# StruSR: Structure-Aware Symbolic Regression with Physics-Informed Taylor Guidance

**Yunpeng Gong[1], Sihan Lan[1], Can Yang[1], Kunpeng Xu[2], Min Jiang[1]**

[1]School of Informatics, Xiamen University, Xiamen, China
[2]Université de Sherbrooke, Québec, Canada

## Abstract

Symbolic regression aims to find interpretable analytical expressions by searching over mathematical formula spaces to capture underlying system behavior, particularly in scientific modeling governed by physical laws. However, traditional methods lack mechanisms for extracting structured physical priors from time series observations, making it difficult to capture symbolic expressions that reflect the system's global behavior. In this work, we propose a structure-aware symbolic regression framework—StruSR—that leverages trained Physics-Informed Neural Networks (PINNs) to extract locally structured physical priors from time series data. By performing local Taylor expansions on the outputs of the trained PINN, we obtain derivative-based structural information to guide symbolic expression evolution. To assess the importance of expression components, we introduce a masking-based attribution mechanism that quantifies each subtree's contribution to structural alignment and physical residual reduction. These sensitivity scores steer mutation and crossover operations within genetic programming, preserving substructures with high physical or structural significance while selectively modifying less informative components. A hybrid fitness function jointly minimizes physics residuals and Taylor coefficient mismatch, ensuring consistency with both the governing equations and the local analytical behavior encoded by the PINN. Experiments on benchmark PDE systems demonstrate that StruSR improves convergence speed, structural fidelity, and expression interpretability compared to conventional baselines, offering a principled paradigm for physics-grounded symbolic discovery.

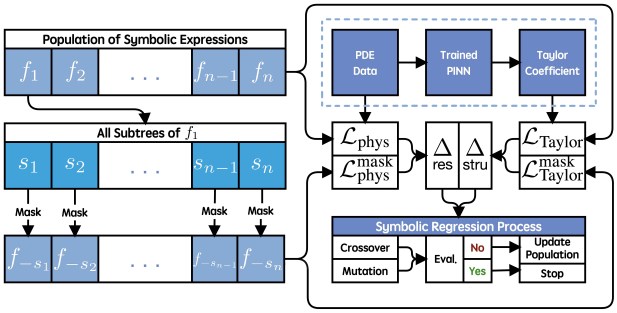



Figure 1: Overview of the proposed Taylor-structure-guided symbolic regression framework. From the left, a population of symbolic expressions is maintained and each expression is decomposed into subtrees. Masking individual subtrees allows evaluating their contributions to physics residual loss ($\mathcal{L}_{phys}$) and Taylor-based structural loss ($\mathcal{L}_{Taylor}$), the latter derived from local Taylor coefficients extracted from a Physics-Informed Neural Network (PINN) trained on PDE data. Sensitivity scores computed from these loss variations guide crossover and mutation in genetic programming, preserving structurally and physically important subexpressions. The symbolic regression process iterates with population updates until convergence or stopping criteria are met.

## Introduction

In physical field simulations, partial differential equations (PDEs) are employed to model the spatiotemporal evolution of physical quantities, where the time series [38, 22, 39, 19, 40, 45, 43, 29] at specific spatial locations reflect the system's dynamic characteristics; symbolic regression, by searching for the underlying analytical structure of the governing PDEs, enables interpretable and generalizable modeling of complex physical systems. Symbolic regression (SR) aims to discover closed-form mathematical expressions that describe the underlying relationships in data. With the rapid advancement of deep learning [16, 11, 9, 12, 15, 10, 42, 14, 8, 13, 21, 20], numerous new methods have emerged for solving partial differential equations. Unlike black-box models such as deep neural networks, SR offers human-interpretable outputs, making it particularly valuable in scientific modeling and systems governed by physical laws. In recent years, symbolic regression has seen renewed interest through integration with neural architectures and evolutionary algorithms, yet a persistent challenge remains: how to guide the search toward expressions that are not only accurate but also structurally aligned with physical principles.

Most existing SR methods focus on data fitting accuracy, often overlooking whether the discovered expression reflects the correct physical structure. This shortcoming becomes especially pronounced in the context of discovering governing equations or partial differential equations (PDEs), where expressions must satisfy physical constraints and exhibit meaningful derivative behavior. While PINNs [30] offer a promising way to approximate solutions to differen-

tial equations by embedding physics into training objectives, their outputs are inherently non-symbolic and lack direct interpretability.

To address this gap, we propose a novel framework that leverages the structural insights encoded in trained PINNs to guide the symbolic regression process. The key idea of our approach is to leverage the local Taylor expansions of a trained PINN solution as structural "ground truth" for guiding symbolic expression discovery. By comparing the Taylor coefficients of candidate expressions against those extracted from the PINN, we can quantitatively assess structural similarity in a manner that is both principled and robust. This allows us to inject domain knowledge—derived from a continuous, differentiable, and physically meaningful PINN representation—into the symbolic search process in a principled way.

From a deeper perspective, a trained PINN should not be regarded merely as a numerical predictor for PDE solutions—it acts as a global encoder of physical structure. Through minimizing residuals of differential operators across the entire domain, PINNs embed physical consistency into the learned parameter space. Although the analytic solution remains hidden within the black-box network, its local structure can be made accessible via automatic differentiation. This gives rise to local Taylor expansions of the form:

$$T_{\text{PINN}}(x; x_0) = \sum_{k=0}^{K} \frac{u^{(k)}(x_0)}{k!}(x - x_0)^k, \tag{1}$$

which serve as a physically informed, differentiable surrogate for the unknown ground-truth expression. Although the local Taylor expansions do not equate to the analytic solution, they provide a locally valid proxy that respects both the governing PDE and the global data distribution. From this view, the Taylor expansion becomes a conceptual and computational bridge—connecting the numeric expressivity of PINNs with the symbolic abstraction of SR.

In contrast to direct derivative estimation from discrete data, which is highly sensitive to noise and lacks structural context, Taylor coefficients extracted from PINNs are globally consistent and physically grounded. This makes them especially suitable as guiding signals during symbolic evolution, where preserving critical structural components is essential for interpretability and convergence.

Building on this insight, we introduce a masking-based attribution mechanism that evaluates the contribution of each expression subtree to both the structural and physical fidelity of the model. These sensitivity scores are then used to inform the mutation and crossover strategies within a genetic programming (GP) framework. Substructures that align closely with the PINN-derived Taylor pattern are preserved, while those with low structural or physical importance are prioritized for modification. In addition, a hybrid fitness function combines physical residual loss with Taylor-based structural loss to jointly optimize both physical consistency and symbolic alignment.

Our approach transforms symbolic regression from a purely heuristic or data-driven procedure into a physics-guided, structurally-informed discovery process. Experiments on benchmark PDEs demonstrate that our method improves convergence speed, reduces overfitting, and produces symbolic expressions that are more consistent with the governing equations. Moreover, our framework offers a flexible plug-in mechanism for incorporating PINN knowledge into symbolic search, opening the door to new forms of hybrid neuro-symbolic modeling.

In summary, this work makes the following contributions:

- Proposes a novel framework that leverages local Taylor expansions extracted from Physics-Informed Neural Networks (PINNs) as structured, derivative-based priors to guide symbolic expression discovery through both initialization and loss alignment.
- Introduces a dual-sensitivity attribution mechanism that quantifies the physical and structural importance of symbolic subtrees, enabling fine-grained evaluation and evolution control.
- Integrates sensitivity-aware mutation and crossover strategies into genetic programming, along with a composite objective function that jointly optimizes for physical consistency and structural fidelity.

This fusion of PINN-informed priors and symbolic regression paves the way for more interpretable, physically meaningful machine-discovered models in scientific computing and data-driven equation discovery.

## Related Work

Partial Differential Equations (PDEs) are fundamental tools for modeling continuous processes in natural and engineering systems. Existing methods for solving PDEs can be broadly categorized into two families: numerical discretization approaches, such as Finite Difference Methods (FDM)[31, 36, 35] and Finite Element Methods (FEM)[6, 32, 28], which aim to approximate numerical solutions with high precision; and analytical modeling approaches, particularly Symbolic Regression (SR), which seeks to discover explicit mathematical expressions that reveal the underlying physical laws, offering high interpretability and structural transferability[34, 25, 27]

Traditional symbolic regression methods are mostly based on Genetic Programming (GP), which evolves expression trees to fit data[26]. Representative works such as Eureqa [7], GPTree [1] have demonstrated strong expressiveness in physical modeling and equation discovery tasks. However, these methods typically lack structural awareness and do not incorporate structural attribution mechanisms or modeling priors, often resulting in overly complex, physically implausible, or poorly convergent expressions. In recent years, approaches like RAG-SR [44] NetGP [2], HD-TLGP [3], SP-GPSR [4] and PhySO [33] have attempted to combine gradient-based optimization with neural guidance to enhance search efficiency. Nonetheless, they largely remain within a "structure-blind" search paradigm and fail to incorporate physical constraints effectively.

In parallel, neural networks have made significant progress in the numerical solution of PDEs. Physics-

Informed Neural Networks (PINNs)[30], in particular, embed PDE residuals into the loss function, enabling the network to learn continuous solutions under unsupervised or weakly supervised settings. A series of follow-up works, including DeepRitz[41], Galerkin PINNs[37], and XPINN[18], have improved the modeling of solution accuracy, boundary conditions, and generalization. Despite their success, these methods remain fundamentally black-box approximators and cannot generate explicit, interpretable symbolic expressions or reusable structural information.

In recent years, several studies have attempted to bridge PDE solvers (especially PINNs) with symbolic modeling, aiming toward more interpretable and transparent analytical modeling pipelines[5]. A representative direction is to use PINNs to generate training data and then perform symbolic regression using differentiable program architectures, as in [24]. Another approach embeds symbolic branches (e.g., $\sin$, $\exp$) directly into the network architecture, trained under physics constraints—e.g., Physics-Informed Symbolic Networks (PISN) [23]. Alternatively, some methods fit hidden differential components inside PINNs and then apply symbolic regression (e.g., AI-Feynman) to extract interpretable formulas [34]. These methods demonstrate the potential of combining symbolic expressiveness with PDE solution accuracy, yet most remain focused on final expression output or symbolic embedding, without structural attribution or controllable symbolic evolution.

In the field of symbolic regression, TaylorGP [17] is a representative method that incorporates structural guidance mechanisms. It performs Taylor expansions at multiple sampling points of the target function to extract local structural attributes such as variable separability, parity, and monotonicity. These structural cues are then used to guide the initialization of symbolic expressions and constrain evolutionary operations such as mutation and crossover. The method shows preliminary success in improving structural control, demonstrating the positive role of structural information in symbolic modeling. However, TaylorGP lacks a dynamic attribution mechanism and global consistency modeling capability, and it is not suitable for PDE settings where only equation-based constraints are available.

To address these limitations, we propose StruSR—a structurally guided symbolic modeling framework that, for the first time, leverages PDE numerical solvers (PINNs) as structure information extractors rather than direct solvers. Trained over global data and physical constraints, PINNs produce smooth solutions that exhibit global consistency. The Taylor expansion derived from such solutions, though local in form, reflects structural characteristics under global constraints, providing a physically consistent and globally coherent structural prior. Furthermore, StruSR introduces a structure-sensitive attribution mechanism based on masked gradients to identify key substructures and guide symbolic mutation and crossover at fine granularity. This enables a transition from expression-level to structure-level guidance. StruSR effectively integrates the advantages of numerical and symbolic modeling, demonstrating strong performance in PDE-driven symbolic expression recovery tasks.

**Algorithm 1:** Structure-Guided Symbolic Regression via PINN Taylor Expansion

**Input:**
Trained PINN model $u(x)$, Collocation points $\{x_i\}_{i=1}^N$, Differential operator $\mathcal{N}[\cdot]$, Taylor order $K = 5$, Population size $P$, mutation rate $p_{\text{mut}}$, crossover rate $p_{\text{cross}}$, Structural weight $\lambda$, sensitivity balance $\beta$.
**Output:**
Best symbolic expression $f^*(x)$

Initialize population $\mathcal{G} = \{f_1, f_2, \ldots, f_P\}$ with random symbolic expressions;
Compute PINN Taylor expansions $\{T_{(u)}(x; x_i)\}_{i=1}^N$ at collocation points;
**while** *termination criteria not met* **do**
  **foreach** *expression* $f \in \mathcal{G}$ **do**
    Compute structural loss $\mathcal{L}_{\text{Taylor}}(f; x_i)$;
    Compute physics loss $\mathcal{L}_{\text{phys}}(f) = \frac{1}{N}\sum_{i=1}^N (\mathcal{N}[f](x_i))^2$;
    Evaluate fitness: $\mathcal{F}(f) = \mathcal{L}_{\text{phys}}(f) + \lambda \cdot \mathcal{L}_{\text{Taylor}}(f)$;
    **foreach** *subtree* $s_j$ *of* $f$ **do**
      Construct masked expression $f_{-s_j}(x)$ by replacing $s_j$ with constant 1;
      Compute sensitivity scores:;
        $\Delta_j^{\text{struct}} = \mathcal{L}_{\text{Taylor}}(f_{-s_j}) - \mathcal{L}_{\text{Taylor}}(f)$;
        $\Delta_j^{\text{res}} = \text{MSE}(\mathcal{N}[f_{-s_j}]) - \text{MSE}(\mathcal{N}[f])$;
        $\Delta_j^{\text{total}} = \beta \cdot \Delta_j^{\text{res}} + (1-\beta) \cdot \Delta_j^{\text{struct}}$;
  Select parents from $\mathcal{G}$ based on fitness;
  **foreach** *parent pair* $(f_A, f_B)$ **do**
    **if** *rand()* $< p_{cross}$ **then**
      Sample $s_A$ from $f_A$ using a sensitivity-based probability distribution;
      Sample $s_B$ from $f_B$;
      Swap $s_A$ and $s_B$ to create offspring;
  **foreach** *offspring* $f$ **do**
    **if** *rand()* $< p_{mut}$ **then**
      Sample $s_j$ using a sensitivity-based probability distribution;
      Replace $s_j$ with a random subtree from the symbol library;
  Update population $\mathcal{G}$;
**return** *expression* $f^*(x) = \arg\min_f \mathcal{F}(f)$

**Algorithm Description.** This algorithm aims to integrate the local Taylor expansion structure derived from a Physics-Informed Neural Network (PINN) into the genetic programming (GP) process for symbolic regression, thereby enhancing the physical consistency and structural interpretability of the generated expressions. The core idea is to introduce a dual sensitivity evaluation mechanism—based on both structural alignment and residual fidelity—during the evolution process, enabling fine-grained guidance over substructures in candidate expressions.

As illustrated in Fig. 1, a pre-trained PINN model is used to extract $K$-order Taylor expansions (default $K = 5$) at multiple anchor points $\{x_0^{(i)}\}$ sampled from the domain. These expansions encode the local derivative behavior of the target differential equation around each point, collectively providing a structurally consistent and physically grounded reference for guiding symbolic expression discovery.

The evolution begins with a population of randomly initialized symbolic expressions. For each candidate expres-

sion, two key losses are computed: (1) the *physics residual loss* $\mathcal{L}_{\text{phys}}$, which quantifies the mean squared residual when the expression is substituted into the target differential operator $\mathcal{N}[\cdot]$ across all collocation points; and (2) the *structure loss* $\mathcal{L}_{\text{Taylor}}$, measuring the squared difference between the Taylor coefficients of the expression and the PINN expansion. These two metrics are jointly incorporated into a hybrid fitness function $\mathcal{F}(f)$ that reflects both physical correctness and structural alignment.

To assess the local significance of each subcomponent, we introduce a masking attribution mechanism. For each subtree $s_j$ in a symbolic expression $f(x)$, we temporarily mask it by replacing it with a neutral constant (e.g., 1), forming a masked expression $f_{-s_j}(x)$. The changes in structure loss and physics residual due to this masking are recorded as *structural sensitivity* $\Delta_j^{\text{struct}}$ and *residual sensitivity* $\Delta_j^{\text{res}}$, respectively. A total sensitivity score $\Delta_j^{\text{total}}$ is computed via weighted summation controlled by hyperparameter $\beta$, reflecting the joint importance of subtree $s_j$ with respect to both structural and physical objectives.

During genetic operations, sensitivity scores define a probability distribution that guides both crossover and mutation, balancing structural preservation with search diversity. For crossover, each parent pair $(f_A, f_B)$ samples a subtree $s_A$ from $f_A$ using this sensitivity-based distribution, giving higher selection probability to less critical regions, and swaps it with a subtree $s_B$ from $f_B$. This reduces the risk of disrupting key structures while maintaining exploration. For mutation, a subtree $s_j$ is sampled from the same sensitivity-based distribution and replaced with a randomly generated subtree from the symbol library, allowing mutations to occur mainly in non-critical parts while preserving structurally important components.

The evolutionary process repeats until convergence or other termination criteria are met. At each generation, expressions are evaluated by a hybrid fitness function that balances physics residual fidelity and structural alignment. The best-performing expression $f^*(x)$ is finally selected as the solution.

## Methodology

We propose a structure-aware symbolic regression framework that integrates structural priors extracted from Physics-Informed Neural Networks (PINNs) into the evolutionary process of Genetic Programming (GP). At the heart of this approach lies the use of local Taylor expansions derived from PINNs as a source of ground-truth structural guidance. This mechanism promotes the discovery of symbolic expressions that are not only accurate with respect to data but also structurally aligned with underlying physical principles.

Let $u(x)$ denote a trained PINN model that approximates the solution to a target equation. Around a selected reference point $x_0$, we extract a $K$-order Taylor expansion:

$$T_{(u)}(x; x_0) = \sum_{k=0}^{K} \frac{u^{(k)}(x_0)}{k!}(x - x_0)^k \tag{2}$$

Here, $u^{(k)}(x_0)$ denotes the $k$-th derivative of $u(x)$ evaluated at $x_0$, and $k!$ is the factorial normalization. Likewise, a candidate symbolic expression $f(x)$ can be expanded in the same manner. The structure loss is then defined as the mean squared error between the Taylor coefficients of $f(x)$ and $u(x)$:

$$\mathcal{L}_{\text{Taylor}}(f; x_0) = \sum_{k=0}^{K} \left( \frac{f^{(k)}(x_0)}{k!} - \frac{u^{(k)}(x_0)}{k!} \right)^2 \tag{3}$$

This loss function provides a robust and interpretable measure of structural alignment. Since Taylor expansions are naturally ordered by monomial degree $(x - x_0)^k$, structural information is encoded directly in the coefficient vector. When a derivative term is theoretically absent in $f(x)$, its corresponding coefficient should ideally be zero. However, due to numerical estimation errors, such coefficients are often small but non-zero. We set the Taylor expansion order $K = 5$ to balance expressiveness and numerical stability: lower orders may underfit local structure, while higher orders are prone to derivative noise and overfitting.

To identify which subtrees in a symbolic expression contribute most to the structure and data fidelity, we introduce a *masking attribution* mechanism. For each subtree $s_j$ in a candidate expression $f(x)$, we construct a masked version $f_{-s_j}(x)$ by replacing $s_j$ with a neutral constant (e.g., 1). We then evaluate two forms of sensitivity:

**Structural sensitivity** quantifies the change in structure loss after masking:

$$\Delta_j^{\text{struct}} = \mathcal{L}_{\text{Taylor}}(f_{-s_j}; x_i) - \mathcal{L}_{\text{Taylor}}(f; x_i). \tag{4}$$

We define the Taylor structure loss $\mathcal{L}_{\text{Taylor}}$ at point $x_i$ as:

$$\mathcal{L}_{\text{Taylor}}(f; x_i) = \sum_{k=0}^{K} \left( \frac{f^{(k)}(x_i)}{k!} - \frac{u^{(k)}(x_i)}{k!} \right)^2, \tag{5}$$

where $u(\cdot)$ is obtained from the trained PINN, and $f(\cdot)$ is obtained from the candidate symbolic expression.

**Residual sensitivity** quantifies the change in physical residual error, defined by the violation of the governing differential equation when substituting the symbolic expression into the physics operator $\mathcal{N}[\cdot]$:

$$\Delta_j^{\text{res}} = \text{MSE}(\mathcal{N}[f_{-s_j}](x)) - \text{MSE}(\mathcal{N}[f](x)) \tag{6}$$

Here, $\mathcal{N}[\cdot]$ represents the differential operator associated with the target PDE, and the residual is evaluated point-wise over a collocation set $x$. The mean squared error measures how well the expression satisfies the physical constraint. We integrate both sensitivities into a single scalar score:

$$\Delta_j^{\text{total}} = \beta \cdot \Delta_j^{\text{res}} + (1 - \beta) \cdot \Delta_j^{\text{struct}} \tag{7}$$

The balancing parameter $\beta \in [0, 1]$ controls the trade-off between residual accuracy and structural fidelity. Subtrees with low $\Delta_j^{\text{total}}$ are deemed structurally and functionally unimportant, and thus are prioritized for modification during evolution.

In crossover, we adopt a sensitivity-guided strategy that prioritizes protecting beneficial symbolic patterns while

Table 1: Summary of tested differential equations with their deterministic conditions.

| Name | PDE | Deterministic Conditions |
|---|---|---|
| Advection | $\frac{\partial \phi}{\partial t} + u\frac{\partial \phi}{\partial x} = 0$ | $u(x,0) = \sin(x), \quad x \in [0,1],\ t \in [0,1]$ |
| Diffusion | $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - e^{-t}\sin(\pi x)(1-\pi^2)$ | $u(x,0) = \sin(\pi x), \quad x \in [-1,1]$<br>$u(-1,t) = u(1,t) = 0, \quad t \in [0,1]$ |
| Poisson2D | $\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = 30x_1^2 - 7.8x_1 + 1$ | $u(\mathbf{x}) = 2.5x_1^4 - 1.3x_1^3 + 0.5x_2^2 - 1.7x_2, \quad \mathbf{x} \in \partial[-1,1]^2$ |
| Poisson3D | $\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \frac{\partial^2 u}{\partial x_3^2} = 30x_1^2 - 7.8x_2 + 1$ | $u(\mathbf{x}) = 2.5x_1^4 - 1.3x_2^3 + 0.5x_3^2, \quad \mathbf{x} \in \partial[-1,1]^3$ |
| Heat2D | $\frac{\partial u}{\partial t} - \left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}\right) = -30x_1^2 + 7.8x_2 + t$ | $u(\mathbf{x},t) = 2.5x_1^4 - 1.3x_2^3 + 0.5t^2, \quad \mathbf{x} \in \partial[-1,1]^2,\ t \in [0,1]$<br>$u(\mathbf{x},0) = 2.5x_1^4 - 1.3x_2^3, \quad \mathbf{x} \in [-1,1]^2$ |
| Heat3D | $\frac{\partial u}{\partial t} - \left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \frac{\partial^2 u}{\partial x_3^2}\right) = -30x_1^2 + 7.8x_2 - 2.7$ | $u(\mathbf{x},t) = 2.5x_1^4 - 1.3x_2^3 + 0.5x_3^2 - 1.7t, \quad \mathbf{x} \in \partial[-1,1]^3,\ t \in [0,1]$<br>$u(\mathbf{x},0) = 2.5x_1^4 - 1.3x_2^3 + 0.5x_3^2, \quad \mathbf{x} \in [-1,1]^3$ |
| Wave2D | $\frac{\partial^2 u}{\partial t^2} - \left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}\right) = -u^3 + (0.25 - 4x_1^2)$ | $u(\mathbf{x},0) = \exp(x_1^2)\sin(x_2), \quad \mathbf{x} \in [-1,1]^2$<br>$u(\mathbf{x},t) = \exp(x_1^2)\sin(x_2)e^{-0.5t}, \quad \mathbf{x} \in \partial[-1,1]^2,\ t \in [0,1]$ |
| Wave3D | $\frac{\partial^2 u}{\partial t^2} - \left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \frac{\partial^2 u}{\partial x_3^2}\right) = u^2 - \left(4x_1^2 + 4x_3^2 + 2.75\right)$ | $u(\mathbf{x},0) = \exp(x_1^2 + x_3^2)\cos(x_2), \quad \mathbf{x} \in [-1,1]^3$<br>$u(\mathbf{x},t) = \exp(x_1^2 + x_3^2)\cos(x_2)e^{-0.5t}, \quad \mathbf{x} \in \partial[-1,1]^3,\ t \in [0,1]$ |

Table 2: Mean absolute error (MAE) comparison between StruSR and representative symbolic regression baselines on recovering closed-form solutions for a set of PDE benchmarks. Each entry reports the average test MAE and its standard deviation over 10 independent runs.

| Name | StruSR (Ours) | RAG-SR | NetGP | HD-TLGP | PhySO |
|---|---|---|---|---|---|
| Advection | 1.21E-13 ± 3.57E-15 | 2.03E-13 ± 4.22E-14 | **9.59E-14 ± 1.27E-14** | 7.91E-11 ± 6.89E-12 | 3.18E-10 ± 1.10E-10 |
| Diffusion | **6.38E-6 ± 1.33E-6** | 8.91E-6 ± 1.18E-6 | 1.79E-5 ± 2.12E-5 | 2.23E-5 ± 3.97E-5 | 3.41E-3 ± 7.59E-4 |
| Poisson2D | **7.32E-5 ± 2.71E-6** | 8.03E-5 ± 3.21E-6 | 5.69E-1 ± 7.94E-2 | 1.87E-2 ± 2.40E-3 | 7.13E+0 ± 1.43E+0 |
| Poisson3D | **5.62E-3 ± 1.41E-3** | 5.77E-3 ± 1.52E-3 | 2.67E+0 ± 1.08E-1 | 1.85E-1 ± 1.69E-2 | 6.11E+0 ± 1.48E+0 |
| Heat2D | **1.53E-6 ± 7.81E-7** | 1.61E-6 ± 8.12E-7 | 3.16E+0 ± 1.32E-1 | 2.21E-2 ± 1.34E-2 | 1.18E+1 ± 1.21E+0 |
| Heat3D | 1.09E-4 ± 1.29E-5 | **9.27E-5 ± 1.42E-5** | 2.74E+0 ± 3.32E-1 | 8.71E-2 ± 2.91E-3 | 1.13E+1 ± 1.69E+0 |
| Wave2D | **4.33E-5 ± 2.41E-6** | 4.79E-5 ± 2.89E-6 | 9.47E+0 ± 2.19E-1 | 4.32E-4 ± 3.57E-4 | 2.01E+0 ± 1.39E-1 |
| Wave3D | **7.09E-6 ± 1.64E-6** | 8.92E-6 ± 1.52E-6 | 1.36E+1 ± 4.03E+0 | 9.71E-4 ± 2.35E-5 | 1.34E+1 ± 2.29E+0 |

maintaining diversity in the search. Instead of deterministically selecting the lowest-sensitivity subtrees for swapping, we construct a softmax distribution over all subtrees based on $\Delta_j^{\text{total}}$. This biases the selection toward less sensitive regions of each parent, yet assigns a small probability to higher-sensitivity subtrees, preventing the search from repeatedly modifying the same locations and improving overall exploration. This approach preserves structural alignments inherited from PINN guidance while reducing the risk of premature convergence, ultimately enhancing the interpretability of the evolved expressions.

Similarly, for mutation, a candidate expression is mutated with probability $p_{\text{mut}}$. When triggered, a subtree $s_j$ is sampled from the same softmax distribution, favoring low-sensitivity components but allowing occasional changes in other regions to ensure population diversity. The selected subtree is then replaced with a randomly generated one from the symbol library, introducing variability while largely preserving structurally critical components.

Finally, to balance physical fidelity and symbolic expressiveness, we define a hybrid fitness function that jointly penalizes physics violations and structural mismatch:

$$\mathcal{F}(f) = \mathcal{L}_{\text{phys}}(f) + \mathcal{L}_{\text{Taylor}}(f) \tag{8}$$

Here, $\mathcal{L}_{\text{phys}}(f)$ denotes the physics residual loss, computed as the mean squared deviation of $f$ when substituted into the governing differential equation operator $\mathcal{N}[\cdot]$, i.e.,

$$\mathcal{L}_{\text{phys}}(f) = \frac{1}{N}\sum_{i=1}^{N}\left(\mathcal{N}[f](x_i)\right)^2 \tag{9}$$

where $\{x_i\}_{i=1}^N$ are collocation points and $\mathcal{N}[f]$ represents the target PDE/ODE form. The second term $\mathcal{L}_{\text{Taylor}}(f)$ quantifies the structural mismatch between the Taylor expansion of the candidate expression and that of the reference PINN at selected anchor points. Candidate expressions are selected based on this composite objective, facilitating symbolic discovery that is both accurate and grounded in physical priors.

In summary, our framework transforms the symbolic regression pipeline by incorporating Taylor-expansion-based structural signals into GP evolution. It promotes the generation of symbolic models that exhibit both empirical accuracy and physics-guided interpretability.

# Experiments

## Benchmark Setup

To comprehensively evaluate the accuracy and interpretability of the proposed method, we construct a PDE benchmark suite consisting of eight representative partial differential

Table 3: Comparison of ground-truth analytical solutions and symbolic solutions identified by StruSR on classic PDEs.

| Name | Ground Truth $u_{\text{true}}$ | Symbolic Solution $\hat{u}$ by StruSR |
|---|---|---|
| **Advection** | $\sin(x_1 - t)$ | $\sin(x_1 - 1.001t) + 0.001\sin(2x_1)$ |
| **Diffusion** | $e^{-t}\sin(\pi x)$ | $(1.022 - 0.482t + 0.124t^2)\sin(3.145x - 0.005t)$ |
| **Poisson2D** | $x_1^2(2.5x_1^2 - 1.3x_1) + 0.5x_2(x_2 - 1.4) - x_2$ | $x_1^2(2.500x_1^2 - 1.301x_1) + 0.504x_2(x_2 - 1.396) - x_2$ |
| **Poisson3D** | $2.5x_1^4 - 1.3x_2^3 + 0.5x_3^2$ | $2.500x_1^4 - 1.300x_2^3 + 0.501x_3^2 - 2.158 \times 10^{-5}$ |
| **Wave2D** | $\exp(x_1^2)\sin(x_2)\,e^{-0.5t}$ | $(1.000 - 0.500t + 0.120t^2)(1.000 + x_1^2)\sin(x_2)$ |
| **Wave3D** | $\exp(x_1^2 + x_3^2)\sin(x_2)e^{-0.5t}$ | $\exp(-0.5001t + 1.000x_1^2 + 0.999x_3^2)\sin(x_2)$ |
| **Heat2D** | $0.5t^2 + 2.5x_1^4 - 1.3x_2^3$ | $0.500^2 + 2.501x_1^4 - 1.299x_2(x_2 + 0.020)x_2$ |
| **Heat3D** | $-1.7t + 2.5x_1^4 - 1.3x_2^3 + 0.5x_3^2$ | $-1.701t + 2.500x_1^4 - 1.300x_2^3 + 0.499x_3^2$ |

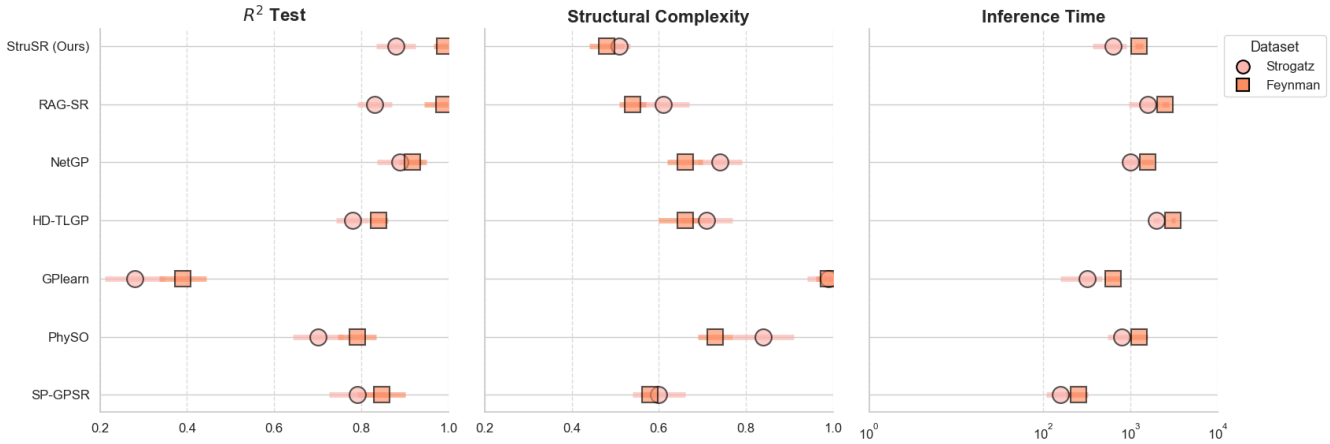



Figure 2: Performance comparison of symbolic regression methods across three evaluation dimensions on two benchmark suites (Strogatz and Feynman). Each row corresponds to a specific baseline algorithm, while the circular and square markers represent results on the Strogatz and Feynman datasets, respectively. Subfigure (a) reports the test $R^2$ score (higher is better), indicating predictive accuracy. Subfigure (b) shows the normalized structural complexity (lower is better), reflecting the compactness of the learned expressions. Subfigure (c) presents the inference time (lower is better), measuring computational efficiency.

equations, covering one, two, and three-dimensional cases. These tasks are designed to reflect various physical structures and boundary conditions. For each equation, the goal is to recover the underlying analytical solution based on given boundary constraints and sparse observations. Tab. 1 summarizes the governing equations along with their deterministic conditions, including initial and boundary constraints.

## Quantitative Performance Comparison

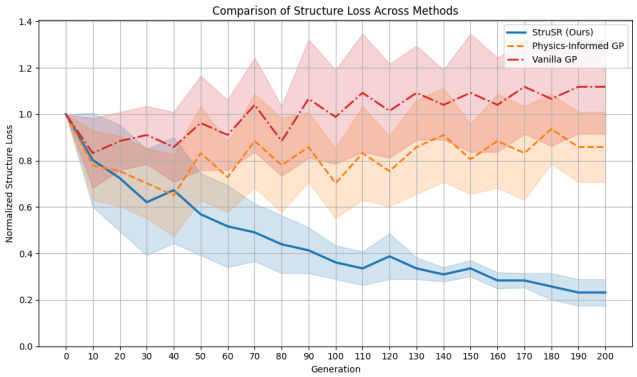
We first compare StruSR against representative symbolic regression baselines on a suite of PDE recovery tasks. Competing methods include RAG-SR, NetGP, HD-TLGP, and PhySO. Evaluation is based on the Mean Absolute Error (MAE) computed on the test set. Each result reports the average and standard deviation over 10 independent runs to ensure statistical robustness.

As shown in Tab. 2, StruSR consistently achieves the lowest or second-lowest MAE across most PDE benchmarks. This advantage is particularly evident in structurally complex scenarios such as Poisson2D/3D, Wave2D, and Heat2D. We attribute this superior performance to StruSR's structure-aware modeling framework, which effectively integrates Taylor priors and guided evolution to align expression generation with the underlying physical structure.

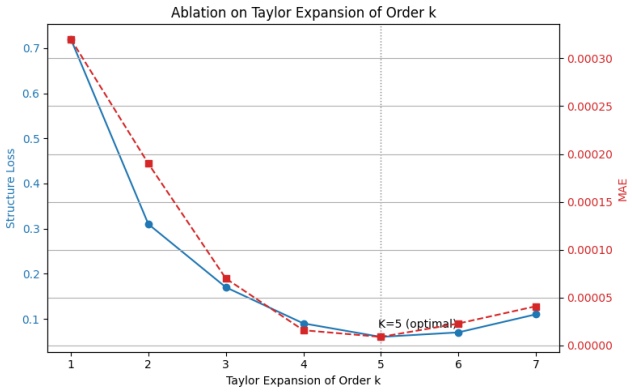
Moreover, StruSR maintains competitive accuracy in 3D problems, demonstrating strong generalization to high-dimensional tasks. While RAG-SR shows promising results on specific cases like Heat3D, it falls short in overall consistency. Other baselines, such as NetGP and PhySO, exhibit higher variance across tasks, indicating limited robustness in structural expression discovery.

## Analytical Structure Matching

To assess structural alignment between predicted and ground-truth solutions, we further compare symbolic expressions identified by StruSR with the original closed-form equations. Tab.3 shows that the discovered formulas exhibit strong structural fidelity, often matching the ground truth up to minor coefficient variations or algebraic transformations. These results validate StruSR's ability to not only approxi-




(a) Normalized structure loss across generations.




(b) Effect of Taylor expansion order $K$ on structure loss and MAE.

Figure 3: Ablation studies on structural supervision: (a) convergence of normalized structure loss across different methods; (b) impact of Taylor expansion order $K$ on performance.

mate numerical behavior but also recover semantically faithful symbolic forms.

## General Symbolic Modeling Performance

To assess the general modeling capability of StruSR beyond PDE recovery, we further evaluate it on two widely-used symbolic regression benchmarks: the Strogatz system and the Feynman equation dataset [34]. Following the standard symbolic regression evaluation protocol [44], each dataset is randomly split into 75% training and 25% testing subsets. All methods are evaluated based on 10 independent runs. We report three metrics: test $R^2$ score (higher is better), normalized structural complexity (lower is better), and average inference time (lower is better).

Fig.2(a) shows that StruSR achieves competitive or leading prediction accuracy on both datasets, demonstrating strong symbolic function approximation capabilities. This performance stems from its structure-aware expression evolution strategy, which allows the model to efficiently explore semantically meaningful subspaces. In terms of symbolic expression complexity, Fig.2(b) evaluates the number of nodes in the output expression trees. All complexity values are normalized to the $[0, 1]$ interval for each task, where 1 corresponds to the most complex expression in the set. StruSR consistently produces more compact and interpretable formulas, reflecting the effectiveness of its structural guidance mechanisms in constraining unnecessary redundancy during evolution. Fig.2(c) presents the average inference time per expression on a logarithmic scale. StruSR exhibits relatively low inference latency across most tasks, highlighting its potential for real-time and resource-efficient deployment scenarios. These results collectively demonstrate that StruSR not only achieves accurate symbolic modeling, but also provides structurally concise and computationally efficient solutions across diverse application domains.

## Structural Guidance Analysis

To evaluate the effectiveness of structural guidance in symbolic expression evolution, we conduct two complementary studies: structure loss dynamics across methods and an ablation on Taylor expansion order for prior extraction.

Fig.3(a) shows normalized structure loss over generations for three methods: Vanilla GP (red), Physics-Informed GP (orange), and our StruSR (blue). StruSR achieves stable and monotonic structure loss reduction, outperforming both baselines. Vanilla GP shows minimal structural improvement, lacking inductive bias. Physics-Informed GP improves early on but fluctuates later due to the absence of explicit structural preservation. Fig.3(b) evaluates the impact of varying Taylor expansion order $K$. Both structure loss and prediction MAE are minimized at $K = 5$, indicating an optimal trade-off between local fidelity and robustness. Lower orders ($K \leq 2$) lack sufficient derivative information, while higher orders ($K > 5$) risk overfitting and noise. These results validate our dual-level guidance: Taylor-based initialization from a PINN captures globally consistent physical priors, while masking-based attribution enables fine-grained, subtree-level mutation control. Together, they promote stable, interpretable, and physically grounded symbolic modeling.

## Conclusion

We propose StruSR, a structure-aware symbolic regression framework that leverages physics-informed priors derived from Taylor expansions of trained PINNs. By introducing a dual-level structural guidance mechanism—comprising Taylor-based initialization and sensitivity-based attribution—StruSR steers symbolic evolution toward globally consistent, physically meaningful solutions. Extensive experiments on PDE recovery and standard symbolic benchmarks demonstrate that StruSR consistently delivers high predictive accuracy, reduced structural complexity. It reliably discovers semantically faithful expressions with stable and interpretable structural evolution, highlighting the effec-

tiveness of embedding physical inductive biases into symbolic modeling. This work lays a principled foundation for physics-guided symbolic discovery and opens new directions for interpretable, efficient, and generalizable equation modeling.

## References


[1] Achituve, I.; Navon, A.; Yemini, Y.; Chechik, G.; and Fetaya, E. 2021. GP-Tree: A Gaussian process classifier for few-shot incremental learning. In *International conference on machine learning*, 54–65. PMLR.

[2] Cao, L.; Feng, Y.; Jiang, M.; and Tan, K. C. 2025. NetGP: A Hybrid Framework Combining Genetic Programming and Deep Reinforcement Learning for PDE Solutions. In *2025 IEEE Congress on Evolutionary Computation (CEC)*, 1–8. IEEE.

[3] Cao, L.; Liu, Y.; Wang, Z.; Xu, D.; Ye, K.; Tan, K. C.; and Jiang, M. 2024. An interpretable approach to the solutions of high-dimensional partial differential equations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 20640–20648.

[4] Cao, L.; Zheng, Z.; Ding, C.; Cai, J.; and Jiang, M. 2023. Genetic programming symbolic regression with simplification-pruning operator for solving differential equations. In *International Conference on Neural Information Processing*, 287–298. Springer.

[5] Das, J.; Bhaumik, B.; De, S.; and Changdar, S. 2025. Physics-informed neural network with symbolic regression for deriving analytical approximate solutions to nonlinear partial differential equations. *Neural Computing and Applications*, 1–36.

[6] David Müzel, S.; Bonhin, E. P.; Guimarães, N. M.; and Guidi, E. S. 2020. Application of the finite element method in the analysis of composite materials: A review. *Polymers*, 12(4): 818.

[7] Dubčáková, R. 2011. Eureqa: software review.

[8] Gong, Y.; Hou, Y.; Shi, J.; DIEP, K. L.; and Jiang, M. 2026. A Theory-Guided Framework for Few-Shot Cross-Modal Sketch Person Re-Identification.

[9] Gong, Y.; Hou, Y.; Wang, Z.; Lin, Z.; and Jiang, M. 2024. Adversarial learning for neural PDE solvers with sparse data. *arXiv preprint arXiv:2409.02431*.

[10] Gong, Y.; Hou, Y.; Zhang, C.; and Jiang, M. 2023. Beyond augmentation: Empowering model robustness under extreme capture environments. In *IJCNN,2023*. IEEE.

[11] Gong, Y.; Huang, L.; and Chen, L. 2021. Eliminate deviation with deviation for data augmentation and a general multi-modal data learning method. *arXiv preprint arXiv:2101.08533*.

[12] Gong, Y.; Huang, L.; and Chen, L. 2022. Person re-identification method based on color attack and joint defence. In *CVPR, 2022*, 4313–4322.

[13] Gong, Y.; Li, J.; Chen, L.; and Jiang, M. 2022. Exploring color invariance through image-level ensemble learning. *arXiv preprint arXiv:2401.10512*.

[14] Gong, Y.; Zeng, Q.; Xu, D.; Wang, Z.; and Jiang, M. 2024. Cross-modality attack boosted by gradient-evolutionary multiform optimization. *arXiv preprint arXiv:2409.17977*.

[15] Gong, Y.; Zhang, C.; Hou, Y.; Chen, L.; and Jiang, M. 2024. Beyond dropout: Robust convolutional neural networks based on local feature masking. In *IJCNN,2024*. IEEE.

[16] Gong, Y.; Zhong, Z.; Qu, Y.; Luo, Z.; Ji, R.; and Jiang, M. 2024. Cross-modality perturbation synergy attack for person re-identification. *Advances in Neural Information Processing Systems*, 37: 23352–23377.

[17] He, B.; Lu, Q.; Yang, Q.; Luo, J.; and Wang, Z. 2022. Taylor genetic programming for symbolic regression. In *Proceedings of the genetic and evolutionary computation conference*, 946–954.

[18] Jagtap, A. D.; and Karniadakis, G. E. 2020. Extended physics-informed neural networks (XPINNs): A generalized space-time domain decomposition based deep learning framework for nonlinear partial differential equations. *Communications in Computational Physics*, 28(5).

[19] Jin, M.; Zhang, Y.; Chen, W.; Zhang, K.; Liang, Y.; Yang, B.; Wang, J.; Pan, S.; and Wen, Q. 2024. Position: What can large language models tell us about time series analysis. In *41st International Conference on Machine Learning*. MLResearchPress.

[20] Lin, J.; Wang, Z.; Hou, Y.; Tang, Y.; and Jiang, M. 2024. Phy124: Fast physics-driven 4d content generation from a single image. *arXiv preprint arXiv:2409.07179*.

[21] Lin, J.; Zhenzhong, W.; Dejun, X.; Shu, J.; Gong, Y.; and Jiang, M. 2024. Phys4DGen: A Physics-Driven Framework for Controllable and Efficient 4D Content Generation from a Single Image. *arXiv preprint arXiv:2411.16800*.

[22] Lu, W.; Wang, J.; Sun, X.; Chen, Y.; and Xie, X. ???? Out-of-distribution Representation Learning for Time Series Classification. In *The Eleventh International Conference on Learning Representations*.

[23] Majumdar, R.; Jadhav, V.; Deodhar, A.; Karande, S.; Vig, L.; and Runkana, V. 2022. Physics informed symbolic networks. *arXiv preprint arXiv:2207.06240*.

[24] Majumdar, R.; Jadhav, V.; Deodhar, A.; Karande, S.; Vig, L.; and Runkana, V. 2023. Symbolic regression for pdes using pruned differentiable programs. *arXiv preprint arXiv:2303.07009*.

[25] Mundhenk, T.; Landajuela, M.; Glatt, R.; Santiago, C. P.; Petersen, B. K.; et al. 2021. Symbolic regression via deep reinforcement learning enhanced genetic programming seeding. *Advances in Neural Information Processing Systems*, 34: 24912–24923.

[26] Oh, H.; Amici, R.; Bomarito, G.; Zhe, S.; Kirby, R.; and Hochhalter, J. 2023. Genetic programming based symbolic regression for analytical solutions to differential equations. *arXiv preprint arXiv:2302.03175*.

[27] Petersen, B. K.; Landajuela, M.; Mundhenk, T. N.; Santiago, C. P.; Kim, S. K.; and Kim, J. T. 2019. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. *arXiv preprint arXiv:1912.04871*.

[28] Polycarpou, A. C. 2022. *Introduction to the finite element method in electromagnetics*. Springer Nature.

[29] Qin, Y.; Song, D.; Cheng, H.; Cheng, W.; Jiang, G.; and Cottrell, G. W. 2017. A dual-stage attention-based recurrent neural network for time series prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2627–2633.

[30] Raissi, M.; Perdikaris, P.; and Karniadakis, G. E. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378: 686–707.

[31] Robertsson, J. O.; and Blanch, J. O. 2011. Numerical methods, finite difference. In *Encyclopedia of solid earth geophysics*, 883–892. Springer.

[32] Szabó, B.; and Babuška, I. 2021. Finite element analysis: Method, verification and validation.

[33] Tenachi, W.; Ibata, R.; and Diakogiannis, F. I. 2023. Deep symbolic regression for physics guided by units constraints: toward the automated discovery of physical laws. *The Astrophysical Journal*, 959(2): 99.

[34] Udrescu, S.-M.; and Tegmark, M. 2020. AI Feynman: A physics-inspired method for symbolic regression. *Science advances*, 6(16): eaay2631.

[35] Ullah, H.; Hayat, T.; Ahmad, S.; and Alhodaly, M. S. 2021. Entropy generation and heat transfer analysis in power-law fluid flow: Finite difference method. *International Communications in Heat and Mass Transfer*, 122: 105111.

[36] Wahyudi, S.; Lestari, P.; and Gapsari, F. 2021. Application of Finite Difference Methods (FDM) on mathematical model of bioheat transfer of one-dimensional in human skin exposed environment condition. *J Mech Eng Res Develop*, 44(5): 1–9.

[37] Wang, S.; Teng, Y.; and Perdikaris, P. 2021. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5): A3055–A3081.

[38] Wen, Q.; Sun, L.; Yang, F.; Song, X.; Gao, J.; Wang, X.; and Xu, H. 2020. Time series data augmentation for deep learning: A survey. *arXiv preprint arXiv:2002.12478*.

[39] Xu, K.; Chen, L.; and Wang, S. 2024. Kan4drift: Are kan effective for identifying and tracking concept drift in time series? In *NeurIPS Workshop on Time Series in the Age of Large Models*.

[40] Xu, K.; Chen, L.; and Wang, S. 2024. Kolmogorov-arnold networks for time series: Bridging predictive power and interpretability. *arXiv preprint arXiv:2406.02496*.

[41] Yu, B.; et al. 2018. The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems. *Communications in Mathematics and Statistics*, 6(1): 1–12.

[42] Zeng, Q.; Gong, Y.; and Jiang, M. 2023. Cross-Task Attack: A Self-Supervision Generative Framework Based on Attention Shift. In *IJCNN, 2023*. IEEE.

[43] Zhang, C.; Song, D.; Chen, Y.; Feng, X.; Lumezanu, C.; Cheng, W.; Ni, J.; Zong, B.; Chen, H.; and Chawla, N. V. 2019. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 1409–1416.

[44] Zhang, H.; Chen, Q.; Banzhaf, W.; Zhang, M.; et al. 2025. RAG-SR: Retrieval-augmented generation for neural symbolic regression. In *The Thirteenth International Conference on Learning Representations*.

[45] Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; and Jin, R. 2022. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, 27268–27286. PMLR.