Data-driven progressive discovery of physical laws

Mingkun Xia and Weiwei Zhang*

School of Aeronautics, Northwestern Polytechnical University, Xi'an 710072, China;
International Joint Institute of Intelligent Fluid Mechanics, Northwestern Polytechnical University, Xi'an 710072, China;
and National Key Laboratory of Aircraft Configuration Design, Xi'an 710072, China

(Received 18 March 2026; accepted 6 October 2026; published 2 November 2026)

Symbolic regression is a powerful tool for knowledge discovery, enabling the extraction of interpretable mathematical expressions directly from data. However, conventional symbolic discovery typically follows an end-to-end, "one-step" process, which often generates lengthy and physically meaningless expressions when dealing with real physical systems, leading to poor model generalization. This limitation fundamentally stems from its deviation from the basic path of scientific discovery: Physical laws do not exist in a single form but follow a hierarchical and progressive pattern from simplicity to complexity. Motivated by this principle, we propose Chain of Symbolic Regression (CoSR), a framework that models the discovery of physical laws as a chain of symbolic knowledge. This knowledge chain is formed by progressively combining multiple knowledge units with clear physical meanings along a specific logic, ultimately enabling the precise discovery of the underlying physical laws from data. CoSR fully recapitulates the progressive discovery path from Kepler's third law to the law of universal gravitation in classical mechanics and is applied to three types of problems: turbulent Rayleigh-Bénard convection, viscous flows in a circular pipe, and laser-metal interaction, demonstrating its ability to improve classical scaling theories. Finally, CoSR showcases its capability to discover new knowledge in the complex engineering problem of aerodynamic coefficient scaling for different aircraft.

DOI: 10.1103/f2kp-pp4q

I. INTRODUCTION

The essence of physical discovery lies in distilling universal laws from natural phenomena and then articulating them through concise mathematical formulations, thereby constructing a framework for understanding the physical world [1,2]. This endeavor is seldom a one-off event; rather, it is a cumulative, progressive process that begins with first principles, advances via the tireless work of successive generations, and gradually builds toward more systematic and precise theoretical edifices. For example, the rigorous geometric system derived axiomatically from Euclidean geometry; Newton's law of universal gravitation synthesized from Kepler's laws of planetary motion; and the extension of Euler's equations for ideal fluids to the Navier-Stokes equations governing real fluids. This trajectory—from simplicity to complexity—epitomizes the intellectual triumph of scientific discovery.

While this progressive paradigm has achieved remarkable success throughout scientific history, it has relied more heavily on human intuition and deductive reasoning. As scientific problems grow in both dimensionality and nonlinear coupling, their complexity now approaches the boundary of what human cognition can directly process. Consequently, approaches based on intuition and deduction are confronting a bottleneck. To address this challenge, advances in data-driven paradigms, particularly machine learning, now offer powerful tools for uncovering hidden patterns in high-dimensional data [3,4].

Symbolic regression, as a class of data-driven methods characterized by its "white-box" nature, offers a distinct pathway for scientific knowledge discovery [2,5–7]. By systematically searching the space of mathematical expressions, this approach can directly identify differential equations [8,9] or extract closed-form expressions from observational data [10–13], thereby addressing previously intractable problems in modeling complex systems and advancing the automation of scientific discovery. However, conventional symbolic regression typically employs an end-to-end, one-step discovery strategy, which faces limitations when applied to complex physical systems characterized by high dimensionality, multivariate interactions, and multiscale features [14]. Owing to the combinatorial explosion, symbolic regression struggles to discover physical laws involving intricate hierarchical couplings. This manifests primarily in two typical issues: On the one hand, pursuing high fidelity to data often generates excessively lengthy and uninterpretable expressions, which tend to overfit and thus exhibit poor extrapolation capability; on the other hand, pursuing parsimony and interpretability often comes at the cost of omitting certain physical information, resulting in models that lack accuracy and generalizability. These issues are difficult to reconcile within a one-step discovery process.

*Contact author: aeroelastic@nwpu.edu.cn

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

The fundamental limitation of this one-step methodology stems from its conceptualization of scientific discovery as a static, one-off process, and it thereby overlooks the hierarchical and progressive pattern in which scientific knowledge itself is constructed. The progressive discovery paradigm aligns more closely with the inherent logic of theoretical development, offering greater systematic coherence and higher value for engineering applications. For instance, when confronting the limitations of an existing model in terms of accuracy or applicability, a progressive strategy can be employed. This involves incrementally extending the original theory within different regimes, resulting in a new theory that subsumes the old while achieving enhanced generality and precision [15,16]. Alternatively, one can introduce asymptotic scaling corrections to the base scaling laws derived from classical methods like dimensional analysis, systematically collapsing multiscale data from complex systems onto a single, unified curve [17,18]. In machine learning, hierarchical architectures, exemplified by convolutional neural networks, achieve progressive learning by extracting and fusing features layer by layer, advancing from local patterns to global laws [19]. Furthermore, incorporating physical constraints (such as conservation laws and symmetries) into the training process guides the model to maintain physical consistency as its structure and parameters expand [20]. This physical consistency, in turn, enhances reliable knowledge transfer and ensures theoretical compatibility within data-driven modeling frameworks.

Then, can this progressive paradigm of scientific discovery be applied to data-driven knowledge discovery, enabling the discovery of complete theories in typical physics scenarios? Motivated by this concept, we integrate the strengths of both progressive and data-driven paradigms to propose Chain of Symbolic Regression (CoSR), a method for structured, interpretable physical knowledge discovery. This framework internalizes the primary modes of human knowledge discovery—invariance learning, multilayered compression, and scaling transformation—into an assembly process of symbolic knowledge chains, forming a data-driven discovery framework for complex systems. Built upon symbolic regression as a white-box foundation, and drawing inspiration from the forward propagation and hierarchical representation principles of deep neural networks, CoSR models the discovery process of physical laws as a dynamically growing knowledge chain that evolves from simplicity to complexity: Starting from raw data, it progressively extracts intermediate knowledge units with explicit physical meanings to construct hidden layers. The discoveries from each layer serve as augmented feature spaces input to the subsequent layer, ultimately composing complex physical laws and achieving the precise discovery of the knowledge structures embedded within data. This architecture strictly adheres to the progressive cognitive mode of "from simplicity to complexity" inherent in scientific discovery, characterized by clear physical prior constraints and hierarchical features. We demonstrate the potential of CoSR across multiple physics examples with complex hierarchical couplings, showing that CoSR can not only discover classical physical laws but also automatically uncover concise and interpretable new laws from data.

II. METHODS

A. Overview of progressive discovery

We introduce Chain of Symbolic Regression (with its overall architecture and workflow illustrated in Fig. 1), which systematically constructs physical knowledge chains through progressive steps. First, in the invariance learning stage, physical quantities are nondimensionalized on the basis of the Buckingham π theorem, achieving dimensionality reduction while maintaining dimensional consistency. Implicit symbolic regression is then applied to automatically uncover intrinsic constraint relationships within the data, laying the groundwork for subsequent modeling. Second, in the multilayer compression stage, the dimensionless parameters obtained from the previous stage serve as inputs. A hierarchical symbolic regression approach is employed to progressively extract hierarchical structures among variables, constructing intermediate knowledge units with clear physical meanings. Finally, in the scaling transformation stage, the established mathematical relations undergo necessary refinement and simplification via transformed symbolic regression (e.g., "curves to lines" to reduce complexity and "multiple lines to one" to unify scaling laws), yielding a final knowledge representation layer that achieves a trade-off among formal parsimony, physical meaningfulness, and predictive fidelity. The knowledge chain ultimately constructed by CoSR bears structural resemblance to symbolic neural networks [21], yet they differ fundamentally: The knowledge chain based on CoSR is dynamically adapted rather than prespecified; its construction process is bidirectionally driven—both bottom-up for progressive refinement and top-down for simplification and condensation. This framework enables us to precisely extract universal laws embedded within data while clearly revealing the transparent process of physical knowledge assembly, allowing observation and understanding of how complex laws emerge stepwise from simple physical building blocks. The following subsections detail the methodology, including dimensional analysis, hierarchical discovery, implicit discovery, and transformation discovery.

To demonstrate the capability of the CoSR framework in scientific discovery, we conducted validation and application cases across multiple domains. First, by reproducing the discovery process of classical physical laws in celestial mechanics (the law of circular motion $\rightarrow$ Kepler's third law $\rightarrow$ Newton's law of universal gravitation), we validated its effectiveness within classical physics. Furthermore, we extended CoSR to complex problems such as turbulent Rayleigh-Bénard convection, viscous flows in a circular pipe, laser-metal interaction, and aerodynamic coefficient scaling for different aircraft, systematically revealing the knowledge chains that evolve from simplicity to complexity and exploring previously unrecognized physical laws.

B. Dimensional analysis

Dimensional analysis is a methodological approach that uncovers physical laws by examining the fundamental dimensions of physical quantities—such as mass, length, and time. Its core principle lies in the requirement that mathematical expressions of physical laws must remain invariant

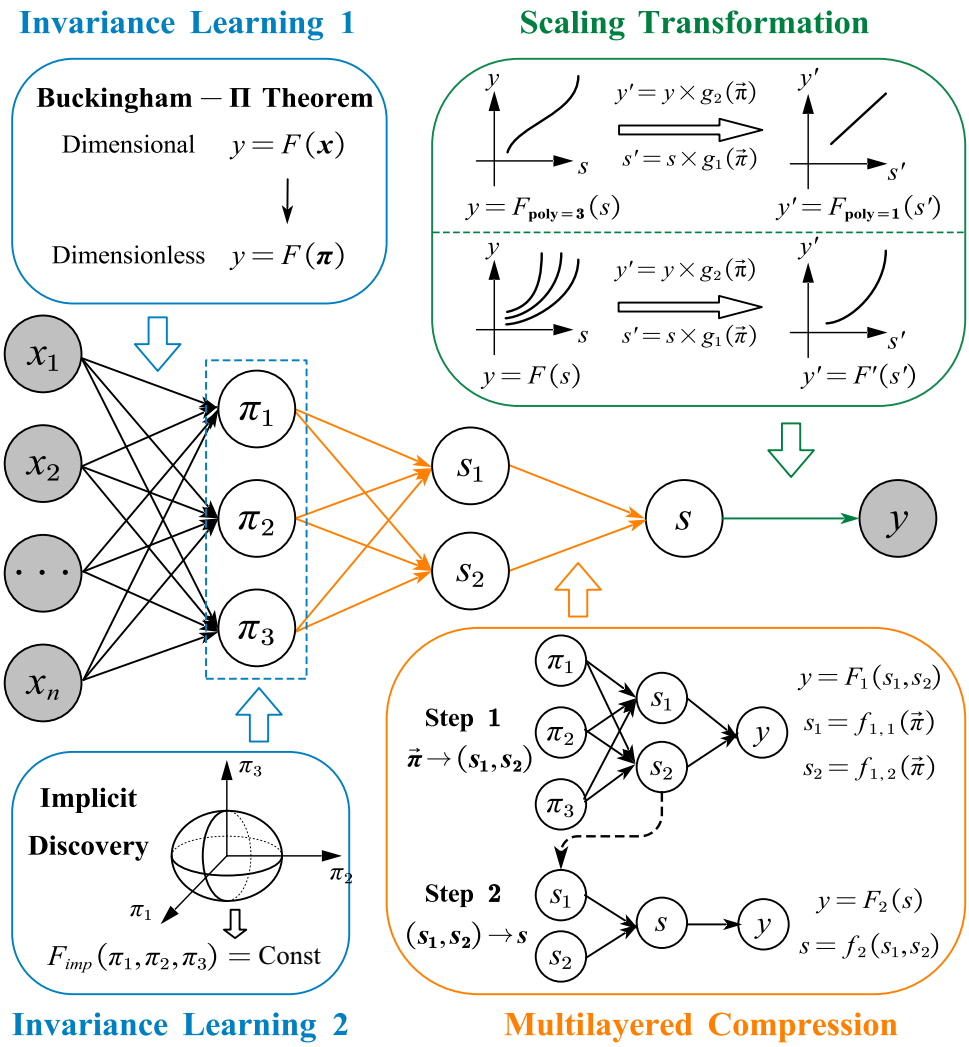


FIG. 1. Schematic illustration of Chain of Symbolic Regression framework. This figure systematically depicts the complete workflow of the CoSR framework, which constructs physical knowledge chains through a series of progressive steps. Invariance learning: Dimensionality reduction of the parameter space via nondimensionalization based on the Buckingham π theorem, followed by implicit symbolic regression to uncover intrinsic constraint relationships within the data. Multilayered compression: Hierarchical symbolic regression through multilevel nested functions to progressively extract knowledge structures. Scaling transformation: Refinement and simplification of expressions via transformation techniques (top: "curves to lines" to reduce complexity; bottom: "multiple lines to one" to unify scaling laws). These progressive discovery modes operate in concert through a physics-guided dynamic switching mechanism, culminating in a knowledge chain that achieves a balance among formal parsimony, physical interpretability, and predictive accuracy.

under changes in the choice of measurement units [22,23]. In the context of knowledge discovery, dimensional analysis operates through three types of constraints.

Buckingham π theorem. A relationship involving n physical variables and k fundamental dimensions, expressed as $f(Q_1, Q_2, \ldots, Q_n) = 0$, can be transformed into a relationship $F(\pi_1, \pi_2, \ldots, \pi_p) = 0$ among $p = n - k$ independent dimensionless π groups. This elevates specific physical problems to the level of universal scaling laws.

Normalization. By selecting characteristic quantities such as a reference velocity or mass, physical variables are expressed as relative ratios—examples include the Mach number and reduced mass. This is a specific form of nondimensionalization.

Dimensional homogeneity. As the most fundamental constraint that any physical equation must satisfy, dimensional homogeneity requires that all terms on both sides of an equation, as well as those being added or subtracted, possess identical dimensional compositions. This ensures that the equation describes an intrinsic physical relationship rather than a numerical coincidence arising from a particular system of units.

Data-driven dimensional analysis has revitalized this classical approach by integrating machine learning and large-scale data mining techniques. Researchers can now automatically identify physically meaningful dimensionless combinations from high-dimensional, complex observational or simulation data—and even uncover key scaling laws that are difficult to anticipate through traditional methods [17,24–27].

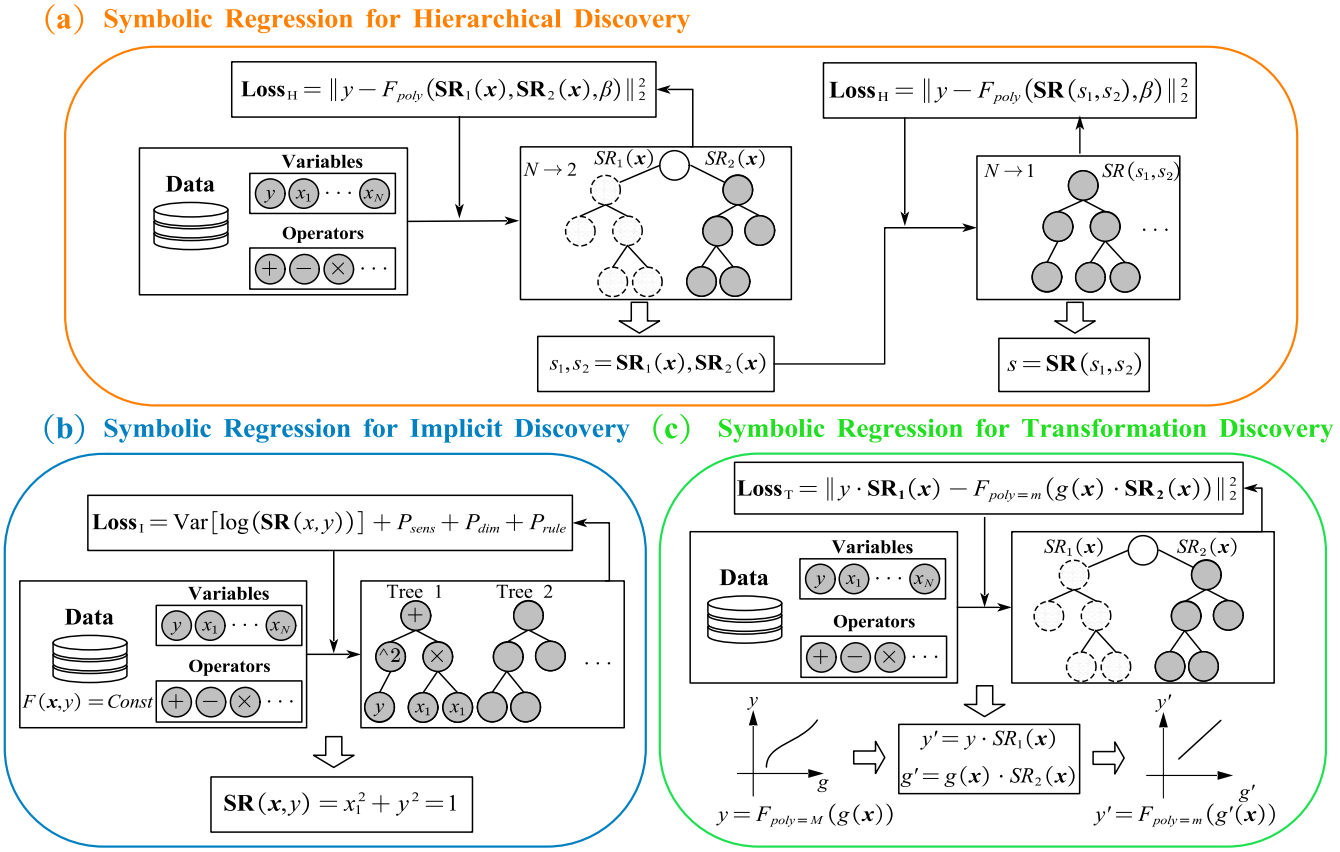


FIG. 2. Three strategies for knowledge discovery via symbolic regression. (a) Hierarchical discovery: Illustration of a layerwise discovery process based on composite function modeling. (b) Implicit discovery: A discovery approach grounded in the identification of implicit relationships. (c) Transformation discovery: Simplifying expressions through mathematical reformulation.

C. Symbolic regression

Symbolic regression is a method for automatically discovering mathematical expressions from data, offering both accurate data fitting and interpretability through concise representations. In this paper, the symbolic regression process of CoSR is implemented using the PySR method [10], which combines evolutionary algorithms with parallelized optimization to achieve high search efficiency and a strong success rate in recovering expressions. PySR provides a flexible application programming interface and allows for customizable loss functions (see Appendix B for the PySR parameter configurations and Sec. 2.2 of the Supplemental Material [28] for the loss functions and pseudocode). Building on this method, we employ CoSR to construct a physical knowledge chain through three complementary modes of discovery: hierarchical, implicit, and transformational. Together, these three modes constitute a complete pathway for knowledge discovery, as illustrated in Fig. 2.

D. Symbolic regression for hierarchical discovery

Hierarchical discovery aims to actively search for composite functions of the form $y = f_1(f_2(\cdots f_n(x)))$, where each subfunction f_i corresponds to an interpretable low-dimensional manifold or physical mechanism [29,30]. This paradigm mirrors the multiscale philosophy in physical modeling, in which a complex macroscopic phenomenon y is decomposed into a chain of interactions among a series of intermediate physical quantities. The process is implemented using the evolutionary symbolic regression method PySR, whose multibranch binary tree structure offers inherent flexibility for extracting two expressions in parallel [Fig. 2(a)]. To assess the validity of these intermediate variables, we construct a loss function based on multivariate polynomial fitting:

$$\mathrm{Loss}_H = \left\| y - F_{\mathrm{poly}}(\mathbf{SR}(x),\ \beta) \right\|_2^2, \tag{1}$$

where x denotes the variables from the previous layer, $\mathbf{SR}(x)$ represents the intermediate variables extracted via symbolic regression, and the number of such variables typically decreases layer by layer (e.g., from three to two, eventually converging to one), forming an inverted pyramidal structure of compression. The target variable is y, and $F_{\mathrm{poly}}(\mathbf{SR}(x), \beta)$ is generally a multivariate polynomial fit of order 1–5. This loss function is designed to select, at each layer, intermediate variables that are both parsimonious and critically influential on the output (if an extracted expression is overly complex, it can be split into two or more components). The newly obtained intermediate variables are then used as inputs in an iterative process to discover subsequent layers, ultimately forming a hierarchical knowledge network composed of knowledge units with clear physical meanings. For further details on

hierarchical discovery, see Sec. 2.2.1 of the Supplemental Material [28].

E. Symbolic regression for implicit discovery

An implicit relation $F(x, y) = \text{Const}$ defines an intrinsic constraint that the system variables must obey, describing an isosurface or a constrained manifold in the state space of the variables, rather than an explicit causal mapping [31–34]. The core idea is that a robust implicit relation should exhibit a function value $F(x, y)$ that fluctuates minimally around a specific constant, while remaining sensitive to perturbations in any of the constituent variables. To achieve this, building upon the work of Zhong *et al.* [34], we propose an improved loss function to guide the search process [Fig. 2(b)]:

$$\text{Loss}_I = \text{Var}[\log_{10}(\mathbf{SR}(x, y))] + P_{\text{sens}} + P_{\text{dim}} + P_{\text{rule}}. \quad (2)$$

Here, the variance minimization term $\text{Var}[\log_{10}(\mathbf{SR}(x, y))]$ enforces that all data points satisfy a constant constraint relationship. The sensitivity term P_{sens} imposes a variable contribution constraint, expressed as $P_{\text{sens}} = \sum_{i=1}^{n} \mathbf{I}[\|\mathbf{SR}(x, y) - \mathbf{SR}(\tilde{x}_i, y)\|_2^2 < \tau]$, where $\tilde{x}_i$ denotes the input after applying random perturbations to the ith feature, τ is the sensitivity threshold, and $\mathbf{I}[\cdot]$ is an indicator function that returns a penalty value when the condition is met. This constraint ensures that each variable plays a substantive role in the expression, thereby preventing trivial solutions. The dimensional term P_{dim} enforces physical plausibility by requiring that the expression $F(x, y)$ adhere to the principle of dimensional homogeneity. The regularity term P_{rule} imposes a constraint on the simplicity of the expression, prohibiting unreasonably large or small constants and nonsensical operations such as $x - x$ or x/x. Implicit discovery need not be applied if no underlying implicit equation is present. A detailed demonstration of implicit function discovery is provided in Sec. 2.2.2 of the Supplemental Material [28].

F. Symbolic regression for transformation discovery

Transformation discovery seeks to reformulate complex relationships into simpler forms within a transformed space through mathematical reconstruction [35–37], thereby reducing the complexity of the expression (if no reduction in complexity is achieved, transformation discovery may be omitted). The problem of transformation discovery is formulated as follows: an Mth order polynomial target relation $y = F_{\text{poly}=M}(g(x))$ is transformed into an mth order ($m < M$) polynomial form $y' = F_{\text{poly}=m}(g'(x))$, specifically $y \cdot \mathbf{SR}_1(x) = F_{\text{poly}=m}(g(x) \cdot \mathbf{SR}_2(x))$. Leveraging the multibranch tree structure of PySR, the two subfunctions $\mathbf{SR}_1(x)$ and $\mathbf{SR}_2(x)$ are automatically discovered in parallel [Fig. 2(c)]. The loss function directly minimizes the discrepancy between the two sides of the equation:

$$\text{Loss}_T = \left\| y \cdot \mathbf{SR}_1(x) - F_{\text{poly}=m}(g(x) \cdot \mathbf{SR}_2(x)) \right\|_2^2. \quad (3)$$

For instance, in Case 2 (turbulent Rayleigh-Bénard convection), $m = 1$ is employed to achieve a linear representation of a complex relationship. In Case 3 (viscous flows in a circular pipe), $m = 2$ is used to realize a unified characterization across different regimes. Additional discussions on alternative transformation forms are provided in Sec. 2.2.3 of the Supplemental Material [28].

III. RESULTS

A. The progressive discovery of the law of universal gravitation

The discovery of the law of universal gravitation (proceeding from Tycho Brahe's observational data, through Kepler's empirical laws and Huygens' analysis of centrifugal force, to Newton's ultimate synthesis) exemplifies the progressive discovery paradigm. This historical trajectory constitutes a complete physical knowledge discovery chain: data observation and induction (Kepler's third law) → dynamical analysis (formula for circular centripetal force) → fundamental interaction (law of universal gravitation). In this benchmark, we employ the CoSR framework to recapitulate this scientific discovery process.

To validate the framework's generality, we curated two representative classes of astronomical datasets: Solar System data [38] and exoplanetary system data [39], including the Trappist-1 and GJ 667 systems, as well as binary star system data [40]. These datasets encompass key observational parameters: the mass pair (M, m), orbital distance R, and orbital period T. To obtain numerical values for the gravitational force F (the output parameter), we computed it from kinematic parameters within the framework of Newton's second law. The target relationship can thus be formulated as $F = f(M, m, R, T)$, as illustrated in Fig. 3(a). It should be noted that this calculation procedure is entirely independent of the specific form of the law of universal gravitation, ensuring that the final form of the gravitational law emerges progressively through symbolic regression.

In implementing the methodology [Fig. 3(b)], we first normalize all physical quantities to comparable orders of magnitude (using the Sun-Earth distance, solar mass and Earth year as units for length, mass, and time, respectively), hereby mitigating numerical disparities that could interfere with symbolic regression. Subsequently, CoSR pursues progressive discovery along two independent chains. For the Solar System data, where the solar mass M remains constant, CoSR applies hierarchical discovery to the parameter set (m, R, T), yielding the dynamical relationship $F = m(4\pi^2/T^2)R$. Concurrently, implicit symbolic regression identifies Kepler's third law: $T^2/R^3 = \text{Const} = (4\pi^2)/GM$. For the exoplanet and binary star system data, symbolic regression operating on the parameter set (M, m, R, T) automatically discovers the reduced mass $\mu = mM/(M + m)$—a structural transformation that substantially simplifies the expression—leading to the dynamical relation for binary systems:

$$F = \mu \frac{4\pi^2}{T^2} R = \frac{mM}{M + m} \frac{4\pi^2}{T^2} R \quad (4)$$

and the generalized Kepler's third law via implicit symbolic regression:

$$\frac{(M + m)T^2}{R^3} = \text{Const} = \frac{4\pi^2}{G}, \quad (5)$$

which reduces to the familiar Kepler's third law in the limit $M \gg m$. Ultimately, CoSR combines the circular motion formula and Kepler's third law to yield the universal law of

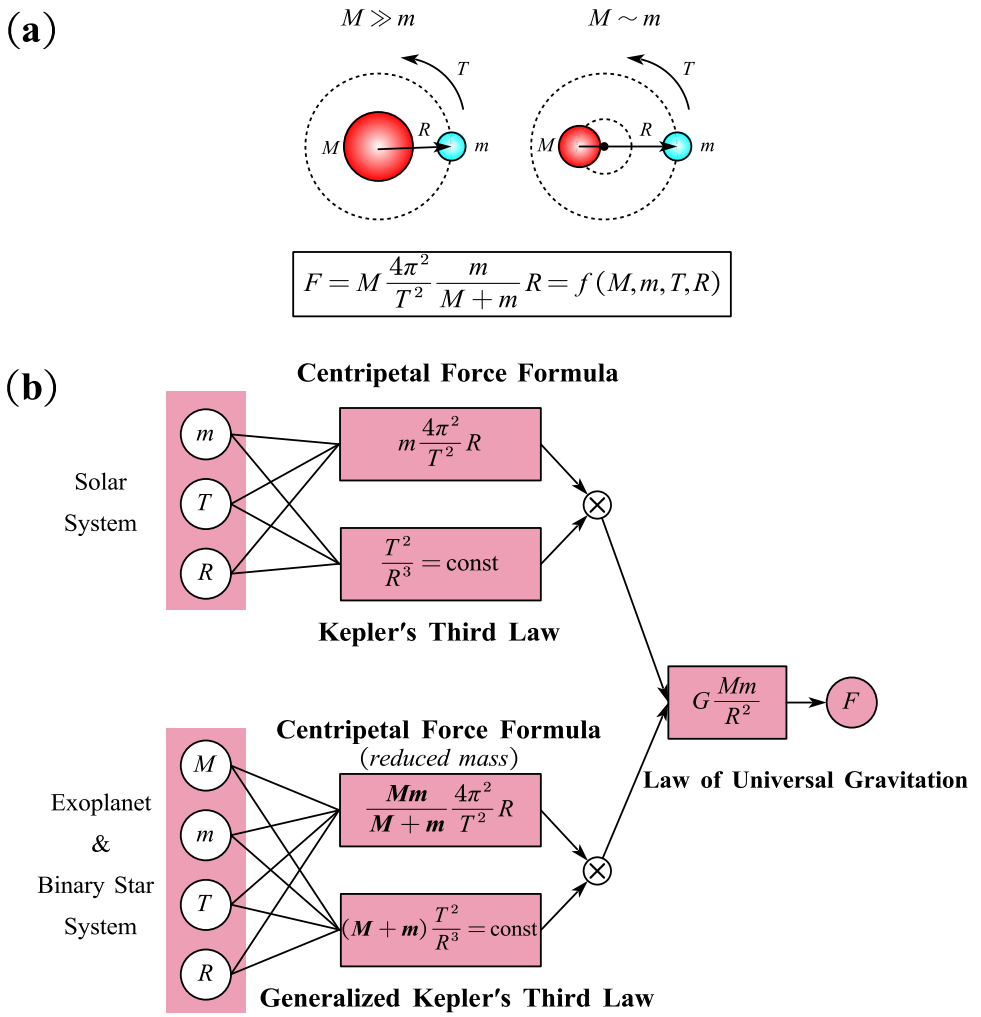


FIG. 3. Progressive discovery pathway of the law of universal gravitation. (a) Problem description: Celestial systems are categorized into two prototypical models based on the mass ratio between the central body and its orbiting companion: planet-star systems (where $M \gg m$) and binary star systems (where $M \sim m$ in order of magnitude). The relationship to be uncovered by CoSR is $F = f(M, m, T, R)$. (b) Progressive symbolic network chain: When applied to Solar System data, the framework extracts dynamical relationships through hierarchical discovery and identifies Kepler's third law via implicit discovery. In exoplanet and binary star systems, it automatically identifies the reduced mass, yielding the corresponding dynamical relationships with the reduced mass and a generalized formulation of Kepler's third law. These two discovery trajectories are ultimately synthesized to obtain the universal law of gravitation, thereby demonstrating the unification of physical laws.

gravitation. This process distills the governing relationship into a more general and parsimonious form, reducing the number of variables from four to three. The consistency of this discovery across both Solar System and extrasolar datasets robustly validates the unified nature of the underlying gravitational law. Further details regarding this incremental discovery benchmark are provided in Sec. 3.1 of the Supplemental Material [28].

This benchmark case demonstrates the effectiveness of the CoSR framework in discovering physical laws. It should be noted that the present benchmark operates within the conceptual framework of Newtonian mechanics; deeper physical principles or concepts—such as the symmetry of forces and the nature of interactions—remain beyond the current scope of symbolic regression discovery. Nevertheless, the framework automatically identifies the key physical quantity of reduced mass and uncovers the generalized Kepler's third law, achieving both deep mining and theoretical extension. Looking ahead, the automated creation of fundamental physical concepts may prove more transformative than the construction of specific mathematical relations alone.

B. Turbulent Rayleigh-Bénard convection

Rayleigh-Bénard convection describes the turbulent motion in a fluid layer heated from below and cooled from above, driven by density differences induced by temperature gradients. The fundamental question concerns the intrinsic relationship between heat transfer efficiency and the governing parameters. The system involves key physical quantities: heat flux q, container height h, temperature difference ΔT, gravitational acceleration g, thermal conductivity λ, thermal expansion coefficient α, kinematic viscosity ν, and thermal diffusivity κ. Heat transfer efficiency is characterized by the

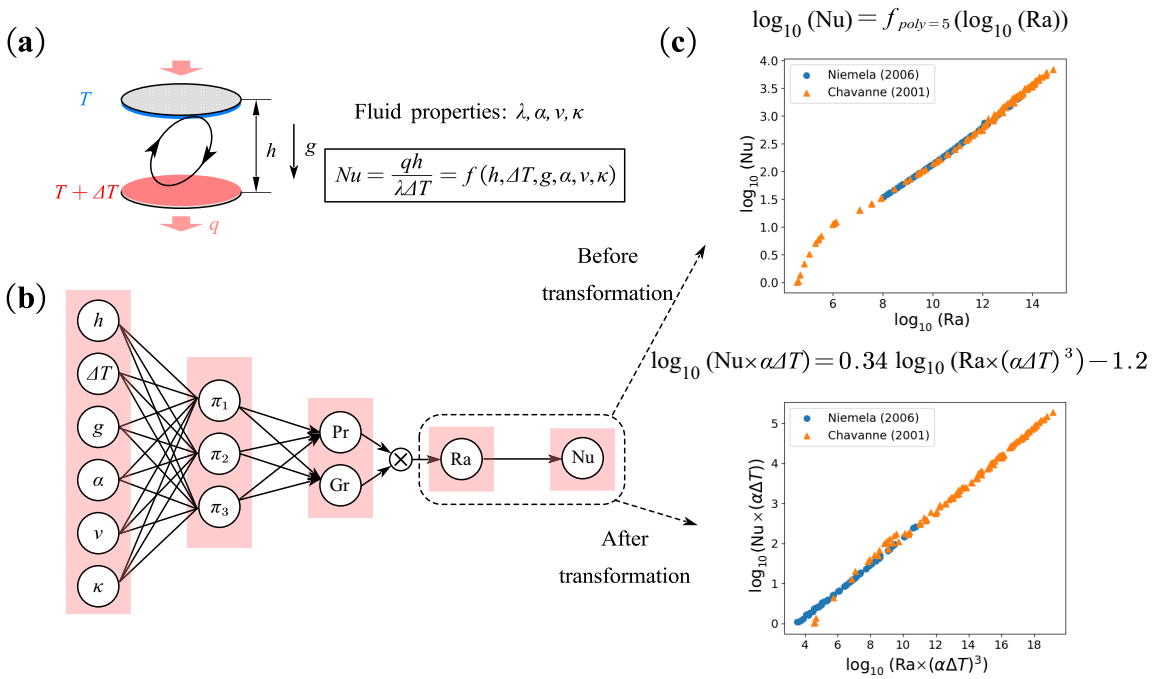


FIG. 4. Progressive discovery pathway for turbulent Rayleigh-Bénard convection. (a) Problem description: Schematic illustration of the Rayleigh-Bénard convection system, along with the input and output parameters. (b) Progressive symbolic network chain: Starting from six raw parameters, hierarchical discovery progressively identifies the Prandtl number, the Grashof number, and ultimately the Rayleigh number through a structured pathway. (c) Refinement via scaling transformation: The conventional nonlinear Nu-Ra relationship is transformed and simplified into a linear scaling law through correction transformations, revealing deeper physical insights.

Nusselt number, defined as $\mathrm{Nu} = qh/(\lambda \Delta T)$. The target relationship can thus be formulated as $\mathrm{Nu} = f(h, \Delta T, g, \alpha, \nu, \kappa)$, as illustrated in Fig. 4(a). This study employs experimental data from two sources [41,42], encompassing a wide range of parameter space.

In implementing the methodology [Fig. 4(b)], we first apply the Buckingham π theorem to reduce the six raw physical quantities (h, ΔT, g, α, ν, κ) by dimensional analysis. These quantities involve three independent fundamental dimensions: length (L), time (T), and temperature (Θ). Hence, the six quantities yield $6 - 3 = 3$ independent dimensionless groups. Choosing h, ν, and ΔT as dimensionally independent repeating variables and solving the system of dimensional exponents yields $\pi_1 = \alpha \Delta T$, $\pi_2 = \nu \alpha \Delta T/\kappa$, and $\pi_3 = gh^3/\nu^2$. The original relationship is therefore simplified to $\mathrm{Nu} = f(\pi_1, \pi_2, \pi_3)$. Subsequently, the CoSR framework progressively identifies the Prandtl number $\mathrm{Pr} = \nu/\kappa$ (characterizing the ratio of momentum diffusivity to thermal diffusivity) and the Grashof number $\mathrm{Gr} = \alpha \Delta T g h^3/\nu^2$ (characterizing the ratio of buoyancy to viscous forces) through hierarchical discovery. Their product yields the Rayleigh number (a composite dimensionless parameter that encapsulates the relative importance of buoyancy, viscous dissipation, and thermal diffusion):

$$\mathrm{Ra} = \mathrm{Gr} \times \mathrm{Pr} = \frac{\alpha \Delta T\, g h^3}{\nu \kappa}. \tag{6}$$

As shown in Fig. 4(c), confronted with the high-order nonlinear relationship between Nu and Ra —expressed as $\log_{10}(\mathrm{Nu}) = f_{\mathrm{poly}=5}(\log_{10}(\mathrm{Ra}))$—the CoSR framework applies symbolic regression to discover the transformation $\log_{10}(\mathrm{Nu} \times T_{\mathrm{Nu}}(\pi_1, \pi_2, \pi_3)) = f_{\mathrm{poly}=1}(\log_{10}(\mathrm{Ra} \times T_{\mathrm{Ra}}(\pi_1, \pi_2, \pi_3)))$. Specifically, it identifies a correction term $T_{\mathrm{Nu}} = \alpha \Delta T$ for Nu and a correction term $T_{\mathrm{Ra}} = (\alpha \Delta T)^3$ for Ra , transforming the nonlinear scaling into a linear scaling law:

$$\log_{10}(\mathrm{Nu} \times \alpha \Delta T) = 0.34 \log_{10}(\mathrm{Ra} \times (\alpha \Delta T)^3) - 1.2. \tag{7}$$

Here, $\alpha \Delta T$ represents the thermal expansion effect induced by the temperature difference between the top and bottom walls. By introducing this parameter (previously overlooked), we establish a more parsimonious relationship between Nu and Ra, further highlighting its critical role in uncovering the physical mechanism of thermal convection. While previous studies have generally understood the physics of Rayleigh-Bénard convection through a nonlinear power-law relationship between Nu and Ra , our symbolic regression analysis reveals a deeper physical essence: a linear power-law relationship involving the synergistic interaction among $\alpha \Delta T$, Nu , and Ra . Notably, although this scaling formula differs in form from that obtained by Xie *et al.* [26] using dimensionless learning, the two formulations are inherently self-consistent. This finding may offer physical insights into complex turbulent Rayleigh-Bénard convection. Alternative formulations are provided in Sec. 3.2 of the Supplemental Material [28].

C. Viscous flows in a circular pipe

Turbulent flow in rough pipes constitutes a core component of industrial transport systems, and its flow resistance characteristics directly influence energy consumption and operational efficiency, which has been a long-standing research focus in fluid mechanics [43–48]. This section

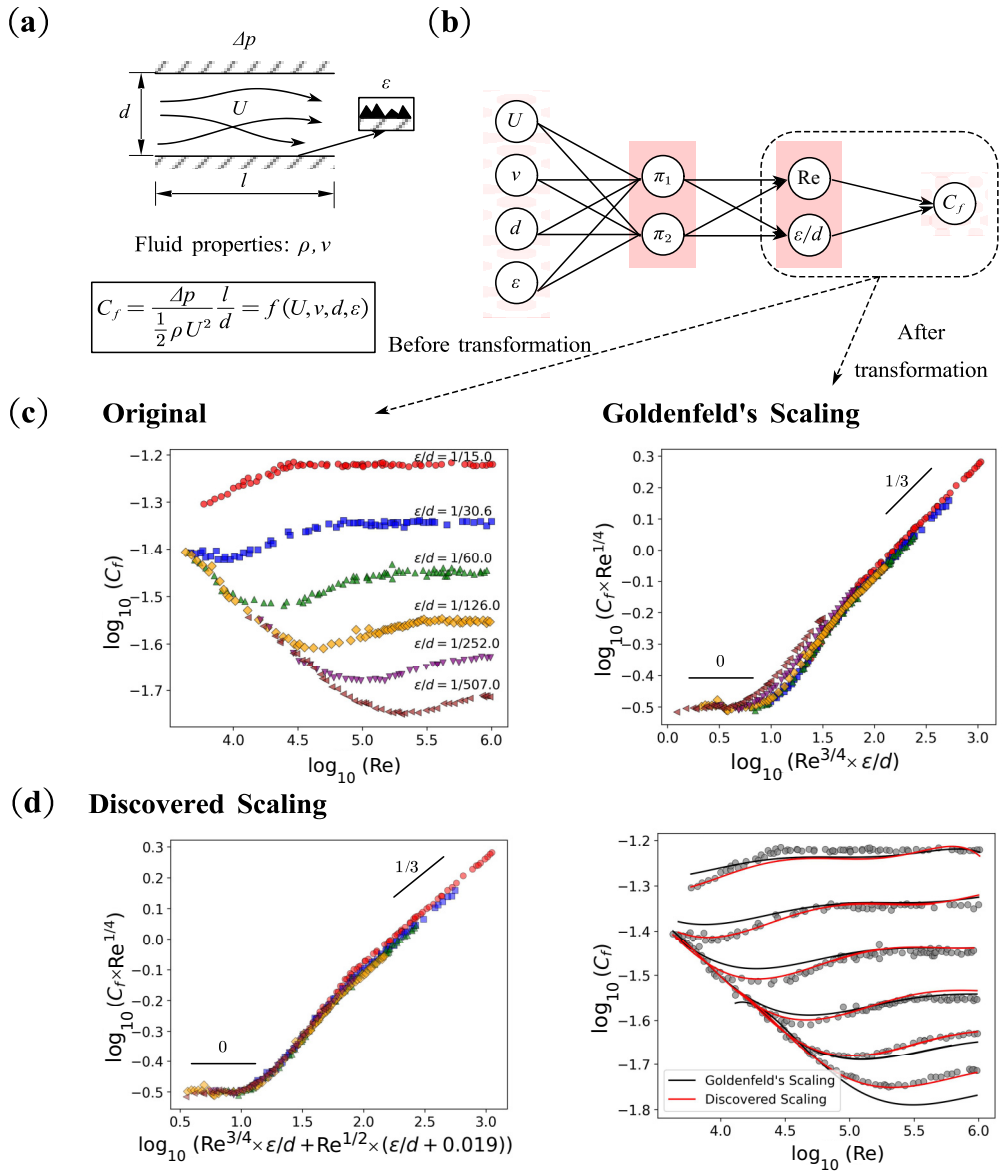


FIG. 5. Progressive discovery pathway for rough-wall pipe flow. (a) Problem description: Schematic illustration of rough-wall pipe flow, along with the input and output parameters. (b) Progressive symbolic network chain: Structured discovery pathway from raw parameters through dimensional analysis to the optimal dimensionless combination (Re, ε/d). (c), (d) Refinement via scaling transformation: Through transformations, the conventional C_f–Re–(ε/d) relationship (typically requiring complex piecewise power-law descriptions) is reconstructed into a parsimonious and unified scaling form. This approach not only reproduces the classic Goldenfeld's scaling law but further yields a scaling form with superior data collapse. Based on this formulation, reconstructing C_f using polynomials achieves improved prediction accuracy, particularly in the transitional turbulent regime.

systematically analyses flow characteristics in the turbulent regime using the classic experimental dataset of Nikuradse [49]. The system involves key physical quantities: pressure drop Δp, mean flow velocity V, fluid kinematic viscosity ν, pipe diameter d, and wall roughness height ε. Flow resistance is characterized by the Darcy friction factor $C_f = \Delta p d/(0.5\rho V^2 L)$, with the target relationship formulated as $C_f = f(V, \nu, d, \varepsilon)$, as illustrated in Fig. 5(a).

Following dimensional reduction of the four raw physical quantities, CoSR automatically identifies the Reynolds number $\mathrm{Re} = Vd/\nu$ and relative roughness ε/d as the optimal parameter combination from among numerous possible dimensionless groups [Fig. 5(b)]. Owing to the highly nonlinear and regime-dependent nature of conventional friction factor relationships, a direct formulation of the form $C_f = f(\mathrm{Re}, \varepsilon/d)$ proves inadequate. Through transformation discovery, CoSR constructs a unified scaling form $C_f f_1(\mathrm{Re}, \varepsilon/d) = G(f_2(\mathrm{Re}, \varepsilon/d))$, yielding

$$C_f\,\mathrm{Re}^{1/4} = G(\mathrm{Re}^{3/4} \times \varepsilon/d). \tag{8}$$

This is Goldenfeld's scaling form (a Widom-type scaling law [43]), which successfully unifies the previously complex piecewise power-law regimes into a parsimonious scaling representation [Fig. 5(c)]. From a physical mechanism perspective, this relation captures the scaling behavior of two established regimes. In the fully rough regime (large Re),

the system tends to follow the Strickler law, where C_f scales as $(\varepsilon/d)^{1/3}$. Within the framework of Eq. (8), this 1/3 slope corresponds to $G(x) \sim x^{1/3}$, which leads to the cancellation of Re on both sides of the equation, yielding C_f independent of Re. In the hydraulically smooth regime, the system follows the Blasius law, where C_f becomes independent of ε/d and scales as $\mathrm{Re}^{-1/4}$, corresponding to $G(x) \to$ constant in Eq. (8). This scaling law appropriately bridges these two asymptotic regimes, providing a unified framework for understanding cross-scale behavior in rough-wall pipe flow.

Importantly, the discovery pathway itself holds significant value: Rather than relying on traditional theoretical derivation and assumptions, this form was automatically discovered through symbolic regression, substantially accelerating the cognitive process from data to physical laws.

Building upon Goldenfeld's scaling, we further employ symbolic regression to discover a better scaling form that achieves superior data collapse and prediction performance:

$$C_f\,\mathrm{Re}^{1/4} = G\big(\mathrm{Re}^{3/4} \times \varepsilon/d + \mathrm{Re}^{1/2}(\varepsilon/d + 0.019)\big). \quad (9)$$

This formulation extends Goldenfeld's scaling while preserving the asymptotic behavior of the two established regimes. In the fully rough regime ($\mathrm{Re} \to \infty$), $\mathrm{Re}^{3/4}\varepsilon/d$ dominates over $\mathrm{Re}^{1/2}(\varepsilon/d + 0.019)$, recovering the original Goldenfeld form. In the hydraulically smooth limit ($\varepsilon/d \to 0$), the argument of G reduces to $0.019\,\mathrm{Re}^{1/2}$. Although this argument depends on Re, the function G has a zero slope (i.e., remains constant) in the small-argument range corresponding to smooth pipes. Consequently, $C_f\mathrm{Re}^{1/4}$ tends to a constant value, consistent with the Blasius law. The constant 0.019 plays a role analogous to the shape coefficient C_s in Tao's scaling [44], representing a lumped empirical parameter that accounts for the geometric irregularity of roughness elements and improves the data collapse in the transitional regime. The additional correction term substantially improves data collapse in the transitional turbulent regime, reducing the mean relative prediction error from 3.55% (Goldenfeld's scaling) to 2.88% (discovered scaling). Section 3.3 of the Supplemental Material [28] presents our scaling model alongside the classical Tao's scaling, demonstrating superior collapse performance of the discovered formulation.

D. Laser-metal interaction

Keyhole dynamics arising from laser-metal interactions constitute a critical physical process in military and industrial applications, including laser materials processing, high-energy laser weapons, and aerospace propulsion systems [50–52]. Quantifying keyhole dynamics is essential for optimizing processing quality and controlling process stability. This study investigates the intrinsic physical relationship between keyhole depth e and laser parameters coupled with material properties, using experimental data from the literature [26] encompassing three materials: titanium alloy (Ti6Al4V), aluminium alloy (Al6061), and stainless steel (SS316), as illustrated in Fig. 6(a). Key input parameters include effective laser power ηP, laser scanning speed V_s, laser beam radius r_0, and material property parameters: thermal diffusivity α, material density ρ, heat capacity C_p, and the temperature difference between melting point and ambient conditions $T_1 - T_0$. The target relationship can be formulated as $e^* = e/r_0 = f(\eta P, V_s, r_0, \alpha, \rho, C_p, T_1 - T_0)$.

In the progressive discovery process, CoSR first reduces the seven raw physical quantities to three dimensionless groups through dimensional analysis [Fig. 6(b)]. These groups are then progressively transformed via hierarchical discovery into a system of dimensionless parameters with clear physical meanings. Among these, the keyhole number

$$\mathrm{Ke} = \frac{\eta P}{(T_1 - T_0)\,\rho C_p \sqrt{\alpha V_s r_0^3}} \quad (10)$$

characterizes the dynamic balance between laser energy input and the material's thermal absorption capacity, while the material characteristic number

$$X = \frac{\alpha/r_0}{\sqrt{C_p(T_1 - T_0)}} \quad (11)$$

reveals the relative importance of the material's thermal diffusion capability (α/r_0) versus its heat storage characteristics [$C_p(T_1 - T_0)$]. Notably, the material characteristic number X discovered by the framework achieves a natural physical classification of the three experimental materials: Al6061 > Ti6Al4V (for both r_0 values) > SS316 [Fig. 6(c)]. This classification, which has not been reported in previous studies, reflects differences in the thermal response characteristics of these materials under laser irradiation.

Building upon this foundation, CoSR further constructs [Fig. 6(d)], through hierarchical discovery, a modified keyhole number Ke^* (a non-power-product scaling parameter):

$$\mathrm{Ke}^* = \frac{\mathrm{Ke}}{1 + \mathrm{Ke} \times X}, \quad (12)$$

where the prediction model based on Ke^* achieves a higher coefficient of determination ($R^2 = 0.982$) compared to the Ke model ($R^2 = 0.977$). For all three materials (Ti6Al4V, Al6061, and SS316), the Ke^* model reduces both the mean absolute error and the upper error bound. The most evident improvement is observed for aluminium alloy (Al6061), which is characterized by its exceptional thermal diffusion properties. For this material, the upper error bound decreases by approximately 43.3% [Fig. 6(e)]. Analysis reveals that when the product $\mathrm{Ke} \times X$ is small (corresponding to stainless steel and titanium), keyhole dynamics can be approximately described by the classical keyhole number Ke; however, when $\mathrm{Ke} \times X$ reaches the order of 10^{-1} (corresponding to aluminium alloy), the prediction model based on Ke^* outperforms the classical Ke model (see Sec. 3.4 of the Supplemental Material [28] for further comparison). The modified keyhole number Ke^* may offer insights into the intrinsic regulatory mechanisms through which material thermal properties govern keyhole dynamics.

E. Aerodynamic coefficient scaling for different aircraft

In aerospace engineering, the study of aerodynamic characteristics constitutes a core component of aircraft design. The axial force A represents the aerodynamic force acting along the body axis of an aircraft. This coefficient is fundamentally determined by six physical quantities: angle of attack α, freestream velocity V_∞, speed of sound c, freestream density ρ_∞, characteristic length d, and freestream

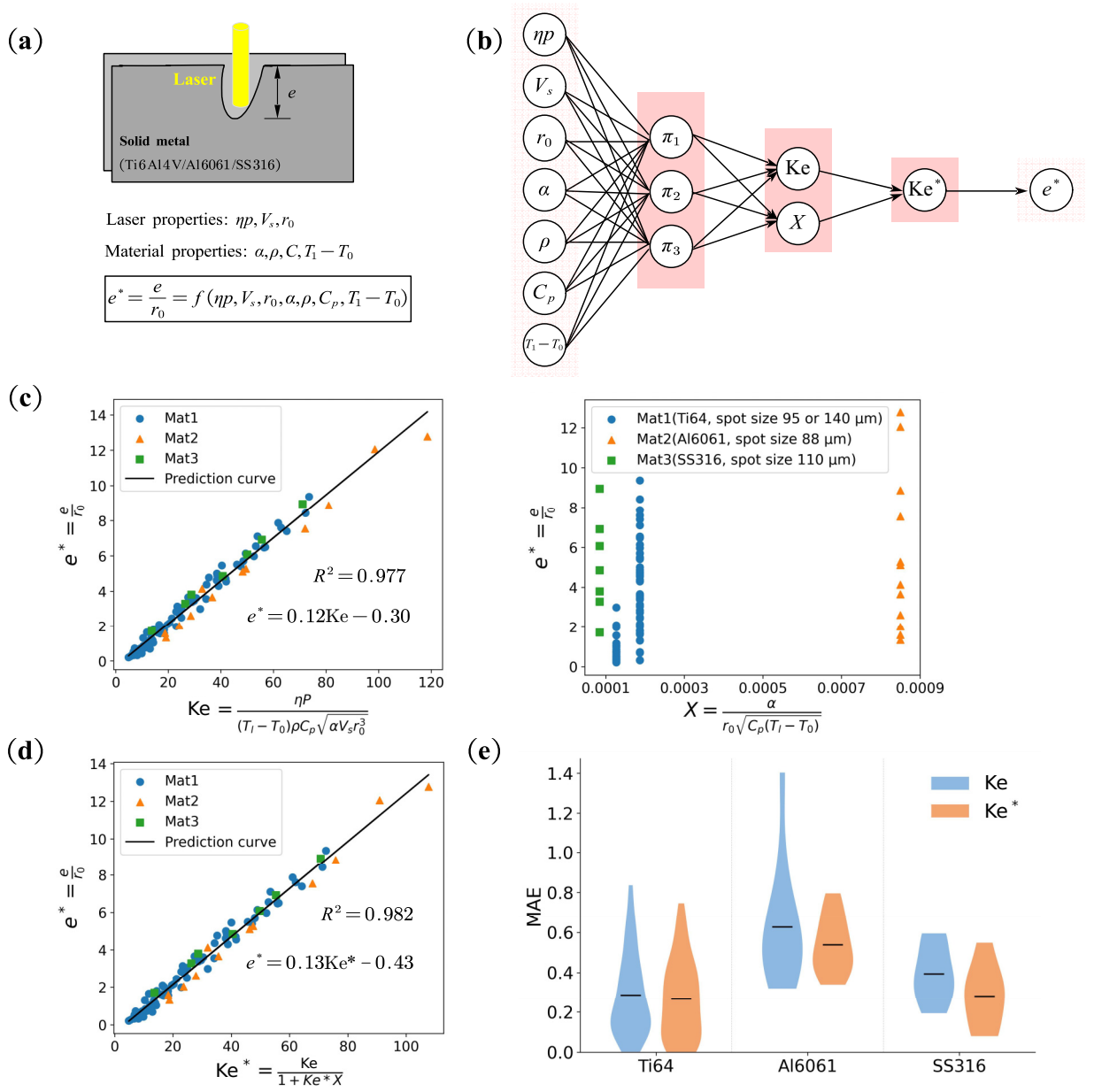


FIG. 6. Progressive discovery pathway for laser-metal interaction. (a) Problem description: Schematic illustration of laser-metal interaction, along with the input and output parameters. (b) Progressive symbolic network chain: Complete discovery pathway from seven raw parameters to key dimensionless groups. (c) Discovery of key parameters: Systematic presentation of two core dimensionless parameters automatically discovered by the framework: the keyhole number Ke and the material characteristic number X. (d) Dominant parameter extraction: Combination of Ke and X to discover the modified keyhole number Ke*. (e) Absolute prediction error distributions for the two keyhole numbers: Comparison demonstrating the superior predictive accuracy of the modified keyhole number Ke* over the classical Ke across three materials, with particularly notable improvement for Al6061.

dynamic viscosity μ_∞ [Fig. 7(a)]. Nondimensionalization of the axial force yields the axial force coefficient $C_A = A/(0.5\rho_\infty V_\infty^2 d^2)$, with the target relationship formulated as $C_A = f(\alpha, V_\infty, c, \rho_\infty, d, \mu_\infty)$. Through dimensional analysis, these six quantities reduce to three key dimensionless parameters: angle of attack α (dimensionless geometric parameter), Mach number $M = V_\infty/c$ (characterizing compressibility effects), and Reynolds number $\mathrm{Re} = \rho_\infty V_\infty d/\mu_\infty$ (characterizing viscous effects), thereby simplifying the target relationship to $C_A = f(\alpha, M, \mathrm{Re})$. This study systematically investigates the intrinsic scaling laws governing this relationship using hypersonic simulation data for sharp-cone and blunt-body configurations from Lin *et al.* [53].

Taking the sharp-cone geometry as an example, CoSR successfully identifies two core dimensionless parameters and constructs a complete scaling relationship through its progressive discovery process. The framework first identifies the dominant term

$$S_1 = \alpha^2 + \frac{c_1}{M^2}, \tag{13}$$

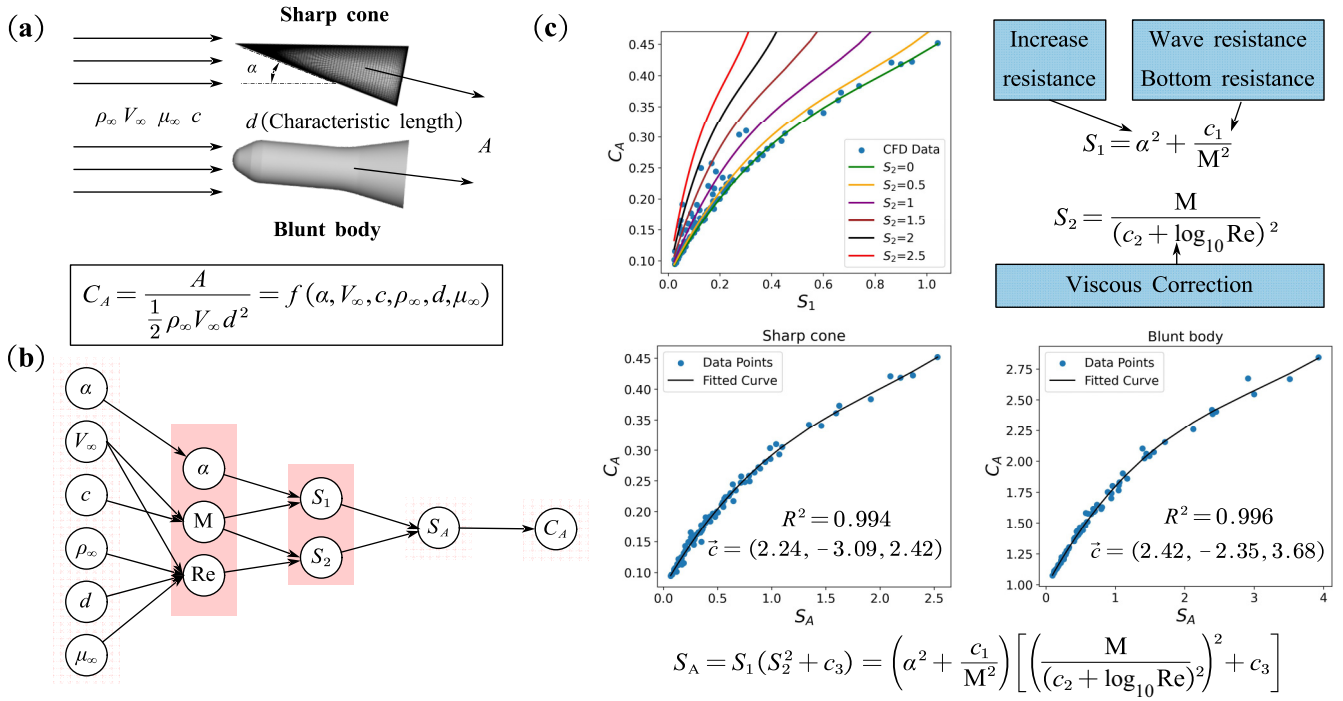


FIG. 7. Progressive discovery pathway for aerodynamic coefficient scaling across different aircraft configurations. (a) Problem description: Schematic illustration of two aircraft geometries (sharp cone and blunt body) along with the input and output parameters. (b) Progressive symbolic network chain: Structured discovery pathway from six raw physical parameters to three core dimensionless groups (α, M, Re) and subsequently to the axial force scaling parameters. (c) Discovery of key parameters: Systematic presentation of the dominant term S_1, the viscous correction term S_2, and the complete scaling relationship $S_A = S_1(S_2^2 + c_3)$ automatically identified by the framework.

where the α^2 term reflects lift-induced drag, the c_1/M^2 term captures shock wave and base drag effects, and c_1 is a constant generated by symbolic regression. The dominant scaling parameter S_1 satisfies the hypersonic invariance principle (flow characteristics tend to stabilize at extremely high Mach numbers in inviscid flow fields), and it can be derived theoretically [53]. Concurrently, the framework discovers the viscous correction term

$$S_2 = \frac{M}{(c_2 + \log_{10} \mathrm{Re})^2}, \tag{14}$$

where c_2 is a constant generated by symbolic regression. The viscous correction term S_2 is functionally similar to the viscous interaction parameter, effectively capturing viscous effects under hypersonic conditions. Finally, the framework establishes a non-power-product scaling relationship:

$$\begin{aligned} S_A &= S_1\left(S_2^2 + c_3\right) \\ &= \left(\alpha^2 + \frac{c_1}{M^2}\right)\left(\left(\frac{M}{(c_2 + \log_{10} \mathrm{Re})^2}\right)^2 + c_3\right), \end{aligned} \tag{15}$$

where c_3 is a constant generated by symbolic regression. When Reynolds number effects are neglected ($\mathrm{Re} \to \infty$, so $S_2 \to 0$), the expression reduces to $S_A \to c_3 S_1$. Because the outer function is a polynomial and the scaling parameter S_1 forms a linear cluster, the constant c_3 can be absorbed into S_1 via $\tilde{S}_1 = c_3 S_1$ without loss of generality. Thus, in the redefined sense, S_A reduces to $\tilde{S}_1$ (i.e., S_1) when Reynolds number effects are neglected, ensuring physical consistency. In this formulation, the viscous correction term represents the friction drag component, while c_3 corresponds to the pressure drag component [Figs. 7(b) and 7(c)].

Notably, for different aerodynamic geometries, the functional form of the scaling law S_A remains essentially identical; only the constant vector $\vec{c} = (c_1, c_2, c_3)$ varies. These constants can be optimized using algorithms such as L-BFGS-B and carry distinct physical interpretations. For instance, parameter fitting yields $\vec{c} = (2.24, -3.09, 2.42)$ for the sharp cone and $\vec{c} = (2.42, -2.35, 3.68)$ for the blunt body [Fig. 7(c)]. Parameter analysis reveals that (1) the larger c_1 value for the blunt body reflects its greater combined shock wave and base drag; (2) the smaller c_2 value for the sharp cone indicates a larger friction drag component compared to the blunt body; and (3) the substantially larger c_3 value for the blunt body arises from its greater pressure drag in the absence of viscous interaction effects. The negative values of c_2 do not imply a singularity at finite positive Reynolds numbers, because the formula is intended for turbulent flows where $\log_{10} \mathrm{Re}$ is large, and the divergence never occurs in the physically relevant regime. These results demonstrate the capability of the framework in addressing complex aerodynamic problems, providing interpretable theoretical foundations and novel pathways for knowledge discovery. Section 3.5 of the Supplemental Material [28] compares some discovery pathways (including power-product scaling and end-to-end

symbolic regression), showing that the end-to-end method often overfits and lacks physical interpretability, whereas CoSR recovers the correct structure.

IV. DISCUSSION

Existing approaches typically treat the target function F as a single expression to be discovered through direct search. This strategy not only triggers combinatorial explosion and overfitting but also yields a multitude of candidate "truth formulas" that defy interpretation—a phenomenon we term the "Rashomon effect" of knowledge discovery [54]. In this paper, we introduce CoSR, a framework grounded in the philosophy that complex physical laws are not monolithic expressions but rather structured systems assembled from multiple simple knowledge units with clear physical meanings through specific logical relationships. This epistemological shift has led us to abandon the traditional path of directly searching for the global function $y = F(x)$ in favor of a structured decomposition strategy that transforms high-dimensional search spaces into tractable hierarchical subtasks. Through three sequential tiers, CoSR systematically constructs symbolic knowledge chains: from dimensionally constrained dimensionality reduction via the Π theorem and implicit function discovery, to hierarchical discovery driving multilevel knowledge unit extraction, and finally to transformation and reconstruction—together constituting a unified and self-consistent methodology for scientific discovery in complex systems.

By decomposing symbolic regression into structured layers, this methodology reduces high-dimensional complexity while establishing a reproducible pathway for systematically uncovering unknown laws from data. This stepwise integration of information across layers enables the framework to move beyond surface-level correlations, excavating the inherent laws hidden behind variables and achieving structured representation of previously unresolved physical mechanisms. Rather than proposing a symbolic regression algorithm, we harness the potential of existing methods through a progressive strategy that simulates the "assembly" process of physical knowledge. This allows us to observe how complex laws emerge stepwise from simple physical building blocks—a perspective that ultimately serves the broader vision of AI for Science.

At the application level, this paper systematically demonstrates the framework's capability to address real-world scientific problems through five representative cases. In classical mechanics, the framework not only recapitulated the discovery pathway from Kepler's third law to the law of universal gravitation but also extracted the reduced mass representation and automatically discovered the generalized Kepler's law applicable to binary systems. In turbulent systems, the discovered linear scaling law not only satisfies the established $1/3$ scaling exponent but also offers theoretical perspectives on energy transfer mechanisms in turbulent thermal transport. In rough pipe flow studies, the constructed Widom-type scaling law successfully bridges the two classical regimes corresponding to the Strickler and Blasius laws, replacing previous piecewise scaling representations and achieving unified characterization of all three turbulent regimes in rough-wall pipe flow. In laser-metal interaction, the framework automatically identifies a material characteristic number X that enables natural physical classification of three materials and discovers a modified keyhole number Ke^* that reduces prediction errors for aluminium alloys. In aerodynamic coefficient scaling problems, a parameterized non-power-product scaling function was established, whose parameters (c_1, c_2, c_3) quantitatively reveal the essential distinctions between sharp and blunt bodies in terms of shock wave intensity, viscous effects, and base drag, achieving unified characterization of aerodynamic mechanisms across different geometries.

Although CoSR has demonstrated remarkable knowledge discovery capabilities across multiple physical systems, achieving the precise discovery of physical knowledge from existing data, certain limitations remain to be addressed. First, when processing high-dimensional complex systems, symbolic regression may generate multiple mathematically equivalent yet physically inconsistent expressions; currently, selection among candidates still relies on researchers' physical intuition in some instances. Second, efficient coordination among the three discovery modes still depends on experience-based strategies. While the proposed progressive pathway of "invariance + hierarchy + transformation" has yielded promising results, future work requires developing more intelligent symbolic regression discovery strategies to achieve autonomous optimization of discovery pathways. Finally, while symbolic regression can discover new physical laws, it is constrained to uncovering relationships within predefined human-specified concepts (such as "force" in gravitation) rather than autonomously generating new concepts or paradigms. Future research should strive toward realizing a complete scientific discovery chain from concept creation to theoretical system construction.

In summary, CoSR successfully transforms the scientific discovery philosophy of "from simplicity to complexity" into a practicable methodology, offering technical pathways and theoretical paradigms for automated scientific discovery. As research progresses, this framework holds promise for playing a significant role in broader scientific discovery tasks, ultimately advancing from assistive tool to AI scientist.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grants No. 92152301, No. U2441211, and No. U23B6009) and the National Key R&D Program of China (Grant No. 2022YFB4300200).

Methodology, writing, visualization, and validation: M.X. Conceptualization, project administration, methodology, and funding acquisition: W.Z.

The authors declare no competing interests.

DATA AVAILABILITY

All data and codes used in this paper are publicly available on GitHub [55].

TABLE I. Extracted expressions under different noise levels.

Case	Target	0.01	0.05	0.1	0.15	0.2
Law of universal gravitation	F_1	$\frac{mM}{M+m}\frac{R}{T^2}$	$\frac{mM}{M+m}\frac{R}{T^2}$	×	×	×
	F_2	$(M+m)\frac{T^2}{R^3}=\text{Const}$	$(M+m)\frac{T^2}{R^3}=\text{Const}$	×	×	×
	F_3	$G\frac{mM}{R^2}$	$G\frac{mM}{R^2}$	×	×	×
Turbulent Rayleigh-Bénard convection	(S_1, S_2)	Gr, Pr	Gr, Pr	Ra	Ra	Ra, $\frac{gh^3}{\nu^2}$
	S_A	Ra	Ra	Ra	Ra	×
	$(S_{\text{lc}}, S_{\text{rc}})$	$\alpha\Delta T$, $(\alpha\Delta T)^3$	$\alpha\Delta T$, $(\alpha\Delta T)^3$	$\alpha\Delta T$, $(\alpha\Delta T)^3$	$\alpha\Delta T\frac{\nu}{\kappa}$, $\left(\alpha\Delta T\frac{\nu}{\kappa}\right)^3$	×
Viscous flows in a circular pipe	(S_1, S_2)	Re, $\frac{\varepsilon}{d}$	Re, $\frac{\varepsilon}{d}$	Re, $\frac{U\varepsilon}{\nu}$	Re, $\frac{U\varepsilon}{\nu}$	×
	$(S_{\text{lc}}, S_{\text{rc}})$	$\text{Re}^{1/4}$, $\text{Re}^{3/4}\frac{\varepsilon}{d}$	$\text{Re}^{1/4}$, $\text{Re}^{3/4}\frac{\varepsilon}{d}$	×	×	×
Laser-metal interaction	(S_1, S_2)	Ke, X	Ke, $\frac{\alpha V_s}{r_0 C_p(T_1-T_0)}$	Ke, $\frac{\alpha}{V_s r_0}$	$\frac{V_s r_0^2 \rho C_p(T_1-T_0)}{\eta P}$, $\frac{\alpha}{V_s r_0}$	×
	S_A	Ke*	Ke	Ke	Ke	×
Aerodynamic coefficients scaling	(S_1, S_2)	$\alpha^2+\frac{2.13}{M^2}$, $\frac{M}{(\log_{10}\text{Re}-3.53)^2}$	$\alpha^2+\frac{2.96}{M^2}$, $\frac{M}{(\log_{10}\text{Re})^2}$	$\alpha^2+\frac{1.95}{M^2}$, $\frac{\log_{10}\text{Re}}{M}$	×	×
	S_A	$S_A = S_1(S_2^2+4.00)$	$S_A = S_1(S_2^2+0.05)$	×	×	×

APPENDIX A: ROBUSTNESS ANALYSIS

Robustness analysis is crucial for comprehensively assessing the applicability boundaries and limitations of the method. To thoroughly evaluate the stability and generalizability of the proposed framework under various perturbations and data conditions, we conducted a series of robustness tests: Noise sensitivity analysis, variable addition/removal tests, sample size sensitivity analysis, and analysis of the influence of different dimensionless quantities on results.

1. Noise sensitivity analysis

To evaluate the robustness of the CoSR framework in noisy environments, we systematically tested the final extraction results for all problems at different noise levels, as shown in Table I. For noise sensitivity analysis, we further added Gaussian noise to the simulation data as follows:

$$y_{\text{noisy}} = y + \epsilon, \quad \epsilon \sim \mathcal{N}(0, (\sigma y)^2), \tag{A1}$$

where σ is the noise scaling factor. We sequentially tested five noise levels: $\sigma = 0.01$, 0.05, 0.1, 0.15, and 0.2.

As shown in Table I, the noise tolerance of CoSR varies across cases. For turbulent Rayleigh-Bénard convection, the method is highly robust: The main Ra-related expression can still be recovered even at $\sigma = 0.2$, although some other subexpressions degrade or fail at the highest noise level. For viscous pipe flow and aerodynamic coefficient scaling, reliable expressions are generally obtained for $\sigma \leqslant 0.05$, while extraction fails at $\sigma \geqslant 0.15$ in most cases. For laser-metal interaction, key dimensionless quantities such as Ke remain identifiable at $\sigma = 0.1$–0.15, but the S_A expression fails at $\sigma = 0.2$. In contrast, the universal gravitation case is the most noise sensitive: Correct expressions are only recovered at $\sigma = 0.01$ and 0.05, and extraction fails when $\sigma > 0.05$. This high sensitivity is mainly attributed to the limited data volume—only 24 planetary records—and the requirement for high-precision astronomical observations. Overall, these results demonstrate that CoSR is generally robust to noise, with its tolerance depending on data quantity, problem complexity, and the precision required for the target law.

2. Variable addition/removal tests

In the modeling of complex systems, limitations in prior knowledge or experience often mean that the selected variable set is not perfectly appropriate; it may include redundant variables or omit crucial ones. To address these two issues, systematic testing and analysis strategies are required.

When crucial variables are omitted, the loss of symbolic regression typically increases significantly, sometimes by orders of magnitude, as shown in Table II. For most cases, the resulting errors exceed the typical 5% mean relative error (MRE) threshold allowed in engineering applications, or the loss increases by at least one order of magnitude relative to the full-variable case. This indicates that missing variables severely affect model accuracy and practicality. For the viscous circular pipe flow case, removing U or ν makes dimensional analysis impossible; therefore, the corresponding loss and MRE cannot be computed and are marked as N/A.

When redundant variables are present, some variables may be essentially irrelevant to the system output. For each case, we added one synthetic redundant variable that was independently sampled from a uniform distribution, $x_{\text{irr}} \sim \mathcal{U}(0, 1)$, and was statistically independent of both the physical input variables and the target output. In such cases, global sensitivity analysis methods, such as the Sobol method, can be used to quantitatively assess variable importance and thereby screen the key variable set before modeling. The surrogate model required by the Sobol analysis is constructed using symbolic regression. As shown in Fig. 8, this approach effec-

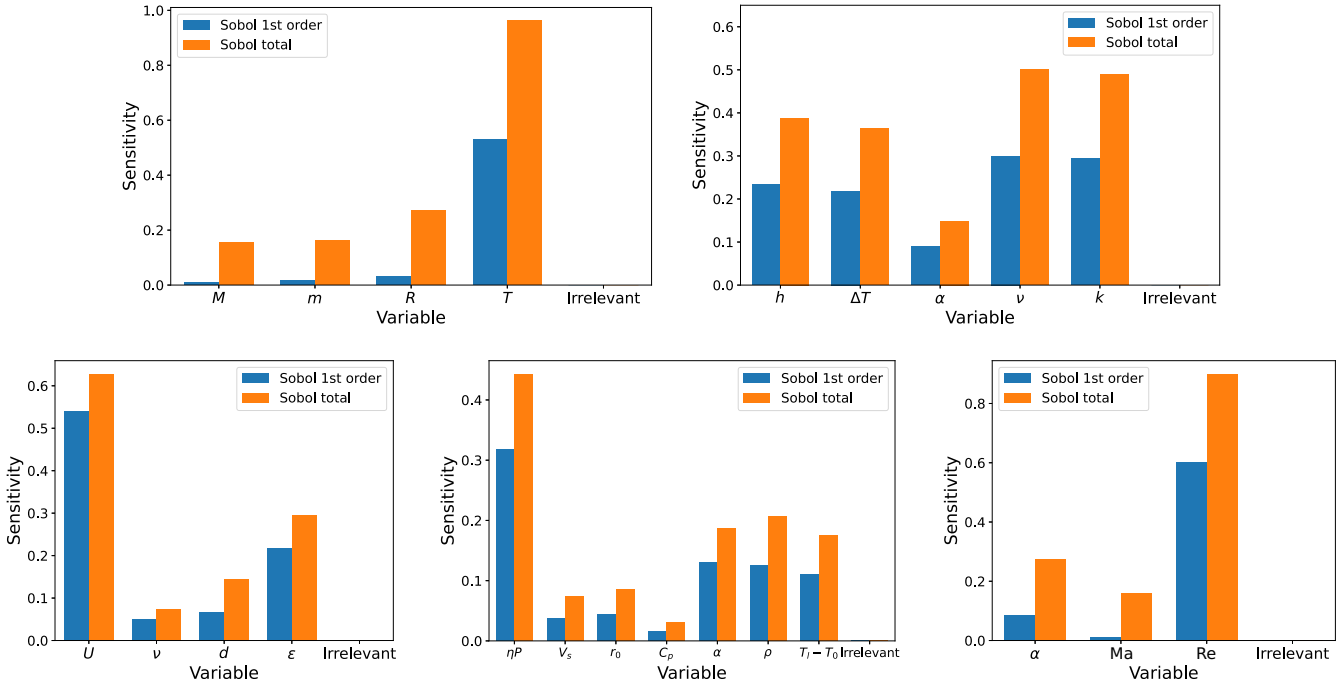


FIG. 8. Sobol sensitivity analysis after adding redundant variables. From top to bottom and left to right: the law of universal gravitation, turbulent Rayleigh-Bénard convection, viscous flows in a circular pipe, laser-metal interaction, and aerodynamic coefficients scaling.

tively identifies and eliminates irrelevant variables, improving model parsimony and generalization capability.

3. Sample size sensitivity analysis

To evaluate the sensitivity of the CoSR framework to sample size, we conducted a systematic analysis across all five cases. For each case, we performed Latin hypercube sampling from the full dataset at ratios of 0.8, 0.6, 0.4, and 0.2, and repeated the experiments 10 times with different random seeds. Since each case involves multiple discovery targets, a run was considered successful only if all target expressions were correctly recovered; failure of any single target was counted as a failure. As shown in Fig. 9, the accuracy rate generally decreases as the data size decreases, but the sensitivity varies considerably among cases. Case 4 is the most robust. Cases 2 and 3 are also relatively robust at 0.8, and remain above or near 50% at 0.6, but degrade rapidly at smaller data sizes. For Case 5, the accuracy rate drops from 70% at 0.8 to 50% at 0.6 and 20% at 0.4, and no full success is observed at 0.2. Notably, for Case 5 (aerodynamic coefficient scaling), even at data sizes of 0.2 and 0.4, the leading S_1 term can still be discovered, but the viscous correction term can no longer be identified, indicating a loss of information required to recover the complete expression. Case 1 is the most sample sensitive: Its accuracy rate falls from 80% at 0.8 to 20% at 0.6, and nearly 0% at 0.4 and 0.2. This is mainly because the universal gravitation case contains only 24 planetary data points; when the number of samples falls below about 20, the governing gravitational law can no longer be reliably discovered. Overall, most cases retain an accuracy rate of at least 50% at a data size of 0.6, demonstrating that CoSR can still extract meaningful expressions under moderate data reduction. These results indicate that the amount of data required by CoSR depends on both the complexity of the target law and the number of available samples, and that sufficient data are particularly important for recovering higher-order correction terms and for cases with inherently limited observations.

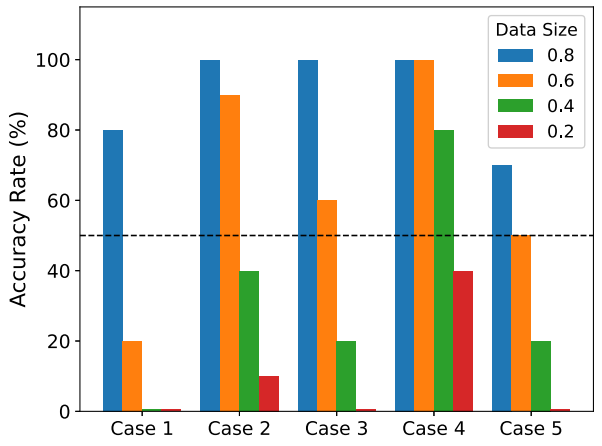


FIG. 9. Accuracy rate of discovery for all cases under different data sizes. The dashed line indicates the 50% success level.

4. Analysis of the influence of different dimensionless quantities on results

In the discovery framework, although dimensional analysis (Buckingham π theorem) can determine the number of independent dimensionless quantities, it cannot uniquely specify their exact mathematical forms. For a system involving n physical quantities and k fundamental dimensions, there exist, in theory, infinitely many equivalent sets consisting of $p = n - k$ independent dimensionless numbers. We investigated the impact of different choices of dimensionless terms

TABLE II. Variable removal tests. "Full" denotes the full variable set; "$-X$" denotes removal of variable X.

Case	Input	Loss	MRE
Law of universal gravitation	Full	1.49×10^{-9}	0.12%
	$-M$	7.71×10^{-5}	28.19%
	$-m$	7.05×10^{-5}	26.96%
	$-T$	2.40×10^{-4}	49.74%
	$-R$	8.88×10^{-5}	30.26%
Turbulent Rayleigh-Bénard convection	Full	0.0027	2.13%
	$-h$	0.0880	12.07%
	$-\Delta T$	0.2224	19.19%
	$-g$	0.0879	12.06%
	$-\alpha$	0.2224	12.07%
	$-\nu$	0.0112	4.31%
	$-\kappa$	0.0497	9.07%
Viscous flows in a circular pipe	Full	0.0013	2.44%
	$-U$	N/A	
	$-\nu$		
	$-d$	0.0177	9.00%
	$-\varepsilon$	0.0247	10.63%
Laser-metal interaction	Full	0.1351	0.86%
	$-\eta P$	3.9905	44.51%
	$-V_s$	0.4899	15.60%
	$-r_0$	3.6843	42.77%
	$-C_p$	2.6485	36.26%
	$-\alpha$	1.1796	24.20%
	$-\rho$	4.6277	47.93%
	$-(T_1 - T_0)$	2.6485	36.26%
Aerodynamic coefficients scaling	Full	3.98×10^{-5}	2.97%
	$-\alpha$	1.30×10^{-3}	16.99%
	$-M$	4.93×10^{-3}	33.09%
	$-\mathrm{Re}$	4.44×10^{-4}	9.93%

on the final results; these terms can take various forms, including complex power-law product forms, additive combinations, and other related formulations. Table III summarizes the results obtained using different dimensionless terms. For the universal gravitation case, we used dimensional consistency constraints rather than dimensionality reduction, so the choice of dimensionless basis functions is not involved. Moreover, for the turbulent Rayleigh-Bénard convection and viscous pipe flow cases, the correction terms can also be discovered under different dimensionless bases. For the other cases, adopting different dimensionless basis functions changes the data distribution and thus the discovery difficulty. These results demonstrate that the CoSR framework is robust to the choice of dimensionless representation, and that the physically meaningful dimensionless groups can be consistently identified even when the input representation is changed.

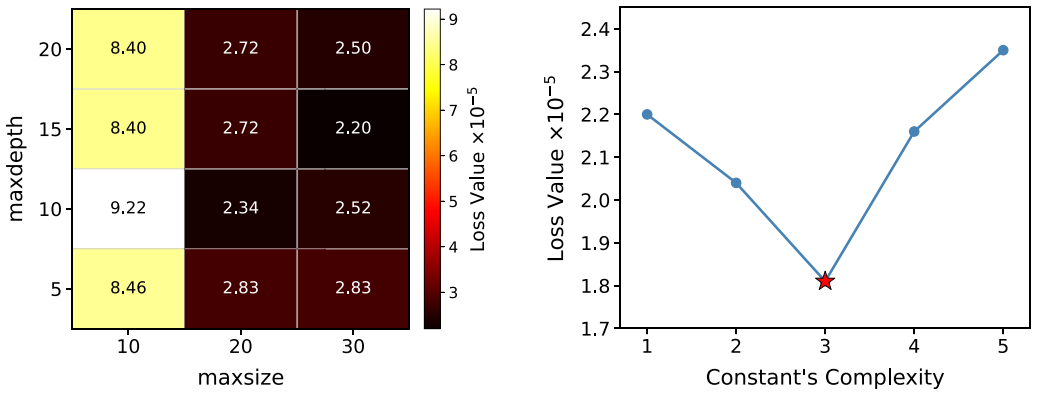


FIG. 10. Complexity hyperparameter sweep results for the aerodynamic case. (a) Two-dimensional heatmap of `maxsize` vs `maxdepth`; darker colors indicate lower loss. (b) Relationship between `constant_complexity` and loss; the arrow points to the minimum at a value of 3.

TABLE III. Results obtained using different dimensionless terms for each case.

Case	Dimensionless quantities	Result
Turbulent Rayleigh-Bénard convection	$\pi_1 = \frac{\alpha^2(\Delta T)^2\nu^2}{gh^3}$, $\pi_2 = \frac{gh^3}{\kappa^2}$, $\pi_3 = \frac{gh^3\kappa}{\alpha\Delta T\nu^3}$	$s_1 = \mathrm{Ra}$, $s_2 = \frac{\alpha^2(\Delta T)^2\nu^2}{gh^3}$, $S_A = s_1$, $S_{\mathrm{lc}} = \pi_1^{3/4}\pi_2^{1/4}\pi_3^{1/2}$, $S_{\mathrm{rc}} = \pi_1^{9/4}\pi_2^{3/4}\pi_3^{3/2}$
	$\pi_1 = \frac{\alpha\Delta T\nu^2}{gh^3}$, $\pi_2 = \frac{gh^3}{\nu^2}$, $\pi_3 = \alpha\Delta T + \frac{gh^3}{\nu\kappa}$	$S_A = (\pi_3 - \pi_1\pi_2)\pi_1\pi_2 = \mathrm{Ra}$, $S_{\mathrm{lc}} = \pi_1\pi_2$, $S_{\mathrm{rc}} = (\pi_1\pi_2)^3$
Viscous flows in a circular pipe	$\pi_1 = \frac{U^2\varepsilon d}{\nu^2}$, $\pi_2 = \frac{U\varepsilon}{\nu}$	$s_1 = \mathrm{Re}$, $s_2 = \frac{\varepsilon}{d}$, $S_{\mathrm{lc}} = s_1^{1/4}$, $S_{\mathrm{rc}} = s_1^{3/4}s_2$
	$\pi_1 = \frac{U^2 d(\varepsilon+d)}{\nu^2}$, $\pi_2 = \frac{U^2\varepsilon d}{\nu^2}$	$s_1 = \sqrt{\pi_1 - \pi_2} = \mathrm{Re}$, $s_2 = \frac{\varepsilon}{d}$, $S_{\mathrm{lc}} = s_1^{1/4}$, $S_{\mathrm{rc}} = s_1^{3/4}s_2$
Laser-metal interaction	$\pi_1 = \frac{\eta P}{V_s r_0^2 \rho C_p (T_1 - T_0)}$, $\pi_2 = \sqrt{\frac{\alpha}{V_s r_0}}$, $\pi_3 = \frac{r_0^2\rho C_p^{1.5}(T_1-T_0)^{1.5}}{\eta P}$	$s_1 = \mathrm{Ke}$, $s_2 = X$, $S_A = \frac{s_1}{1+s_1 s_2}$
	$\pi_1 = \frac{C_p^{0.5}(T_1-T_0)^{0.5}}{V_s}$, $\pi_2 = \frac{\sqrt{\alpha}\eta P}{V_s^{3.5} r_0^{2.5}\rho}$, $\pi_3 = \frac{\sqrt{\alpha} V_s^{2.5} r_0^{1.5}\rho}{\eta P}$	$s_1 = \mathrm{Ke}$, $s_2 = X$, $S_A = \frac{s_1}{1+s_1 s_2}$
Aerodynamic coefficients scaling	$\pi_1 = \alpha^2 - M$, $\pi_2 = M$, $\pi_3 = \log_{10}\mathrm{Re}$	$s_1 = \pi_1 + \pi_2 + \frac{2.14}{\pi_2^2}$, $s_2 = \frac{\pi_2}{\pi_3^2}$, $S_A = s_1(s_2^2 + 0.079)$
	$\pi_1 = \frac{\alpha}{M}$, $\pi_2 = \alpha M$, $\pi_3 = \log_{10}\mathrm{Re}$	$s_1 = \pi_1\pi_2$, $s_2 = \frac{\pi_1}{\pi_2}$, $s_3 = \frac{1}{\pi_3^2}$, $S_{A1} = s_1 + 2.24 s_2$, $S_{A2} = \frac{s_3^2}{s_2}$, $S_A = S_{A1}(S_{A2}^2 + 0.075)$

APPENDIX B: PYSR HYPERPARAMETER SELECTION GUIDE

To achieve data-driven knowledge discovery for complex physical systems, this study selected the open-source symbolic regression library PySR as the core computational engine. To balance exploration capability with model parsimony, we suggested easy-to-follow guidelines for key parameters such as operator sets, complexity, and penalty terms (see Table IV), aiming to achieve an optimal balance between the breadth of the search space, the depth of model complexity, and physical interpretability. The configuration of the loss function is detailed in Sec. 2.2 of the Supplemental Material [28].

The binary operators default to include +, −, *, and / to cover fundamental mathematical relationships, while the retention of the power operator ^ is crucial for specifically discovering power-product scaling relationships when

TABLE IV. PYSR code API configuration guide.

	Dimension	API configuration
Operator set	Unary operators	`sqrt, square, cube, exp, log10, sin, cos`
	Binary operators	`+, −, *, /, ^`
Complexity	Max complexity and tree depth	Exploration: `maxsize = 30, maxdepth = 15`; conciseness: `maxsize = 20, maxdepth = 10`
	Constant complexity	3–5
Penalty terms	Symbol nesting penalty	`''sqrt'': {''square'': 0}, ''cube'': {''cube'': 0}`
Other configurations	Early stopping condition	`stop_if(loss, complexity) = loss < 1e−10 && complexity < 20`
	Model selection	Automatic: best; manual: backtracking output list

exploring more complex non-power-product relationships. The unary operators focus on `sqrt`, `square`, and `cube`, as they directly correspond to common geometric relationships and integer-order scaling laws in physical systems; extended functions such as `exp`, $\texttt{log}_{10}$, `sin`, and `cos` are used to capture specific physical mechanisms (e.g., exponential growth and periodicity). Complexity control adopts a dual-mode strategy: An unrestricted or higher upper limit (`maxsize` = 30, `maxdepth` = 15) is used for fully exploring possible multilayer structures or complex relationships; stricter limits (`maxsize` = 20, `maxdepth` = 10) are used to prioritize conciseness and interpretability of expressions. The weight of constant complexity is kept at a relatively low level (3–5) to prioritize the discovery of universal laws dominated by variable relationships, avoiding models that lose physical clarity due to over-reliance on fitted constants. Symbol nesting penalties (e.g., prohibiting unnecessary nesting of `sqrt` and `square`) are used to suppress mathematically redundant or high-order combinations, guiding the algorithm to output more concise forms. Model selection is based not only on optimal screening along the Pareto front but is also supplemented by manual backtracking through the list of candidate expressions to uncover promising models that may contain novel physical insights (on exploratory test cases). Additionally, an adaptive early stopping condition terminates the search early when the algorithm has found a sufficiently accurate and concise solution, thereby intelligently conserving computational resources. This comprehensive set of configurations collectively forms a robust search framework that balances exploration capability, physical interpretability, and computational efficiency.

Apart from the symbol library (e.g., whether to include trigonometric or exponential functions) and nesting rules, which can be manually preset based on physical priors (such as flow periodicity or boundary-layer decay characteristics), the three core hyperparameters related to complexity—`maxsize`, `maxdepth`, and constant complexity—are difficult to determine accurately from prior knowledge alone. These three directly control the structural capacity of the expression. If their values are too low, underfitting occurs (failure to capture the true physical laws); if too high, overfitting is likely (introduction of spurious numerical terms). Therefore, this study adopts a grid search strategy, using the aerodynamic case as a representative example, to quantitatively evaluate the influence of each parameter on the fitting loss and thereby determine the default configuration. Taking the aerodynamic case as an example, as shown in Fig. 10(a), the joint sweep shows that the low-loss region is concentrated at `maxsize` = 20–30 and `maxdepth` = 10–15, with the optimal combination being (30, 15), corresponding to a loss of 2.20×10^{-5}. Fixing this combination, we further swept constant's complexity [Fig. 10(b)]. The loss exhibits a U-shaped trend and reaches its minimum at a value of 3 (1.81×10^{-5}), indicating that this value best balances the expressive capacity of the constant term and generalization performance. Accordingly, the default configuration for this case is determined as `maxsize` = 30, `maxdepth` = 15, and `constant_complexity` = 3.

[1] H. Wang *et al.*, Scientific discovery in the age of artificial intelligence, Nature (London) **620**, 47 (2023).

[2] M. Schmidt and H. Lipson, Distilling free-form natural laws from experimental data, Science **324**, 81 (2009).

[3] R. Vinuesa *et al.*, Decoding complexity through machine learning is redefining scientific discovery, Commun. Phys. **9**, 168 (2026).

[4] S. L. Brunton, B. R. Noack, and P. Koumoutsakos, Machine learning for fluid mechanics, Annu. Rev. Fluid Mech. **52**, 477 (2020).

[5] P. Langley, Data-driven discovery of physical laws, Cognit. Sci. **5**, 31 (1981).

[6] D. Angelis, F. Sofos, and T. E. Karakasidis, Artificial intelligence in physical sciences: Symbolic regression trends and perspectives, Arch. Computat. Methods Eng. **30**, 3845 (2023).

[7] J. Guo and W.-J. Yin, Harnessing data using symbolic regression methods for discovering novel paradigms in physics, Sci. China Phys. Mech. Astron. **67**, 267301 (2024).

[8] S. L. Brunton, J. L. Proctor, and J. N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proc. Natl. Acad. Sci. USA **113**, 3932 (2016).

[9] Z. Chen, Y. Liu, and H. Sun, Physics-informed learning of governing equations from scarce data, Nat. Commun. **12**, 6136 (2021).

[10] M. Cranmer, Interpretable machine learning for science with PySR and SymbolicRegression.jl, arXiv:2305.01582.

[11] S.-M. Udrescu and M. Tegmark, AI Feynman: A physics-inspired method for symbolic regression, Sci. Adv. **6**, eaay2631 (2020).

[12] S. Forrest, Genetic algorithms: Principles of natural selection applied to computation, Science **261**, 872 (1993).

[13] R. Guimerà, I. Reichardt, A. Aguilar-Mogas, F. A. Massucci, M. Miranda, J. Pallarès, and M. Sales-Pardo, A Bayesian machine scientist to aid in the solution of challenging scientific problems, Sci. Adv. **6**, eaav6971 (2020).

[14] W. Zhang, M. Xia, and J. Kou, A scientometric investigation of artificial intelligence for fluid mechanics: Emerging topics and active groups, Prog. Aerosp. Sci. **157**, 101130 (2025).

[15] Z. Yang, X. Shan, X. I. Yang, and W. Zhang, Data-enabled discovery of specific and generalisable turbulence closures, J. Fluid Mech. **1016**, R1 (2025).

[16] B. M. De Silva, D. M. Higdon, S. L. Brunton, and J. N. Kutz, Discovery of physics from data: Universal laws and discrepancies, Front. Artif. Intell. **3**, 25 (2020).

[17] Y. Yuan and A. Lozano-Durán, Dimensionless learning based on information, Nat. Commun. **16**, 9171 (2025).

[18] J.-F. Arnoldi, A. L. Jackson, I. Peralta-Maraver, and N. L. Payne, A universal thermal performance curve arises in biology and ecology, Proc. Natl. Acad. Sci. USA **122**, e2513099122 (2025).

[19] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature (London) **521**, 436 (2015).

[20] R. Yu and R. Wang, Learning dynamical systems from data: An introduction to physics-guided deep learning, Proc. Natl. Acad. Sci. USA **121**, e2311808121 (2024).

[21] S. Kim, P. Y. Lu, S. Mukherjee, M. Gilbert, L. Jing, and V. Čeperić, Integration of neural network-based symbolic regression in deep learning for scientific discovery, IEEE Trans. Neural Netw. Learn. Syst. **32**, 4166 (2020).

[22] E. Buckingham, On physically similar systems; Illustrations of the use of dimensional equations, Phys. Rev. **4**, 345 (1914).

[23] G. I. Barenblatt, *Scaling, Self-Similarity, and Intermediate Asymptotics: Dimensional Analysis and Intermediate Asymptotics* (Cambridge University Press, Cambridge, 1996), Vol. 14.

[24] J. Bakarji, J. Callaham, S. L. Brunton, and J. N. Kutz, Dimensionally consistent learning with Buckingham Pi, Nat. Comput. Sci. **2**, 834 (2022).

[25] L. Zhang and G. He, Mutual-information-based dimensional learning: Objective algorithms for identification of relevant dimensionless quantities, Comput. Methods Appl. Mech. Eng. **440**, 117922 (2025).

[26] X. Xie, A. Samaei, J. Guo, W. K. Liu, and Z. Gan, Data-driven discovery of dimensionless numbers and governing laws from scarce measurements, Nat. Commun. **13**, 7562 (2022).

[27] M. Xia, H. Lin, and W. Zhang, Hierarchical dimensionless learning: A physics-data hybrid-driven approach for discovering dimensionless parameter combinations, Eng. Appl. Artif. Intell. **170**, 114190 (2026).

[28] See Supplemental Material at http://link.aps.org/supplemental/10.1103/f2kp-pp4q for a summary table of case studies, methodological details and pseudocode for three types of symbolic regression discovery methods, and supplementary notes for five cases.

[29] L. Foppa, T. A. Purcell, S. V. Levchenko, M. Scheffler, and L. M. Ghiringhelli, Hierarchical symbolic regression for identifying key physical parameters correlated with bulk properties of perovskites, Phys. Rev. Lett. **129**, 055301 (2022).

[30] I. Icke and J. C. Bongard, Modeling hierarchy using symbolic regression, in *2013 IEEE Congress on Evolutionary Computation* (IEEE, Piscataway, NJ, 2013), pp. 2980–2987.

[31] P. Y. Lu, R. Dangovski, and M. Soljačić, Discovering conservation laws using optimal transport and manifold learning, Nat. Commun. **14**, 4744 (2023).

[32] K. Kaheman, J. N. Kutz, and S. L. Brunton, SINDy-PI: A robust algorithm for parallel implicit sparse identification of nonlinear dynamics, Proc. R. Soc. A **476**, 20200279 (2020).

[33] M. Schmidt and H. Lipson, Symbolic regression of implicit equations, in *Genetic Programming Theory and Practice VII*, edited by R. Riolo, UM. O'Reilly, and T. McConaghy (Springer, Boston, MA, 2009), pp. 73–85.

[34] J. Zhong, J. Yang, Y. Chen, W.-L. Liu, and L. Feng, Mining implicit equations from data using gene expression programming, IEEE Trans. Emerg. Topics Comput. **10**, 1058 (2021).

[35] K. Champion, B. Lusch, J. N. Kutz, and S. L. Brunton, Data-driven discovery of coordinates and governing equations, Proc. Natl. Acad. Sci. USA **116**, 22445 (2019).

[36] B. Lusch, J. N. Kutz, and S. L. Brunton, Deep learning for universal linear embeddings of nonlinear dynamics, Nat. Commun. **9**, 4950 (2018).

[37] D. Wadekar *et al.*, Augmenting astrophysical scaling relations with machine learning: Application to reducing the Sunyaev–Zeldovich flux–mass scatter, Proc. Natl. Acad. Sci. USA **120**, e2202074120 (2023).

[38] NASA, Planets factsheet, https://nssdc.gsfc.nasa.gov/planetary/factsheet/, accessed: 2017.

[39] NASA, Exoplanet archive, https://exoplanetarchive.ipac.caltech.edu/, accessed: 2017.

[40] B. Novakovic, Orbits of five visual binary stars, arXiv:0712.4242.

[41] J. Niemela and K. R. Sreenivasan, Turbulent convection at high Rayleigh numbers and aspect ratio 4, J. Fluid Mech. **557**, 411 (2006).

[42] X. Chavanne, F. Chilla, B. Chabaud, B. Castaing, and B. Hebral, Turbulent Rayleigh–Bénard convection in gaseous and liquid He, Phys. Fluids **13**, 1300 (2001).

[43] N. Goldenfeld, Roughness-induced critical phenomena in a turbulent flow, Phys. Rev. Lett. **96**, 044503 (2006).

[44] J. Tao, Critical instability and friction scaling of fluid flows through pipes with rough inner surfaces, Phys. Rev. Lett. **103**, 264502 (2009).

[45] Z.-S. She, Y. Wu, X. Chen, and F. Hussain, A multi-state description of roughness effects in turbulent pipe flow, New J. Phys. **14**, 093054 (2012).

[46] O. Reynolds, XXIX. An experimental investigation of the circumstances which determine whether the motion of water shall be direct or sinuous, and of the law of resistance in parallel channels, Philos. Trans. R. Soc. London **174**, 935 (1883).

[47] J. Jiménez, Turbulent flows over rough walls, Annu. Rev. Fluid Mech. **36**, 173 (2004).

[48] W.-J. Wang, S. Li, W. Huang, Z. Han, and W.-H. Wang, A unified friction factor formulation: Bridging laminar and turbulent friction factor with critical points analysis, Phys. Fluids **36**, 014116 (2024).

[49] J. Nikuradse, Laws of flow in rough pipes, Technical Report, National Advisory Committee for Aeronautics, Washington, DC, 1950.

[50] C. Zhao, Q. Guo, X. Li, N. Parab, K. Fezzaa, W. Tan, L. Chen, and T. Sun, Bulk-explosion-induced metal spattering during laser processing, Phys. Rev. X **9**, 021052 (2019).

[51] Z. Gan, O. L. Kafka, N. Parab, C. Zhao, L. Fang, O. Heinonen, T. Sun, and W. K. Liu, Universal scaling laws of keyhole stability and porosity in 3D printing of metals, Nat. Commun. **12**, 2379 (2021).

[52] J. Ye, S. A. Khairallah, A. M. Rubenchik, M. F. Crumb, G. Guss, J. Belak, and M. J. Matthews, Energy coupling mechanisms and scaling behavior associated with laser powder bed fusion additive manufacturing, Adv. Eng. Mater. **21**, 1900185 (2019).

[53] H. Lin, X. Wang, and W. Zhang, Scaling function learning: A sparse aerodynamic data reconstruction method for various aircraft shapes, Chin. J. Aeronaut. **39**, 103797 (2025).

[54] L. Breiman, Statistical modeling: The two cultures (with comments and a rejoinder by the author), Stat. Sci. **16**, 199 (2001).

[55] https://github.com/Alexia369/Chain-of-Symbolic-Regression.