Nexus
Research article

Worth of prior knowledge for enhancing deep learning

Hao Xu, Yuntian Chen, and Dongxiao Zhang
*Correspondence: ychen@eitech.edu.cn (Y.C.); dzhang@eitech.edu.cn (D.Z.)

IN BRIEF

In deep learning, prior knowledge is essential for mitigating shortcomings of data-driven models. However, there still exist several challenges, including the evaluation of the worth of prior knowledge, the multiobjective optimization of data and rule loss, and the maximization of the effect of knowledge. In this work, we present a framework to enable efficient evaluation of the worth of knowledge quantitatively by the derived rule importance, which deepens the understanding of the nexus between data and knowledge. It is discovered that there exist sophisticated relationships between data and rules, including dependence, synergistic, and substitution effects. Meanwhile, the worth of prior knowledge differs in the in-distribution and out-of-distribution scenarios. The proposed framework can be applied to improve the performance of informed machine learning, as well as to distinguish improper prior knowledge. Experiments have proven that our framework can shed light on diverse fields encompassing physics, chemistry, geoscience, and engineering.

Nexus
Research article

Worth of prior knowledge for enhancing deep learning

Hao Xu,[1] Yuntian Chen,[2,*] and Dongxiao Zhang[2,3,*]
[1]BIC-ESAT, ERE, and SKLTCS, College of Engineering, Peking University, Beijing 100871, P.R. China
[2]Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, Zhejiang 315200, P.R. China
[3]Lead contact
*Correspondence: ychen@eitech.edu.cn (Y.C.); dzhang@eitech.edu.cn (D.Z.)
https://doi.org/10.1016/j.ynexs.2024.100003

BROADER CONTEXT

Knowledge can be viewed as the intricate relationship between data in both temporal and spatial dimensions in data science. While there are occasional instances where such relationships can be expressed through concise mathematical formulations, in most cases, knowledge is represented as a collection of correlations among data points. In deep learning, prior knowledge is essential for mitigating shortcomings of data-driven models, such as data dependence, generalization ability, and compliance with constraints. However, there still remain several challenges: (1) how to evaluate the worth of prior knowledge, (2) what the relationship is between data and rules, and (3) how to make prior rules work better. This work presents a framework to enable efficient evaluation of the worth of knowledge quantitatively by the derived rule importance, which deepens the understanding of the nexus between data and knowledge. It is significant for improving the performance of informed machine learning.

ABSTRACT

Knowledge constitutes the accumulated understanding and experience that humans use to gain insight into the world. In deep learning, prior knowledge is essential for mitigating shortcomings of data-driven models, such as data dependence, generalization ability, and compliance with constraints. Here, we present a framework to enable efficient evaluation of the worth of knowledge by the derived rule importance. Through quantitative experiments, we assess the influence of data volume and estimation range on the worth of knowledge. Our findings elucidate the complex relationship between data and knowledge, including dependence, synergistic, and substitution effects. Our model-agnostic framework can be applied to a variety of common network architectures, providing a comprehensive understanding of the role of prior knowledge in deep learning models. It also offers practical utility for knowledge identification and model construction within interdisciplinary research by improving the performance of informed machine learning and distinguishing improper prior knowledge.

INTRODUCTION

The emergence of deep learning techniques has revolutionized the field of scientific research, resulting in a series of remarkable achievements.[1–3] Deep learning excels at uncovering potential relationships in high-dimensional space from abundant available data. However, data-driven models still face certain challenges, such as data dependence,[4] generalization ability,[5] and compliance with constraints.[6] In response to this, informed machine learning has become increasingly popular, enabling prior knowledge to be incorporated into the learning process.[7,8] As illustrated in Figure 1A, various types of knowledge can be integrated into a machine learning model, such as functional relations,[9] logic rules,[10] differential equations,[11] invariance,[12] and algebraic relations,[13] which are usually incorporated through the modification of model structure or loss function. From a data science standpoint, knowledge can be viewed as the intricate relationship between data in both temporal and spatial dimensions. While there are occasional instances where such relationships can be elegantly expressed through concise mathematical formulations, in the vast majority of cases, knowledge is represented as a collection of correlations among data points. For knowledge to be incorporated into a machine learning model, it needs to be formalized, meaning that it has to be structured in a manner that can be expressed mathematically. Therefore, various knowledge discovery techniques have been proposed to extract formalized knowledge from data.[14] In this case, the formalized knowledge that can be integrated into a machine learning model is referred to as rules.

Informed machine learning has been deployed in a variety of problem domains, such as the solution of partial differential equations (PDEs),[15] quantification of fluid flow,[4] time series prediction,[16] and robot control.[17] It usually functions as a bridge between the knowledge of diversified fields and informatics. Two main approaches in the field of informed machine learning, depending on whether the rules are strictly obeyed, are soft constraint[18] and hard constraint.[19] Despite its promise, the worth of knowledge is currently only vaguely understood, which limits our ability to comprehend the relationship between data and knowledge. Figure 1B provides a clear example of the divergent effects of data and rules in the context of in-distribution and out-of-distribution (i.e., distribution shift) prediction tasks. The data-driven model performs well in in-distribution scenarios but poorly in out-of-distribution tasks. In contrast, the rule-driven model performs better in out-of-distribution tasks. Notably, the informed machine learning model combining data and rules achieves a moderate prediction ability in both in-distribution and out-of-distribution scenarios, which implies that there exists an underlying nexus between data and rules.

To elucidate the importance of data and rules, it is useful to consider the perspective of learning high-dimensional sample distributions (as illustrated in Figure 1C). Pure data-driven learning faces challenges in extrapolating beyond the training data distribution. Conversely, the incorporation of rules possessing more universal features can contribute to an improved generalization ability to a certain extent. Although rules can reduce the optimization space of distributions, they cannot directly yield an accurate distribution unless a unique solution can be obtained from the prior rules. In this way, informed machine learning uses training data to locate the correct sample distribution within the restricted high-dimensional space created by the rules. The combination of data and rules can theoretically result in high efficiency and good inference ability for both in-distribution and out-of-distribution tasks. However, incorporating multiple prior rules can lead to a high risk of model collapse due to intricate internal interactions, making

Published by The Hong Kong Polytechnic University in association with Cell Press, an imprint of Elsevier Inc.

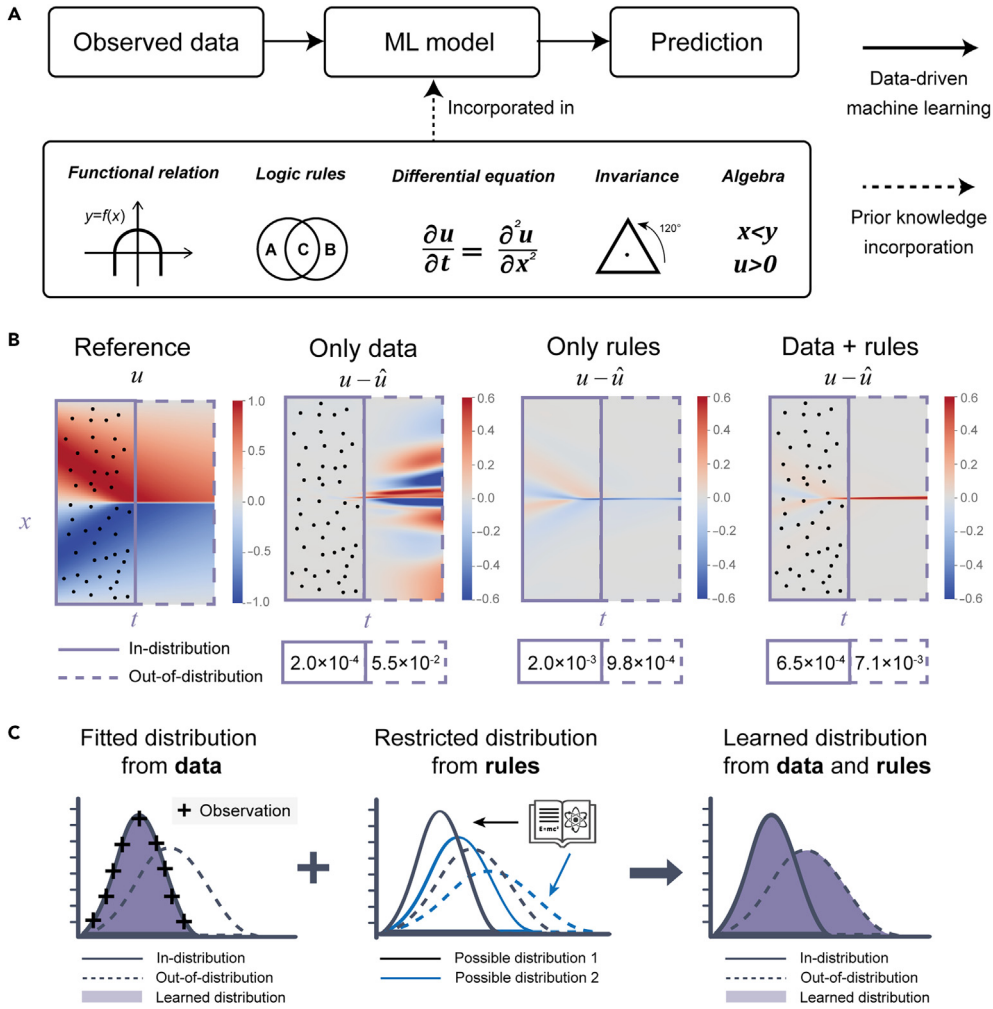

Figure 1. Overview of informed machine learning

(A) The information flow in conventional data-driven machine learning and informed machine learning.

(B) An example of Burgers' shock equation for informed machine learning. The black dots refer to the training data. u and $\hat{u}$ are references and predictions, respectively. The mean squared error of in-distribution and out-of-distribution scenarios is displayed in solid and dotted purple lines, respectively.

(C) The explanation for the function of data and rules in informed machine learning from the aspect of the sample distribution. Here, rules refer to the formalized knowledge that can be incorporated into the machine learning model.

convergence of the training process more difficult. To address this, it is necessary to measure the importance of each integrated rule to guide model construction and maximize the value of knowledge. In this work, our primary focus is on quantitatively measuring the worth of knowledge to uncover the underlying principles of data and rules. Our contributions can be summarized as addressing the following three main questions.

How to evaluate the worth of knowledge?

In this work, we propose a framework for quantitatively measuring the effect of prior rules in informed machine learning. We introduce the concept of rule importance (*RI*) to effectively address the issue of assigning the contribution of integrated rules in informed deep learning.

What is the relationship between data and rules?

Through quantitative experiments, we have evaluated the influence of data volume and estimation range on the worth of knowledge and uncovered the complex relationship between data and knowledge, comprising dependence, synergism, and substitution effects.

How to make prior rules work better

Our measurement of *RI* can be used to facilitate adjustment of the regularization parameter in informed machine learning to prevent non-convergence during the training process and maximize the value of knowledge. It can also be used to identify improper prior rules.

RESULTS

Derivation and calculation process of *RI*

In this work, the *RI* is proposed to measure the worth of knowledge. In essence, the importance of rules is assigned by their respective marginal contribution, where the profit of the rule coalition is defined as the improvement of the model's predictive accuracy. As detailed in the experimental procedures section, the definition *RI* involves the contribution assignment in the condition involving both cooperation and competition. Therefore, we derive a specialized framework to evaluate the *RI*. In the context of this work, the marginal contribution of rule i when incorporating into S is defined to be

$$\delta_i^{RI}(s) = -[\log_{10}(MSE(s)) - \log_{10}(MSE(s\setminus\{i\}))], s \in S_i, \quad \text{(Equation 1)}$$

where $s\setminus\{i\}$ is the new set produced by removing rule i from rule coalition s, and $MSE(s)$ and $MSE(s\setminus\{i\})$ refer to the mean squared error (MSE) of the model trained with rule coalition s and $s\setminus\{i\}$, respectively. Considering that the change of MSE usually strides across orders of

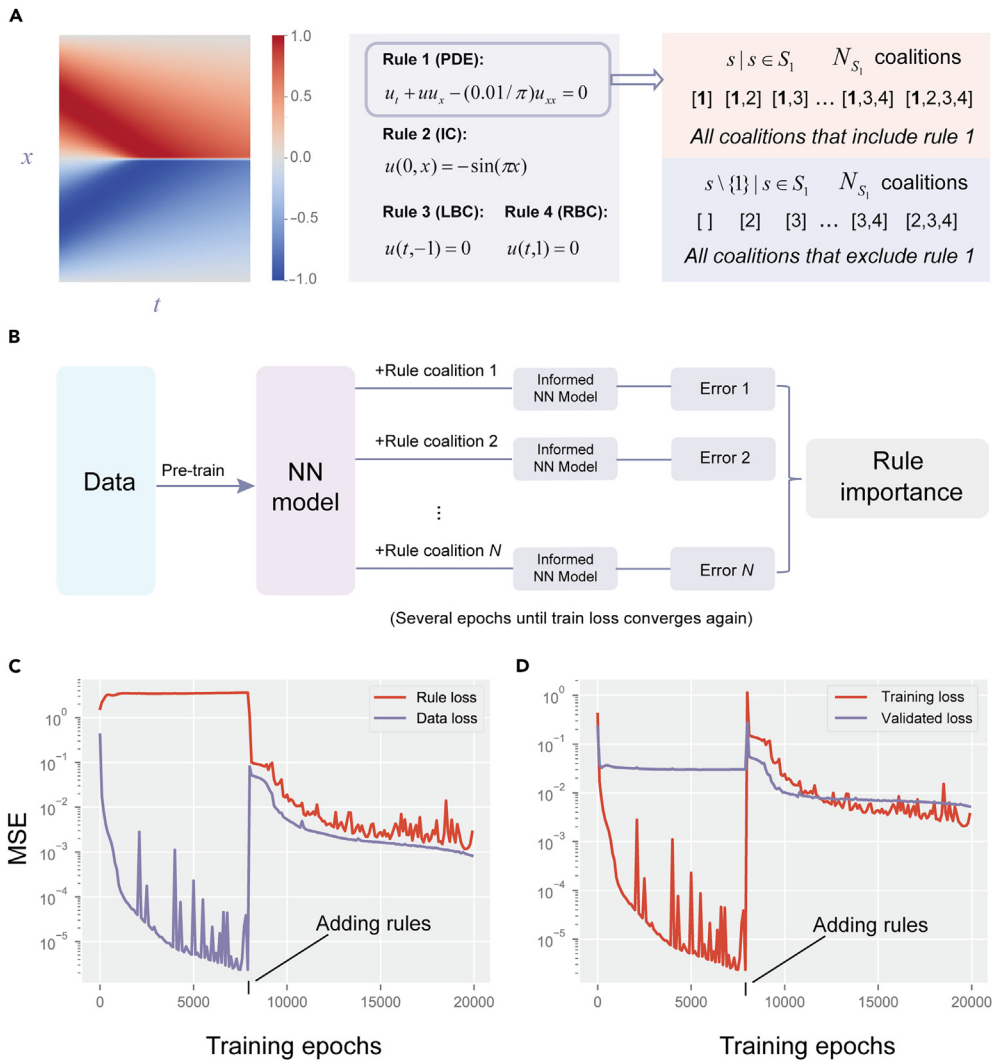

Figure 2. Illustration of the calculation process for rule importance (*RI*)

(A) An example of calculating the importance of the PDE rule for Burgers' shock equation. Here, PDE refers to the partial differential equation, LBC refers to the left boundary condition, RBC refers to the right boundary condition, and IC refers to the initial condition.

(B) The flow chart for calculating the *RI*.

(C) The rule loss and data loss on the training data. The rules are added to the network at 8,000 epochs where convergence has been achieved in the pre-train stage.

(D) The training loss and validating loss during the training process. The training loss only contains the data loss in the pre-train stage and contains both data loss and rule loss after the rules are incorporated into the network. The validated loss is the mean squared error between the predicted and observed validating data.

magnitude, the logarithm of MSE is adapted to measure the contribution. Equation 1 can be simplified into

$$\delta_i^{RI}(s) = \log_{10}\left(\frac{MSE(s\ \{i\})}{MSE(s)}\right), s \in S_i. \quad \text{(Equation 2)}$$

Notably, different from the Shapley value,[20,21] our framework permits the existence of negative contributions due to the inclusion of improper rules or their coalition, which may actually lead to a decrease in overall predictive accuracy. Then, the *RI* is calculated by the average of the defined marginal contribution, which can be written as

$$RI_i = \frac{1}{N_{S_i}} \sum_{s \in S_i} \delta_i^{RI}(s), \quad \text{(Equation 3)}$$

where RI_i is the importance of rule i and N_{S_i} is the size of S_i. For each rule i, there are two possible situations, namely presence or absence, which can be represented by the binary code 1 or 0, respectively. Therefore, for n considered rules, the possible coalition can be represented by a binary sequence with the length of n. Consequently, the N_{S_i} is kept the same as 2^{n-1}, where n is the number of considered rules. Finally, the formula for calculating the *RI* can be obtained as follows:

$$RI_i = \frac{1}{2^{n-1}} \sum_{s \in S_i} \log_{10}\left(\frac{MSE(s\ \{i\})}{MSE(s)}\right). \quad \text{(Equation 4)}$$

For comparison, we also define the full importance (*FI*) in this work, which refers to the importance of a given rule when all the other rules exist. The formula for *FI* can be expressed as

$$FI_i = \log_{10}\left(\frac{MSE(s_f\ \{i\})}{MSE(s_f)}\right), \quad \text{(Equation 5)}$$

where FI_i is the *FI* of rule i and s_f is the rule coalition that includes all rules.

The calculation process of *RI* is visualized in Figure 2A, where four rules are incorporated as an example. For the PDE rule (rule 1), all coalitions that include and exclude this rule are considered. For each

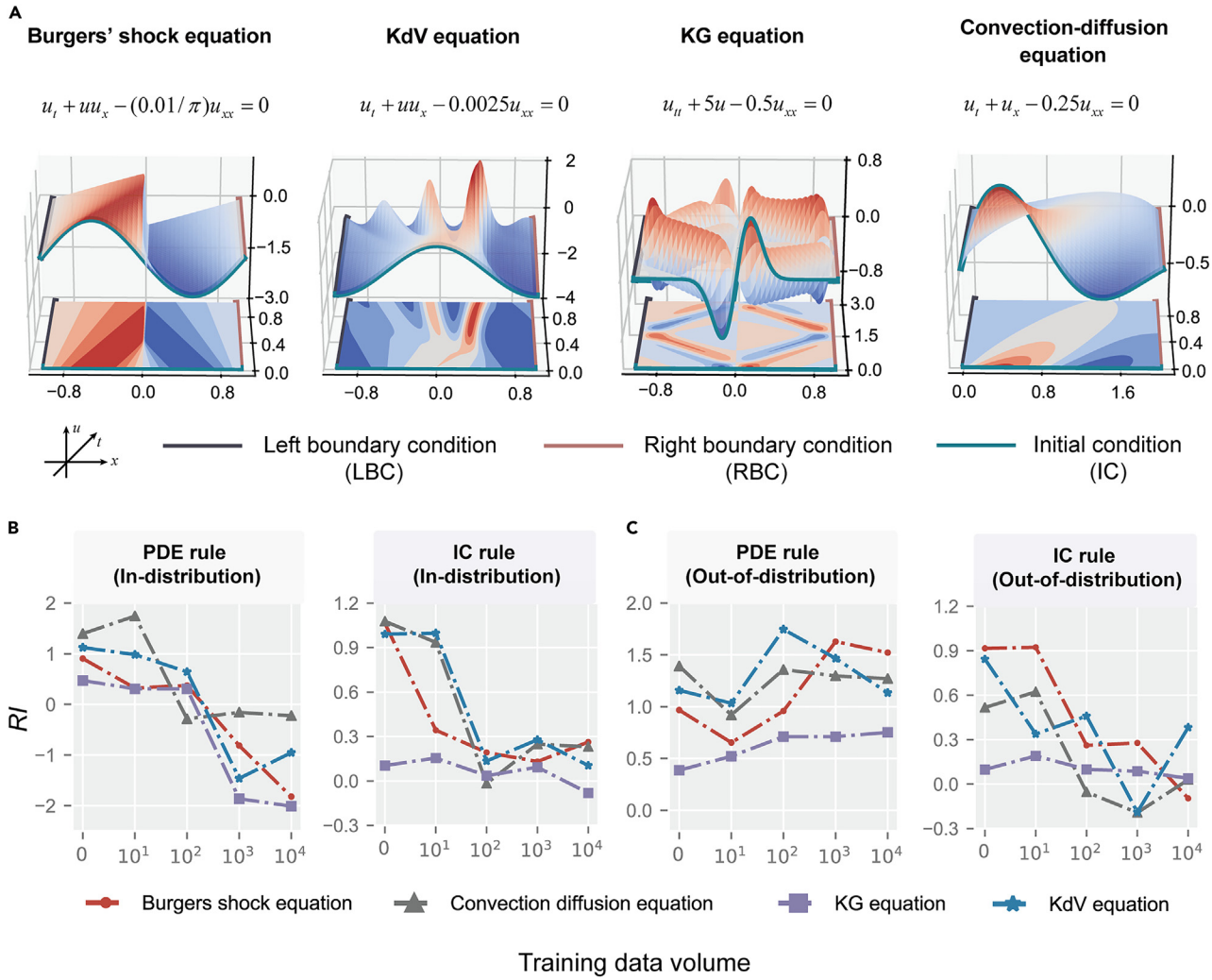

Figure 3. Numerical experiments to reveal the relationship between data and rules in the context of in-distribution and out-of-distribution scenarios

(A) The information of the four canonical physical processes that are guided by governing equations. The involved rules include PDEs, ICs, and BCs.

(B) The importance of PDE rules (global rules) and IC rules (local rules) varies with the data volume in the case of in-distribution in the four physical processes.

(C) The importance of PDE rules (global rules) and IC rules (local rules) varies with the data volume in the case of out-of-distribution in the four physical processes. The *RI* values are calculated using the framework established in this study.

possible coalition, a machine learning model is trained to obtain the respective *MSE*(*s*) and *MSE*(*s*\{*i* = 1}). In brief, the influence of the rule in all relevant rule coalitions is employed to calculate the marginal contribution as its importance. The whole calculation process is provided in Figure 2B. Due to the computational cost associated with training multiple models, a disturbance-based approach is employed. First, a neural network model is trained without any rules until convergence, which serves as a baseline. Then, relevant rule coalitions are incorporated into the pre-trained model, respectively, to disturb the convergence. The model is trained for several epochs until a new convergence is established. Finally, the MSE of each informed model on the testing data is measured to calculate the *RI* by Equation 4. When calculating the *RI* in the scenario of Figure 2A, the observation dataset is split into training, validating, and testing datasets, and a deep neural network model is considered. In the pre-training phase, the training data loss continues to decrease and tends to converge, whereas the rule loss remains elevated. At this time, the rule coalition is added to the model and disturbs the convergence (Figure 2C). After the incorporation of the rule coalition, the data loss presents a sudden enlargement and starts to decrease together with the rule loss. Meanwhile, as shown in Figure 2D, the validating loss achieves a convergence in the pre-train stage rapidly. After adding rules, a new convergence is achieved after certain epochs, and the convergence value is less than that in the pre-train stage. Therefore, in practice, the parameters of the pre-trained network and optimizer are saved, and the involved rule coalitions are added to the pre-trained network to continue training with the constraint of rules. Compared with training a complete physical constraint model for each rule alliance, this strategy can save considerable time by experiencing the long pre-train process only once.

Inherent principles behind data and rules

In this section, we conduct a series of systematic experiments to elucidate the fundamental principles underlying data and rules. At this stage, four canonical physical processes that can be described by explicit governing equations are taken as representative examples, which are illustrated in Figure 3A. The involved rules cover the governing PDEs, boundary conditions (BCs), and initial conditions (ICs). The rules are incorporated into the model through the loss function,[7] and the experimental settings are detailed in section S1. The generalization ability of models trained with both data and rules is depicted in Figure S8 and indicates that these models have satisfactory predictive capabilities in both in-distribution and out-of-distribution scenarios. In the following

part, we will provide a summary of qualitative and consistent insights that have been further explicated from the perspective of the sample distribution. These insights will also be validated via numerical experiments.

***In-distribution prediction: Larger data volume, lower* RI**

First, we examined the in-distribution prediction scenario, wherein the distributions of test data and training data are either similar or identical. Constructing the surrogate model is a typical in-distribution task since it aims to recover attributes in the whole domain through a few scattered observation data. Here, different volumes of data, including 0 (no data), 10, 10^2, 10^3, and 10^4, are randomly sampled as the training dataset, and the test data volume is 10,000. For the four canonical physical processes, the importance of each integrated rule is calculated and illustrated in Figures 3B and S6A.

Through the quantification of *RI*, the efficacy of different types of rules and their sensitivity to the data volume are reflected explicitly. Notably, we observe a diminishing impact of *RI* with increasing data volume, with distinct decay patterns depending on the type of rule (as illustrated in Figure 3B). Our findings have uncovered counter-intuitive results in which, when the volume of data exceeds a certain threshold, the effect of PDEs may become negative. From the perspective of high-dimensional sample distribution, these findings can be elucidated since the model essentially interpolates on the learned sample distribution from training data for in-distribution prediction. In this scenario, a larger volume of data enables the model to learn a more accurate distribution without the need to integrate rules to limit the range of alternative distributions.

***Out-of-distribution prediction: Larger data volume, higher global* RI, *lower local* RI**

In real-world scenarios, physical processes often exhibit evolving distributions of quantities, which poses a challenge in predicting future quantities. Therefore, we investigate the influence of data volume on out-of-distribution prediction, which involves inferring a different sample distribution, such as future prediction. The experimental settings are provided in section S1.

In out-of-distribution prediction, we find that the influence of data volume varies with rule type. Rules that globally restrict the entire domain (e.g., PDE rules) are termed global rules, while rules that locally restrict the observable area (e.g., IC rules) are termed local rules. It is found that the importance of global rules increases as the data volume rises (Figure 3C). Conversely, the importance of local rules diminishes with a larger data volume (Figure 3C). This cognition is distinctive, but reasonable, as the unobservable domain outside of the distribution is difficult to learn with observable data. Under this circumstance, more observed data do not necessarily improve the predictive ability of the model but instead increase the risk of overfitting. Therefore, global rules will play a more critical role in instructing the model by restricting possible distributions globally. In contrast, the function of local rules coincides with observed data and this accounts for the declining *RI* with data volume.

The interactions between rules

In this section, we aim to explore the interactions among multiple rules through complex scenarios, such as the solution of multivariable equations and two-dimensional PDE, which involves more rules. For solving multivariable equations, the following example is provided:

$$\begin{cases} c = |\sin(a) - \cos(b) & (\text{rule } 1) \\ d = \log\left((a-b)^2+1\right) & (\text{rule } 2) \\ e = 0.5\left(1+c^2\right) & (\text{rule } 3) \\ f = \exp(-e) & (\text{rule } 4) \end{cases}, \quad \text{(Equation 6)}$$

where the independent variables are $a \in [0, \pi]$ and $b \in [-\pi, 0]$ and the dependent variables are *c*, *d*, *e*, and *f*. Another implicit rule, $e, f > 0$ (rule 5), is also considered. The form of the complex two-dimensional PDE is written as

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^4 u}{\partial y^4} = \left(2 - x^2\right)e^{-y}, \quad \text{(Equation 7)}$$

where the definition domains are $x \in [0, 1]$ and $y \in [0, 1]$. There are six conditions for solving this PDE (termed as rule 1), namely

$$\begin{array}{llll} u_{yy}(x,0) = x^2 & (\text{rule } 2) & u_{yy}(x,1) = x^2/e & (\text{rule } 3) \\ u(x,0) = x^2 & (\text{rule } 4) & u(x,1) = x^2/e & (\text{rule } 5) \\ u(0,y) = 0 & (\text{rule } 6) & u(1,y) = e^{-y} & (\text{rule } 7) \end{array}. \quad \text{(Equation 8)}$$

From our evaluation framework, several inherent interactions between rules are revealed explicitly, i.e., dependence, synergism, and substitution effects.

Dependence

This work reveals the existence of extensive dependencies among rules and variables. Two types of dependence are identified: inner dependence between rules, and outer dependence between rules and dependent variables.

The inner dependence can be explicitly revealed by comparing the *FI*, which is defined as the importance of a specific rule when all other rules are present, with the calculated *RI*. In the task of solving two-dimensional PDEs, as illustrated in Figure 4A, the *RI* of most rules is lower than *FI*, especially the PDE rule (rule 1), indicating that most rules have dependence and need other relying rules to be effective. Given that the proposed *RI* essentially measures the marginal contribution of the rule across all coalitions, the low *RI* of the PDE rule implies that the incorporation of the PDE rule is ineffective in most scenarios since it depends highly on the other rules (i.e., high dependence). Meanwhile, the high *FI* underscores the PDE rule's ability to yield its full impact when complemented by other rules. Moreover, it is observed that the inclusion of observation data affects inner dependence (Figure 4A).

The outer dependence is illuminated by evaluating the importance of each involved rule for every dependent variable. As illustrated in Figure 4B, the outer dependence in the task of solving multivariable equations is transparent, where a higher *RI* corresponds to a larger dependence. For example, variable *d* depends heavily on rule 2, while variable *f* is slightly dependent on rules 1, 3, and 4. This result is consistent with the intuitive analysis of the rules, demonstrating the reliability of the proposed method and indicating that the quantitative measurement of *RI* facilitates an explicit investigation of outer dependence.

Synergism

Synergism is a specific form of interaction in which multiple rules work together to produce an effect that is greater than the sum of their individual effects. In our study, we observed evidence of synergism through the importance of different numbers of other relying rules, which is defined as

$$RI_i^r(v) = \frac{1}{N_{S_i^r}} \sum_{s \in S_i^r} \log_{10}\left(\frac{MSE(s \setminus \{i\})}{MSE(s)}\right), \quad \text{(Equation 9)}$$

where RI_i^r is the importance of rule *i* with the number of other relying rules *r* and S_i^r refers to the set of rules coalition containing the rule *i* and other relying rules *r*. As illustrated in Figures 4C and 4D, synergism is discovered in both cases. Specifically, in the task of solving multivariable equations, RI^r shows an evident increase for rules 1, 2, 3, and 4, which indicates that a synergism effect exists in these rules. Conversely, rule 5 does not participate in the synergism. Similarly, in the task of solving two-dimensional equations, synergism is observed in the PDE rule. However, this synergy effect is only apparent when sufficient dependency rules exist (over five relying rules) and is not affected by the inclusion of observation data. The RI^r for other rules is displayed in Figures S9 and S10.

Substitution

The substitution effect refers to the phenomenon in which the function of one rule can be substituted by either the data or other rules. This effect is illustrated in Figure 4C, where the RI^r of rule 5 is subtle overall and even decreases with a larger number of relying rules. We have discovered that rules regarding domain are more susceptible to being substituted. The substitution effect can arise when there are sufficient data or redundancy among rules or when specific rules only apply to certain conditions or domains. Overall, apprehending the substitution effect can assist to simplify the model and improve its efficiency by reducing redundant or replaceable rules.

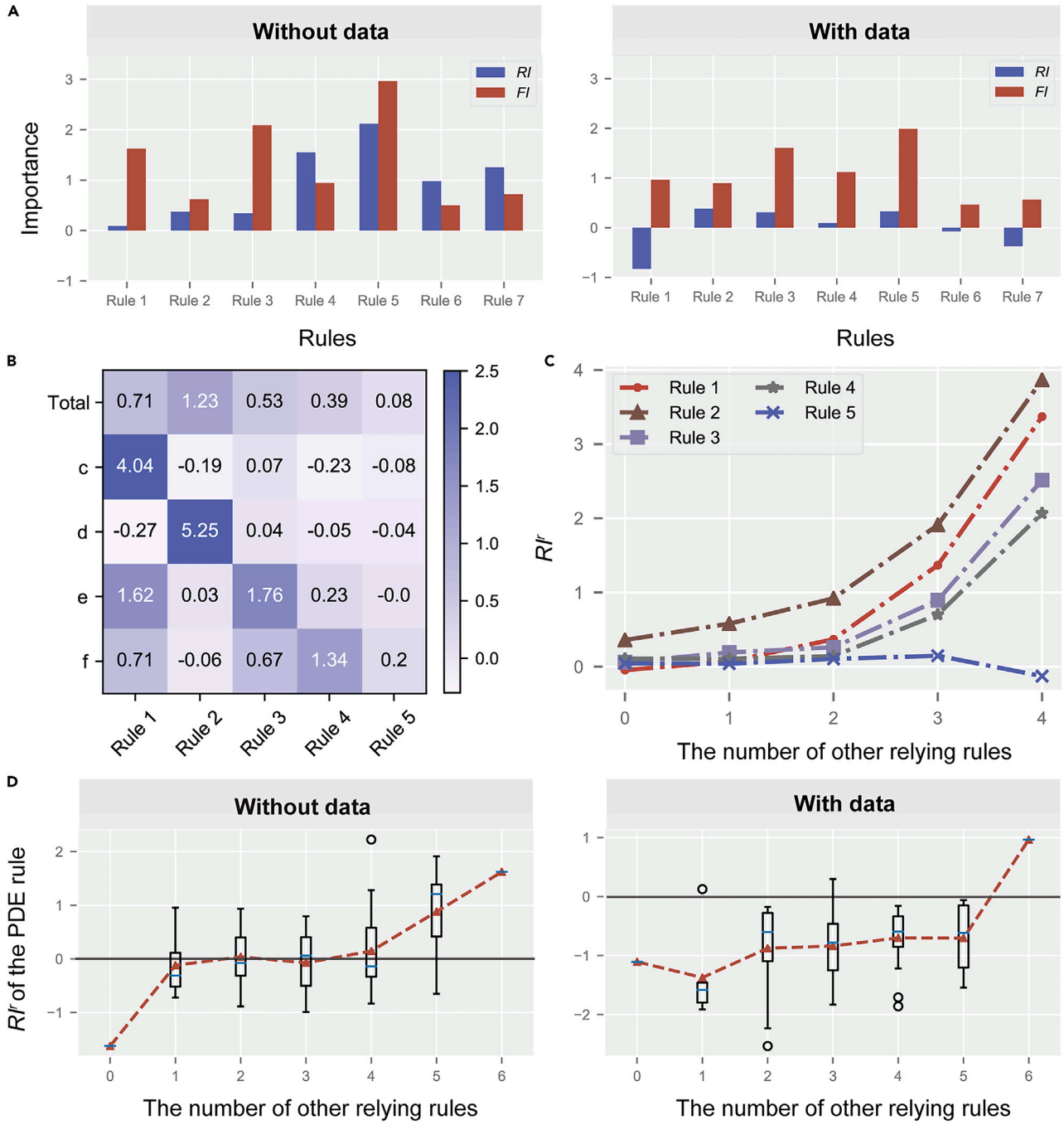

Figure 4. Numerical experiments to explore the interactions among multiple rules

(A) The *RI* and full importance (*FI*) of each involved rule when solving a two-dimensional PDE with and without data.

(B) The *RI* of the involved rules is calculated by each dependent variable when solving multivariable equations without data.

(C) The importance of the involved rules under different numbers of relying rules when solving multivariable equations without data.

(D) The importance of the PDE rule under different numbers of relying rules when solving two-dimensional PDE with and without data. The boxplot refers to all calculated importance with a given number of relying rules. The blue lines in the boxplot refer to the median.

Practical applications

In current machine learning research, it is widely acknowledged that incorporating prior knowledge can enhance the predictive ability of an informed machine learning model.[7] However, the selection of appropriate prior knowledge and the quantification of its worth remain challenging, hindering the practical application of such techniques, particularly in scenarios with multiple rules. Our proposed framework manages to solve the abovementioned problem and offers a solution at the pragmatic level. In this work, we present two examples of how our framework can be used to adjust the regularization parameter and distinguish improper prior rules.

One of the key challenges in informed machine learning, particularly in physics-informed neural networks (PINNs), is the difficulty of adjusting the weights of rules during the training process.[22,23] Our framework provides a simple, yet efficient, strategy to address this problem, whereby we increase the weights of rules with positive importance and decrease the weights of rules with negative importance. The detailed algorithm for this process is provided in Figure S11. We adopt

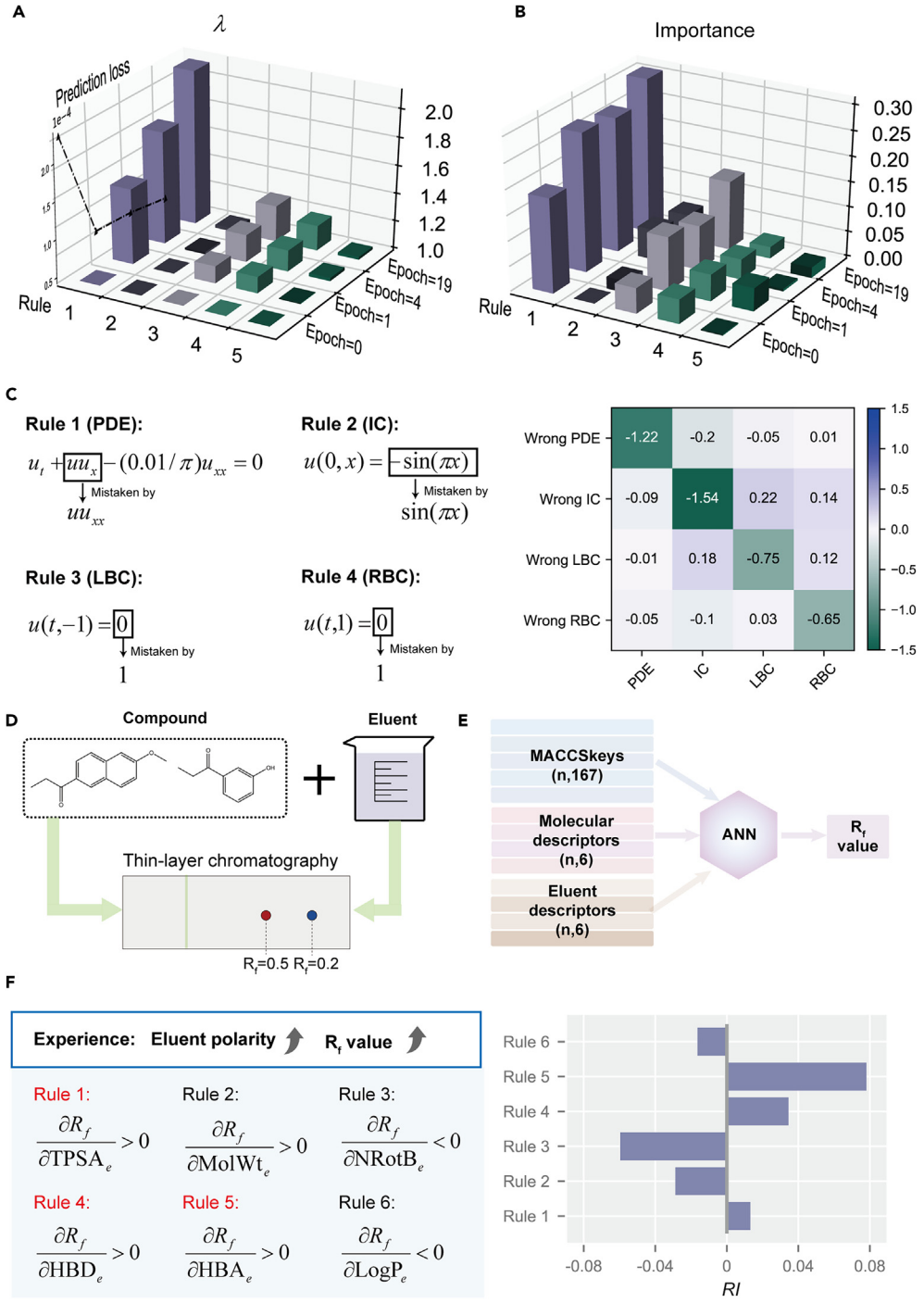

Figure 5. Practical application of rule importance

(A) The change in the weights λ_i of rules under different iterations. The left side is the prediction loss measured by mean squared error. The example of solving multivariable equations is utilized here.

(B) The change of rule importance during the optimization process.

(C) The correct rules and the corresponding mistaken part (left), and the change in the importance of each rule with wrong rules (right). Here, a situation in which only one rule error is considered.

(D) The diagram for thin-layer chromatography.

(E) The feature characterization and model construction for R_f value prediction model.

(F) The extracted rules from experience (left), where the rules in red are identified to be effective by rule importance (right).

the example of solving multivariable equations to demonstrate the effectiveness of our technique, and the optimized results are presented in Figures 5A and 5B. The prediction error shows a significant decrease during the optimization (Figure 5A), indicating its efficacy. Meanwhile, the importance of most rules increases after the optimization, which indicates that the adjustment assists to enhance the value of knowledge (Figure 5B). The comparison with other optimization methods is provided in Table S1.

This framework can also assist in identifying improper or wrong rules that may negatively impact the model. In real-world applications, it is often challenging to ensure that all rules are appropriate for the given problem, which can interfere with model training. A numerical example is provided in Figure 5C in which each rule is mistaken by a wrong rule with an imperceptible error in turn. The Burgers' shock equation is utilized here as an example, and the change in the *RI* of each rule in the presence of wrong rules is also displayed in Figure 5C. Several interesting findings can be obtained from the results. First, the wrong rules present a fully negative effect, leading to a significant decline in importance. This striking negative importance highlights the wrong rule, which provides an explicit way to identify the improper rules. Second, the wrong rules will affect the importance of the other rules. For example, the importance of the PDE rule (rule 1) decreases with the existence of other improper rules. In contrast, the importance of the initial rule (rule 2) increases when the BCs are wrong but decreases with a wrong PDE rule. This phenomenon implies the existence of different competitive and promotional relationships between rules.

The proposed framework offers practical utility for knowledge identification and model construction within interdisciplinary research. To illustrate its effectiveness, we provide an example from the field of chemical informatics. In this context, we demonstrate how *RI* can enhance the transformation of experiential insights into actionable knowledge, thereby improving the performance of deep learning models when confronted with real-world data.

In Figure 5D, we present an example related to thin-layer chromatography, a commonly used technique for measuring the polarity of compounds in organic experiments. In practical operations, the retardation factor value (R_f value), calculated as the ratio of the developed distance of solutes to eluents, plays a crucial role in reaction monitoring, product identification, and establishing chromatography conditions for subsequent purification.[24] However, the experimental process can be time consuming and labor intensive. Consequently, bridging the gap between experimental chemistry and informatics to predict experimental results using deep learning methods becomes essential. To ensure data conformance and integrality, the experimental data are collected through an automatic high-throughput platform.[25,26] The feature characterization and model construction are depicted in Figure 5E. Molecular features are represented using MACCS keys and descriptors, while eluent features are represented by their weighted descriptors.

In contrast to the conventional mathematical models that can provide explicit rules, practical scenarios often involve experiential insights. For instance, in this context, domain experts might know that an eluent with higher polarity tends to result in a larger R_f value for the same compound. However, this experiential knowledge is not formalized and cannot be directly incorporated into the model. Therefore, we tentatively propose six potential rules based on this experiential insight, as shown in Figure 5F, but their effectiveness remains uncertain. Our framework is employed to calculate the *RI* of each rule (Figure 5F), which clearly shows that the rule relating to the total polar surface area, hydrogen bond acceptors, and hydrogen bond donors of the eluent is more effective, as indicated by its positive *RI*. This process allows us to extract effective rules, thereby generating new insights into how eluent descriptors influence R_f values. Experimental results validate that incorporating these rules significantly enhances model performance, reducing the MSE of the testing dataset for new compounds from 0.052 to 0.036.

DISCUSSION

In this study, we have introduced an interpretable framework for evaluating the worth of knowledge in deep learning. The proposed measurements of *FI* and *RI* provide insights into the relationship between data and rules. We have also investigated the dependence, synergies, and substitution of rules, which can assist to identify important knowledge components. Additionally, the importance of rules can be utilized in practical applications, such as improving the performance of informed machine learning and identifying improper rules, which in turn can enhance the worth of knowledge. The utilization of the proposed framework in real-world scenarios with actual data indicates that the *RI* can effectively transform experiential insights into actionable knowledge, thereby improving the performance of the model. Additionally, we have substantiated the significance of this framework through its application to a practical engineering problem related to two-dimensional heterogeneous subsurface flow (section S2.3). These experiments prove that our framework can shed light on diverse fields encompassing physics, chemistry, geoscience, and engineering. Moreover, we evaluated the effectiveness of the framework in the presence of noisy data and discovered that prior rules tend to have a higher importance in the presence of moderate data noise (section S2.4). Meanwhile, it is revealed that the choice of collocation points, including the location and volume, has an impact on the model training (section S2.1).

A deep understanding of the nexus between knowledge and data plays a crucial role in the construction of informed machine learning models for practical problems. In essence, the prior knowledge assists the models to pre-learn certain patterns or principles of the target process or phenomenon; therefore, informed machine learning techniques such as PINN have broad scientific and engineering applications across multiple domains. However, there are still some limitations in their practical use. In real-world scenarios, there often exist various forms of rules, which leads to a complex multiobjective optimization problem when training PINN and thereby impeding its accuracy and application. Therefore, the framework proposed in this paper aims to improve the applicability and accuracy of informed machine learning models by measuring the importance of each rule to assist in adjusting hyperparameters. From this perspective, the framework can be applied to the optimization of most scientific and engineering problems involving informed machine learning models. Additionally, for some more complex problems, there may not exist natural rules but only empirical models and potential patterns. However, whether they are effective is unknown. Therefore, our framework can be employed to evaluate and filter out useful prior knowledge to enhance the predictive capability of deep learning models, which effectively expands the applicability of informed machine learning.

On the other hand, our exploration into *RI* has illuminated the imperfections inherent in informed machine learning. Firstly, the experiments in our work have revealed that an increase in the volume of data will diminish the significance of local rules in the out-of-distribution scenario, highlighting the necessity for more general rules. However, disciplines such as chemistry and biology often lack readily available general rules akin to governing equations. While our study identifies this issue, efficient methods for extracting general rules from large-scale datasets in these fields remain elusive, warranting further research. Meanwhile, our framework may take a relatively long time (0.5–2 h) for the computation of *RI*. Therefore, in the future, we will further speed up the calculation process of *RI* through a Monte Carlo-type method. Additionally, we aim to extend the proposed framework to measure the importance of rules related to mathematics and physics, such as invariance and logic rules, and investigate their functions in deep learning. Overall, we believe that evaluating the worth of knowledge is critical not only for interpretability but also for improving the security and reliability of informed machine learning models. This approach is especially relevant for future research on large-scale models, in which model security is a growing concern.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Dongxiao Zhang (dzhang@eitech.edu.cn).

Materials availability

This study did not generate new unique materials.

Data and code availability

The dataset utilized in this study has been deposited in the GitHub repository: https://github.com/woshixuhao/Worth_of_Knowledge/tree/main/Data. All original code has been deposited in the GitHub repository: https://github.com/woshixuhao/Worth_of_Knowledge/tree/main/Code. https://doi.org/10.5281/zenodo.10083605.

The theoretical basis for *RI*

For the simplest case, in which a single rule is integrated into the model, the profit of integrating this rule can be easily measured by the improvement of the network's predictive accuracy. However, informed machine learning usually involves multiple prior rules, which will incur complex interactions between rules. Under this circumstance, the measurement of *RI* draws inspiration from the Shapley value (see section S1.5), which is initially proposed to solve the problem of determining individual contribution in the coalitional game.[20,21] The rule coalitions are not coalitional games in the conventional sense; however, the marginal contribution of each rule can be assigned in our framework where negative contributions are permitted. For the calculation of *RI*, several descriptions are put forward beforehand to describe the process explicitly, which can be summarized as follows.

1. The basic predictive performance of the model without any rules is seen as the baseline.

In this work, rules are incorporated into the model to improve predictive accuracy. Therefore, the basic predictive performance of the model without any rules is seen as the baseline. In other words, the predictive accuracy of the model trained with only data (or even no data) is employed for comparison to measure the effect of incorporating rules.

2. The profit of the rule coalition is defined as the improvement of the model's predictive accuracy.

In this work, the profit of the rule coalition is embodied by the improvement of the model's predictive accuracy. The definition means that negative contributions may exist due to the inclusion of improper rules or their coalition, which may actually lead to a decrease in overall predictive accuracy. Therefore, the calculated *RI* can be either positive or negative, as rules may improve or decrease the predictive accuracy of the model. It is important to note that this problem is inherently complex since the alliance of rules may not always enhance the profit. As a result, there may be complex interactions between rules involving both cooperation and competition, which is carefully considered in our analysis.

3. The importance of rules is assigned by their respective marginal contributions.

Similar to the Shapley value, the proposed *RI* is also assigned by respective marginal contributions, which is calculated in a specialized manner that permits the negative contribution and can handle the situation involving both cooperation and competition. This means that the measurement of rule *i* is essentially the average value of the difference in the model's predictive accuracy between all rule coalitions including and excluding rule *i*.

SUPPLEMENTAL INFORMATION

It can be found online at https://doi.org/10.1016/j.ynexs.2024.100003.

ACKNOWLEDGMENTS

This work was supported and partially funded by the National Center for Applied Mathematics Shenzhen (NCAMS), the Shenzhen Key Laboratory of Natural Gas Hydrates (grant no. ZDSYS20200421111201738), the SUSTech – Qingdao New Energy Technology Research Institute, and the National Natural Science Foundation of China (grant no. 62106116).

AUTHOR CONTRIBUTIONS

H.X., Y.C., and D.Z. conceived the project, designed and performed research, and wrote the paper, and H.X. implemented workflow, created code, visualized results, and analyzed and curated data.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 6, 2023
Accepted: January 14, 2024
Published Online: March 8, 2024

REFERENCES

1. Zhang, Y.D., Morabito, F.C., Shen, D., et al. (2021). Advanced deep learning methods for biomedical information analysis: An editorial. Neural Network. 133, 101-102.
2. Wang, J., Zhu, H., Wang, S.H., et al. (2021). A Review of Deep Learning on Medical Image Analysis. Mobile Network. Appl. 26, 351-380.
3. Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596, 583-589.
4. Raissi, M., Yazdani, A., and Karniadakis, G.E. (2020). Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. Science 367, 1026-1030.
5. Fu, J., Liu, P., and Zhang, Q. (2020). Rethinking Generalization of Neural Models: A Named Entity Recognition Case Study. In AAAI 2020 - 34th AAAI Conference on Artificial Intelligence.
6. Vonrueden, L., Mayer, S., Beckh, K., et al. (2021). Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. IEEE Trans. Knowl. Data Eng. 1-20.
7. Karniadakis, G.E., Kevrekidis, I.G., Lu, L., et al. (2021). Physics-informed machine learning. Nat. Rev. Phys. 3, 422-440.
8. Diligenti, M., Roychowdhury, S., and Gori, M. (2017). Integrating prior knowledge into deep learning. In Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017.
9. Fung, G.M., Mangasarian, O.L., and Shavlik, J.W. (2003). Knowledge-based support vector machine classifiers. In Advances in Neural Information Processing Systems.
10. Towell, G.G., and Shavlik, J.W. (1994). Knowledge-based artificial neural networks. Artif. Intell. 70, 119-165.
11. 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. J. Comput. Phys. 378, 686-707.
12. Wang, R., Walters, R., and Yu, R. (2020). Incorporating symmetry into deep dynamics models for improved generalization. Preprint at arXiv. https://doi.org/10.48550/arXiv.2002.03061.
13. Muralidhar, N., Islam, M.R., Marwah, M., et al. (2019). Incorporating prior domain knowledge into deep neural networks. In Proceedings - 2018 IEEE International Conference on Big Data.
14. Xu, H., Chang, H., and Zhang, D. (2020). DLGA-PDE: Discovery of PDEs with incomplete candidate library via combination of deep learning and genetic algorithm. J. Comput. Phys. 418, 109584.
15. Lu, L., Meng, X., Mao, Z., et al. (2021). DeepXDE: A deep learning library for solving differential equations. SIAM Rev. 63, 208-228.
16. Chen, Y., and Zhang, D. (2021). Theory-guided deep-learning for electrical load forecasting (TgDLF) via ensemble long short-term memory. Adv. Appl. Energy 1, 100004.
17. Cully, A., Clune, J., Tarapore, D., et al. (2015). Robots that can adapt like animals. Nature 521, 503-507.
18. Yang, Y., and Perdikaris, P. (2019). Adversarial uncertainty quantification in physics-informed neural networks. J. Comput. Phys. 394, 136-152.
19. Chen, Y., Huang, D., Zhang, D., et al. (2021). Theory-guided hard constraint projection (HCP): A knowledge-based data-driven scientific machine learning method. J. Comput. Phys. 445, 110624.
20. Shapley, L.S. (1971). Cores of convex games. Int. J. Game Theor. 1, 11-26.
21. Aumann, R., and Hart, S. (1992). Handbook of Game Theory with Economic Applications.
22. Rong, M., Zhang, D., and Wang, N. (2022). A Lagrangian dual-based theory-guided deep neural network. Complex Intell. Syst. 8, 4849-4862.
23. Du, M., Chen, Y., and Zhang, D. (2022). AutoKE: An automatic knowledge embedding framework for scientific machine learning. IEEE Trans. Artif. Intell. 1-16.

24. Sherma, J., and Fried, B. (2003). Handbook of Thin-Layer Chromatography (Revised and Expanded).

25. Xu, H., Zhang, D., and Mo, F. (2022). High-throughput automated platform for thin layer chromatography analysis. STAR Protoc. 3, 101893.

26. Xu, H., Lin, J., Liu, Q., et al. (2022). High-throughput discovery of chemical structure-polarity relationships combining automation and machine-learning techniques. Chem 8, 3202-3214.