Electronic Research Archive

ERA, 33(9): 5158–5178.
DOI: 10.3934/era.2025231
Received: 20 December 2024
Revised: 16 June 2025
Accepted: 26 June 2025
Published: 01 September 2025

https://www.aimspress.com/journal/ERA

Research article

Reducing the number of input variables through symbolic regression

Dejan Brkić[1,2,*], Pavel Praks[2], Martin Marek[3,4], Uroš Ilić[1] and Zoran Stajić[1]

[1] Faculty of Electronic Engineering, University of Niš, 18000 Niš, Serbia
[2] IT4Innovations, VSB-Technical University of Ostrava, 708 00 Ostrava, Czech Republic
[3] ENET Centre, VSB – Technical University of Ostrava, 708 00 Ostrava, Czech Republic
[4] Department of Technical Studies, College of Polytechnics Jihlava, 586 01 Jihlava, Czech Republic

* **Correspondence:** Email: dejan.brkic@elfak.ni.ac.rs, dejan.brkic@vsb.cz.

Abstract: Symbolic regression, a type of machine learning technique, can efficiently disregard variables that are not significant to the final output, even if they were initially preselected as inputs. Various input parameters are tested in the three examples presented here, where the outputs are modeled using symbolic regression: estimating the middle plasma torch temperature used for waste gasification, the active energy of a solar power plant, and the diameter of a pipe with a known flow and pressure drop through it. Final highly accurate formulas are produced after numerous attempts with lower performances. The process for rejecting the parameters without or with limited influence is automatic and can be performed without human intervention and supervision. The results obtained using symbolic regression are easily interpretable by human experts. This approach shows how to use machine learning-based modeling as an additional tool for sensitivity analysis.

Keywords: symbolic regression; sensitivity analysis; significant input variables; waste gasification; plasma torch; solar plant; electricity production; pipe diameter

Nomenclature:

Section 2.1:
T, middle plasma torch temperature (K)
T, normalized temperature in the reactor (°C)
K, base constant of the nozzle (dimensionless)
U, voltage of the electric arc (V)

P, power of the plasma torch (kW)

f_p, filling pressure (bar)

R^2, goodness of fit (dimensionless)

Section 2.2:
h, time (h)
T_a, average temperature (K)
e, active energy (kWh)
R^2, goodness of fit (dimensionless)
i_D, direct irradiation (J/m^2)
i, irradiation (J/m^2)

Section 2.3:
d, pipe diameter (m)
Q, flow (m^3/s)
ν, kinematic viscosity of fluid (m^2/s)
ε, relative roughness of pipe surface (dimensionless)
Re, Reynolds number (dimensionless)
L, pipe length (m)
ρ, fluid density (kg/m^3)
π, Ludolph number (≈ 3.1415)
Π_1 and Π_2, newly defined dimensionless parameters

1. Introduction

Symbolic regression [1,2], as a type of machine learning, is used in this article for sensitivity analysis to automatically ignore input parameters with no or with minimal impact on the final results. Unimportant parameters are not always easily observable in large datasets. It is essential to highlight that regression tools can discern what is of interest even if such information is concealed within big data. Symbolic regression is a specialized application within symbolic computation that focuses on finding mathematical models to fit the data by searching for both the structure and parameters of equations. While symbolic computation involves a broad range of symbolic operations on mathematical expressions, symbolic regression specifically uses these techniques to discover the equations that describe data patterns.

The methodology above described was developed on the basis of three real-life simulations from recently completed or ongoing projects (including one PhD research project). The simulations described here are as follows:

1) Estimating the middle plasma torch temperature used for waste gasification [3–5], where one parameter was rejected as a valid input and three were kept producing in final very accurate but still simple formula;

2) The active energy of a solar power plant [6], where one parameter among many is identified as the most important; and

3) The diameter of a pipe with a known flow and pressure drop through it [7–9], where one

parameter is rejected as not important.

Symbolic regression [10] is a method based on artificial intelligence that can find the mathematical expressions which best describe the data without the need for a predefined model structure as an assumption, although certain functions can be selected or avoided. It searches for the optimal equation and variables, often using evolutionary algorithms, and produces results that are both accurate and interpretable. Highly accurate formulas are produced following multiple attempts with lower performance. The final formulas, as the results of symbolic regression that align with the input–output relationships within a given dataset, are easily interpretable by human experts.

The Eureqa software tool [11] was used in this article to obtain the required symbolic regression formulas. Of course, software tools which can work on supercomputers are acknowledged for large datasets, such as PySR "High-Performance Symbolic Regression in Python and Julia" [12–14] or AI Feynman [15]. Various tools can be also used, and it is expected that they can produce equally accurate and simple formulations.

As applied here, symbolic regression proved to be an efficient tool for sensitivity analysis, effectively identifying and excluding input variables with little or no influence on the final outcome. Although multiple linear regression was also tested, it failed to capture the complexity of the problem as accurately as symbolic regression. Nonetheless, when used for sensitivity analyses, it reaffirmed the results obtained from symbolic regression.

Symbolic regression in the way shown in this article can be used for preselection of important input parameters if they are given as numerical datasets. It can automatically select parameters with influence that cannot be easily predicted or envisaged in advance from a large set of data.

The article demonstrates that the process for rejecting unimportant input parameters, either without or with limited influence, is automatic and can be performed without human intervention and supervision. The common denominator of the analyzed examples is the energy domain and the elimination of unnecessary input parameters that do not have or have only limited influence on the final outcome. The machine models obtained using symbolic regression are easily interpretable by human experts. The validity of the formulas can be verified through physical trend analysis [16]. This approach shows how machine learning modeling can be used as an additional tool for sensitivity analysis.

This article is structured into three main sections: 1) The introduction, 2) the methods section detailing the testing methodology across three applications, 3) state-of-the-art and selected case studies in a literature review, and 4) the conclusion. Additionally, it includes supplementary data. The linkage among all three problems is in their solution using symbolic regression; i.e., in the elimination of unnecessary input parameters which do not have or have only limited influence on the final outcome.

2. Methods: The testing methodology across three applications

The ability of artificial intelligence to reject unimportant parameters (i.e., parameters with limited influence or without it) is explored here. In this way, machine learning modeling is used as an additional tool for sensitivity analyses, i.e., as a method to reduce the number of input variables. Additionally, in symbolic regression, fixed combinations of variables are often observed in complex expressions. These combinations can be grouped and replaced with composite variables, allowing the dimensionality of the problem to be reduced [17]. Candidate groupings can be generated and tested for functional dependence to ensure the validity. Once validated, these substitutions can be applied within any symbolic regression framework to improve performance and formula recovery.

The general algorithm applied in this study is given in Figure 1.

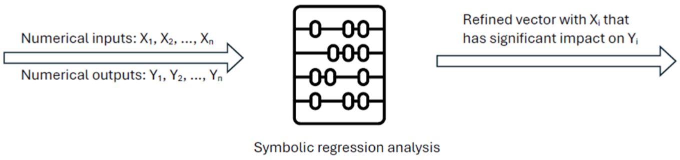

Numerical inputs: $X_1, X_2, ..., X_n$
Numerical outputs: $Y_1, Y_2, ..., Y_n$
Symbolic regression analysis
Refined vector with X_i that has significant impact on Y_i

Figure 1. Algorithm for reducing the number of input variables through symbolic regression.

The methodology is tested on the following three applications:

1) middle plasma torch temperature,
2) the active energy output of a solar power plant, and
3) a pipe diameter.

2.1. Middle plasma torch temperature

A plasma torch (also known as a plasma arc, plasma gun, plasma cutter, or plasmatron) is a device for generating a focused stream of plasma [18–20]. For the purpose of a study on the production of syngas from waste in the Czech Republic aimed at delivering mobile solutions to urban settlements, various parameters influencing the plasma torch used for this purpose were examined. The analyzed facility for waste gasification is equipped with a direct current (DC) low-temperature plasma torch. For the analysis of this waste to energy facility, the middle plasma torch temperature needs to be evaluated as an input parameter of its future digital twin, based on a software tool developed in house (https://shinyenet.vsb.cz/, accessed 13 June 2025). This approach facilitates the analysis of real processes within a virtual environment.

However, the middle plasma torch temperature depends on various parameters, each of them with a level of influence which was unknown in advance.

The experimental design of the middle plasma torch temperature is based on an approximate linear model of the physical conditions of an electric arc within a typical plasma torch. Utilizing standard thermodynamic formulations, the electric power applied to the arc volume (assumed to be cylindrical with predefined parameters) is analyzed. The data collection process considers variations in the plasma gas filling pressures and applied electric power, derived from the arc voltage and current. Through this decomposition, tabular variations of the physical conditions within the electric arc's plasma and the surrounding plasma torch are obtained. A linear model is applied to all tabulated variations of the physical quantities.

It is important to note that the feedback effect of temperature on the fundamental thermodynamic parameters is not incorporated. Consequently, certain regions and operational conditions may deviate from realistic behavior, potentially leading to deformation of the arc ratios and physical parameters in practical applications. Refinement of the model to include temperature feedback on the thermodynamic conditions and parameters is the focus of ongoing research.

2.1.1. Setup, input parameters, and numerical data analysis

The facility for the gasification of waste in the Czech Republic which is examined in this article uses a plasma torch operating with DC, commonly known as a DC plasma torch. It is a device in which a direct current is conducted through a cathode and an anode, resulting in the generation of a powerful electric field. This field induces the ionization of the gas located between the cathode and anode in the form of an arc, ultimately producing low-temperature plasma [21]. For the facility used in this study, waste gasification should be performed in a chamber with a temperature from 750 °C to 1100 °C in the presence of air or steam while the energy is provided through plasma. The main product is a mixture of H_2, CO, and CH_4 as useful gases associated with the undesired N_2 and CO_2 (further CO_2 can be used for CO production). The plasma torch in this study provides plasma with a middle temperature from a minimum of 1295 K to a maximum of 19,750 K according to an estimation based on real physical experiments (which is in agreement with the literature [22] and which should be reduced further to the required temperature prescribed for the gasification chamber).

The pilot plasma torch for this study is given in Figure 2 and a schematic of the plasma-generating system is given in Figure 3.

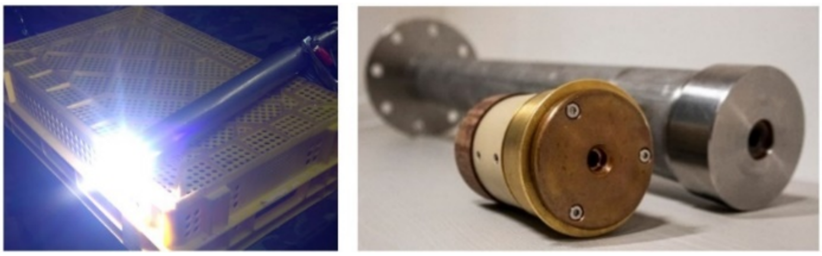


Figure 2. Pilot plasma torch used in this study.

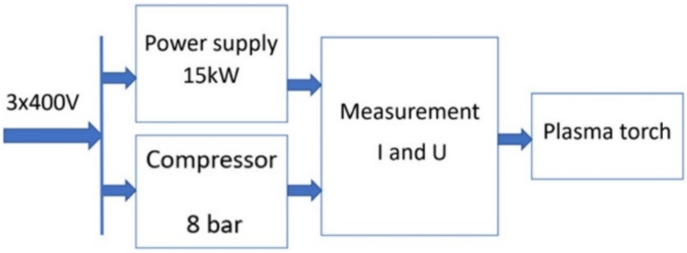


Figure 3. Setup for low-temperature plasma generation.

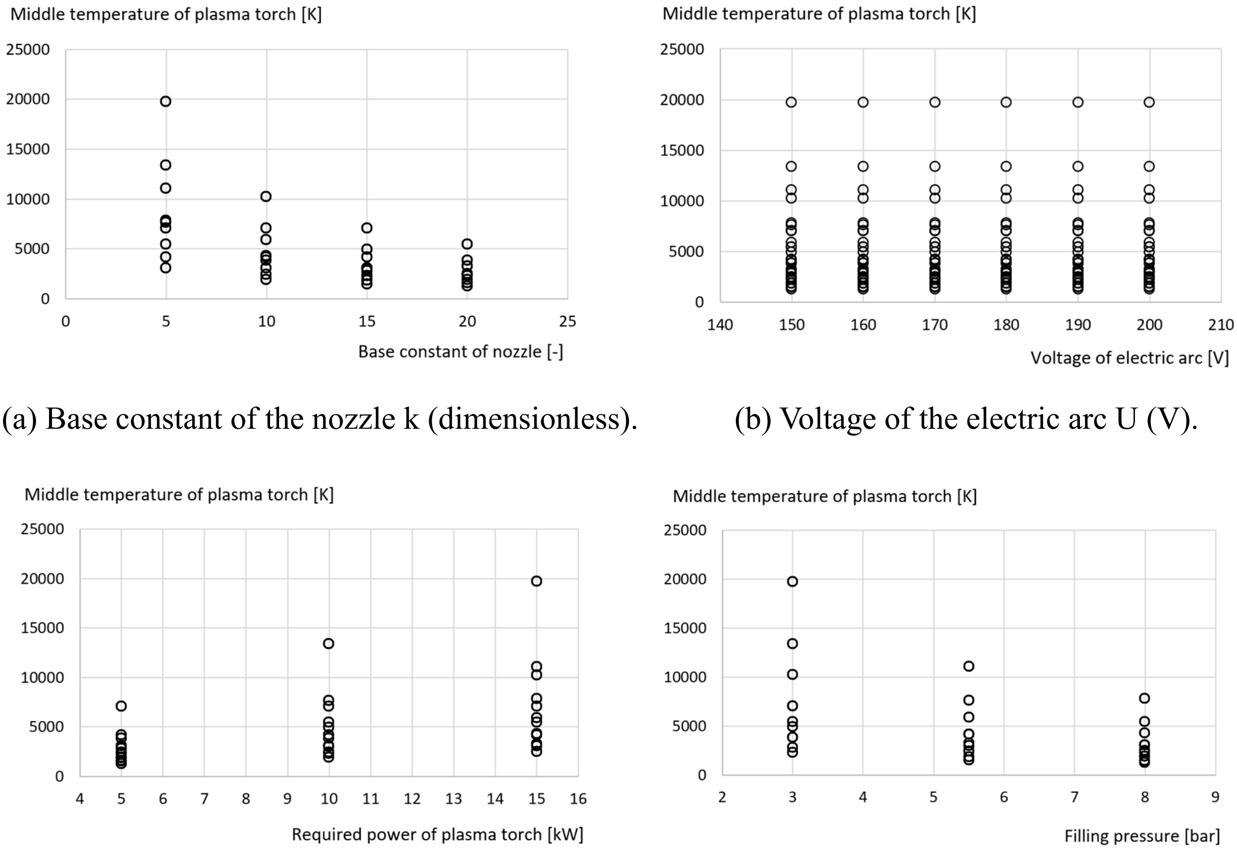
The analyzed combinations of the preselected input parameters are based on the available physical equipment. The maximal output power of the analyzed plasma torch is 15 kW (it can be set to lower values of 10 kW and 5 kW), while the gas filling the space between the electrodes is ambient air or steam delivered at a maximum pressure of 8 bar = 0.8 MPa (it can be set to lower values of 5.5 bar or 3 bar) via a compressor station. This pressurized air is directed through nozzles to sustain the plasma flow and prevent any electrical short circuits between the anode and cathode (set to 5, 10, 15 or 20). Voltage of electric arc of plasma is also measured (150 V, 160 V, 170 V, 180 V, 190 V or 200 V).

To summarize, the values of the output parameter of middle plasma torch temperature (T, in K) were evaluated for the following combinations of input parameters (given in discrete values), which served also as the input dataset for the symbolic regression and multiple linear regression:

1) Base constant of the nozzle, k (dimensionless0 (5, 10, 15, 20);
2) Voltage of the electric arc, U (V) (150 V–200 V; step, 10 V);

3) Required power of the plasma torch, P (kW) (it operates only ay 5 kW, 10 kW, or 15 kW);
4) Filling pressure. f_p (bar) (3 bar, 5.5 bar, 8 bar).

Figure 4 shows the preselected input parameters with possible influence on the middle temperature of the plasma torch. Even without deeper sensitivity analysis, Figure 4(b) shows that the output does not depend on the voltage of the electric arc U (it does not have any influence on the output parameter). Subsequently, the selected symbolic regression tool Eureqa also rejected voltage of yjr electric arc U from the final formulation. Moreover, this finding is confirmed also by multiple linear regression.



(a) Base constant of the nozzle k (dimensionless).

(b) Voltage of the electric arc U (V).

(c) Required power of the plasma torch P (kW).

(d) Filling pressure f_p (bar).

Figure 4. Sensitivity of the input parameters of the plasma torch.

The simplest approach with all possible equidistant combinations of the input parameters was used without any problem in constructing the dataset for symbolic regression and multiple linear regression. The data were split in subsets for training and validation. This dataset possesses a suitable structure and size, eliminating the necessity for employing more complex distributions such as Sobol quasi-Monte Carlo sampling, typically used for the construction of a small but robust quasi-random dataset to provide effective inputs for machine learning (for example, for symbolic regression).

Table 1 (the whole dataset is given in Supplementary File A) contains essential data which are necessary for generating the required expression for the middle plasma torch temperature, T.

Table 1. Data for the symbolic regression of the plasma torch temperature[1].

No.	Combinations[2]	A, k (–)	B, U (V)	C, P (kW)	D, f_p (bar)	T (K)[3]
1	A1, B1, C1, D1	A1 = 5	B1 = 150	C1 = 5	D1 = 3	7048.9
2	A1, B1, C1, D2	A1 = 5	B1 = 150	C1 = 5	D2 = 5.5	4162.9
3	A1, B1, C1, D3	A1 = 5	B1 = 150	C1 = 5	D3 = 8	3080.7
4	A1, B1, C2, D1	A1 = 5	B1 = 150	C2 = 10	D1 = 3	13,398.1
5	A1, B1, C2, D2	A1 = 5	B1 = 150	C2 = 10	D2 = 5.5	7626.1
6	A1, B1, C2, D3	A1 = 5	B1 = 150	C2 = 10	D3 = 8	5461.6
7	A1, B1, C3, D1	A1 = 5	B1 = 150	C3 = 15	D1 = 3	19,747.3
8	A1, B1, C3, D2	A1 = 5	B1 = 150	C3 = 15	D2 = 5.5	11,089.3
9	A1, B1, C3, D3	A1 = 5	B1 = 150	C3 = 15	D3 = 8	7842.6
10	A1, B2, C1, D1	A1 = 5	B2 = 160	C1 = 5	D1 = 3	7048.9
⋮	⋮	⋮	⋮	⋮	⋮	⋮
214	A4, B6, C3, D1	A4 = 20	B6 = 200	C3 = 15	D1 = 3	5461.6
215	A4, B6, C3, D2	A4 = 20	B6 = 200	C3 = 15	D2 = 5.5	3297.1
216	A4, B6, C3, D3	A4 = 20	B6 = 200	C3 = 15	D3 = 8	2485.4

[1]Note: The whole dataset is given as Supplementary File A.

[2]Note: The first 81 combinations of the dataset are generated in MS Excel using: = TEXTJOIN (“,”,,{“A”, “B”, “C”, “D”} & MID (BASE(ROW()-1, 3, 4), {1, 2, 3, 4}, 1) + 1). Parameter A has 4 discrete values, B has 6 values, C has 3 values, and D has 3 values: which gives A·B·C·D = 4·6·3·3 = 216 combinations.

[3]Note: The temperature T (K) of the plasma stream is reduced to the actual temperature in the gasification chamber t (°C) using the formula: t = 0.01917·(T – 273.15) + 726.59 (this relation is based on simple linear regression); The middle temperature in gasification chamber t is given in (°C), while the middle temperature of the plasma torch T is in K. The only reason for this is that the raw data are provided in different units for temperature (°C and K are linearly dependent, where temperature in K = temperature in °C + 273.15).

2.1.2. Regression formulations

It is expected that the same input parameters would have the same influence as for any other plasma torch. Leveraging the measurements and using the regression approach as shown here, similar highly accurate formula can be obtained for any similar plasma torch.

The symbolic and multiple linear regression approaches were tested.

Symbolic regression: Symbolic regression gives Eq (1) in this case, which is recommended for use as being simple and accurate.

$$\mathrm{T} = 700 + \frac{19050 \cdot \mathrm{P}}{\mathrm{k} \cdot \mathrm{f_p}}. \tag{1}$$

Symbolic regression (Eq (1)) provides an adequate solution for the middle plasma torch temperature T (in K) which depends on the required power of the plasma torch P (kW), the base constant of the nozzle k (dimensionless), and the filling pressure f_p (bar).

Equation (1) is not only very simple for coding but is also suitable for human eyes (formulas obtained by machine learning very often can become relatively complex for hand use outside computer codes if they are not transformed in a smart way, which is not always achievable). The formula is very accurate and, at discrete points from the dataset used, introduces the relative error Δ‰, which is less than 0.283‰, i.e., 0.0283% (the error among the tested points can be slightly larger). The relative error is defined as Δ‰ = 1000‰ × |y – f(x)|/y, where f(x) is T from Eq (1) and y is the real measured value of T listed in Table 1.

The plasma torch in this study provides low-temperature plasma with a middle plasma torch temperature T from a minimum of 1295.3 K to a maximum of 19,750.0 K according to real physical experiments and Eq (1) gives very accurate predictions only in this range, with an almost perfect fit, as shown in Figure 5.

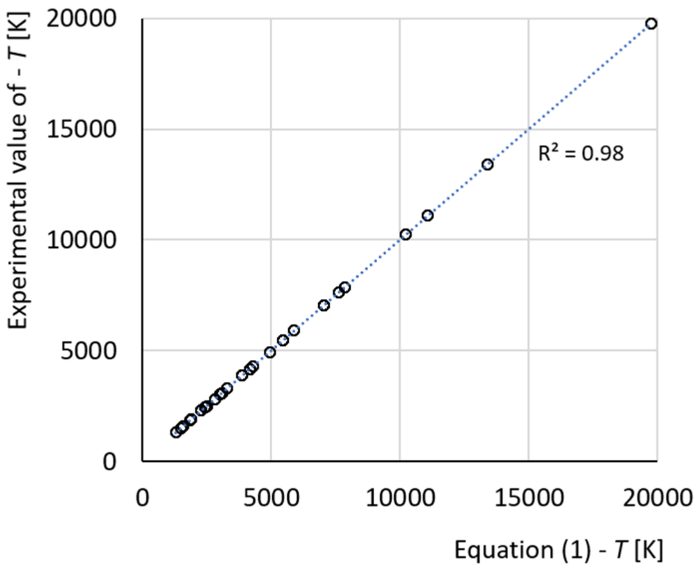


Figure 5. R^2 value of the symbolic regression model from Eq (1).

Equation (1) was run in Eureqa, a software tool for symbolic regression. It is expected that when the Eureqa software is used in multiple trials, the resulting formula may not be identical in each repetition but similar due to the stochastic nature of symbolic regression. Eureqa, a popular symbolic regression software tool, was used for this study, as the size of the input dataset is manageable for a personal computer, where Eureqa can run without difficulties.

The symbolic regression-derived Eq (1) captures the key physical parameters governing the middle plasma torch temperature T (in K). The required power input P (in kW) directly contributes to the thermal energy delivered to the plasma, thereby elevating the temperature. This aligns with the fundamental principle that increased electrical power enhances the energy density of the plasma arc, enhancing ionization and thermal excitation within the plasma column [18].

- The nozzle base constant k (dimensionless) reflects the geometric and material characteristics of the nozzle, such as the throat diameter and wall conductivity, which influence the arc's constriction and heat dissipation. Optimized nozzle designs have been shown to significantly affect the temperature and velocity profiles of plasma jets [19].
- The filling pressure fp (in bar) governs the plasma density and flow regime; higher pressures typically increase convective cooling and collision frequency, which can broaden the arc and reduce peak temperatures [20].

Thus, the inverse dependence on k and f_p reflects their moderating influence on the thermal concentration, while the direct proportionality with P underscores its role as the primary energy source.

The dataset from Supplementary File A used for symbolic regression was split into the training dataset (50%) and the validation dataset (50%) (rows of the dataset were randomly shifted before splitting), while functions for the buildings blocks were constants, coefficients, plus (+), minus (–), multiplication (*), division (/), sinus (sin), cosine (cos), exponential function (exp), logarithm (log), power (^), and square root (sqrt). The default absolute error was used as the error metric for evaluations of the quality of the models discovered.

As shown in Table 2, the task was terminated at 3 min and 18 s after starting the process of finding the best solution when Eq (1) was generated (the software generated 23.08 million formulas every second and needed less than 10 s to reach high accuracy; it did not show all the results but only the most important among them, which are listed in Supplementary File B). Five equations among those listed in Supplementary File B are given in Table 3; Model No. 5 is also given in Eq (1).

Table 2. Reduction in the mean absolute error (MAE) over time for symbolic regression.

Time (seconds)	MAE
0.1	2000
1	1250
2	1000
3	480
5	250
6	~0

The properties of Models 1–5 from Table 3 are given in Figure 6. It can be seen that the symbolic regression generates six models with varying levels of complexity and precision. As expected, very simple models are not accurate, and vice versa. However, the discovered symbolic regression model of Eq (1) fits the given dataset very well, as both the maximum and mean squared errors are very close to zero, and R^2 is ideal at 1.00.

Table 3. Selected symbolic regression models.

Model No.	Solution	Time[1]
1	T = 392.967373821864 × P	10:32:41 AM
2	T = 39212.8038477366/k	10:32:54 AM
3	T = 211561.739609413/(k × fp)	10:33:23 AM
4	T = 21146.7482988027 × P/(k × fp)	10:33:38 AM
5	T = 699.708454771734 + 19047.6190475889 × P/(k × fp)	10:35:31 AM

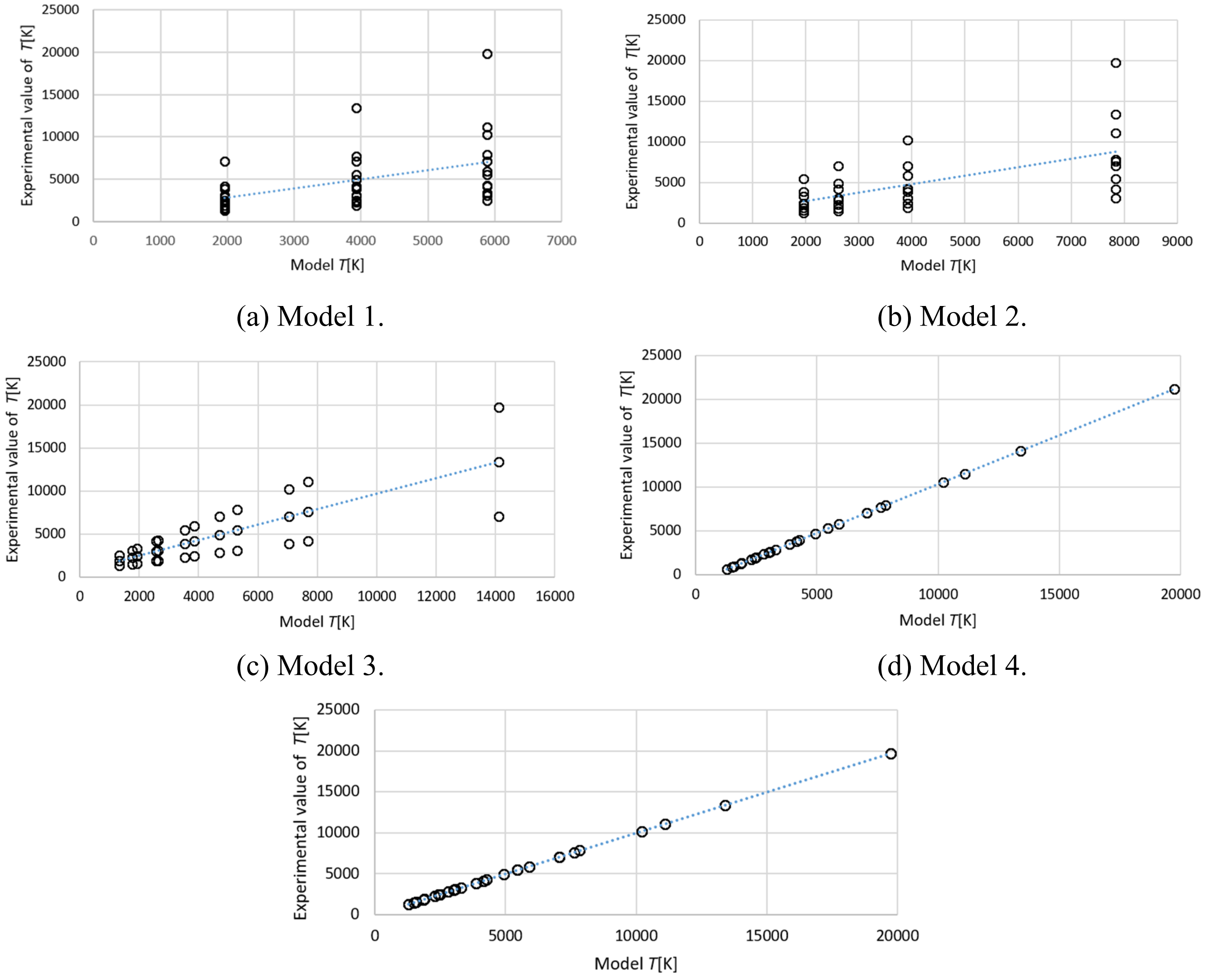
[1]Note: Started at 10:32:40 AM.

The symbolic regression provides a variety of interpretable models with various levels of complexity and precision; Models 1–5 can be seen in Figure 6. For a visual evaluation of the performance of the discovered symbolic regression models, predicted vs. observed plots are presented in Figure 6(a) for Model 1, Figure 6(b) for Model 2, through to Figure 6(e) for Model 5.

We followed the recommendation of Piñeiro [23] to plot the predicted values (the simulated model) on the *x*-axis vs. the observed data on the *y*-axis.

The *x*-axis of the plot represents the predicted (simulated) values of the temperature T, while the *y*-axis represents the observed values of temperature T. Each point on the plot corresponds to a pair of predicted and observed values of temperature T.

For example, Figure 6(a) identifies a systematic error in the simple Model 1, which depends only on one parameter, namely P (power), which has three values (i.e., 5, 10, and 15 kW). Consequently, the corresponding output of Model 1 also has three values; see the *x*-axis of Figure 6(a) for Model 1. It is not surprising that the Model 1 has a low explanatory power: the R^2 value of 0.21 indicates that only 21% of the variance in T is explained by the parameter P in the model. This relatively low explanatory power suggests that while P has a measurable impact on T, there are likely other significant factors influencing T that are not captured by this model.



(a) Model 1.

(b) Model 2.

(c) Model 3.

(d) Model 4.

(e) Model 5.

Figure 6. Properties of the models from Table 3.

A similar situation is seen for Model 2. The model depends only on the parameter k and explains approximately 39% of variability in T.

Model 3, expressed as T = 211561.739609413/(k × fp), demonstrates a significant relationship between the dependent variable T and the independent variables k and fp. With an R^2 value of 0.67, Model 3 explains approximately 67% of the variance in T, indicating a relatively strong explanatory power.

The regression model, defined by the equation T = 21146.7482988027 × P/(k × fp), exhibits an exceptional fit with an R^2 value of 1.000. This indicates that the model explains 100% of the variance in the dependent variable T, suggesting a perfect relationship between T and the predictors P, k, and fp. Such a high R^2 value is rare and implies that the chosen predictors are highly effective in capturing the dynamics influencing T. The same is situation is also seen for Model 5.

The plots of Models 4 and 5 are characterized by a 45-degree line, which represents this exceptional fit: All points lie on this 45-degree line, indicating perfect predictions of the temperature T.

The complexity and error of the models from Table 3 are given in Table 4.

Table 4. Symbolic regression models with different sizes (i.e., complexity) and precision.

Model No.	1	2	3	4	5
Coefficients	1	1	1	1	2
Size	3	4	6	8	10
R^2 goodness of fit	0.14	0.32	0.66	0.98	1
Maximal error[1]	13,852.82	11,904.77	7055.201	1399.421	4.37×10^{-6}

[1] Note: Maximal error is calculated as max(|yi – f(xi)|), where f(x) is from the models and y is the real measured value of the temperature T listed in Table 1, while i is the counter.

Multiple linear regression: MS Excel and the Scikit learn library in Python were used for multiple linear regression. Both of them gave approximately the same results.

The coefficients of multiple linear regression are given in Table 5.

Table 5. Coefficients of multiple linear regression.

Parameter	MS Excel	Python
Intercept	10,157.88	10,157.9
K	-392.897	-392.9
U	-3.6×10^{-16}	1.64×10^{-14}
P	423.3806	423.38
f_p	-826.72	-826.72

The multiple linear regression formulas in MS Excel and Python have identical structures and nearly identical coefficient values, as given in Eq (2)

$$\mathrm{T} = \begin{cases} 10157.88 - 392.897 \cdot \mathrm{k} - 3.6 \times 10^{-16} \cdot \mathrm{U} + 423.3806 \cdot \mathrm{P} - 826.72 \cdot \mathrm{f_p} \\ 10157.9 - 392.9 \cdot \mathrm{k} + 1.64 \times 10^{-14} \cdot \mathrm{U} + 423.38 \cdot \mathrm{P} - 826.72 \cdot \mathrm{f_p} \end{cases}. \quad (2)$$

As shown in Figure 7, the R^2 value of the multiple linear regression model in Eq (2) is 0.7522, which means that it explains approximately 75% of the variability in the output value.

Multiple linear regression is not a reliable method for predictions in this case because negative temperatures are detected, as can be seen in Figure 7, which cannot be expected in reality (e.g., the real experimental temperature of 1493.4 K was simulated as –238 K, and 1294.9 K as –2203 K by multiple linear regression).

Understanding the importance of the input parameters in a predictive model is crucial for better model interpretation and feature selection. For linear models such as multiple linear regression, the importance of the coefficients assigned to each input parameter can be directly examined. These coefficients indicate the relative importance of each feature input parameter in predicting the target variable (temperature in this case). It can be seen that the coefficient of the voltage of the electric arc U is very close to zero in both cases: -3.6 × 10^{-16} in MS Excel, and 1.64 × 10^{-14} in Python. Thus, multiple linear regression shows that this parameter is not important, as the value of the corresponding coefficient is very close to zero.

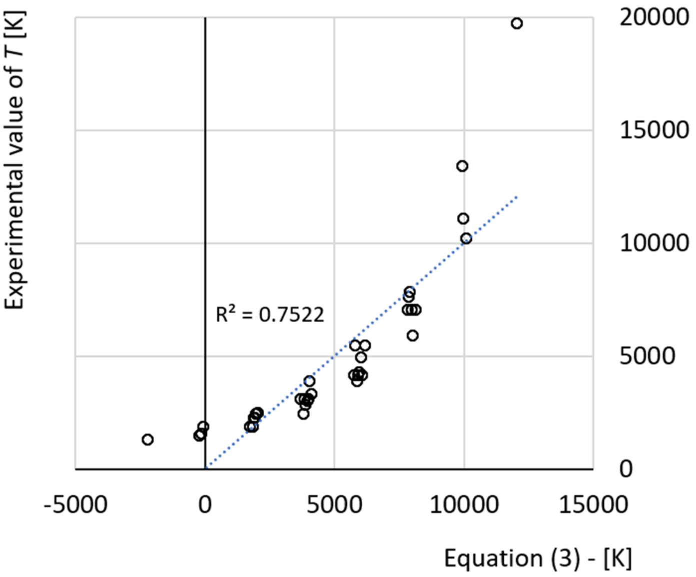


Figure 7. R^2 value of the multiple linear regression model from Eq (2).

2.2. Active energy of a solar plant

A growing number of companies and households in Serbia have been installing rooftop solar power plants, becoming prosumers in the process. These installations offer substantial savings on operational costs, making them highly profitable investments [24].

With the development of the electricity market in Serbia and recent amendments to the Energy Law, the concept of the active buyer has been introduced; see "Serbia adopts changes to Law on Energy–introduces active buyers, dynamic tariffs, auxiliary services market" (https://balkangreenenergynews.com/serbia-adopts-changes-to-law-on-energy-introduces-active-buyers-dynamic-tariffs-auxiliary-services-market/, accessed 28 November 2024). A key characteristic of an active buyer is the obligation to independently manage their balancing responsibility within the electricity market. This means that users who can accurately predict the energy output of their production facilities, such as solar power plants, will gain a competitive advantage through active participation in the open market.

Consequently, forecasting solar power generation is a critical task that needs to be addressed. Solar power generation forecasts rely on meteorological data obtained from various sources. These datasets often contain a large number of parameters collected by meteorological stations. Utilizing all these parameters without a prior analysis of their relevance to the prediction of active energy would be counterproductive, as it would increase the computational load on the artificial intelligence model used for prediction, thereby extending its execution time.

The active energy data used in this study were obtained from the cloud server of a solar power plant owned by a company in Vladičin Han, while the meteorological data were retrieved from the Copernicus Climate Data Store (https://cds.climate.copernicus.eu/datasets, accessed on 5 September 2024) specifically for the region surrounding this town in southern Serbia. The study covers a two-year period from 1 January 2022 to 31 December 2023, with a data sampling rate of one hour.

The input parameters of the observed plant consist of timestamp data, meteorological variables (the average, maximum, and minimum temperatures, irradiance and direct irradiance, various cloud cover types, and precipitation), and the active energy output of the solar power plant at each recorded time point. These parameters are presented in Table 6, where the output parameter is the active energy.

The active energy should be modeled using symbolic regression.

Table 6. Solar plant data used for symbolic regression[1].

No.	Date and time	Temperature (K)			Irradiation (J/m^2)		Cloudiness (%)				Precipitation	Active energy
		T_a Average	Max.	Min.	i	Direct i_D	C	C_H High	C_M Medium	C_L Low	p (m)	e (kWh)
1	1 January 2022 7:00	274.7	277.0	275.2	81,344	93,824	0.6867676	0.39279175	0.2174987	0.35183	0	0.7378
2	1 January 2022 8:00	277.3	279.7	277.6	365,120	420,288	0.5285644	0.40185547	0.1205749	0.231872	0	2.2754
3	1 January 2022 9:00	278.7	282.0	280.1	718,592	824,384	0.3558349	0.16433716	0.1318664	0.187652	0	2.7132
4	1 January 2022 12:00	281.8	282.7	282.3	733,056	836,864	0.7489013	0	0.3848877	0.704895	6.68×10^{-6}	3.4585
5	1 January 2022 13:00	281.6	282.2	281.5	470,848	539,008	0.4190673	0	0.0021972	0.418518	1.43×10^{-6}	4.2651
6	1 January 2022 14:00	281.6	281.4	280.7	199,872	229,248	0.3752441	0	0.0363464	0.350708	8.15×10^{-5}	4.7678
7	1 January 2022 15:00	281.1	280.5	279.8	68,032	77,952	0.5561828	0	0.0067443	0.556182	1.41×10^{-4}	0.9494
8	2 January 2022 7:00	279.1	279.6	278.4	82,368	94,144	0.4236450	0	0	0.423645	2.38×10^{-6}	0.4320
9	2 January 2022 8:00	280.3	280.7	279.9	386,944	441,600	0.3910827	0	0	0.391082	4.77×10^{-7}	2.3703
10	2 January 2022 12:00	284.5	282.6	282.5	1,042,240	1,185,792	0.4821167	0	0	0.482116	0	5.9633
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
8417	31 December 2023 14:00	283.2	282.3	280.9	543,552	618,240	0.2179870	0.21420288	0.0076293	0	0	7.7548
8418	31 December 2023 15:00	282.7	280.5	276.8	162,880	185,920	0.2550964	0.21636963	0.0672302	0	0	3.4112
8419	31 December 2023 16:00	278.3	276.6	276.3	128	128	0.0270996	0.01400756	0.0087280	0.004364	0	0.4057

[1] Note: The whole dataset is given in Supplementary File C.

2.2.1. Symbolic regression formula for active energy

The aim is to predict the values in the column ‘Active energy (kWh)’ (denoted by the symbol ‘e’), using all the available features of the given time-series dataset for symbolic regression (Table 6). It means that information extracted from the features date and time, temperature, irradiation, cloudiness and precipitation is used for symbolic regression.

For example, date and time information from the given time-series dataset are extracted using Python. Consequently, values corresponding to the given year ‘y’, month ‘m’, day ‘d’, hour ‘h’, and minute ‘m’ are used for regression analyses. An example of Python script for extraction of the year, month, day, hour, and minute from a string is given in Supplementary File D.

Symbolic regression is used as an automated tool to discover an unknown symbolic function f, which uses the values of the given features as inputs to simulate the values of the active energy e: Active energy (in kWh) = f(m, d, y, h, Temper, maxtemper, mintemper, Irradiation, DirectIrrad, Cloudiness, Highcloudcover, Mediumcloudcover, Lowcloudcover, Precipitation). For example, the variables ‘Temper’, ‘maxtemper’, and ‘mintepler’ correspond to the columns “T_a Average”, “max.”, and “min.” in the column “Temperature (K)”, respectively.

The list of all formulations obtained is given in Supplementary File E. The best discovered options are given in Eqs (3)–(5).

$$e = 1.97704663673864 \times 10^{-5} \cdot i, \tag{3}$$

$$e = 6.78260970805625 \times 10^{-8} \cdot i \cdot T_a, \tag{4}$$

$$e = 2.289424896529 \times 10^{-10} \cdot i \cdot {T_a}^2. \tag{5}$$

The goodness of fit R^2 for Eqs (3)–(5) are, respectively, 0.82800994, 0.83303623, and 0.83563064. This means that the model describes (fits) 83% of the variability of the dataset, meaning that the simple model given in Eq (3) is sufficiently accurate and that all other parameters apart from irradiation i in Table 6 are not of influence. It is obvious that irradiation i is the most influential input parameter on the active energy of a solar plant.

2.2.2. Refining parameters using symbolic regression

To check which parameters are of influence if the most important detected parameter (irradiation, i) is not available, it should be omitted from the input dataset and the symbolic regression analysis should be repeated.

The symbolic regression is again used as an automated tool to discover an unknown symbolic function f, which uses values of all the given features, except the irradiation i, as inputs to simulate the values of the active energy e: Active energy (in kWh) = f(m, d, y, h, Temper, maxtemper, mintemper, DirectIrrad, Cloudiness, Highcloudcover, Mediumcloudcover, Lowcloudcover, Precipitation).

The most promising solution has a similar logic as the previous solution, reconfirming the influence of irradiation through using only the direct irradiation i_D as given in Eq (6) with a goodness of fit $R^2 = 0.81485$.

$$e = 1.65813936399725 \times 10^{-5} \cdot i_D. \tag{6}$$

When both types of irradiation are omitted from the initial dataset, the next best model identifies hours h (0–24 h) and temperature t_a, where the goodness of fit decline compared with the previous models to $R^2 = 0.648086$. The model is given in Eq (7)

$$e = T_a - 271.079573155232 - 15.1602066014215 \times \sin(0.440729896457029 \times \mathrm{h}). \quad (7)$$

The discovered symbolic regression model uses a sine function, which uses the angle expressed in radians as the input.

The extracted time-dependent part of the automatically discovered law of Eq (7) is visualized in Figure 8. The discovered time-dependent law reflects that the analyzed time-series data are stored mainly within the time interval of 6–17 hours. Detection of the pattern for hours (0–23 h) also reconfirms that the irradiation is the most important parameter in the data and time features, as the irradiation is maximal during the day and minimal during the night. Figure 8 was printed using a Python script, as given in Supplementary File D.

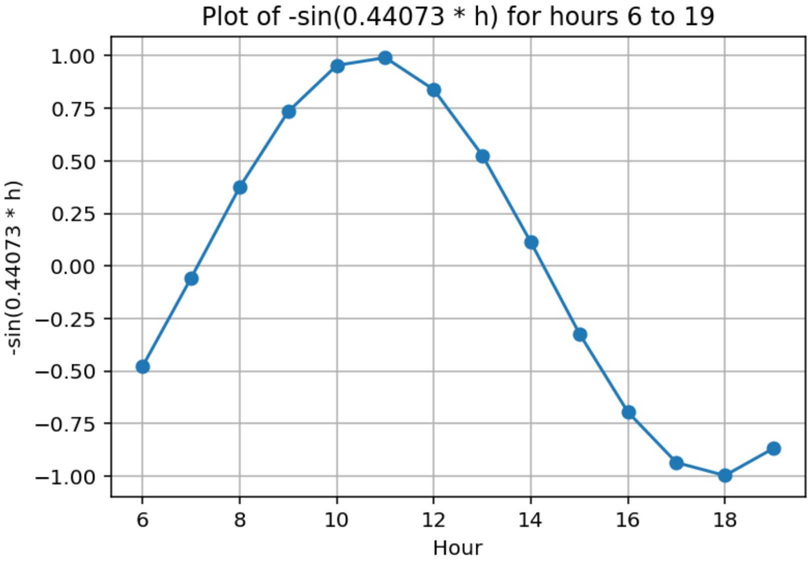


Figure 8. The extracted time-dependent part of the automatically discovered law given by Eq (7), which significantly contributes to the model of irradiation.

To compute the phase shift for the equation sin(0.44073 h), obtained through symbolic regression, we need to understand the general form of a sine function, which can be expressed as $y = \sin(bx + c)$. The phase shift is given by $-c/b$. In the equation sin(0.44073 h), the c term is zero, as symbolic regression penalizes complex equations with a large number of constants. Consequently, there is no phase shift. However, Figure 8 is useful, as it effectively demonstrates the physics of modeling irradiation for 6 to 19 hours.

2.3. Pipe diameter

It was discovered in [7] through symbolic regression that the kinematic viscosity of water ν does not have an influence on the diameter d of the pipe for the known pressure drop Δp, the flow rate Q, the absolute roughness of the pipe ε, the pipe length L, and water density ρ. This means that in Eq (8), the kinematic viscosity of water ν cannot be omitted but can be set to any value between 0 °C ($\nu = 1.785 \times 10^{-6}$ m^2/s) and 100 °C ($\nu = 2.94 \times 10^{-7}$ m^2/s)

$$\mathrm{d} = \frac{4 \cdot \mathrm{Q}}{\nu \cdot \mathrm{Re} \cdot \pi}, \tag{8}$$

where

$$\mathrm{Re} = \frac{1.4446}{\left(\frac{1}{\sqrt{\Pi_1}}\right)^{0.37972} \cdot \left(\Pi_2 + \left(\frac{1}{\sqrt{\Pi_1}}\right)^{0.7}\right)^{0.051}},$$

$$\Pi_1 = \frac{128}{\pi^3} \cdot \frac{\Delta \mathrm{p} \cdot \mathrm{Q}^3}{\rho \cdot \mathrm{L} \cdot \nu^5},$$

$$\Pi_2 = \frac{\pi \cdot \nu \cdot \varepsilon}{4 \cdot \mathrm{Q}}.$$

For example [8], for $\Delta p = 52{,}875.9$ Pa, $Q = 0.051$ m^3/s, $\varepsilon = 0.5 \times 10^{-3}$ m, L = 1000 m, and ρ = 1000 kg/m^3, the diameter of the pipe varies from d = 0.2569 m to 0.2547 m, which is a minor variation, for values for the kinematic viscosity of water from 0 °C ($\nu = 1.785 \times 10^{-6}$ m^2/s) to 100 °C ($\nu = 2.94 \times 10^{-7}$ m^2/s), respectively.

The whole dataset for the pipe diameter problem is given in [7].

3. State of the art and selected case studies: Literature review

Orzechowski et al. [25] provides the state-of the art of machine learning approaches, including linear and Lasso regression, gradient and adaptive boosting, and extreme gradient boosting. Analyses of 100 regression benchmarks from open-source repositories show that symbolic regression performs strongly compared with state-of-the-art gradient boosting algorithms.

Moreover, Aldeia and França [26] provide a benchmark of another 100 equations with applications to physics. Their results also show that symbolic regression is an interesting alternative from 'white box' linear regression to 'black box' neural network models, as symbolic regression provides accurate, explainable models.

Classical machine learning methods, such as gradient boosting libraries like XGBoost, are faster and even overperform deep learning for tabular data in terms of precision [27]. Moreover, another difficulty with deep learning models is that there are 'black boxes', making them difficult to interpret. Unfortunately, the learned parts of neural networks typically are linear piecewise approximations, which extrapolate only linearly [28]. Thus, to provide interpretable machine learning, symbolic regression, which is highly interpretable, has been introduced [13]. The drawback of symbolic regression is its computational speed [26]. According to Orzechowski et al. [25], symbolic regression models are among the slowest machine learning methods. Nevertheless, the recent review by Tonda [14] shows that the aforementioned open-source tool PySR provides high-performance symbolic regression in Python and Julia.

Makke and Chawla [2] highlighted the interpretability of symbolic regression, particularly as machine learning (ML) models, such as ChatGPT-4, grow in complexity and size. According to them, symbolic regression is likely to play a central role in the future of machine learning, especially in the scientific domain.

Stajić et al. [29] and Praks et al. [30] successfully used PySR and SymbolicRegression.jl for modeling and predicting dynamic systems. These applications also include modeling gas prices [29] and gasification of alternative fuels [30]. In [30], the criterion for dynamic systems is used to evaluate the complexity of symbolic regression by utilizing approximate entropy and sample entropy.

Recently, a classical and symbolic regression methods was used by Praks et al. [31] to provide fast yet reliable and interpretable machine learning approximations for predicting open porous media reservoir simulations. In this article [31], regression methods were also employed, as they offer suitable interpretability and effectively approximate the analyzed processes.

In this study, the Eureqa software tool [32] with its default settings is used, as recommended for general-purpose applications and to avoid the potential bias introduced by manual parameter tuning. Eureqa is a commercial, closed-source software designed for symbolic regression. Technical details about the software are provided in [32], which focuses primarily on the underlying methodology. Key aspects include the use of genetic programming to evolve mathematical expressions, the application of Pareto optimization to balance model's accuracy and complexity, and the representation of equations as expression trees. However, this study does not specify fixed values for parameters such as the population size, number of generations, crossover probability, or mutation probability. Instead, it notes that new equations are formed by recombining previous equations and probabilistically varying their subexpressions. For a broader overview of Eureqa's capabilities and its application to real-world datasets, such as radon measurement evaluation, the readers are referred to the review by Dubčáková [11].

4. Conclusions

Symbolic regression can be efficiently used as a tool for sensitivity analyses for identifying and excluding irrelevant input parameters from the models. The methodology was successfully tested using real three cases in the energy domain: Estimating the middle plasma torch temperature used for waste gasification in the Czech Republic, predicting the active energy output of a solar power plant in Serbia, and a pipe diameter study. The linkage among all three problems is their solution using symbolic regression (i.e., in eliminating unnecessary input parameters which do not have or have limited influence on the final outcome). In the case of estimating the middle plasma torch temperature, the proposed methodology is tested against a traditional method: multiple linear regression. The results show that the quality of symbolic regression's predictions outperform the traditional multiple linear regression. In the case of predicting active energy output, the symbolic regression automatically discovers a time-dependent law, which can be easily implemented. The results of the regression analyses together with a visualization of the time-dependent model demonstrate that the discovered law significantly contributes to modeling the irradiation. As less significant parameters are identified and excluded, the refining symbolic regression model of irradiation can be based only on hours. In all cases, symbolic regression reduces the computational complexity and improves the model's interpretability, as it is shown that some input parameters can be completely omitted from the model, while maintaining acceptable accuracy of predictions.

The advantage of symbolic regression is its ability to discover the underlying functional forms directly from the data. This contrasts with linear and Lasso regression, which are limited to linear relationships, and gradient boosting methods, which rely on additive models. Moreover, symbolic regression performs very well on small datasets, as traditional machine learning methods can fail due to overfitting or a lack of sufficient data to train complex models [33].

Recently, Shmuel et al. [34] proposed integrating symbolic regression as a feature engineering method within a machine learning or deep learning model to enhance its performance on 'out-of-distribution' (OOD) data. This refers to data collected at different times, possibly under varying conditions or in different environments, compared with the data used to create the model.

Future research includes identifying and ranking of the most impactful input variables influencing the models by employing global sensitivity analysis. For instance, Praks et al. [31] implemented global sensitivity analysis in conjunction with symbolic regression to accelerate and interpret complex physical modeling processes, drawing on predictive outputs from the open-source porous media reservoir simulator, OPM Flow.

Use of AI tools declaration

The authors declare they have not used artificial intelligence (AI) tools in the creation of this article.

Acknowledgments

Dejan Brkić, Uroš Ilić, and Zoran Stajić want to acknowledge the following. This work has been supported by the Ministry of Science, Technological Development, and Innovation of the Republic of Serbia (grant number 451-03-136/2025-03/200102). Dejan Brkić and Pavel Praks received support from the European Union under the REFRESH - Research Excellence For REgion Sustainability and High-tech Industries project number CZ.10.03.01/00/22_003/0000048 via the Operational Programme Just Transition and the Ministry of Education, Youth and Sports of the Czech Republic through e–INFRA CZ (ID: 90254).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. D. Angelis, F. Sofos, T. E. Karakasidis, Artificial intelligence in physical sciences: Symbolic regression trends and perspectives, *Arch. Comput. Methods Eng.*, **30** (2023), 3845–3865. https://doi.org/10.1007/s11831-023-09922-z
2. N. Makke, S. Chawla, Interpretable scientific discovery with symbolic regression: A review, *Artif. Intell. Rev.*, **57** (2024), 2. https://doi.org/10.1007/s10462-023-10622-0
3. M. Marek, D. Brkić, P. Praks, T. Kozubek, J. Frantík, Experimental analysis of magnetic focusing of the plasma arc of a cutting torch, *Materials*, **18** (2025), 1811. https://doi.org/10.3390/ma18081811
4. R. K. Mohanta, D. Kumawat, G. Ravi, Effect of chamber pressure on the output characteristics of a low-pressure DC plasma torch, *J. Appl. Phys.*, **134** (2023), 153302. https://doi.org/10.1063/5.0160624

5. E. Sanjaya, A. Abbas, Plasma gasification as an alternative energy-from-waste (EFW) technology for the circular economy: An environmental review, *Resour. Conserv. Recycl.*, **189** (2023), 106730. https://doi.org/10.1016/j.resconrec.2022.106730
6. T. M. Pavlović, I. S. Radonjić, D. D. Milosavljević, L. S. Pantić, A review of concentrating solar power plants in the world and their potential use in Serbia, *Renewable Sustainable Energy Rev.*, **16** (2012), 3891–3902. https://doi.org/10.1016/j.rser.2012.03.042
7. D. Brkić, P. Praks, R. Praksová, T. Kozubek, Symbolic regression approaches for the direct calculation of pipe diameter, *Axioms*, **12** (2023), 850. https://doi.org/10.3390/axioms12090850
8. D. Brkić, Z. Stajić, M. Živković, Sizing pipes without iterative calculus: Solutions for head loss, flow discharge and diameter, in *2023 24th International Carpathian Control Conference (ICCC)*, IEEE, (2023), 71–76. https://doi.org/10.1109/ICCC57093.2023.10178917
9. D. Brkić, Revised friction groups for evaluating hydraulic parameters: Pressure drop, flow, and diameter estimation, *J. Mar. Sci. Eng.*, **12** (2024), 1663. https://doi.org/10.3390/jmse12091663
10. G. Kronberger, B. Burlacu, M. Kommenda, S. M. Winkler, M. Affenzeller, *Symbolic Regression*, CRC Press, 2024. https://doi.org/10.1201/9781315166407
11. R. Dubčáková, Eureqa: Software review, *Genet. Program. Evolvable Mach.*, **12** (2011), 173–178. https://doi.org/10.1007/s10710-010-9124-z
12. F. Llorella, J. A. Cebrián, A. Corbi, A. M. Pérez, Fostering scientific methods in simulations through symbolic regressions, *Phys. Educ.*, **59** (2024), 045010. https://doi.org/10.1088/1361-6552/ad3cad
13. M. Cranmer, Interpretable machine learning for science with pysr and symbolicregression.jl, preprint, preprint, arXiv:2305.01582.
14. A. Tonda, Review of PySR: High-performance symbolic regression in Python and Julia, *Genet. Program. Evolvable Mach.*, **26** (2025), 7. https://doi.org/10.1007/s10710-024-09503-4
15. S. M. Udrescu, M. Tegmark, AI Feynman: A physics-inspired method for symbolic regression, *Sci. Adv.*, **6** (2020), eaay2631. https://doi.org/10.1126/sciadv.aay2631
16. K. Wang, T. Shen, J. Wei, J. Liu, W. Hu, An intelligent framework for deriving formulas of aerodynamic forces between high-rise buildings under interference effects using symbolic regression algorithms, *J. Build. Eng.*, **99** (2025), 111614. https://doi.org/10.1016/j.jobe.2024.111614
17. P. Kahlmeyer, M. Fischer, J. Giesen, Dimension reduction for symbolic regression, in *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI, (2025), 17707–17714. https://doi.org/10.1609/aaai.v39i17.33947
18. S. Nguyen-Kuok, The arc plasma torches, in *Theory of Low-Temperature Plasma Physics*, Springer, (2017), 285–366. https://doi.org/10.1007/978-3-319-43721-7_8
19. N. Yu, Y. Yang, R. Jourdain, M. Gourma, A. Bennett, F. Fang, Design and optimization of plasma jet nozzles based on computational fluid dynamics, *Int. J. Adv. Manuf. Technol.*, **108** (2020), 2559–2568. https://doi.org/10.1007/s00170-020-05568-4
20. A. A. Safronov, V. E. Kuznetsov, O. B. Vasilieva, Y. D. Dudnik, V. N. Shiryaev, AC plasma torches. Arc initiation systems. Design features and applications, *Instrum. Exp. Tech.*, **62** (2019), 193–200. https://doi.org/10.1134/S0020441219020246
21. M. Skakov, A. Miniyazov, T. Tulenbergenov, I. Sokolov, G. Zhanbolatova, A. Kaiyrbekova, et al., Hydrogen production by methane pyrolysis in the microwave discharge plasma, *AIMS Energy*, **12** (2024), 548–560. https://doi.org/10.3934/energy.2024026

22. J. Deng, J. Zhang, Q, Zhang, S. Xu, Effects of induction coil parameters of plasma torch on the distribution of temperature and flow fields, *Alexandria Eng. J.*, **60** (2021), 501–510. https://doi.org/10.1016/j.aej.2020.09.022

23. G. Piñeiro, S. Perelman, J. P. Guerschman, J. M. Paruelo, How to evaluate models: Observed vs. predicted or predicted vs. observed?, *Ecol. Modell.*, **216** (2008), 316–322. https://doi.org/10.1016/j.ecolmodel.2008.05.006

24. Z. Čorba, D. Milićević, B. Dumnić, B. Popadić, The experiences of the realization of PV power plants after implementation of the prosumers status, *J. Process. Energy Agric.*, **27** (2023), 13–15. https://doi.org/10.5937/jpea27-43506

25. P. Orzechowski, W. La Cava, J. H. Moore, Where are we now?: A large benchmark study of recent symbolic regression methods, in *Proceedings of the Genetic and Evolutionary Computation Conference*, Association for Computing Machinery, (2018), 1183–1190. https://doi.org/10.1145/3205455.3205539

26. G. S. I. Aldeia, F. O. de França, Interpretability in symbolic regression: A benchmark of explanatory methods using the Feynman data set, *Genet. Program. Evolvable Mach.*, **23** (2022), 309–349. https://doi.org/10.1007/s10710-022-09435-x

27. V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, G. Kasneci, Deep neural networks and tabular data: A survey, *IEEE Trans. Neural Networks Learn. Syst.*, **35** (2024), 7499–7519. https://doi.org/10.1109/TNNLS.2022.3229161

28. M. Cranmer, A. Sanchez-Gonzalez, P. Battaglia, R. Xu, K. Cranmer, D. Spergel, et al., Discovering symbolic models from deep learning with inductive biases, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., **33** (2020), 17429–17442.

29. L. Stajić, R. Praksová, D. Brkić, P. Praks, Estimation of global natural gas spot prices using big data and symbolic regression, *Resour. Policy*, **95** (2024), 105144. https://doi.org/10.1016/j.resourpol.2024.105144

30. P. Praks, M. Lampart, R. Praksová, D. Brkić, T. Kozubek, J. Najser, Selection of appropriate symbolic regression models using statistical and dynamic system criteria: Example of waste gasification, *Axioms*, **11** (2022), 463. https://doi.org/10.3390/axioms11090463

31. P. Praks, A. Rasmussen, K. O. Lye, J. Martinovič, R. Praksová, F. Watson, et al., Sensitivity analysis of parameters for carbon sequestration: Symbolic regression models based on open porous media reservoir simulators predictions, *Heliyon*, **10** (2024), e40044. https://doi.org/10.1016/j.heliyon.2024.e40044

32. M. Schmidt, H. Lipson, Distilling free-form natural laws from experimental data, *Science*, **324** (2009), 81–85. https://doi.org/10.1126/science.1165893

33. C. Wilstrup, J. Kasak, Symbolic regression outperforms other models for small data sets, preprint, arXiv:2103.15147.

34. A. Shmuel, O. Glickman, T. Lazebnik, Machine and deep learning performance in out-of-distribution regressions, *Mach. Learn.: Sci. Technol.*, **5** (2024) 045078. https://doi.org/10.1088/2632-2153/ada221

Supplementary: Data are given as an electronic annex which consists of 1) the integral version of Table 1 with the complete dataset for middle plasma torch temperature used for regression (Supplementary File A) and an integral version of Table 3 with the history of the models obtained by Eureqa (Supplementary Material B; 2) the integral version of Table 6 with the complete dataset for the active energy output of a solar power plant used for symbolic regression (Supplementary File C), and the Python script for extraction of the year, month, day, hour, and minute from a string and Python script to print the automatically extracted time-dependent contribution for modeling the active energy using symbolic regression (Supplementary File D, and the history of all models obtained by Eureqa (Supplementary File E).