

Discovering symbolic expressions with parallelized tree search

Kai Ruan¹, Ze-Feng Gao¹, Yike Guo², Hao Sun^{1,3,*}, Ji-Rong Wen^{1,3,*}, Yang Liu^{4,5,*}

¹Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

²Department of Computer Science and Engineering, HKUST, Hong Kong, China

³Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China

⁴School of Engineering Science, University of Chinese Academy of Sciences, Beijing, China

⁵State Key Laboratory of Nonlinear Mechanics, Institute of Mechanics, Chinese Academy of Sciences, Beijing, China

*Corresponding authors

Abstract

Symbolic regression plays a crucial role in modern scientific research thanks to its capability of discovering concise and interpretable mathematical expressions from data. A grand challenge lies in the arduous search for parsimonious and generalizable mathematical formulas, in an infinite search space, while intending to fit the training data. Existing algorithms have faced a critical bottleneck of accuracy and efficiency over a decade when handling problems of complexity, which essentially hinders the pace of applying symbolic regression for scientific exploration across interdisciplinary domains. To this end, we introduce a parallelized tree search (PTS) model to efficiently distill generic mathematical expressions from limited data. Through a series of extensive experiments, we demonstrate the superior accuracy and efficiency of PTS for equation discovery, which greatly outperforms the state-of-the-art baseline models on over 80 synthetic and experimental datasets (e.g., lifting its performance by up to 99% accuracy improvement and one-order of magnitude speed up). PTS represents a key advance in accurate and efficient data-driven discovery of symbolic, interpretable models (e.g., underlying physical laws) and marks a pivotal transition towards scalable symbolic learning.

Introduction

Over the centuries, scientific discovery has never departed from the use of interpretable mathematical equations or analytical models to describe complex phenomena in nature. Pioneering scientists discovered that behind many sets of empirical data in the real world lay succinct governing equations or physical laws. A famous example of this is Kepler's discovery of three laws of planetary motion using Tycho Brahe's observational data, which laid the foundation for Newton's discovery of universal gravitation. Automated extraction of these natural laws from data, as a class of typical symbolic regression (SR) problems [1], stands at the forefront of data-driven scientific exploration in natural sciences and engineering applications [2–10]. However, uncovering parsimonious closed-form equations that govern complex systems (e.g., nonlinear dynamics) is always challenging. Data revolution, rooted in advanced machine intelligence, has offered an alternative to tackle

this issue. Although well-known regression methods [11] have been widely applied to identify the coefficients of given equations in fixed forms, they are no longer effective for natural systems where our prior knowledge of the explicit model structure is vague.

Attempts have been made to develop evolutionary computational methods to uncover symbolic formulas that best interpret data. Unlike traditional linear/nonlinear regression methods which fit parameters to equations of a given form, SR allows free combination of mathematical operators to obtain an open-ended solution and thus makes it possible to automatically infer an analytical model from data (e.g., by simultaneously discovering both the form of equations and controlling parameters). Monte Carlo sampling [12] and evolutionary algorithms, such as genetic programming (GP), have been widely applied to distill mathematical expressions and governing laws that best fit available measurement data for nonlinear dynamical systems [1, 13–18]. However, this type of approach is known to scale poorly to problem’s dimensionality, exhibits sensitivity to hyperparameters, and generally suffers from extensive computational cost in an extensively large search space.

Another remarkable breakthrough leverages sparse regression, in a restricted search space based on a pre-defined library of candidate functions, to select an optimal analytical model [19]. Such an approach quickly became one of the state-of-art methods and kindled significant enthusiasm in data-driven discovery of ordinary or partial differential equations [9, 20–26] as well as state estimation [27]. However, the success of this sparsity-promoting approach relies on a properly defined candidate function library that operates on a fit-complexity Pareto front. It is further restricted by the fact that the linear combination of candidate functions is usually insufficient to express complicated mathematical formulas. Moreover, when the library size is overly massive, this approach generally fails to hold the sparsity constraint.

Deep learning has also been employed to uncover generic symbolic formulas from data, e.g., recurrent neural networks with risk-seeking policy gradient formulation [28], variational grammar autoencoders [29], neural-network-based graph modularity [30, 31], pre-trained transformers [32–34], etc. Another notable work has leveraged symbolic neural networks to distill physical laws of dynamical systems [35, 36], where commonly seen mathematical operators are employed as symbolic activation functions to establish intricate formulas via weight pruning. Nevertheless, this framework is primarily built on empirical pruning of the weights, thus exhibits sensitivity to user-defined thresholds and may fall short to produce parsimonious equations for noisy and scarce data.

Very recently, Monte Carlo tree search (MCTS) [37, 38], which gained acclaim for powering the decision-making algorithms in AlphaGo [39], AlphaZero [40] and AlphaTensor [41], has shown a great potential in navigating the expansive search space inherent in SR [42]. This method uses stochastic simulations to meticulously evaluate the merit of each node in the search tree and has empowered several SR techniques [43, 44]. However, the conventional application of MCTS maps each node to a unique expression, which tends to impede the rapid recovery of accurate symbolic representations. This narrow mapping may limit the strategy’s ability to efficiently parse through the complex space of potential expressions, thus presenting a bottleneck in the quest for swiftly uncovering underlying mathematical equations.

SR involves evaluating a large number of complex symbolic expressions composed of various operators, variables, and constants. The operators and variables are discrete, while the value of the coefficients is continuous. The NP-hardness of a typical SR process, noted by many scholars [28, 30, 45], has been formally established [46]. This nature makes the algorithm need to traverse various possible combinations, resulting in a huge and even infinite search space and thus facing the problem of combinatorial explosion. Unfortunately, all the existing SR methods evaluate each candidate expression independently, leading to significantly low computational efficiency, and although some works considered caching evaluated expressions to prevent recomputation, they do not reuse these results as subtree values for evaluating deeper expressions. Consequently, the majority of these

methods either rely on meta-heuristic search strategies, narrow down the search space based on specific assumptions, or incorporate pre-trained models to discover relatively complex expressions, which make the algorithms prone to producing specious results (e.g., local optima). Therefore, the efficiency of candidate expression evaluation is of paramount importance. By enhancing the efficiency of candidate expression evaluation, it becomes possible to design new SR algorithms that are less reliant on particular optimization methods, while increasing the likelihood of directly finding the global optima in the grand search space. Thus, we can improve the SR accuracy (in particular, the symbolic recovery rate) and, meanwhile, drastically reduce the computational time.

To this end, we propose a novel parallelized tree search (PTS) model to automatically distill symbolic expressions from limited data. Such a model is capable of efficiently evaluating potential expressions, thereby facilitating the exploration of hundreds of millions of candidate expressions simultaneously in parallel within a mere few seconds. In particular, we propose a parallel symbolic regression network (PSRN) as the cornerstone search engine, which (1) automatically captures common subtrees of different math expression trees for shared evaluation that avoids redundant computations, and (2) capitalizes on graphics processing unit (GPU) based parallel search that results in a notable performance boost. It is notable that, in a standard SR process for equation discovery, many candidate expressions share common subtrees, which leads to repeatedly redundant evaluations and, consequently, superfluous computation. To address this issue, we propose a strategy to automatically cache common subtrees for shared evaluation and reuse, effectively circumventing significantly redundant computation. We further execute PTS on a GPU to perform large-scale evaluation of candidate expressions in parallel, thereby augmenting the efficiency of the evaluation process. To our astonishment, the synergy of these two techniques could yield up to four orders of magnitude efficiency improvement in the context of expression evaluation. In addition, to expedite the convergence and enhance the capacity of PSRN for exploration and identification of more intricate expressions, we amalgamate it with the MCTS strategy which identifies a set of admissible base expressions as tokenized input. The remarkable efficacy and efficiency of the proposed PTS model have been demonstrated on a variety of benchmark and lab test datasets. The results show that PTS surpasses evidently several existing baseline methods, achieving higher symbolic recovery rate and efficiency.

Results

Symbolic expression discovery with parallel evaluation

SR aims to discover concise, precise, and human-interpretable mathematical expressions hidden within data, offering significant value in aiding scientists to decipher the underlying meanings of unknown systems. Mathematically speaking, SR can be cast as the process of finding an expression $f : \mathbb{R}^m \rightarrow \mathbb{R}$ that satisfies $\mathbf{y} = f(\mathbf{X})$ given data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, where $\mathbf{X} \in \mathbb{R}^{n \times m}$, $\mathbf{y} \in \mathbb{R}^{n \times 1}$, n represents the number of data samples and m the number of independent variables. Here, the form of f is usually constructed by a finite set of math symbols based on a given token library $\mathcal{O} = \{+, -, \times, \div, \sin(\cdot), \exp(\cdot), \dots, \text{const.}\}$. To make the underlying expression interpretable and parsimonious, SR algorithms are required to swiftly produce a Pareto front, balancing the error and complexity in underlying expressions. In other words, there is a need for a new SR algorithm that is both computationally efficient and exhibits a high symbolic recovery rate (e.g., also referred as accuracy; see [Supplementary Note 2.4](#) for specific definition). However, existing SR algorithms are faced with significant bottlenecks associated with low efficiency and poor accuracy in finding complex mathematical expressions. To this end, we propose a novel parallelized tree search (PTS) model, as shown in Fig. 1, to automatically discover mathematical expressions from limited data.

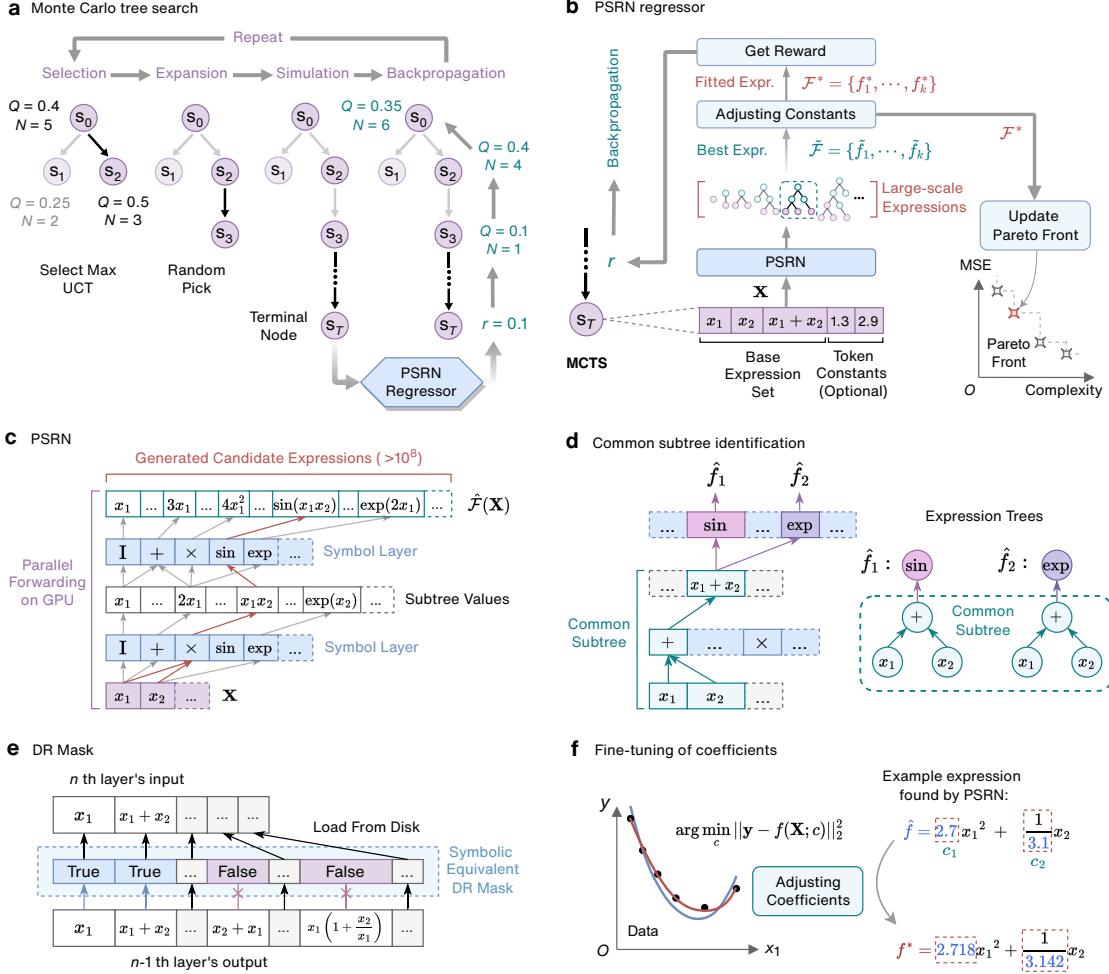


Fig. 1: Overview of the proposed PTS model. **a**, Monte Carlo tree search (MCTS) for automatic discovery of admissible tokenized input. Each node in the MCTS model represents a base expression set denoted as s_i . The search begins from the root node and iteratively performs selection, expansion, simulation, and backpropagation steps until a stopping condition is met or the specified number of epochs is completed. Once a terminal node is reached, the corresponding base expression set s_T is fed into PSRN regressor for expression discovering. The obtained reward r is then used to guide the MCTS backpropagation. **b**, Schematic of PSRN regressor for discovering optimal expressions from data. The base expression set s_T from terminal nodes, along with optionally sampled token constant values, is fed into PSRN for forward computation. After evaluating the error of large-scale candidate expressions, PSRN provides the optimal (or top- k) candidate expressions denoted as $\tilde{\mathcal{F}}$. Subsequently, the least squares method is employed for the identification and fine-tuning of the coefficients, resulting in adjusted expressions \mathcal{F}^* , which are then utilized for updating the Pareto front as computing the complexity and reward. **c**, Forward computation in PSRN. The base expression set and sampled constants, along with the corresponding data tensor \mathbf{X} , are fed into the network. The network comprises multiple Symbol Layers (e.g., typically three, but only two are shown for simplicity). Each layer provides all possible subtree values resulting from one-operator computations among all input expressions. This process can be efficiently parallelized on a GPU for rapid computation. Upon completion of PSRN’s forward computation, a myriad of candidate expressions $\hat{\mathcal{F}}$ (e.g., up to hundreds of millions) corresponding to the base expression set s_T , along with their associated error tensors on the given data, are generated. Then, the expression with the minimal error (or top- k expressions) will be selected. **d**, Schematic of common subtree identification. Expressions sharing common subtrees, such as $\sin(x_1 + x_2)$ and $\exp(x_1 + x_2)$, leverage identical subtree computation outcomes. This approach circumvents extensive redundant calculations, thereby increasing the efficiency of SR. **e**, Schematic of duplicate removal mask (DR Mask). We designed the DR Mask layer to mask the output tensor of the penultimate layer in PSRN, thereby excluding subtree expressions that are symbolically equivalent. This significantly reduces PSRN’s usage of GPU memory. **f**, Estimation and fine-tuning of constant coefficients. The least squares method is used to fine-tune the coefficients of the preliminary expressions discovered by PSRN using the complete data set.

In particular, we develop a PSRN regressor as the core search engine, depicted in Fig. 1**b–c**, to enhance significantly the efficiency of candidate expression evaluation. The architecture of PSRN is designed parallelism-friendly, thereby enabling rapid GPU-based parallel computation (see Fig. 1**c**). The PSRN is empowered by automatically identifying and reusing intermediate calculation of common subtrees (see Fig. 1**d**).

To further bolster the model’s discovery capability, we also integrate MCTS to locate a set of admissible base expressions as token input to PSRN for exploring deeper expressions (see Fig. 1**a**). Specifically, the PTS model revolves around PSRN continually activating MCTS iterations, starting with a root node with several available independent variables. During these iterations, the set of base expressions within each node are expanded by progressively incorporating more complex base expressions. When the search reaches a terminal node, the token constants are sampled and, together with the base expression set, are input into PSRN for evaluation and search of optimal symbolic expressions (see Fig. 1**b**). Typically, PSRN can rapidly (e.g., within a matter of seconds) identify the expression with the smallest error or a few best candidates from hundreds of millions of symbolic expressions. This represents a significant speed improvement compared to existing approaches which independently evaluate candidate expressions. Additionally, we design a duplicate removal mask step (e.g., DR Mask) for PSRN to reduce memory usage (see Fig. 1**e**). In the PSRN regressor, the coefficients of the most promising expressions are identified and fine-tuned based on least squares estimation (see Fig. 1**f**). The rewards on the given data are computed and then back-propagated for subsequent MCTS searches. As the search progresses, the Pareto front representing the optimal set of expressions is continuously updated to report the final result. Further details of the proposed model are provided in [Methods](#).

Symbolic regression benchmarks

We firstly demonstrate the efficacy of PTS on recovering specified mathematical formulas given multiple benchmark problem sets (including Nguyen [47], Nguyen-c [48], R [49], Livermore [45] and Feynman [30], as described in [Supplementary Note 2.1](#)), commonly used to evaluate the performance of SR algorithms. Each SR puzzle consists of a ground truth equation, a set of available math operators, and a corresponding dataset. These benchmark data sets contains various math expressions, e.g., $x^3 + x^2 + x$ (Nguyen-1), $3.39x^3 + 2.12x^2 + 1.78x$ (Nguyen-1c), $(x + 1)^3/(x^2 - x + 1)$ (R-1), $1/3 + x + \sin(x^2)$ (Livermore-1), $x_1^4 - x_1^3 + x_1^2 - x_2$ (Livermore-5), and $x_1x_2x_3 \log(x_5/x_4)$ (Feynman-9), which are listed in detail in [Supplementary Tables S.1–S.2](#). Our objective is to uncover the Pareto front of optimal mathematical expressions that balance the equation complexity and error. The performance of PTS is compared with four baseline methods, e.g., symbolic physics learner (SPL) [42], neural-guided genetic programming (NGGP) [45], deep generative symbolic regression (DGSR) [34] and Pysr [50]. For the Nguyen-c dataset, each model is run for at least 20 independent trials with different random seeds, while for all other puzzles, 100 distinct random seeds are utilized. We mark the successful case if the ground truth equation lies in the discovered Pareto front set.

The SR results of different models, in terms of the symbolic recovery rate and computational time, are depicted in Fig. 2. It can be seen that the proposed PTS method evidently outperforms the baseline methods for all the benchmark problem sets in terms of recovery rate, meanwhile maintaining the highest efficiency (e.g., expending the minimal computation time) that achieves up to two orders of magnitude speedup. The detailed results for each SR problem set are listed in [Supplementary Tables S.7–S.12](#). An intriguing finding is that PTS achieves an impressive symbolic recovery rate of 99% on the R benchmark expressions, while the baseline models almost fail (e.g., recovery rate < 2%). We attribute this to the mathematical nature of the R benchmark expressions, which are all rational fractions like $(x + 1)^3/(x^2 - x + 1)$ (R-1). Confined to the sampling interval

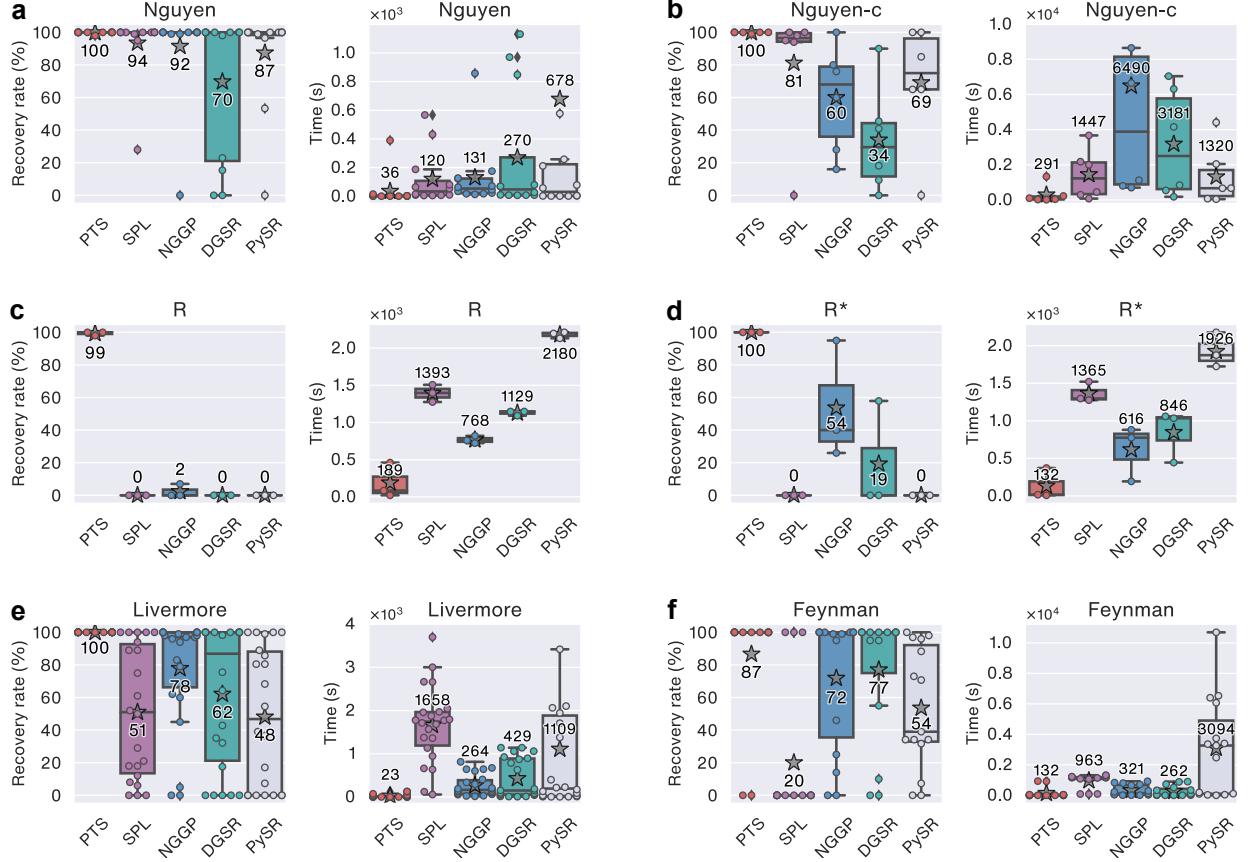


Fig. 2: Performance of various models on symbolic regression benchmarks. We employ two evaluation metrics, namely, symbolic recovery rate and average runtime, to assess the performance of each algorithm. Our PTS approach achieves the highest recovery rate across various benchmark problem sets, meanwhile expending the minimal computation time, compared to the baseline methods (e.g., SPL [42], NGGP [45], DGSR [34] and PySR [50]). This highlights the substantial superiority of PTS. Note that the star marks the mean value over each problem set.

$x \in [-1, 1]$, the properties of the R expressions bear a high resemblance to polynomial functions, resulting in intractable local minima that essentially lead to the failure of NGGP and DGSR. This issue can be alleviated given a larger interval, e.g., $x \in [-10, 10]$, as illustrated in the R* dataset (see [Supplementary Tables S.10](#)). Notably, the performance of DGSR based on pre-trained language models stems from the prevalence of polynomial expressions in the pre-training corpora, while NGGP collapses on account of its limited search capacity in an enormously large search space. In contrast, owing to the direct and parallel evaluation of multi-million expression structures, PTS possesses the capability of accurately and efficiently recovering complex expressions.

Discovery of chaotic dynamics

Nonlinear dynamics is ubiquitous in nature and typically governed by a set of differential equations. Distilling such governing equations from limited observed data plays a crucial role in better understanding the fundamental mechanism of dynamics. Here, we test the proposed PTS model to discover a series of multi-dimensional autonomous chaotic dynamics (e.g., Lorenz attractor [51] and its variants [52]). The synthetic datasets of these chaotic dynamical systems are described in [Supplementary Note 2.2](#). It is noted that we only measure the noisy trajectories (e.g., with 1%

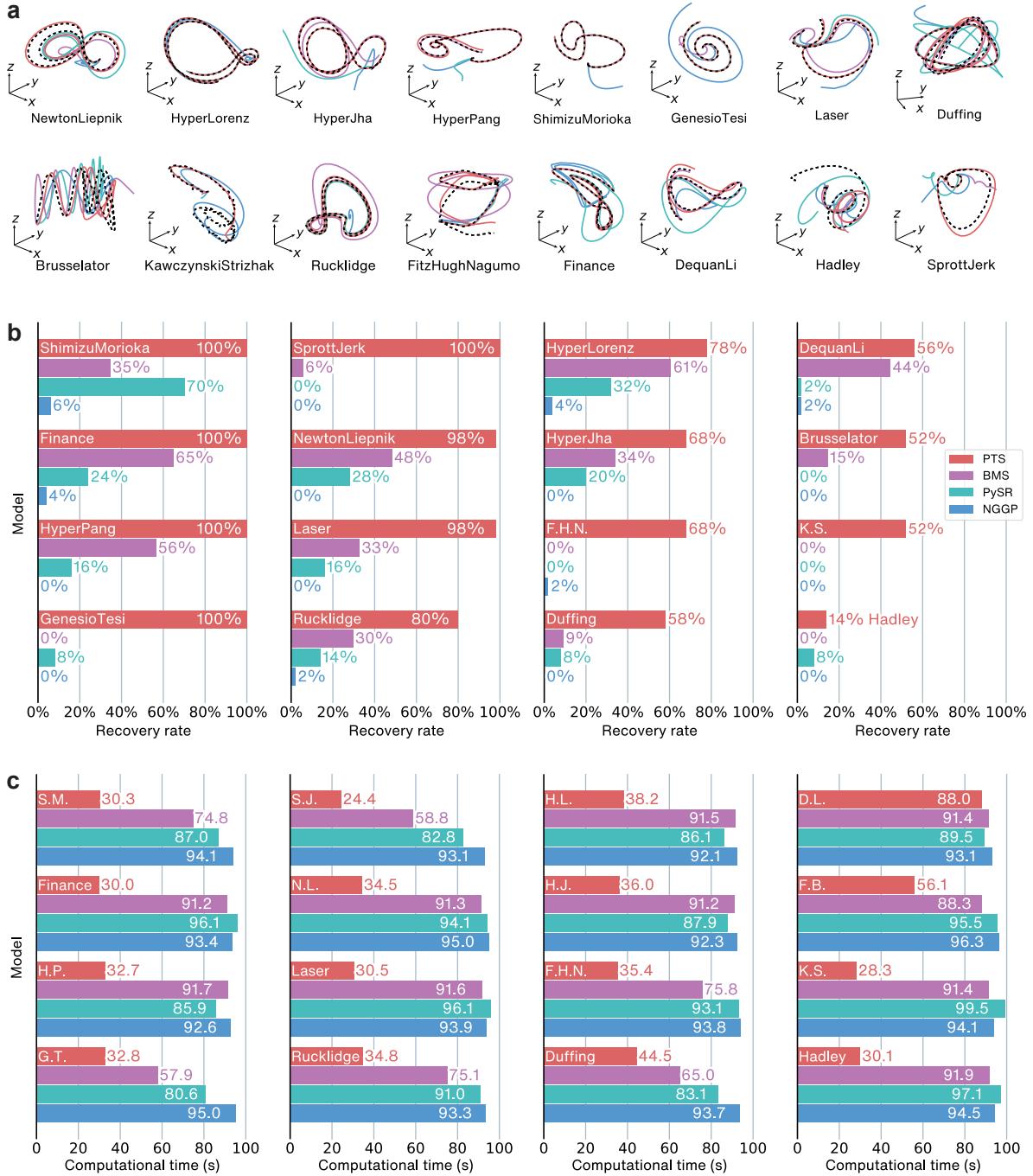


Fig. 3: Data-driven discovery of nonlinear chaotic dynamics by different models including PTS, BMS, PySR and NGGP. A large number of operators (e.g., $+$, \times , $-$, \div , \sin , \cos , \exp , \cosh , \tanh , abs , sign , etc.) are used to simulate the situation in which scientists explore unknown systems in the real world with no prior knowledge. Given the same budget of computational time, PTS has a higher probability of finding the true governing equations among a nearly infinite number of possible expression structures. Each model's runtime is capped at around 100 seconds by limiting the number of iterations. **a**, Examples of predicted trajectories (e.g., solid lines) by different models compared with the ground truth (e.g., dashed lines). **b**, Average recovery rates of SR algorithms on 16 nonlinear chaotic dynamics datasets. Since the dynamics of each system is governed by multiple coupled differential equations, the discovery is conducted independently to compute the average recovery rate for each equation, among which the minimum rate is taken to represent each model's overall capability. **c**, Average computational time of SR algorithms on 16 nonlinear chaotic dynamics datasets.

Gaussian noise added to the clean data) while determining the velocity states via smoothed numerical differentiation. We compare PTS with three pivotal baseline models (namely, Bayesian machine scientist (BMS) [12], PySR [50] and NGGP [45]) and run each model on 50 different random seeds to calculate the average recovery rate for each dataset. Here, we dropped DGSR [34] for comparison given its poor performance in the previous benchmark tests, and included BMS for comparison since it is specifically designed to perform discovery of equations with coefficients. Considering the noise effect, the criterion for successful equation recovery in this experiment is defined as follows: the discovered Pareto front covers the structure of the ground truth equation (allowing for a constant bias term). Since the dynamics of each system is governed by multiple coupled differential equations (e.g., 3~4 as shown in [Supplementary Table S.3–S.6](#)), the discovery is conducted independently to compute the average recovery rate for each equation, among which the minimum rate is taken to represent each model’s overall capability.

Our main focus herein is to investigate whether SR methods can successfully recover the underlying differential equations without any *a priori* knowledge under the limit of a short period of computational time (e.g., one minute). In our experiments, we set the candidate binary operators as $+$, $-$, \times , and \div , while the candidate unary operators include \sin , \cos , \exp , \log , \tanh , \cosh , abs , and sign . Fig. 3 depicts the results of discovering the closed-form governing equations for 16 chaotic dynamical systems. The experiments demonstrate that the proposed PTS approach can achieve a much higher symbolic recovery rate (see Fig. 3b), while maintaining a much lower computational cost as depicted in Fig. 3c), enabling to identify more accurately the underlying governing equations, even under noise effect, to better describe the chaotic behaviors (see Fig. 3a). This substantiates the capability and efficiency of our method on data-driven discovery of governing laws for more complex chaotic dynamical systems beyond the Lorenz attractor. More detailed results are listed in [Supplementary Fig. S.3](#) and [Supplementary Table S.13](#).

Electro-mechanical positioning system

Real-world data, replete with intricate noise and nonlinearity, may hinder the efficacy of SR algorithms. To further validate the capability of our PTS model in uncovering the governing equations for real-world dynamical systems (e.g., mechanical devices), we test its performance on a set of lab experimental data of an electro-mechanical positioning system (EMPS) [53], as shown in Fig. 4a–b. The EMPS setup is a standard configuration of a drive system used for prismatic joints in robots or machine tools. Finding the governing equation of such a system is crucial for designing better controllers and optimizing system parameters. The dataset was bifurcated into two even parts, serving as the training and testing sets, respectively. The reference governing equation is given by $M\ddot{q} = -F_v\dot{q} - F_c\text{sign}(\dot{q}) + \tau - c$ [53], where q , \dot{q} , and \ddot{q} represent joint position, velocity, and acceleration. Here, τ is the joint torque/force; c , M , F_v , and F_c are all constant parameters in the equation. In EMPS, there exists friction that dissipates the system energy. Based on this prior knowledge, we include the sign operator to model such a mechanism. Hence, the candidate math operators we use to test the SR algorithms read $\{+, -, \times, \div, \sin, \cos, \exp, \log, \text{sign}\}$.

We compare our PTS model with three pivotal baseline models, namely, PySR [50], NGGP [45], and BMS [12]. We execute each SR model 20 trials on the training dataset to ascertain the Pareto fronts. Subsequently, we select the discovered equation from the candidate expression set of each Pareto front based on the test dataset which exhibits the highest reward value delineated in Eq. (8). The reward is designed to balance the prediction error and the complexity of the discovered equation, which is crucial to derive the governing equation that is not only as consistent with the data as possible but also parsimonious and interpretable. Finally, we select among 20 trials the discovered equation with the median reward value to represent each SR model’s average performance. The

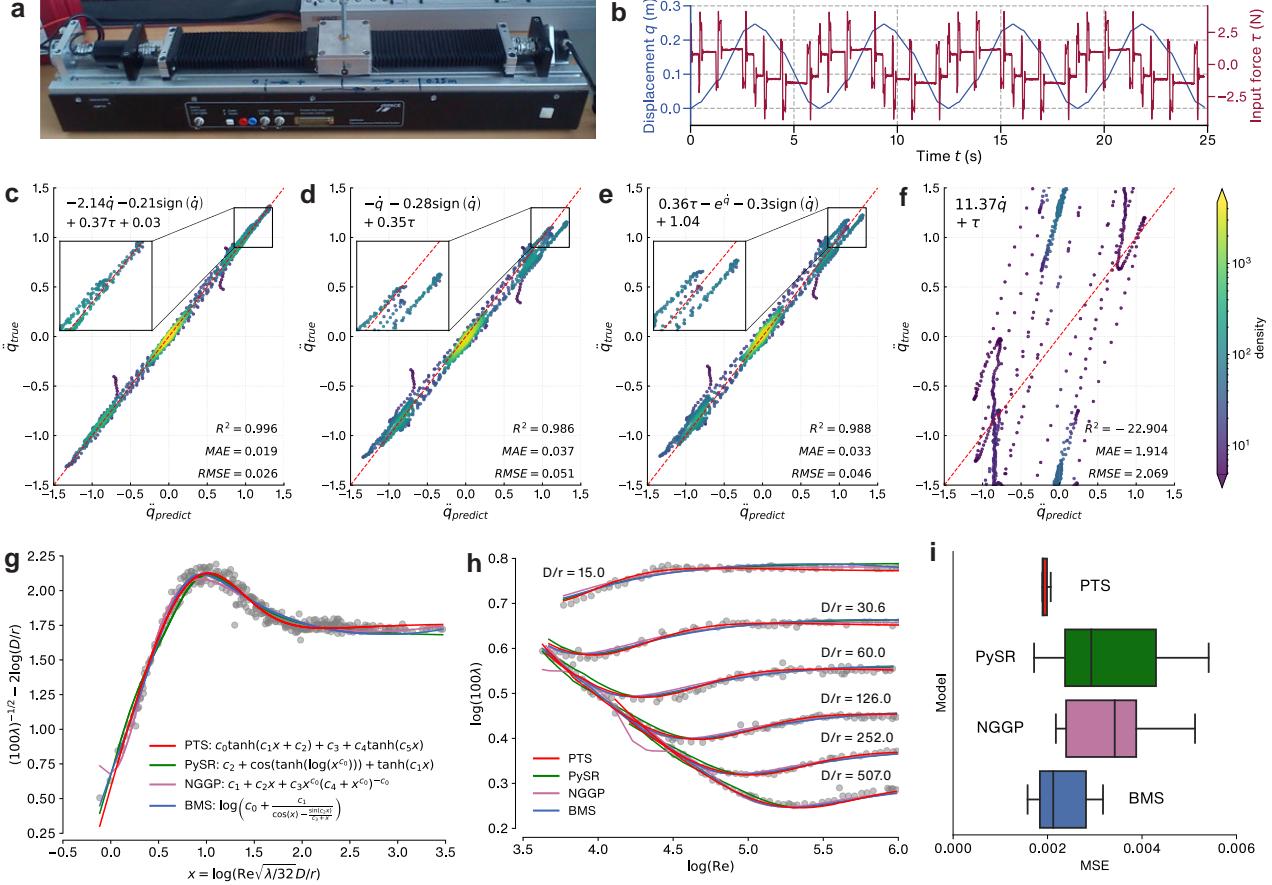


Fig. 4: Discovering the underlying physical laws with experimental data. **a**, The setup of the EMPS experiment. **b**, Collected displacement and input force data. **c**, The prediction performance along with the governing equation discovered by PTS. **d**, The prediction performance along with the governing equation discovered by PySR. **e**, The prediction performance along with the governing equation discovered by BMS. **f**, The prediction performance along with the governing equation discovered by NGGP. **g**, The transformed (or collapsed) Nikuradse's dataset and discovered equations of turbulent friction by different SR methods. The legend shows the median reward models (see Eq. (8)), among 20 trials, obtained by PTS (red), PySR (green) [50], NGGP (purple) [45], and BMS (blue) [12]. **h**, The fitting performance of each SR method on the original Nikuradse's data. **i**, The prediction mean square error (MSE) of each model. Notably, our PTS method excels at fitting turbulent friction data within a predefined time budget of 1.5 minutes, meanwhile discovering a more parsimonious equation.

runtime of each model is limited to around 1.5 minutes. The results demonstrate that our PTS model achieves the best performance in successfully discovery of the underlying governing equation (see Fig. 4c–f).

Governing equation of turbulent friction

Uncovering the intrinsic relationship between fluid dynamics and frictional resistance has been an enduring pursuit in the field of fluid mechanics, with implications spanning engineering, physics, and industrial applications. In particular, one fundamental challenge lies in finding a unified formula to quantitatively connect the Reynolds number (Re), the relative roughness r/D , and the friction factor λ , based on experimental data. The Reynolds number, a dimensionless quantity, captures the balance between inertial and viscous forces within a flowing fluid, which is a key parameter in determining the flow regime, transitioning between laminar and turbulent behaviors. The relative

roughness, a ratio between the size of the irregularities and the radius of the pipe, characterizes the interaction between the fluid and the surface it flows over. Such a parameter has a substantial influence on the flow's energy loss and transition to turbulence. The frictional force, arising from the interaction between the fluid and the surface, governs the dissipation of energy and is crucial in determining the efficiency of fluid transport systems. The groundbreaking work [54] dated in the 1930s stepped out the first attempt by meticulously cataloging in the lab the flow behavior under friction effect. The experimental data, commonly referred to as the Nikuradse dataset, offers insights into the complex interplay between these parameters (Re and r/D) and the resultant friction factor (λ) in turbulent flows. We herein test the performance of the proposed PTS model in uncovering the underlying law that governs the relationship between fluid dynamics and frictional resistance based on the Nikuradse dataset.

We firstly transform the data by a data collapse approach [55], a common practice used in previous studies [56, 57]. Our objective is to find a parsimonious closed-form equation given by $\bar{\lambda} = \bar{h}(x)$, where $\bar{\lambda} = \lambda^{-1/2} + 2 \log(r/D)$ denotes the transformed friction factor, $x = Re\sqrt{\lambda/32}(D/r)$ an intermediate variable, and \bar{h} the target function to be discovered. We consider three baseline models for comparison, namely, PySR [50], NGGP [45], and BMS [12]. The candidate operators used in these models read $\{+, \times, -, \div, \sin, \cos, \exp, \log, \tanh, \cosh, \square^2, \square^3\}$. We run each SR model for 20 independent trials with different random seeds under the time budget of 1.5 minutes and choose the identified expression with the median reward as the representative result. For each trial, we report the expression with the highest reward (see Eq. (8)) on the discovered Pareto front. Fig. 4g illustrates discovered governing equations for turbulent friction. The fitting performance of each SR method and the prediction error distribution are shown in Fig. 4h-i. It can be observed that our PTS model achieves the best performance.

Model performance analysis

The performance of the proposed PTS model depends on several factors, including the noise level of the given data, model hyper-parameters, and whether certain modules are used. Herein, we present an analysis of these factors as well as the model efficiency and the memory footprint.

Model ablation study. We conduct three cases of model ablation study to evaluate the role of certain modules. First, we investigate how much improvement the use of MCTS for automatic discovery of admissible tokenized input brings. Second, we conduct a sensitivity analysis of the token constants range. Third, we investigate the extent of the benefit of using DR Mask. The results of the ablation experiments are shown in Fig. 5a-c.

Firstly, we evaluate the symbolic recovery rate to observe the impact of replacing MCTS with random generation of input tokens for PSRN. The tests are conducted on the Nguyen-7/12, R-1/2/3 and Livermore-1/3/5/12/13/15/18/22 benchmark expressions, selected because of their higher complexity. It can be observed in Fig. 5a that the recovery rate diminishes if MCTS is removed from PTS, indicating its vital role in admissible token search that signals the way forward for expression exploration. Secondly, to investigate the sensitivity of our model to the randomly sampled token constants, we set the range to $[0, 1]$, $[0, 3]$, and $[0, 10]$ respectively, and performed the experiments on the Nguyen-c benchmark expressions. The result in Fig. 5b shows that when the token constants are sampled from the range excluding the ground truth, the model needs to spend extra search effort to retain the accuracy (e.g., recovery rate). Last but not least, we test the efficacy of DR Mask for memory saving. The result in Fig. 5c illustrates that DR Mask is able to save the graphic memory around 50%, thus improving the capacity of the operator set (e.g., more operators could be included to represent complex expressions).

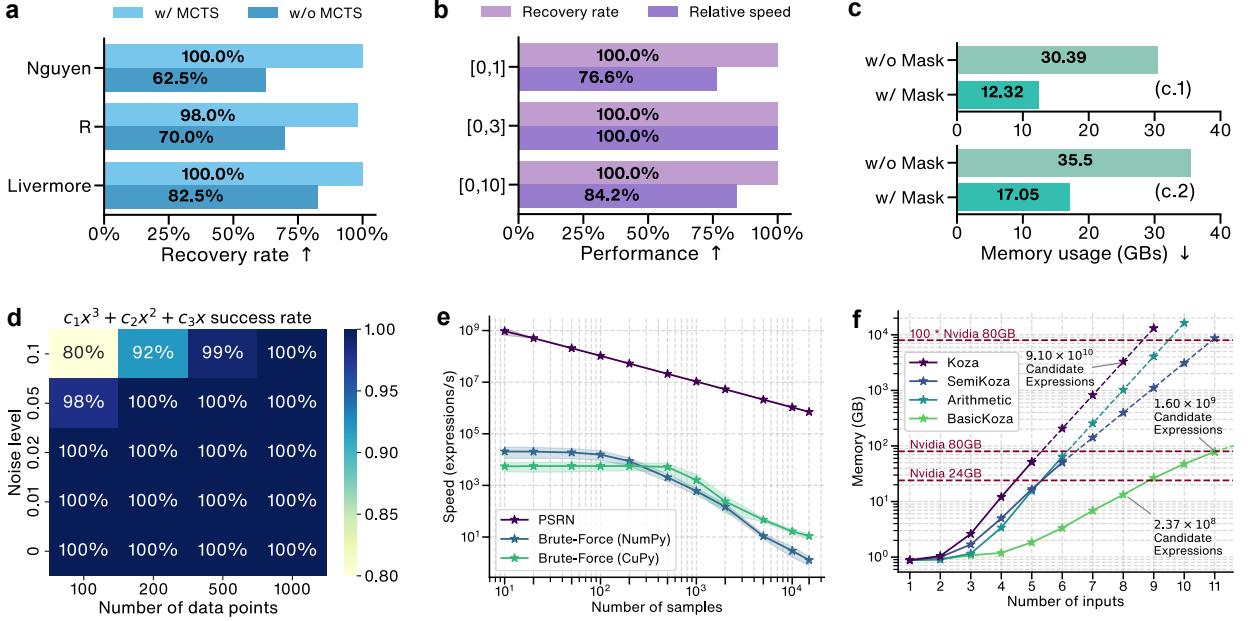


Fig. 5: Ablation study of the proposed PTS model. **a**, Ablation of MCTS. **b**, Ablation of token constants range. **c**, Ablation of DR Mask on two different PTS configurations: (c.1) a 4-input, 3-layer PSRN with $\mathcal{O}_{\text{Koza}}$ library (e.g., $\{+, \times, -, \div, \text{identity}, \sin, \cos, \exp, \log\}$), and (c.2) a 5-input, 3-layer PSRN with $\mathcal{O}_{\text{SemiKoza}}$ library (e.g., $\{+, \times, \text{SemiSub}, \text{SemiDiv}, \text{identity}, \text{neg}, \text{inv}, \sin, \cos, \exp, \log\}$). **d**, Robustness of PTS to noise. **e**, Expression search efficiency of the PSRN module. **f**, Space complexity of PSRN with three Symbol Layers, with respect to the number of input slots and memory footprints. We incorporate four operator sets: $\mathcal{O}_{\text{Koza}}$, $\mathcal{O}_{\text{SemiKoza}}$, $\mathcal{O}_{\text{Arithmetic}}$ (e.g., $\{+, \times, -, \div, \text{identity}\}$), and $\mathcal{O}_{\text{BasicKoza}}$ (e.g., $\{+, \times, \text{identity}, \text{neg}, \text{inv}, \sin, \cos, \exp, \log\}$) for comparison. With the expansion of the memory footprint, our model demonstrates scalability by evaluating a greater number of candidate expressions in a single forward pass, leading to enhanced performance.

Robustness to noise. To test the robustness of our PTS model to measurement noise, we conducted experiments with different levels of noise (e.g., Gaussian-type) and data availability for the target equation $f(x) = 0.3x^3 + 0.5x^2 + 2x$ where $x \in [-1, 1]$ [42]. The operators used were $\{+, -, \times, \div\}$. Since noise may cause inaccurate constant fitting, our evaluation criterion is set to consider the equation uncovered successfully as long as the correct equation form is found. The heatmap in Fig. 5d shows the effectiveness of our model, which indicates that PTS has a fairly high level of robustness against data noise and scarcity.

Expression search efficiency. We compare the efficiency of PTS in the context of evaluation of large-scale candidate expressions, in comparison with two brute-force methods (e.g., NumPy [58] that performs CPU-based evaluation serially, and CuPy [59] that operates on GPUs in parallel based on batches). Note that PTS possesses the capability of automatic identification and evaluation of common subtrees, while NumPy and CuPy do not have such a function. Assuming independent input variables $\{x_1, \dots, x_5\}$ with operators $\{+, \times, \text{identity}, \text{neg}, \text{inv}, \sin, \cos, \exp, \log\}$ for a maximum tree depth of 3, the complete set of generated expressions is denoted by $\hat{\mathcal{F}}$. We consider the computational time required to evaluate the loss values of all the expressions, e.g., $\|\mathbf{y} - \hat{f}(\mathbf{X})\|_2^2$ where $\hat{f} \in \hat{\mathcal{F}}$, under different sample sizes (e.g., $10^1 \sim 10^4$ data points).

The result shows that PTS can quickly evaluate all the corresponding expressions, clearly surpassing the brute-force methods (see Fig. 5e). When the number of samples is less than 10^4 , the search speed of PTS is about 4 orders of magnitude faster, exhibiting an unprecedented increase in efficiency, thanks to the reuse of common subtree evaluation in parallel. Notably, when the number

of samples is big, downsampling of the data is suggested in the process of uncovering the equation structure in order to take the speed advantage of PTS during forward propagation. In the coefficient estimation stage, all samples should be used. This could further increase the efficiency of PTS while maintaining accuracy.

Space complexity. We categorize the operators used in PTS into three types: unary, binary-squared, and binary-triangled, which are represented by u , b_S , and b_T , respectively. Binary-squared operators represent non-commutative operators (e.g., $-$ and \div) depending on the order of operands, which requires ω_{i-1}^2 space on GPU during PSRN forward propagation (here, ω_{i-1} represents the input size of the previous symbol layer). Binary-triangled operators represent commutative operators (e.g., $+$ and \times) or the memory-saving version of non-commutative operators that only take up $\omega_{i-1}(\omega_{i-1} + 1)/2$ space (e.g., the SemiSub and SemiDiv symbols, described in [Methods: Symbol layer](#), in the Feynman benchmark which are specifically designed to save memory and support operations in only one direction).

With the number of operators in each category denoted by N_u , N_{b_S} and N_{b_T} , and the number of independent variables and layers of PTS denoted by m and l , respectively, the number of floating-point values required to be stored in PTS can be analyzed. Ignoring the impact of DR Mask, there is a recursive relationship between the tensor dimension of the $(i - 1)$ -th layer (e.g., ω_{i-1}) and that of the i -th layer (e.g., ω_i), namely,

$$\omega_i = N_u \omega_{i-1} + N_{b_S} \omega_{i-1}^2 + N_{b_T} \omega_{i-1} \frac{\omega_{i-1} + 1}{2} \leq \kappa \omega_{i-1}^2,$$

where $\kappa = N_u + N_{b_T} + N_{b_S}$. Thus the complexity of the number of floating-point values required to be stored by an l -layer PSRN can be calculated as $O(\kappa^{2^{l-1}} \omega_0^{2^l})$, where ω_0 represents the number of input slots.

Clearly, the memory consumption of PSRN increases dramatically with the number of layers. If each subtree value is a single-precision floating-point number (e.g., 32-bit), with the input dimension of 20 and operator set $\{+, -, \times, \div, \text{identity}, \sin, \cos, \exp, \log\}$, the required memory will reach over 10^5 GBs when the number of layers is 3, which requires a large compute set. Hence, finding a new strategy to relax the space complexity and alleviate the memory requirement is needed to further scale up the proposed model. Fig. 5f illustrates the graphic-memory footprint of various three-layered PSRN architectures, each characterized by a different operator set and the number of input slots. While the rise in memory demands serves as a constraint, this escalation is directly tied to the scalable model's ability to evaluate a greater number of candidate expressions within a single forward pass. This result also shows that PSRN follows the scaling law (e.g., the model capacity and size scale with the number of token inputs, given the fixed number of layers). The detailed hardware settings are found in [Supplementary Note 2.5](#).

Discussion

This paper introduces a new SR method, called PTS, to automatically and efficiently discover parsimonious equations to interpret the given data. In particular, we propose a PSRN architecture as the core search engine, which (1) automatically captures common subtrees of different symbolic expression trees for shared evaluation to expedite the computation, and (2) capitalizes on GPU-based parallel search with a notable performance boost. By recognizing and exploiting common subtrees in a vast number of candidate expressions, PSRN effectively bypasses the inefficiency and redundancy of independently evaluating each candidate. Such a strategy not only increases the

likelihood of directly finding the global optima in the grand search space but also significantly expedites the discovery process. Furthermore, resorting to high-performance GPUs, the proposed PTS model marks a pivotal transition towards a more rapid and efficient SR paradigm. When coupled with MCTS, the model’s ability to delve into complex expressions is further magnified, showing a promising potential to push the boundaries of solving more complex SR problems.

The efficacy of PTS has been extensively evaluated in discovering a variety of complex mathematical equations based on both synthetic and experimental datasets, including multiple benchmark problem sets (e.g., Nguyen, Nguyen-c, R, Livermore, and Feynman), nonlinear chaotic dynamics, and two datasets collected via lab experiments (e.g., EMPS and turbulent friction). We have demonstrated that PTS possesses a powerful capability in general-purpose SR, exhibiting a remarkably superior recovery rate and speed. The performance of PTS exceeds comprehensively several representative baseline models, achieving up to two orders of magnitude efficiency improvement while maintaining a much better accuracy. Moreover, even for a target equation in a very complex form, the PTS model is still capable of swiftly and reliably uncovering the ground truth directly, rather than being led astray by the ambiguous patterns present in the data. Consequently, PTS excels in discovering accurate and parsimonious expressions from very limited data in a short time of budget.

Despite its demonstrated efficacy and potential, the PTS model is faced with several challenges that need to be addressed in the future. Firstly, the PSRN module has a rapidly increasing demand for memory while increasing the number of symbol layers (see [Model performance analysis: Space complexity](#)). Currently, a brute-force implementation of PSRN can only directly handle expressions with a parse tree depth ≤ 3 under conventional settings. Otherwise, such a method relies on MCTS, which locates advanced tokenized input, to extend PSRN’s capacity to interpret deeper parse trees. This bottleneck impedes the PTS’s exploration of much deeper expressions. An ongoing work to tackle this issue lies in designing a learnable score-based sampling strategy to select a finite number of optimal subtrees for the generation of candidate expressions at a deeper layer. This has the potential to deepen the symbol layers meanwhile saving the graphic memory requirement. Secondly, it should be noted that our model is of limited heuristic guidance. This arises from the fact that MCTS primarily serves to offer directional cues and steer PSRN away from re-searching the base expressions set that has already undergone forward computation. It lacks the continuous ability to analyze expressions within the Pareto front. This shortfall implies that over an extended runtime, the advantage of our PTS method compared to heuristic techniques (e.g., GP) tends to diminish. However, this sheds light on the path for our future work on incorporating a Pareto front analysis component (e.g. meta-heuristic algorithms) into PTS to further improve its performance. Lastly, the current work does not take into account our prior knowledge or dimensional constraints (e.g., the unit of a physical quantity). Another exciting ongoing work by the authors attempts to integrate the units of the input variables to identify expressions that comply with dimensional constraints. Such a strategy has the potential to conserve the graphic memory requirement and, at the same time, expedite the search. We intend to continue addressing these challenges methodically in our forthcoming research.

Methods

The NP-hardness of SR has been noted in existing studies [28, 30, 45], followed by a formal proof recently [46]. This implies the absence of a known solution that can be determined in polynomial time for solving SR problems across all instances, and reflects that the search space for SR is indeed vast and intricate. Almost all the existing SR methods involve a common procedure, which is to assess the quality of candidate expressions $\hat{\mathcal{F}} = \{\hat{f}_1, \hat{f}_2, \dots\}$ based on the given data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$,

where $\mathbf{X} \in \mathbb{R}^{n \times m}$, $\mathbf{y} \in \mathbb{R}^{n \times 1}$. Here, \hat{f} is usually constructed by a given operator set \mathcal{O} (e.g., $\{+, \times, -, \div, \sin, \cos, \exp, \log\}$). Typically, this evaluation is guided by the mean square error (MSE) expressed as:

$$\text{MSE} = \frac{1}{n} \|\hat{f}(\mathbf{X}) - \mathbf{y}\|_2^2. \quad (1)$$

During the search process, a large number of candidate expressions need to be evaluated, while the available operators and variables are limited, leading to the inevitable existence of vast repeated sub-expressions. Existing methods evaluate candidate expressions sequentially and independently. As a result, the common intermediate expressions are repeatedly calculated, leading to significant computational burden (see [Supplementary Fig. S.1](#)). By reducing the amount of repeated computation, the search process can be significantly accelerated.

Symbol layer

A mathematical expression can be equivalently represented as a parse tree [13], where the internal nodes denote mathematical operators and the leaf nodes the variables and constants. The crux of the aforementioned computational issue lies in the absence of temporary storage for subtree values and parallel evaluation. Consequently, the common subtrees in different candidate expressions are repeatedly evaluated, resulting in significant computational wastage.

To this end, we introduce the concept of Symbol Layer (see Fig. 1c), which consists of a series of mathematical operators. The Symbol Layer serves to transform the computation results of shallow expression parse trees into deeper ones. From the perspective of avoiding redundant computations, the results of the Symbol Layer are cached and can be utilized by parse trees with greater heights. From the view of parallel computation, the Symbol Layer can leverage the power of GPU to compute the results of common subtrees in parallel, significantly improving the overall speed (see [Supplementary Fig. S.1](#)). We categorize the mathematical operators in the Symbol Layer into three types: (1) unary operators, (e.g., \sin and \exp); (2) binary-squared operators (e.g., $-$ and \div), representing non-commutative operators; (3) binary-triangled operators, representing commutative operators (e.g., $+$ and \times) or a variant of non-commutative operators with low-memory footprint (e.g., SemiSub and SemiDiv that output $x_i - x_j$ and $x_i \div x_j$ for $i \leq j$, respectively). Note that the binary-triangled operators only take up half of the space compared with the binary-squared operators. We denote these three types of operators as u , b_S , and b_T , respectively. Mathematically, a Symbol Layer located at the l -th level can be represented as follows:

$$\mathbf{h}^{(l)} = \left(\left\|_{i=1}^{N_u} \mathbf{u}_i (\mathbf{h}^{(l-1)}) \right\| \right) \left\| \left(\left\|_{i=1}^{N_{b_S}} \mathbf{b}_{Si} (\mathbf{h}^{(l-1)}) \right\| \right) \left\| \left(\left\|_{i=1}^{N_{b_T}} \mathbf{b}_{Ti} (\mathbf{h}^{(l-1)}) \right\| \right) \right\|, \quad (2)$$

where

$$\mathbf{u}(\mathbf{h}) = \left\|_i u(h_i), \quad \mathbf{b}_S(\mathbf{h}) = \left\|_{i,j} b_S(h_i, h_j), \quad \mathbf{b}_T(\mathbf{h}) = \left\|_{i \leq j} b_T(h_i, h_j). \quad (3) \right.\right.$$

Here, $\|$ represents the concatenation operation; N_u , N_{b_S} and N_{b_T} denote the numbers of unary operators, binary-squared operators and binary-triangled operators.

For example, let us consider independent variables $\{x_1, x_2\}$ and dependent variable z , with the task of finding an expression that satisfies $z = f(x_1, x_2)$. When the independent variables $\{x_1, x_2\}$ are fed into a Symbol Layer, we obtain the combinatorial results (e.g., $\{x_1, x_2, \sin(x_1), \sin(x_2), \dots, x_1 + x_2, x_1 + x_2, x_2 + x_2, \dots\}$). Notably, each value in the output tensor of the Symbol Layer corresponds to a distinct sub-expression. This enables PTS to compute the common subtree values just once,

avoiding redundant calculations. Additionally, the Symbol Layer can be leveraged on the GPU for parallel computation, further enhancing the speed of expression searches significantly.

When establishing a Symbol Layer initially, an inherent offset tensor Θ is inferred and stored within the layer for subsequent lazy symbolic deducing during the backward pass (see Extended Data Fig. 1). For the l -th Symbol Layer, its offset tensor can be represented as follows:

$$\Theta^{(l)} = \left(\left\|_{i=1}^{N_u} \Theta_u \right\| \right) \parallel \left(\left\|_{i=1}^{N_{b_S}} \Theta_{b_S} \right\| \right) \parallel \left(\left\|_{i=1}^{N_{b_T}} \Theta_{b_T} \right\| \right), \quad (4)$$

where

$$\Theta_u = \begin{bmatrix} 1 & \dots & \omega \\ \emptyset & \dots & \emptyset \end{bmatrix}_{2 \times \omega}, \quad (5a)$$

$$\Theta_{b_S} = \left\|_{j=1}^{\omega} \begin{bmatrix} 1 & \dots & \omega \\ j & \dots & j \end{bmatrix}_{2 \times \omega^2}, \quad (5b)$$

$$\Theta_{b_T} = \left\|_{j=1}^{\omega} \begin{bmatrix} 1 & \dots & \omega - j + 1 \\ j & \dots & j \end{bmatrix}_{2 \times \frac{\omega(\omega+1)}{2}}. \quad (5c)$$

Here, ω represents the output dimension (e.g., the number of symbolic parse trees) of the previous layer.

Each position in the output tensor $\mathbf{h}^{(l)}$ of the Symbol Layer corresponds to a unique column in the offset tensor $\Theta^{(l)}$. In other words, it corresponds to two child indices (for unary operators, only one), representing the left and right child nodes of the current output tensor position from the previous layer. These offset tensors are used for recursive backward symbol deduction after the position of a minimum MSE value is found.

Parallel symbolic regression network

By stacking multiple Symbol Layers, we can then construct a PSRN as shown in Fig. 1c. This network takes in a set of admissible base expressions, denoted by s , along with their corresponding data tensor \mathbf{X} , and is capable of utilizing GPU for rapid parallel forward computation on the data (e.g., typically within a mere few seconds). Based on the operators set \mathcal{O} , a multitude of distinct subtree values are efficiently computed layer by layer. Once the calculations of the final layer are completed, hundreds of millions of expression parse tree values $\hat{\mathcal{F}}(\mathbf{X}) = \{\hat{f}_1(\mathbf{X}), \hat{f}_2(\mathbf{X}), \dots\}$ could be obtained. These generated candidate expressions make use of intermediate results of common subtrees, thus avoiding extensive redundant computations. Subsequently, these candidate expression values are used to compute the MSE defined in Eq. (1) for each \hat{f}_i ($i = 1, 2, \dots, |\mathcal{F}|$), in order to identify the position of the minimum value in the error tensor. Then, the pre-generated offset tensor Θ in each layer, containing the indices of each node's children, is leveraged to efficiently trace where their constituent sub-expressions originate and recover the top-level operators among them. This enables recursive inference of the optimal candidate expression(s) in a layer-wise manner. Extended Data Fig. 1 provides a more detailed elucidation of the implementation specifics of PSRN.

As the number of layers increases, the required GPU memory also grows rapidly. In [Model performance analysis: Space complexity](#), the space complexity of PSRN is discussed in detail. For the experiments, we use a 3-layer PSRN configuration by default. In the SR benchmark tasks, we employ a 5-input PSRN with the operator set $\mathcal{O}_{\text{Koza}} = \{+, \times, -, \div, \text{identity}, \sin, \cos, \exp, \log\}$, except for the Feynman expression set which uses a 6-input PSRN with the operator set $\mathcal{O}_{\text{SemiKoza}} =$

$\{+, \times, \text{SemiSub}, \text{SemiDiv}, \text{identity}, \text{neg}, \text{inv}, \text{sin}, \text{cos}, \text{exp}, \text{log}\}$. Such a setting aims to conserve GPU memory for handling more input variables. The distinguishing feature of the $\mathcal{O}_{\text{SemiKoza}}$ operator set is that it treats division and subtraction as binary-triangled operators and allows only one direction of operation (see [Methods: Symbol layer](#)). This trade-off reduces expressive power but conserves GPU memory, which enables to tackle larger scale SR tasks.

PSRN regressor with Monte Carlo tree search

Given the data tensor \mathbf{X} and the corresponding base expression set s , we obtain the values of expressions for all parse trees with depths up to l , namely, $\{\hat{f}(\mathbf{X})|d(\hat{f}) \leq l\}$, where l denotes the number of Symbol Layers in PSRN (e.g., $l = 3$ for typical cases). On a graphics card with 80GB of memory, we can allocate space for up to 5 inputs for PSRN with the operators set $\mathcal{O}_{\text{Koza}}$. Apart from the independent variables, the remaining input slots can be used for more complex sub-expressions. For example, given independent variables x_1 and x_2 , the additional slots can be employed to try other base expressions like $x_1 + x_1$, $x_1 \times x_2$, and $\sin(x_1)$, which enables the network to derive deeper parse trees based on these additional inputs of the slot. Therefore, the task of identifying admissible inputs can be approached as a search problem. The algorithm is anticipated to progressively expand the base expression set until all the slots are used, resulting in a terminal state $s_{\mathcal{T}} = \{x_1, x_2, g_1(x_1, x_2), g_2(x_1, x_2), g_3(x_1, x_2)\}$. Then, a forward propagation is conducted by PSRN using $s_{\mathcal{T}}$ as the base expression set to enrich and diversify the generation of potential expression trees. Such an approach could facilitate the exploration of deeper expressions (e.g., $\{\hat{f}(\mathbf{X})|d(\hat{f}) > l\}$) and broadens the spectrum of candidate expressions.

Here, we utilize MCTS to tackle this search problem. MCTS is a decision-making algorithm designed for exploring vast combinatorial spaces represented as search trees. This approach follows the best-first search principle, relying on evaluations from stochastic simulations. MCTS consists of four iterative steps: selection, expansion, simulation, and backpropagation (see Fig. 1a) described in the following.

- (1) **Selection:** Starting from the initial base expression set s_0 as the root node, the algorithm iteratively chooses a new node with the highest Upper Confidence Bounds applied to Trees (UCT) value [38] defined as

$$\text{UCT}(s, a) = Q(s, a) + c\sqrt{\frac{\ln N(s)}{N(s, a)}}, \quad (6)$$

where a represents the action of inserting a new base expression f' into the current node s . $Q(s, a)$ is the expected value associated with node s and action a . $N(s)$ is the total number of visits to node s . $N(s, a)$ is the number of visits to node s after taking action a . c is an exploration parameter. Each node in the search tree denotes a distinct base expression set.

- (2) **Expansion:** If a node s_t is expandable, MCTS expands it by randomly selecting an unvisited node s_{t+1} . During the expansion of node s_t , a new base expression f' is then generated by combining the base expressions present in s_t . The new base expression set is obtained by adding f' to s_t , expressed as

$$s_{t+1} = s_t \cup \{f'\}. \quad (7)$$

- (3) **Simulation:** Following the expansion, if the current node is non-terminal, a playout sequence is initiated that continues through successive states until a terminal node, denoted by $s_{\mathcal{T}}$, is reached. Then, the PSRN conducts a forward propagation, combining all the base expressions

in s_T and finding the expression \tilde{f} that minimizes the MSE in Eq. (1). The expression's reward is calculated by

$$r = \frac{\eta^\alpha}{1 + \sqrt{\text{MSE}}}, \quad (8)$$

where η is a discount factor promoting concise trees (e.g., 0.99 by default), and α represents the complexity of f . Here, we consider the number of operators in an expression as a proxy for its complexity. If token constants are enabled, \tilde{f} is first optimized through a least squares optimization step to obtain f^* , which is discussed in [Methods: Coefficients tuning](#).

- (4) **Backpropagation:** The reward of the expression is backpropagated along the original path, updating the expected values Q and visit counts N of the involved nodes.

The algorithm involves a series of epochs, represented by $\mathcal{N}_e = |s_T| - |s_0|$, the difference between the number of elements in a terminal state s_T and the initial state s_0 . Within each epoch, a fixed number of simulations is performed. At the end of each epoch, the node with the highest expected value Q becomes the root node for the next epoch. This iterative process facilitates systematic exploration of the search space and leverages the accumulated knowledge from previous simulations.

Coefficients tuning

When handling expressions with coefficients, we pre-select a sampling range for token constants and reserve multiple input slots of the network. Before executing each PSRN forward propagation, several constants are sampled and fed into the PSRN along with the base expressions (Fig. 1b). The sampled token constants can be combined with various operators within the network to form a more diverse set of constants. For example, 1.6 and 2.3 can be combined as $1.6 + 2.3$, $\exp(2.3)$, $2.3/\sin(1.6)$, etc. This allows the network to rapidly increase the range of coefficients it can represent as the number of network layers deepens. After each forward propagation, we obtain expressions $\tilde{\mathcal{F}} = \{\tilde{f}_1, \tilde{f}_2, \dots, \tilde{f}_k\}$ that best fit the given data, where k is a pre-defined number, denoted as

$$\tilde{\mathcal{F}} = \text{PSRN}_{\text{const}}(\mathbf{X}, \mathbf{y}). \quad (9)$$

We parse the searched expression \tilde{f} , take out its coefficients as the initial values, and apply least-squares (LS) optimization to derive the optimized expression f^* given by

$$f_i^* = \text{LS}(\tilde{f}_i, \mathbf{X}, \mathbf{y}), \quad i = 1, \dots, |\tilde{\mathcal{F}}|. \quad (10)$$

It is noted that, due to PSRN's tendency to prefer relatively complex expressions when searching equations containing coefficients, we introduce a linear regression step for the given independent variables before each MCTS begins. This step was implemented to expedite the process of discovering simpler expressions using PTS.

Duplicate removal mask

One major limitation of standard PSRN lies in its high demand for graphics memory. When using binary operators such as addition and multiplication, the graphics memory required for each layer grows quadratically with respect to the tensor dimensions of the previous layer. Therefore, reducing the output dimension of the penultimate layer can significantly reduce the required graphics memory. We propose a technique called Duplicate Removal Mask (DR Mask) as shown in Fig. 1e. Before the last PSRN layer, we employ the DR Mask to remove repeated terms and generate a dense input set. Specifically, assuming the input variables x_1, x_2, \dots, x_m are independent, we extract the output

expressions from the last second layer, parse them using the symbolic computing library SymPy, and remove duplicate expressions to obtain a mask of unique expressions. SymPy’s efficient hash value computation facilitates the comparison of mathematical expressions for symbolic equivalence. Typically, this process takes less than a minute. Once the DR Mask is identified, it is reusable for PSRNs with the same architecture (e.g., total number of network layers and operator sets). During the search process, the expressions marked by the mask are extracted from the penultimate layer’s output tensor and then passed to the final layer for the most graphics-intensive binary operator computations. The percentage of graphics memory saved by the DR Mask technique depends on the input size, the number of layers, and the operators used in the PSRN architecture. Detailed results are shown in [Model ablation study](#).

Baseline models

We consider five main baseline SR algorithms used in the comparasion analysis: DGSR [34], NGGP [45], PySR [50], BMS [12] and SPL [42]. They represent a selection of SR algorithms, ranging from recent advancements to established, powerful methods. The introduction and detailed settings of the baseline models are found in [Supplementary Note 2.3](#).

Data availability

All the datasets used to test the methods in this study are available on GitHub at <https://github.com/intell-sci-comput/PTS>.

Code availability

All the source codes used to reproduce the results in this study are available on GitHub at <https://github.com/intell-sci-comput/PTS>.

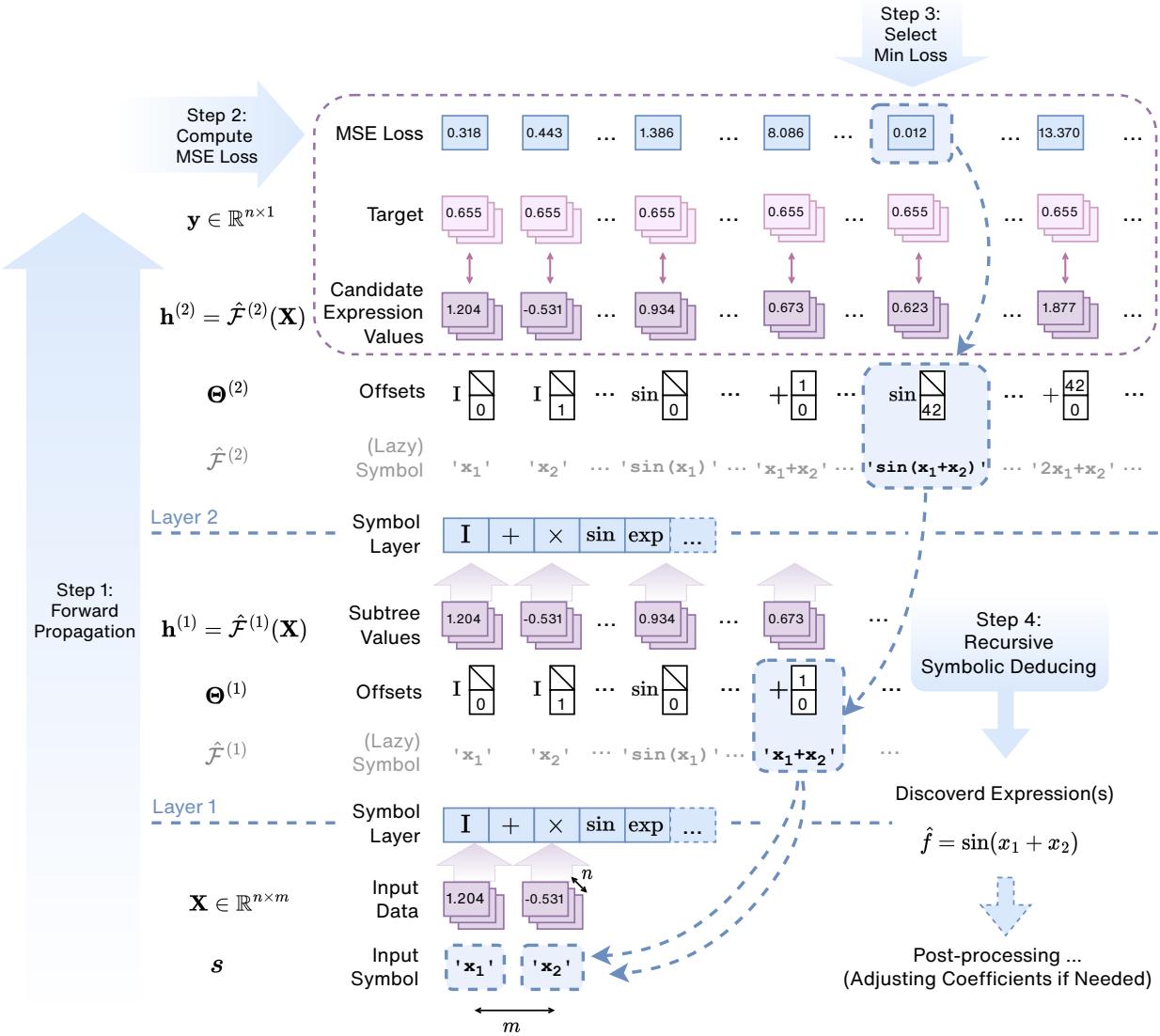
Acknowledgement: The work is supported by the National Natural Science Foundation of China (No. 92270118, No. 62276269, and No. 62206299), the Beijing Natural Science Foundation (No. 1232009), and the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB0620103). In addition, H.S and Y.L. would like to acknowledge the support from the Fundamental Research Funds for the Central Universities (No. 202230265 and No. E2EG2202X2).

Author contributions: K.R., H.S., Y.L. contributed to the ideation and design of the research; K.R. performed the research; H.S. and J.R.W. supervised the project; all authors contributed to the research discussions, writing, and editing of the paper.

Corresponding authors: Hao Sun (haosun@ruc.edu.cn), Ji-Rong Wen (jrwen@ruc.edu.cn) and Yang Liu (liuyang22@ucas.ac.cn).

Competing interests: The authors declare no competing interests.

Supplementary information: The supplementary information is attached.



Extended Data Fig. 1: The detailed process of PSRN forward propagation for obtaining the optimal expression. Note that, to maintain simplicity, token constants are not shown here and the illustration is limited to an example with only two layers. Step 1: The input data \mathbf{X} and the base expression set s are fed into PSRN for forward propagation. Based on the designated operator categories, a large number of distinct subtree values \mathbf{h} , e.g., $\hat{\mathcal{F}}(\mathbf{X})$ are rapidly calculated layer by layer (purple). Step 2: The MSE is computed between the final layer's output and the broadcast target tensor \mathbf{y} , resulting in the loss tensor. Step 3: The position of the minimum value in the loss tensor is selected. Step 4: Starting from the optimal position, a recursive symbolic deducing is performed using the offset tensor Θ generated when building the network, ultimately yielding the optimal expression. If necessary, the coefficients of this expression will be adjusted in post-processing.

Supplementary Information

1 Background

As far as we know, existing SR algorithms inevitably rely on a necessary step, namely evaluating large-scale candidate expressions $\hat{\mathcal{F}} = \{\hat{f}_1, \hat{f}_2, \dots\}$ on the given data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ to obtain the error. While some methods involving the use of neural networks can generate candidate expressions using GPUs, the process of obtaining the error for each mathematical expression still depends on sequential and independent evaluations using CPUs. When the number of expressions to be evaluated becomes sufficiently large, there will be a significant amount of redundant evaluations of common subtrees inevitably. In Figure S.1, we illustrate the fundamental difference between the PSRN module of PTS and existing methods in terms of symbolic expression evaluation.

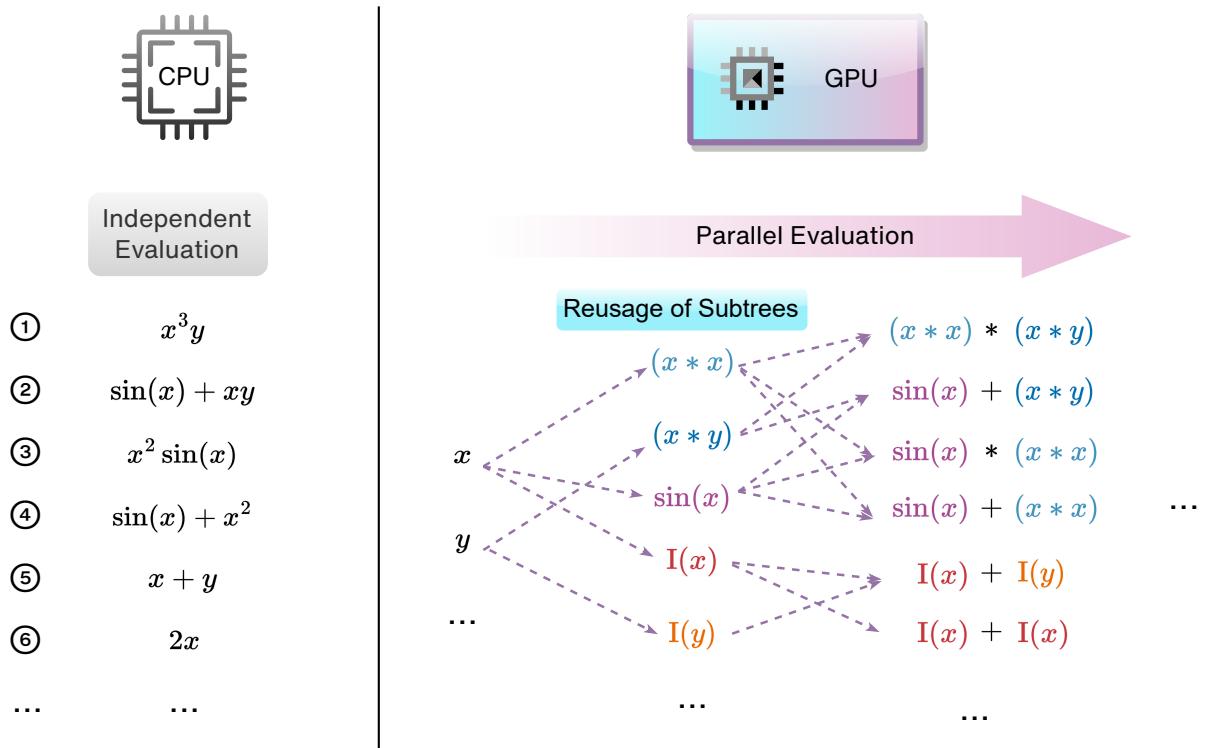


Figure S.1: Comparison between CPU independent evaluation and GPU parallel evaluation. The PSRN significantly reduces unnecessary computational workload through the reuse of common subtree values when operating on GPUs. Subtrees bearing identical values are indicated by the same color, while the dashed lines trace the propagation of these subtree values throughout the PSRN architecture.

2 Methodology

2.1 Benchmark Symbolic Regression Problems

We employed the Nguyen, Nguyen-c [47, 48], R, R* [49], Livermore [45], and Feynman [30] benchmark problem sets to evaluate the symbolic recovery rate and recovery speed of the algorithms. The specific benchmark problem settings were consistent with those described in the NGGP [45] and DGSR [34]. Table S.1 and S.2 shows these SR benchmark problems.

Table S.1: Nguyen, Nguyen-c, R and R* Benchmark Problems. U denotes uniform sampling over the interval, while E denotes equidistant sampling over the interval. The three parameters (a, b, c) represent the lower bound, upper bound, and the number of sampling points within the interval. The operator set used is $\{+, \times, -, \div, \sin, \cos, \exp, \log\}$.

Benchmark	Expression	Dataset	Tokens
Nguyen-1	$x_1^3 + x_1^2 + x_1$	$U(-1, 1, 20)$	$\{x_1\}$
Nguyen-2	$x_1^4 + x_1^3 + x_1^2 + x_1$	$U(-1, 1, 20)$	$\{x_1\}$
Nguyen-3	$x_1^5 + x_1^4 + x_1^3 + x_1^2 + x_1$	$U(-1, 1, 20)$	$\{x_1\}$
Nguyen-4	$x_1^6 + x_1^5 + x_1^4 + x_1^3 + x_1^2 + x_1$	$U(-1, 1, 20)$	$\{x_1\}$
Nguyen-5	$\sin(x_1^2) \cos(x_1) - 1$	$U(-1, 1, 20)$	$\{x_1\}$
Nguyen-6	$\sin(x_1) + \sin(x_1 + x_1^2)$	$U(-1, 1, 20)$	$\{x_1\}$
Nguyen-7	$\log(x_1 + 1) + \log(x_1^2 + 1)$	$U(0, 2, 20)$	$\{x_1\}$
Nguyen-8	$\sqrt{x_1}$	$U(0, 4, 20)$	$\{x_1\}$
Nguyen-9	$\sin(x_1) + \sin(x_2^2)$	$U(0, 1, 20)$	$\{x_1, x_2\}$
Nguyen-10	$2 \sin(x_1) \cos(x_2)$	$U(0, 1, 20)$	$\{x_1, x_2\}$
Nguyen-11	$x_1^{x_2}$	$U(0, 1, 20)$	$\{x_1, x_2\}$
Nguyen-12	$x_1^4 - x_1^2 + \frac{1}{2}x_2^2 - x_2$	$U(0, 1, 20)$	$\{x_1, x_2\}$
Nguyen-1c	$3.39x_1^3 + 2.12x_1^2 + 1.78x_1$	$U(-1, 1, 20)$	$\{x_1, \text{const}\}$
Nguyen-2c	$0.48x_1^4 + 3.39x_1^3 + 2.12x_1^2 + 1.78x_1$	$U(-1, 1, 20)$	$\{x_1, \text{const}\}$
Nguyen-5c	$\sin(x_1^2) \cos(x_1) - 0.75$	$U(0, 2, 20)$	$\{x_1, \text{const}\}$
Nguyen-8c	$\sqrt{1.23x_1}$	$U(0, 4, 20)$	$\{x_1, \text{const}\}$
Nguyen-9c	$\sin(1.5x_1) + \sin(0.5x_2^2)$	$U(0, 1, 20)$	$\{x_1, x_2, \text{const}\}$
Nguyen-10c	$\sin(1.5x_1) \cos(0.5x_2)$	$U(0, 1, 20)$	$\{x_1, x_2, \text{const}\}$
R-1	$(x_1 + 1)^3 / (x_1^2 - x_1 + 1)$	$E(-1, 1, 20)$	$\{x_1\}$
R-2	$(x_1^5 - 3x^3 + 1) / (x_1^2 + 1)$	$E(-1, 1, 20)$	$\{x_1\}$
R-3	$(x_1^6 + x_1^5) / (x_1^4 + x_1^3 + x_1^2 + x_1 + 1)$	$E(-1, 1, 20)$	$\{x_1\}$
R-1*	$(x_1 + 1)^3 / (x_1^2 - x_1 + 1)$	$E(-10, 10, 20)$	$\{x_1\}$
R-2*	$(x_1^5 - 3x^3 + 1) / (x_1^2 + 1)$	$E(-10, 10, 20)$	$\{x_1\}$
R-3*	$(x_1^6 + x_1^5) / (x_1^4 + x_1^3 + x_1^2 + x_1 + 1)$	$E(-10, 10, 20)$	$\{x_1\}$

Table S.2: Livermore and Feynman Benchmark Problems. U denotes uniform sampling over the interval, while E denotes equidistant sampling over the interval. The three parameters (a, b, c) represent the lower bound, upper bound, and the number of sampling points within the interval. The operator set used is $\{+, \times, -, \div, \sin, \cos, \exp, \log\}$.

Benchmark	Expression	Dataset	Tokens
Livermore-1	$1/3 + x_1 + \sin(x_1^2)$	$U(-10, 10, 1000)$	$\{x_1\}$
Livermore-2	$\sin(x_1^2) \cos(x_1) - 2$	$U(-1, 1, 20)$	$\{x_1\}$
Livermore-3	$\sin(x_1^3) \cos(x_1^2) - 1$	$U(-1, 1, 20)$	$\{x_1\}$
Livermore-4	$\log(x_1 + 1) + \log(x_1^2 + 1) + \log(x_1)$	$U(0, 2, 20)$	$\{x_1\}$
Livermore-5	$x_1^4 - x_1^3 + x_1^2 - x_2$	$U(0, 1, 20)$	$\{x_1, x_2\}$
Livermore-6	$4x^4 + 3x^3 + 2x^2 + x_1$	$U(-1, 1, 20)$	$\{x_1\}$
Livermore-7	$\sinh(x_1)$	$U(-1, 1, 20)$	$\{x_1\}$
Livermore-8	$\cosh(x_1)$	$U(-1, 1, 20)$	$\{x_1\}$
Livermore-9	$x_1^9 + x_1^8 + x_1^7 + x_1^6 + x_1^5 + x_1^4 + x_1^3 + x_1^2 + x_1$	$U(-1, 1, 20)$	$\{x_1\}$
Livermore-10	$6 \sin(x_1) \cos(x_2)$	$U(0, 1, 20)$	$\{x_1, x_2\}$
Livermore-11	$x_1^4/(x_1 + x_2)$	$U(-1, 1, 50)$	$\{x_1, x_2\}$
Livermore-12	x_1^5/x_2^3	$U(-1, 1, 50)$	$\{x_1, x_2\}$
Livermore-13	$x_1^{\frac{1}{3}}$	$U(0, 4, 20)$	$\{x_1\}$
Livermore-14	$x_1^3 + x_1^2 + x_1 + \sin(x_1) + \sin(x_1^2)$	$U(-1, 1, 20)$	$\{x_1\}$
Livermore-15	$x_1^{\frac{1}{5}}$	$U(0, 4, 20)$	$\{x_1\}$
Livermore-16	$x_1^{\frac{2}{5}}$	$U(0, 4, 20)$	$\{x_1\}$
Livermore-17	$4 \sin(x_1) \cos(x_2)$	$U(0, 1, 20)$	$\{x_1, x_2\}$
Livermore-18	$\sin(x_1^2) \cos(x_1) - 5$	$U(-1, 1, 20)$	$\{x_1\}$
Livermore-19	$x_1^5 + x_1^4 + x_1^2 + x_1$	$U(-1, 1, 20)$	$\{x_1\}$
Livermore-20	$\exp(-x_1^2)$	$U(-1, 1, 20)$	$\{x_1\}$
Livermore-21	$x_1^8 + x_1^7 + x_1^6 + x_1^5 + x_1^4 + x_1^3 + x_1^2 + x_1$	$U(-1, 1, 20)$	$\{x_1\}$
Livermore-22	$\exp(-\frac{1}{2}x_1^2)$	$U(-1, 1, 20)$	$\{x_1\}$
Feynman-1	$x_1 x_2$	$U(1, 5, 20)$	$\{x_1, x_2\}$
Feynman-2	$\frac{x_1}{2(1+x_2)}$	$U(1, 5, 20)$	$\{x_1, x_2\}$
Feynman-3	$x_1 x_2^2$	$U(1, 5, 20)$	$\{x_1, x_2\}$
Feynman-4	$1 + \frac{x_1 x_2}{1-(x_1 x_2/3)}$	$U(0, 1, 20)$	$\{x_1, x_2\}$
Feynman-5	$\frac{x_1}{x_2}$	$U(1, 5, 20)$	$\{x_1, x_2\}$
Feynman-6	$\frac{1}{2} x_1 x_2^2$	$U(1, 5, 20)$	$\{x_1, x_2\}$
Feynman-7	$\frac{3}{2} x_1 x_2$	$U(1, 5, 20)$	$\{x_1, x_2\}$
Feynman-8	$\frac{x_1}{e^{\frac{x_4 x_5}{x_2 x_3}} + e^{-\frac{x_4 x_5}{x_2 x_3}}}$	$U(1, 3, 50)$	$\{x_1, x_2, x_3, x_4, x_5\}$
Feynman-9	$x_1 x_2 x_3 \log \frac{x_5}{x_4}$	$U(1, 5, 50)$	$\{x_1, x_2, x_3, x_4, x_5\}$
Feynman-10	$x_1 (x_3 - x_2) \frac{x_4}{x_5}$	$U(1, 5, 50)$	$\{x_1, x_2, x_3, x_4, x_5\}$
Feynman-11	$\frac{x_1 x_2}{x_5 (x_3^2 - x_4^2)}$	$U(1, 3, 50)$	$\{x_1, x_2, x_3, x_4, x_5\}$
Feynman-12	$\frac{x_1 x_2^2 x_3}{3 x_4 x_5}$	$U(1, 5, 50)$	$\{x_1, x_2, x_3, x_4, x_5\}$
Feynman-13	$x_1 (e^{\frac{x_4 x_5}{x_2 x_3}} - 1)$	$U(1, 5, 50)$	$\{x_1, x_2, x_3, x_4, x_5\}$
Feynman-14	$x_5 x_1 x_2 (\frac{1}{x_4} - \frac{1}{x_3})$	$U(1, 5, 50)$	$\{x_1, x_2, x_3, x_4, x_5\}$
Feynman-15	$x_1 (x_2 + x_3 x_4 \sin(x_5))$	$U(1, 5, 50)$	$\{x_1, x_2, x_3, x_4, x_5\}$

2.2 Chaotic Dynamics Dataset

We use a nonlinear chaotic dynamics dataset called dysts [52]. Considering the widespread presence of chaotic dynamics in the real world, such as in finance and meteorology, we believe that this dataset act as a robust benchmark for evaluating SR algorithms' ability to uncover underlying physical laws. Using this dataset avoids solely focusing on toy examples like Lorenz Attrator [51] (see Fig S.3), enabling a more objective and general assessment of SR algorithms' performance. By doing so, we can better understand the algorithm's capabilities in uncovering underlying dynamical equations.

We conducted tests using chaotic dynamics NewtonLiepnik [60], HyperLorenz [52], HyperJha [52], HyperPang [61], ShimizuMorioka [62], GenesioTesi [63], Laser [64], Duffing [65], Brusselator [66], KawczynskiStrizhak [67], Rucklidge [68], FitzHughNagumo [69], Finance [70], DequanLi [71], Hadley [72] and SprottJerk [73]. Tables S.3, S.4, S.5 and S.6 show the governing equations of these chaotic dynamics. For systems with additional external force, we set the coefficients of their external force terms to 0 (i.e., autonomous chaotic dynamics). All other parameters were kept at their default values from the dysts problem set. The trajectory data is generated by solving numerical equations, and during this process, we sampled 1000 points, with every 100 points representing one period, and introduced 1% Gaussian noise.

In this experiment, a large operator set $\{+, \times, -, \div, \sin, \cos, \exp, \log, \text{abs}, \text{sign}, \tanh, \cosh\}$, are used to simulate the situation in which scientists explore unknown systems in the real world with no prior knowledge. The models being compared were allocated a time budget of approximately 90 seconds. After the algorithms completed their search, we assessed whether the Pareto front produced by each algorithm contained expressions with the same structure (allowing for a constant bias term) as the ground truth expression. We calculated the average recovery rate of the derivatives (e.g., \dot{x} , \dot{y} , \dot{z} , which were determined using data affected by noise.) based on 50 repeated experiments, and then selected the minimum recovery rate among these derivatives to represent the overall recovery rate of the whole system.

2.3 Configurations of Models

PTS: The PTS model for the symbolic regression benchmark employs $\mathcal{O}_{\text{Koza}}$ (i.e., $\{+, \times, -, \div, \text{identity}, \sin, \cos, \exp, \log\}$) operators and, for 5-dimensional Feynman expressions, set to $\mathcal{O}_{\text{SemiKoza}}$ (i.e., $\{+, \times, \text{SemiSub}, \text{SemiDiv}, \text{identity}, \text{neg}, \text{inv}, \sin, \cos, \exp, \log\}$) for saving graphic memory. It has a structure of 3 layers, accepts 5 input tokens (for 5-dimensional Feynman expressions, set to 6), and performs down-sampling when the number of samples exceeds 40. During each epoch, 400 MCTS simulations are conducted. The model tries 2 token constants within a range of $[0, 3]$, repeating the process 3 times in a MCTS simulation. It retrieves the top 10 expressions with minimum error in each PSRN forward pass and uses a sampling interval of 0.1 for token constants. For chaotic dynamics, the operators expand to include $\{+, \times, -, \div, \sin, \cos, \exp, \log, \tanh, \cosh, \text{sign}, \text{abs}\}$, and down-sampling is set to 200. With a similar configuration, during each epoch, 3 MCTS simulations are conducted (For DequanLi, used 10). It tries 2 token constants (For 4-dimensional chaotic dynamics, set to 1 for saving graphic memory) within a range of $[0, 3]$, repeating the process once in a MCTS simulation, and retrieves the top 30 expressions with minimum error in each PSRN forward pass. For EMPS experiment, the operators used are $\{+, \times, -, \div, \sin, \cos, \exp, \log, \text{sign}\}$, and down-sampling is set to 200. During each epoch, 5 MCTS simulations are conducted. For the turbulence experiment, the operators used are $\{+, \times, -, \div, \sin, \cos, \exp, \log, \tanh, \cosh, \text{pow2}, \text{pow3}\}$, and down-sampling is set to 100. During each epoch, 5 MCTS simulations are conducted. In the experiments described, when there is not enough graphic memory, $-$ and \div operators are replaced

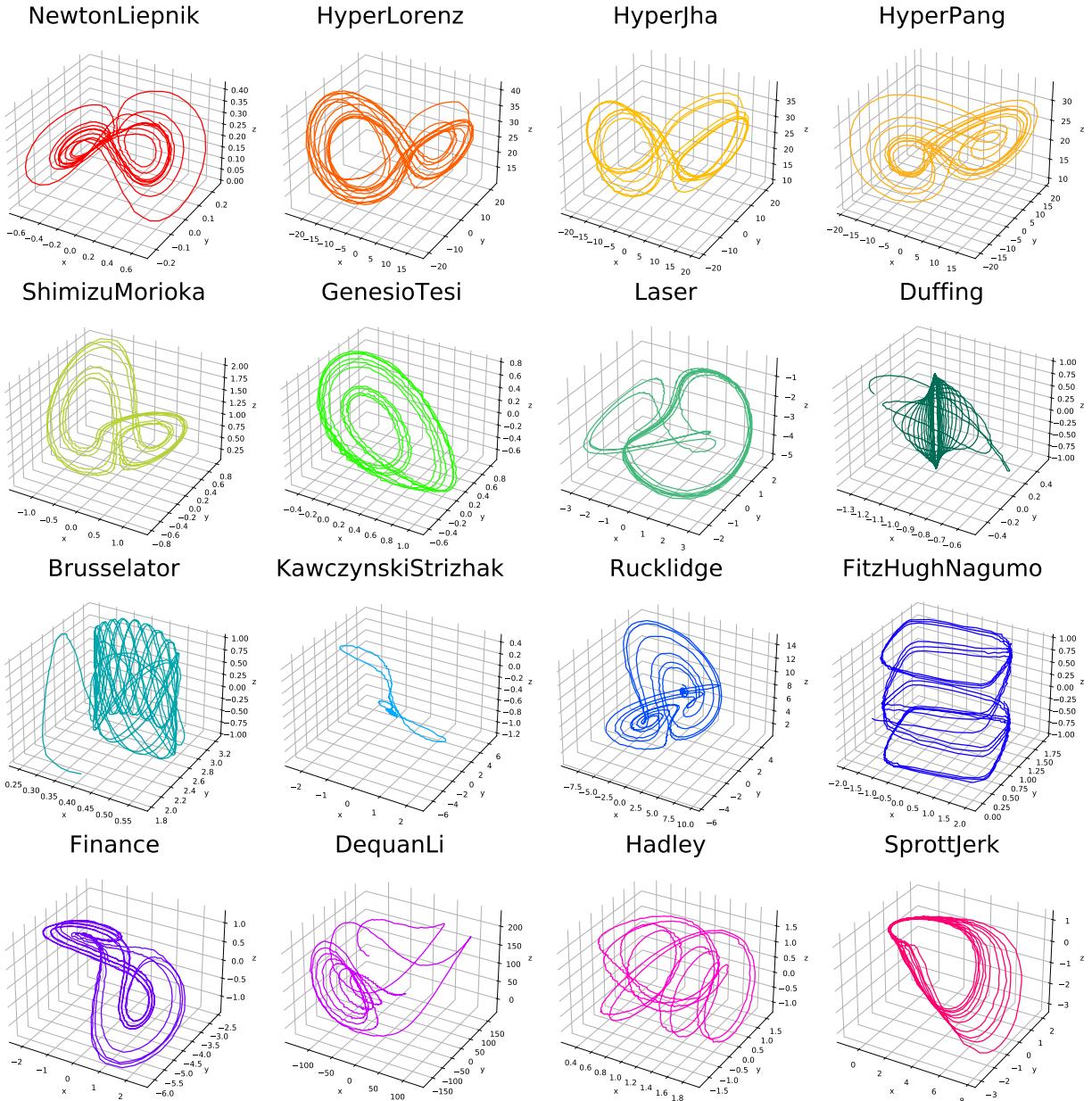


Figure S.2: Visualization of 16 different nonlinear chaotic systems, each with 1% Gaussian noise added. For systems with 4 dimensions, only the first 3 dimensions are shown.

Table S.3: Chaotic dynamics 1 to 4.

Chaotic Dynamic	Governing Equations	Parameters
NewtonLiepnik	$\dot{x} = -ax + y + 10yz$ $\dot{y} = -x - 0.4y + 5xz$ $\dot{z} = bz - 5xy$	$a : 0.4, b : 0.175$
HyperLorenz	$\dot{x} = a(y - x) + w$ $\dot{y} = -xz + cx - y$ $\dot{z} = -bz + xy$ $\dot{w} = dw - xz$	$a : 10, b : 2.667, c : 28, d : 1.1$
HyperJha	$\dot{x} = a(y - x) + w$ $\dot{y} = -xz + bx - y$ $\dot{z} = xy - cz$ $\dot{w} = -xz + dw$	$a : 10, b : 28, c : 2.667, d : 1.3$
HyperPang	$\dot{x} = a(y - x)$ $\dot{y} = -xz + cy + w$ $\dot{z} = xy - bz$ $\dot{w} = -d(x + y)$	$a : 36, b : 3, c : 20, d : 2$

Table S.4: Chaotic dynamics 5 to 8.

Chaotic Dynamic	Governing Equations	Parameters
ShimizuMorioka	$\dot{x} = y$ $\dot{y} = x - ay - xz$ $\dot{z} = -bz + x^2$	$a : 0.85, b : 0.5$
GenesioTesi	$\dot{x} = y$ $\dot{y} = z$ $\dot{z} = -cx - by - az + x^2$	$a : 0.44, b : 1.1, c : 1$
Laser	$\dot{x} = a(y - x) + byz^2$ $\dot{y} = cx + dxz^2$ $\dot{z} = hz + kx^2$	$a : 10.0, b : 1.0, c : 5.0,$ $d : -1.0, h : -5.0, k : -6.0$
Duffing	$\dot{x} = y$ $\dot{y} = -\delta y - \beta x - \alpha x^3$ $\dot{z} = \omega$	$\alpha : 1.0, \beta : -1.0, \delta : 0.1, \omega : 1.4$

Table S.5: Chaotic dynamics 9 to 12.

Chaotic Dynamic	Governing Equations	Parameters
Brusselator	$\dot{x} = a + x^2y - (b + 1)x$ $\dot{y} = bx - x^2y$ $\dot{z} = w$	$a : 0.4, b : 1.2, w : 0.81$
KawczynskiStrizhak	$\dot{x} = \gamma(y - x^3 + 3\mu x)$ $\dot{y} = -2\mu x - y - z + \beta$ $\dot{z} = \kappa(x - z)$	$\beta : -0.4, \gamma : 0.49, \kappa : 0.2, \mu : 2.1$
Rucklidge	$\dot{x} = -ax + by - yz$ $\dot{y} = x$ $\dot{z} = -z + y^2$	$a : 2.0, b : 6.7$
FitzHughNagumo	$\dot{x} = x - x^3/3 - y + curr$ $\dot{y} = \gamma(x + a - by)$ $\dot{z} = \omega$	$a : 0.7, b : 0.8, curr : 0.965,$ $\gamma : 0.08, \omega : 0.04365$

Table S.6: Chaotic dynamics 13 to 16.

Chaotic Dynamic	Governing Equations	Parameters
Finance	$\dot{x} = (1/b - a)x + z + xy$ $\dot{y} = -by - x^2$ $\dot{z} = -x - cz$	$a : 0.001, b : 0.2, c : 1.1$
DequanLi	$\dot{x} = a(y - x) + dxz$ $\dot{y} = kx + fy - xz$ $\dot{z} = cz + xy - \epsilon x^2$	$a : 40, c : 1.833, d : 0.16, \epsilon : 0.65, f : 20, k : 55$
Hadley	$\dot{x} = -y^2 - z^2 - ax + af$ $\dot{y} = xy - bxz - y + g$ $\dot{z} = bxy + xz - z$	$a : 0.2, b : 4, f : 9, g : 1$
SprottJerk	$\dot{x} = y$ $\dot{y} = z$ $\dot{z} = -x + y^2 - \mu z$	$\mu : 2.017$

with SemiSub and SemiDiv to reduce the graphic memory needed.

DGSR: DGSR is designed for symbolic regression with a two-step process [34]. In the first step, it pre-trains an encoder-decoder model on a vast set of mathematical expressions, aiming to capture the posterior distribution. The second step involves refining this distribution during inference to generate symbolic expressions. The framework incorporates a permutation-invariant encoding mechanism and an autoregressive decoding process, leveraging transformer architecture to handle the complexities of SR. We employed the code from <https://github.com/samholt/DeepGenerativeSymbolicRegression>, specifically the version as of September 12, 2023, with default hyperparameters. Since the authors did not provide a pre-trained model for searching expressions containing coefficients, we retrained a constant expression search model with hardcoded constants of 1, 5, and 10, using the operator library $\mathcal{O}_{\text{Koza}}$, following the settings in the repository’s codebase. All other hyperparameters remain at their default values. The operators used are the same as those in PTS.

NGGP: The NGGP [45] model proposes a novel methodology in the realm of symbolic regression by integrating neural-guided search with genetic programming. This hybrid technique leverages a neural network to seed the initial population for the genetic algorithm, aiming to capitalize on the neural network’s pattern recognition capabilities to enhance the search process. The model operates by cycling between generating candidate expressions through a neural sequence generator and refining these candidates using a stateless genetic programming component. We utilized the code from <https://github.com/dso-org/deep-symbolic-optimization>. For symbolic regression benchmark tasks, we employed default hyperparameters, and use operator library $\mathcal{O}_{\text{Koza}}$. In physics data tasks with a specified time budget, we used the same operator sets as PTS and 5 GP iterations, 20000 training samples in each training epoch to avoid excessively long training sessions, and batch size is selected to be 50. The operators used are the same as those in PTS.

PySR: PySR [50] is an open-source symbolic regression library that employs a multi-population genetic algorithm for equation discovery. It features an evolutionary approach through an evolve-simplify-optimize loop, focusing on the enhancement of constant optimization in empirical expressions. PySR utilizes tournament selection and incorporates elements of simulated annealing within its algorithmic structure, aiming to improve the search process. The library operates on a highly optimized Julia backend, SymbolicRegression.jl, which is integral to its performance and efficiency in handling SR tasks. We use the code available at <https://github.com/MilesCranmer/PySR>. For different experiments, we used the same operator sets as PTS and restrict the approximate time budget by setting the number of iterations. In benchmark symbolic regression problems, the number of iterations is selected to be 10000. In chaotic dynamics experiments and EMPS experiment, the number of iterations is selected to be 30. In turbulence experiment, the number of iterations is selected to be 135. The operators used are the same as those in PTS.

BMS: The BMS model [12] introduces a novel Bayesian approach to automated model discovery, aiming to find interpretable mathematical expressions from data. It employs a Markov chain Monte Carlo (MCMC) algorithm to navigate the space of possible models, effectively sampling from the posterior distribution over mathematical expressions. BMS uses a maximum entropy principle to learn a prior over expressions from a large empirical corpus, typically compiled from scientific literature, such as Wikipedia. We use the code available at <https://bitbucket.org/riguimera/machine-scientist/>. The prior file used in turbulent experiments is nv1.np5.2017-10-18 18_07_

`35.227360.dat` given in the repository. For EMPS, we used prior file with `nv3.np3`. For chaotic dynamics, we utilized `nv3.np3`, `nv4.np8` and `nv5.np7` corresponding to the number of variables involved. The model parameters `nsample`, `thin`, and `burnin` are selected to be 100, 10, and 500 respectively for limiting the time. The operators used are the same as those in PTS.

SPL: The SPL model [42] employs a Monte Carlo tree search (MCTS) agent for traversing expression trees in pursuit of the most suitable expressions that align with the fundamental physics. This model adeptly incorporates existing knowledge and specific constraints within its grammatical structure, showcasing a strong capability to identify intricate expressions effectively. We utilized the code from <https://github.com/isds-neu/SymbolicPhysicsLearner>. And we used the result from Sun *et al.* [42] and Xu *et al.* [44].

2.4 Evaluation Metrics

Here, we will introduce the definitions of some key evaluation metrics:

Complexity: The number of operators in an expression. A higher complexity often suggests a greater risk of overfitting when the expression is used in a predictive model. This concept is consistent with the principle of Occam’s razor, which posits that, all else being equal, simpler explanations are generally preferable to more complex ones.

Recovery Rate: The proportion of instances where SR algorithms identify symbolically equivalent expressions compared to the ground truth, measured across numerous independent trials with different random seeds. In the symbolic regression benchmark test, the found expressions must be symbolically equivalent to the ground truth expression (validated using SymPy [74]). For symbolic regression benchmark expressions containing constants, the constants are trimmed to two decimal places before comparison. In chaotic dynamic discovery task, due to the presence of Gaussian noise, an expression is considered successfully recovered if it has the same structural form as the ground truth expression (allowing for a constant bias term).

2.5 Hardware and Environment

The experiments were performed on servers with Nvidia A100 (80GB) and Intel(R) Xeon(R) Platinum 8380 cpus @ 2.30GHz. Note that for environments running the PTS model, a minimum Pytorch version of 2.0.0 is required. Versions below this threshold may be incompatible with the `torch.topk` operation and may result in computational issues.

3 Results

The following table shows the detailed recovery rates and time cost on the symbolic regression benchmark problem sets and 16 chaotic dynamics.

3.1 Symbolic Regression Benchmark Problem

Table S.7, S.8, S.9, S.10, S.11, S.12 show the recovery rate and time cost on each symbolic regression benchmark problem set.

Table S.7: Symbolic regression results on Nguyen problem set.

Benchmark	Recovery rate					Time cost (s)				
	PTS	SPL	NGGP	DGSR	PySR	PTS	SPL	NGGP	DGSR	PySR
Nguyen-1	100%	100%	100%	100%	100%	1.10	8.78	13.45	10.01	2.10
Nguyen-2	100%	100%	100%	100%	99%	1.11	7.30	18.20	8.25	2.17
Nguyen-3	100%	100%	100%	100%	100%	2.83	81.29	34.86	8.45	4.15
Nguyen-4	100%	99%	100%	100%	97%	3.31	567.06	71.65	27.16	258.13
Nguyen-5	100%	95%	99%	98%	100%	2.43	431.23	175.45	69.51	577.04
Nguyen-6	100%	100%	99%	0%	100%	2.95	64.65	122.97	7.29	1.96
Nguyen-7	100%	100%	100%	15%	100%	15.29	14.99	126.60	970.35	78.67
Nguyen-8	100%	100%	100%	100%	53%	1.12	5.59	63.27	68.88	212.03
Nguyen-9	100%	100%	100%	100%	100%	2.06	5.74	17.81	8.19	2.91
Nguyen-10	100%	100%	100%	100%	100%	2.08	53.24	45.06	79.70	4.43
Nguyen-11	100%	100%	100%	23%	100%	2.10	10.16	22.55	847.79	57.95
Nguyen-12	98%	28%	0%	0%	0%	389.74	187.90	857.18	1130.91	6933.52

Table S.8: Symbolic regression results on Nguyen-c problem set.

Benchmark	Recovery rate					Time cost (s)				
	PTS	SPL	NGGP	DGSR	PySR	PTS	SPL	NGGP	DGSR	PySR
Nguyen-1c	100%	100%	76%	45%	65%	24.77	452.73	804.64	181.72	666.11
Nguyen-2c	100%	94%	16%	18%	0%	41.19	295.77	1122.97	541.87	2047.44
Nguyen-5c	100%	95%	28%	41%	85%	15.48	2178.89	8649.82	4158.69	671.19
Nguyen-8c	100%	100%	100%	90%	100%	103.62	77.89	701.43	845.45	73.60
Nguyen-9c	99%	98%	60%	0%	100%	1332.79	2001.40	21005.76	7043.42	46.42
Nguyen-10c	100%	0%	80%	9%	65%	226.26	3672.39	6653.39	6315.63	4413.90

Table S.9: Symbolic regression results on R problem set.

Benchmark	Recovery rate					Time cost (s)				
	PTS	SPL	NGGP	DGSR	PySR	PTS	SPL	NGGP	DGSR	PySR
R-1	100%	0%	0%	0%	0%	21.47	1507.01	759.25	1148.08	2196.87
R-2	98%	0%	0%	0%	0%	460.96	1397.70	722.63	1091.35	2212.13
R-3	100%	0%	7%	0%	0%	84.87	1275.47	820.86	1147.61	2132.06

Table S.10: Symbolic regression results on R* problem set.

Benchmark	Recovery rate					Time cost (s)				
	PTS	SPL	NGGP	DGSR	PySR	PTS	SPL	NGGP	DGSR	PySR
R*-1	100%	0%	26%	0%	0%	16.37	1296.93	880.24	1034.55	2178.47
R*-2	100%	0%	40%	0%	0%	371.57	1521.33	775.00	1059.25	1724.86
R*-3	100%	0%	95%	58%	0%	8.02	1275.47	193.51	444.13	1874.46

Table S.11: Symbolic regression results on Livermore problem set.

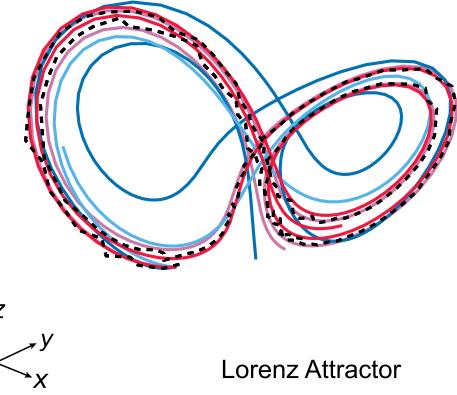
Benchmark	Recovery rate					Time cost (s)				
	PTS	SPL	NGGP	DGSR	PySR	PTS	SPL	NGGP	DGSR	PySR
Livermore-1	100%	94%	100%	76%	89%	12.06	56.39	77.48	895.84	29.98
Livermore-2	100%	29%	100%	100%	55%	2.48	1568.58	100.95	58.38	1378.44
Livermore-3	100%	50%	96%	100%	86%	33.18	121.77	205.71	84.68	398.74
Livermore-4	100%	61%	100%	100%	100%	15.81	661.39	29.00	42.79	6.56
Livermore-5	100%	100%	100%	0%	99%	8.42	1714.67	65.75	1060.11	131.11
Livermore-6	100%	8%	98%	100%	0%	13.82	1958.34	189.15	152.36	2063.86
Livermore-7	100%	18%	0%	0%	0%	1.16	1362.29	643.53	1140.42	36.12
Livermore-8	100%	6%	0%	0%	0%	1.16	946.12	672.93	1140.30	89.14
Livermore-9	100%	21%	79%	43%	0%	43.14	1375.24	405.70	924.97	2103.65
Livermore-10	100%	75%	5%	35%	68%	10.63	1773.06	813.33	1057.65	3418.58
Livermore-11	100%	0%	100%	18%	80%	2.07	1684.00	66.84	12.43	5028.28
Livermore-12	100%	100%	62%	100%	39%	2.07	1795.80	117.98	9.24	5396.28
Livermore-13	100%	12%	97%	98%	80%	8.26	2665.69	53.60	93.35	172.28
Livermore-14	100%	100%	99%	0%	100%	2.41	1875.82	143.00	52.13	9.92
Livermore-15	100%	0%	98%	100%	37%	122.61	2662.18	133.97	187.01	8.18
Livermore-16	100%	0%	97%	0%	7%	127.50	3700.33	184.83	786.57	216.37
Livermore-17	100%	89%	45%	64%	100%	2.07	637.18	642.14	629.29	227.37
Livermore-18	100%	18%	60%	100%	0%	72.95	1132.71	600.62	146.83	1706.93
Livermore-19	100%	89%	100%	100%	100%	2.32	2043.58	23.67	7.36	4.55
Livermore-20	100%	100%	100%	100%	17%	1.09	3003.36	14.77	14.73	16.52
Livermore-21	100%	52%	94%	100%	0%	13.18	1970.23	294.23	58.87	1944.50
Livermore-22	100%	100%	83%	32%	0%	6.15	1776.83	337.37	872.14	4.25

Table S.12: Symbolic regression results on Feynman problem set.

Benchmark	Recovery rate					Time cost (s)				
	PTS	SPL	NGGP	DGSR	PySR	PTS	SPL	NGGP	DGSR	PySR
Feynman-1	100%	100%	100%	100%	100%	2.25	85.68	12.22	6.71	2.30
Feynman-2	100%	0%	99%	100%	88%	2.25	1181.64	45.09	41.25	30.54
Feynman-3	100%	100%	100%	100%	98%	2.25	88.65	11.84	2.88	40.27
Feynman-4	100%	0%	0%	0%	0%	40.08	1061.49	711.07	725.39	6074.75
Feynman-5	100%	100%	100%	95%	97%	2.25	78.55	12.44	3.08	109.72
Feynman-6	100%	0%	100%	100%	96%	2.25	1316.97	122.17	35.03	190.46
Feynman-7	100%	0%	100%	100%	73%	2.24	1301.57	69.20	15.54	93.19
Feynman-8	0%	0%	0%	0%	0%	933.74	1224.00	920.51	898.20	10694.80
Feynman-9	100%	0%	95%	55%	34%	2.27	1084.67	295.75	517.40	3359.48
Feynman-10	100%	0%	100%	100%	39%	2.26	1192.10	45.36	23.52	3267.69
Feynman-11	100%	0%	25%	100%	71%	2.26	1136.84	800.94	342.79	6531.40
Feynman-12	100%	0%	46%	95%	34%	29.06	1181.74	661.46	331.04	2475.89
Feynman-13	0%	0%	14%	10%	32%	920.71	1164.30	832.85	880.70	6402.28
Feynman-14	100%	0%	99%	100%	34%	2.26	1189.21	225.26	100.04	3404.15
Feynman-15	100%	0%	100%	100%	7%	40.84	1152.71	46.11	11.41	3736.32

True Equations:

$$\begin{aligned}\dot{x} &= -10x + 10y \\ \dot{y} &= -xz + 28x - y \\ \dot{z} &= xy - 2.667z\end{aligned}$$



Discovered Equations:

$$\begin{aligned}\dot{x} &= -9.95x + 9.96y \\ \dot{y} &= -0.98xz + 27.32x - 0.76y + 0.04 \\ \dot{z} &= 0.99xy - 2.65z - 0.04\end{aligned}$$

$$\begin{aligned}\dot{x} &= 9.9555y - 9566.4353\sin(1.0 \cdot 10^{-5}x\text{sign}(z)) \\ \dot{y} &= -0.9864xz + 27.5126x - 0.9018y \\ \dot{z} &= 0.9811xy + \frac{0.2724xy}{z} - 2.6189z - 0.7271\end{aligned}$$

$$\begin{aligned}\dot{x} &= -9.9565x + 9.9565y \\ \dot{y} &= -0.0286x|x||z| + 7.2958y + \frac{0.0381|x||z|}{x} \\ \dot{z} &= xy - 2.6694z\end{aligned}$$

$$\begin{aligned}\dot{x} &= 0.0041xy - 9.9511x - 0.0041y^2 + 9.9522y + 0.067 \\ \dot{y} &= -0.0311xz^2 + 0.8968xz \\ \dot{z} &= 0.9601xy + 0.0173x - 0.9601z - 37.6876\end{aligned}$$

Figure S.3: PTS discovered true equation structure of Lorenz Attractor in the same time budget.

3.2 Chaotic Dynamics

Before performing on these 16 nonlinear chaotic dynamics, we first test the performance of our algorithm on the well-known Lorenz attractor [51] and compare it with other algorithms as an example. The equation for the Lorenz attractor is given below.

$$\dot{x} = \sigma(y - x) \quad (11)$$

$$\dot{y} = x(\rho - z) - y \quad (12)$$

$$\dot{z} = xy - \beta z \quad (13)$$

We set the candidate binary operators as $+$, $-$, \times , and \div , while the candidate unary operators were \sin , \cos , \exp , \log , abs , sign , \tanh , \cosh . We simulated and generated trajectory samples of the Lorenz attractor, incorporating 1% Gaussian noise. We generated 1000 training sample points using the default initial state of the Lorenz attractor from the dysts library. Subsequently, we changed the initial state to $(1, 1, 1)^\top$ and regenerated 1000 test sample points for evaluation purposes. We applied symbolic regression using PTS, BMS [12], PySR [50], and NGGP [45] to the synthesized data, and the time is limited to around 30 seconds by limiting their hyperparameters. As depicted in Fig. S.3, employing the same operator set and time constraint, our model successfully identified the equation structure of the Lorenz attractor.

Table S.13 shows the average recovery rate and time cost on other 16 chaotic dynamics. Our PTS model is capable of identifying the underlying control equations within chaotic dynamics data more rapidly and accurately under a given time budget. Generally speaking, the ability to precisely recover the most complex symbolic expressions within a system of equations determines the overall effectiveness of the system's reconstruction. We have observed that these challenging expressions can be discovered by the PTS with a significantly higher recovery rate, demonstrating the superiority of our method.

Table S.13: Recovery rate and average time cost of chaotic dynamics task

Chaotic Dynamic	Variant	Recovery Rate (%)				Avg. Time Cost			
		PTS	BMS	PySR	NGGP	PTS	BMS	PySR	NGGP
DequanLi	\dot{x}	56.0	44.4	2.0	4.0	106.43	91.35	91.24	92.94
	\dot{y}	100.0	72.2	92.0	12.0	79.61	91.46	92.43	92.9
	\dot{z}	94.0	70.4	100.0	2.0	78.04	91.31	84.96	93.33
Duffing	\dot{x}	100.0	90.9	100.0	100.0	38.83	44.55	70.32	93.27
	\dot{y}	58.0	9.1	8.0	0.0	42.78	91.45	93.49	96.13
	\dot{z}	100.0	65.5	100.0	50.0	51.84	58.89	85.38	91.69
Finance	\dot{x}	100.0	68.5	24.0	6.0	28.5	91.37	104.8	94.28
	\dot{y}	100.0	64.8	62.0	4.0	29.25	91.39	96.57	92.9
	\dot{z}	100.0	66.7	100.0	82.0	32.28	90.95	86.97	93.1
Brusselator	\dot{x}	52.0	14.5	0.0	0.0	63.81	91.58	100.67	97.52
	\dot{y}	96.0	52.7	50.0	2.0	48.71	91.38	97.24	96.97
	\dot{z}	100.0	58.2	100.0	46.0	55.89	81.86	88.7	94.39
FitzHughNagumo	\dot{x}	68.0	0.0	0.0	2.0	47.55	91.24	111.04	94.38
	\dot{y}	100.0	81.5	100.0	54.0	27.88	91.25	91.33	93.81
	\dot{z}	100.0	79.6	100.0	94.0	30.73	44.99	77.03	93.16
GenesioTesi	\dot{x}	100.0	96.4	100.0	100.0	29.19	39.18	71.51	95.11
	\dot{y}	100.0	96.4	100.0	100.0	29.29	43.55	72.8	94.1
	\dot{z}	100.0	0.0	8.0	0.0	39.84	91.0	97.48	95.87
Hadley	\dot{x}	30.0	0.0	8.0	0.0	32.61	91.56	107.73	94.68
	\dot{y}	14.0	7.5	58.0	0.0	24.02	91.39	94.34	94.11
	\dot{z}	100.0	35.8	80.0	2.0	33.78	92.69	89.35	94.67
HyperJha	\dot{w}	96.0	74.5	100.0	18.0	31.17	91.2	86.57	91.82
	\dot{x}	100.0	64.3	100.0	40.0	34.38	90.86	83.18	92.26
	\dot{y}	68.0	33.9	20.0	0.0	49.04	91.26	92.56	92.57
	\dot{z}	100.0	69.1	100.0	18.0	29.28	91.58	89.17	92.72
HyperLorenz	\dot{w}	94.0	60.7	100.0	30.0	34.28	91.46	82.69	92.42
	\dot{x}	96.0	82.1	100.0	42.0	36.52	91.4	84.33	92.36
	\dot{y}	78.0	66.1	32.0	4.0	52.54	91.63	91.61	92.85
	\dot{z}	100.0	62.5	100.0	20.0	29.44	91.65	85.81	90.81
HyperPang	\dot{w}	100.0	69.1	100.0	82.0	29.66	90.76	86.75	91.89
	\dot{x}	100.0	69.1	100.0	52.0	31.43	91.4	82.49	92.73
	\dot{y}	100.0	56.4	16.0	0.0	37.72	91.81	87.71	93.34
	\dot{z}	100.0	56.4	100.0	22.0	32.0	92.71	86.67	92.37
KawczynskiStrizhak	\dot{x}	52.0	0.0	0.0	0.0	32.34	91.51	110.81	94.53
	\dot{y}	100.0	89.1	40.0	12.0	28.07	91.54	97.16	94.27
	\dot{z}	100.0	79.6	100.0	76.0	24.4	91.23	90.61	93.63
Laser	\dot{x}	98.0	34.5	16.0	0.0	32.33	91.61	94.95	94.18
	\dot{y}	100.0	32.7	92.0	2.0	28.44	91.68	90.94	93.88
	\dot{z}	100.0	63.6	76.0	6.0	30.62	91.53	102.34	93.53
NewtonLiepnik	\dot{x}	100.0	69.6	28.0	0.0	39.41	91.25	98.09	94.6
	\dot{y}	98.0	48.2	54.0	2.0	34.53	91.41	94.87	95.67
	\dot{z}	100.0	51.8	100.0	0.0	29.51	91.32	89.34	94.84
Rucklidge	\dot{x}	80.0	29.6	14.0	2.0	49.62	91.01	105.27	94.81
	\dot{y}	96.0	96.3	100.0	100.0	27.01	42.79	74.24	92.06
	\dot{z}	100.0	68.5	90.0	22.0	27.74	91.48	93.51	93.18
ShimizuMorioka	\dot{x}	100.0	92.7	100.0	100.0	29.38	41.1	70.33	94.3
	\dot{y}	100.0	34.5	70.0	6.0	37.19	91.23	96.18	93.89
	\dot{z}	100.0	65.5	100.0	8.0	24.31	91.98	94.63	94.01
SprottJerk	\dot{x}	100.0	94.3	100.0	100.0	24.98	41.22	72.13	92.72
	\dot{y}	100.0	100.0	100.0	100.0	23.88	44.13	72.81	93.69
	\dot{z}	100.0	5.7	0.0	0.0	24.19	91.15	103.6	92.85

References

- [1] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.
- [2] Baicheng Weng, Zhilong Song, Rilong Zhu, Qingyu Yan, Qingde Sun, Corey G Grice, Yanfa Yan, and Wan-Jian Yin. Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts. *Nature Communications*, 11(1):3513, 2020.
- [3] Cristina Cornelio, Sanjeeb Dash, Vernon Austel, Tyler R Josephson, Joao Goncalves, Kenneth L Clarkson, Nimrod Megiddo, Bachir El Khadir, and Lior Horesh. Combining data and theory for derivable scientific discovery with AI-Descartes. *Nature Communications*, 14(1):1777, 2023.
- [4] Digvijay Wadekar, Leander Thiele, Francisco Villaescusa-Navarro, J Colin Hill, Miles Cranmer, David N Spergel, Nicholas Battaglia, Daniel Anglés-Alcázar, Lars Hernquist, and Shirley Ho. Augmenting astrophysical scaling relations with machine learning: Application to reducing the sunyaev–zeldovich flux–mass scatter. *Proceedings of the National Academy of Sciences*, 120(12):e2202074120, 2023.
- [5] Yanzhang Li, Hongyu Wang, Yan Li, Huan Ye, Yanan Zhang, Rongzhang Yin, Haoning Jia, Bingxu Hou, Changqiu Wang, Hongrui Ding, et al. Electron transfer rules of minerals under pressure informed by machine learning. *Nature Communications*, 14(1):1815, 2023.
- [6] Dion Häfner, Johannes Gemmrich, and Markus Jochum. Machine-guided discovery of a real-world rogue wave model. *Proceedings of the National Academy of Sciences*, 120(48):e2306275120, 2023.
- [7] Yuanyuan Chong, Yaoyuan Huo, Shuang Jiang, Xijun Wang, Baichen Zhang, Tianfu Liu, Xin Chen, TianTian Han, Pieter Ernst Scholtz Smith, Song Wang, and Jun Jiang. Machine learning of spectra-property relationship for imperfect and small chemistry data. *Proceedings of the National Academy of Sciences*, 120(20):e2220789120, 2023.
- [8] Hendrik Jung, Roberto Covino, A Arjun, Christian Leitold, Christoph Dellago, Peter G Bolhuis, and Gerhard Hummer. Machine-guided path sampling to discover mechanisms of molecular self-organization. *Nature Computational Science*, pages 1–12, 2023.
- [9] Matthew Golden, Roman O Grigoriev, Jyothishraj Nambisan, and Alberto Fernandez-Nieves. Physically informed data-driven modeling of active nematics. *Science Advances*, 9(27):eabq6120, 2023.
- [10] Xiaoli Chen, Beatrice W Soh, Zi-En Ooi, Eleonore Vissol-Gaudin, Haijun Yu, Kostya S Novoselov, Kedar Hippalgaonkar, and Qianxiao Li. Constructing custom thermodynamics using deep learning. *Nature Computational Science*, 4:66–85, 2023.
- [11] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied Logistic Regression*, volume 398. John Wiley & Sons, 2013.
- [12] Roger Guimerà, Ignasi Reichardt, Antoni Aguilar-Mogas, Francesco A Massucci, Manuel Miranda, Jordi Pallarès, and Marta Sales-Pardo. A bayesian machine scientist to aid in the solution of challenging scientific problems. *Science Advances*, 6(5):eaav6971, 2020.
- [13] Stephanie Forrest. Genetic algorithms: principles of natural selection applied to computation. *Science*, 261(5123):872–878, 1993.

- [14] Josh Bongard and Hod Lipson. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 104(24):9943–9948, 2007.
- [15] Daniel L Ly and Hod Lipson. Learning symbolic representations of hybrid dynamical systems. *Journal of Machine Learning Research*, 13(1):3585–3618, 2012.
- [16] Markus Quade, Markus Abel, Kamran Shafi, Robert K Niven, and Bernd R Noack. Prediction of dynamical systems by symbolic regression. *Physical Review E*, 94(1):012214, 2016.
- [17] Harsha Vaddireddy, Adil Rasheed, Anne E Staples, and Omer San. Feature engineering and symbolic regression methods for detecting hidden physics from sparse sensor observation data. *Physics of Fluids*, 32(1), 2020.
- [18] He Ma, Arunachalam Narayanaswamy, Patrick Riley, and Li Li. Evolving symbolic density functionals. *Science Advances*, 8(36):eabq0279, 2022.
- [19] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.
- [20] Samuel H Rudy, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017.
- [21] Zichao Long, Yiping Lu, Xianzhong Ma, and Bin Dong. PDE-Net: Learning pdes from data. In *International Conference on Machine Learning*, pages 3208–3216, 2018.
- [22] Kathleen Champion, Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116(45):22445–22451, 2019.
- [23] Zhenlin Wang, Xun Huan, and Krishna Garikipati. Variational system identification of the partial differential equations governing the physics of pattern-formation: Inference under varying fidelity and noise. *Computer Methods in Applied Mechanics and Engineering*, 356:44–74, 2019.
- [24] Zhao Chen, Yang Liu, and Hao Sun. Physics-informed learning of governing equations from scarce data. *Nature Communications*, 12(1):6136, 2021.
- [25] Chengping Rao, Pu Ren, Qi Wang, Oral Buyukozturk, Hao Sun, and Yang Liu. Encoding physics to learn reaction–diffusion processes. *Nature Machine Intelligence*, 5(7):765–779, 2023.
- [26] Ting-Ting Gao and Gang Yan. Autonomous inference of complex network dynamics from incomplete and noisy data. *Nature Computational Science*, 2(3):160–168, 2022.
- [27] Kevin Course and Prasanth B Nair. State estimation of a physical system with unknown governing equations. *Nature*, 622(7982):261–267, 2023.
- [28] Brenden K Petersen, Mikel Landajuela Larma, Terrell N. Mundhenk, Claudio Prata Santiago, Soo Kyung Kim, and Joanne Taery Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. In *International Conference on Learning Representations*, 2021.
- [29] Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International Conference on Machine Learning*, pages 1945–1954, 2017.

- [30] Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020.
- [31] Silviu-Marian Udrescu, Andrew Tan, Jiahai Feng, Orisvaldo Neto, Tailin Wu, and Max Tegmark. AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. *Advances in Neural Information Processing Systems*, 33:4860–4871, 2020.
- [32] Luca Biggio, Tommaso Bendinelli, Alexander Neitz, Aurelien Lucchi, and Giambattista Pascaniolo. Neural symbolic regression that scales. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 936–945, 2021.
- [33] Pierre-Alexandre Kamienny, Stéphane d’Ascoli, Guillaume Lample, and François Charton. End-to-end symbolic regression with transformers. *Advances in Neural Information Processing Systems*, 35:10269–10281, 2022.
- [34] Samuel Holt, Zhaozhi Qian, and Mihaela van der Schaar. Deep generative symbolic regression. In *The Eleventh International Conference on Learning Representations*, 2023.
- [35] Subham Sahoo, Christoph Lampert, and Georg Martius. Learning equations for extrapolation and control. In *International Conference on Machine Learning*, pages 4442–4450, 2018.
- [36] Zichao Long, Yiping Lu, and Bin Dong. PDE-Net 2.0: Learning pdes from data with a numeric-symbolic hybrid deep network. *Journal of Computational Physics*, 399:108925, 2019.
- [37] Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In *International Conference on Computers and Games*, pages 72–83. Springer, 2006.
- [38] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European Conference on Machine Learning*, pages 282–293. Springer, 2006.
- [39] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.
- [40] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.
- [41] Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022.
- [42] Fangzheng Sun, Yang Liu, Jian-Xun Wang, and Hao Sun. Symbolic physics learner: Discovering governing equations via monte carlo tree search. In *The Eleventh International Conference on Learning Representations*, 2023.
- [43] Pierre-Alexandre Kamienny, Guillaume Lample, Sylvain Lamprier, and Marco Virgolin. Deep generative symbolic regression with monte-carlo-tree-search. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

- [44] Yilong Xu, Yang Liu, and Hao Sun. Reinforcement symbolic regression machine. In *The Twelfth International Conference on Learning Representations*, 2024.
- [45] Terrell Mundhenk, Mikel Landajuela, Ruben Glatt, Claudio P Santiago, Brenden K Petersen, et al. Symbolic regression via deep reinforcement learning enhanced genetic programming seeding. *Advances in Neural Information Processing Systems*, 34:24912–24923, 2021.
- [46] Marco Virgolin and Solon P Pissis. Symbolic regression is NP-hard. *Transactions on Machine Learning Research*, 2022.
- [47] Nguyen Quang Uy, Nguyen Xuan Hoai, Michael O’Neill, Robert I McKay, and Edgar Galván-López. Semantically-based crossover in genetic programming: application to real-valued symbolic regression. *Genetic Programming and Evolvable Machines*, 12:91–119, 2011.
- [48] James McDermott, David R White, Sean Luke, Luca Manzoni, Mauro Castelli, Leonardo Vanneschi, Wojciech Jaskowski, Krzysztof Krawiec, Robin Harper, Kenneth De Jong, et al. Genetic programming needs better benchmarks. In *Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation*, pages 791–798, 2012.
- [49] Krzysztof Krawiec and Tomasz Pawlak. Approximating geometric crossover by semantic back-propagation. In *Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation*, pages 941–948, 2013.
- [50] Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv preprint arXiv:2305.01582*, 2023.
- [51] Edward N Lorenz. Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, 20(2):130–141, 1963.
- [52] William Gilpin. Chaos as an interpretable benchmark for forecasting and data-driven modelling. In *Thirty-fifth Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.
- [53] Alexandre Janot, Maxime Gautier, and Mathieu Brunot. Data set and reference models of EMPS. In *Workshop on Nonlinear System Identification Benchmarks*, 2019.
- [54] Johann Nikuradse et al. Laws of flow in rough pipes. Technical Report (No. NACA-TM-1292), 1950.
- [55] Ludwig Prandtl. Recent results of turbulence research. Technical Report (No. NACA-TM-720), 1933.
- [56] Nigel Goldenfeld. Roughness-induced critical phenomena in a turbulent flow. *Physical Review Letters*, 96(4):044503, 2006.
- [57] Jianjun Tao. Critical instability and friction scaling of fluid flows through pipes with rough inner surfaces. *Physical Review Letters*, 103:264502, 2009.
- [58] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

- [59] Ryosuke Okuta, Yuya Unno, Daisuke Nishino, Shohei Hido, and Crissman Loomis. Cupy: A NumPy-compatible library for NVIDIA GPU calculations. In *Proceedings of the Thirty-first Conference on Neural Information Processing Systems*, 2017.
- [60] RB Leipnik and TA Newton. Double strange attractors in rigid body motion with linear feedback control. *Physics Letters A*, 86(2):63–67, 1981.
- [61] Shouquan Pang and Yongjian Liu. A new hyperchaotic system from the lü system and its control. *Journal of Computational and Applied Mathematics*, 235(8):2775–2789, 2011.
- [62] T Shimizu and N Morioka. On the bifurcation of a symmetric limit cycle to an asymmetric one in a simple model. *Physics Letters A*, 76(3-4):201–204, 1980.
- [63] Roberto Genesio and Alberto Tesi. Harmonic balance methods for the analysis of chaotic dynamics in nonlinear systems. *Automatica*, 28(3):531–548, 1992.
- [64] A Aboee, HA Yaghini-Bonabi, and Mohammad Reza Jahed-Motlagh. Analysis and circuitry realization of a novel three-dimensional chaotic system. *Communications in Nonlinear Science and Numerical Simulation*, 18(5):1235–1245, 2013.
- [65] Georg Duffing. *Erzwungene Schwingungen bei veränderlicher Eigenfrequenz und ihre technische Bedeutung*. Number 41-42. Vieweg, 1918.
- [66] Ilya Prigogine. *From Being to Becoming Time and Complexity in the Physical Sciences*. W.H. Freeman, C1980, 1980.
- [67] Peter E Strizhak and Andrzej L Kawczynski. Complex transient oscillations in the belousov-zhabotinskii reaction in a batch reactor. *The Journal of Physical Chemistry*, 99(27):10830–10833, 1995.
- [68] Alastair Michael Rucklidge. Chaos in models of double convection. *Journal of Fluid Mechanics*, 237:209–229, 1992.
- [69] Richard FitzHugh. Impulses and physiological states in theoretical models of nerve membrane. *Biophysical Journal*, 1(6):445–466, 1961.
- [70] Guoliang Cai and Juanjuan Huang. A new finance chaotic attractor. *International Journal of Nonlinear Science*, 3(3):213–220, 2007.
- [71] Dequan Li. A three-scroll chaotic attractor. *Physics Letters A*, 372(4):387–393, 2008.
- [72] George Hadley. On the cause of the general trade winds. *Philosophical Transactions of the Royal Society*, 34:58–62, 1735.
- [73] JC Sprott. Simplest dissipative chaotic flow. *Physics Letters A*, 228(4-5):271–274, 1997.
- [74] Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, 2017.