Archives of Computational Methods in Engineering (2023) 30:3845–3865
https://doi.org/10.1007/s11831-023-09922-z

REVIEW ARTICLE

Artificial Intelligence in Physical Sciences: Symbolic Regression Trends and Perspectives

Dimitrios Angelis[1] · Filippos Sofos[1] · Theodoros E. Karakasidis[1]

Received: 4 December 2022 / Accepted: 27 March 2023 / Published online: 19 April 2023

Abstract

Symbolic regression (SR) is a machine learning-based regression method based on genetic programming principles that integrates techniques and processes from heterogeneous scientific fields and is capable of providing analytical equations purely from data. This remarkable characteristic diminishes the need to incorporate prior knowledge about the investigated system. SR can spot profound and elucidate ambiguous relations that can be generalizable, applicable, explainable and span over most scientific, technological, economical, and social principles. In this review, current state of the art is documented, technical and physical characteristics of SR are presented, the available programming techniques are investigated, fields of application are explored, and future perspectives are discussed.

1 Introduction

Data science has been the driving force for the dawn of the fourth industrial revolution, bringing the big-data concept in focus of most science and engineering applications. It has been stated that the global datasphere will get as high as 175 Zettabytes by the year 2025 while it was merely 33 Zettabytes in 2018 [1]. As a result, researchers from most disciplines have been motivated on exploring ways to deploy this vast amount of data. The main idea is the identification of patterns and/or hidden equations that govern these datasets. Data mining, apart from the exploitation of well-established statistical and numerical methods, has also been directed towards the extraction of mathematical expressions, which can be utilized for the establishment of new and the verification of existing physical laws. Therefore, the question that now arises is: Should we acknowledge the era we experience as the era of big data, and if so, what is the effect on science and technology?

Novel data-driven approaches are now exploited in materials science, among others, [2] (see Fig. 1) and have opened the road to the introduction of Materials Informatics [3, 4] and relevant approaches whose primary concern is the discovery of novel materials at reasonable computational cost. Furthermore, in order to enhance this initiative, there exist databases (e.g., Inorganic Crystal Structure Database [5], Open Quantum Material Database [6], The Cambridge Structural Database [7], AFLOWLIB [8]) to provide adequate support for scientists and engineers. However, materials science isn't the only field that has benefited from the advent of big data. The fourth paradigm of science, under the framework of Artificial Intelligence (AI) to facilitate the procedure [9], has substantially evolved in fields such as bioinformatics, particle physics, space research, medical imaging, construction applications, and more. To support this observation, the universal recognition of big data and data-driven methods is further becoming clear by the vast increase of published articles on the topic [10].

Nevertheless, big data would be hard to handle without the incorporation of AI, or, in other words, the assemblage of methods and skills that allow humans and machines to execute assignments that only intelligent entities can (e.g., perceive, reason, and act) [11]. Having been introduced as a branch of computer science, AI is further comprised of techniques with their own categories of algorithms (for example, Machine Learning and Particle Swarm Optimization [12]), suggesting a statistical, predictive framework that

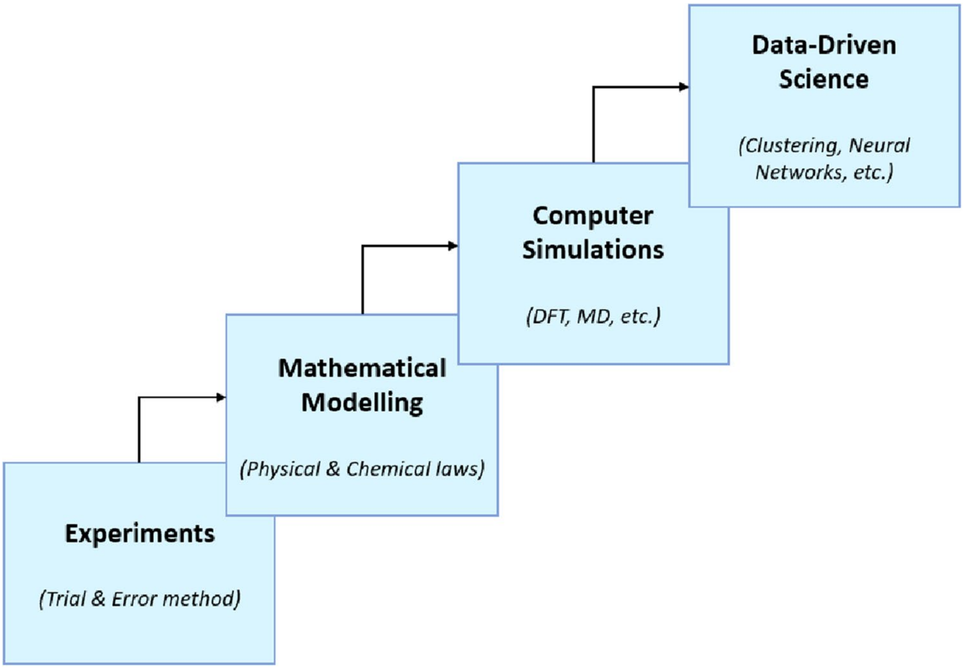
✉ Filippos Sofos
fsofos@uth.gr

Dimitrios Angelis
dimangelis@uth.gr

Theodoros E. Karakasidis
thkarak@uth.gr

[1] Condensed Matter Physics Laboratory, Department of Physics, University of Thessaly, Lamia 35100, Greece

Fig. 1 The four paradigms of science: empirical, theoretical, computational and data-driven



can be bound to a specific problem both in research and in everyday life.

Machine Learning (ML) is an AI technique that has acquired major interest for data analysis tasks, based on its ability to learn from experience and, therefore, provides accurate approximations and/or predictions on the underlying patterns [13]. Oftentimes, common statistical analysis has been misidentified to ML, however, the former aims on the accurate description of observations while the latter is constructing algorithms whose focus lies on successful data classification and approximations in order to predict the outcome [14]. ML is widely recognized as an effective tool for research and applied purposes, in fields like additive manufacturing [15, 16, 17, 18], materials science [14, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29], autonomous driving [30, 31, 32], solar cells [33, 34, 35, 36, 37, 38, 39], chemistry [40, 41, 42], welding industry [43], solar radiation [44, 45, 46, 47, 48, 49, 50, 51] and many more.

Notwithstanding the data-driven techniques evolution, barriers have arisen when it comes to the interpretation of the underlying physics, that has somehow to be extracted from data patterns. A successful attempt to bind physical laws to the solution of partial differential equations has been made with the incorporation of Physics-Informed Neural Networks (PINNs) [52] and some of their new implementations, such as PI-GP [53] and B-PINNs [54]. From another point of view, during the development of a NN the output triggered by an input value is predetermined, and that means the mapping could be abstract [55]. Although this is a procedure that machines can deal with, it might be extremely difficult for humans to make a sense out of it [56], since this “black-box” model [57] does not adhere to the principles of explainable and generalizable AI. Several problems may also come up when these models are used in the field, for instance, when a manufacturer of an autonomous car cannot apprehend the choice that the car will do in infrequent critical situations (as in protecting the driver or pedestrians in an imminent crash), concerns might appear on the applicability of the algorithm [58]. The absence of a physical interpretation in data-driven black-box models, has been the driving force for a change of heart.

This review has been focused on presenting an ML-based method, Symbolic Regression (SR), which has been developed on Evolutionary Computing principles. SR is differentiated from a purely data-driven black-box model, as it is equipped with the ability to generate symbolic expressions (analytical equations) without considering prior constraints and provides a physics-inspired overlook. The ever growing adaptation of SR in various scientific fields has been captured by the rapid growth of related published papers in the last decade and more (see Fig. 2) and it would be beneficial to dive deep into its remarkable characteristics that have made it a new trend in physics-based computations. In the sections that follow, there will be a brief introduction on various ML types (Sect. 2) adopted in science and engineering, a comprehensive analysis of SR and the respective algorithms incorporated (Sect. 3), mapping of scientific fields where SR has been successfully employed and others that have shown a promising prospect (Sect. 4) and, finally, we sum up with a

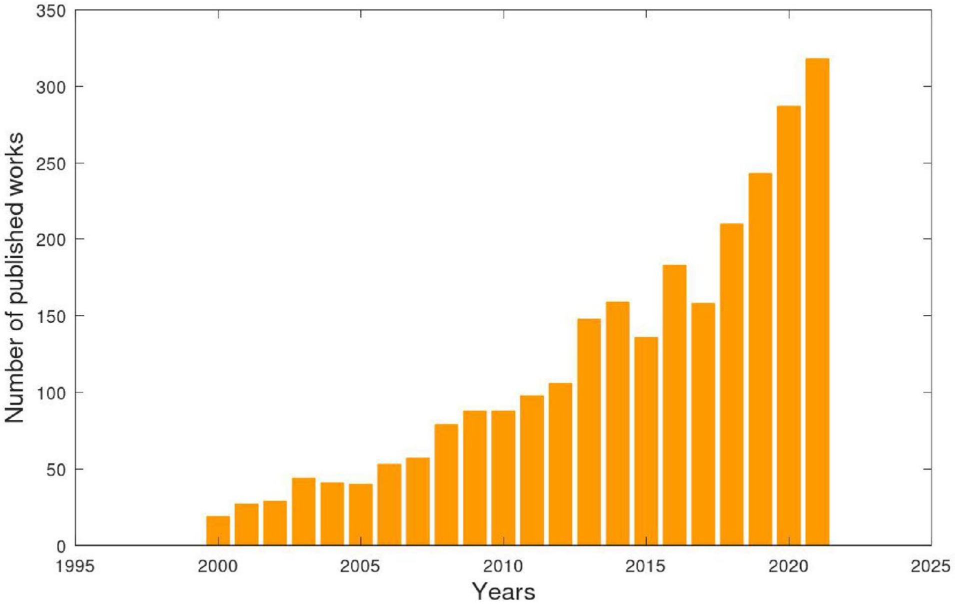


Fig. 2 Publications related to symbolic regression from 2000 to 2021

concluding discussion and present the future perspectives of SR (Sect. 5).

2 Machine Learning

The present section is organized as follows. Firstly, categories and useful information about ML will be introduced. Secondly, there will be a brief presentation of several ML algorithms that researchers are most familiar with. Finally, more complex approximations such as Deep Learning will be discussed.

2.1 Machine Learning Categories

Machine Learning can be subdivided into three major categories, (i) Supervised Learning (SL), (ii) Unsupervised Learning (UL) and (iii) Reinforcement Learning (RL) [59], although the latter isn't always recognized as a separate division. In SL, the training procedure involves data manipulation based on labeled input/output pairs, while the implied algorithms seek for determining a hidden function able to map input behavior to the desired output. On the other hand, in UL, no labels are specified on the input data and the algorithmic procedures urge to reveal the implied data interconnections [60]. In RL, the model does not require input data as it constructs its own by self-training and, furthermore, it is self-challenged to achieve higher accuracy metrics [61]. A combination of SL and UL has also been proposed in Semi-Supervised Learning (SSL) in which both labeled and unlabeled data are utilized. SSL focuses on the identification of how the learning procedure may be affected by a mixture of labeled and unlabeled data and the construction of algorithms capable of exploiting this scheme [62].

A successful ML algorithm implementation is tightly paired with the quality and quantity of available data. In cases where data extracted from various databases and/or literature sources poses no viable option, attention is drawn into experimental or simulation output (e.g., Density-Functional Theory (DFT), Molecular Dynamics (MD), etc.) [23], posing a four-way route to acquire data. Attention has to be drawn also on data representation, as it may vary from discrete (e.g., texts) to continuous (e.g., vectors and tensors) or weighted graphs [22]. Still, data availability is not enough; it is imperative to follow a common format [63] and, most of the times, a pre-processing step is required [23].

However, an inherent disadvantage of such approaches is bound to the fact that ML methods are prone to overfitting. Overfitting occurs when an algorithm is "finely" trained on a specific dataset, which in turn results on high statistical errors when applied to a different dataset. In such case, the proposed algorithm becomes unsuitable for further use, or in other words, ungeneralizable. To overcome this issue, various techniques have been proposed, such as hold-out, k-fold cross-validation, and regularization [16].

2.2 Machine Learning Algorithms

There has been a wealth of ML algorithms being established throughout the years, from simple linear models to laborious deep learning architectures. Some of the most widely adopted are Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Decision Trees (DT) based. The idea behind ANN is that it reacts the same way as the neural networks of human brain, with its abilities spanning to applications such as classification, regression, learning and generalization [64]. SVM models are utilized for classification, while its regression counterpart is the Support Vector Regression (SVR) model [28]. DT methods employ tree-form graphs and have been utilized for classification tasks. DTs are susceptible to complexity and overfitting issues, and various alternatives have been constructed, such as Random Forest (RF) and Gradient Boost (GB), by employing different trees in a forest or a serial weighted manner, respectively [14]. There are many references in the literature for these models, which are not going to be covered here (see, for example, [65]).

Nevertheless, although implementations based on Shallow Learning algorithms, such as ANNs, SVMs, and DTs, are quite effective in numerous fields (e.g., materials science), some issues may occur in demanding applications, such as poor accuracy over DFT simulations [22]. This has opened the discussion on establishing more robust methods, such as Deep Learning (DL).

2.3 Deep Learning

While Shallow Learning methods may be effective and accurate on dealing with data over a small set of computational nodes, DL is capable of exploiting big data by mapping it to multiple layers in order to extract information and make predictions [23], even on noisy data [66]. DL models embed mathematical concepts (linear algebra, probability theory) and programming techniques in hidden layers that span over a number of thousands or more [12, 21], making them perfectly fit in applications ranging from processing videos and images [67, 68, 69, 70], speech recognition [71], and bio-informatics [72], among others. Nevertheless, physical interpretation of the outcome still lacks, and this would benefit their application in physical sciences, where interpretability has a central role. In such cases, it would be beneficial to adopt models that produce meaningful results (i.e., mathematical expressions), which could spot correlations with existing empirical relations and propose an analytical approach bound to physical laws.

3 Symbolic Regression

Symbolic Regression is a type of regression analysis in which a mathematical function that describes a given dataset is derived. While conventional regression methods (e.g., linear, quadratic, etc.) have their independent variable(s) predetermined and try to adjust a number of numerical coefficients in order to achieve perfect fit, SR attempts to find the parameters and equations simultaneously [55].

Derived from the superset of AI available methods, SR is usually implemented by evolutionary algorithms. At the same time, the most widely adopted concepts for SR construction are adopted from Genetic Programming (GP) [73]. In this section, the basic features of GP will be presented, an analytical development of the basic SR procedure is to be exemplified, several features that characterize SR superiority will be presented, while available SR programming techniques will be evaluated.

3.1 Genetic Programming Fundamentals

Genetic Programming (GP) [74] is an Evolutionary Algorithm (EA) instance [75] which in turn is a subset of Evolutionary Computing (EC) [76] (see Fig. 3). Moreover, GP provides the framework to express data behavior through mathematical equations, by exploring the available mathematical space in an evolutionary process. It is a fact that GP can find applicability in most regression-based science and engineering problems. The procedure that GP follows, includes the construction of different symbolic expressions, on which a comparison is made on its parts. The expressions that do not comply with accuracy and complexity measures being set are discarded, while those that appear as a potential solution to the problem are combined and form an output expression able to produce the desired outcome. The most common way to visualize a symbolic expression is a tree-structure with nodes and branches. Currently there are numerous programming options that

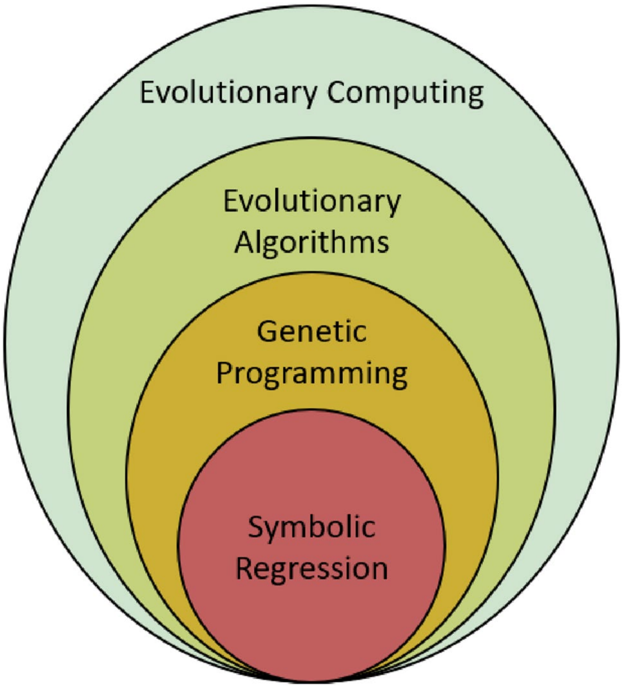


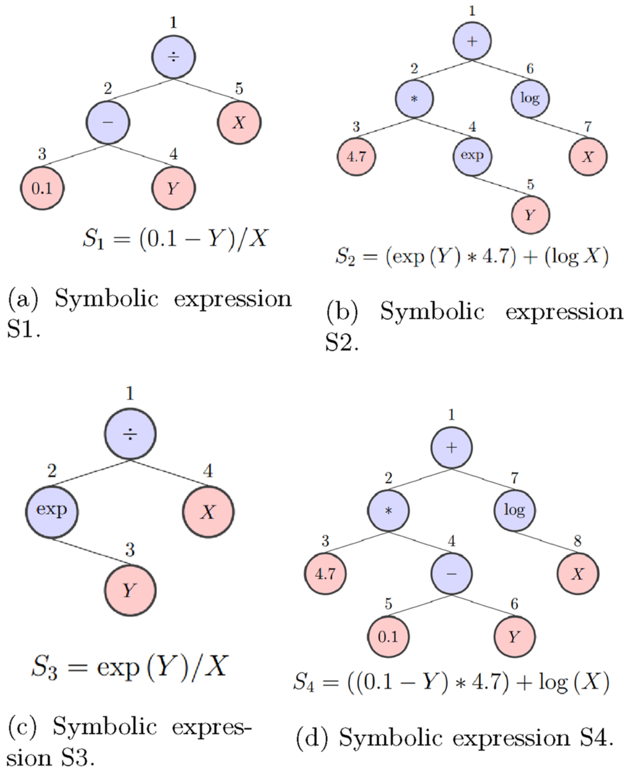
Fig. 3 From evolutionary computing to symbolic regression

can implement GP calculations, such as the Glyph package in Python [77] and GPTIPS toolbox in MatLab [78].

3.2 GP-SR Procedure

The GP tree-structure scheme consists of primitive functions and terminal nodes. While primitive functions could appear in any possible form (e.g., $+$, $-$, $*, \div, \log$, etc.), terminal nodes correspond to data inputs (e.g., X, Y) and numeric constants (e.g., 0.1) that may be needed to construct a symbolic expression. The final combination gives the desired expression in the form of a rooted tree. For example, in Fig. 4a, a tree-structure that corresponds to the symbolic expression S_1 is presented. The numbers that appear above each node are irrelevant to the procedure and serve only as a reference point. Nodes with number 1 and 2 are primitive functions (e.g., $\div$, $-$) while nodes 3,4,5 are filled with a numerical constant (e.g., 0.1) and input variables (e.g., X, Y). Configurations of primitive functions and terminals are drawn hierarchically in GP, while at the same time they serve as "individuals" from a population that contains a plurality of those.

The process taking place aims on finding the proper number of nodes/terminals that achieve the best fit over a given dataset and correspond to a final equation. For that reason, GP evolves by exploring every possible implementation between the primitive functions and terminals. To gain insight on the complexity implied, it should be noted that the search takes place in an infinite space that includes all available mathematical operators and numerical constants. Therefore, to search for the optimal fit, comparison measures are required, and these are related to the nature of the problem investigated. During an equation search, each possible implementation is rated by the number of data it can effectively handle, and the one that seems more likely to contribute on the final equation is evaluated on the error it produces. A common error metric usually incorporated in SR algorithms is the sum of the squares (*SSE*), derived from the differences between the predicted outcome and the tabulated output value [56], given by:



$S_1 = (0.1 - Y)/X$

(a) Symbolic expression S1.

$S_2 = (\exp(Y) * 4.7) + (\log X)$

(b) Symbolic expression S2.

$S_3 = \exp(Y)/X$

(c) Symbolic expression S3.

$S_4 = ((0.1 - Y) * 4.7) + \log(X)$

(d) Symbolic expression S4.

Fig. 4 Symbolic expression examples

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \,, \tag{1}$$

or due to the fact that the search occurs on various instances, the error could be measured as an average (Mean Squared Error (*MSE*)), with a mathematical formula of:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \,, \tag{2}$$

where, n is the total number of inspected situations, i corresponds to the individual situation, $\hat{Y}$ is the predicted value of the configuration and Y is the correct answer.

Starting from the top, from an infinite pool of choices, GP creates the first population randomly, while the size of the population is predetermined by the user. Each configuration inside the population could possibly contribute as a poor fitness parameter, affecting the whole dataset. However, some configurations might appear more effective than others. By continuously examining the effect of each one of these implementations, the process discovers promising candidates that produce small error, and employs them in future implementations, while those with poor performance are abandoned. Furthermore, the selection of those parts (sub-configurations) are random (e.g., sub-configuration 2-3-4 from Fig. 4a), resulting that way on a generated symbolic expression that differs in terms of tree shape and depth compared to the parental expression.

For instance, let there be another symbolic expression S_2 as shown in Fig. 4b. Additionally, let S_2 be deemed efficient by the comparison on fitness and therefore sustain the combination as described above. Finally, let the node with number 2 from Fig. 4a (sub-configuration 2-3-4) and the node with number 4 from Fig. 4b (sub-configuration 4-5) be randomly selected for the combination. Then, the configurations in the symbolic expression S_1 and S_2 have their sub-configurations exchanged. When the exchanging procedure is completed, two new configurations have emerged by the parts of their parents, while the parental configurations are concurrently removed from the procedure. Thus, new symbolic expressions are obtained, S_3 and S_4, which are depicted in Fig. 4c and d respectively. Although their shape has changed and contributed to the creation of new expressions, the fact that their depth remained the same is purely coincidental, as the investigation of possible configurations and combination of those, occurs on thousands of expressions. It should also be noted that the exchange takes place in parallel and by an iterative manner, random parts of promising expressions are combined till the final goal of obtaining a robust equation.

The evolutionary procedure proceeds step by step, as the average error is decreased due to the removal of poor-performing expressions. Eventually, at a predetermined point, the GP sequence finalizes, and the equation is exported. Reasons for termination could be either the definition of a maximum step on the algorithm (i.e., a limit set on the created populations by the user), a point where the statistical error of an individual equals to zero (optimal equation) or a point where the statistical error is lower than a given constant (e.g., 0.01). However, a mixture of those is preferred to cover every aspect. Finally, it should be noted that it is not mandatory for GP to produce one equation as it can be defined to export a number of suggestions, usually with a ranging complexity between equations. Here, it has to be beard in mind that low complexity might indicate poor error performance, while high complexity value could be prone to overfitting (Pareto front) [56].

In other words, GP initiates the procedure by creating an initial population filled by random symbolic expressions, with dimensions that vary according to the user configuration. Random mutations take place to minimize the possibility of the algorithm being trapped in local minima [79]. A way to visualize symbolic expressions is by a tree-structure form that contains primitive functions and terminal constants. Additionally, by producing a great number of expressions, GP compares them on the basis of how thy fitting on the given dataset. Candidate expressions that achieve small error are selected for further investigation. It should be reminisced that the nature of the comparison varies according to the problem's domain, while in the present case, the *SSE*, or the *MSE* are employed. Furthermore, GP surveys all possible sub-configurations and combines them in a way that creates a combined superset, whereas the parental expressions are discarded, and the final equation(s) are exported. A flowchart that employs GP evolution is illustrated in Fig. 5.

3.3 Programming Techniques

During the past decades, the extreme computational cost of SR implementations has been the main reason that posed barriers on its wide adaptation. Nowadays, as hardware has evolved in fast parallel architectures, the road for new SR approaches has opened, suggesting both efficiency and calculation speed.

At present, there are ample methods in the literature that propose different SR implementations. To mention a few, there exist models that employ a Monte Carlo tree search (MCTS) algorithm [80], a matrix-based encoding process [55], and advanced pre-processing schemes applied before algorithmic training [81] or even before regression begins

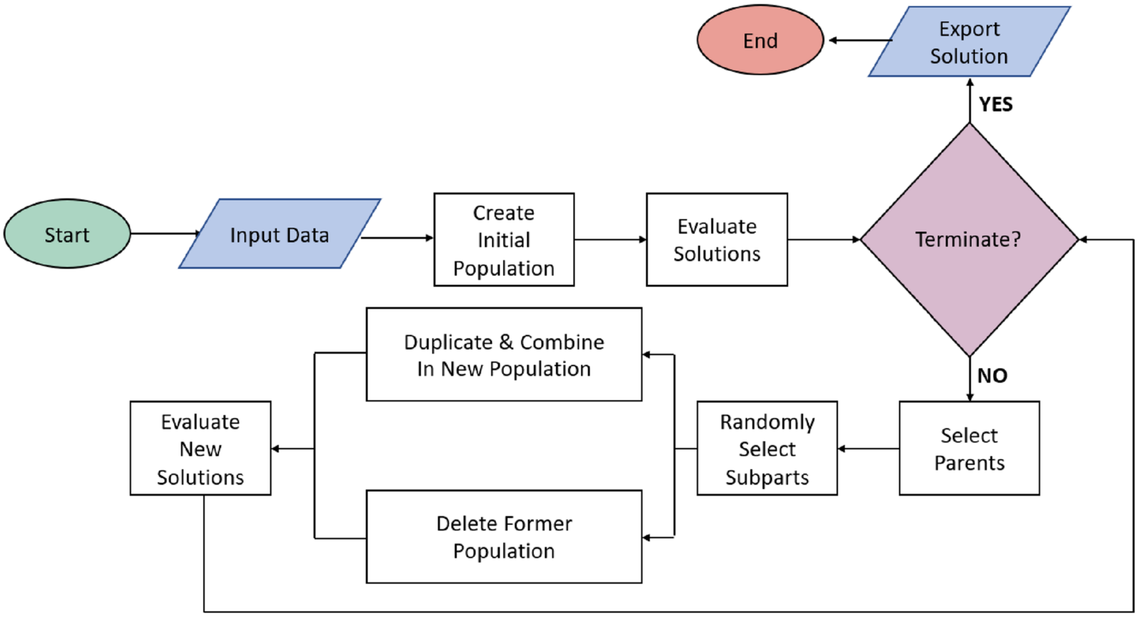


Fig. 5 Genetic programming based symbolic regression flowchart

[82]. Others have generated algorithms such as nearest neighbor indexing [83] or non-evolutionary techniques such as the FFX algorithm [73]. The latter, although extremely fast, produces non-interpretable equations [84]. In addition, there also exist Mixed-Integer Non-Linear Programming (MINLP) formulations [85, 86], models that identify the SR problem as a linguistic [87], others that incorporate probabilistic features such as probabilistic framework [88] or probabilistic grammars [89], Bayesian approaches [90] and more [91, 92].

The concept of increased accuracy is the focal point of GP approaches as it constitutes a key element on the applicability of ML models [93]. There have been efforts on modifying the basic GP-SR procedure [18, 94, 95, 96, 97, 98], while, on the other hand, some argue that GP-based procedures lead to abstract mathematical formulas that make no sense [83]. An intriguing idea that is also supported is the restriction of the search space into a set of symbols, by incorporating several constraints into the algorithm (usually by accepting prior knowledge about the system) [55, 58, 89, 99, 100]. Towards this direction, one could enforce a certain pattern to the generated expressions, such as monotonicity or symmetry and, as a result, increase the accuracy, reduce the computational cost, while, in parallel, adhere to well-established physical empirical or analytical expressions that already exist in the literature and need modifications according to the problem under investigation [101].

However, researchers should be aware of the limitations implied in such approaches. For instance, a non-GP approach [84] which constraints the derived formula to a set of mathematical symbols, has produced accurate predictions with low computational cost, but, on the downside, the algorithm has shown limited applicability and adaptability, since it cannot employ more complex cases, where periodicity [84] or exponentiality is needed. Hybrid methods have also been proposed, such as Deterministic/GP-SR models [102], neural-guided/GP [103] and others [99, 104, 105, 106]. Moreover, Neural Network based architectures have been employed [107, 108], while others have gone a step further and merged NN-based models with several DL features [109, 110].

Promising results have been obtained through methods that incorporate Bayesian Neural Networks (BNN) [111]. Bayesian statistics [112] in contrast to conventional statistics (also known as frequentist statistics) do not consider a fixed parameter, but they rather identify it as a random variable which can be described with a probability distribution. BNN functions as a typical Neural Network with the exception that the parameters are distributions, instead of a fixed value, and the training occurs via Bayesian inference. This is an important feature of BNN as it provides the ability to quantify uncertainties, meaning that the algorithm incorporates confidence intervals instead of a single point. Moreover, Bayesian inference considers every plausible scenario that could happen, and it marginalizes the parameters over the most possible outcome. For example, in image processing, if an image appears distorted, frequentist statistical inference has no power to rationalize and make a valid sense out of it as it has no available space to work. It is what it is. On the other hand, Bayesian inference, model the problem by navigating through probabilities. Thus, it overcomes the issue generated by the distorted (noisy) image [112].

At present, there are various programming techniques to implement SR under GP principles, either heuristic or

Table 1 List of various SR approaches

Short description	Year	References
Shape-constrained	2022	[100]
PS-Tree (GP modification)	2022	[97]
DoME (deterministic)	2022	[84]
Bayesian	2022	[90]
Tensorial sparse	2022	[92]
Mixed-integer non-linear programming	2022	[86]
Probabilistic framework	2022	[88]
Functional-Hybrid model	2022	[104]
TaylorGP (GP modification)	2022	[96]
GSR (matrix based encoding scheme)	2022	[55]
Symbolic Physics Learner (MCTS)	2021	[80]
Physically constrained	2021	[99]
GP-GOMEA (GP modification)	2021	[116]
Temporal regression	2021	[91]
SymbolicGPT (probabilistic language)	2021	[87]
Large scale pre-training	2021	[81]
Pre-regression	2021	[82]
Hybrid Neural-Guided/GP	2021	[103]
Probabilistic grammars	2021	[89]
Neural Network based	2021	[109]
GP modification	2021	[94]
Bayesian Neural Network	2021	[111]
AI Feynman 2.0 (graph modularity)	2020	[108]
GP modification	2020	[18]
Multi-task SISSO	2019	[117]
AI Feynman (Neural Network)	2019	[107]
DSR (Deep Learning)	2019	[110]
Positional CGP (GP modification)	2018	[95]
IT (search space constrain)	2018	[58]
SISSO	2018	[118]
Mixed-integer non-linear programming	2017	[85]
Hybrid	2013	[102]
FFX (fast & deterministic)	2011	[73]
Nearest Neighbor Indexing	2010	[83]
Eureqa (software)	2009[a]	[113]
HeuristicLab (software)	2002[a]	[114]

[a] First launched

effective only in the implied data region, being presented either as stand-alone codes or in user-friendly platforms (e.g., Eureqa [113], HeuristicLab [114]), free or commercial. A relevant list is presented in Table 1. Finally, future studies on SR can exploit the open-source characteristics of SRBench [115] initiative, which is a promising benchmark that can provide access in different datasets, perform algorithmic comparisons and result analysis, among others.

3.4 Pros and Cons

In contrast to other non-linear regression methods, SR does not require a priori knowledge of the studied system, as it is completely data-driven [119]. Of outmost importance is the fact that SR can identify ambiguous relations in datasets and therefore provide a more profound solution [80]. There may be cases that governing equations that describe a system are partially known [120], and this is also a field where SR can apply, providing a deeper understanding. Towards generalizable AI, it provides a closed-form mathematical expression easier to incorporate at models similar to the one under investigation (e.g., finite element solver) [119].

Equally important is the ability of SR to be bound on and validate physical laws [121]. For instance, Newton's law of gravitation was somehow rediscovered [122], while another study has focused on rediscovering conservation laws [123], validated by a data-driven approach and given by a symbolic formula. However, care has to be taken when aiming on potential scientific discoveries, as oversimplified datasets and lack of evaluation metrics may lead to false results [124].

The investigation of complex and nonlinear dynamic systems demands deep understanding of the physical behavior in order to provide a reliable model [125]. Identifying differential equations via data-driven models, is a way to provide governing equations directly from data instead of physical laws, as equations in complex physical systems are scarce [126]. Nevertheless, issues may arise due to data scarcity or low fidelity, as oftentimes noise is present [80]. Noise can be, though, anticipated with Bayesian approaches. SR has also proven to be more effective than several ML models in small datasets [127] and by generating simple and physics-based relations, has an impact in its employability.

In contrast, the main drawback of SR is by no doubt the computational time needed for evaluating thousands or more equations [128]. This is one of the reasons that SR is better suited in applications where the number of input parameters is as small as possible [129]. To confront this barrier, a common strategy is to first identify the most important factors in the dataset (this can be done by other ML models, such as RF [101]) and then apply SR. However, this technique may affect the obtained results when accuracy is the main question.

From another point of view, as interpretability derives from the need to understand and trust an ML model, verify or manipulate it [130], and SR claims to be interpretable, at least compared to other black-box approaches, the degree of straightforwardness usually accounts for the size of generated expressions. Bloat is a common side effect that arises by GP, in which the results tend to suffer from a burst of complexity, while improvements on achieved fitting remain slight [131], though promising [131].

4 Application in Science and Technology

Current computational techniques in most fields of science and engineering have been able to find a balance between computational efficiency and cost, as the rise of data-driven models, along with the introduction of parallel hardware architectures, constitute a novel framework that can be exploited to obtain accurate results faster. Taking further into account the inherent interpretability of SR models, the perspective of obtaining meaningful equations is welcomed by numerous fields in theoretical and applied science. An extended review on materials science applications will be made first, as it is an interdisciplinary field that enters most science and engineering fields, through material properties prediction and novel materials discovery, such as electrical and mechanical engineering, construction applications, biophysics and energy applications, to mention a few.

4.1 Materials Science

Material science, from the sub-atomic to the macroscopic level, is currently undergoing a major shift towards full digitalization and automation and has opened new perspectives for innovation. Incorporation of databases, multi-scale computations, and experiments are integrated with the aim of reducing the time and cost of design and manufacturing of materials. AI techniques are now focused on finding new and/or predicting the properties of existing materials [25], which will make possible the discovery of novel, tailor-made materials. As materials investigation has been mainly conducted with expensive and complex experimental methods, under the particular researcher's intuition [132], and theoretical analysis is widely based on empirical relations, by leveraging big data and AI methods, a new computational paradigm emerges, to pose as a catalyst for materials development, along with exploiting available simulation techniques. To this end, data from multiple-scales of microstructures can be embedded with physics-based descriptions, to reveal physical concepts such as thermodynamics, kinetics, functional and mechanical properties.

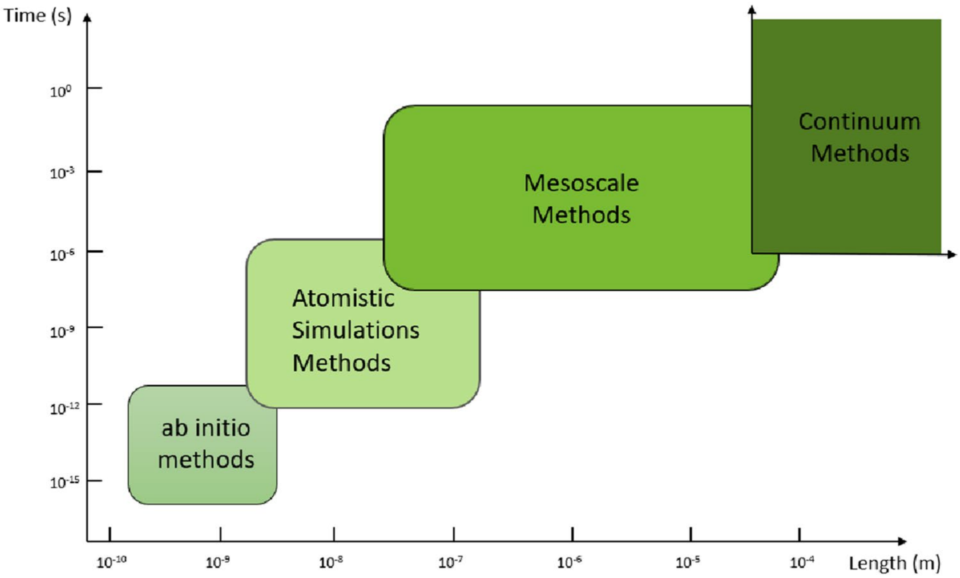


Fig. 6 Multiscale modelling

4.1.1 Multiscale Modelling

Research on materials (solids or fluids) takes place at different scales of length and time, with each scale incorporating features from the former (see Fig 6).

Starting from the atomic scale, ab-initio methods (first principles) are performed by quantum mechanics (QM) calculations in order to obtain a form that describes the energy of a system, the potential energy surface (PES). Calculations are derived directly from physical laws and do not require the incorporation of any experimental data or assumptions. However, these are based on finding solutions to the Schrödinger's equation, which may not be practical for most real-world systems. This can be partly anticipated by the incorporation of the density functional theory (DFT) [133]. Albeit capable of achieving quantum-accuracy, the usage of atomistic methods is limited in terms of the accessible computational time and simulation size. To overcome these barriers, a number of particle methods have arisen.

Particle methods include, among others, the methods of Molecular Dynamics (MD), Dissipative Particle Dynamics (DPD) and Smoothed Particle Hydrodynamics (SPH). These methods are appropriate for different size and time scales and share the same features with purely meshfree and Lagrangian nature. A particle in MD, DPD, and SPH acts as both a material point and as an approximation point, that is, the particle is regarded as a single atom or molecule in MD in atomistic scales, a small cluster of atoms or molecules in DPD in meso-scales, and a very small region in SPH in macro-scales [134].

By joining multiple techniques and methods, multiscale modeling has been developed and, during the past few years, it succeeded to integrate ML with simulations to create surrogate models [135, 136]. Research has shown that the construction of ML interatomic potentials (MLIPs) trained over ab-initio MD (AIMD) simulation results, could extend the possibilities of materials research by bridging DFT with MD and finite element (FE) simulations [137]. Furthermore, SR approaches thrive on interatomic potential by providing simple solutions that significantly decrease the risk of overfitting. Specifically, a SR model [138] generates fast and accurate many-body interatomic potentials formed by fundamental physical principles, which in turn are flexible enough to perform in multi-scale.

For example, a novel technique for bridging between micro- and macro-scale, Ref. [139], where a combination of computer simulations and ML models is suggested. Specifically, FE simulations were conducted for the generation of input data to enter the ML models, and, subsequently, ML algorithms develop a macroscopic model trained by a microscopic. The fact that traditional methods work in conjunction with ML models, provides the ability to bridge across scales and therefore provide more accurate predictions. However, it is also stated that analytical functions are appealing for the investigation of properties in micro- as well as in macro-scale [139]. Thus, SR appears as a prominent tool for multi scale investigation.

4.1.2 Properties Prediction

The development of material properties databases in a systematic way has altered materials research, as researchers opt for ML models to extract information out of them [26]. Both material properties and their relation to processing conditions, are translated to form new computational models [140]. There are cases where constitutive models express how a material responds in different conditions, which in turn produces a stress–strain relation to generate the governing laws [141]. For instance, data-driven plasticity formulas were generated in studies [142, 143], while in another approach [144], elastic solid models were constructed in similar fashion.

Moreover, several applications focus on equation generation from scratch towards the prediction of Lennard–Jones fluid diffusion [56, 145, 146] or the electrical conductivity of ionic liquids [101]. While others, have equipped SR to investigate lattice thermal conductivity [147], the critical temperature in superconductors [148] or the yield strength of polycrystalline metals by key physical quantities [149].

4.1.3 Discovery and Design

A concept familiar with material discovery, is material stability. There exist abundant element combinations one could perform towards information from the periodic table and produce a material, at least theoretically, since the additional criterion of stability, i.e., a measure to estimate if the hypothetical material is feasible, has to be taken into account. The role of SR in materials discovery and design has been evaluated [150] and it has been proven to be a suitable tool for the identification of compound representations (also known as descriptors [151]) and the creation of new that correlate with materials stability [152].

On the other hand, materials discovery and design, aim to fully exploit property prediction and produce materials with target behavior. Nevertheles, this might be a challenging task, as the targeted property could appear only in unique structures and, in addition, some properties ought to have a perfect alignment in order to achieve a high performance [153]. Therefore, it is vital to identify the parameters that govern the functionality and their dependencies, in order to optimize them [154].

A successful example of incorporating SR in such cases has been presented ref [152]. Without any prior knowledge on the problem's domain, SR has generated accurate descriptors that define MXene stability. Furthermore, another study [155] has implemented SR and generated simple and meaningful descriptors that has ultimately contributed on the discovery of new catalysts. These applications illustrate the fact that SR can produce accurate descriptors without any chemical or other knowledge on the system and eventually accelerate the discovery of novel materials.

4.2 Engineering Principles

In this section, various engineering sub-principles are to be evaluated on the applicability of SR-related methods, such as civil and construction, chemical, petroleum and natural gas engineering, mechanical and computer engineering. Main focus is the applicability of AI methods, and SR is capable to provide symbolic expressions to be used at hand and pose as a fast solution to real-life problems.

4.2.1 Construction and Building Materials

Starting from a subset of civil engineering, in hydraulics (liquid flows through pipes), the Colebrook equation for flow friction is a familiar model among engineers that is also being adopted by adjacent engineering disciplines [156]. Several models of the Colebrook equation via SR were generated in study [156], where the obtained results presented to be accurate enough. As the authors state, their approaches are only valid for a turbulent regime due to the fact that a transition from a laminar to a turbulent regime is not efficiently described by the Colebrook equation. They supplement, in their previous works, approaches were made by genetic algorithms and neural networks to model this transition unefficiently. In a subsequent study [157], simpler equations were discovered that unify laminar and turbulent hydraulic regimes and therefore diminishing the need to account for changes in flow patterns at separate laminar or turbulent flow models [157]. However, their dataset were generated by sampling through already established equations and not experimental data and therefore their application is hindered [157].

From another point of view, concrete, the main construction material for most higher-scale applications, has been also a popular subject of AI research. A number of published papers has been focused on the construction of models to estimate the seismic peak drift ratio [158, 159], the penetration depth into concrete blocks [160], the shear capacity of steel fiber-reinforced concrete beams, tracing fire response of concrete structures [161] or the seismic response through a fragility analysis [162], while others aim on the accurate description of remaining fatigue life [163, 164] or bearing-type bolted connections' shear resistance. One should note that, while the investigation of previously noted instances were conducted by modelling measurements, in several occasions [160, 163, 165] the generated equations outperformed conventional employed formulas.

4.2.2 Chemical Engineering

Drag coefficient has a crucial role in gas-solid flows, as it provides an analytical view of the hydrodynamics therein [166]. At the industrial scale, Computational Fluid Dynamics (CFD) simulations are being employed towards this investigation [167]. Moreover, CFD simulations depend on the Euler–Euler or Euler–Lagrange models [167], which both of them need to possess efficient drag correlation models in order to take into consideration gas-particle interactions [168]. Current employment of SR techniques, are found in studies about the investigation of fundamental principles in the drag coefficient [169], construction of brand new drag correlations which can be used as input to CFD models [167] or a simple drag model [166] which have proven to outperform standard formulations.

Further notations highlight the prospect of SR employment to the catalysis field in order to obtain physic-based models [79], or its incorporation into a method to obtain a whey protein fouling prediction model in plate heat exchanger, by formulating a parameter that needs to be re-adjusted when a slight change on the solution of whey protein or process conditions take place [170]. Others, have successfully identified physical relations of fluids and kinetic laws of chemical reactions [171] or generated expression in order to predict the particle size distribution during fluidization [172].

4.2.3 Petroleum and Natural Gas Engineering

In petroleum engineering, oil viscosity is of significant value, with its calculation being conducted by experimental measurements or empirical formulas at different pressure regions [173]. However, the former occasionally suffer from inadequate measurement supply while the latter generates insufficiently outcomes [173]. To overwhelm this situation, in study [173], there is presented a SR approach in which correlation models were constructed across every pressure region directly from data points, which in turn managed to outperform current models. On the other hand, a SR application about predicting the rate of penetration in drilling of hydrocarbon reservoirs [174], resulted on an expression to be overpowered by RF and ANN estimations [174].

Moreover, foam induced by a surfactant solution and nitrogen, finds room of application in tasks such as oil recovery, acid diversion and aquifer remediation, with its mobility being generally characterized in terms of pressure drop [175]. Capturing the physical behavior of the system and classifying relative variables according to their significance to steady state pressure drop, was accomplished by generating analytical expressions in study [175], by accepting no prior knowledge regarding the underlying physical behavior. Others, have focused on modelling oil production [176] or estimating the multiple fractured horizontal wells flow performance [177].

In contrast to other hydrocarbon-based materials (e.g., oil, coal), natural gas constitutes a cheaper and cleaner option [178] to meet our energy demands. Similar to petroleum engineering, estimating the viscosity is one of the top priorities in natural gas studies, as it can be utilized to efficiently synthesize models about production, transportation or gas storage systems [179]. To this end, SR studies regarding the prediction of dynamic viscosity [180] or pure and impure viscosity [179] appear most appealing. Supplementary applications, deal with models construction towards hydrate formation temperature estimation [181], estimation of equilibrium water dewpoint temperature [182] and the prediction of the gas compressibility factor [178].

4.2.4 Mechanical Engineering

Objectives of control systems could be summarized into maintaining a process, which could be affected by external parameters, and a transition from one process to another [183]. In order to do so, the control system often manages other parameters (e.g., pressure, temperature, etc.) to reach or preserve a certain status [183]. In this regard, several studies have equipped SR to generate analytic functions towards a control system design [128, 184].

Moreover, SR has been utilized to derive models capable of describing the underlying connection between alloy composition, cooling time and hardness, in welding heat-affected zone of low alloy steel [185]. Additionally, SR has been employed to estimate constitutive model parameters in an alloy research [186], or even facilitated to create expressions to search for the optimized components during machine tool design by finding the modal mass distribution matrix, which is usually hard to obtain [187].

Furthermore, SR has provided constitutive formulas of material behavior in aluminum alloys [119] or been incorporated into a mergent of techniques where SR estimated the calibration parameters of a physics-based model [188]. Calibration of model parameters that depend on processing conditions may pose a major obstacle [140]. These parameters are occasionally fitted, in order to reach an agreement with measurements; for that reason, the degree of importance of other parameters on the calibration parameters is not completely established [188]. To overcome this issue, two techniques have been proposed, the explicit and implicit [140].

In the explicit method, the calibration parameters are primarily optimized, while a formula for the prediction of the optimized values using SR, is generated next. Then, a combination of the generated expressions on calibration parameters and a physics-based constitutive model takes place, in order to create a hybrid approach. In the implicit method, no optimization of calibration parameters occurs as they undergo a tree-based GP procedure on the first steps. In addition, no extra combination of expressions similar to the explicit method are performed, as they are already combined in a multi-tree GP, where each individual has a number of trees that correspond to the number of calibration parameters. Note, that the authors recommend the implicit method for further use, while they also note that although the implicit method may be more computationally expensive, the remarkably higher accuracy cannot be ignored.

4.2.5 Computer Engineering

Computer and Information science has much to offer to SR programming, from the view of constructing and suggesting new techniques to improve the applicability of the genetic algorithms on existing physical and industrial problems. For example, a novel fault detection mechanism has been constructed with SR symbolic techniques and found to achieve better results than traditional methods, such as the support vector machine and pattern recognition neural network algorithms [189]. Moreover, another industrial application refers to enhancing models of learning behavior that present, better learning response than manual and experienced learners [190]. Techniques such as the narrowing of the search space through a semantic cluster library have given promising results [191], while a statistical-based SR algorithm has been proposed that uses statistical information to improve its performance [192].

On the other hand, SR improvement may be also achieved by decomposing the problem under investigation into several subproblems [193]. There are also cases where SR has been bound to reinforcement learning, and has been able to deal with dynamic tasks, with back-propagation capability [194] or even a dynamic process formulation [195]. Finally, since GP problems oftentimes require tons of computational time to complete, the evaluation time has been used as an estimate of model complexity and a new method is proposed to control it [196].

4.3 Other Fields

4.3.1 Physics and Astronomy

Data from astronomical observations is undoubtedly rich and AI methods are well-posed to its exploitation. For example, galaxy clusters turn out to be the most immense structures in the universe [197], as they contain several galaxies, that further include dark matter, black holes and more [198]. Moreover, they operate by mechanisms regarding the evolution and formation of those, whose details are not yet fully understood [198]. Thus, various approaches have emerged for their investigation, such as searching of expressions capable to unify properties of galaxy clusters to their masses [197], studying the galaxy-halo connection [199], modelling the assembly bias [129] and estimating the total mass of a subhalo [198].

In addition, SR has been applied in exoplanet transit spectroscopy modelling [200], as observations of planetary transits at different wavelengths are often investigated as a method to gain knowledge of the structure and composition of an exoplanet's atmosphere [200]. Further applications worth noting, include rediscovering orbital anomalies from observations of position and velocity by generating a model of dynamics [201], predicting gravitational waveform surrogates [202], reconstructing the duality parameter in an approach to predict the cosmic distance duality relation with strongly lensed gravitational wave events [203], analyzing solar activity in a solar cycle and successfully revealing underlying governing laws regarding magnetic wave generation [204].

4.3.2 Energy and the Environment

Green growth is the way towards a sustainable future. Although SR cannot be considered as the basic means of directing the world towards green transition, it can be easily implemented to contribute to a greener environment. For example, an SR-based study supports that green transition is highly likely to be embraced by developed countries, while underdeveloped or still developing countries follow a model where they choose economic growth over green [205]. The current work avoids classification of countries in developed or underdeveloped and focuses on ways to enhance the green environment's harnessing systems and fundamental components that it is accounted for.

Energy from the sun is currently covering a substantial amount of global energy demands. From a ML point of view, applications such as solar radiation prediction [206, 207] and photovoltaic power prediction [208] can open the pathway towards better energy management. Apart from renewable applications that incorporate solar radiation, wind energy is also harnessed to produce a significant amount of energy resources. Wind speed analysis [209, 210] and wind power generation efficiency [211] studies, could eventually produce an improved version of wind turbines, by exploiting novel computational ML models.

However, from an environmentalist point of view, an energy surplus should not lead on an excessive use of available energy, by continuously increasing the income of generated power to adhere to our costly way of life; damage should be also minimized. Damage minimization can occur via energy consumption modelling [90, 212, 213], energy management [214], carbon emission studies [215, 216], exhaust emission [217] and modelling of air quality [91]. All these applications can contribute towards a green energy transition and securing a healthier planet.

4.3.3 Medical Sciences

The majority of SR applications in medical science, follow a similar pattern, starting from the identification of important features between a dataset (usually measurements or patients' medical history) and followed by the establishment of suitable models towards forecasting, prognostication or successful diagnosis.

For example, in a study about Parkinson disease [218], SR was able to find important features that relate to gait changes. Similarly, in study [219], a comprehensible risk model was generated to predict survival rate of breast cancer patients. Further applications are found in analyzing measurements towards hepatocellular carcinoma diagnosis (liver cancer) [220], estimating hemoglobin and glucose levels in blood by modelling key features from fingertip videos [221], pairing patients that show similar characteristics on the way to radiotherapy dose reconstruction and therefore improve the design of radiation treatments [222] or analyzing measurements of different body areas towards human walking modelling [223].

Additionally, SR incorporation has led to the enhancement of previous estimator models, adding mathematical interpretability to previously adopted "black-box" ML models [224]. More specifically, SR has been exploited to create mathematical formulas about transformations of covariates from patients' medical records and then those formulas were used in the Cox model [225], resulting in even higher prediction accuracy compared to solely applying the Cox model. In another study about pregnancies which develop pre-eclampsia, SR has outperformed models based on logistic regression by identifying relations between important features [226].

4.3.4 Financial

The Covid-19 pandemic wreaked havoc to previous macroeconomic models, and thereupon the need to establish accurate estimates is now, more than ever, evident [227]. Macroeconomic models are often utilized as a means to guide political and financial decisions [228] and by integrating SR into those approaches, possible relations between variables might come up to light. Early studies have focused on the recognition of those interactions in large datasets that contain different observations of several economic quantities [228], while succeeding models are centered in the prediction of crude oil price [229] and economic growth forecasting by investigating expectations of agents. Agents' expectations regarding the economy's condition are high-valued for economic modelling due to the fact that they contain explicit, multi-variate information about market [230] and usually obtained via tendency surveys (business and consumer surveys) [231]. As a result, approaches are made via SR to form a link between survey data and a successful economic growth model [227, 232].

Another application of SR via GP has been in econometric modelling [233] where the conventional ”exchange equation” was reinvented. However, GP has not revealed its full prospective in the field, as being underused in macroeconomic modelling [227]. Researchers think that the polymeric generated expressions, who differ by those European Commission presents [232], will eventually escalate the situation by proving their superiority and update the currently equipped models.

5 Conclusion and Future Perspectives

Symbolic Regression has emerged as a method that bridges the gap between data, ML models and scientific theories, providing analytical equations at hand, and this can be specifically applicable in cases where only empirical and numerical approaches have been established. Barriers that may appear lie in the increased computational cost, equation complexity, and efficient programming approaches.

It is imperative that new computational tools be exploited in the future in order to reduce the computational cost of SR. As advanced NN architectures and functions are constantly being proposed for complex DL tasks, they should also be embedded in an SR framework. For example, the incorporation of Generative Adversarial Networks along with the SR algorithm, would facilitate the discovery of hidden physics in small datasets. Data pre-processing is also an open issue, since there is no standard procedure to deal with raw data, as is the case with common ML tasks. Moreover, parallel, GPU-based implementations would accelerate the SR process, reducing computational time, which, in cases of many-input datasets, is prohibitive.

The transformational impact of SR on science and technology can be very high. Notwithstanding the fact that results coming from SR investigation in various fields, both in research and applied sciences, are still preliminary, the method has shown great potential, extracting models with performance comparable to state-of-the-art empirical relations widely used in the literature.

Novel algorithms, efficient data preprocessing methods, data mining, and new function prediction, bound to physical laws and as simple as possible, should be further prospected to guide future scientific research. Following disciplines that adhere to generalizability, interpretability, and applicability, SR is the AI branch that will be on the focus of primary research in the years to come and its evolution will be towards meeting the requirements of practical applications.

We should keep in mind, however, that SR is not directed to replace well-established theoretical research and mathematical approaches. It is more likely to be bound on data science techniques and suggest an alternative method of explaining and discovering hidden patterns and behaviors in data coming from various sources. The concept of obtaining an analytical expression to describe physical phenomena and processes only by diving into a dataset is practical and can be intriguing, but, on the other hand, we believe that many things are yet to be done on the direction of ensuring that this expression is physically explainable and lie on firm basis.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11831-023-09922-z.

Funding Open access funding provided by HEAL-Link Greece. F.S. has received research support in the context of a grant received by the Center of Research Innovation and Excellence of U.Th., funded by the Special Account for Research Grants of U.Th.

Declarations

Conflict of interest D.A. and T.E.K have no financial interests to declare. On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Rydning DR-JG-J, Reinsel J, Gantz J (2018) The digitization of the world from edge to core. Framingham: International Data Corporation **16**
2. Agrawal A, Choudhary A (2016) Perspective: materials informatics and big data: realization of the “fourth paradigm'' of science in materials science. APL Mater 4(5):053208. https://doi.org/10.1063/1.4946894
3. Frydrych K, Karimi K, Pecelerowicz M, Alvarez R, Dominguez-Gutiérrez FJ, Rovaris F, Papanikolaou S (2021) Materials informatics for mechanical deformation: a review of applications and challenges. Materials. https://doi.org/10.3390/ma14195764
4. Lopez-Bezanilla A, Littlewood PB (2020) Growing field of materials informatics: databases and artificial intelligence. MRS Communications 10(1):1–10. https://doi.org/10.1557/mrc.2020.2
5. Belsky A, Hellenbrandt M, Karen VL, Luksch P (2002) New developments in the inorganic crystal structure database (ICSD): accessibility in support of materials research and design. Acta Crystallogr Sect B 58(3 Part 1):364–369. https://doi.org/10.1107/S0108768102006948
6. Kirklin S, Saal JE, Meredig B, Thompson A, Doak JW, Aykol M, Rühl S, Wolverton C (2015) The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. Npj Comput Mater 1(1):1–15. https://doi.org/10.1038/npjcompumats.2015.10
7. Allen FH (2002) The Cambridge structural database: a quarter of a million crystal structures and rising. Acta Crystallogr Sect B 58(3 Part 1):380–388. https://doi.org/10.1107/S0108768102003890
8. Curtarolo S, Setyawan W, Wang S, Xue J, Yang K, Taylor RH, Nelson LJ, Hart GLW, Sanvito S, Buongiorno-Nardelli M, Mingo N, Levy O (2012) Aflowlib.org: a distributed materials properties repository from high-throughput ab initio calculations. Comput Mater Sci 58:227–235. https://doi.org/10.1016/j.commatsci.2012.02.002
9. Li S, Liu Y, Chen D, Jiang Y, Nie Z, Pan F (2022) Encoding the atomic structure for machine learning in materials science. Wiley Interdiscip Rev: Comput Mol Sci 12(1):1558. https://doi.org/10.1002/wcms.1558
10. Zhou T, Song Z, Sundmacher K (2019) Big data creates new opportunities for materials research: a review on methods and applications of machine learning for materials design. Engineering 5(6):1017–1026. https://doi.org/10.1016/j.eng.2019.02.011
11. Patrick H (1992) Winston. Artificial Intelligence. Addison-Wesley (now 3rd Edition)
12. Xu X, Aggarwal D, Shankar K (2022) Instantaneous property prediction and inverse design of plasmonic nanostructures using machine learning: current applications and future directions. Nanomaterials. https://doi.org/10.3390/nano12040633
13. Frank M, Drikakis D, Charissis V (2020) Machine-learning methods for computational science and engineering. Computation. https://doi.org/10.3390/computation8010015
14. Chowdhury MA, Hossain N, Ahmed Shuvho MB, Fotouhi M, Islam MS, Ali MR, Kashem MA (2021) Recent machine learning guided material research—a review. Comput Condens Matter 29:00597. https://doi.org/10.1016/j.cocom.2021.e00597
15. Guo S, Agarwal M, Cooper C, Tian Q, Gao RX, Guo W, Guo YB (2022) Machine learning for metal additive manufacturing: towards a physics-informed data-driven paradigm. J Manufact Syst 62:145–163. https://doi.org/10.1016/j.jmsy.2021.11.003
16. Meng L, McWilliams B, Jarosinski W, Park H-Y, Jung Y-G, Lee J, Zhang J (2020) Machine learning in additive manufacturing: a review. Jom 72(6):2363–2377. https://doi.org/10.1007/s11837-020-04155-y
17. Qi X, Chen G, Li Y, Cheng X, Li C (2019) Applying neural-network-based machine learning to additive manufacturing: Current applications, challenges, and future perspectives. Engineering 5(4):721–729. https://doi.org/10.1016/j.eng.2019.04.012
18. Wang C, Tan XP, Tor SB, Lim CS (2020) Machine learning in additive manufacturing: state-of-the-art and perspectives. Addit Manuf 36–101538. https://doi.org/10.1016/j.addma.2020.101538
19. Liu Y, Zhao T, Ju W, Shi S (2017) Materials discovery and design using machine learning. J Materiomics 3(3):159–177. https://doi.org/10.1016/j.jmat.2017.08.002
20. Pilania G (2021) Machine learning in materials science: from explainable predictions to autonomous design. Comput Mater Sci 193:110360. https://doi.org/10.1016/j.commatsci.2021.110360
21. Dimiduk DM, Holm EA, Niezgoda SR (2018) Perspectives on the impact of machine learning, deep learning, and artificial intelligence on materials, processes, and structures engineering. Integr Mater Manuf Innov 7(3):157–172. https://doi.org/10.1007/s40192-018-0117-8
22. Wei J, Chu X, Sun X-Y, Xu K, Deng H-X, Chen J, Wei Z, Lei M (2019) Machine learning in materials science. InfoMat 1(3):338–358. https://doi.org/10.1002/inf2.12028

23. Guo K, Yang Z, Yu C-H, Buehler MJ (2021) Artificial intelligence and machine learning in design of mechanical materials. Mater. Horiz. 8:1153–1172. https://doi.org/10.1039/D0MH01451F
24. Sajid S, Haleem A, Bahl S, Javaid M, Goyal T, Mittal M (2021) Data science applications for predictive maintenance and materials science in context to industry 4.0. Mater Today 45, 4898–4905. https://doi.org/10.1016/j.matpr.2021.01.357. Second International Conference on Aspects of Materials Science and Engineering (ICAMSE 2021)
25. Morgan D, Jacobs R (2020) Opportunities and challenges for machine learning in materials science. Annu Rev Mater Res 50(1):71–103. https://doi.org/10.1146/annurev-matsci-070218-010015
26. Hart GL, Mueller T, Toher C, Curtarolo S (2021) Machine learning for alloys. Nat Rev Mater 6(8):730–755. https://doi.org/10.1038/s41578-021-00340-w
27. DeRousseau MA, Laftchiev E, Kasprzyk JR, Rajagopalan B, Srubar WV (2019) A comparison of machine learning methods for predicting the compressive strength of field-placed concrete. Constr Build Mater 228:116661. https://doi.org/10.1016/j.conbuildmat.2019.08.042
28. Nguyen H, Vu T, Vo TP, Thai H-T (2021) Efficient machine learning models for prediction of concrete strengths. Constr Build Mater 266:120950. https://doi.org/10.1016/j.conbuildmat.2020.120950
29. Kang Y, Li L, Li B (2021) Recent progress on discovery and properties prediction of energy materials: simple machine learning meets complex quantum chemistry. J Energy Chem 54:72–88. https://doi.org/10.1016/j.jechem.2020.05.044
30. Shalev-Shwartz S, Shammah S, Shashua A (2016) Safe, multi-agent. reinforcement learning for autonomous driving arXiv. https://doi.org/10.48550/ARXIV.1610.03295
31. Bolte J-A, Bar A, Lipinski D, Fingscheidt T (2019) Towards corner case detection for autonomous driving. In: 2019 IEEE Intelligent Vehicles Symposium (IV), pp. 438–445. https://doi.org/10.1109/IVS.2019.8813817
32. Okuyama T, Gonsalves T, Upadhay J (2018) Autonomous driving system based on deep q learnig. In: 2018 International Conference on Intelligent Autonomous Systems (ICoIAS), pp. 201–205. https://doi.org/10.1109/ICoIAS.2018.8494053
33. Zhong S, Yap BK, Zhong Z, Ying L (2022) Review on y6-based semiconductor materials and their future development via machine learning. Crystals. https://doi.org/10.3390/cryst12020168
34. Zhang L, He M, Shao S (2020) Machine learning for halide perovskite materials. Nano Energy 78:10538. https://doi.org/10.1016/j.nanoen.2020.105380
35. Li F, Peng X, Wang Z, Zhou Y, Wu Y, Jiang M, Xu M (2019) Machine learning (ml)-assisted design and fabrication for solar cells. Energy Environ Mater 2(4):280–291. https://doi.org/10.1002/eem2.12049
36. Mahmood A, Wang J-L (2021) Machine learning for high performance organic solar cells: current scenario and future prospects. Energy Environ Sci 14(1):90–105. https://doi.org/10.1039/D0EE02838J
37. Li J, Pradhan B, Gaur S, Thomas J (2019) Predictions and strategies learned from machine learning to develop high-performing perovskite solar cells. Adv Energy Mater 9(46):1901891. https://doi.org/10.1002/aenm.201901891
38. Padula D, Simpson JD, Troisi A (2019) Combining electronic and structural features in machine learning models to predict organic solar cells properties. Mater Horiz 6(2):343–349. https://doi.org/10.1039/C8MH01135D
39. Sahu H, Ma H (2019) Unraveling correlations between molecular properties and device parameters of organic solar cells using machine learning. J Phys Chem Lett 10(22):7277–7284. https://doi.org/10.1021/acs.jpclett.9b02772
40. Artrith N, Butler KT, Coudert F-X, Han S, Isayev O, Jain A, Walsh A (2021) Best practices in machine learning for chemistry. Nat Chem 13(6):505–508. https://doi.org/10.1038/s41557-021-00716-z
41. Janet JP, Liu F, Nandy A, Duan C, Yang T, Lin S, Kulik HJ (2019) Designing in the face of uncertainty: exploiting electronic structure and machine learning models for discovery in inorganic chemistry. Inorg Chem 58(16):10592–10606. https://doi.org/10.1021/acs.inorgchem.9b00109
42. Townsend J, Micucci CP, Hymel JH, Maroulas V, Vogiatzis KD (2020) Representation of molecular structures with persistent homology for machine learning applications in chemistry. Nat Commun 11(1):1–9. https://doi.org/10.1038/s41467-020-17035-5
43. R RM, Jagan A, Pavithran L, Shrivastava A, Selvaraj SK (2021) Intelligent welding by using machine learning techniques. Mater Today 46, 7402–7410. https://doi.org/10.1016/j.matpr.2020.12.1149.3rd International Conference on Materials, Manufacturing and Modelling
44. Quej VH, Almorox J, Arnaldo JA, Saito L (2017) Anfis, svm and ann soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment. J Atmos Solar-Terr Phys 155:62–70. https://doi.org/10.1016/j.jastp.2017.02.002
45. Ramli MAM, Twaha S, Al-Turki YA (2015) Investigating the performance of support vector machine and artificial neural networks in predicting solar radiation on a tilted surface: Saudi Arabia case study. Energy Convers Manag 105:442–452. https://doi.org/10.1016/j.enconman.2015.07.083
46. Cornejo-Bueno L, Casanova-Mateo C, Sanz-Justo J, Salcedo-Sanz S (2019) Machine learning regressors for solar radiation estimation from satellite data. Solar Energy 183:768–775. https://doi.org/10.1016/j.solener.2019.03.079
47. Feng Y, Gong D, Zhang Q, Jiang S, Zhao L, Cui N (2019) Evaluation of temperature-based machine learning and empirical models for predicting daily global solar radiation. Energy Convers Manag 198:111780. https://doi.org/10.1016/j.enconman.2019.111780
48. Zhou Y, Liu Y, Wang D, Liu X, Wang Y (2021) A review on global solar radiation prediction with machine learning models in a comprehensive perspective. Energy Convers Manag 235:113960. https://doi.org/10.1016/j.enconman.2021.113960
49. Narvaez G, Giraldo LF, Bressan M, Pantoja A (2021) Machine learning for site-adaptation and solar radiation forecasting. Renew Energy 167:333–342. https://doi.org/10.1016/j.renene.2020.11.089
50. Fan J, Wu L, Zhang F, Cai H, Zeng W, Wang X, Zou H (2019) Empirical and machine learning models for predicting daily global solar radiation from sunshine duration: a review and case study in China. Renew Sustain Energy Rev 100:186–212. https://doi.org/10.1016/j.rser.2018.10.018
51. Ümit Ağbulut, Gürel AE, Biçen Y (2021) Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. Renew Sustain Energy Rev 135:110114. https://doi.org/10.1016/j.rser.2020.110114
52. Cai S, Mao Z, Wang Z, Yin M, Karniadakis GE (2022) Physics-informed neural networks (pinns) for fluid mechanics: a review. Acta Mech Sin. https://doi.org/10.1007/s10409-021-01148-1
53. Raissi M, Perdikaris P, Karniadakis GE (2017) Machine learning of linear differential equations using gaussian processes. J Comput Phys 348:683–693. https://doi.org/10.1016/j.jcp.2017.07.050
54. Yang L, Meng X, Karniadakis GE (2021) B-pinns: Bayesian physics-informed neural networks for forward and inverse pde

problems with noisy data. J Comput Phys 425:109913. https://doi.org/10.1016/j.jcp.2020.109913

55. Tohme T, Liu D, Youcef-Toumi K (2022) GSR: a generalized symbolic regression approach. arXiv. doi:1048550/ARXIV.2205.15569
56. Papastamatiou K, Sofos F, Karakasidis TE (2022) Machine learning symbolic equations for diffusion with physics-based descriptions. AIP Adv 12(2):025004. https://doi.org/10.1063/5.0082147
57. Asadzadeh MZ, Gänser H-P, Mücke M (2021) Symbolic regression based hybrid semiparametric modelling of processes: an example case of a bending process. Appl Eng Sci 6:100049. https://doi.org/10.1016/j.apples.2021.100049
58. de Olivetti França, F (2018) A greedy search tree heuristic for symbolic regression. Inf Sci 442–443:18–32. https://doi.org/10.1016/j.ins.2018.02.040
59. Raschka S (2015) Python machine learning. Packt publishing Ltd, Birmingham, UK.
60. Chatzilygeroudis K, Hatzilygeroudis I, Perikos I (2021) Machine learning basics. In: Intelligent computing for interactive system design: Statistics, digital signal processing, and machine learning in practice (1st ed.) Association for Computing Machinery, New York, NY, USA, pp. 143–193. https://doi.org/10.1145/3447404.3447414
61. Alexiadis A (2019) Deep multiphysics: coupling discrete multiphysics with machine learning to attain self-learning in-silico models replicating human physiology. Artif Intell Med 98:27–34. https://doi.org/10.1016/j.artmed.2019.06.005
62. Hu F, Hao Q (2012) Intelligent sensor networks: the integration of sensor networks, signal processing and machine learning. Taylor & Francis. https://doi.org/10.1201/b14300
63. Sofos F, Stavrogiannis C, Exarchou-Kouveli KK, Akabua D, Charilas G, Karakasidis TE (2022) Current trends in fluid research in the era of artificial intelligence: a review. Fluids. https://doi.org/10.3390/fluids7030116
64. Nazemi E, Dinca M, Movafeghi A, Rokrok B, Choopan Dastjerdi MH (2019) Estimation of volumetric water content during imbibition in porous building material using real time neutron radiography and artificial neural network. Nucl Instrum Methods Phys Res Sect A 940:344–350. https://doi.org/10.1016/j.nima.2019.06.052
65. Ben Chaabene W, Flah M, Nehdi ML (2020) Machine learning prediction of mechanical properties of concrete: critical review. Constr Build Mater 260:119889. https://doi.org/10.1016/j.conbuildmat.2020.119889
66. Poulinakis K, Drikakis D, Kokkinakis IW, Spottswood SM (2023) Machine-learning methods on noisy and sparse data. Mathematics. https://doi.org/10.3390/math11010236
67. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJ, Bottou L, Weinberger KQ (eds.) Advances in neural information processing systems, vol. 25. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
68. Farabet C, Couprie C, Najman L, LeCun Y (2013) Learning hierarchical features for scene labeling. IEEE Trans Pattern Anal Mach Intell 35(8):1915–1929. https://doi.org/10.1109/TPAMI.2012.231
69. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
70. Jiang Y-G, Wu Z, Wang J, Xue X, Chang S-F (2018) Exploiting feature and class relationships in video categorization with regularized deep neural networks. IEEE Trans Pattern Anal Mach Intell 40(2):352–364. https://doi.org/10.1109/TPAMI.2017.2670560
71. Hinton G, Deng L, Yu D, Dahl GE, Mohamed A-R, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Processing Mag 29(6):82–97. https://doi.org/10.1109/MSP.2012.2205597
72. Leung MKK, Xiong HY, Lee LJ, Frey BJ (2014) Deep learning of the tissue-regulated splicing code. Bioinformatics 30(12):121–129. https://doi.org/10.1093/bioinformatics/btu277
73. McConaghy T (2011). In: Riolo R, Vladislavleva E, Moore JH (eds) FFX: fast, scalable, deterministic symbolic regression technology. Springer, New York, NY, pp 235–260
74. Koza JR (1994) Genetic programming as a means for programming computers by natural selection. Stat Comput 4(2):87–112. https://doi.org/10.1007/BF00175355
75. Back T (1996) Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms. Oxford University Press, Oxford
76. Eiben AE, Smith JE et al (2003) Introduction to evolutionary computing. Springer, New York. https://doi.org/10.1007/978-3-662-44874-8
77. Quade, M., Gout, J., Abel, M.: Glyph: Symbolic Regression Tools. arXiv (2018). https://doi.org/10.48550/ARXIV.1803.06226
78. 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, vol. 1, pp. 77–80 (2010). Citeseer
79. Liu C-Y, Senftle TP (2022) Finding physical insights in catalysis with machine learning. Curr Opin Chem Eng 37:100832. https://doi.org/10.1016/j.coche.2022.100832
80. Gilpin W (2021). Chaos as an interpretable benchmark for forecasting and data-driven modelling. https://doi.org/10.48550/ARXIV.2110.05266
81. Biggio, L., Bendinelli, T., Neitz, A., Lucchi, A., Parascandolo, G.: Neural symbolic regression that scales. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 936–945. PMLR (2021). https://proceedings.mlr.press/v139/biggio21a.html
82. Udrescu S-M, Tegmark M (2021) Symbolic pregression: discovering physical laws from distorted video. Phys Rev E 103:043307. https://doi.org/10.1103/PhysRevE.103.043307
83. McRee, R.K.: Symbolic regression using nearest neighbor indexing. In: Proceedings of the 12th Annual Conference Companion on Genetic and Evolutionary Computation, pp. 1983–1990. Association for Computing Machinery, New York, NY, USA (2010). https://doi.org/10.1145/1830761.1830841
84. Rivero D, Fernandez-Blanco E, Pazos A (2022) Dome: a deterministic technique for equation development and symbolic regression. Exp Syst Appl 198:116712. https://doi.org/10.1016/j.eswa.2022.116712
85. Austel V, Dash S, Gunluk O, Horesh L, Liberti L, Nannicini G, Schieber B (2017). Globally Optimal Symbolic Regression arXiv. https://doi.org/10.48550/arXiv.1710.10720
86. Engle MR, Sahinidis NV (2022) Deterministic symbolic regression with derivative information: General methodology and application to equations of state. AIChE Journal 68(6):17457. https://doi.org/10.1002/aic.17457
87. Valipour, M., You, B., Panju, M., Ghodsi, A.: SymbolicGPT: a generative transformer model for symbolic regression. arXiv (2021). https://doi.org/10.48550/ARXIV.2106.14131

88. Gong C, Bryan J, Furcoiu A, Su Q, Grobe R (2022) Evolutionary symbolic regression from a probabilistic perspective. SN Comput Sci 3(3):1–15. https://doi.org/10.1007/s42979-022-01094-0
89. Brence J, Todorovski L, Džeroski S (2021) Probabilistic grammars for equation discovery. Knowl-Based Syst 224:107077. https://doi.org/10.1016/j.knosys.2021.107077
90. Vázquez D, Guimerá R, Sales-Pardo M, Guillén-Gosálbez G (2022) Automatic modeling of socioeconomic drivers of energy consumption and pollution using bayesian symbolic regression. Sustain Prod Consum 30:596–607. https://doi.org/10.1016/j.spc.2021.12.025
91. Lucena-Sánchez E, Sciavicco G, Stan IE (2021) Feature and language selection in temporal symbolic regression for interpretable air quality modelling. Algorithms. https://doi.org/10.3390/a14030076
92. Wang C, Zhang Y, Wen C, Yang M, Lookman T, Su Y, Zhang T-Y (2022) Symbolic regression in materials science via dimension-synchronous-computation. J Mater Sci Technol 122:77–83. https://doi.org/10.1016/j.jmst.2021.12.052
93. Virgolin, M., Medvet, E., Alderliesten, T., Bosman, P.A.N.: Less is More: A call to focus on simpler models in genetic programming for interpretable machine learning. arXiv (2022). https://doi.org/10.48550/ARXIV.2204.02046
94. Kubalík J, Derner E, Babuška R (2021) Multi-objective symbolic regression for physics-aware dynamic modeling. Exp Syst Appl 182:115210. https://doi.org/10.1016/j.eswa.2021.115210
95. Wilson D, Miller JF, Cussat-Blanc S, Luga H (2018). Positional Cartesian genetic programming arXiv. https://doi.org/10.48550/ARXIV.1810.04119
96. He, B., Lu, Q., Yang, Q., Luo, J., Wang, Z.: Taylor Genetic Programming for Symbolic Regression. arXiv (2022). https://doi.org/10.48550/ARXIV.2205.09751
97. Zhang H, Zhou A, Qian H, Zhang H (2022) Ps-tree: a piecewise symbolic regression tree. Swarm Evolut Comput 71:101061. https://doi.org/10.1016/j.swevo.2022.101061
98. Virgolin, M., Alderliesten, T., Witteveen, C., Bosman, P.A.N.: Scalable genetic programming by gene-pool optimal mixing and input-space entropy-based building-block learning. In: Proceedings of the Genetic and Evolutionary Computation Conference. GECCO ’17, pp. 1041–1048. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3071178.3071287
99. Reinbold PA, Kageorge LM, Schatz MF, Grigoriev RO (2021) Robust learning from noisy, incomplete, high-dimensional experimental data via physically constrained symbolic regression. Nat Commun 12(1):1–8. https://doi.org/10.1038/s41467-021-23479-0
100. Kronberger G, de Franca FO, Burlacu B, Haider C, Kommenda M (2022) Shape-constrained symbolic regression-improving extrapolation with prior knowledge. Evolut Comput 30(1):75–98. https://doi.org/10.1162/evco_a_00294
101. Karakasidis TE, Sofos F, Tsonos C (2022) The electrical conductivity of ionic liquids: numerical and analytical machine learning approaches. Fluids. https://doi.org/10.3390/fluids7100321
102. Icke, I., Bongard, J.C.: Improving genetic programming based symbolic regression using deterministic machine learning. In: 2013 IEEE Congress on Evolutionary Computation, pp. 1763–1770 (2013). https://doi.org/10.1109/CEC.2013.6557774
103. Mundhenk, T.N., Landajuela, M., Glatt, R., Santiago, C.P., Faissol, D.M., Petersen, B.K.: Symbolic regression via neural-guided genetic programming population seeding. arXiv (2021). https://doi.org/10.48550/ARXIV.2111.00053
104. Narayanan H, Cruz Bournazou MN, Guillén Gosálbez G, Butté A (2022) Functional-hybrid modeling through automated adaptive symbolic regression for interpretable mathematical expressions. Chem Eng J 430:133032. https://doi.org/10.1016/j.cej.2021.133032
105. Rad, H.I., Feng, J., Iba, H.: GP-RVM: Genetic Programing-based Symbolic Regression Using Relevance Vector Machine. arXiv (2018). https://doi.org/10.48550/ARXIV.1806.02502
106. de Veloso Melo V, Banzhaf W (2018) Automatic feature engineering for regression models with machine learning: an evolutionary computation and statistics hybrid. Inf Sci 430–431:287–313. https://doi.org/10.1016/j.ins.2017.11.041
107. Udrescu S-M, Tegmark M (2019). Ai feynman: a physics-inspired method for symbolic regression. https://doi.org/10.48550/ARXIV.1905.11481
108. Udrescu, S.-M., Tan, A., Feng, J., Neto, O., Wu, T., Tegmark, M.: Ai feynman 2.0: pareto-optimal symbolic regression exploiting graph modularity. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 4860–4871. Curran Associates, Inc., (2020). https://proceedings.neurips.cc/paper/2020/file/33a854e247155d590883b93bca53848a-Paper.pdf
109. Kim S, Lu PY, Mukherjee S, Gilbert M, Jing L, Čeperić V, Soljačić M (2021) Integration of neural network-based symbolic regression in deep learning for scientific discovery. IEEE Trans Neural Netw Learn Syst 32(9):4166–4177. https://doi.org/10.1109/TNNLS.2020.3017010
110. Petersen BK, Landajuela M, Mundhenk TN, Santiago CP, Kim SK, Kim JT (2019). Deep symbolic regression: recovering mathematical expressions from data via risk-seeking policy gradients. https://doi.org/10.48550/ARXIV.1912.04871
111. Cranmer M, Tamayo D, Rein H, Battaglia P, Hadden S, Armitage PJ, Ho S, Spergel DN (2021) A bayesian neural network predicts the dissolution of compact planetary systems. Proc Natl Acad Sci 118(40):2026053118. https://doi.org/10.1073/pnas.2026053118
112. McElreath R (2020) Statistical rethinking: A bayesian course with examples in R and STAN. Chapman and Hall, CRC texts in statistical science. https://doi.org/10.1201/9780429029608
113. Dubčáková R (2011) Eureqa: software review. Springer, New York. https://doi.org/10.1007/s10710-010-9124-z
114. Wagner S, Kronberger G, Beham A, Kommenda M, Scheibenpflug A, Pitzer E, Vonolfen S, Kofler M, Winkler S, Dorfer V, Affenzeller M (2014). In: Klempous R, Nikodem J, Jacak W, Chaczko Z (eds) Architecture and design of the HeuristicLab optimization environment. Springer, Heidelberg, pp 197–261. https://doi.org/10.1007/978-3-319-01436-4_10
115. La Cava, W., Orzechowski, P., Burlacu, B., de França, F.O., Virgolin, M., Jin, Y., Kommenda, M., Moore, J.H.: Contemporary Symbolic Regression Methods and their Relative Performance. arXiv (2021). https://doi.org/10.48550/ARXIV.2107.14351
116. Virgolin M, Alderliesten T, Witteveen C, Bosman PAN (2021) Improving model-based genetic programming for symbolic regression of small expressions. Evolut Comput 29(2):211–237. https://doi.org/10.1162/evco_a_00278
117. Ouyang R, Ahmetcik E, Carbogno C, Scheffler M, Ghiringhelli LM (2019) Simultaneous learning of several materials properties from incomplete databases with multi-task SISSO. J Phys: Mater 2(2):024002. https://doi.org/10.1088/2515-7639/ab077b
118. Ouyang R, Curtarolo S, Ahmetcik E, Scheffler M, Ghiringhelli LM (2018) Sisso: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. Phys Rev Mater 2:083802. https://doi.org/10.1103/PhysRevMaterials.2.083802
119. Kabliman E, Kolody AH, Kommenda M, Kronberger G (2019) Prediction of stress-strain curves for aluminium alloys using symbolic regression. AIP Conf Proc 2113(1):180009. https://doi.org/10.1063/1.5112747
120. Vaddireddy H, Rasheed A, Staples AE, San O (2020) Feature engineering and symbolic regression methods for detecting

hidden physics from sparse sensor observation data. Phys Fluids 32(1):015113. https://doi.org/10.1063/1.5136351

121. Schmidt M, Lipson H (2009) Distilling free-form natural laws from experimental data. Science 324(5923):81–85. https://doi.org/10.1126/science.1165893
122. Lemos, P., Jeffrey, N., Cranmer, M., Ho, S., Battaglia, P.: Rediscovering orbital mechanics with machine learning. arXiv (2022). https://doi.org/10.48550/ARXIV.2202.02306
123. Liu Z, Tegmark M (2021) Machine learning conservation laws from trajectories. Phys Rev Lett 126:180604. https://doi.org/10.1103/PhysRevLett.126.180604
124. Matsubara, Y., Chiba, N., Igarashi, R., Taniai, T., Ushiku, Y.: Rethinking symbolic regression datasets and benchmarks for scientific discovery. arXiv (2022). https://doi.org/10.48550/ARXIV.2206.10540
125. Haider, C., de França, F.O., Burlacu, B., Kronberger, G.: Using shape constraints for improving symbolic regression models. arXiv (2021). https://doi.org/10.48550/ARXIV.2107.09458
126. Cao W, Zhang W (2022) Data-driven and physical-based identification of partial differential equations for multivariable system. Theor Appl Mech Lett 12(2):100334. https://doi.org/10.1016/j.taml.2022.100334
127. Wilstrup, C., Kasak, J.: Symbolic regression outperforms other models for small data sets. arXiv (2021). https://doi.org/10.48550/ARXIV.2103.15147
128. Danai K, La Cava WG (2021) Controller design by symbolic regression. Mechan Syst Signal Process 151:107348. https://doi.org/10.1016/j.ymssp.2020.107348
129. Wadekar, D., Villaescusa-Navarro, F., Ho, S., Perreault-Levasseur, L.: Modeling assembly bias with machine learning and symbolic regression. arXiv (2020). https://doi.org/10.48550/ARXIV.2012.00111
130. Aldeia GSI, de França FO (2022) Interpretability in symbolic regression: a benchmark of explanatory methods using the Feynman data set. Genetic Program Evol Mach. https://doi.org/10.1007/s10710-022-09435-x
131. Bomarito GF, Leser PE, Strauss NCM, Garbrecht KM, Hochhalter JD (2022) Bayesian model selection for reducing bloat and overfitting in genetic programming for symbolic regression. Association for Computing Machinery, New York, NY, USA, pp 526–529. https://doi.org/10.1145/3520304.3528899
132. Dunn A, Wang Q, Ganose A, Dopp D, Jain A (2020) Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. Npj Comput Mater 6(1):1–10. https://doi.org/10.1038/s41524-020-00406-3
133. Behler J, Parrinello M (2007) Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys Rev Lett 98:146401. https://doi.org/10.1103/PhysRevLett.98.146401
134. Ye T, Pan D, Huang C, Liu M (2019) Smoothed particle hydrodynamics (sph) for complex fluid flows: recent developments in methodology and applications. Phys Fluids 31(1):011301. https://doi.org/10.1063/1.5068697
135. Mishin Y (2021) Machine-learning interatomic potentials for materials science. Acta Mater 214:116980. https://doi.org/10.1016/j.actamat.2021.116980
136. Yang Y, Zhao L, Han C-X, Ding X-D, Lookman T, Sun J, Zong H-X (2021) Taking materials dynamics to new extremes using machine learning interatomic potentials. J Mater Inform 1(2):10. https://doi.org/10.20517/jmi.2021.001
137. Stephenson D, Kermode JR, Lockerby DA (2018) Accelerating multiscale modelling of fluids with on-the-fly Gaussian process regression. Microfluidics and Nanofluidics 22(12):1–12. https://doi.org/10.1007/s10404-018-2164-z
138. Hernandez A, Balasubramanian A, Yuan F, Mason SA, Mueller T (2019) Fast, accurate, and transferable many-body interatomic potentials by symbolic regression. Npj Comput Mater 5(1):1–11. https://doi.org/10.1038/s41524-019-0249-1
139. Reimann D, Nidadavolu K, ul Hassan H, Vajragupta N, Glasmachers T, Junker P, Hartmaier A (2019) Modeling macroscopic material behavior with machine learning algorithms trained by micromechanical simulations. Front Mater. https://doi.org/10.3389/fmats.2019.00181
140. Kronberger G, Kabliman E, Kronsteiner J, Kommenda M (2022) Extending a physics-based constitutive model using genetic programming. Appl Eng Sci 9:100080. https://doi.org/10.1016/j.apples.2021.100080
141. Zhang X, Chen Z, Liu Y (2017) Chapter 6—constitutive models. In: Zhang X, Chen Z, Liu Y (eds) The material point method. Academic Press, Oxford, pp 175–219. https://doi.org/10.1016/B978-0-12-407716-4.00006-5
142. Bomarito GF, Townsend TS, Stewart KM, Esham KV, Emery JM, Hochhalter JD (2021) Development of interpretable, data-driven plasticity models with symbolic regression. Comput Struct 252:106557. https://doi.org/10.1016/j.compstruc.2021.106557
143. Versino D, Tonda A, Bronkhorst CA (2017) Data driven modeling of plastic deformation. Comput Methods Appl Mech Eng 318:981–1004. https://doi.org/10.1016/j.cma.2017.02.016
144. Wang M, Chen C, Liu W (2022) Establish algebraic data-driven constitutive models for elastic solids with a tensorial sparse symbolic regression method and a hybrid feature selection technique. J Mech Phys Solids 159:104742. https://doi.org/10.1016/j.jmps.2021.104742
145. Sofos F, Charakopoulos A, Papastamatiou K, Karakasidis TE (2022) A combined clustering/symbolic regression framework for fluid property prediction. Phys Fluids 34(6):062004. https://doi.org/10.1063/5.0096669
146. Alam TM, Allers JP, Leverant CJ, Harvey JA (2022) Symbolic regression development of empirical equations for diffusion in Lennard-Jones fluids. J Chem Phys 15(1):014503. https://doi.org/10.1063/5.0093658
147. Loftis C, Yuan K, Zhao Y, Hu M, Hu J (2021) Lattice thermal conductivity prediction using symbolic regression and machine learning. J Phys Chem A 125(1):435–450. https://doi.org/10.1021/acs.jpca.0c08103
148. Xie S, Quan Y, Hire A, Deng B, DeStefano J, Salinas I, Shah U, Fanfarillo L, Lim J, Kim J et al (2022) Machine learning of superconducting critical temperature from Eliashberg theory. Npj Comput Mater 8(1):1–8. https://doi.org/10.1038/s41524-021-00666-7
149. Jiang L, Fu H, Zhang H, Xie J (2022) Physical mechanism interpretation of polycrystalline metals' yield strength via a data-driven method: a novel Hall-Petch relationship. Acta Mater 231:117868. https://doi.org/10.1016/j.actamat.2022.117868
150. Xiong J, Zhang T, Shi S (2020) Machine learning of mechanical properties of steels. Sci China Technol Sci 63(7):1247–1255. https://doi.org/10.1007/s11431-020-1599-5
151. Seko A, Togo A, Tanaka, I.. (2018). In: Tanaka I (ed) Descriptors for machine learning of materials data. Springer, Singapore, pp 3–23. https://doi.org/10.1007/978-981-10-7617-6_1
152. He M, Zhang L (2021) Machine learning and symbolic regression investigation on stability of mxene materials. Comput Mater Sci 196:110578. https://doi.org/10.1016/j.commatsci.2021.110578
153. Hautier G (2019) Finding the needle in the haystack: materials discovery and design through computational ab initio high-throughput screening. Comput Mater Sci 163:108–116. https://doi.org/10.1016/j.commatsci.2019.02.040

154. Wang Y, Wagner N, Rondinelli JM (2019) Symbolic regression in materials science. MRS Commun 9(3):793–805. https://doi.org/10.1557/mrc.2019.85
155. Weng B, Song Z, Zhu R, Yan Q, Sun Q, Grice CG, Yan Y, Yin W-J (2020) Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts. Nat Commun 11(1):1–8. https://doi.org/10.1038/s41467-020-17263-9
156. Praks P, Brkić D (2018) Symbolic regression-based genetic approximations of the colebrook equation for flow friction. Water. https://doi.org/10.3390/w10091175
157. Milošević M, Brkić D, Praks P, Litričin D, Stajić Z (2022) Hydraulic losses in systems of conduits with flow from laminar to fully turbulent: a new symbolic regression formulation. Axioms. https://doi.org/10.3390/axioms11050198
158. Hamidia M, Ganjizadeh A (2022) Post-earthquake damage evaluation of non-ductile rc moment frames using surface crack patterns. Struct Control Health Monitor. https://doi.org/10.1002/stc.3024
159. Mansourdehghan S, Dolatshahi KM, Asjodi AH (2022) Data-driven damage assessment of reinforced concrete shear walls using visual features of damage. J Build Eng 53:104509. https://doi.org/10.1016/j.jobe.2022.104509
160. Imran Latif QBA, Memon ZA, Mahmood Z, Qureshi MU, Milad A (2022) A machine learning model for the prediction of concrete penetration by the ogive nose rigid projectile. Appl Sci. https://doi.org/10.3390/app12042040
161. Naser MZ (2019) Heuristic machine cognition to predict fire-induced spalling and fire resistance of concrete structures. Autom Constr 106:102916. https://doi.org/10.1016/j.autcon.2019.102916
162. Rezaei H, Zarfam P, Golafshani EM, Amiri GG (2022) Seismic fragility analysis of rc box-girder bridges based on symbolic regression method. Structures 38:306–322. https://doi.org/10.1016/j.istruc.2021.12.058
163. Gan L, Wu H, Zhong Z (2022) Integration of symbolic regression and domain knowledge for interpretable modeling of remaining fatigue life under multistep loading. Int J Fatigue 161:106889. https://doi.org/10.1016/j.ijfatigue.2022.106889
164. Ren J, Zhang L, Zhao H, Zhao Z, Wang S (2022) Determination of the fatigue equation for the cement-stabilized cold recycled mixtures with road construction waste materials based on data-driven. Int J Fatigue 158:106765. https://doi.org/10.1016/j.ijfatigue.2022.106765
165. Ben Chaabene W, Nehdi ML (2021) Genetic programming based symbolic regression for shear capacity prediction of sfrc beams. Constr Build Mater 280:122523. https://doi.org/10.1016/j.conbuildmat.2021.122523
166. Sonolikar RR, Patil MP, Mankar RB, Tambe SS, Kulkarni BD (2017) Genetic programming based drag model with improved prediction accuracy for fluidization systems. Int J Chem React Eng 15(2):20160210. https://doi.org/10.1515/ijcre-2016-0210
167. Ma L, Guo Q, Li X, Xu S, Zhou J, Ye M, Liu Z (2022) Drag correlations for flow past monodisperse arrays of spheres and porous spheres based on symbolic regression: effects of permeability. Chem Eng J 445:136653. https://doi.org/10.1016/j.cej.2022.136653
168. Tang Y, Peters EAJF, Kuipers JAM, Kriebitzsch SHL, van der Hoef MA (2015) A new drag correlation from fully resolved simulations of flow past monodisperse static arrays of spheres. AIChE J 61(2): 688–698. https://doi.org/10.1002/aic.14645
169. El Hasadi YMF, Padding JT (2023) Do logarithmic terms exist in the drag coefficient of a single sphere at high Reynolds numbers? Chem Eng Sci 265:118195. https://doi.org/10.1016/j.ces.2022.118195
170. Alhuthali S, Delaplace G, Macchietto S, Bouvier L (2022) Whey protein fouling prediction in plate heat exchanger by combining dynamic modelling, dimensional analysis, and symbolic regression. Food Bioprod Process 134:163–180. https://doi.org/10.1016/j.fbp.2022.05.009
171. Neumann P, Cao L, Russo D, Vassiliadis VS, Lapkin AA (2020) A new formulation for symbolic regression to identify physico-chemical laws from experimental data. Chem Eng J 387:123412. https://doi.org/10.1016/j.cej.2019.123412
172. Farizhandi AAK, Zhao H, Chen T, Lau R (2020) Evaluation of material properties using planetary ball milling for modeling the change of particle size distribution in a gas-solid fluidized bed using a hybrid artificial neural network-genetic algorithm approach. Chemical Engineering Science 215:115469. https://doi.org/10.1016/j.ces.2020.115469
173. Bahonar E, Chahardowli M, Ghalenoei Y, Simjoo M (2022) New correlations to predict oil viscosity using data mining techniques. Journal of Petroleum Science and Engineering 208:109736. https://doi.org/10.1016/j.petrol.2021.109736
174. Hashemizadeh A, Bahonar E, Chahardowli M, Kheirollahi H, Simjoo M (2022) A data-driven approach to estimate the rate of penetration in drilling of hydrocarbon reservoirs. PREPRINT (version 1), available at Research Square. https://doi.org/10.21203/rs.3.rs-1740481/v1
175. Thorat R, Bruining H (2016) Determination of the most significant variables affecting the steady state pressure drop in selected foam flow experiments. Journal of Petroleum Science and Engineering 141:144–156. https://doi.org/10.1016/j.petrol.2015.12.001
176. Yang G, Li X, Wang J, Lian L, Ma T (2015) Modeling oil production based on symbolic regression. Energy Policy 82:48–61. https://doi.org/10.1016/j.enpol.2015.02.016
177. MoradiDowlatabad M, Jamiolahmady M (2018) New approach for predicting multiple fractured horizontal wells performance in tight reservoirs. Journal of Petroleum Science and Engineering 162:233–243. https://doi.org/10.1016/j.petrol.2017.12.040
178. Kamari A, Gharagheizi F, Mohammadi AH, Ramjugernath D (2016) A corresponding states-based method for the estimation of natural gas compressibility factors. Journal of Molecular Liquids 216:25–34. https://doi.org/10.1016/j.molliq.2015.12.103
179. Izadmehr M, Shams R, Ghazanfari MH (2016) New correlations for predicting pure and impure natural gas viscosity. Journal of Natural Gas Science and Engineering 30:364–378. https://doi.org/10.1016/j.jngse.2016.02.026
180. Abooali D, Khamehchi E (2014) Estimation of dynamic viscosity of natural gas based on genetic programming methodology. Journal of Natural Gas Science and Engineering 21:1025–1031. https://doi.org/10.1016/j.jngse.2014.11.006
181. Abooali D, Khamehchi E (2019) New predictive method for estimation of natural gas hydrate formation temperature using genetic programming. Neural Comput Appl 31(7):2485–2494. https://doi.org/10.1007/s00521-017-3208-0
182. Rostami A, Shokrollahi A (2017) Accurate prediction of water dewpoint temperature in natural gas dehydrators using gene expression programming approach. J Mol Liq 243:196–204. https://doi.org/10.1016/j.molliq.2017.08.045
183. Kerlin TW, Upadhyaya BR (2019) Chapter 8—reactor control. In: Kerlin TW, Upadhyaya BR (eds) Dynamics and control of nuclear reactors. Academic Press, Cambridge, pp 89–104. https://doi.org/10.1016/B978-0-12-815261-4.00008-1
184. Shmalko E, Diveev A (2021) Control synthesis as machine learning control by symbolic regression methods. Appl Sci. https://doi.org/10.3390/app11125468
185. Geng X, Mao X, Wu H-H, Wang S, Xue W, Zhang G, Ullah A, Wang H (2022) A hybrid machine learning model for predicting continuous cooling transformation diagrams in welding heat-affected zone of low alloy steels. J Mater Sci Technol 107:207–215. https://doi.org/10.1016/j.jmst.2021.07.038

186. Shen J, Kotha S, Noraas R, Venkatesh V, Ghosh S (2022) Developing parametrically upscaled constitutive and crack nucleation models for the ti64 alloy. Int J Plast 151:103182. https://doi.org/10.1016/j.ijplas.2021.103182
187. Liu H, Lin H, Jiang X, Mao X, Liu Q, Li B (2019) Estimation of mass matrix in machine tool's weak components research by using symbolic regression. Comput Ind Eng 127:998–1011. https://doi.org/10.1016/j.cie.2018.11.033
188. Kabliman E, Kolody AH, Kronsteiner J, Kommenda M, Kronberger G (2021) Application of symbolic regression for constitutive modeling of plastic deformation. Appl Eng Sci 6:100052. https://doi.org/10.1016/j.apples.2021.100052
189. Hale WT, Safikou E, Bollas GM (2022) Inference of faults through symbolic regression of system data. Comput Chem Eng 157:107619. https://doi.org/10.1016/j.compchemeng.2021.107619
190. Nembhard DA, Sun Y (2019) A symbolic genetic programming approach for identifying models of learning-by-doing. Comput Ind Eng 131:524–533. https://doi.org/10.1016/j.cie.2018.08.020
191. Jeong H, Kim JH, Choi S-H, Lee S, Heo I, Kim KS (2022) Semantic cluster operator for symbolic regression and its applications. Adv Eng Softw 172:103174. https://doi.org/10.1016/j.advengsoft.2022.103174
192. Amir Haeri M, Ebadzadeh MM, Folino G (2017) Statistical genetic programming for symbolic regression. Appl Soft Comput 60:447–469. https://doi.org/10.1016/j.asoc.2017.06.050
193. Mousavi Astarabadi SS, Ebadzadeh MM (2018) A decomposition method for symbolic regression problems. Appl Soft Comput 62:514–523. https://doi.org/10.1016/j.asoc.2017.10.041
194. Žegklitz J, Pošík P (2019) Symbolic regression in dynamic scenarios with gradually changing targets. Appl Soft Comput 83:105621. https://doi.org/10.1016/j.asoc.2019.105621
195. Derner E, Kubalík J, Ancona N, Babuška R (2020) Constructing parsimonious analytic models for dynamic systems via symbolic regression. Appl Soft Comput 94:106432. https://doi.org/10.1016/j.asoc.2020.106432
196. Sambo AS, Azad RMA, Kovalchuk Y, Indramohan VP, Shah H (2021) Evolving simple and accurate symbolic regression models via asynchronous parallel computing. Appl Soft Comput 104:107198. https://doi.org/10.1016/j.asoc.2021.107198
197. Wadekar, D., Thiele, L., Villaescusa-Navarro, F., Hill, J.C., Cranmer, M., Spergel, D.N., Battaglia, N., Anglés-Alcázar, D., Hernquist, L., Ho, S.: Augmenting astrophysical scaling relations with machine learning : application to reducing the SZ flux-mass scatter. arXiv (2022). https://doi.org/10.48550/ARXIV.2201.01305
198. Shao H, Villaescusa-Navarro F, Genel S, Spergel DN, Anglés-Alcázar D, Hernquist L, Davé R, Narayanan D, Contardo G, Vogelsberger M (2022) Finding universal relations in subhalo properties with artificial intelligence. Astrophys J 927(1):85. https://doi.org/10.3847/1538-4357/ac4d30
199. Delgado AM, Wadekar D, Hadzhiyska B, Bose S, Hernquist L, Ho S (2022) Modelling the galaxy-halo connection with machine learning. Mon Not R Astron Soc 515(2):2733–2746. https://doi.org/10.1093/mnras/stac1951
200. Matchev KT, Matcheva K, Roman A (2022) Analytical modeling of exoplanet transit spectroscopy with dimensional analysis and symbolic regression. Astrophys J 930(1):33. https://doi.org/10.3847/1538-4357/ac610c
201. Manzi, M., Vasile, M.: Orbital anomaly reconstruction using deep symbolic regression. In: 71st International Astronautical Congress (2020)
202. Barsotti D, Cerino F, Tiglio M, Villanueva A (2022) Gravitational wave surrogates through automated machine learning. Class Quantum Grav 39(8):085011. https://doi.org/10.1088/1361-6382/ac5ba1
203. Arjona R, Lin H-N, Nesseris S, Tang L (2021) Machine learning forecasts of the cosmic distance duality relation with strongly lensed gravitational wave events. Phys Rev D 103:103513. https://doi.org/10.1103/PhysRevD.103.103513
204. Shepherd SJ, Zharkov SI, Zharkova VV (2014) Prediction of solar activity from solar background magnetic field variations in cycles 21–23. Astrophys J 795(1):46. https://doi.org/10.1088/0004-637x/795/1/46
205. Yang G, Sun T, Wang J, Li X (2015) Modeling the nexus between carbon dioxide emissions and economic growth. Energy Policy 86:104–117. https://doi.org/10.1016/j.enpol.2015.06.031
206. Pan I, Pandey DS, Das S (2013) Global solar irradiation prediction using a multi-gene genetic programming approach. J Renew Sustain Energy 5(6):063129. https://doi.org/10.1063/1.4850495
207. Al-Hajj R, Assi A, Fouad M, Mabrouk E (2021) A hybrid lstm-based genetic programming approach for short-term prediction of global solar radiation using weather data. Processes. https://doi.org/10.3390/pr9071187
208. Massaoudi, M., Chihi, I., Sidhom, L., Trabelsi, M., Refaat, S.S., Oueslati, F.S.: Enhanced Evolutionary Symbolic regression via genetic programming for PV power forecasting. arXiv (2019). https://doi.org/10.48550/ARXIV.1910.10065
209. Abdellaoui, I.A., Mehrkanoon, S.: Symbolic regression for scientific discovery: an application to wind speed forecasting. In: 2021 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 01–08 (2021). https://doi.org/10.1109/SSCI50451.2021.9659860
210. Valsaraj P, Thumba DA, Asokan K, Kumar KS (2020) Symbolic regression-based improved method for wind speed extrapolation from lower to higher altitudes for wind energy applications. Appl Energy 260:114270. https://doi.org/10.1016/j.apenergy.2019.114270
211. Pan X, Zhang J, Li C, Pan X, Song J (2019) Analysis of China's regional wind power generation efficiency and its influencing factors. Energy Environ 30(2):254–271. https://doi.org/10.1177/0958305X18788820
212. Rueda R, Cuéllar MP, Pegalajar MC, Delgado M (2019) Straight line programs for energy consumption modelling. Appl Soft Comput 80:310–328. https://doi.org/10.1016/j.asoc.2019.04.001
213. Wenninger S, Kaymakci C, Wiethe C (2022) Explainable long-term building energy consumption prediction using qlattice. Appl Energy 308:118300. https://doi.org/10.1016/j.apenergy.2021.118300
214. Kefer K, Hanghofer R, Kefer P, Stöger M, Hofer B, Affenzeller M, Winkler S (2022) Simulation-based optimization of residential energy flows using white box modeling by genetic programming. Energy Build 258:111829. https://doi.org/10.1016/j.enbuild.2021.111829
215. Pan X, Uddin MK, Ai B, Pan X, Saima U (2019) Influential factors of carbon emissions intensity in oecd countries: evidence from symbolic regression. J Clean Prod 220:1194–1201. https://doi.org/10.1016/j.jclepro.2019.02.195
216. Liu H, Zhang Z (2022) Probing the carbon emissions in 30 regions of china based on symbolic regression and tapio decoupling. Environ Sci Pollut Res 29(2):2650–2663. https://doi.org/10.1007/s11356-021-15648-x
217. Domínguez-Sáez A, Rattá GA, Barrios CC (2018) Prediction of exhaust emission in transient conditions of a diesel engine fueled with animal fat using artificial neural network and symbolic regression. Energy 149:675–683. https://doi.org/10.1016/j.energy.2018.02.080
218. Hughes, J.A., Houghten, S., Brown, J.A.: Gait model analysis of parkinson's disease patients under cognitive load. In: 2020 IEEE Congress on Evolutionary Computation (CEC), pp. 1–8 (2020). https://doi.org/10.1109/CEC48606.2020.9185621

219. Alaa AM, Gurdasani D, Harris AL, Rashbass J, van der Schaar M (2021) Machine learning to guide the use of adjuvant therapies for breast cancer. Nat Mach Intell 3(8):716–726. https://doi.org/10.1038/s42256-021-00353-8
220. Goyal, R.: A symbolic regression approach to hepatocellular carcinoma diagnosis using hypermethylated cpg islands in circulating cell-free dna. medRxiv (2022). https://doi.org/10.1101/2022.01.25.22269799
221. Golap MA-U, Raju SMTU, Haque MR, Hashem MMA (2021) Hemoglobin and glucose level estimation from ppg characteristics features of fingertip video using mggp-based model. Biomed Signal Process Control 67:102478. https://doi.org/10.1016/j.bspc.2021.102478
222. Virgolin, M., Alderliesten, T., Bel, A., Witteveen, C., Bosman, P.A.N.: Symbolic regression and feature construction with gp-gomea applied to radiotherapy dose reconstruction of childhood cancer survivors. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 1395–1402. Association for Computing Machinery, (2018). https://doi.org/10.1145/3205455.3205604
223. Dasgupta P, Hughes JA, Daley M, Sejdić E (2021) Is human walking a network medicine problem? An analysis using symbolic regression models with genetic programming. Comput Methods Progr Biomed 206:106104. https://doi.org/10.1016/j.cmpb.2021.106104
224. Wilstrup C, Cave C (2022) Combining symbolic regression with the cox proportional hazards model improves prediction of heart failure deaths. BMC Med Inform Decis Mak 22(1):1–7. https://doi.org/10.1186/s12911-022-01943-1
225. Cox DR (1972) Regression models and life-tables. J R Statistical Soc: Ser B (Methodol) 34(2):187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x
226. Wilstrup, C., Hedley, P.L., Rode, L., Placing, S., Wøjdemann, K.R., Shalmi, A.-C., Sundberg, K., Christiansen, M.: Symbolic regression analysis of interactions between first trimester maternal serum adipokines in pregnancies which develop pre-eclampsia. medRxiv (2022). https://doi.org/10.1101/2022.06.29.22277072
227. Claveria O, Monte E, Torra S (2022) A genetic programming approach for economic forecasting with survey expectations. Appl Sci. https://doi.org/10.3390/app12136661
228. Kronberger G, Fink S, Kommenda M, Affenzeller M (2011) Macro-economic time series modeling and interaction networks. In: Di Chio C, Brabazon A, Di Caro GA, Drechsler R, Farooq M, Grahl J, Greenfield G, Prins C, Romero J, Squillero G, Tarantino E, Tettamanzi AGB, Urquhart N, Uyar AŞ (eds) Appl Evol Comput. Springer, Berlin, Heidelberg, pp 101–110
229. Drachal, K.: Analysis of Bayesian Symbolic Regression Applied to Crude Oil Price. In: Sinteza 2022 - International Scientific Conference on Information Technology and Data Related Research, pp. 3–13 (2022). https://doi.org/10.15308/Sinteza-2022-3-13
230. Claveria O, Monte E, Torra S (2017) Using survey data to forecast real activity with evolutionary algorithms. a cross-country analysis. J Appl Econ 20(2):329–349. https://doi.org/10.1016/S1514-0326(17)30015-6
231. Claveria O, Monte E, Torra S (2019) Empirical modelling of survey-based expectations for the design of economic indicators in five european regions. Empirica 46(2):205–227. https://doi.org/10.1007/s10663-017-9395-1
232. Claveria O, Monte E, Torra S (2020) Economic forecasting with evolved confidence indicators. Econ Modell 93:576–585. https://doi.org/10.1016/j.econmod.2020.09.015
233. Koza, J.R.: A genetic approach to econometric modeling. In: Sixth World Congress of the Econometric Society, Barcelona, Spain, vol. 27 (1990)