arXiv:2603.03346v1 [physics.geo-ph] 24 Feb 2026

基于物理约束的符号回归：从实验数据中发现多峰水分保持曲线的闭式方程

Yejin Kim[1] Hyoung Suk Suh[1,*]凯斯西储大学土木与环境工程系，美国
y.kim@case.edu hssuh@case.edu

摘要

利用多峰孔径分布描述多孔材料的非饱和行为颇具挑战，因为标准水力学模型往往无法捕捉其复杂的多尺度特征。一种常见的变通方法是叠加单峰保持函数，每个函数针对特定孔径范围进行调整；然而，这种方法需要针对每种模式分别进行参数识别，因此会限制模型的可解释性和泛化能力，尤其是在数据稀疏的情况下。在本研究中，我们提出了一种本质上不同的方法：一种用于元建模的物理约束机器学习框架，可直接从实验数据中自动发现多峰水分保持曲线的闭式数学表达式。数学表达式以二叉树表示，并通过遗传编程不断演化；同时，将物理约束嵌入损失函数，引导符号回归器找到符合物理规律且数学上稳健的解。结果表明，所提出的框架能够发现闭式方程，有效表征具有不同孔隙结构的多孔材料的持水特性。为支持第三方验证、应用和扩展，我们在开源代码库中公开了完整实现。

1 引言

持水特性在多孔介质非饱和流建模中起着核心作用，因为它决定了不同条件下流体的分布与滞留情况。水分保持曲线能够描述这种行为，它将基质吸力与饱和度联系起来，并作为水力-力学耦合模拟宏观模型中的关键本构方程 [Borja, 2004, Griffiths and Lu, 2005, Song and Borja, 2014, Kim et al., 2025]。Brooks、Van Genuchten、Kosugi、Fredlund 和 Xing 提出的半经验模型 [Brooks, 1965, Van Genuchten, 1980, Kosugi, 1994, Fredlund and Xing, 1994], 常用于此类模拟，因为它们能够以较少的参数，通过简单的解析表达式描述水分保持曲线。然而，这些模型假设相应的多孔介质具有单峰孔径分布，因此不适用于准确描述具有复杂或多峰孔隙结构材料的非饱和行为 [Li 等人，2014, Zhou 等人， 2017]。一种常见的解决方法是组合多个单峰模型，每个模型对应一个不同的孔径范围 [Durner, 1994, Priesack and Durner,2006], ，但这种方法可能会使参数识别过程变得复杂，从而限制其适用性，尤其是在可用数据稀少时。

水分保持特性在多孔介质非饱和流建模中发挥着核心作用，因为它决定了流体在不同条件下的分布与滞留情况。这种行为可通过水分保持曲线来描述，该曲线将基质吸力与饱和度联系起来，并作为水力-力学耦合模拟中宏观模型的关键本构方程 [Borja, 2004, Griffiths and Lu, 2005, Song and Borja, 2014, Kim et al., 2025]。半经验模型（例如 [Brooks, 1965, Van Genuchten, 1980, Kosugi, 1994, Fredlund and Xing, 1994], 提出的模型）常用于这些模拟，因为它们以相对较少的参数提供简洁的解析表达式来描述水分保持曲线。然而，这些模型依赖于这样一种假设：相应的多孔介质具有单峰孔径分布，因此不适用于准确描述孔隙结构复杂或呈多峰分布材料的非饱和行为 [Li et al., 2014, Zhou et al., 2017]。一种常见的变通方法是组合多个单峰模型，每个模型对应一个不同的孔径范围 [Durner, 1994, Priesack and Durner,2006], ，但这种方法可能会使参数识别过程更加复杂，从而限制其适用性，尤其是在可用数据稀少时。

*通讯作者。

克服上述挑战的一种可能途径，是采用数据驱动方法，直接从实验数据中学习复杂行为，而无需预先指定函数形式或明确标定材料参数。在各种方法中，深度神经网络因其高度灵活且表达能力强而广受青睐。研究表明，它们在再现多种本构模型方面潜力巨大，例如弹性能量泛函 [Vlassis et al., 2020, Linden et al., 2023], 塑性屈服面和硬化规律 [Vlassis and Sun, 2021, Suh et al., 2023],渗透率 [Singh et al., 2020, Tembely et al., 2020]，以及水分保持特性 [Liu et al., 2019, Heider et al., 2021]。尽管神经网络具有强大的能力来表征这类复杂的非线性关系，但在实际工程应用中，它们尚未成为主流。这在很大程度上是由于其缺乏可解释性；神经网络的黑箱特性往往使训练模型的可信度受到质疑，这可能阻碍了其更广泛的应用 [Murdoch et al., 2019, Fan et al., 2020, Wing, 2021]。

尽管人们已采取多种措施来提高机器学习方法的可解释性[Saleem 等，2022，Li 等，2022，Liu 等， 2024], 正如 [Agarwal 等， 2021], 所指出的那样，为提高可解释性而设计的模型往往缺乏实现高预测准确度所需的表达能力（即，可解释性与准确度之间的权衡）。相比之下，符号回归旨在发现最符合给定数据集的数学表达式，只要用于搜索表达式空间的组合优化取得成功，就能够同时实现较高的可解释性和准确性 [Angelis 等，2023，Cranmer，2023，Makke 和 Chawla， 2024]。然而，这种最佳结果在实践中往往难以实现。符号回归的一个主要局限是其可扩展性：随着输入空间的维度或目标函数的复杂度增加，候选表达式的数量会呈指数级增长，使搜索过程在计算上变得十分繁重 [Petersen 等，2019，Udrescu 和 Tegmark， 2020]。此外，符号回归对数据中的噪声和离群值尤其敏感，这可能导致过拟合，并发现无法推广到训练数据集之外的表达式 [Bomarito 等，2022，de Franca 和 Kronberger， 2023]。为解决这些问题，近期研究 [Bahmani 等，2024，Suh 等， 2024] 提出了将神经网络与符号回归相结合的混合框架：首先训练神经网络来逼近数据，随后通过符号回归提取神经网络函数的闭式表达式。尽管这种方法在解决上述问题方面取得了一定成功，但有一点被忽视了：所得数学表达式或其神经网络对应形式可能产生在物理上不合理的学习函数（例如，学习得到的持水模型预测出超过 1 的饱和值，或表现出非单调行为）。

因此，本研究提出了一种针对多孔介质水分保持的元建模方法，采用多目标符号回归器，从实验数据中发现多模态持水曲线的闭式数学表达式，同时保持物理一致性。该框架将数学表达式表示为二叉树，并利用遗传编程使其在连续多代中不断演化。多目标优化策略通过最小化预测误差来驱动这一演化，同时借助专门的损失项施加物理约束，包括单调性、极限约束和模式约束。该方法能够以数据驱动的方式发现最能描述观测数据的数学表达式，同时严格满足施加的物理约束。即使所得方程具有一定的复杂度，它们仍然透明、易于解释且可进行解析处理，与黑箱神经网络模型形成鲜明对比。因此，发现的闭合关系可以像传统持水模型一样，无缝集成到现有的水力模拟代码中。为促进透明度、可重复性和进一步开发，完整实现已公开发布于开源代码库。为确保清晰并聚焦研究主题，本研究专门针对主干燥（即排水）过程中的持水行为建模，刻意不考虑滞后效应，并计划在未来工作中处理这些复杂问题。

2 方法

本节首先在第 2.1 节将从实验数据中发现持水模型的问题表述为多目标回归任务。随后，我们概述模型必须满足的关键物理约束，例如单调性、吸力压力极端情况下的极限条件以及饱和度界限，以准确反映非饱和多孔介质的行为（第 2.2 节）。最后，在第 2.3 节中，我们

提出一种用于多模态持水曲线的物理约束符号回归（PCSR）框架，并在训练过程中通过惩罚施加物理约束。稍后将在第 3 节通过一系列示例展示所提出框架的性能。

2.1 学习问题

本文所考虑的机器学习任务是：给定一组数据对 $\{(s^{(i)},\ S_w^{(i)})\}_{i=1}^{n_{\text{data}}}$，其中 n_{data}表示实验测量的数量，学习一个函数 $\hat{}\ S_w(s)$，将吸力压力$s \in \mathbb{R}$ 映射到相应的含水饱和度 $S_w \in \mathbb{R}$。我们的目标是学习一个模型，使其不仅能够拟合观测数据，还能很好地泛化到未见过的基质吸力水平，准确表征多孔材料的持水行为。这要求确保所学习的函数符合底层物理规律。此外，我们假设目标函数的形状先验已知，并由指定的峰数表征，记为 N_{mode}。该参数可从实验观测或材料特性中识别，作为结构先验，并引导符号搜索，从而在不限定僵化数学架构的前提下缓解过拟合。

为此，可以将该问题表述为一个多目标优化任务，其中总损失函数 $\mathcal{L}$ 由三个主要部分组成：用于确保拟合实验数据的数据损失项 $\mathcal{L}_{\text{data}}$ 、用于施加必要物理约束的物理损失项 $\mathcal{L}_{\text{phys}}$ ，以及用于使所学习函数达到预期峰数的形状引导损失项 $\mathcal{L}_{\text{mode}}$ ：

$$\mathcal{L} = \underbrace{\frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} \left[\hat{S}_w^{(i)}(s^{(i)}) - S_w^{(i)} \right]^2}_{=\mathcal{L}_{\text{data}}} + \mathcal{L}_{\text{phys}} + \mathcal{L}_{\text{mode}}, \tag{1}$$

其中，上方带帽符号的量对应于所学习模型的预测值，而 N_{mode} 表示所学习函数的目标模式数。如果采用神经网络表示模型，例如，所有相关的 $\mathcal{L}_{\text{phys}}$ 项都必须构造为（部分）可微函数，以便在优化过程中将其纳入反向传播。相比之下，本研究采用的符号回归方法不要求损失分量可微。因此，无需为 $\mathcal{L}_{\text{phys}}$定义特定的函数形式；相反，通过自定义惩罚策略纳入相关的物理约束。这些约束将在第 2.2节概述，实现细节将在第 2.3节介绍。

虽然函数 $\hat{}\ S_w(s)$ 是根据一组观测数据对推断得到的，但因多孔材料类型不同，实验测量中的吸力压力 s 和含水饱和度 S_w在取值范围和尺度上都可能存在显著差异。直接使用此类原始数据可能会影响机器学习模型的稳定性和效率。为此，我们通过参考点 $(s_{\text{min}}, S_{w,\text{max}})$ 和 $(s_{\text{res}}, S_{w,\text{res}})$，将每组数据对 (s, S_w) 映射到归一化坐标 (s^*, S_w^*)。具体而言，首先对吸力压力 s 进行对数变换，再将其线性映射至 $s^* \in [0,1]$；而含水饱和度 S_w 则依据其最小值和最大值进行归一化，得到 $S_w^* \in [0,1],$，如图 1所示。这里，$S_{w,\text{max}}$ 和 $S_{w,\text{res}}$ 分别表示最大饱和度和残余饱和度，而 s_{min} 和 s_{res} 分别对应于引发气相侵入所需的最小吸力，以及达到残余饱和度时的吸力。利用这一映射，映射空间中的数据损失项可表示为，

$$\mathcal{L}_{\text{data}} = \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} \left[\hat{S}_w^{*(i)}(s^{*(i)}) - S_w^{*(i)} \right]^2, \tag{2}$$

而形状引导损失分量可确定为:

$$\mathcal{L}_{\text{mode}} = \left(\hat{N}_{\text{mode}} - N_{\text{mode}} \right)^2, \tag{3}$$

请注意， N_{mode} 与映射无关。在本研究中，所有训练均在映射空间中进行，并将$S_{w,\ \text{max}}$ 设为 1，以确保干燥过程从完全饱和条件开始。需要指出的是，这些映射参数是根据观测数据范围确定的辅助量，用于引导搜索过程，而非固有材料参数。

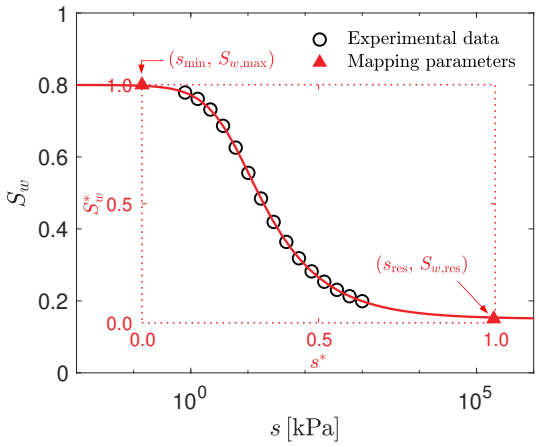


图 1：数据映射过程示意图。黑色空心符号表示示例性持水数据，黑色坐标轴表示原始（s， S_w）空间，红色坐标轴表示映射后的（s^*， S_w^*）空间。红色方框的左上角对应参考点（$s_{\min}$， $S_{w,\ \max}$），右下角对应原始空间中的（s_{res}，$S_{w,\ \min}$）；它们在变换空间中分别映射到（**0**， **1**）和（**1**， **0**）。此映射确保学习得到的函数（红色曲线）位于 $s^* \in [0,1]$ 和$S_w^* \in [0,1]$ 内。

映射参数或数据的微小扰动可能会影响发现的方程形式，尤其是考虑到基于遗传编程的符号回归具有随机性。尽管如此，需要指出的是，本研究的主要目标并非确定唯一的真实符号表达式，而是发现一个能够最佳表征给定数据集、且严格满足施加的物理约束的闭式表达式。尽管符号形式不可避免地存在差异，这种方法仍能确保模型具有物理一致性和可解释性。

2.2 水分保持曲线的物理约束

要确保学习得到的函数符合物理规律且适用于真实的多孔系统，就必须满足一组基于热力学原理的基本约束。这些约束规定了任何有效的持水模型在表征主干燥路径时都应满足的基本条件。假设学习得到的函数是连续的，以下列出这些关键条件：

单调性约束。 </style}随着 s 增加， S_w必须降低，这是干燥持水曲线的一项关键要求。换言之，饱和度应是基质吸力的单调递减函数：

$$\frac{\mathrm{d}S_w}{\mathrm{d}s} \leq 0. \tag{4}$$

这可确保随着基质吸力增加，水不会自发进入孔隙空间，否则将违背热力学原理。

极限约束。 </style}水分保持曲线的干燥路径在基质吸力范围的两端必须呈现特定的极限行为。在处，饱和度应保持在其最大值， $_{\max}$不变。此外，在湿端，饱和度的变化率应趋近于零，即

$$S_w(s_{\min}) = S_{w,\max} \ ; \ \left.\frac{\mathrm{d}S_w}{\mathrm{d}s}\right|_{s=s_{\min}} = 0. \tag{5}$$

相反，在极干燥条件（$s \geq s_{\text{res}}$）下，由于大多数孔隙中的水已排出，仅剩少量水分，水饱和度应趋近于残余水平。这意味着，在 $s = s_{\text{res}}$ 处，饱和度

趋近于其残余值，而 S_w 的变化率应趋近于 0：

$$S_w(s_{\text{res}}) = S_{w,\text{res}} \ ; \ \left.\frac{\mathrm{d}S_w}{\mathrm{d}s}\right|_{s=s_{\text{res}}} = 0. \tag{6}$$

有界性约束。无论参考点的位置如何，吸力-饱和度关系都必须限制在物理范围内。具体而言，饱和度 S_w 必须满足：

$$0 \leq S_w \leq 1. \tag{7}$$

该约束确保饱和度不超过 1，也不低于 0；这符合最大含水量对应完全饱和条件的概念，同时避免了含水量为负等非物理情况。

由于我们在映射空间中训练模型，物理约束 [式（4）-（7）] 必须相应地重新表述。令撇号表示对归一化吸力 s^*的导数，则映射空间中的单调性约束变为：

$$S_w^{*\prime}(s^*) \leq 0, \tag{8}$$

因为 $\mathrm{d}S_w^*/\mathrm{d}S_w$ 和 $\mathrm{d}s/\mathrm{d}s^*$ 均为非负数。同样，映射空间中湿端的极限约束可表示为：

$$S_w^*(0) = 1 \ ; \ S_w^{*\prime}(0) = 0, \tag{9}$$

而干端对应的约束为，

$$S_w^*(1) = 0 \ ; \ S_w^{*\prime}(1) = 0. \tag{10}$$

值得注意的是，如果两个参考点处的饱和度满足式（7），则由于其他约束的作用，映射空间中的有界性约束将自动得到满足。这意味着在模型训练过程中无需显式施加有界性约束。

2.3 用于发现闭式水分保持模型的物理约束符号回归器

符号回归旨在发现一个显式数学表达式，使其最符合给定数据集，而无需预先规定其形式。搜索过程通常涉及探索由预定义的变量、常数和数学运算符构成的大量候选表达式空间，每个候选表达式均以二叉树表示（见图 2）。此处，每个发现的表达式中的节点总数反映其复杂度，而复杂度与准确性共同构成评估模型性能的关键指标。在本研究中，采用遗传编程框架 [Koza，1994年、Wagner 和 Affenzeller，2005年、Schmidt 和 Lipson，2009] 在这一空间中进行搜索，并通过进化操作迭代优化候选个体群体，从而提高预测准确度和物理一致性。

训练首先随机生成一个候选表达式的初始群体。随后，该群体经历一个迭代进化过程，包含三项主要操作：选择、变异和交叉。选择是指从当前群体中挑选表现较好的个体。变异如图 3a所示，对单个候选个体进行小幅随机修改，以增加群体的多样性。交叉如图 3b所示，将两个个体的遗传信息组合起来，生成同时继承双亲特征的后代。

标准符号回归（SR）旨在最大限度地减少模型预测与数据之间的差异 [例如，使用均方误差（MSE）损失], 本研究则在此基础上，将第 2.2 节所述的基于物理的约束以及先验模态信息纳入学习过程。在我们的 PCSR 中，标准均方误差损失被自定义损失函数取代，该函数明确纳入了物理约束和模态约束。具体而言，每项约束都作为单独的损失项纳入，用于量化施加的约束与模型行为之间的差异；这些损失项则作为惩罚项整合进总损失函数，以引导训练过程中的搜索。

与训练物理信息神经网络类似 [Raissi 等，2019，Cai 等，2021，Haghighat 等，2021], 我们引入总计 n_{col} 个配点 $\boldsymbol{s}^*_{\text{col}} = \{s^{*(j)}{}_{\text{col}}\}_{j=1}^{n_{\text{col}}}$，这些配点均匀分布在

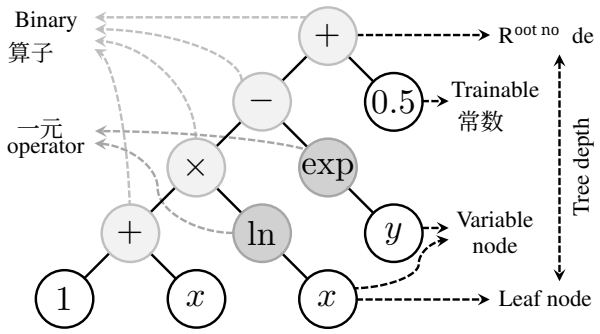


图2：以二叉树表示的示例符号方程。该树表示一个表达式（$1 + x$）$\ln(x) - \exp(y) + 0.5$，深度为 4，节点总数为 12。

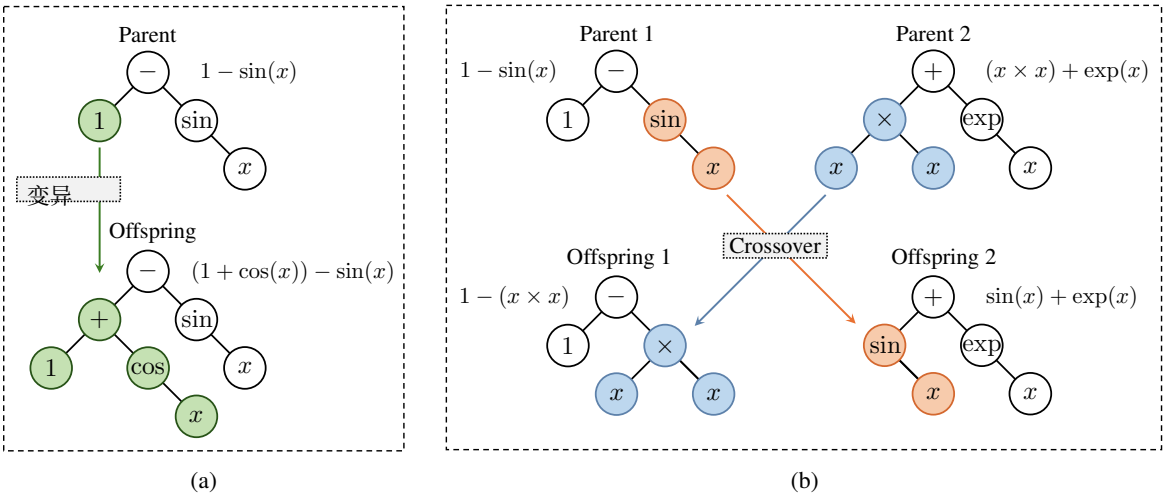


图3：演化符号回归算法中 (a) 变异和 (b) 交叉操作的示意图。

算法 1 用于物理约束符号持水模型发现的自定义损失函数。

输入： 映射空间中的符号表达式树 $\hat{}\, S_w^*(s)$、归一化数据集 $\{(s^{*(i)},\ S_w^{*(i)})\}_{i=1}^{n_{\text{data}}}$、规定的峰数 N_{mode}以及配点 $\boldsymbol{s}_{\text{col}}^* = \{s^{*(j)}{}_{\text{col}}\ \}_{j=1}^{n_{\text{col}}}$。

输出： 总损失 $\mathcal{L}$。

1. 计算 $\mathcal{L}$ data。

$\mathcal{L}_{\text{data}} \leftarrow \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} \left[\hat{}S_w^{*(i)}(s^{*(i)}) - S_w^{*(i)}\right]^2$ ▷ Eq. (2)

2. 计算 $\mathcal{L}$ phys。

如果 任意($\hat{S}_w^{*\prime}(\boldsymbol{s}_{\text{col}}^*) > 0$) **那么**

$\mathcal{L}_{\text{phys}} \leftarrow \mathcal{L}_{\text{phys}} + \frac{1}{\text{nnz}(\hat{}S_w^{*\prime}(\boldsymbol{s}_{\text{col}}^*)>0)} \sum_{j=1}^{n_{\text{col}}} \left[\text{ReLU}\left(\hat{}S_w^{*\prime}(s^{*(j)}{}_{\text{列}})\right)_2\right]$ ▷ Eq. (8)

else

$\mathcal{L}_{\text{phys}} \leftarrow 0$

结束如果

if $|\hat{S}_w^*(0)| < 1$**则**

$\mathcal{L}_{\text{phys}} \leftarrow \mathcal{L}_{\text{phys}} + \left[\hat{}S_w^*(0) - 1\right]^2$ ▷ Eq. $(9)_1$

结束 if

if $|\hat{S}_w^{*\prime}(0)| < 0$**则**

$\mathcal{L}_{\text{phys}} \leftarrow \mathcal{L}_{\text{phys}} + \left[\hat{}S_w^{*\prime}(0)\right]^2$ ▷ Eq. $(9)_2$

结束 if

if $|\hat{S}_w^*(1)| < 0$**则**

$\mathcal{L}_{\text{phys}} \leftarrow \mathcal{L}_{\text{phys}} + \left[\hat{}S_w^*(1)\right]^2$ ▷ Eq. $(10)_1$

结束如果

if $|\hat{S}_w^{*\prime}(1)| < 0$**则**

$\mathcal{L}_{\text{phys}} \leftarrow \mathcal{L}_{\text{phys}} + \left[\hat{}S_w^{*\prime}(1)\right]^2$ ▷ Eq. $(10)_2$

结束如果

3. 评估 $\mathcal{L}$ mode。

$\hat{}N_{\text{mode}} \leftarrow$ 模式数量函数($\hat{}S_w^*(\boldsymbol{s}_{\text{col}}^*)$)

如果不是 $\hat{N}_{\text{mode}} == N_{\text{mode}}$ **那么**

$\mathcal{L}_{\text{mode}} \leftarrow \left(\hat{}N_{\text{mode}} - N_{\text{mode}}\right)^2$ ▷ Eq. (3)

else

$\mathcal{L}_{\text{mode}} \leftarrow 0$

结束条件判断

4. 返回总损失 $\mathcal{L}$。**返回**

$\mathcal{L} \leftarrow \mathcal{L}_{\text{data}} + \mathcal{L}_{\text{phys}} + \mathcal{L}_{\text{mode}}$ 式（1）

该映射域，并作为明确施加约束的特定位置。如算法 1 所述，自定义损失函数的构造首先是在归一化数据集上计算数据损失（$\mathcal{L}_{\text{data}}$），以确保拟合实验观测。随后，通过检查配点处的预测是否违反物理约束，计算物理损失（$\mathcal{L}_{\text{phys}}$）。通过惩罚偏离方程 (8) 的情况来施加单调性约束：在每个不满足单调性条件的配点处，将惩罚项 $[\text{ReLU}(\hat{}S_w^{*\prime}(s^{*(j)}\ \text{col}))]^2$ 加入物理损失。此处，ReLU(•) 表示修正线性单元，它确保仅对导数大于零的情况进行惩罚，从而量化非单调性的程度。类似地，通过对域端点处（$s^* = 0$ 和 $s^* = 1$）的预期值及导数偏差施加额外惩罚，纳入极限约束，从而确保最大饱和度和残余饱和度下的物理一致性。随后，使用方程 (3) 计算模式损失（$\mathcal{L}_{\text{mode}}$）。此处，$\text{nummodes}(\hat{}S_w^*(\boldsymbol{s}_{\text{col}}^*))$ 通过统计学习函数（N_{mode}）的二阶导数 $S_w^{*\prime\prime}$ 中由负变正的符号变化次数，确定其模式数；该次数对应由凹到凸的拐点数。最后，自定义损失函数返回方程 (1) 中定义的总损失 $\mathcal{L}$。

为进一步探究施加的物理约束如何抑制非物理解并有效缓解过拟合，尤其是在存在含噪数据时，我们在附录A中对原始SR与受约束变体进行了详细的比较分析。

整个 PCSR 实现均构建于 Julia 软件包`SymbolicRegression.jl` [Cran-mer之上， 2023], 并以开放源代码形式发布，以支持第三方应用与扩展（参见开放研究）。

3 结果

本节全面评估所提出的 PCSR 框架在各种情景下发现水分保持模型闭式方程的表现。为建立全面的基准，我们将 PCSR 的性能与原始 SR 以及成熟的半经验模型进行比较 [Van Genuchten， 1980年， Durner， 1994]。在第3.1节中，我们将 PCSR 应用于两个不同的实验数据集，这些数据集代表具有单峰孔径分布的土壤，并考察其发现物理一致方程的能力。第3.2 节将评估扩展至多峰水分保持曲线，包括实验测得的双峰（$N_{\text{mode}} = 2$）数据集，以及模式数更高的合成生成数据集（例如， $N_{\text{mode}} = 3$和 4），以考察 PCSR 在捕捉复杂、多尺度持水行为方面的可扩展性和稳健性，以及其强制执行规定模式数的有效性。

本研究中的所有符号回归均采用一组统一的超参数设置进行训练：训练迭代次数固定为 100，候选表达式的种群规模设为 50，且每个候选表达式允许的最大复杂度限制为 150 个节点。模型可用的二元运算符限定为 $\{+, \times\}$，一元运算符则包括 {sin(•)、cos(•)、exp(•)、ln(•)}。对于 PCSR，配点数量固定为 $n_{\text{col}} = 100$。

需要注意的是，本研究并未采用传统的训练-测试集划分策略。所提出的框架旨在针对给定的实验数据集发现物理上合理且拟合效果最佳的闭式方程（即元建模），而非构建能够推广至不同数据集的模型。因此，所有可用的测量数据均被充分用作训练数据；同时，通过施加物理约束，有效控制过拟合。这些约束限制了可接受的函数空间，并防止模型找到可能拟合局部噪声或波动的非物理解。此外，由于我们的主要目标是确保发现的方程满足物理约束，而非防止过参数化（例如，最小化表达式复杂度），最终得到的方程可能非常复杂，但仍与物理规律相符。

3.1 单峰水分保持曲线

我们考虑了两个不同的水分保持数据集：一个对应级配不良砂，其水分保持曲线呈现陡峭斜率 [Song， 2014], 另一个对应黏壤土，其水分保持曲线为 [Tuller 和 Or， 2001]。对于每种情况，我们按照第 2.1 节所述，将实验数据映射到归一化空间以进行训练，使用以下映射参数：对于级配不良砂，$s_{\min} = 10^{-2}, S_{w,\max} = 1.0$，以及 $s_{\text{res}} = 10^5$；对于黏壤土， $s_{\min} = 10^{-1}, S_{w,\max} = 1.0, s_{\text{res}} = 10^6$，以及 $S_{w,\text{res}} = 0.035$。随后，将逆映射应用于

训练后模型的输出，从而得到模型预测。

图 4 展示了使用带模式约束（红色实线）和不带模式约束（黄色虚线）的 PCSR 发现的水分保持模型，以及使用原始 SR（绿色实线）和 [VanGenuchten， 1980] 模型（蓝色点线）得到的模型。图中，实验训练数据以黑色空心圆表示。半经验模型能够很好地拟合级配不良砂的实验数据（图 4a），但可能无法准确捕捉黏壤土在进气值或残余饱和度附近的持水特性（图4b）。尽管原始 SR 得到的模型均方误差损失极低，且与实验数据高度吻合，但它们往往会过拟合，在训练数据集之外的泛化能力有限。将物理损失 $\mathcal{L}_{\text{phys}}$ 纳入原始 SR（即不含$\mathcal{L}_{\text{mode}}$的 PCSR）后，可以有效抑制过度波动，确保所得曲线遵循第 2.2 节所述的物理约束。然而，如图 4b中的子图所示，曲线仍可能出现轻微过拟合，无法体现单峰孔径分布。相比之下，同时包含物理约束和模式约束的所提出的框架（即含 $\mathcal{L}_{\text{mode}}$的 PCSR）将目标模式数作为额外输入（例如，本例中为 $N_{\text{mode}} = 1$），有效避免了对稀疏实验数据的过拟合，得到的水分保持模型仅呈现一个由凹转凸的拐点。

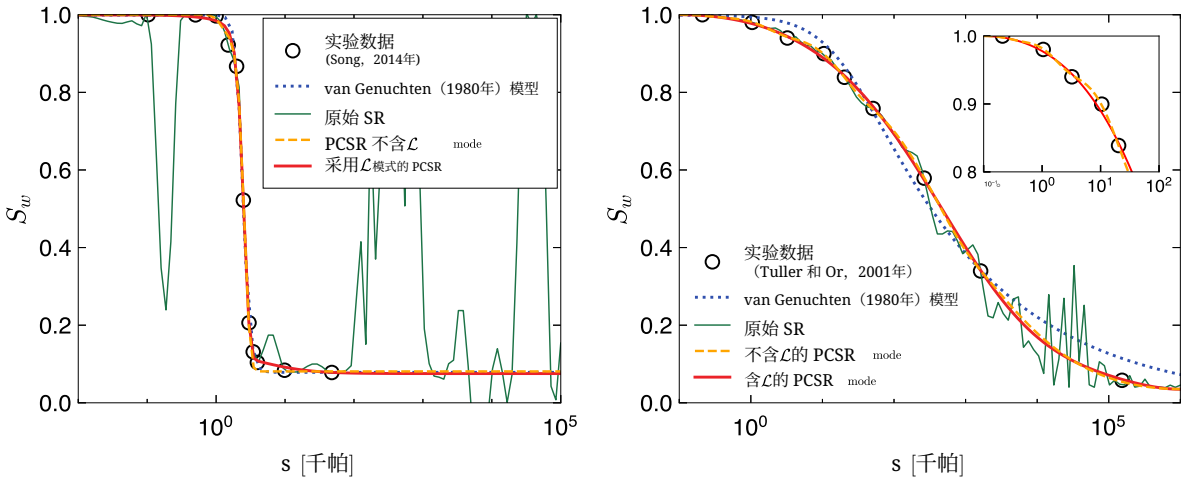


图4：采用所提出的方法从实验单峰水分保持数据中发现的符号模型，并与 [Van Genuchten、 1980] 模型、原始 SR 以及不含模态约束的 PCSR 进行比较。

表1总结了使用 PCSR（含模式约束和不含模式约束）针对每个单峰水分保持数据集发现的最佳拟合数学表达式，对应于图4a-ˆ4b所示的结果。对于每种情况，我们还给出了训练所得模型呈现的模式数（N_{mode}）、符号表达式的复杂度以及相应的总损失 $\mathcal{L}$。未采用模态损失项训练的模型ˆ ˆ均无法达到 $N_{\text{mode}} = 1$，而纳入 $\mathcal{L}_{\text{mode}}$ 的模型通常能够稳定达到 $N_{\text{mode}} = 1$，由此确认所识别的水分保持曲线为单峰曲线。此外，最佳拟合数学表达式的总损失值足够低，复杂度也低于 150 的最大上限，这表明所提出的方法能够高效地为单峰数据集识别出简洁、符合物理规律且峰数一致的水分保持曲线。

图5 展示了PCSR在 $\mathcal{L}_{\text{mode}}$下发现的六个具有代表性的数学表达式，它们各自呈现出不同水平的复杂度，用以说明其对损失以及所呈现模式数的影响。此处，数据（$\mathcal{L}_{\text{data}}$）损失与物理（$\mathcal{L}_{\text{phys}}$）损失分别以堆叠条ˆ形图中的白色和灰色区段表示，对应左侧主y轴；而所呈现的模式数 N_{mode} 则以右侧次y轴上的红色条形表示。结果表明，随着复杂度水平提高， $\mathcal{L}_{\text{data}}$ 与 $\mathcal{L}_{\text{phys}}$ 均趋于下降，表明随着二叉树中节点数量的增加，所发现的数学表达式能更好地描述持水行为。这凸显了准确性与

Table 1: Discovered closed-form equations corresponding to those illustrated in Figures 4a–4b.

	Model	Discovered expression	$\hat{N}_{\text{mode}}$	Complexity	Loss
(a)	PCSR without $\mathcal{L}_{\text{mode}}$	$S_w = \sin(\sin((\cos(\cos(\exp(\exp((\sin(s) - 0.32) \times 64.88) \times -0.25))) + 0.85) \times 4.51) \times (\cos(\cos(\exp(\exp((\sin(s) - 0.27) \times 64.88) \times -0.25))) + 0.76)) \times 1.00$	2	35	3.751×10^{-4}
	PCSR with $\mathcal{L}_{\text{mode}}$	$S_w = \sin(\sin(\exp(\exp(s \times 97.45 \times s) \times -1.46\text{E}{-}05 + 0.25)) \times 1.44) \times 0.94 + \exp(\cos((\sin(s + s) + s) \times (s + \sin(s + s)) \times \sin(s + s)) \times 2.69) \times 0.005$	1	42	2.198×10^{-4}
(b)	PCSR without $\mathcal{L}_{\text{mode}}$	$S_w = \cos(\cos(\sin(\sin(\cos(s \times (s - 2.12) \times -12.45) \times s) \times -1.70)) \times 0.46) \times (\sin(\cos(\cos(s \times -3.14) + 4.75)) \times 0.75 + 0.55)$	4	31	4.790×10^{-6}
	PCSR with $\mathcal{L}_{\text{mode}}$	$S_w = \cos(\sin(((s + 0.64) \times \sin(s \times 1.88) \times 1.42 + s) \times (s - 0.23))) \times \cos(\sin(s \times 1.88) \times (s + 0.65))$	1	28	3.023×10^{-5}

简洁性。此外，当 $\mathcal{L}_{\text{mode}}$ 被纳入PCSR时，数学表达式始终保持$\hat{}N_{\text{mode}} = 1$，表明模态损失项能够有效地使所发现的水分保持曲线保持预期的单峰形状。

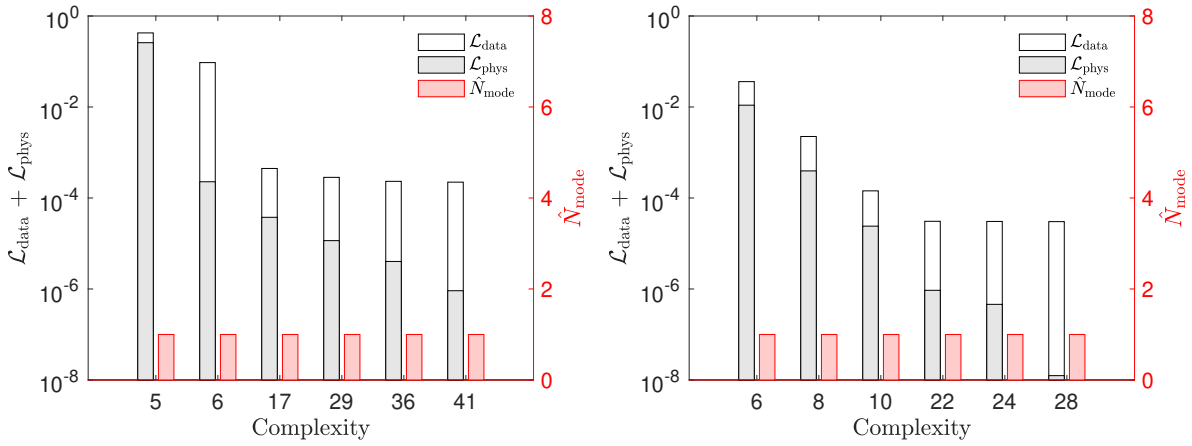


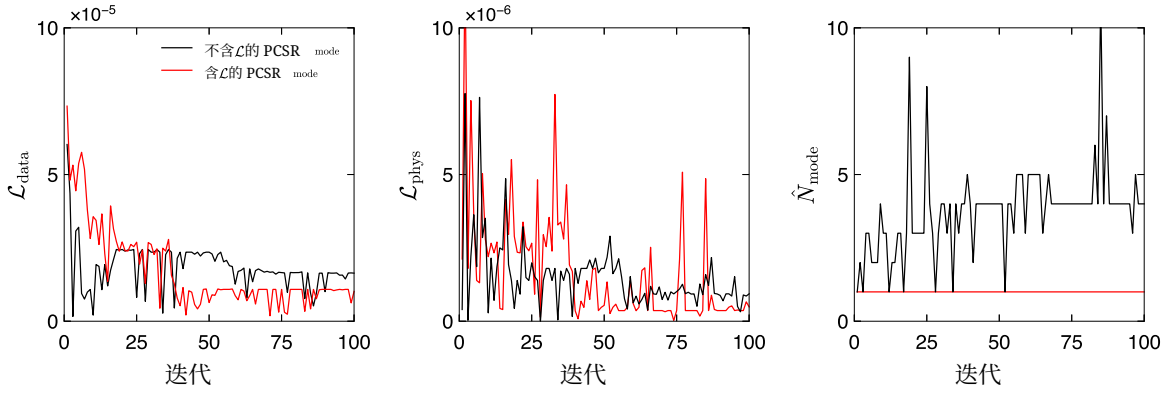
图5：数据损失分量与物理损失分量，以及PCSR在 $\mathcal{L}_{\text{mode}}$ 下以不同复杂度水平最优拟合图4所示数据时所呈现的模式数。

我们还比较了在有无 $\mathcal{L}_{\text{mode}}$条件下训练的PCSR模型的学习曲线，如图6所示。图6a 和6b各包含三个子图，展示 $\mathcal{L}_{\text{data}}$、 $\mathcal{L}_{\text{phys}}$的演变，以及不以$\hat{}\mathcal{L}_{\text{mode}}$、而以 N_{mode} 在训练迭代过程中的变化。黑色实线表示不含模态损失项的PCSR学习曲线，红色实线表示包含形状引导模态损失项的PCSR结果。在两种模型中， $\mathcal{L}_{\text{data}}$ 和 $\mathcal{L}_{\text{phys}}$ 均随训练推进呈下降趋势，但由于符号回归的进化特性，可观察到一些噪声。尽管如此，总体损失水平已足够低，模型之间的差异可以忽略。两种$\hat{}$模型之间的差异可见于训练过程中 N_{mode} 的变化。在不含 $\mathcal{L}_{\text{mode}}$ 的情况下训练的PCSR无法达到单峰水分保持曲线的目标模式数，往往通过提高$\hat{}$中的表达式复杂度来试图降低数据损失和物理损失。这导致过拟合，并使 N_{mode} 显著偏离规定值，如图6a 和6b所示。相比之下，由带有模态损失项的PCSR$\hat{}$获得的候选表示在整个训练过程中始终得到 $N_{\text{mode}} = 1$ ，同时降低另外两个损失分量的值。

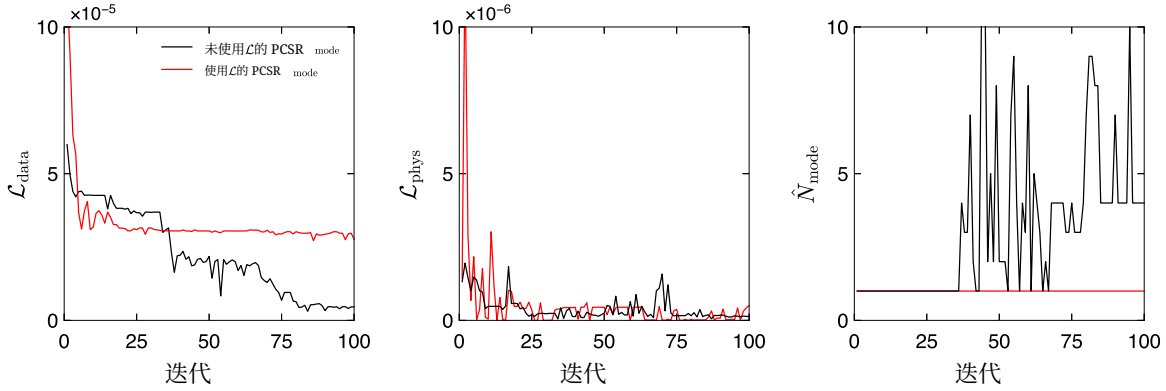
结果表明，在符号回归训练中同时引入物理损失项和模态损失项，能够发现可描述单峰水分保持行为的闭式数学表达式。与其他方法相比，我们的 PCSR 可避免过拟合和伪峰，始终得到符合物理规律的模型，同时保持指定的模式数。在证明所提出方法对单峰情况有效后，接下来我们将考察其在更复杂的多峰水分保持数据集上的表现。

3.2 多峰水分保持曲线

本节首先考察使用双峰水分保持数据（$N_{\text{mode}} = 2$）训练的 PCSR 模型，采用四个从文献中收集的不同训练数据集。如图7所示，这些数据集被收集来测试我们的 PCSR 模型的可扩展性和适用性，因为它们的曲线形状差异显著，且覆盖不同的吸力范围。这些数据集分别为：（1）含砾粉质砂和（2）级配良好的含砂砾石 [Li 等人， 2014]，（3）水稻土 [Zhai等人， 2017], 以及（4）90% 砂和 10% 高岭土的混合物 [Zhou 等人， 2017]。对于符号回归模型训练，各数据集的映射参数设置如下：（1） $s_{\text{min}} = 10^{-2}$, $S_{w,\ \text{max}} = 1.0$, $s_{\text{res}} = 10^{5}$，以及 $S_{w,\ \text{res}} = 0.1$；（2） $s_{\text{min}} = 10^{-1}$,$S_{w,\ \text{max}} = 1.0$, $s_{\text{res}} = 10^{6}$，以及$S_{w,\ \text{res}} = 0.1$；（3） $s_{\text{min}} = 10^{-3}$, $S_{w,\ \text{max}} = 1.0$, $s_{\text{res}} = 10^{6}$，以及 $S_{w,\ \text{res}} = 0.35$；以及



(a)



(b)

图6：PCSR 不使用和使用$\mathcal{L}_{\text{mode}}$ 图中所示数据4的损失分量学习曲线及预测模态数。

(4) $s_{\min} = 10^{-2}$, $S_{w,\max} = 1.0$, $s_{\text{res}} = 10^5$和$S_{w,\text{res}} = 0.05$。

图7 展示了使用 PCSR 发现的最佳拟合持水模型，包括使用和不使用 $\mathcal{L}_{\text{mode}}$的结果，并与实验结果以及使用 [Durner， 1994] 模型拟合的曲线进行了比较。值得注意的是，此处使用的符号和曲线的颜色与样式，与前文图4中所述完全相同。使用原始 SR 发现的水分保持曲线（绿色实线）基于给定数据集给出了相当准确的预测，但存在明显的显著波动，与单峰结果（图4）类似。这些问题可能源于仅关注最小化给定数据点与模型预测之间的差异；如 PCSR 不使用 $\mathcal{L}_{\text{mode}}$ 所发现的持水曲线（黄色实线）所示，引入物理约束即可有效解决这些问题。尽管曲线拟合有所改善（见表2 和3），但 PCSR^在不使用模态损失项时发现的数学表达式，会高估 N_{mode}的值——具体而言，图7a–7d所示曲线的值分别为 3、9、4 和 12——因此无法匹配目标模式数 $N_{\text{mode}} = 2$。相比之下，所提出的框架（即使用 $\mathcal{L}_{\text{mode}}$的 PCSR）成功发现了具有所^需双峰形状的持水曲线，得到 $N_{\text{mode}} = 2$，由此表明其与目标模态数精确吻合。

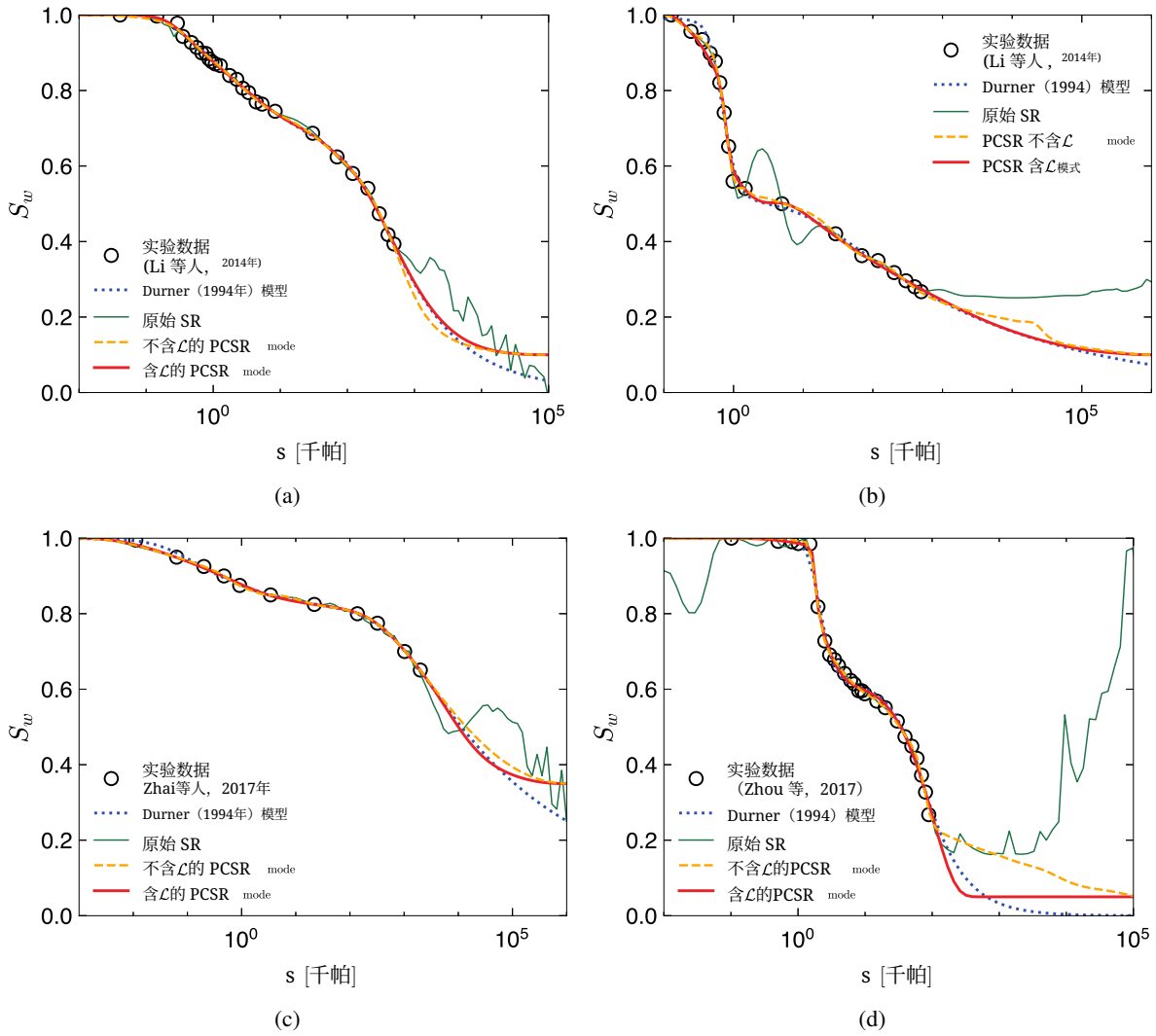


图7：使用所提出的方法从实验双峰水分保持数据中发现的符号模型，与 [Durner， 1994]模型、原始 SR 以及无模态约束的 PCSR 进行比较。

随后，图8 展示了我们的 PCSR 框架在另外两个曲线形状更复杂的持水数据集上的应用，具体而言，这两个数据集分别具有 $N_{\text{mode}} = 3$ 和4。由于这些多模态持水情形的实验数据并不易得，因此使用 [Durner， 1994]模型合成生成了这两个数据集。

尽管合成数据集经过设计，能够清晰展现曲线的多模态特征，但使用原始 SR 发现的水分保持曲线会产生带噪声的预测，且不满足物理约束。相比之下，无论是否包含 $\mathcal{L}_{\text{mode}}$，PCSR 模型生成的水分保持曲线在视觉上都相似。然而，如图8b所示，未施加模态约束的水分保持曲线仍无法准确捕捉进气压力。此外，不纳入模态约束会导致发现的水分保持曲线与目标模式数不符，如表 4所示。相比之下，纳入 $\mathcal{L}_{\text{mode}}$ 后，PCSR 能够可靠地发现物理上合理的符号表达式，不仅能很好地拟合给定的数据点，还能符合目标模式数。

本节的比较分析表明，加入模态损失项的 PCSR框架能够有效应对基于稀疏数据集对多峰水分保持曲线进行建模所面临的挑战。采用此方法发现的水分保持曲线能够始终预测正确的模式数，并生成符合物理规律的符号表达式，无需显式进行参数识别，也不需要预先设定数学形式。这些结果凸显了所提出的 PCSR框架在模拟具有复杂多尺度孔隙结构材料的持水行为方面的有效性。

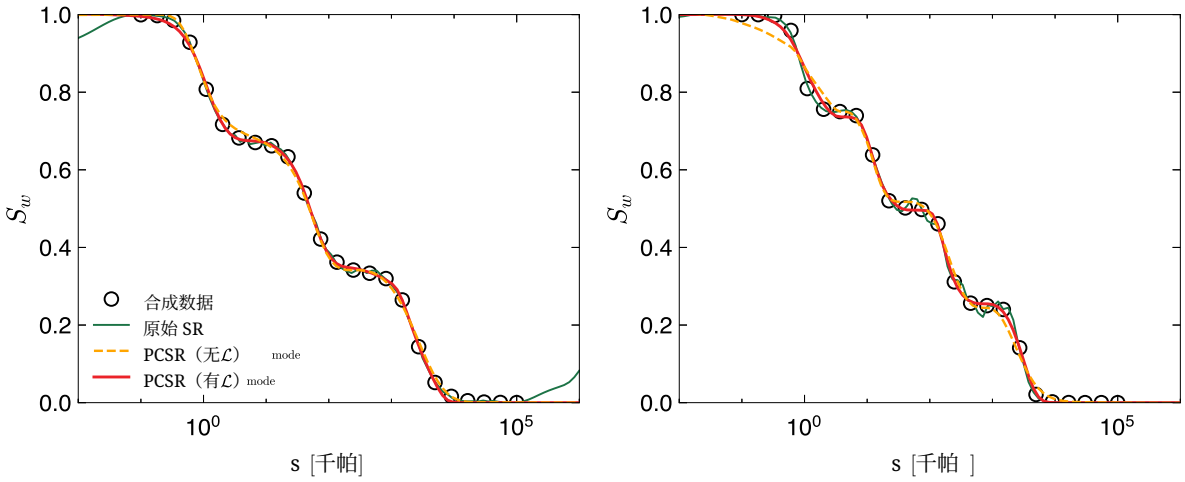


图 8：从合成生成的多模态持水数据中发现的符号模型，与 [Durner、 1994]模型、原始符号回归以及未施加模态约束的 PCSR 进行比较。

4 结论与未来展望

在本研究中，我们提出了一个物理约束符号回归（PCSR）框架，旨在直接从稀疏实验数据中发现水分保持曲线的闭式方程。通过将物理一致的水分保持曲线的数据驱动发现视为多目标优化问题，我们利用由多个物理约束项构成的目标函数，对符号回归任务进行了形式化。假设目标材料孔径分布的模式数事先已知，我们的框架还加入了模态约束，以限制所学习函数中由凸转凹的拐点数量，从而有效引导回归过程。我们在单峰和多模态水分保持数据集上对 PCSR 框架进行了严格评估。结果表明，加入这些额外约束后，发现的数学表达式不仅能够准确拟合实验数据，还能以热力学一致的方式描述持水行为，同时保留所需的模式数。这一优势在原始

Table 2: Discovered closed-form equations corresponding to those illustrated in Figures 7a-7b.

	Model	Discovered expression	$\hat{N}_{\text{mode}}$	Complexity	Loss
(a)	PCSR without $\mathcal{L}_{\text{mode}}$	$S_w = \cos(\sin(\sin(\sin(s \times 1.01 + s + \sin(\sin(\exp((s \times 2.72 + s) \times 0.23) + \sin(s \times (s + \sin(s) + (s \times 2.70 + s) \times 0.27 + s + s + \sin(\sin(s) + (s + s + \sin(\sin(\sin(\sin(s + s + s \times (s + 1.18) - 0.51)) + s + s + \sin(s) - 2.24 + s) + s + s \times s + s + s + \sin(s + s)) - 0.53) \times \exp((s \times 2.73 + s \times 0.27) \times -0.23 + 0.85) + s + s + s - 1.59) - 0.78)))) - 1.41) + s - 0.55)) + \exp(\cos((s + 0.20) \times -0.57) + 1.38) \times 0.03 + 0.48) \times 1.00$	3	137	4.555×10^{-5}
	PCSR with $\mathcal{L}_{\text{mode}}$	$S_w = \sin(\exp(\sin(\exp(\cos(s \times \sin(s + s + s + 1.74)) + 0.93)) + \cos(\sin(\exp(\cos(\cos(\exp(\cos(\exp(\cos(\sin(s) + \exp(\sin(s)) + s)))) + s + \sin(\exp(\cos(\exp(\cos(\cos(\cos(\exp(\cos(\cos(s \times (s - 1.06 + s - 0.19)) + \exp(\cos(\sin(\exp(\cos(s \times (s + 0.92 + s \times 1.94)) + 0.92)))) + s))) + \cos(\cos(\sin(s) + (s + 0.68) \times s \times 1.97))) + \exp(\cos(s + \sin(\sin(s)) + s + 0.50)) + \exp(s \times s))))) + \exp(\cos(\sin(s)) + 0.86)))) + \sin(\exp(\cos(\sin(s \times (s + 0.78) \times \cos(\sin(s + 2.95)))) + 0.88)))) \times (s + s \times 1.95\backslash\text{left}())) - 0.19) \times 0.35)$	2	141	4.407×10^{-5}
(b)	PCSR without $\mathcal{L}_{\text{mode}}$	$S_w = \cos(\sin(\sin(s \times 1.55) + \sin(\sin(\sin(s)) \times 2.36) + (\sin(\sin(\sin(s \times s \times 1.68 + s) + (s + \sin(s) + s) \times 3.10 + s - 1.02)) + s) \times (\sin((\sin(s) + \sin(s) + s \times 0.97 + \sin(\sin(\sin((s + s + s) \times 2.05 + s - 1.29) + \sin(s - 1.30 + (\sin(s + s) + 2.04) \times ((s + s + s) \times 1.04 + s))) \times (\sin(s + s + (s + s + s + \sin(s)) \times 1.42) + s + \sin(\sin(s + s))))) \times 1.36 + s + \sin(s + s + \sin(\sin(s)))) + \sin(s) + s))) \times \cos((\sin(\sin(s \times 1.02 + s)) + \sin(s)) \times 0.97)$	9	143	2.450×10^{-5}
	PCSR with $\mathcal{L}_{\text{mode}}$	$S_w = \cos(\cos(\sin(\sin(s \times 1.21) + \sin(\sin(s + s)))) \times (\sin((s + \sin(s \times 1.37) + \sin(s + s)) \times \cos(\sin((\sin(\sin(\sin((\sin(s) + s + s - 0.45) \times 1.07) \times \exp(\sin(\sin(\sin(\sin(s + s + s - 0.45)) \times -1.70) \times \exp(\sin(\sin(\sin((s + s + 1.07) \times 3.96) + s \times 1.02 + \sin(s)) \times -1.68) + 1.04)) + 0.99) + s) + s) + \sin(s)) \times -1.73 + \cos(\cos(s))))) + \sin(\sin(s + s)) + \sin(s))) \times 1.20 - 0.20$	2	108	1.862×10^{-4}

Table 3: Discovered closed-form equations corresponding to those illustrated in Figures 7c-7d.

	Model	Discovered expression	$\hat{N}_{\text{mode}}$	Complexity	Loss
(c)	PCSR without $\mathcal{L}_{\text{mode}}$	$S_w = \sin(\exp(\sin(\sin(\sin(\sin(s \times s)) \times (s + \sin(\sin(s)) \times (s + \sin((s + s) \times 1.12)))))) - 0.03) + \sin(s + \sin(\sin(\cos(s \times 1.69) \times s \times (s + \sin(\sin(s + s \times s + \sin(\sin((s + \sin(s) + 1.48) \times 1.70 \times (\sin(s - 0.55) + \exp(s)) \times \cos(s) \times (\exp(\sin(s \times 1.00 + s + s)) - 0.52)) + \sin(s \times (s + 4.81)) \times -0.31))) + 0.27) \times s \times (\sin(\exp(\sin(s \times (s + s + 1.41)))) - 0.53) \times (s + s + 1.43) + s))) \times \sin(s) \times -0.91) \times 1.21$	4	123	1.812×10^{-5}
	PCSR with $\mathcal{L}_{\text{mode}}$	$S_w = \sin((\sin(s) + (s - 2.02) \times \cos((\cos((s + s \times s \times \cos(s) \times 0.63) \times (s \times -1.01 + \cos(\cos(s \times 1.00) \times \sin((s + (\sin(s + s \times 0.99) - 4.27) \times -1.02) \times (s + \sin(s + s - 0.28)))))) - 1.08) \times (\exp(s) + s))) \times (\cos(\sin((\cos(\cos(\sin((s \times \sin(s) + \sin(s) + s + s) \times s)) - 0.91) - 1.07) \times (s - 4.76)) + \cos((s + (s - 1.44) \times (\cos(s \times s + \sin(s + 0.45)) + 0.05)) \times 1.59 \times s)) - 1.00))$	2	112	3.046×10^{-5}
(d)	PCSR without $\mathcal{L}_{\text{mode}}$	$S_w = (\cos(\sin(\sin(s + \exp((\cos(s \times (s \times s + 0.97 + s) \times s \times s) + 1.75) \times \sin(\exp(\exp(\sin(\exp(s) \times 0.89) + \sin(\sin(\exp(\exp(\sin(\sin(\sin(\sin(\sin(\exp(\sin(\sin(\exp(\sin(\exp(\sin(s + 1.26 + (\sin(\exp(s - 0.42) \times \sin(s)) + \sin(s) + s) \times \sin(s \times 0.71)) + \sin((\sin(\sin(\sin(s) + \exp(s))) + s) \times 2.27) \times 1.00)) + 1.21))) + \sin(\exp(s + s \times 1.26))) \times 0.43 + s + 0.36)) + 1.20) + \sin((s + \sin(\sin(s + \exp(\sin(s))))) \times 2.40) \times 1.13)))) + 1.37)))))) \times (s - 0.65)) + 0.59) + 0.39) - 0.18) \times 1.23 - 0.001$	12	134	3.321×10^{-5}
	PCSR with $\mathcal{L}_{\text{mode}}$	$S_w = \sin(\exp((s \times -0.88 + \exp(\sin(s \times s \times s \times 1.69 + s \times (s + s) \times \exp((s \times 0.47 + \exp((s \times s \times -77.22 + 3.26) \times 6.57 + \sin(s) \times (((s + 6.57) \times 1.11 + \cos(s)) \times (s \times s \times -77.22 + 3.26) + \cos(s \times s \times -0.52 \times s) + \sin(s + (s + s) \times -18.56) - 17.47) \times 1.25) + s) \times s \times \sin(s) \times s \times -1.00 + s \times \exp((\sin(s) \times (s - 4.27) - 0.12) \times s \times 1.77) \times s \times -17.47 \times (s \times 4.05 + 0.87)) \times \exp(s \times -1.77) \times 26.08 - 0.40) \times 7.98) \times -0.50 \times s) \times s \times 3.41 + 0.37)) \times 1.01$	2	139	1.207×10^{-4}

Table 4: Discovered closed-form equations corresponding to those illustrated in Figures 8a-8b.

	Model	Discovered expression	$\hat{N}_{\text{mode}}$	Complexity	Loss
(a)	PCSR without $\mathcal{L}_{\text{mode}}$	$S_w = \cos((\sin((\sin(s) + (\cos(\sin((\cos(s) + s + 0.36) \times \sin(s)) + 0.38) + 0.07) \times \cos((\cos(\exp(s)) + 1.10) \times (\cos(s \times (\cos(\cos((\cos(s \times 0.39 + (s + 2.81) \times \cos(s + \sin((\cos(\cos(\sin(s)) + 2.02) + 3.12) \times s))) + 0.13) \times ((\cos(\exp(\cos(s \times 1.94))) \times \cos(s \times \cos((\cos(\cos(\cos(s)) + 2.02) + 3.12) \times (s + 0.36) + 0.05)) + 2.31) \times (s \times \cos(\exp(s)) + 1.00) - 0.36))) + 2.33) \times (\exp(\sin(\sin(s))) + 1.05)) + s \times 0.89 + 0.98))) \times (\cos(s) + 1.40)) + 0.69) \times \sin(s + s \times 0.49) + 0.04)$	6	127	1.026×10^{-4}
	PCSR with $\mathcal{L}_{\text{mode}}$	$S_w = \exp(\exp((\sin(\exp(s) \times (s - 0.21) \times 2.31) + \sin(\sin(\sin((\exp(\cos(s) \times s) + s) \times \exp(\sin(\sin(s \times 0.79))) + s \times 5.17) - 4.25) + s \times 3.30)) \times -2.64 + \sin(\sin(s))) \times s \times (s \times (\sin(s) + \cos(\sin((s + \cos(\sin(((\sin(s \times 3.29 + \exp(s \times 0.52) \times \exp(s + 1.00) + \sin(\sin(s \times 5.15 - 2.70))) + 0.96) \times -2.63 + \sin(\sin(\sin(s))) + 1.55) \times s)) + 0.20) \times \sin(s \times 5.19 + ((s + \exp(\sin(\sin(s)))) \times \exp(\sin(s \times (s + 0.27))) + s \times 3.30) \times 1.02 - 4.21)))) + s + s \times s) \times -0.87) \times 1.00$	3	136	3.861×10^{-5}
(b)	PCSR without $\mathcal{L}_{\text{mode}}$	$S_w = (\cos(\cos(\cos(s + s \times \cos(s) \times ((\cos(\exp(s) \times s \times \exp(s \times s) \times 0.12) - 0.29) \times \cos(\cos(\cos(\exp(\cos(\cos(\sin(\cos(s \times \ln(\cos(\sin(s \times \exp(s \times 0.44 + \cos(s \times (s \times 2.05 \times s \times -0.12 + 0.28) \times 3.90))) + s))))) + 1.00 + \exp(\exp(s \times 0.45 + \cos(\cos(s) \times 4.13 \times (s \times 1.29 - 0.19))) \times 0.94 - 0.31)))) \times -0.53 + \ln(\cos(\sin(s \times \exp(\cos(s \times (s \times \exp(\cos(s)) \times s \times -0.13 + 0.29) \times 3.91) + s \times 0.43)) + s)))) + \sin(s))))) - 0.54) \times 3.17$	8	123	5.784×10^{-4}
	PCSR with $\mathcal{L}_{\text{mode}}$	$S_w = \exp(\exp((s + \cos(\exp(s \times s \times 0.67 + 1.56)) \times -0.38) \times 7.37) \times \sin(\sin(s) \times \sin(s \times \cos(\exp(\cos((\cos(\exp(\cos(\exp(s \times \sin(\cos(\exp((\cos(\cos(\cos((\exp(s \times 1.98) \times s \times 1.98 + s) \times s)) + (\cos(s) + \cos((s + s) \times -2.66)) \times (\exp(\exp(\cos(s \times s))) + s)) + s \times \cos(\exp(\cos(\cos(\exp(s \times \exp(s \times s + s) \times (s - 0.17) \times s) \times -1.38 - 1.10)))) \times 3.44) \times s + 0.56) \times (s - 1.10))) \times s + \exp(s \times \sin(s)))) \times -0.58)) + \cos(\cos(\exp(\cos(\cos(\exp(s))) + \sin(s) + 0.21)) \times \exp(\exp(\sin(s))))) \times (s - 1.11)))))))$	4	139	1.361×10^{-4}

符号回归模型往往会过拟合，无法重现正确的曲线形状，尤其是在多模态行为复杂时。由于发现的模型以闭式方程的形式表达，因此可以像传统水分保持模型一样，直接集成到现有的水力模拟工作流程中，从而有望改善非均质且数据稀疏条件下的模拟效果。本研究侧重于多孔介质持水行为的元建模，而非解释发现的表达式中各个参数或组成部分的物理意义；未来研究可以探索嵌入策略，以建立符号结构与物理性质之间的联系。未来工作还将致力于扩展 PCSR，使其能够考虑滞后水分保持行为，并探索其在更广泛的岩土工程和环境问题中的应用，同时提高符号回归器的效率。

开放研究

本研究期间生成和/或分析的源代码及数据集已公开存储于 ZenodoKim 和 Suh [2026]。

致谢

本研究获得韩国政府（MSIT）资助的韩国国家研究基金会（NRF）项目经费（编号：RS-2024-00360509），以及凯斯西储大学的创业启动基金支持。

附录A. 物理约束对所发现符号表达式的影响

本附录考察引入物理约束如何影响由符号回归（SR）发现的闭式方程。本研究考虑的三项约束——单调性、极限约束和有界性——在无噪声和含噪设置下均进行了评估，并与无约束（原始）SR进行比较。在每种情况下，均由专门设计用于突出相应约束效果的基准方程生成数据集。

A.1 单调性约束

为考察单调性约束的效果，本示例考虑如下基准方程：

$$y(x) = x + \sin(Ax) + \epsilon, \tag{A.1}$$

其满足$\mathrm{d}y/\mathrm{d}x \geq 0$ 对于 $A = 1$，但包含$\mathrm{d}y/\mathrm{d}x < 0$当 $A > 1$时的区域。此处， ϵ 表示测量或过程不确定性的加性噪声项。通过选取 $A = 1.1$ 和 $\epsilon = 0$，我们生成了均匀采样数据（图 A.1a）。随后将原始 SR 和带单调性约束$\mathrm{d}y/\mathrm{d}x \geq 0$ 的 SR 用于识别最拟合数据的数学表达式。

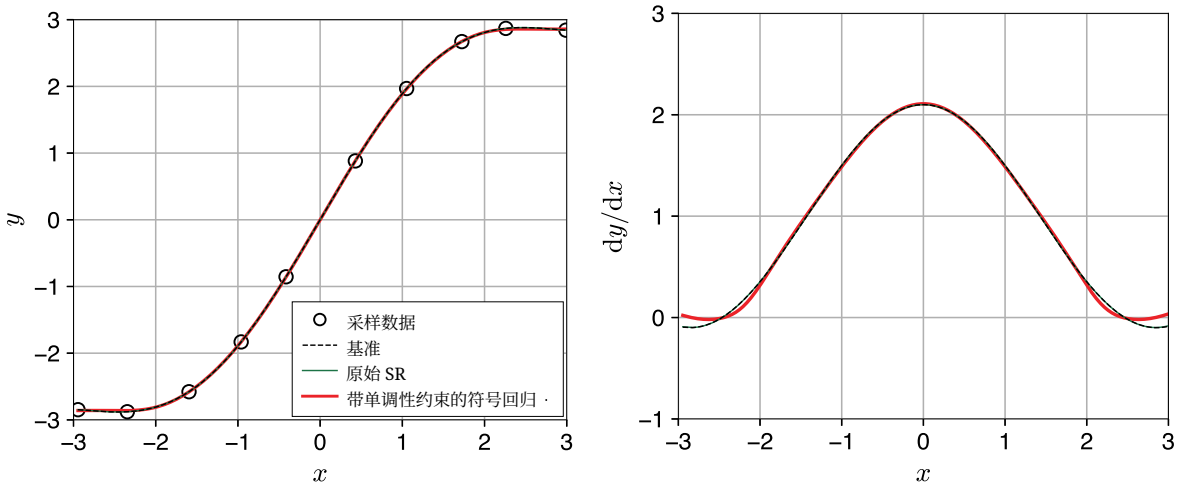
为评估系统结构识别的稳健性，我们有意限制了搜索空间：二元运算符仅限于 $\{+, \times\}$，一元运算符仅限于 {指数函数}，因此排除了基准方程中包含的三角函数。如表A.1所示，原始 SR 和采用单调性约束的 SR 都通过生成由嵌套指数项构成的复杂表达式来弥补这些限制。

尽管原始 SR 和采用单调性约束的 SR 所生成的曲线在视觉上都近似于基准函数，但二者的导数行为存在显著差异。如图 A.1b所示，原始 SR 模型存在负斜率区域，违反了单调性。相比之下，受约束的 SR 模型在整个定义域内都保持非负导数。重要的是，尽管符号表示存在差异，受约束的模型仍正确捕捉了预期行为。这表明，发现的表达式存在差异，源于运算符限制下符号回归的灵活性，而非所学习模型的不稳定性或不可靠性。

引入高斯过程噪声后（$\epsilon = 0.05$；图 A.2），原始 SR 会更积极地适应含噪数据，导致负导数区域扩大。相比之下，采用单调性约束的 SR 严格保持单调行为，有效抑制了噪声导致的违反单调性现象。

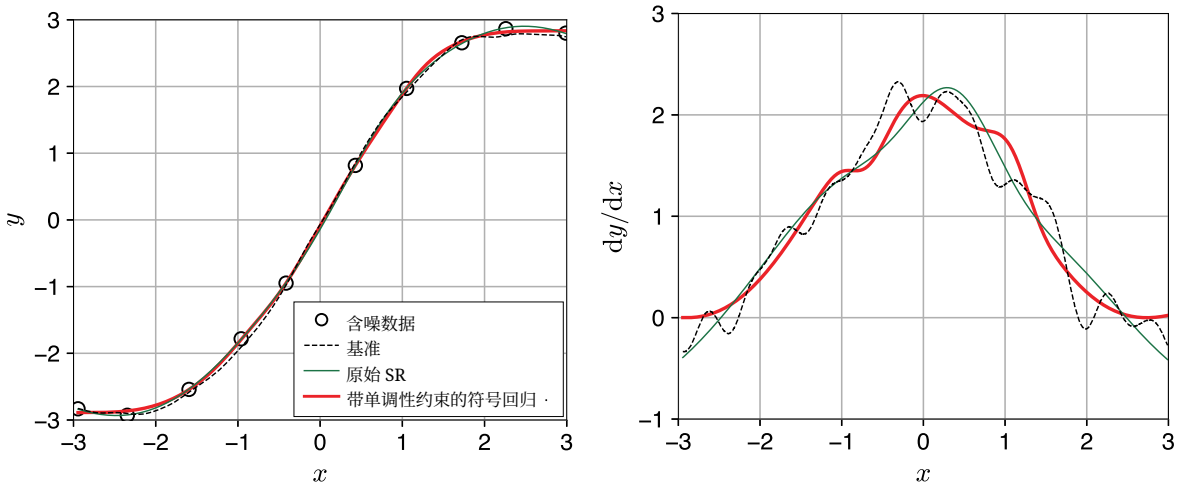
表 A.1：最拟合由式（A.1）中基准表达式所生成数据的发现的方程，对应于图 A.1 所示。

模型	**发现的表达式**
原始 SR	$y = 2 \times x \times (1.5181 \times \exp(x \times (-0.0948 \times x + 6.829 \times 10^{-6})) + (0.0327 \times x \times x - 9.953 \times 10^{-6} \times x - 0.4680)) + 1.5125 \times 10^{-6}$
带单调性约束的符号回归。	$y = ((\exp(((\exp(((x \times (-0.176)) \times (x + (\exp((x \times x) \times 0.469) \times (x \times 0.434)))) + 0.465) + ((x + x) \times x)) \times (-0.108)) + 0.502) + 0.718) \times x) - 0.001$



(a) (b)

图 A.1：无噪声情形下单调性约束（$\mathrm{d}y/\mathrm{d}x \geq 0$）的影响。(a) 基准方程和采样数据（黑色空心圆），与原始 SR（绿色实线曲线）以及带有单调性约束的 SR（红色实线曲线）的对比。(b) 相应方程的一阶导数，展示了原始 SR 中的单调性违反，以及约束模型对单调性的保持。



(a) (b)

图 A.2：高斯过程噪声存在时单调性约束（$\mathrm{d}y/\mathrm{d}x \geq 0$）的影响。(a) 含噪采样数据以及原始 SR 和带单调性约束的符号回归学习得到的方程。(b) 相应方程的一阶导数，说明原始 SR 中的单调性违反以及约束模型中单调性的保持。

A.2 极限约束

为研究极限约束的影响，我们在从以下基准中采样得到的数据集上，分别训练了原始 SR 和带有极限约束的 SR:

$$y(x) = B(x + C) + \cos\left((D(x + C)\right) + \epsilon. \tag{A.2}$$

此示例考虑了对于 $x < 0.2$ 的极限约束 $y = 0$，同时选择 $B = 0.4$，$C = 0.4$，$D = 3.0$，并且 $\epsilon = 0$使得来自公式（A.2）的采样数据有意违反了施加的条件。

如图 A.3所示，原始 SR 能够准确拟合采样数据，但会为$x < 0.2$产生非零值，从而违反极限约束。相比之下，带有极限约束的 SR 会严格执行此条件，同时在其他区域保持良好的数据拟合效果。

对于采样数据包含高斯过程噪声（$\epsilon = 0.2$）的情况，如图A.4所示，原始 SR 会强烈适应噪声波动，导致明显违反施加的极限约束。相比之下，带有极限约束的符号回归能有效抑制极限区域内的过拟合，确保发现的表达式符合施加的条件。

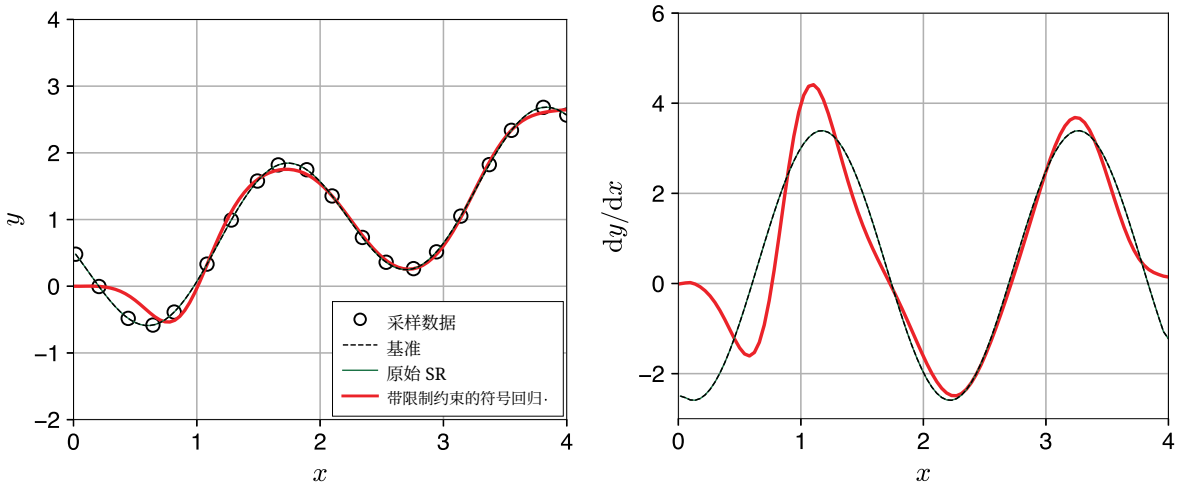
A.3 有界性约束

本节通过考察以下基准方程，探究有界性约束的影响:

$$y(x) = E(x^2 - F)(x^2 - G) + \epsilon. \tag{A.3}$$

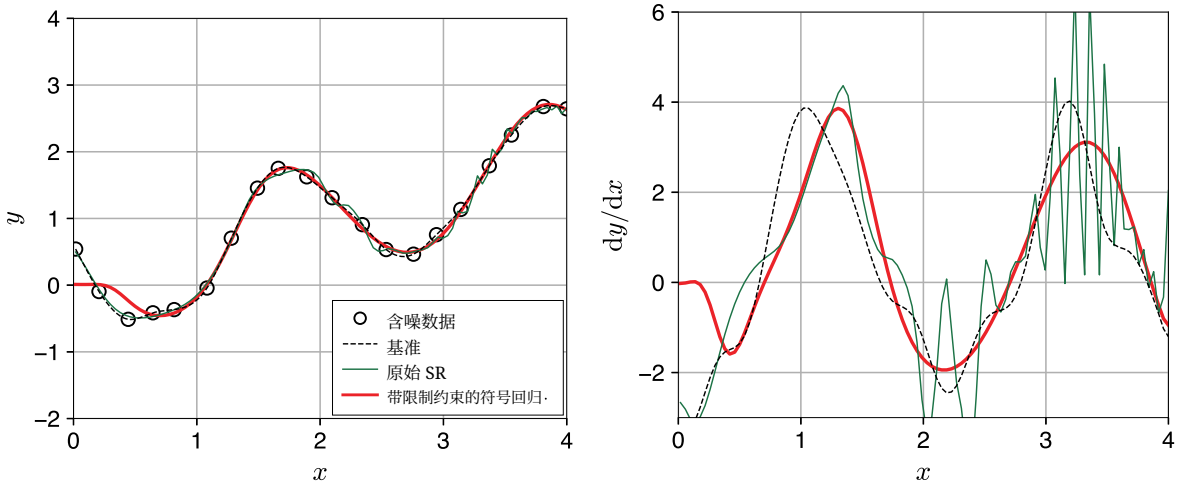
基于此设定，我们考虑有界性约束 $0 \leq y \leq 1$，同时特意选择参数（$E = 0.125$、$F = 4.0$、$G = 2.25$、$\epsilon = 0$），使得由此得到的基准方程——以及采样数据——违反该条件。

如图A.5a所示，原始 SR 会发现最符合采样数据的表达式，且其越界行为与采样数据一致；而采用有界性约束的符号回归得到的方程取值范围为 $0 \leq y \leq 1$，满足施加的有界性约束。向数据集中加入高斯过程噪声后，也观察到类似的模式（$\epsilon = 0.5$）。如图A.5b所示，原始 SR 仍会重现由噪声放大的越界偏差，而受约束的模型则始终保持施加的界限。



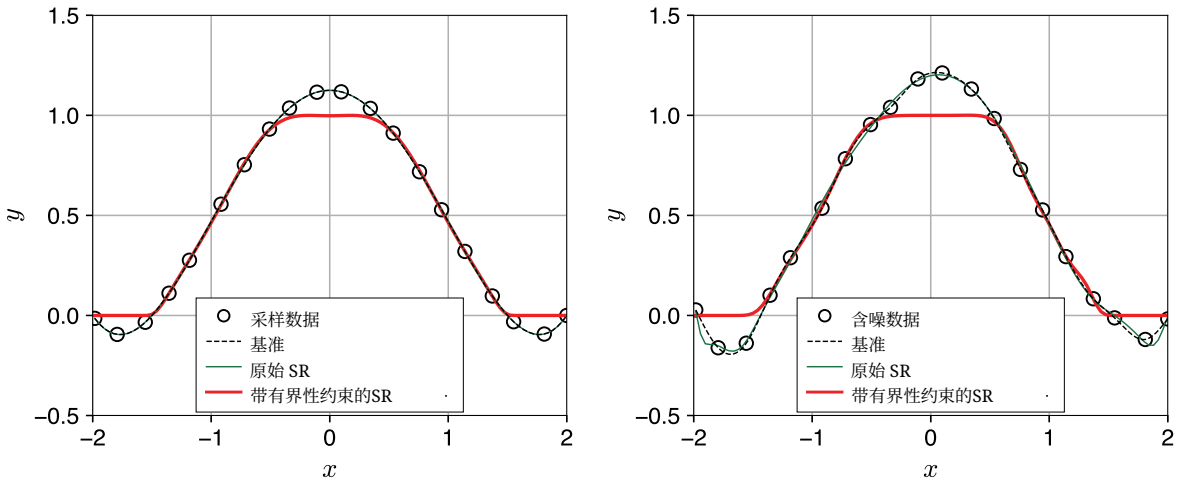
(a) (b)

图 A.3：无噪声情形下极限约束（$y = 0$ 对 $x < 0.2$）的影响。(a) 基准方程和采样数据与原始 SR 及带极限约束的符号回归的比较。(b) 相应方程的一阶导数，展示了约束模型对极限行为的严格执行。



(a) (b)

图 A.4：存在高斯过程噪声时极限约束（$y = 0$ 对 $x < 0.2$）的影响。(a) 含噪采样数据以及由原始 SR 和带极限约束的符号回归学习得到的方程。(b) 相应方程的一阶导数，表明施加极限约束后对噪声的敏感性降低。



(a) (b)

图 A.5：有界性约束（$0 \leq y \leq 1$）在 (a) 无噪声情形和 (b) 存在高斯过程噪声时的效果。

这些结果表明，具有物理依据的约束，尤其是源自热力学原理的约束，可以有效地纳入符号回归框架。通过将此类约束嵌入学习过程，所提出的方法能够直接从数据中实现模型的数据驱动发现，同时防止出现非物理行为，即使存在噪声也是如此。

参考文献

Rishabh Agarwal、Levi Melnick、Nicholas Frosst、Xuezhou Zhang、Ben Lengerich、Rich Caruana 和 Geof- frey E Hinton。神经加性模型：利用神经网络实现可解释机器学习。神经信息处理系统进展，34：4699–4711，2021年。

Dimitrios Angelis、Filippos Sofos 和 Theodoros E Karakasidis。《物理科学中的人工智能：符号回归趋势与展望》。《计算工程方法档案》，30(6)：3845–3865，2023 年。

Bahador Bahmani、Hyoung Suk Suh 和 WaiChing Sun。通过神经多项式方法实现的符号回归发现可解释的弹塑性模型。《应用力学与工程中的计算机方法》，422：116827，2024 年。

Geoffrey F Bomarito、Patrick E Leser、Nolan CM Strauss、Karl M Garbrecht 和 Jacob D Hochhalter。用于符号回归遗传编程中减少代码膨胀和过拟合的贝叶斯模型选择。载于《遗传与进化计算会议论文集》，第 526–529 页，2022 年。

Ronaldo I Borja。剑桥黏土塑性。第五部分：用于部分饱和多孔介质三相变形与应变局部化分析的数学框架。《应用力学与工程中的计算机方法》，193(48-51)：5301–5338，2004 年。

Royal Harvard Brooks. 多孔介质的水力特性. 科罗拉多州立大学，1965年。

Shengze Cai、Zhicheng Wang、Sifan Wang、Paris Perdikaris 和 George Em Karniadakis. 用于热传递问题的物理信息神经网络.《传热学报》，143(6):060801，2021年。

Miles Cranmer。使用 PySR 和符号回归进行科学领域的可解释机器学习。*arXiv* 预印本 *arXiv:2305.01582*，2023年。

Fabricio Olivetti de Franca 和 Gabriel Kronberger。利用等式饱和减少符号回归模型的过参数化。发表于遗传与进化计算会议论文集，第 1064–1072 页，2023年。Wolfgang Durner。针对具有非均质孔隙结构的土壤进行水力传导率估算。水资源研究，30(2):211–223，1994年。Fenglei Fan、Jinjun Xiong 和 Ge Wang。论人工神经网络的可解释性。*arXiv* 预印本 *arXiv:2001.02522*，2(10)，2020年。Delwyn G Fredlund 和 Anqing Xing。土水特征曲线方程。加拿大岩土工程杂志，31(4):521–532，1994年。DV Griffiths 和 N Lu。采用稳态入渗或蒸发的非饱和边坡稳定性分析：弹塑性有限元方法。国际岩土力学数值与解析方法杂志，29(3): 249–267，2005年。Ehsan Haghighat、Maziar Raissi、Adrian Moure、Hector Gomez 和 Ruben Juanes。用于固体力学反演和代理建模的物理信息深度学习框架。应用力学与工程中的计算机方法，379:113741，2021年。Yousef Heider、Hyoung Suk Suh 和 WaiChing Sun。借助自主设计与改进的神经网络实现的离线多尺度非饱和孔隙力学模型。国际岩土力学数值与解析方法杂志，45(9):1212–1237，2021年。Taegu Kim、Tae Sup Yun 和 Hyoung Suk Suh。ChatGPT 能否为岩土工程应用实现有限元模型?国际岩土力学数值与解析方法杂志，2025年。Yejin Kim 和 Hyoung Suk Suh。代码与数据：用于多模态水分保持曲线建模的物理约束符号回归（版本 1.2） [软件]。*Zenodo*，页面 https://doi.org/10.5281/zenodo.18760423，2026年。Ken’ichirou Kosugi。用于土壤持水的三参数对数正态分布模型。水资源研究，30(4):891–901，1994年。John R Koza。作为通过自然选择对计算机编程的手段的遗传编程。统计与计算，4:87–112，1994年。Xianmin Li、Jin Hui Li 和 Li Ming Zhang。利用基于物理的参数预测双峰土水特征曲线和渗透率函数。计算机与岩土工程，57:85–96，2014年。Xuhong Li、Haoyi Xiong、Xingjian Li、Xuanyu Wu、Xiao Zhang、Ji Liu、Jiang Bian 和 Dejing Dou。可解释深度学习：解释、可解释性、可信度及其他。知识与信息系统，64(12):3197–3234，2022年。Lennart Linden、Dominik K Klein、Karl A Kalina、Jörg Brummund、Oliver Weeger 和 Markus Kästner。神经网络与超弹性：物理约束指南。固体力学与物理学杂志，179:105363，2023年。Siyan Liu、Arsalan Zolfaghari、Shariar Sattarin、Amirmasoud Kalantari Dahaghi 和 Shahin Negahban。神经网络在多孔介质多相流中的应用：毛细压力和相对渗透率曲线预测。石油科学与工程杂志，180:445–455，2019年。Ziming Liu、Yixuan Wang、Sachin Vaidya、Fabian Ruehle、James Halverson、Marin Solja˘ci´c、Thomas Y Hou 和 Max Tegmark。KAN：Kolmogorov-Arnold网络。*arXiv* 预印本 *arXiv:2404.19756*，2024年。

Nour Makke 和 Sanjay Chawla。利用符号回归进行可解释的科学发现：综述。人工智能评论，57(1):2，2024。W James Murdoch、Chandan Singh、Karl Kumbier、Reza Abbasi-Asl 和 Bin Yu。可解释机器学习中的定义、方法与应用。美国国家科学院院刊，116(44): 22071–22080，2019。Brenden K Petersen、Mikel Landajuela、T Nathan Mundhenk、Claudio P Santiago、Soo K Kim 和 Joanne T Kim。深度符号回归：通过风险寻求策略梯度从数据中恢复数学表达式。*arXiv* 预印本 *arXiv:1912.04871*，2019。E Priesack 和 W Durner。多模态非饱和导水率函数的闭式表达式。渗流带期刊，5(1):121–124，2006。Maziar Raissi、Paris Perdikaris 和 George E Karniadakis。物理信息神经网络：一种用于求解涉及非线性偏微分方程的正向和逆向问题的深度学习框架。计算物理学杂志，378:686–707，2019。Rabia Saleem、Bo Yuan、Fatih Kurugollu、Ashiq Anjum 和 Lu Liu。解释深度神经网络：全局解释方法综述。神经计算，513:165–180，2022。Michael Schmidt 和 Hod Lipson。从实验数据中提炼自由形式自然定律。科学，324 (5923):81–85，2009。Vijay Kumar Singh、Devendra Kumar、PS Kashyap、Pramod Kumar Singh、Akhilesh Kumar 和 Sudhir Kumar Singh。基于土壤物理性质，使用不同数据驱动算法对土壤渗透性进行建模。水文学杂志，580:124223，2020。Xiaoyu Song 和 Ronaldo I Borja。有限变形范围内非饱和流的数学框架。国际工程数值方法杂志，97(9):658–682，2014。Young-Suk Song。不同相对密度下非饱和砂的吸应力。工程地质，176: 1–10，2014。Hyoung Suk Suh、Chulmin Kweon、Brian Lester、Sharlotte Kramer 和 WaiChing Sun。一个公开可用的、用于水平集塑性的 PyTorch-Abaqus UMAT 深度学习框架。材料力学，184:104682，2023。Hyoung Suk Suh、Jun Young Song、Yejin Kim、Xiong Yu 和 Jinhyun Choo。数据驱动发现可变形多孔介质的可解释水分保持模型。*Acta Geotechnica*，19(6):3821–3835，2024。Moussa Tembely、Ali M AlSumaiti 和 Waleed Alameri。基于深度学习视角预测多孔介质渗透率：从网络建模到直接模拟。计算地球科学，24(4): 1541–1556，2020。Markus Tuller 和 Dani Or。变饱和多孔介质的水力传导率：角形孔隙空间中的薄膜与角流。水资源研究，37(5):1257–1276，2001。Silviu-Marian Udrescu 和 Max Tegmark。AI Feynman：一种受物理启发的符号回归方法。科学进展，6(16):eaay2631，2020。M Th Van Genuchten。用于预测非饱和土水力传导率的闭式方程。美国土壤科学学会期刊，44(5):892–898，1980。Nikolaos N Vlassis 和 WaiChing Sun。针对具有水平集硬化的可解释弹塑性模型，对热力学信息神经网络进行 Sobolev 训练。应用力学与工程中的计算机方法，377:113695，2021。

Nikolaos N Vlassis、Ran Ma 和 WaiChing Sun。计算力学的几何深度学习：第一部分：各向异性超弹性。应用力学与工程中的计算机方法，371:113299，2020。Stefan Wagner 和 Michael Affenzeller。HeuristicLab：一个通用且可扩展的优化环境。载于自适应与自然计算算法：*2005*年在葡萄牙科英布拉举行的国际会议论文集，第538–541页。施普林格，2005年。

Jeannette M Wing. 可信人工智能。*ACM*通讯，64(10):64–71，2021年。

Qian Zhai、Harianto Rahardjo、Alfrendo Satyanaga 等。双峰土水特征曲线对渗透函数估算的影响。工程地质，230:142–151，2017年。

Hu Zhou、Sacha Jon Mooney 和彭新华。通过结合土壤水分保持曲线与X射线计算机断层扫描方法研究双峰土壤孔隙结构。美国土壤科学学会期刊，81(6):1270–1278，2017年。