深度学习会有自己的“牛顿力学”吗

一篇 2026 年论文试图把散落的经验、定律与极限，组织成“学习力学”这门新科学

2026 年 4 月 23 日，一篇题目近乎宣言式的论文出现在 arXiv 上：*There Will Be a Scientific Theory of Deep Learning*。这不是一篇提出新模型或刷新基准的技术论文，也不是传统意义上的文献综述。它更像一场公开押注。作者们押注的是：深度学习并不只是一套靠经验、算力和大规模试错堆出来的工程工艺；它正在缓慢长出一种真正的科学理论。作者给这门尚未成熟的理论起了一个名字，叫 **learning mechanics**，中文可以译作**学习力学**。

这个名字并不低调。它故意把自己放在物理学的影子下面，暗示深度学习未来也许会拥有类似经典力学、统计力学那样的地位：不是解释每一个细节，而是抓住系统中最重要、最稳定、最可预测的量。论文的野心，恰恰不在于“彻底看懂一个神经元”，而在于回答另一种更像科学的问题：**当网络、数据、目标函数和优化规则相互作用时，哪些宏观现象是稳定可测的，哪些规律可以被数学预测，哪些超参数其实只是表象，哪些看似不同的模型其实在学同一种东西？**

这也是这篇论文最有意思的地方。今天人们谈深度学习理论，常常会陷入两种极端。第一种极端是悲观：既然大型模型如此复杂，理论几乎注定追不上实践。第二种极端是轻率：把一切零散理论工作都匆忙打包成“我们已经快理解深度学习了”。这篇论文的立场夹在中间。它既没有宣称深度学习的统一理论已经到来，也不接受“永远不会有理论”这种放弃式判断。它真正想说的是：**理论不是突然完整降临的，而是先以一组彼此呼应的局部规律出现；当这些规律开始共享语言、共享对象、共享预测方式时，一门科学才算真正开始成形。**

从这个角度看，作者们并不是在宣布胜利，而是在尝试界定战场。

一、我们到底缺的是什么

要理解这篇论文的雄心，首先要理解它对“缺什么”的判断。深度学习当然并非完全没有理论。普适逼近、PAC 学习、泛化界、优化理论、核方法、统计物理传统，这些都是真实存在的理论成果。问题在于，它们大多擅长回答的是旧问题，或者只在过于简化的系统里给出漂亮答案。今天的神经网络却不是那样运作的。

现代深度学习最令人不安的事实是：它一方面极其可测，另一方面又极其难懂。神经网络不像天气、细胞或宇宙那样把控制方程藏在自然界里。它的组成部分明明白白摆在桌上：架构、数据集、损失函数、初始化、学习率、batch size、优化器，连每一步梯度都可以被记录下来。从这个意义上说，深度学习不是“黑盒”，至少它不是那种我们连内部变量都拿不到的黑盒。

可悖论恰恰在这里。**深度学习的问题不是缺少可观测量，而是可观测量多到几乎淹没解释。** 你可以追踪每个参数、每个激活值、每个 Hessian 特征值，但这不等于你因此知道训练为什么这样走、而不是那样走。一个系统的可见性，和一个系统的可理解性，并不是同一回事。天文学

也能观测到大量数据，但只有当开普勒和牛顿把这些数据组织成简洁关系时，天文学才开始变成一种可预测的科学。

作者们的核心判断是：深度学习正在接近这样一个时刻。不是因为我们终于能逐层逐点解释所有参数，而是因为一些**粗粒度、聚合性的统计规律**开始反复出现。它们不解释每一个像素或每一个 token 如何被处理，却能解释训练过程中什么先学会、什么后学会，哪些量如何随规模变化，哪些不同模型最终会趋向相似表示，哪些超参数实际上是在控制同一种底层动力学。

这篇论文最关键的转折，是把“深度学习理论”从‘证明什么可能’转成‘描述什么真实发生’。 这个转向听起来像措辞变化，实际上意义很大。它把研究目标从最坏情况保证、形式可证的边界，部分迁移到了平均情形、经验可检验的规律、以及可被实验反驳的定量预测。这更像自然科学，而不那么像经典理论计算机科学。

二、为什么作者坚持要叫它“学习力学”

“学习力学”这个比喻不是装饰性的。它是全文的组织原则。

在物理学里，力学研究的是对象如何在约束和作用力下运动。作者说，神经网络训练过程也可以这样看：参数在高维空间里移动，梯度扮演“力”，数据和架构决定损失地形，优化器决定运动方式，最后系统在某些局部区域停下来。这样的类比当然不是严格同构，但它足够强，强到可以指导研究方法。

如果你接受这点，那么很多原本分散的理论工作就会突然连成一线。物理学会用可解模型建立直觉，深度学习也有自己的可解玩具模型；物理学会取极限来剥离细节，深度学习也有无限宽、无限深、小初始化、小学习率等极限；物理学里存在像 Boyle 定律、Ohm 定律那样先于完整理论出现的经验规律，深度学习里也有 scaling laws、edge of stability 这样的可重复关系；物理学研究无量纲参数和控制参数，深度学习现在也在系统研究学习率、batch size、宽度、深度如何改变动力学；物理学最激动人心的一刻，往往是发现不同系统共享同一普适行为，而深度学习如今也开始在不同架构、不同模态、不同随机种子下看到类似的表示和归纳偏置。

这就是作者所谓五条“证据链”。

它们分别是：**可解析的理想化设定、可处理的极限、简单经验定律、超参数理论、跨系统的普适现象。** 单看任何一条，都不足以构成一门科学；但当五条线都在朝相近方向推进时，作者认为这已经不是偶然的拼盘，而是一种学科轮廓。

这个判断很大胆，也很容易惹人不快。因为它隐含着一个主张：过去很多人以为深度学习太复杂，不可能出现像“物理”那样的理论；但也许真正的问题不是深度学习不可理论化，而是我们长期试图用不对的理论姿势去理解它。

三、第一条证据：玩具模型不是逃避现实，而是逼近现实

很多人一听到“可解模型”就会皱眉。深度线性网络？核回归？这离 GPT、扩散模型和多模态系统差得也太远了。用这种玩具模型得出的结论，真的有资格指导现代深度学习吗？

这是论文最先要为自己辩护的地方。作者的回答是：在成熟科学里，最有价值的模型往往不是最逼真的，而是**最简单但足以暴露关键机制的模型**。氢原子不是化学世界的全部，但它是量子力学的训练场。谐振子不能代表所有物体运动，但它是理解振动、近似和稳定性的母语。

深度学习理论里类似的工作马主要有两个。

第一个是**深度线性网络**。它把非线性激活拿掉，于是网络对输入是线性的，但对参数仍然是高度非线性的。这一点非常关键。它保留了“深”的乘法结构，因此训练动力学仍有层间耦合、鞍点结构、时间尺度分离、初始化依赖等现象。作者特别强调，Saxe 等人的经典结果表明，在合适条件下，这类网络会按奇异值大小**依次**学会任务中的不同模式。大的、简单的、信号更强的模式先被掌握，小的、复杂的、容易混入噪声的模式后被掌握。

这是一个极其重要的直觉。因为它告诉我们，训练不是把函数整体一下子塞进网络，而更像是在一堆模式中按“可学性”和“强度”排队。**网络对任务的学习，天然带着一种贪心的低秩偏置。**这和很多经验观察互相照应，比如神经网络往往先学简单模式、后学复杂模式，先学信号、后学噪声。

如果要找一个更直观的类比，可以把它想成一块正在显影的底片。最粗、对比最强的轮廓总是先浮现，细微纹理和噪声则晚得多。深度线性网络并不告诉你所有真实网络都严格这样，但它告诉你：这种“显影顺序”不是错觉，它在某些可解系统里就是训练动力学的结果。

第二个工作马是**参数线性化与神经切线核（NTK）**。这里的简化不是把输入映射线性化，而是在初始化附近把模型对参数的变化做一阶展开。这样一来，网络虽然对数据仍然非线性，但对参数变成线性，于是训练动力学能被核回归描述。这个框架的价值在于，它把架构如何编码归纳偏置这件事，变成了一个可以直接计算的对象：NTK 的特征结构。

对非理论读者来说，最值得抓住的不是公式，而是它提供的视角：**如果一个复杂网络在训练中并没有离初始化太远，那么它的很多行为其实更像是在固定特征上做线性拟合。**一旦进入这个视角，double descent、泛化误差、不同目标函数的学习顺序，就开始具备可预测性。

当然，问题也随之而来：真实的大模型真的会长期停留在线性化附近吗？作者并没有夸大。论文明确承认，这类模型在解释“特征真正发生变化”的 rich regime 时是不够的。它们最大的价值，不是提供完整答案，而是提供**可计算的基准点**。正因为我们知道线性化世界会发生什么，我们才更能分辨真实网络究竟在哪些地方偏离了它。

这也是整篇论文反复出现的一个方法论：先找到一个足够干净的参考系，再研究现实系统如何偏离这个参考系。没有参考系，复杂性只是噪音；有了参考系，复杂性才开始显露结构。

四、第二条证据：极限不是逃跑，而是把细节洗掉

深度学习理论中最著名的极限，大概就是“无限宽”。很多人对它的印象并不好，因为一听就觉得不现实：现实网络再大也有限，为什么要拿无限对象说事？

论文给出的辩护很物理。热力学里的无限粒子、流体力学里的连续介质、量子场论里的连续极限，本来也都不字面真实。极限的作用，从来不是宣称世界真有无穷多个粒子，而是找到一个**在宏观上更容易看清结构的描述**。神经网络也一样。

在无限宽极限下，一类著名现象是所谓的 **lazy learning**：隐藏表示几乎不变，训练主要像在线性化框架里调整最后的组合方式。与之相对的是 **rich learning**：表示本身显著演化，网络真正“学到新特征”。这个区分之所以重要，是因为它切中了一个长期模糊的问题：深度学习到底只是大型核方法，还是它真正通过训练制造了新的表示结构？

作者并不否认 lazy regime 在理论上极其干净，也不否认它确实解释了很多现象。可他们更强调的是，现代深度学习的关键能力大多来自 rich regime。换句话说，**真正让大模型强大的，不只是它能在固定特征上拟合，而是它能在训练过程中重写自己的特征坐标系。**

这时，参数化理论就成了关键。论文把 μP 视为一个里程碑，因为它不是单纯讨论“无限宽会发生什么”，而是讨论“怎样随宽度一起缩放超参数，才能在变宽时保住有意义的特征学习”。这比传统“只看无限极限”更接近工程实践。它提供的不是一种形而上的解释，而是一套可迁移的操作规则：在小模型上调参，再把规律迁移到大模型上。

这里有一个很像物理的思想转折。过去大家把宽度、深度、学习率看成许多互相缠绕的独立旋钮；现在越来越多工作表明，其中一些旋钮也许只是同一底层动力学的不同刻度。**如果某些超参数可以通过合适极限被“吸收”掉，那么复杂系统就会露出更简单的骨架。**

作者甚至提出一种“离散化假说”：有限宽、有限深、有限学习率的网络，可能可以被理解为某种连续极限系统的离散近似。这个想法还远没被证明，但它很抓人。因为它把“模型变大为什么更强”从一件经验事实，转成了一个更可检验的问题：也许更大的模型不是神秘地获得新能力，而只是更好地逼近了某种连续动力学。

如果这个想法最终站住脚，深度学习中的许多规模效应就会出现全新解释。那时“更大模型更强”不再只是经验法则，而会像数值分析里“网格更细误差更小”那样，变成某种离散误差控制问题。

五、第三条证据：先有“气体定律”，再有完整理论

对深度学习实践者来说，论文里最熟悉的一类内容，恐怕是各种经验定律。比如 scaling laws，比如 edge of stability。这些东西已经足够有名，以至于很多人甚至把它们看成大模型时代少数真正“有工程价值的理论”。

论文的高明之处在于，它并没有因为这些规律还缺完整解释，就把它们贬成经验 hack。恰恰相反，它把它们视为科学开始成熟的信号。

在近代科学史里，经验定律经常先于底层理论出现。开普勒先总结出行星轨道规律，牛顿后来才解释为什么。Boyle、Ohm、Hooke 这些名字之所以重要，不是因为他们给出了万物终极原理，而是因为他们找到了**在复杂系统中稳定复现的数量关系**。一门科学最早成熟的标志，往往不是“已经理解一切”，而是“已经知道哪些量之间总会这样联系”。

深度学习里的 scaling laws 就有这种气质。大模型损失随参数、数据、算力按幂律改善，这种现象跨架构、跨模态、跨任务地反复出现。它最震撼人的地方，不只是“可拟合”，而是**如此粗糙的宏观量，居然被如此简单的函数描述。** 这意味着在纷繁的训练细节之下，也许真的有少数主导变量在控制整体行为。

但论文也十分克制。作者明确承认，今天没有任何理论能稳健地从第一性原理预测这些幂律指数。我们知道有幂律，却不知道指数为什么是那个数；我们知道某些模型共享斜率，却不知道数据分布、架构和优化器各自贡献多少；我们知道这些关系对工程极其有用，却还不能说自己真正“懂了”它们。

这正是论文最像科学宣言的一点：**经验规律不是理论的替代品，而是理论必须解释的对象。** 真正的危险不是目前还解释不出 scaling exponents，而是把“能拟合”误当成“已经理解”。作者不断把读者往后推一步，逼你问：这些幂律从哪来？它们到底编码了数据的什么结构？为什么不同任务里指数不同？有没有一套先验测量可以在训练前预言它们？

另一条经验规律是 edge of stability。大量实验显示，梯度下降训练时的曲率会朝一个和学习率相关的边界靠拢。最动人的地方在于，这个现象最初看起来像纯经验怪象，后来却被越来越多理论工作部分解释成一种**隐式曲率正则化**。换句话说，学习率和 batch size 不是单纯决定“走多快”，而是在暗中塑造优化轨迹经过什么样的地形。

这是一个很深的转折。它意味着优化超参数并非外在控制杆，而是系统动力学本身的一部分。过去人们说“调学习率很重要”，更多像经验工艺；现在更可能的说法是：**学习率在选择网络会沿着什么样的几何路径学习。**

六、第四条证据：超参数也许没我们想的那么神秘

如果说现代深度学习最像炼丹的部分，超参数调优一定名列前茅。学习率、batch size、动量、初始化尺度、宽度、深度、输出 multiplier……它们彼此耦合，轻微变化就可能毁掉训练。很多工程师对“理论会解决超参数问题”天然怀疑，因为实践中这实在太混乱了。

论文在这里给出的不是乐观口号，而是一种更有节制的主张：超参数不一定能被全部消灭，但其中相当一部分正在被**解缠**。也就是说，我们开始知道哪些调节其实是在改变同一个无量纲控制量，哪些缩放能保留近似不变的训练轨迹，哪些参数的作用可以用更简单的等效系统来表达。

最经典例子是 batch size 和 learning rate 的联动缩放。在线性 scaling rule 背后，一条重要理论路线把 SGD 视为某种随机微分方程的离散化。这样一来，“把 batch size 翻倍、learning rate 也翻倍、步数减半，轨迹近似不变”就不再只是经验口诀，而是某种连续极限下的近似对称性。

这种解释的价值并不只是“更好记”。它改变的是问题本身。以前你面对的是一堆孤立旋钮，现在你开始问：**哪些旋钮是在共同控制噪声强度，哪些是在控制曲率正则化，哪些是在决定是否进入 feature learning regime？**

μP 的贡献则更激进。它把“不同宽度模型的最优学习率不一致”这件令人头疼的工程事实，变成一个参数化问题：如果你选对了随宽度缩放的规则，那么最优超参数会在跨宽度时趋于稳定。这件事的哲学意味很强。它暗示过去我们以为是“每个模型都要从头调”的复杂性，其中一部分其实是因为我们在错误坐标系里看问题。

这很像物理学里从有量纲参数换到无量纲参数之后，原本杂乱的数据突然塌缩到同一条曲线上。**当不同系统能被同一组缩放规律整理起来时，理论就不只是解释，它开始压缩现实。**

当然，论文也没有承诺乌托邦。作者明确提出一个开放问题：能不能把超参数真正压到接近零？还是说有些超参数是不可约的？这是一个异常诚实的问题。因为它承认，复杂系统并不保证最终会被“洗成”纯净公式。有些细节也许会被吸收进理论，有些细节则可能永远是工程设计的一部分。

七、第五条证据：最令人不安也最令人兴奋的，是“普适性”

如果前面四条还比较像“理论家在整理工具箱”，那么第五条证据才真正带有一种科学革命的味道：**不同系统也许在学习同样的东西。**

这条主张最初听起来近乎夸张。卷积网络和 Transformer 长得不像，图像和文本的数据结构差别巨大，不同随机种子训练出的模型参数也不会逐项相同。为什么我们还要谈“普适表示”或“普适归纳偏置”？

作者给出的理由有三层。

第一层是**功能上的普适性**。很多任务上，不同架构在给定足够数据、算力和合适 recipe 后，会收敛到相近性能。视觉任务里卷积和 Transformer 已经越来越像这样的关系；扩散模型工作中，甚至出现不同架构给相同噪声种子生成近乎相同图像分布的结果。这说明架构差异未必总是决定性的，有时它们只是在通往同一解族的不同路径。

第二层是**数据结构上的普适性**。图像、音频、文本看似完全不同，但都展现出某些共享统计结构：幂律谱、稀疏性、多尺度层级、Zipf 分布、组合性。这里最重要的不是“它们一样”，而是“它们都有可被一般学习系统利用的结构”。如果深度学习之所以成功，是因为它在许多自然数据里反复碰到类似结构，那么理论最终就不可能只是一套模型理论，它还必须部分变成一套**数据理论**。

第三层，也是最刺激的一层，是**表示上的普适性**。越来越多工作发现，不同初始化、不同宽度、不同任务、甚至不同模态的大模型，内部表示会朝相似方向靠近。论文提到一种颇有诗意的说法：随着模型规模和性能上升，表示也许在趋向一种“柏拉图式”的结构。这个说法当然还带着强烈猜测色彩，但它捕捉到一个惊人的可能性：**模型之所以相似，不一定是因为它们抄了彼此，而是因为数据本身把所有足够强的学习系统都推向了某些相似表示。**

如果这是真的，意义就非常大。因为它意味着我们也许不需要为每个模型重造一遍理论。只要抓住那些在不同系统中重复出现的机制，理论就有机会获得迁移性。

但这里也埋着重要边界。论文专门提醒，表示相似性高度依赖你用什么度量去比较。不同 similarity metric 可能得出不同结论。也就是说，**“模型学到了同样东西”这句话本身还需要被严密定义。** 这不是鸡蛋里挑骨头，而是理论成熟前必须经历的一步：把口号拆成可检验对象。

八、这篇论文真正的争议，不在技术细节，而在它是否过早命名

如果只看社交平台和论坛反应，这篇论文的争议其实很集中。Hacker News 上最典型的一类评论认为，这篇文章像是在“立旗”，把一个仍然碎片化的研究方向，提前包装成新学科。有评论直言它像“学术版的我早就说过”，甚至怀疑标题本身是一种引 citation 的操作。另一类较温和的评论则说：深度学习理论当然存在，只是目前很分裂；这篇论文更像是在**标准化语言**，而不是首次发现某个新大陆。

这类质疑并不肤浅。因为它击中了整篇论文最脆弱也最关键的一点：**作者到底是在发现一门新科学，还是在替许多旧工作重新命名并组装叙事？**

我认为公允的回答是：两者都有。

从严格知识增量上说，这篇论文当然没有单独发明深度线性网络、NTK、scaling laws、μP 或表示相似性研究。它的核心工作确实是组织、综合、命名、建立研究纲领。从这个意义上说，怀疑者说它更像“宣言”而不是“定理”，并没有错。

但问题在于，科学史上很多关键文本本来就不是靠单点新结果取胜的。它们的作用是把一堆还没互相承认同类的工作，拉进同一个句法里。**命名不是表面文章；命名是在决定哪些现象以后要被当作同一个问题去研究。**

这正是“学习力学”这个名字的真正赌注。它不是证明五条证据已经统一，而是在建议人们今后以统一方式提问：训练动力学的有效变量是什么？哪些宏观量可预测？哪些现象是普适的？哪些超参数可被消去？哪些极限是正确参考系？一旦一代研究者真的开始沿这个纲领工作，这个名字就会反过来塑造学科本身。

因此，外部讨论里最有价值的分歧不是“这篇论文标题是否夸张”，而是另一个更深的问题：**深度学习理论的未来，究竟该继续以零散问题为单位推进，还是应当主动追求一种像物理学那样共享对象、共享实验风格、共享近似语言的统一科学？**

这篇论文的回答极其明确：必须追求后者。

九、它如何回应“根本不可能形成理论”的反对意见

论文第四部分很值得读，因为它正面处理了几类常见悲观论。

第一类悲观论是：研究了这么多年还没成功，说明本来就没有统一理论。作者的回应并不神秘。他们说，深度学习真正压倒性成功其实是近年的事，我们今天可研究的经验系统，比十年前丰富得多；而且过去几年出现的一些现象，比如大模型表示的趋同，本来就不可能在旧时代显现。换句话说，**理论之所以过去不够，不一定是因为理论能力不够，也可能是因为现实系统还没把可解释现象暴露出来。**

第二类悲观论是：今天理论能处理的对象和 LLM 相比太原始了，差距大到不可能补齐。作者承认这点，但强调近中期理论未必需要“一步到位解释整个 LLM”。即便只是解释超参数缩放、训练不稳定、数据归因、优化器行为，这种局部理论也已经有现实价值。这里的策略很科学：不要先要求一张总地图，先把若干关键地形测清。

第三类悲观论是：真正重要的是高层行为和能力，不是微观训练动力学。作者用了一个很巧的分层比喻来回答：深度学习也许需要自己的“物理、生物、心理学”。学习力学是物理层，机械可解释性更像生物层，而能力、人格、目标等高层行为更像心理层。**不是只有最高层重要，而是没有底层可解析结构，高层研究容易悬空。**

第四类悲观论更尖锐：我们需要的不是深度学习理论，而是数据理论。作者的回答是：两者都要。因为模型之所以能泛化，必然依赖数据里的结构；而模型如何把这些结构学出来，又是另一个同样不可省略的问题。这个回答很重要，因为它防止“学习力学”滑向纯模型中心主义。它承认，未来真正成熟的理论必然同时包含**数据的结构理论**与**学习过程的动力学理论**。

最后还有一种近年特有的悲观论：AI 也许会比人更早看懂自己，那我们何必费力建立人类可读理论？作者的反应相当直接：第一，理论已经有近端用处；第二，AI 很可能是帮助人类推进理论，而不是突然独立给出完整答案；第三，如果目标涉及安全、治理和监督，人类仍需要自己的可解释框架。这个回答说不上无懈可击，但它至少澄清了一点：这篇论文并不把理论只当成好奇心，而是把它看成一种**保持人类认知主导权的工具**。

十、最值得认真对待的十个开放问题

如果说整篇论文哪部分最像面向下一代研究者的招募书，那一定是它列出的开放方向。这里面最抓人的问题，不是“还能再证明什么”，而是“哪些现象一旦被解释，整个领域的语言就会变掉”。

其中最核心的几个，我认为有四个。

第一个是：**有没有既保留深层非线性参数动力学、又保留非线性函数学习的可解模型？** 这是对当前两大玩具模型传统的直接追问。深度线性网络擅长解释参数动力学，核方法擅长解释函数空间结构，但真正现代网络最关键的地方往往是两种非线性同时存在。谁先找到一个足够一般、又不完全失控的桥梁模型，谁就可能重写很多问题的提法。

第二个是：**自然数据的可学习结构到底是什么？** 这也许是最深的一问。因为 scaling laws、普适表示、跨模态迁移，最后都指向同一个地方：数据本身含有哪些低维、层级、稀疏、组合或频谱结构，足以让一个统一学习算法在各处奏效？如果没有这个问题的答案，学习力学就只能描述“模型如何动”，却不能解释“模型为什么总能在这些世界里学到东西”。

第三个是：**深度学习是否在隐式最小化某种“函数复杂度”？** 这看似抽象，实际上非常核心。人们早就在不同场合说过简单性偏置、谱偏置、最大间隔偏置、隐式正则化，但它们是否只是局部现象，还是某个更一般原则的不同投影？如果神经网络确实倾向于在低损失函数中选取更“简单”的解，那么“简单”究竟该如何数学定义？这会直接连接到可解释性中那些关于稀疏特征、组合结构和电路的争论。

第四个是：**大模型为什么会在不同训练方式下学出相似表示？** 这是最接近“统一理论可能成真”的试金石。因为一旦相似表示能被稳健地定义并从第一性原理推出，理论的可迁移性就不再只是希望，而会变成方法。

除此之外，论文提出的“能否消灭所有超参数”“能否先验预测 scaling law 指数”“有限网络是否只是无限极限的离散近似”“现代优化器为什么比 SGD 更适合大模型”等问题，也都非常硬。它们的共同特点是：一旦被解决，不只会给出新解释，还会**直接改变训练实践**。

这再次说明，作者想建的不是书斋式理论，而是一种最终能下沉到工程栈里的科学。

十一、它和机械可解释性的关系，也许比它自己想的还重要

论文专门花了一节谈 mechanistic interpretability，并提出一个非常有传播力的比喻：如果机械可解释性想成为深度学习的“生物学”，那么学习力学就该成为它的“物理学”。

这个比喻有两层意思。

第一层是互补性。机械可解释性擅长发现局部机制、特征、电路和模块级行为，像在显微镜下解剖生物器官；学习力学则更关心训练整体如何流动、哪些统计量在什么条件下守恒或变形，像在研究器官为何会在某种环境和约束下长成那样。前者偏局部结构，后者偏宏观动力学。

第二层是方法差异。论文指出，机械可解释性到今天仍更偏定性，很多判断依赖人类研究者的视觉与语义解释；学习力学则追求可重复的定量预测。作者显然希望两者形成循环：mechanistic interpretability 提供需要解释的明确现象，learning mechanics 提供更坚硬的数学骨架。

我觉得这部分尤其值得重视。因为大模型研究现在一个很明显的问题是：一边有人在做微观电路归因，一边有人在做宏观 scaling 与优化理论，但两边常常像在不同国家说不同语言。**如果“学习力学”这个纲领真正有价值，它最大的价值之一，可能就是为这两种研究提供中介层。**

没有这个中介层，机械可解释性容易变成现象学收藏；没有微观现象输入，学习力学又可能变成离现实过远的抽象动力学。作者把这两者设想成共生关系，是整篇论文中最现实、也最有战略眼光的一点。

十二、这篇论文真正改变的，也许不是我们知道了什么，而是我们该怎样提问

讨论到这里，可以回到最初那个问题：深度学习会不会有自己的“牛顿力学”？

严格说，论文当然没有证明会。它甚至连“统一理论的雏形已经足够坚固”都没有证明。它做的是更早一步、也更难的一件事：试图说服读者，**我们已经看到足够多彼此呼应的规律，值得把深度学习当成一种正在形成中的经验科学，而不只是工程配方和局部数学技巧的松散总和。**

这不是小事。因为一旦你接受这个框架，很多问题的优先级会被重排。

你不再首先问“这个理论能不能给我一个严格上界”，而是问“它能不能准确预测一个可测宏观量”。你不再首先问“这个模型是不是现实”，而是问“它是不是一个好参考系”。你不再把 scaling laws 当成拟合技巧，而是把它们当成必须解释的对象。你不再把超参数只看成炼丹旋钮，而是把它们当成潜在的控制变量。你不再把不同模型间的相似性当成巧合，而是当成可能存在普适类的信号。

这正是科学开始成熟时最先发生的变化：不是答案先齐了，而是问题先被重新整理。

如果非要给这篇论文一个最简洁的评价，我会说：它并没有告诉我们深度学习的统一理论已经在哪里，但它认真地画出了一张地图，指出哪些局部地形其实正在连成大陆。

这张地图当然可能有误。也许五条证据最终不会汇成一条主河，而只是若干彼此相关但不统一的分支理论。也许“学习力学”会被证明是个过大的名字。也许深度学习最终更像化学而不是力学，更像生态学而不是经典物理。所有这些都仍然可能。

但即便如此，这篇论文仍然值得认真对待。因为它逼着整个领域面对一个长期被回避的问题：**我们到底满足于把深度学习继续当作一门成功却半盲的工艺，还是准备把它推进为一种能够解释、预测、压缩并最终指导工程实践的科学？**

作者们押注后者。

这个赌注未必很快兑现。但它至少把争论从“有没有魔法”推进到了“什么算解释”。而对任何一门年轻科学来说，这往往就是最重要的第一步。

参考资料

1. Jamie Simon, Daniel Kunin et al. *There Will Be a Scientific Theory of Deep Learning*. arXiv, 2026-04-23.

https://arxiv.org/abs/2604.21691

1. 论文 HTML 版本。

https://arxiv.org/html/2604.21691v1

1. Learning Mechanics 项目主页。

https://learningmechanics.pub/

1. Imbue 配套介绍文章 *There Will Be a Scientific Theory of Deep Learning*，2026-04-24，2026-05-08 更新。

https://imbue.com/blog/2026-04-24-podcast-episode-38-jamie-simon-and-daniel-kunin

1. Hacker News 讨论串 *There Will Be a Scientific Theory of Deep Learning*。

https://news.ycombinator.com/item?id=47893779

1. Reddit `r/MachineLearning` 讨论串 *There Will Be a Scientific Theory of Deep Learning [R]*。

https://www.reddit.com/r/MachineLearning/comments/1sun588/there*will*be*a*scientific*theory*of_deep/