	-	−0.593	0.017	0.805	0.013	yes	yes
$[(u,\nu),(\nu,f)]$	-	-	-	-	-	0.872	0.045	−0.482	0.074	-	-	−0.224	0.058	0.973	0.011	yes	no
$[(u,f),(f,d)]$	-	-	-	-	-	−0.665	0.052	−0.744	0.044	-	-	−0.503	0.036	0.863	0.020	yes	yes
$[(d,f),(f,\nu)]$	-	-	-	-	-	0.860	0.016	−0.510	0.026	-	-	−0.868	0.036	0.491	0.063	yes	yes
$[(u,d,\nu),(\nu,\kappa)]$	0.577	0.577	−0.577	0.707	−0.707	0.602	0.043	0.552	0.065	−0.570	0.053	0.698	0.011	−0.716	0.011	yes	yes
$[(u,d,f),(f,\nu)]$	-	-	-	-	-	0.580	0.002	0.660	0.014	−0.476	0.023	0.817	0.004	−0.576	0.006	yes	yes
$[(u,\kappa,f),(\kappa,\nu)]$	-	-	-	-	-	0.492	0.012	0.703	0.017	−0.512	0.026	0.625	0.012	−0.780	0.009	yes	yes

Table 2 also reports some identified sets that locally comply with the condition of invariance (see Supplementary Note S3 for more details), together with some randomly selected sets for which the procedure succeeds at finding the invariance property, but for which the results are close to the trivial solution, meaning that at least one of the evaluated exponents approaches either 0 or ± 1, thus is not reliable. Following the same approach above, Table 2 labels groups with low standard deviations as found and all the groups in which none of the evaluated exponents approaches 0 or ± 1 as reliable. Remarkably, the analytically evident couple (u, d), sets of two couples $[(u, \nu), (\nu, \kappa)]$ and $[(d, \nu), (\nu, \kappa)]$, and set comprising a triplet and a couple $[(u, d, \nu), (\nu, \kappa)]$, are identified. Furthermore, eight additional couples are found that locally comply with the condition of invariance, with six of them being reliable, and only one couple is not found. In the case of the selected sets of two feature couples, five more sets of two couples locally comply with the condition of invariance, with only two of them not being reliable. Finally, the two additional sets comprising a triplet and a couple are reliably found to represent a local invariance.

3.3.2. Use of Classification Models

In a second attempt at reducing the number of input variables, we aim at the construction of mixed optimized features, such as power combinations of the normalized primitive variables for classification, according to Equation (13). In the case of the Gnielinski correlation, we set class 1 for samples with $\overline{\mathrm{Nu}} < 500$ and class 2 for samples with $\overline{\mathrm{Nu}} \geq 500$. Once the optimization routine finds the Pareto front, the mixed features are created using the point of the Pareto front with the least overlapping of the two classes according to the Battacharyya distance. Figure 9a reports the PDF binning of the training set data over the two classes against the normalized flow velocity u_{norm}, while Figure 9b shows the same PDFs against the normalized mixed feature $\tilde{y}_1$; still, when represented against the primitive variable, the two classes exhibit significant overlap, whereas they appear distinctly separated when plotted against the mixed feature. Two bet-fitting GEV distributions are computed for the two classes according to Equation (17) for samples in class 1 with factors $\mu = 0.276$, $\sigma = 0.112$, $\omega = 0.223$; for samples in class 2 with factors $\mu = 0.510$, $\sigma = 0.091$, $\omega = 0.000$. Figure 9c shows the PDFs over the binned data of the testing set reported against the same mixed feature $\tilde{y}_1$, along with the GEV fittings computed on the training set. Notably, the classes are still well separated, with a good agreement between the GEV distributions and the testing set densities. Furthermore, we attempt to create two mixed normalized features $\tilde{y}_1, \tilde{y}_2$ with the same classes. Figure 10a shows the PDF two dimensional binning of the training set data over the two classes against the normalized flow velocity u_{norm} and friction factor f_{norm}. Figure 10b shows the same PDFs against the mixed features $\tilde{y}_1, \tilde{y}_2$ constructed according to Equation (13) by power combination of the five relevant features and choosing the point of the Pareto front with the least overlapping of the classes. Similarly to the one-dimensional case, the classes appear again well separated when plotted against the mixed features, whereas there is a wide overlap when plotted against the primitive variables. Figure 10c shows the PDFs over the binned data of the testing set, reported against the same mixed features $\tilde{y}_1, \tilde{y}_2$; the two classes still appear well separated. Finally, we aim at the construction of $m = 1$ mixed feature for separating the data samples into three classes: class 0 for $\overline{\mathrm{Nu}} < 400$, class 1 for $400 \leq \overline{\mathrm{Nu}} < 900$, class 2 for $\overline{\mathrm{Nu}} \geq 900$. The multi-objective optimization is performed aiming at concurrently maximizing the pairwise distances between the classes. Figure 11 shows the separation in classes reported against the normalized flow velocity u_{norm} and the new mixed feature $\tilde{y}_1$: in line with previous cases, the considerable overlap observed when representing classes against the primitive feature is significantly reduced when using the optimized mixed feature. The bet-fitting GEV distributions for the three classes are computed according to Equation (17) over the training set samples; specifically, the GEV associated with Class 0 has factors $\mu = 0.245$, $\sigma = 0.083$, $\omega = 0.346$, the GEV for Class 1 has factors $\mu = 0.430$, $\sigma = 0.053$, $\omega = 0.143$, and the GEV for Class 2 has factors $\mu = 0.599$, $\sigma = 0.074$, $\omega = 0.017$. When plotting the PDFs over the binned data of the testing set against the

mixed feature, the classes appear well separated, and there is good agreement between the GEV distributions and the testing set densities.

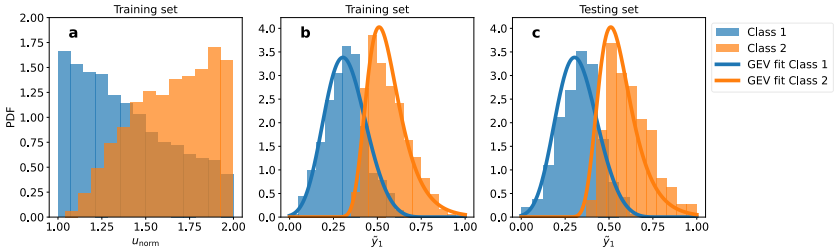

Figure 9. One dimensional example: (**a**) PDFs over binned data of the training set for the two classes ($\overline{\text{Nu}} < 500$ and $\overline{\text{Nu}} \geq 500$) reported against the normalized flow velocity. (**b**) PDFs over binned data of the training set for the two classes reported against the mixed feature y_1, constructed according to Equation (13) and choosing the point of the Pareto front with the least overlapping of the two classes according to the Bhattacharyya distance, along with a GEV analytical fitting of the two binnings. (**c**) PDFs over binned data of the testing set for the two classes reported against the same mixed feature y_1 together with the same GEV fittings of the (**b**) subfigure. The mixed variable $\tilde{y}_1$ shown here is referred exclusively to this Gnielinski one dimensional optimization.

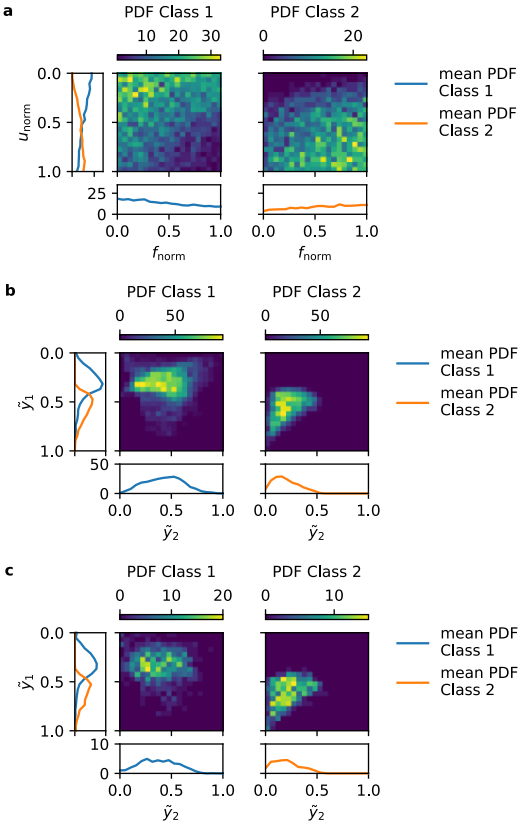

Figure 10. Two dimensional example for the classification on the Gnielinski correlation: (**a**) PDFs over binned data of the training set for the two classes ($\overline{\text{Nu}} < 500$ and $\overline{\text{Nu}} \geq 500$) reported against the normalized flow velocity u_{norm} and friction factor f. (**b**) PDFs over binned data of the training set for the two classes reported against the mixed features $\tilde{y}_1, \tilde{y}_2$, constructed according to Equation (13) and choosing the point of the Pareto front with the least overlapping of the two classes according to the Bhattacharyya distance. (**c**) PDFs over binned data of the testing set for the two classes reported against the same mixed features $\tilde{y}_1, \tilde{y}_2$. The mixed variables $\tilde{y}_1, \tilde{y}_2$ shown here are referred exclusively to this Gnielinski two dimensional optimization.

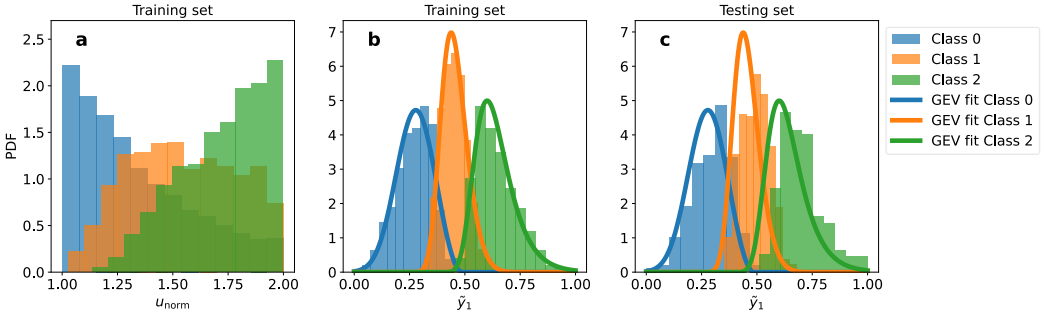

Figure 11. One dimensional example for the ternary classification on the Gnielinski correlation. (**a**) PDFs over binned data of the training set for the three classes ($\overline{\text{Nu}} < 400$, $400 \leq \overline{\text{Nu}} < 900$, and $\overline{\text{Nu}} \geq 900$) reported against the normalized flow velocity u_{norm}. (**b**) PDFs over binned data of the training set for the three classes reported against the mixed feature $\tilde{y}_1$, constructed according to Equation (13) and choosing the point of the Pareto front with the least overlapping of the three classes according to the Bhattacharyya distance, along with a GEV analytical fitting of the three binnings. (**c**) PDFs over binned data of the testing set for the three classes reported against the same mixed feature $\tilde{y}_1$ together with the same GEV fittings of the **b** subfigure. The mixed variable $\tilde{y}_1$ shown here is referred exclusively to this Gnielinski one-dimensional optimization.

3.4. Newton's Law of Universal Gravitation

The module of the interacting force F_{g} for two objects of masses m_1, m_2 is expressed via Newton's law of universal gravitation

$$F_{\text{g}} = \frac{G m_1 m_2}{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}, \tag{20}$$

where G is the gravitational constant, (x_1, y_1, z_1) and (x_2, y_2, z_2) are the coordinates of the centers of their masses.

In this case, our aim is to identify the invariance of the noised $\overline{F_{\text{g}}}$ with respect to groups $(x_1 - x_2)$, $(y_1 - y_2)$, $(z_1 - z_2)$. To this end, we employ the procedure presented in Section 2.2.3 for identification of general group form, focusing specifically on groups with functional dependence $\alpha_1 x_i + \alpha_2 x_j$. As usual, we obtain access to the gradient $\nabla \overline{F_{\text{g}}}(x_1, x_2, y_1, y_2, z_1, z_2, m_1, m_2)$ by means of automatic differentiation over a DNN approximating $\overline{F_{\text{g}}}(x_1, x_2, y_1, y_2, z_1, z_2, m_1, m_2)$. Figure 12a shows the model predictions over the testing set, while in Figure 12b, the corresponding loss across epochs is depicted; specifically, the model is highly predictive, with coefficient of determination $R^2 = 0.962$ and no evidence of overfitting is observed. The trained model is thus fed to the optimization algorithm. The average $\mu_{\alpha_1}, \mu_{\alpha_2}$ estimates of the coefficients α_1, α_2, together with their standard deviations $\sigma_{\alpha_1}, \sigma_{\alpha_2}$, are reported in Table 3. Specifically, the procedure correctly identifies the groups (x_1, x_2), (y_1, y_2), (z_1, z_2) with estimates $\mu_{\alpha_1}, \mu_{\alpha_2}$ of $(-0.679, 0.721)$, $(-0.711, 0.702)$, $(-0.683, 0.722)$, respectively, compared to the true normalized coefficients $(-0.707, 0.707)$ for all cases. Table 3 flags groups with low standard deviations as found and all the groups in which none of the evaluated exponents approaches 0 or 1 as reliable. Specifically, other variables couples show high variance over the corresponding average estimates, meaning that no further invariant group is identified.

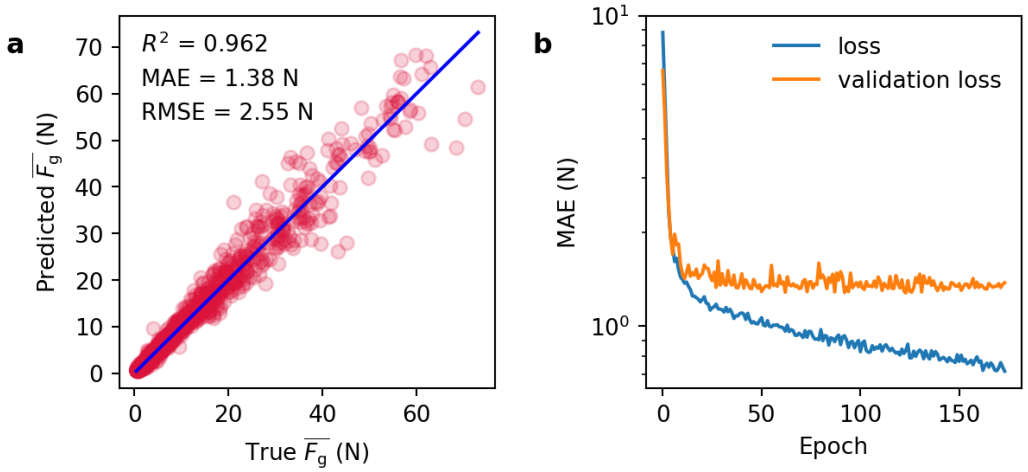

Figure 12. Results of the DNN regression model for the noised gravitational force F_g. (**a**) Predictions over the testing set, and (**b**) corresponding loss curves for the DNN model. Model performances are shown in terms of coefficient of determination R^2, mean absolute error (MAE), and root mean squared error (RMSE).

Table 3. Normalized exponents α_1, α_2 for the Newton's law of universal gravitation, together with their average estimates over 20 evaluations μ_{α_1}, μ_{α_2}, and the corresponding standard deviations σ_{α_1}, σ_{α_2}. Found groups/sets refer to low standard deviation, reliable groups refer to average far from 1 and 0.

Group	α_1	α_2	μ_{α_1}	σ_{α_1}	μ_{α_2}	σ_{α_2}	Found	Reliable
(x_1, x_2)	0.707	−0.707	−0.679	0.111	0.721	0.083	yes	yes
(y_1, y_2)	0.707	−0.707	−0.711	0.023	0.702	0.023	yes	yes
(z_1, z_2)	0.707	−0.707	−0.683	0.030	0.722	0.044	yes	yes
(m_1, m_2)	-	-	0.105	0.556	0.759	0.322	no	-
(m_1, x_1)	-	-	0.000	0.000	1.000	0.000	no	-
(x_1, y_1)	-	-	−0.604	0.369	0.335	0.622	no	-

4. Conclusions and Final Remarks

In this study, we have implemented two innovative methodologies for searching optimal variables to describe physical data, making use of both regression and classification models applied to the data. Specifically, leveraged on well-suited datasets for machine learning, with the goal of predicting a property of interest.

In particular, the methodology introduced for regression tasks has the ambition to find groups of variables that are valid over all the parameter space spanned by the available data. However, due to various factors such as noise in the data, this method might not converge, even if the group exists. Conversely, the methodology introduced for classification tasks gives up on this ambition in favor of the greater robustness of an optimization framework, aiming to only find possible separations in the parameter space. This can be useful, for instance, to find optimal areas in the parameters space with the highest values of a properly defined objective function of interest, even if the group does not govern globally such objective function. Due to their different objectives, the two methods generally do not produce the same groups.

More precisely, the procedure based on the regression model introduced here enables the identification of invariant groups and/or sets of variable groups with respect to which the property of interest is invariant. We have demonstrated its effectiveness on noised data generated by three well known functional relationships: the Dittus–Boelter correlation, the Gnielinski correlation, and Newton's law of universal gravitation. It is noteworthy that we did not apply any particular noise reduction technique to the three datasets; instead, our methods directly handle the noisy data as-is, demonstrating their robustness. Also, we

acknowledge that direct measurement of the Nusselt number would not be feasible in real world scenarios, and that the calculation method of the examined properties of interest (i.e., Nusselt number and gravitational force) is already established. However, our selection of such examples is only aimed at showcasing the effectiveness of our methodologies in sufficiently simple cases. For the former two cases, the procedure has accurately detected groups/sets in power form, while for the latter case, a generalized algorithm successfully identified groups with a linear form. It is worth stressing that, in all the examples, our procedure did not end up with any false negative, i.e., whenever a group exists, it is correctly identified. Interestingly, the methodology is potentially applicable to any functional form, as illustrated in Section 2.2.3. Additionally, the effectiveness of identifying a group in a situation where that group is already known—like in this study—depends on a tolerance threshold set between the identified group and the actual one. In real world scenarios, the identified groups serve as candidates of mixed variables, whose validity has to be verified a posteriori.

The procedure based on classification of the physical property values, allows for the determination of an appropriate set of exponents to combine all the primitive variables in power form, thus constructing mixed features optimized for the classification task. Specifically, we have shown its effectiveness on the Dittus–Boelter and on the Gnielinski correlations. Indeed, the methodology effectively enables the separation of classes even with just one mixed feature, whereas a single primitive variable fails to achieve class separation. Furthermore, we also provide examples with one and two mixed variables, together with separation in two or three classes.

It is worth stressing that the methodologies presented in this study are blessed by generality and, as such, are not restricted to the selected case studies, and are agnostic in terms of data origin, being it numerical or experimental. Therefore, potential applications can be envisioned in various other fields in the future. In particular, we notice that the identification of effective good variables is not only interesting per se to possibly gain a deeper insight on a physical system, but can also be practically advantageous when designing experiments. Indeed, the methodology to detect groups/sets in regression facilitates efficient group/set-level adjustments rather than individually fine-tuning variables within the groups/sets themselves.

Moreover, the classification procedure can enable the reduction of the numerous original primitive variables to a minimal set of optimized variables concerning a specific physical property of interest. Notably, combinations of variables yielding identical mixed feature values exhibit similar performance in terms of the property being classified. As a result, it is possible to find alternative combinations of primitive variables without compromising the overall performance of a given system under study. From a more practical standpoint, the methodologies described here can help generalizing optimal system conditions, thus helping decreasing the most resource-intensive components while properly re-balancing the others. As an example, these general methodologies may thus hold the potential to save resources, such as costly reagents (e.g., Bonke et al. [24] for solar fuel production) or expensive materials as in perovskite solar cells optimization [33]. An additional advantage associated with the correct identification of a reduced set of ruling variables is their possible use for driving sequential learning or Bayesian optimization processes [34,35].

Finally, we believe that an interesting development of the presented methodologies to be pursued in the future, shall be in the direction of a possible handling of systems ruled not only by numerical parameters, but also categorical ones.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/make6030077/s1, Supplementary Note S1: Properties of liquids; Supplementary Note S2: DNN structure; Supplementary Note S3: Local invariance; Supplementary Note S4: Mixed optimized variables for classification.

Author Contributions: Conceptualization, E.C.; methodology, E.C. and G.T.; software, E.C., G.T. and G.B.; validation, G.B. and G.T.; formal analysis, E.C., G.T. and G.B.; investigation, E.C., G.T. and G.B.; resources, E.C.; data curation, G.B. and G.T.; writing—original draft preparation, G.B.; writing—review and editing, G.B., G.T. and E.C.; visualization, G.B. and G.T.; supervision, E.C.; project administration, E.C.; funding acquisition, E.C. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge funding under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3—Call for tender No. 1561 of 11.10.2022 of Ministero dell'Università e della Ricerca (MUR); funded by the European Union—NextGenerationE

Data Availability Statement: Processed datasets and trained models of this study are publicly available in Zenodo at DOI: 10.5281/zenodo.10406490. The codes used to obtain the results of this study are publicly available in github at https://github.com/giuliobarl/GoodPhysVariables accessed on 11th July 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine Learning
DNN	Deep Neural Network
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
PDF	Probability Density Functions
GEV	Generalized Extreme Value

References

1. Rappaz, M.; Bellet, M.; Deville, M.O.; Snyder, R. *Numerical Modeling in Materials Science and Engineering*; Springer: Berlin/Heidelberg, Germany, 2003.
2. Chen, B.; Huang, K.; Raghupathi, S.; Chandratreya, I.; Du, Q.; Lipson, H. Automated discovery of fundamental variables hidden in experimental data. *Nat. Comput. Sci.* **2022**, *2*, 433–442. [CrossRef] [PubMed]
3. Floryan, D.; Graham, M.D. Data-driven discovery of intrinsic dynamics. *Nat. Mach. Intell.* **2022**, *4*, 1113–1120. [CrossRef]
4. Eva, B.; Ried, K.; Müller, T.; Briegel, H.J. How a Minimal Learning Agent can Infer the Existence of Unobserved Variables in a Complex Environment. *Minds Mach.* **2023**, *33*, 185–219. [CrossRef] [PubMed]
5. Chiavazzo, E. Approximation of slow and fast dynamics in multiscale dynamical systems by the linearized Relaxation Redistribution Method. *J. Comput. Phys.* **2012**, *231*, 1751–1765. [CrossRef]
6. Chiavazzo, E.; Karlin, I.V. Quasi-equilibrium grid algorithm: Geometric construction for model reduction. *J. Comput. Phys.* **2008**, *227*, 5535–5560. [CrossRef]
7. Rayleigh, L., VIII. On the question of the stability of the flow of fluids. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1892**, *34*, 59–70. [CrossRef]
8. Buckingham, E. On physically similar systems; illustrations of the use of dimensional equations. *Phys. Rev.* **1914**, *4*, 345. [CrossRef]
9. Curtis, W.; Logan, J.D.; Parker, W. Dimensional analysis and the pi theorem. *Linear Algebra Its Appl.* **1982**, *47*, 117–126. [CrossRef]
10. Chiavazzo, E.; Covino, R.; Coifman, R.R.; Gear, C.W.; Georgiou, A.S.; Hummer, G.; Kevrekidis, I.G. Intrinsic map dynamics exploration for uncharted effective free-energy landscapes. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E5494–E5503. [CrossRef]
11. Chiavazzo, E.; Gear, C.W.; Dsilva, C.J.; Rabin, N.; Kevrekidis, I.G. Reduced models in chemical kinetics via nonlinear data-mining. *Processes* **2014**, *2*, 112–140. [CrossRef]
12. Lin, K.K.; Lu, F. Data-driven model reduction, Wiener projections, and the Koopman-Mori-Zwanzig formalism. *J. Comput. Phys.* **2021**, *424*, 109864. [CrossRef]
13. McRee, R.K. Symbolic regression using nearest neighbor indexing. In Proceedings of the 12th Annual Conference Companion on Genetic and Evolutionary Computation, New York, NY, USA, 7–11 July 2010; pp. 1983–1990.
14. Stijven, S.; Minnebo, W.; Vladislavleva, K. Separating the wheat from the chaff: On feature selection and feature importance in regression random forests and symbolic regression. In Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation, Dublin, Ireland, 12–16 July 2011; pp. 623–630.
15. McConaghy, T. FFX: Fast, scalable, deterministic symbolic regression technology. In *Genetic Programming Theory and Practice IX*; Springer: New York, NY, USA, 2011; pp. 235–260.
16. Arnaldo, I.; O'Reilly, U.M.; Veeramachaneni, K. Building predictive models via feature synthesis. In Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation, New York, NY, USA, 11–15 July 2015; pp. 983–990.

17. Brunton, S.L.; Proctor, J.L.; Kutz, J.N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 3932–3937. [CrossRef]
18. Quade, M.; Abel, M.; Nathan Kutz, J.; Brunton, S.L. Sparse identification of nonlinear dynamics for rapid model recovery. *Chaos Interdiscip. J. Nonlinear Sci.* **2018**, *28*, 063116. [CrossRef]
19. Searson, D.P.; Leahy, D.E.; Willis, M.J. GPTIPS: An open source genetic programming toolbox for multigene symbolic regression. In Proceedings of the International Multiconference of Engineers and Computer Scientists, Hong Kong, China, 17–19 March 2010; Citeseer; Volume 1, pp. 77–80.
20. Dubčáková, R. Eureqa: Software Review. 2011. Available online: https://www.researchgate.net/publication/220286070_Eureqa_software_review (accessed on 5 May 2024).
21. Schmidt, M.; Lipson, H. Distilling free-form natural laws from experimental data. *Science* **2009**, *324*, 81–85. [CrossRef]
22. Udrescu, S.M.; Tegmark, M. AI Feynman: A physics-inspired method for symbolic regression. *Sci. Adv.* **2020**, *6*, eaay2631. [CrossRef] [PubMed]
23. Trezza, G.; Chiavazzo, E. Leveraging composition-based energy material descriptors for machine learning models. *Mater. Today Commun.* **2023**, *36*, 106579. [CrossRef]
24. Bonke, S.A.; Trezza, G.; Bergamasco, L.; Song, H.; Rodríguez-Jiménez, S.; Hammarström, L.; Chiavazzo, E.; Reisner, E. Multi-Variable Multi-Metric Optimization of Self-Assembled Photocatalytic CO2 Reduction Performance Using Machine Learning Algorithms. *J. Am. Chem. Soc.* **2024**, *146*, 15648–15658. [CrossRef]
25. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
26. Al-Helali, B.; Chen, Q.; Xue, B.; Zhang, M. Genetic Programming for Feature Selection Based on Feature Removal Impact in High-Dimensional Symbolic Regression. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *8*, 2269–2282. [CrossRef]
27. Branch, M.A.; Coleman, T.F.; Li, Y. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM J. Sci. Comput.* **1999**, *21*, 1–23. [CrossRef]
28. Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distribution. *Bull. Calcutta Math. Soc.* **1943**, *35*, 99–110.
29. Bhattacharyya, A. On a measure of divergence between two multinomial populations. *Sankhyā Indian J. Stat.* **1946**, *7*, 401–406.
30. Villani, C. *Optimal Transport: Old and New*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 338.
31. Lide, D.R.; Kehiaian, H.V. *CRC Handbook of Thermophysical and Thermochemical Data*; CRC Press: Boca Raton, FL, USA, 2020.
32. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef] [PubMed]
33. Tailor, N.K.; Abdi-Jalebi, M.; Gupta, V.; Hu, H.; Dar, M.I.; Li, G.; Satapathi, S. Recent progress in morphology optimization in perovskite solar cell. *J. Mater. Chem. A* **2020**, *8*, 21356–21386. [CrossRef]
34. Huan, X.; Marzouk, Y.M. Simulation-based optimal Bayesian experimental design for nonlinear systems. *J. Comput. Phys.* **2013**, *232*, 288–317. [CrossRef]
35. Motoyama, Y.; Tamura, R.; Yoshimi, K.; Terayama, K.; Ueno, T.; Tsuda, K. Bayesian optimization package: PHYSBO. *Comput. Phys. Commun.* **2022**, *278*, 108405. [CrossRef]