PROCEEDINGS A

royalsocietypublishing.org/journal/rspa

物理科学符号回归特刊导言

综述

投稿文章

关键词：
符号回归，物理学

通讯作者：
Deaglan J. Bartlett, Harry Desmond,PedroG.Ferreira, Gabriel Kronberger电子邮件：deaglan.bartlett@physics.ox.ac.uk，电子邮件：harry.desmond@port.ac.uk，电子邮件：pedro.ferreira@physics.ox.ac.uk，电子邮件：Gabriel.Kronberger@fh-hagenberg.at

Deaglan J. Bartlett[1]，Harry Desmond[2]，Pedro G. Ferreira[1]，Gabriel Kronberger[3]

[1] 天体物理学，牛津大学，Denys Wilkinson 大楼，Keble 路，牛津 OX1 3RH，英国[2]宇宙学与引力研究所，朴茨茅斯大学，Dennis Sciama 大楼，朴茨茅斯 PO1 3FX，英国 [3] 启发式与进化算法实验室，上奥地利应用科学大学，Softwarepark 11，4232 哈根贝格，奥地利

符号回归（SR）已成为一种强有力的方法，能够从数据中发现可解释的数学关系，为科学发现和高效经验建模开辟了新途径。本文介绍了物理科学符号回归特刊，其灵感源于英国皇家学会于2025年4月举行的研讨会。本特刊收录的论文涵盖多个应用领域，包括自动方程发现、新兴现象建模，以及为计算成本高昂的模拟构建紧凑型模拟器。

导论综述阐述了符号回归的概念基础，将其与传统回归方法进行比较，并概述其在物理科学中的主要应用场景，包括推导有效理论、构建经验函数形式和代理模型。我们总结了若干方法学考量，例如搜索空间设计、算子选择、复杂度控制、特征选择，以及与现代人工智能方法的结合。我们还强调了当前面临的挑战，包括可扩展性、抗噪鲁棒性、过拟合和计算复杂度。最后，我们着重介绍了新兴研究方向，尤其是纳入对称性约束、渐近行为及其他理论信息。总而言之，本特刊中的论文展现了符号回归的迅速进展及其在物理科学各领域日益增长的重要性。

© 作者。由英国皇家学会依据知识共享署名许可协议发布：http://creativecommons.org/licenses/ by/4.0/。该许可允许不受限制地使用，但须注明原作者和来源。

THE ROYAL SOCIETY PUBLISHING

arXiv:2512.15920v1 [cs.LG] 17 Dec 2025

1. 引言：符号回归与英国皇家学会会议

符号回归（SR）正逐渐成为一种强大的计算方法，在推动物理科学各领域研究方面具有巨大潜力。与传统回归技术通过拟合参数来适配预先设定的模型结构不同，这种方法旨在找出最能描述观测数据的底层数学表达式。通过探索浩瀚的可能方程空间，符号回归不仅能够创建高效的预测模型，更有望从经验性证据中发现新的科学定律或关系。

自动寻找能够拟合观测数据的方程，自人工智能研究兴起之初便一直是研究热点。早期用于基于规则发现方程的符号系统和启发式系统包括 Lenat 的 Automatic Mathematician [1], BACON [2], ABACUS [3],或 LAGRANGE [4] ，后者用于代数微分方程。遗传编程（GP）是一种进化算法，于 20 世纪 90 年代引入，并由 Koza 推广 [5]。如今用于符号回归的 GP 实现包括 PySR [6] 和 PyOperon [7]。后来，进化方法与基于规则的方法在 AI Feynman [8],和 AI Descartes [9]等系统中得以结合。依赖深度学习和强化学习的系统包括 EQL [10]和 uDSR [11]。近年来，人们还提出了多种基于端到端学习、Transformer 和基础模型的方法。

符号回归的核心目标是找到一种函数形式 f，使得 $y = f(x_1, x_2, \ldots\ldots, x_n)$ 能够根据给定数据集，最优地描述一组输入变量 $\{x_i\}$ 与输出变量 y之间的关系。与传统回归方法不同，传统方法会预先假定 f 的结构（例如线性结构或特定次数的多项式结构）；符号回归算法则会探索多种数学运算符（例如算术运算、三角函数、指数函数和对数函数）及其组合，以构造候选方程。

符号回归的主要输出不只是一组参数，而是一个或多个明确的数学公式。这种内在的透明性是一项关键优势，因为发现的方程原则上可由人类理解。可解释性使科学家能够深入了解所研究系统的内在机制，根据现有理论验证发现的模型，并有望揭示变量之间此前未知的联系。此外，符号回归模型通过捕捉数据的内在结构，可能展现出更强的泛化能力，尤其是在超出训练数据范围进行外推时。

物理科学领域对符号回归，尤其是方程发现方法的兴趣日益浓厚，这促使我们组织了题为物理科学中的符号回归 的研讨会。该研讨会于2025年4月在伦敦的英国皇家学会举行。与会者包括一批处于机器学习与物理科学交叉领域的研究人员，旨在开发符号回归方法，并利用这些方法直接从实验或观测数据中学习物理定律或简化模型。各领域的应用展现了符号回归的广泛适用性，例如推导金属材料变形过程中的本构定律、加速宇宙学建模，以及根据运动学数据学习暗物质晕轮廓。

在本期特刊中，我们汇集了研讨会报告人的研究论文。这篇导论概述了符号回归的主要用途、不断发展的方法学前沿，以及构成符号回归研究人员未来主要关切的挑战。

2. 用于科学发现的符号回归：对物理定律的探索

符号回归最具雄心的应用，是通过自动化科学发现，直接从实验

或观测数据中提取基本物理定律或新颖的描述性方程。其目标是找到简单而准确的数学公式，捕捉复杂数据集的本质。这与简约原则（即奥卡姆剃刀）相契合，该原则倾向于更简单的解释。符号回归算法通常会采用复杂度惩罚或简洁性先验，引导搜索朝着更精炼、更自然的方程进行。简洁的方程更有可能反映真实的底层关系，而非仅仅是过拟合。这一领域面临诸多挑战，包括数据中的噪声、潜在的观测偏差、方程搜索空间的广阔，以及难以客观界定“简洁性”或“可解释性”。

尽管人们认识到，许多基础原理或许早已确立，因此符号回归有望发现可与广义相对论或量子力学相媲美的全新基本物理定律这一前景有所收窄，但仍存在大量重要机遇。符号回归驱动发现的前沿领域，尤其集中在以涌现现象为特征的领域：复杂行为由许多简单组成部分相互作用而产生，而从微观描述推导宏观定律往往难以实现。物理科学中此类情形的例子包括：

- **天体物理学**：宇宙中存在大量复杂系统，从恒星大气和吸积盘中的等离子体动力学，到星系的形成以及宇宙网的大尺度结构，皆是如此。高度非线性的物理过程耦合了范围极广的尺度，使这类过程极难模拟。符号回归可应用于观测数据（例如光变曲线、光谱、宇宙学巡天）或模拟数据，以发现支配这些现象的新关系，并有望改进恒星演化、星系动力学或系外行星大气性质的模型。

- **凝聚态物理**：该领域充满了复杂的多体系统，它们展现出丰富的涌现行为，例如超导性、拓扑相或复杂的磁有序。符号回归可以帮助研究者根据实验测量（例如输运、光谱数据），或根据详细微观模拟生成的数据，识别能够描述这些宏观性质的有效理论或唯象定律；这些微观模拟过于复杂，无法直接获得解析洞见。

- **工程学**：在复杂系统（例如电动机、化学反应器）的设计中，可以运用符号回归推导方程，根据设计参数预测性能指标；还可以借此寻找材料的本构模型，或为动态过程寻找最优控制策略。

在这些领域中，符号回归未必旨在推翻既有的基本定律，而是要发现特定情境下支配涌现行为的具体数学形式，往往还能为微观物理与宏观可观测量之间建立至关重要的联系。

3. 用于经验建模的符号回归

除了探索基本定律之外，符号回归还是一种非常有效的工具，可用于拟合数据并生成经验公式。在许多科学和工程学领域，首要需求是建立简洁而准确的数学模型，以捕捉观测到的关系，即使该模型并不代表物理定律。符号回归可以生成这类经验模型，它们往往比非参数模型或参数过多的（深度）模型更有洞见，也更易于使用。

在此背景下，符号回归的一项尤为重要的优势是其实现稳健外推的潜力。由于符号回归旨在找出关系的实际函数形式，如果它成功捕捉到底层物理规律的近似形式，所得模型在参数空间中未见过的区域进行外推时，可能比仅基于局部数据模式进行插值的模型更加可靠。这对于在新条件下进行预测，或理解超出实验可及范围的系统行为至关重要。然而，与

所有外推一样，这种能力也必须谨慎对待，因为即使模型在训练数据上表现良好，如果发现的函数形式并不能真正代表底层过程，那么它在该范围之外仍可能表现得极差。简洁性先验和符号回归的最小描述长度原理在此背景下发挥着重要作用。

4. 用于物理系统模拟的符号回归

拟合公式构建的一个重要应用是开发模拟器，也称为代理模型。许多物理现象由基于第一性原理理论（例如量子力学、流体动力学和广义相对论）的复杂、高维且计算密集型模拟来描述。为了进行大规模参数扫描、不确定性量化或实时控制而运行这些模拟，成本可能高得令人难以承受。

符号回归通过为这些复杂模拟生成快速而准确的数学近似，提供了一种极具吸引力的解决方案。这些模拟器以相对简单的方程表示，能够以极高的保真度复现原始模拟的输入输出行为，而计算成本却只需其一小部分。例如，符号回归可用于推导材料性质的解析表达式，而这些性质通常需要通过计算量巨大的量子力学计算才能获得；也可用于创建宇宙学模拟的紧凑模型，或近似求解描述流体流动的偏微分方程。

符号回归所推导的模拟器，其优势不仅在于速度。由于模拟器是显式的数学公式，它能够揭示系统输出对各种输入参数的灵敏度，从而可能发现主导性的物理效应或简化假设，而这些未必能从完整模拟中显而易见。对于快速设计探索、优化和逆问题求解而言，这些对复杂物理学的有效描述极具价值，因为这些任务都需要进行大量模型评估。与神经网络等许多其他模拟工具不同，符号回归的输出是简洁的数学表达式。这些表达式易于用任何编程语言编写，也有望在资源受限硬件（例如嵌入式设备）上运行，例如用于在线控制。

5. 方法论方面

成功应用符号回归需要仔细考虑若干方法论方面。它并非万能灵药，其有效性在很大程度上取决于问题的性质以及搜索的进行方式。

首要关键步骤是策略性地选择问题。当目标是获得一个可能具有可解释性的显式数学模型，且有充分理由相信存在简洁的底层关系时，符号回归最有可能带来显著收益。如果唯一关心的是预测准确性，而可解释性居于次要地位，或者需要考虑大量可能相互关联的输入变量，那么深度神经网络等其他机器学习技术或许更为合适。不过，这些规模大得多的模型需要更多数据量进行训练。符号回归虽然能够处理较小的数据集，但含噪或不足的数据可能会导致错误或过于复杂的方程，并造成过拟合。

所研究物理系统固有的复杂性也会影响符号回归的应用。包含许多相互作用变量的高度复杂系统，可能无法用简单的低维方程来描述。在这种情况下，可能需要进行适当的特征工程、降维，或聚焦于系统中具体且约束明确的方面。遗传编程通过进化选择压力实现隐式特征选择，从而能够自动从高维数据集中识别出与低维方程最相关的输入变量。然而，在必须通过大型模型捕捉大量输入变量，或需要搜索大型数学表达式的情况下，当前符号回归算法的可扩展性也可能成为限制。此时，自动识别函数或构建分层模块化结构方面的新进展值得关注，这有助于生成更精简紧凑的描述。

此外，通过数学构件（变量、常数、算子）及其组合规则来定义和约束搜索空间，是决定符号回归能否成功的关键因素。搜索空间过于受限，可能会排除发现更可能方程的机会；而空间过大，则可能导致搜索在计算上难以实现，或增加找到过于复杂、对数据过拟合的方程的风险。理想情况下，基函数的选择应以关于所建模系统的领域知识为指导。

6. 符号回归不断发展的前沿

符号回归领域不断发展，研究人员持续探索，以提升其能力、应用范围和可靠性。两个尤具前景的未来方向是融入科学先验知识，以及与深度学习、强化学习或大语言模型（LLMs）等新兴人工智能技术协同结合。在这一背景下，符号回归是混合符号、亚符号和神经人工智能系统大有助益的理想应用领域。

符号回归的一项重大进展，是能够超越纯粹的数据驱动方法，将现有科学知识和物理约束直接融入搜索过程。这有助于提高搜索效率，并增强发现的方程的物理合理性。此类先验知识包括：

- 对称性：在候选方程中施加系统已知的对称性（例如平移对称性、旋转对称性和宇称）。
- 守恒定律：要求发现的方程遵循已知的守恒原理（例如能量守恒、动量守恒或质量守恒）。
- 已知的渐近行为或边界条件：引导搜索，使其朝向满足已知极限或条件的函数。
- 不完整或部分理论模型：使用符号回归，在已确立的理论框架中发现缺失项或函数形式。
- 量纲齐次性：确保方程中的所有项都具有一致的物理单位。

明确纳入此类约束，可以有效缩小搜索空间，并引导算法找到更有意义且符合物理规律的解。

将符号回归与大语言模型等基础模型相结合，是一个新兴但极具发展前景的方向。基础模型在海量文本或代码语料库上进行训练，具备出色的自然语言理解、生成乃至初步推理能力。这为增强符号回归带来了多种可能：

- 假设生成：大语言模型可以分析科学文献或问题描述，提出相关的物理变量、可能的函数形式或适用的物理原理，从而帮助确定符号回归的搜索空间并对其加以约束。
- 解释与说明：大语言模型可以将符号回归发现的、往往复杂的数学表达式转化为直观的自然语言解释，使结果更容易被更广泛的科学界受众理解，或帮助提出新的假设。
- 代码生成：大语言模型可以通过生成代码来协助符号回归的实际应用，例如设置符号回归实验、在仿真环境中实现发现的模型，或将符号回归算法与数据采集系统连接起来。
- 衔接数据与背景知识：大语言模型可以基于定性规则或现有理论等背景知识，为符号回归生成或形式化约束，并搜索驱动因素。随后，在大语言模型的帮助下，可以将符号回归生成的紧凑因果模型整合到更广泛的仿真框架中。这种协同有望

弥合纯数据驱动的发现与理论驱动的仿真之间的差距，从而能够对复杂物理现象进行更强大、更灵活的建模。

7. 会议

这场关于物理科学中的符号回归的会议于2025年4月 28th 日和 29th 日举行，呈现了该领域的现状，并试图涵盖广泛的主题。方法论是会议的重点，多场报告介绍了用于搜索和分类的新技术，尤其是引入了一定程度的统计严谨性，而这在该领域此前大多缺失。

Harry Desmond介绍了穷举符号回归背后的主要思想，并倡导使用描述长度作为一种特别有效的表达式排序方法；Roger Guimerá阐释了贝叶斯机器科学家的策略，并指出贝叶斯模型选择与统计物理学之间的相似之处。Cristina Cornelio倡导使用两种算法——AI-Descartes和AI-Hilbert——它们融入了数学中固有的一些公理化方法和形式化证明方法；Geoffrey Bomarito则介绍了在遗传编程搜索最优表达式的过程中引入后验采样的优势。Bogdan Burlacu介绍了一种利用Zobrist 哈希检测重复项的新方法；出于类似的目的，Fabricio Olivetti de Franca提出使用等式图紧凑地存储表达式，从而便于识别递推关系。Etienne Russeil介绍了他的方法：利用多个数据集，从独立实验推导出一般规律。

会上有机会讨论取得了令人瞩目成果的成熟软件包：Miles Cranmer 介绍了 PySR，这是一款广泛使用的、基于遗传编程的符号回归代码；J. Nathan Kutz 倡导使用 SHRED，在潜在空间中应用稀疏回归；William La Cava 则介绍了他的多目标符号回归框架 BRUSH，以及医学领域的模型可解释性。

研讨会的一个重点是深入考察各类应用。除了前几篇论文中介绍的示例外，Deaglan Bartlett 还展示了用于大尺度结构功率谱的解析模拟器，其效果远胜于目前基于神经网络的模拟器。Tariq Yasin 展示了如何利用穷举符号回归（ESR），直接从星系团的弱引力透镜数据中推断解析轮廓。Evgeniya Kabliman 运用符号回归（SR）确定了金属合金材料性质的解析表达式，而 Steven Abel 则成功构建了用于超越标准模型物理的高效模拟器。最后，Andrei Constantin 发现了物理定律数学表述中一项引人入胜的统计性质，这与分析书面文本时发现的齐夫定律类似。

8. 挑战与展望

尽管取得了成功并展现出巨大潜力，符号回归仍面临持续的挑战。由于可能的表达式空间会随着输入数量的增加而迅速扩大，因此将其扩展至高维问题仍然困难。要提取真正的物理洞见，而非重现伪影，必须具备对噪声、异常值和系统测量误差的抗噪鲁棒性。计算限制也制约着搜索大型表达式空间的可行性，因此有必要开发更高效的启发式方法和搜索策略。

识别出数学上正确、但在物理上毫无意义的表达式这一风险依然存在。因此，领域知识对于构建适定问题、选择恰当的算子和约束，以及评估发现的关系是否对应真实机制而非虚假相关，仍然十分重要。通过已有拟合结果、物理对称性或解析驱动的外推行为融入先验知识，也能缩小有效搜索空间；与纯探索式方法相比，这有助于提高效率。

英国皇家学会会议详细探讨了这些问题。与会者强调，均方误差等传统拟合指标不足以评估符号模型的实用性，因为良好的插值并不保证合理的外推或物理合理性。可解释性、简洁性和不确定性下的稳健性等其他标准被认为至关重要，而最小描述长度等方法则为复杂度控制提供了有原则的机制。与会者还持续讨论了数据中的不确定性如何传递至符号表达式，以及为何需要能够提供置信度评估并识别所发现模型可信适用范围的方法。

会议进一步强调，符号回归在计算和组合方面的负担仍是根本性约束。即使采用受限的算子集或限制表达式复杂度，该问题在形式上仍属 NP难，而穷举方法则明显需要在最优性、成本和函数灵活性之间权衡。会议提出，层次化或混合建模方法可以作为一种解决途径：在这类方法中，共享全局结构，而数据集特定参数则在局部变化，从而扩大符号回归的适用范围，同时缓解其中一些挑战。

9. 结论

符号回归是计算科学领域中一种独特且日益强大的工具。它能够从数据中推导出明确的数学模型，因此具有双重作用：既是推动基础科学发现的强大引擎，尤其有助于理解复杂涌现现象；也是用于创建稳健经验公式和高效模拟器的实用工具。符号回归对可解释性和新知识生成的重视，使其有别于许多其他机器学习范式。

物理科学中符号回归的未来发展，需要持续推进算法创新，以应对可扩展性和稳健性方面的挑战；更深入地融合领域特定知识，以指导发现过程；与大语言模型等互补的人工智能技术创造性地协同；并不断拓展其在物理科学实际问题中的应用。随着这些进展逐步实现，符号回归有望在加快发现步伐、深化我们对支配物理世界的复杂数学图景的理解方面发挥日益重要的作用。

参考文献

1. Lenat DB. 1977年 数学中的自动化理论构建。载于 第五届国际人工智能联合会议论文集——第*2*卷 IJCAI’77 第833–842页，美国加利福尼亚州旧金山。Morgan Kaufmann Publishers Inc. 2. Langley P, Bradshaw GL, Simon HA. 1981年 BACON.5：守恒定律的发现。载于第七届国际人工智能联合会议论文集——第*1*卷 IJCAI’81 第121–126页，美国加利福尼亚州旧金山。Morgan Kaufmann Publishers Inc. 3. Falkenhainer BC, Michalski RS. 1986年 *Machine Learning***1**，367–401。（10.1023/a:1022866732136）4. Dzeroski S, Todorovski L. 1995年 发现动力学：从归纳逻辑编程到机器发现。*Journal ofIntelligentInformation Systems* **4**，89–108。（10.1007/bf00962824）5. Koza JR. 1992年 遗传编程：通过自然选择进行计算机编程。MIT出版社。6. Cranmer M. 2023年 使用 PySR 和 SymbolicRegression.jl 开展科学领域的可解释机器学习。（10.48550/ARXIV.2305.01582）7. Burlacu B, Kronberger G, Kommenda M. 2020年 Operon C++：一种用于符号回归的高效遗传编程框架。载于 *2020*年遗传与进化计算会议伴会论文集 GECCO ’20 第1562–1570页，美国纽约州纽约市。计算机协会。（10.1145/3377929.3398099）8. Udrescu SM, Tegmark M. 2020年 AI Feynman：一种受物理学启发的符号回归方法。*Science Advances***6**，eaay2631。（10.1126/sciadv.aay2631）

9. Cornelio C, Dash S, Austel V, Josephson TR, Goncalves J, Clarkson KL, Megiddo N, El Khadir B, Horesh L. 2023年 结合数据与理论，借助 AI-Descartes 实现可推导的科学发现。《自然·通讯》**14**。（10.1038/s41467-023-37236-y）10. Sahoo S, Lampert C, Martius G. 2018年 学习用于外推和控制的方程。载于 Dy J, Krause A 编，第 *35* 届国际机器学习会议论文集第 80 卷 机器学习研究论文集 第4442–4450页。PMLR。11. Landajuela M, Lee CS, Yang J, Glatt R, Santiago CP, Aravena I, Mundhenk T, Mulcahy G, Petersen BK. 2022年 深度符号回归的统一框架。载于 Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A 编，神经信息处理系统进展 第35卷，第33985–33998页。Curran Associates, Inc.