Data Driven Modeling of Thermal Energy System Using Symbolic Regression

*Note: Sub-titles are not captured in Xplore and should not be used

1st Zhaoming Liu
Department of Electrical Engineering
Wuhan University
Wuhan, China
liuzhaoming@whu.edu.cn

2nd Chengxi Liu
Department of Electrical Engineering
Wuhan University
Wuhan, China
liuchengxi@whu.edu.cn

3rd Mingyuan Liu
Department of Electrical Engineering
Wuhan University
Wuhan, China
liumingyuan@whu.edu.cn

***Abstract*—Efficient modeling of thermal energy storage systems is crucial for their optimal control. However, conventional methods often face challenges in balancing model complexity with interpretability. This study explores an innovative application of symbolic regression to develop data-driven models for complex stratified thermal energy storage systems that achieve both high accuracy and high interpretability. This research presents the first systematic application of symbolic regression to full-system dynamic modeling of stratified thermal energy storage. Through a progressive validation of virtual models and final real experimental data, this study demonstrates that symbolic regression can automatically derive explicit mathematical equations describing the temperature evolution of each layer directly from operational data. Experimental results show that with limited dataset sizes, this method achieves significantly better prediction accuracy than black-box models such as LSTM, while generating interpretable models that reveal key physical mechanisms like inter-layer convection in transparent mathematical forms. This enables clear interpretation of the system's intrinsic dynamics and confirms the substantial potential of symbolic regression for modeling complex energy systems.**

***Index Terms*—Symbolic Regression, Thermal Energy Storage, Data driven Modeling, System Identification**

I. Introduction

Thermal Energy Storage(TES) systems play a pivotal role in enhancing the flexibility and efficiency of modern energy systems, enabling the integration of intermittent renewable sources and optimizing energy use in buildings and industrial processes [1]. Under this background, it exhibits temporal dynamics, stratification, internal coupling, and limited observability, particularly in stratified storage configurations. Accurate modeling of TES dynamics is essential for effective control, energy optimization, and integration into larger-scale cyber-physical infrastructures [2]. TES hold significant practical importance, most existing TES models tend to rely on physics-based approaches [3] [4] [5]. Some studies have attempted to model the dynamics of TES systems using black-box methods, such as neural networks [6]; however, these models typically lack interpretability and physical insight.

Identify applicable funding agency here. If none, delete this.

Therefore, when an existing commercial building plans an energy system upgrade, it often faces challenges such as lacking precise dimensional data of the TES unit. This makes it impossible to construct reliable physics-based white-box models, while black-box models struggle to perform effectively with limited dataset sizes. Thus, a growing number of studies have turned their attention to methods that balance physical interpretability with the performance of black-box models. For instance, in [7], to address the requirements of model predictive control (MPC) for heat pump-TES coupled systems, an approach is proposed that, based on the differences in the dynamic characteristics of variables, adopts dynamic models constructed by Sparse Identification of Nonlinear Dynamics with control(SINDyc) and Dynamic Mode Decomposition with control(DMDc) as well as static models built via Least Absolute Shrinkage and Selection Operator(LASSO) regression, and further designs two types of state-space models: coupled fully data-driven models and decoupled hybrid static-dynamic models.

Symbolic Regression (SR) has attracted growing interest in this context. By producing explicit mathematical expressions from data, SR enables models that are potentially more interpretable and physically meaningful than standard black-box methods [8] [9].

In the field of energy, many researchers have also made progress using the SR method. In [10], the researchers proposed a method that integrates SR with Structured Grammatical Evolution to optimize key thermodynamic parameters in the thermal model of lithium-ion batteries, including the drag coefficient, friction factor, and Nusselt number. In the field of thermal energy, SR is also applied. In [11], The researchers proposed a method for constructing heat transfer correlations based on SR. By automatically searching for the optimal function structure and constants, and verifying with experimental data from compact heat exchangers and in-tube flow, it is shown that the prediction errors of the heat transfer correlations generated by this method are significantly lower than those of traditional correlations, providing an effective approach for the algorithmic construction of heat transfer

979-8-3315-8138-1

correlations.

This study aims to address this research gap by applying SR to the modeling of TES systems using both simulation-based and experimental datasets. In doing so, we investigate the extent to which SR can provide interpretable and physically meaningful representations of the dynamic behavior of stratified TES layers, and evaluate its potential as a complementary approach to traditional physics-based and black-box modeling techniques.

II. Methodology: A Physics-Informed Symbolic Regression Framework for TES Modeling

A. Using Symbolic Regression to Model Dynamic System

SR is a machine learning technique based on genetic programming, whose core objective is to automatically discover underlying mathematical expressions from given data without pre-specifying the concrete functional form of the model (e.g., linear, polynomial, etc.). Unlike traditional parameter regression, SR simultaneously searches for both the model structure and the model parameters.

Researchers have previously attempted to apply SR for the identification of dynamic systems. In [12] differential equations were discretized via the Euler's method, transforming the dynamic identification problem into a regression task manageable by SR. This process can be described by the following equations:

$$\frac{\mathrm{d}h_l}{\mathrm{d}t} = f_l(\mathbf{h}, \mathbf{x}) \tag{1}$$

where $\mathbf{h}$ is the vector of state variables describing the system, and $\mathbf{x}$ is the system inputs.

$$h_l(t + \Delta t) \approx h_l(t) + \Delta t \cdot f_l(\mathbf{h}(t), \mathbf{x}(t)) \tag{2}$$

Thus, the next system state can be predicted accordingly, and given the initial state and inputs at each time step, the system evolution can be derived from the SR-based model. Based on this, this study applies this approach to the modeling of TES, with a particular focus on defining the state variables to be identified. Therefore, analyzing the physical mechanisms of TES is essential to this work.

B. Thermal Energy Storage Internal Dynamic Process

According to [3], two commonly used layered modeling approaches have been developed to simulate the internal temperature distribution within TES systems: constant layer volume (CLV) model and the constant layer temperature (CLT) model.

In contrast, the CLV model maintains a fixed thickness for each layer and reflects the system's dynamic behavior by updating the average temperature within each layer. This approach is well-suited for simulations with fixed time-step integration methods. In the CLV, the TES is discretized into several layers of fixed volume, and the temperature of each layer evolves over time. The temperature change of each layer is jointly determined by the following four main physical processes:

$$T_{l,t} = T_{l,t-1} + \Delta T^{\text{conv}}_{l,t} + \Delta T^{\text{cond}}_{l,t} + \Delta T^{\text{buoy}}_{l,t} - \Delta T^{\text{loss}}_{l,t} \tag{3}$$

Among these, the temperature changes induced by heat loss and buoyancy can be neglected in many cases, while the primary sources of temperature variation that require focused attention are the remaining convective exchange and conductive exchange. Conductive heat exchange only becomes relevant when internal heating sources are present. Therefore, in conventional heating configurations, convective exchange remains the dominant factor influencing temperature dynamics as in figure 1. It can be described by the following equation:

$$\begin{aligned}\Delta T^{\text{conv}}_{l,t} = \frac{V^{\text{in}}_{l,t}}{V_l}(T_{l-1,t-1} - T_{l,t-1}) + \frac{V^{\text{out}}_{l+1,t}}{V_l}(T_{l+1,t-1} - \\ T_{l,t-1}) + \frac{V^{p,\text{in}}_{t}}{V_l}(T^{p,\text{in}}_{t} - T_{l,t-1})\end{aligned} \tag{4}$$

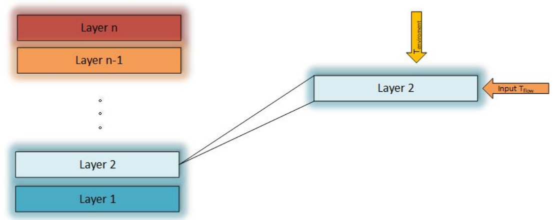


Fig. 1. Explanation for Treating the Separate TES Temperature Layer as PT1.

While the CLV framework provides a structured physics-based description of the dynamics of TES, its practical application still has several limitations. The model relies on a variety of key parameters. When applied to unfamiliar TES systems(for instance, when system aging causes changes to its parameters), these parameters may exhibit significant variations, which in turn render rule-based physical modeling difficult to implement.

Against this background, SR offers a promising approach for automatically discovering compact analytical expressions that capture the layer-wise temperature dynamics from input-output data, while maintaining physical plausibility and model transparency.

C. Specific Method for Applying Symbolic Regression to Thermal Energy Storage Systems

Based on the contents mentioned in the two subsections above, designing a dedicated SR-based dynamic system identification process for TES has become quite straightforward. During the modeling process, according to the physical mechanisms, the input data includes the temperatures of the n layers, denoted as T_1 to T_n (serving as the hidden states of the system), the inlet fluid temperature T_{in}, and its mass flow rate $\dot{m}$. The output target is the temperature change dT_l for each layer per unit time step. The goal of SR is to discover a mathematical expression that describes the relationship between these inputs and dT_l, which can be regarded as an approximation of the

979-8-3315-8138-1

system's ordinary differential equations(ODEs). This process is illustrated in figure 2.

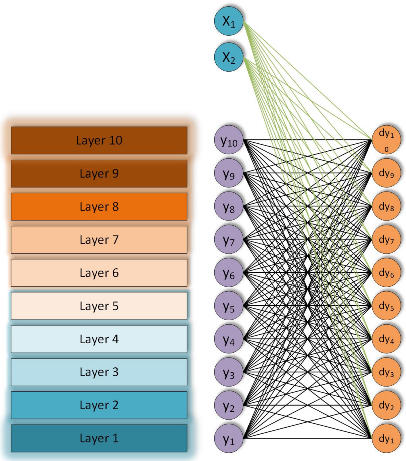


Fig. 2. Detailed Explanation of How to Study TES System Using SR Method.

After obtaining the prediction results, to demonstrate their accuracy, the derived mathematical expressions can be used to predict the response of the same system under different operating conditions. The strategy employed is autoregression, where the state variable values at the current time step are used as inputs for the next time step. This approach enables the reconstruction of the complete prediction trajectory based solely on the initial state and the inputs at each time step.

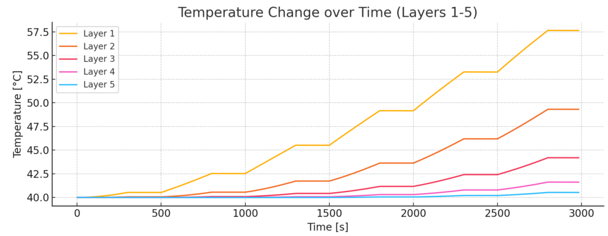
D. Progressive Validation Strategies and Experimental Design

To systematically validate the proposed SR approach for modeling TES systems, this study employs a progressive strategy that evolves from basic dynamic systems to increasingly complex virtual and real TES applications.

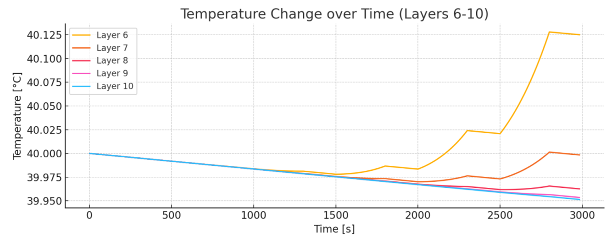
1) Selection of SR Tool: In this study, the SR tool employed is SymbolicRegression.jl, an open-source library written in the Julia programming language. It can be directly integrated and executed within the Julia environment, based on advantages of Julia language, providing a high-performance and flexible toolkit for symbolic model discovery [13].

2) Virtual Hierarchical TES System Modeling: In this section, this study applies the proposed SR method to a complete, hierarchical virtual TES system.

The data for the virtual TES system was generated using the open-source modeling software OpenModelica, with the model derived from the stratified thermal storage model in the AixLib library. The system was discretized into 10 layers of equal volume (nSeg = 10). The temperature and mass flow rate of the inlet fluid were controlled via an external MassFlowSource component to simulate the dynamic behavior of the TES under different operating conditions. With the temperature of the inflowing liquid set to constant, the flow rate is a periodic pulse. The input liquid flows in from the top and out from the bottom. The simulation duration was 3000 seconds with a time step of 1 second. The simulation results are shown in the figure 3.



(a) 1-5



(b) 6-10

Fig. 3. System Responses to Different Input Types in Virtual TES

3) SR Applied to Real Experimental TES Data: Finally, this study validated the effectiveness of the method using experimental data from a real TES system.

The experimental data were obtained from the CoSES laboratory at the Technical University of Munich. The investigated TES unit is a commercial Wolf BSP-800 water storage tank, whose structure is shown in figure 4. The experiments covered one operating mode: charging through external ports. All data were recorded at a measurement frequency of 1 Hz.

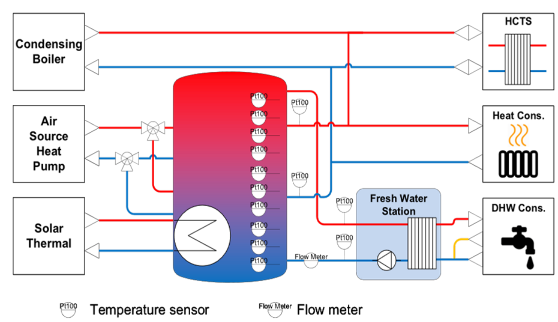


Fig. 4. The Structure of Real TES in TUM CoSES Laboratory.

The modeling process follows the same procedure as the virtual TES. Inputs include layer temperatures, inlet/outlet temperatures, and flow rate. Since there is only one dataset, the first 2/3 is used for training, while the full dataset evaluates autoregressive prediction performance.

III. Results and Analysis

A. Reconstructing the Virtual TES System Using SR

Next, the recognition results of the SR method in the virtual TES system will be presented.

Through learning from the input-output data of the virtual TES system, SR successfully derived corresponding ODEs for

979-8-3315-8138-1

TABLE I
EXPRESSIONS OF EACH LAYER OF TES

Layer	Expression
1	$dx_{1,f} = ((x_{11} - x_1) \times 0.00665759 - 1.8185820) \times x_{12}$
2	$dx_{2,f} = x_{12} \times ((x_1 - x_2) \times 0.0066624)$
3	$dx_{3,f} = x_{12} \times ((x_3 - x_2) \times -0.00665159)$
4	$dx_{4,f} = ((x_4 - x_3) \times -0.0066228) \times x_{12}$
5	$dx_{5,f} = (x_{12} \times (x_4 - x_5)) \times 0.0065444$
6	$dx_{6,f} = (((x_6 - x_5) \times -0.0066680) \times x_{12} - 1.6102656 \times 10^{-5})$
7	$dx_{7,f} = (((x_7 - x_6) \times -0.0066623) \times x_{12} - 1.6345751 \times 10^{-5})$
8	$dx_{8,f} = (-1.6630835 \times 10^{-5} - (x_{10} - x_6) \times (x_{12} \times 0.0013149))$
9	$dx_{9,f} = ((x_8 - x_{10}) \times ((x_{12} \times x_{10}) \times 0.0001371) - 1.6444343 \times 10^{-5})$
10	$dx_{10,f} = ((x_3 \times ((x_{10} - x_8) \times -2.771339 \times 10^{-5})) \times x_{12} - 1.64648411 \times 10^{-5})$

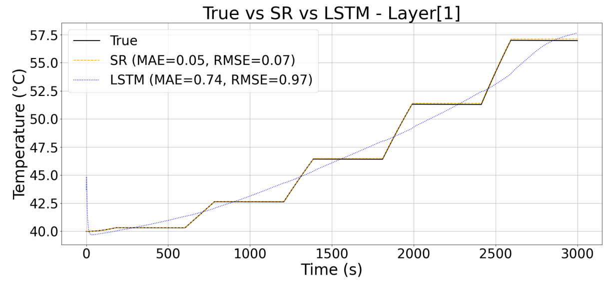
each temperature layer, where x_1 to x_{10} represent the ten temperature layers of the TES from top to bottom, x_{11} denotes the temperature of the inlet fluid, and x_{12} represents the fluid flow rate. As shown in the table I, the obtained expressions demonstrate structural conciseness and consistency. These expressions perfectly capture the core physics of the stratified TES. The dominant structure $(T_{upper} - T_{current}) \times Flow \times Constant$ accurately describes the convective heat transfer process between layers. This indicates that SR is not merely performing simple mathematical fitting, but is rediscovering the physical laws governing the system's evolution from the data. The constant terms likely comprehensively reflect system parameters such as heat loss and heat capacity.

By evaluating their predictive performance on the same training set generated via OpenModelica, we demonstrate the differences in modeling accuracy and generalization capability between the two approaches under identical dataset sizes. For the Test Sets, this study will employ two distinct approaches to generate simulation data: one using inputs with identical patterns but different parameters, and another using inputs with entirely different patterns.

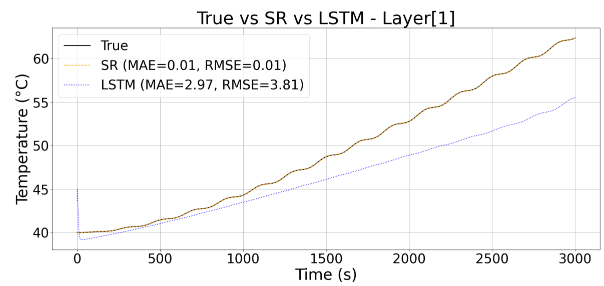
To quantify the performance of SR, we conducted a rigorous comparison against a long-short-term-memory(LSTM) based neural network benchmark model. At each time step, the input of benchmark consists of the initial temperature profile of the TES, $y_0 \in \mathbb{R}^{10}$, and the current control input $x_t = [T_{\text{in}}, \dot{m}] \in \mathbb{R}^2$, where T_{in} represents the inlet temperature and $\dot{m}$ denotes the mass flow rate. These components are concatenated to form a 12-dimensional input vector. The model receives the entire sequence of such inputs and outputs the predicted temperature at each time step. The initial TES state remains constant throughout the sequence, making the model purely input driven. The outputs $y_t \in \mathbb{R}^{10}$ represent the predicted TES temperatures at each layer over time.

The network consists of two stacked LSTM layers, each with 128 hidden units, followed by a fully connected layer that maps the hidden state to a 10-dimensional output. MSE is used as the loss function and the model is trained using the Adam optimizer with a learning rate of 10^{-3} for 200 epochs.

In the simulation, the top layer exhibits the most drastic temperature changes. To better highlight prediction deviations, the first layer's results are presented as a typical example.



(a) Test Set 1



(b) Test Set 2

Fig. 5. Comparison of LSTM and SR in Two Different Datasets

As shown in figure 5, in Test Set 1, the predictions of the SR model almost perfectly overlap with the "ground truth" output (black dotted line) of the virtual system, achieving near-perfect reconstruction. In contrast, while the LSTM model captures the overall trend, it shows significant deviations in dynamic details and fluctuation amplitudes.

In Test Set 2, the SR model demonstrates exceptional generalization capability, maintaining close alignment with the true values even under different input patterns. Conversely, the prediction error of the LSTM model increases substantially, indicating that it primarily learned specific patterns from the training data rather than the essential dynamics of the system, whereas the SR method does the opposite.

B. Using SR to Fully Reconstruct the Real TES System

To validate the modeling capability of SR in real-world noisy environments, this study applies it to typical operational dataset from an experimental thermal storage tank.

SR successfully identified dynamic models for each temperature layer. As shown in the table II, the expressions discovered by SR for most layers exhibit clear structures. Here, x_1 to x_{10} represent the ten temperature layers from bottom to top, x_{11} denotes the temperature of the inlet fluid, x_{12} represents the temperature of the outlet fluid, and x_{13} corresponds to the fluid flow rate. To validate the effectiveness

979-8-3315-8138-1

TABLE II
EXPRESSIONS OF EACH LAYER OF REAL TES

Layer	Expression
1	$dx_{1,t} = ((x2*0.0007897+0.0137505)-x1*0.00147178)$
2	$dx_{2,t} = (x3 + (-0.4399935 - x12)) * 0.00283014$
3	$dx_{3,t} = (((x2 - x4) + -0.5568835) * 0.0003230) * (((x10 * -0.0379826) * (x3 - x4) - (x5 - x4)) * -0.7263389 - 4.6373166)$
4	$dx_{4,t} = (x3 * -0.99811287 - ((x2 - x6) * -0.2030193 - x5)) * 0.0036562$
5	$dx_{5,t} = (x6 * -0.0010864 - ((x7 * -0.0007689 - x4 * 0.0002970) - -0.0043416)) * (x5 - x6)$
6	$dx_{6,t} = (x7 - x5) * 0.0022778$
7	$dx_{7,t} = (x8 - x6) * ((x8 * 1.2086919 - x6) * 0.00012549)$
8	$dx_{8,t} = (x9 - x7) * 0.0028557$
9	$dx_{9,t} = ((x11 - x8) * 7.66218453e - 5) * x13$
10	$dx_{10,t} = x13 * ((x8 - x11) * -7.50019235e - 5)$

of the obtained expressions, the study will similarly verify their accuracy in predicting the dynamic response of the TES.

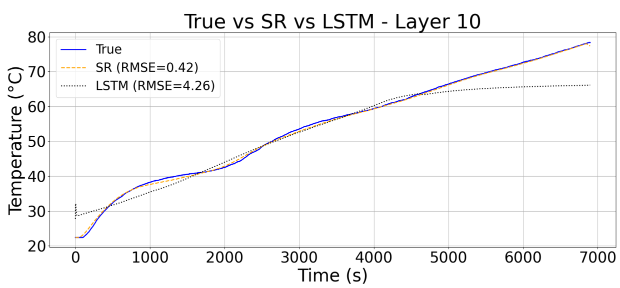


Fig. 6. Prediciton of Layer 10 of Real TES.

It can be observed that, as demonstrated in the virtual dataset phase, the LSTM method performs well within the range covered by the training set. However, beyond the scope of the training data, the predictions of LSTM deviate significantly from the true values. In contrast, the predictions of the SR method almost perfectly align with the actual temperature curves, indicating that the derived mathematical expressions accurately capture the system dynamics during the charging process.

IV. CONCLUSION

The core contribution of this thesis lies in the successful application of SR technology to the core modeling of complex dynamic energy systems, providing a novel and interpretable data-driven paradigm for the analysis and optimization of TES systems.

Traditional TES modeling relies on either complex physical derivations or uninterpretable black-box models. This study breaks through these limitations by achieving a "white-box" modeling of stratified TES systems. By automatically discovering mathematical expressions that describe the temperature dynamics of each layer directly from data, the SR model demonstrates accuracy surpassing the LSTM benchmark in both reconstruction and generalization tests on virtual systems—particularly notable given the limited dataset size. More importantly, it creates a fully transparent physical model directly analyzable and interpretable by engineers, offering a new model identification solution for engineers dealing with TES systems (e.g., some old equipment) where system parameters are hard to obtain directly and measured datasets are small.

Based on the exploration and verification of this study, the application value of the SR method is not limited to stratified TES systems; it can be further extended to more dynamic systems in the energy sector, providing new ideas and feasible approaches for the modeling and analysis of such systems.

ACKNOWLEDGMENT

This article is revised from the first author's master's thesis. He would like to express his gratitude to Sebastian Eichhorn, a PhD student at the Technical University of Munich, for providing valuable inspiration and selfless assistance during his tenure as the supervisor.

REFERENCES

[1] Sarbu, I.; Sebarchievici, C. A Comprehensive Review of Thermal Energy Storage. Sustainability 2018, 10, 191.

[2] Zinsmeister, Daniel, et al. "A benchmarking framework for energy management systems with commercial hardware models." Energy and Buildings 321 (2024).

[3] D. Zinsmeister, P. Tzscheutschler, V. S. Peri ́c, and C. Goebel. "Stratified thermal energy storage model with constant layer volume for predictive control—Formulation, comparison, and empirical validation." In: Renewable Energy 219 (2023).

[4] I. De la Cruz-Loredo, D. Zinsmeister, T. Licklederer, C. E. Ugalde-Loo, D. A. Morales, H. Bastida, V. S. Peri ́c, and A. Saleem. "Experimental validation of a hybrid 1-D multi-node model of a hot water thermal energy storage tank." In: Applied Energy 332 (2023), p. 120556.

[5] S. Kalaiselvam and R. Parameshwaran. "Chapter 11. Review on the Modeing and Simulation of Thermal Energy Storage Systems." In: Dec. 2014.

[6] H. Bastida, I. De la Cruz-Loredo, P. Saikia, and C. E. Ugalde-Loo. "Discrete-time state-of-charge estimator for latent heat thermal energy storage units based on a recurrent neural network." In: Applied Energy 371 (2024), p. 123526.

[7] Brandes, Matthias, et al. "Data-driven modeling of heat pumps and thermal storage units for MPC." Journal of Physics: Conference Series. Vol. 2600. No. 3. IOP Publishing, 2023.

[8] M. Schmidt and H. Lipson. "Distilling free-form natural laws from experimental data." In: science 324.5923 (2009), pp. 81–85.

[9] W. La Cava, B. Burlacu, M. Virgolin, M. Kommenda, P. Orzechowski, F. O. de França, Y. Jin, and J. H. Moore. "Contemporary symbolic regression methods and their relative performance." In: Advances in neural information processing systems 2021.DB1 (2021), p. 1.

[10] I. Rafael, P. A. Estévez, J. R. Vergara, and W. R. Calderón-Muñoz. "Model-based battery thermal parameter optimization using symbolic regression." In: Journal of Energy Storage 73 (2023), p. 109243.

[11] Cai, Weihua, et al. "Heat transfer correlations by symbolic regression." International Journal of Heat and Mass Transfer 49.23-24 (2006): 4352-4359.

[12] S. Gaucel, M. Keijzer, E. Lutton, and A. Tonda. "Learning Dynamical Systems Using Standard Symbolic Regression." In: Genetic Programming. Ed. by M. Nicolau, K. Krawiec, M. I. Heywood, M. Castelli, P. García-Sánchez, J. J. Merelo, V. M. Rivas Santos, and K. Sim. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 25–36.

[13] M. Cranmer. "Interpretable machine learning for science with PySR and SymbolicRegression. jl." In: arXiv preprint arXiv:2305.01582 (2023).

979-8-3315-8138-1