提交版本 — 审稿中。（v1，2026-01-01）

堆叠自编码器进化假说

Hiroyuki Iizuka [1]

通讯作者： Hiroyuki Iizuka (iizuka@chain.hokudai.ac.jp)

1. 人类本性、人工智能与神经科学中心（CHAIN），北海道大学，北12条西7丁目，日本札幌市北区，北海道

arXiv:2602.01026v1 [cs.NE] 1 Feb 2026

摘要。 本研究提出了一种全新的理论框架——堆叠自编码器进化假说。该假说认为，生物进化系统通过多层自编码与解码过程运作，这类似于深度学习中的堆叠自编码器。该假说并不将进化仅仅视为由突变和选择驱动的渐进变化，而是提出自我复制本身会在层级化的抽象层次上压缩并重构遗传信息。这种分层结构使进化系统能够探索多种可能性，不仅限于序列层面，还能跨越越来越抽象的表征层，从而使简单的突变也有可能在这些高阶空间中探索。通过允许更深层潜在层中的突变触发突然且大规模的表型变化，这种机制或可解释自然进化中的间断式进化模式，以及那些看似目标导向的变化。为说明这一机制的合理性，研究开展了人工化学模拟，结果显示层级自编码器结构能够自发涌现。该框架为理解连续与不连续进化变化背后的信息动力学提供了全新视角。

关键词：进化、自编码器、假说

1 引言

自达尔文奠基性的研究以来，进化理论的发展主要建立在通过适应和自然选择实现渐进变化的原则之上（Darwin，1859）。根据这一观点，进化并非通过突然而大规模的转化发生，而是通过长时间积累微小变异，最终产生新的性状和物种。渐进进化模型为解释生命的多样性与复杂性提供了强有力的框架。

然而，化石记录和分子系统发育分析的证据表明，进化变化并非总是以平稳、连续的过程进行。相反，它往往在长期停滞与快速变化阶段之间交替。尤其是，许多有充分记录的进化事件呈现出显著的不连续性，挑战了仅依靠渐进的低水平突变来解释进化的观点。例如，寒武纪大爆发期间快速的形态多样化（McMenamin，2013）、主要身体构型的出现（Valentine，2004），以及达尔文雀或非洲慈鲷等大型适应辐射（Schluter，2000）。这些阶段的特点是形态或功能多样性突然扩展，表明许多性状发生了协同变化，而非彼此独立的变化逐渐累积。这些宏观进化不连续性并不只见于化石记录。进化发育生物学（进化发育生物学）研究同样表明，基因调控网络的重组或发育限制的变化能够引发协同的大规模表型转变（Carroll，2005）。这类不连续性通过迅速拓展形态空间的新区域，并促成仅靠纯粹渐进机制难以实现的转变，在塑造当今生物多样性方面发挥了核心作用。

间断平衡理论体现了这种模式，该理论认为，物种会经历长期稳定期，其间被短暂的重大转变时期打断（Eldredge, Gould, et al., 1972）。类似地，长期以来的跳跃式进化假说，即

进化“飞跃”假说认为，性状的大规模变化可能通过突变或染色体重排突然发生（Goldschmidt, 1940）。这些假说揭示了进化的非连续性，表明进化变化并非总是均匀或渐进的。

尽管这些假说受到广泛关注，但它们大多仍停留在描述层面，对这类非连续变化背后的机制仍缺乏深入理解。例如，间断平衡理论有效地描述了化石记录中观察到的模式，却未能阐明驱动这些变化的进化动力学。同样，尽管偶尔观察到跳跃式进化，但大规模突变产生可存活生物或物种的可能性通常被认为很低，因此很难将跳跃式进化视为进化的主要驱动力。

这类变化的一种常被引用的候选机制是木村分子进化中性理论（Kimura，1979）。该理论认为，大多数分子层面变化对于适合度而言是中性的，因此会在漫长时间内累积，而不表现为表型变化。有人提出，在突发的环境突变或发育重组等特定条件下，这些隐而未显的累积突变可能会同时表达，从而导致快速的表型变化。尽管这种机制为间断变化和跳跃式变化提供了合理解释，但从根本上说，它仍植根于渐进变异的达尔文框架，将进化视为由随机突变和选择驱动的被动随机过程。因此，它不足以充分解释进化转变中有时观察到的明显跃迁或目的性。

相比之下，Imanishi 提出了一种更为主动的进化观（Asquith，2007）。他的理论通常被称为今西进化理论，将进化描绘为一种过程，而非一系列受外部选择压力影响的随机事件；在这一过程中，生物体为了追求与其环境的和谐，主动选择自身的转化，以实现与其环境的和谐

。这一观点旨在解释某些进化转变——这些转变有时被诠释为具有方向性，或仿佛在本质上是有意向的——同时又不假定存在任何字面意义上的目的论。然而，这种生物体能动性和意向性变化的生物学基础仍大多未被阐明；此类过程可能作用于哪些层次，无论在理论上还是经验上，也都尚未得到充分澄清。因此，现有理论虽然认识到进化中的不连续性和方向性，却仍未能对这些现象背后的动态机制提供统一解释。

针对这一空白，本研究提出了一种关于进化动力学的新视角，聚焦于进化系统固有的自组织探索能力与信息动力学。具体而言，本研究假设自我复制过程本身可能体现一种类似于堆叠自编码器的信息处理架构，而这一概念已在人工智能领域取得进展。通过迭代的自我复制，生物体可能自然地压缩并重构遗传信息，从而在其个体发育架构中形成多层潜在空间。发生在这些层级结构化潜在空间中的突变有望促成突然且看似有目的的变化，从而为进化跃迁提供新的解释框架。

2 理论背景

2.1 人工生命与自我复制系统中的进化过程

对进化过程的研究长期以来一直是生物学以及人工生命领域的核心焦点。人工生命研究并不通过将生命还原为其物质组成来探讨生命的本质，而是强调生命系统维持并生成自身组织与过程的机制。值得注意的是，自创生这一概念（Varela，1979）将生命定义为一个持续自我生成的系统；此外，朗顿自我复制环（Langton，1984）等模型，

展示了基于最小规则集的自我复制，为探索类生命行为的涌现与进化提供了基础框架。在此研究方向上，EvoLoop 已证明，自我复制环之间对空间资源的竞争能够产生进化动力学（Salzberg，2003；Sayama，1999）。这里尤为重要的是这样一种观点：进化过程可以自然地从物质层面的自我维持与自我复制机制中涌现，而非由外部强加或预先设定。

这些研究共同凸显了一种进化观，它超越了将进化视为环境优化或目标导向适应过程的传统框架。相反，这些研究强调自我复制是一种动态且持续进行的过程，系统通过这一过程不断维持自身并生成自身的副本，同时保持开放，能够发生转变并与环境互动。这种观点将进化描绘为递归自我生产和变异的开放式过程，而非朝着最优状态线性发展的轨迹。

Tierra（Ray，1992）和 Avida（Ofria 和 Wilke，2004）等实验平台为研究人工生命中的进化过程提供了成熟的计算环境。这些系统在虚拟环境中模拟数字生物体——即自我复制程序，使研究者能够观察多样性和新功能如何通过随机突变涌现。与自创生系统或自我复制环等早期模型不同，这些平台处于稍高层次的抽象水平，将功能行为直接嵌入指令集之中。特别是，Avida 揭示了复制和累积突变如何产生新的行为，为此提供了详细见解。通过再现自我复制与变异的循环，这些平台极大地深化了我们对进化动力学的理解。

尽管这些平台展现出丰富的进化动力学，并已证明新功能能够涌现，但它们采用的遗传表示通常是扁平且

线性的。它们的基因组以指令序列的形式实现，突变通过替换、插入或缺失等局部操作修改单条指令。因此，这些系统中的适应性变化通常被建模为这些序列中小幅局部修改逐渐累积的结果。

与指令序列基因组不同，人工生命研究的另一条脉络明确区分了描述与执行算子。在Ikegami和Hashimoto（1995）提出的“机器–磁带”模型中，自我复制通过读取和写入磁带的机器网络实现；进化变化不仅源于固定基因组的局部扰动，也源于磁带（描述）与机器（算子）之间的耦合动力学。该框架凸显，复制本质上可能包含一个由表征机制与转换机制相互作用所介导的内部编码–解码循环。

除了这些基于模拟的方法之外，还有若干分析框架试图刻画开放式进化背后的生成原理。一个值得注意的例子是 Tria等人（2014）的瓮模型，它将 Kauffman（2000）提出的“相邻可能”概念形式化——即随着新创新不断积累，可达的新状态集合也会扩大。这些观点凸显，开放式进化不仅由变异和选择驱动，也由可能性空间本身的动态扩展所驱动。将本文的假说置于这一更广阔的背景下来看，本文提出的分层信息处理架构等机制，或许能够为生成和探索不断扩展的进化可能性空间提供另一种途径。

基于这一系列研究，本研究提出一个新颖的假说：自我复制本身可能内在地包含一个自我编码与解码的内部过程。具体而言，自我复制过程可以理解为一种分层信息压缩与重构，类似于机器学习中广泛研究的堆叠式自

编码器架构。从这一视角来看，自我复制并非只是复制现有结构，而是一个多层次的信息处理循环：生物体在将自身结构传递给下一代之前，会先在内部对其进行压缩与重构。

下一节将详细阐述堆叠自编码器，并介绍其工作机制。

2.2 堆叠自编码器：深度神经网络中的层次特征学习

堆叠自编码器是深度学习中的基础架构，尤其适用于通过自监督学习获取层次化特征表示（Bengio 等人，2006年；Vincent 等人，2010）。传统上，训练深度神经网络时，若直接堆叠各层，往往会遇到梯度消失等问题。堆叠自编码器通过逐步添加编码层和解码层来解决这一问题，从而支持逐层预训练和层次抽象。如图 1 所示，输入数据经过各层逐步压缩至维度更低的潜在空间，节点数量随之减少，同时提取出越来越抽象的特征（Bengio 等人，2013）。浅层通常捕捉低级特征（例如边缘、纹理、字符级或词级模式），而较深层则编码物体类别、语义语境和抽象表示等高级概念（Masci 等人，2011；Ngiam 等人，2011）。这种架构的一项关键优势是，解码层与编码层同步训练，因此能够根据中间特征表示重构输入。Bengio 等人（2006年）证明，堆叠自编码器能够缓解梯度消失问题，并显著提升深度网络的训练效率，因此成为现代深度学习发展历程中的一个里程碑。

虽然堆叠自编码器采用确定性自编码机制，但其概念

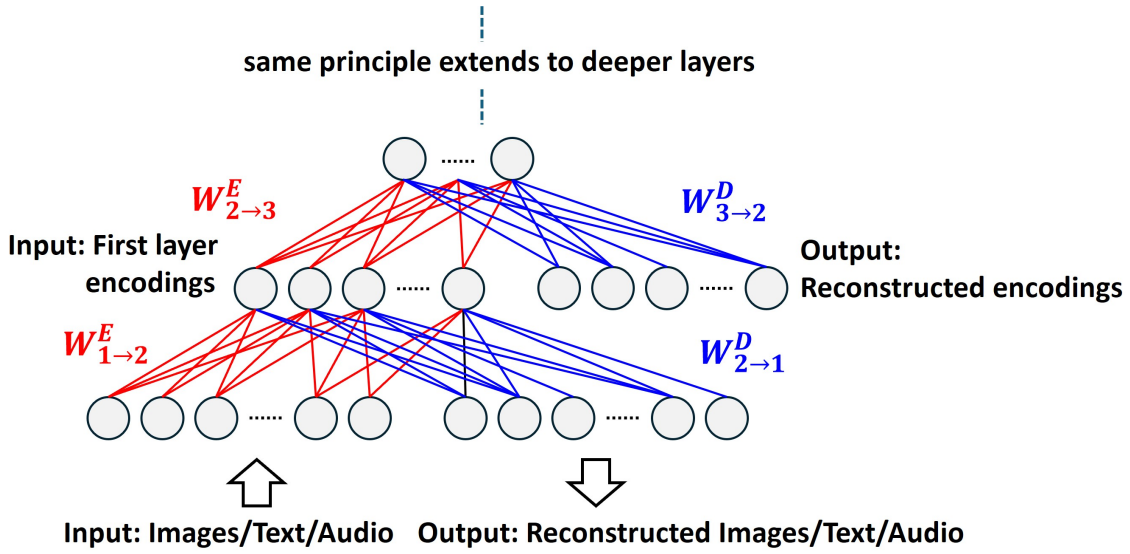


图1：堆叠自编码器神经网络的示意图。红色连接表示用于压缩输入信息的编码权重，蓝色连接表示用于从内部表征重建输入的解码权重。训练逐层进行。

基础在层级表征学习方面与深度置信网络（DBNs）的基础相平行，后者基于概率生成模型（Erhan 等，2010；Hinton 与 Salakhutdinov，2006年）。深度置信网络使用逐层训练的受限玻尔兹曼机（RBMs），而堆叠自编码器则依赖重构目标。尽管存在这些差异，两种架构都旨在发现结构化的多层次表征。本研究提出，进化可能通过堆叠自编码器过程进行，这一概念与确定性自编码器以及深度置信网络等概率生成模型均相容。

3 堆叠自编码器进化假说

正如间断平衡和跳跃式进化等理论所指出的，进化变化并非总是渐进的，有时也会以间断或突发的方式发生，只是发生的时间尺度有所不同。然而，导致这类快速进化转变的潜在遗传机制尚未得到充分阐明。本研究提出了一种新颖的进化

机制——堆叠自编码器进化假说，以期从不同角度弥补这一研究空白。

3.1 假说

堆叠自编码器进化假说的核心前提是，进化由一种基于自编码与解码的自我复制过程驱动，这与自编码器的机制类似。根据这一观点，进化并非仅通过分子复制、突变和选择的简单迭代循环推进，而是通过一种信息过程不断发展：生物体在多个表征层中，于内部维持并分层重构嵌入其分子中的遗传信息，从而实现自我复制。

该假说将自我复制概念化为一个过程：遗传分子在分子相互作用的作用下经历一系列转化，最终重构出原始分子（图 2）。如果假定这些分子携带遗传信息，那么分子转化过程就可以视为遗传信息的转化。在这种情况下，信息会根据存储它的分子，以不同维度的空间表征。这是因为转化后的分子可能比原始分子短，能够容纳的信息更少；反之，也可能更长。至关重要的是，即使新分子只能容纳较少的信息，自我复制过程仍必须设法保留原始内容，例如通过压缩。在自我复制过程中，这种通过一系列转化实现的信息压缩与重构，对应于自编码器中的自编码和解码操作。通过这一压缩与重构过程，遗传信息的关键特征得以提取，并形成越来越抽象的内部表征。

在本文的表述中，我们主要为了具体起见而使用“基因”和“遗传分子”这两个术语。

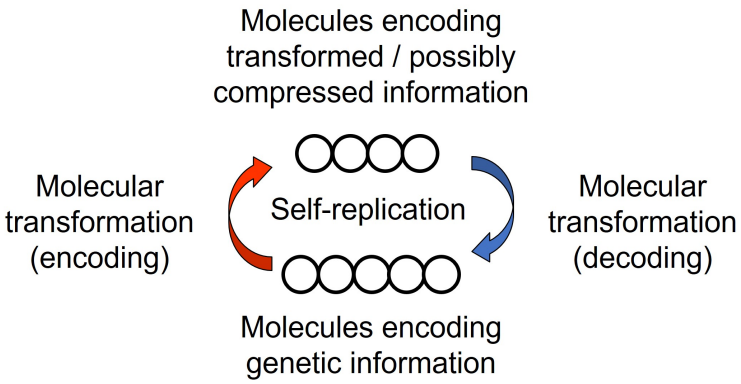


图 2：将自我复制概念化为一系列分子转化。中间分子实现遗传信息的压缩与重构，生成抽象的潜在表征，类似于自编码器架构。

然而，该假说的核心要求并不局限于基于核苷酸的基因本身，而在于存在更低层级的信息基底，这些基底经过明确编码，并参与递归转化和自我复制循环。原则上，此类基底可以由任何能够在不同步骤中被复制、转化和重构的结构化单元构成，例如类似序列的调控模块，或对发育模式、行为模式进行紧凑编码的其他形式。本文聚焦于分子层面的遗传序列，将其作为此类基底最典型的例子；至于向更高组织层级的潜在扩展，则留待未来研究。

由于压缩了信息的分子本身就是分子实体，因此它们可能参与其他自我复制过程（图 3）。当复制通过另一个占用空间更小的分子进行时，信息会得到进一步压缩。如果这个新形成的分子也参与类似过程，自我复制链便会继续延伸，从而形成连续的信息压缩阶段。这种递归复制途径类似于深度学习中堆叠自编码器的架构：遗传信息在连续层级中被编码到抽象空间，每一层都具有更高程度的抽象。值得注意的是，由于每一层都由一个独立的分子在物理上实现，突变原则上可能发生在这一层级结构的任何层面——无论是

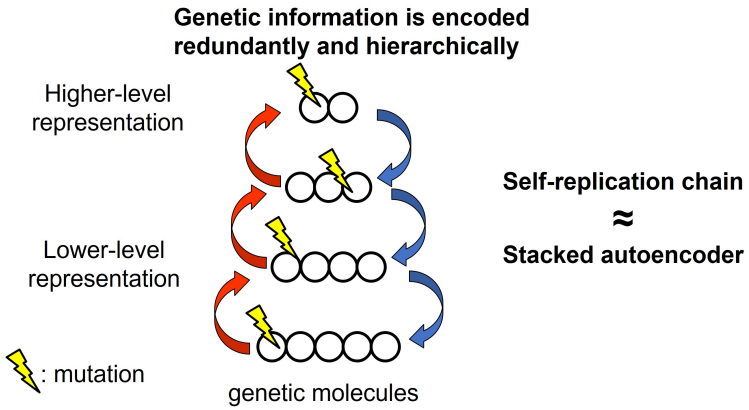


图3：在复制层级结构的不同层级引入的突变，会影响物理序列（浅层）或抽象潜在表征（深层）；深层扰动经解码后会转化为具体的分子变化，从而产生不连续的遗传新颖性。

在表层序列本身，还是在更深层、更抽象的潜在空间中。

在这种层级结构中，遗传信息冗余地保存在多个层级中，其中较深层级对应更高阶的潜在空间。在浅层引入的突变表现为对遗传序列的常规物理化学扰动，而发生在深层的突变则是在抽象特征空间中的扰动（图3）。这些深层扰动不一定对应连续的物理变化，而是使系统在更高阶的表征流形中发生迁移，进而可能促成在性质上截然不同的不连续进化创新。通过复制周期中的下游重构过程，这些抽象突变最终会被解码回具体的分子序列，从而在物理层面实现，并受到自然选择的作用。这一机制提供了一条出人意料的直接途径，使抽象潜在空间中的突变能够产生具体的遗传新颖性。

至于这类编码与解码变换如何获得，其机制与人工神经网络所采用的机制有着根本区别。人工神经网络中的层级结构通常通过逐层

自监督训练获得，而进化则通过受选择压力引导的自组织过程构建这些结构。正如先前的人工生命模拟所展示的那样（例如，（Hickinbotham 等人，2016；Ofria 和 Wilke，2004；Ray，1992；Sayama，1999）），随着时间推移，不利于自我复制的分子转化过程往往会被淘汰，而参与有效自我复制环的分子则会自然保留下来。尽管与人工智能采用的直接优化策略相比，这种进化搜索机制可能显得效率低下，但在进化时间尺度上，它仍是一种可行的过程。

这一假说提出，进化系统本质上通过一种自我复制过程运行，该过程涉及类似于堆叠自编码器的机制，从而能够跨多个层级实现高效的特征压缩与重构。这里所主张的并不是细胞内存在字面意义上的类神经网络模块，而是说，只要多个层次的结构在共享的分子表示中得到表达，一种等效的多层编码—解码组织原则上就可能涌现。从这个意义上说，这种架构是概念性的，但它指向的是一类可以实际实现的分子组织，而非纯粹的比喻性类比。因此，这一机制为以下现象提供了一种潜在解释：非连续且大尺度的进化转变——有时似乎呈现出方向性或类似目标导向的模式——可能在不诉诸真正目的论的情况下产生。

尽管本文提出的方案主要是概念性的，但原则上可以通过实证检验这一假说。堆叠式自编码器的解释预测，自我复制应涉及中间的压缩表征，并保留足够的信息以重构高层结构。这一预测可以通过多种方式进行评估：例如，分析对假定的中间表示施加扰动后，是否会使重构结构产生连贯的非局部变化；考察合成或人工化学系统是否存在多阶段信息变换的证据；或比较重构分子状态的统计几何与层级编码模型预期的统计几何是否相符——

这些研究方向为实证或计算验证提供了可能途径，尽管目前直接开展生物学检验仍然颇具挑战。

3.2 使用人工化学对该假说的概念性阐释

本节借助人工化学框架，对所提出的堆叠自编码器进化假说进行具体阐释。人工化学是一种抽象的计算系统，通过由符号分子构成的反应网络，研究自我复制和进化过程。尽管这些模型并未直接纳入生态相互作用或种群水平动力学，而这两者对于全面解释生物进化十分重要，但它们非常适合描述构成本假说核心的分子转化和自编码器机制。将生态过程和选择过程整合进这一框架，可能有助于更全面、更深入地理解进化动力学。

在该模型中，受RNA世界假说（Gilbert，1986）的启发，分子被表示为数值序列。与深度学习模型处理文本或图像等符号输入的方式类似，这里将分子视为信息的存储载体。每个分子都具有某种RNA二重性，既是存储信息的实体，也是促进其他分子发生转化的催化剂。

当分子彼此足够接近时，就会发生化学反应。这些反应并非被建模为简单的结合或交换，而是被假定为实现相当于神经网络单层的非线性变换。具体而言，催化剂分子 A 在其序列中编码权重（以及必要时的偏置），并将其应用于输入分子 B，从而产生新的输出分子 C。转化可能导致序列长度发生变化，对应于压缩或信息扩展。

当此类转化在环境约束下反复发生时，例如分子数量或分子寿命受到限制，

竞争和选择便会自然涌现，从而形成自我复制网络。在这类网络中，分子 A 和分子 B 可能相互催化对方的生成，形成一个双向反应循环，可将其解释为一种自编码器结构。

如果分子 B 比分子 A 短，那么从 A 到 B 的转化便起到编码（压缩）的作用，而从 B 到 A 的反向转化则起到解码（重构）的作用。这种结构还可以进一步按层级扩展：B 转化为更短的分子 C，C 转化为 D，依此类推。通过这样的迭代转化，便会形成多层自编码器网络。在这种层级结构中，信息会在各层之间冗余地保留。较浅的层对应较低层的物理特征，而较深的层则表示更抽象或更具功能性的特征。因此，发生在不同层的突变可能导致一系列进化后果，从细微的物理变异到功能意义的不连续转变。

这一模型并非仅仅是一种概念示意，而是对本假说所提出的分层自编码器机制如何自发形成的具体实例化。正如先前的人工生命研究所示，这种有组织的自组织确实可以通过选择过程实现。本假说认为，这些自编码器分子架构的自发形成，是促成大规模且具有语义意义的进化变化的关键条件；这类变化往往表现为不连续或目标导向。

4 堆叠式自动编码结构的涌现模拟

4.1 模拟设计与实现

以下模拟的目的并非对堆叠自编码器进化假说进行实证验证，而是要表明其两项核心结构

性假设即便在一个极简模型中也可能出现。具体而言，该模拟表明：（i）当反应空间受到约束时，自我复制可以以分子转化的闭环形式出现；（ii）不同长度分子之间的相互作用可以自发产生分层关系，这与该假说所提出的层级编码结构相一致。这些观察旨在阐明该机制在概念上的合理性，而非详细模拟生物进化。

需要指出的是，本次模拟所采用的参数设置和神经网络架构仅作为示例，旨在帮助理解这一假说。这些设计选择本身并无特殊意义，此处展示的实现也仅是其中一种示例。模型之所以简化，是因为其定位是概念性演示，而非完整的进化情景。

在本次模拟中，我们构建了一个简化的化学系统，其中包含三种分子实体，每种实体既是储存信息的序列，也是通过基于神经网络的反应催化分子转化的催化剂。为聚焦于所提机制的核心动力学，该模型省略了涉及多种分子群体的完整进化情景所具有的复杂性。这三种分子分别记为M_{13}、 M_7和 M_3，对应的序列长度分别为 13、7 和 3。对于长度为 13 的分子，第 i个实例的第 k个元素表示为 $M^i_{13,\ k}$，其中每个元素 $M^i_{:,\ k} \in [-1, 1]$ 均从均匀分布中初始化。

在此模型中，分子的“状态”表示为区间 $[-1, 1]$ 内由实值分量构成的向量。这些数值并不对应任何特定的化学性质，而是作为分子内部自由度的一种抽象编码。选择范围 $[-1, 1]$ 是为了确保数学上的一致性，因为它与催化转化中所用双曲正切函数的输出相匹配。这种连续表示使分子能够充当信息承载实体，其状态可由催化算子压缩、转化和重建。这些特定的序列长度

旨在使分子序列能够映射到神经网络参数。除了存储信息外，每个分子还充当催化剂，以实现转化过程（Ikegami & Hashimoto，1995）。从概念上说，这意味着一种类型的分子（例如 M_{13}、M_7 或 M_3）接收另一个分子作为输入，并输出三种类型之一的新分子，从而使反应网络能够支持映射 $M_{13} \to M_7$、$M_7 \to M_3$、$M_7 \to M_{13}$ 等。

实现此类转化的一种直接方法是使用全连接神经网络，其参数数量与输入维度和输出维度的乘积成正比。为避免这一问题，本模型采用卷积神经网络和反卷积神经网络（Krizhevsky 等人，2012；LeCun 等人，1998），以构建一个自包含的分子种类集合。这些架构通过在不同位置重复使用同一卷积核，大幅减少了参数数量，同时仍可通过卷积核大小、填充和步幅灵活控制输出长度。通过适当选择这些参数，每种催化转化所需的权重参数数量便可与三类分子 M_{13}、M_7 和 M_3 的长度保持一致。

这种设计确保每个催化分子都能准确编码其转化函数所需的参数，从而形成一个闭合反应循环，使所有转化都发生在有限的分子种类集合中。

模拟中使用的参数和反应设计总结于图 4。对于催化反应中的非线性变换，采用缩放后的双曲正切函数 tanh($2x$)，以扩展输出范围并促进生成更大的变异。需要再次强调的是，这种反应网络设计仅是一个可能的示例；在现实场景中，分子种类的多样性能够支持构建各种各样的网络，并以多种方式形成环路。

在每个模拟　　步骤中，模型都会计算分子之间所有可能发生的催化反应　　s.

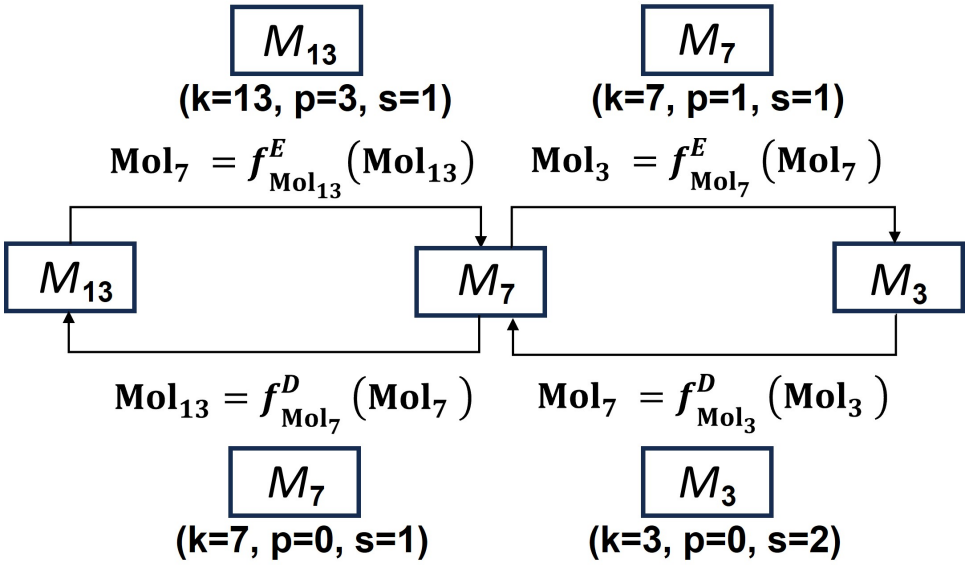


图 4：模拟中的人工化学反应网络。 $M_{\#}$ 表示长度为 # 的分子。箭头表示从 $M_{\#1}$ 到 $M_{\#2}$ 的转化，其中沿箭头放置的 $M_{\#3}$ 充当催化剂。例如，左上方的反应表示，分子M^i_{13}在另一个分子 M^l_{13}的催化作用下转化为分子 M^j_7。催化功能利用卷积网络（编码）或反卷积网络（解码）。$f^E_{M_{\#1}}(M_{\#2})$表示一个由 $M_{\#1}$ 构成、以 $M_{\#2}$ 为输入的卷积网络。催化剂分子的参数 k、 p和 s 分别表示卷积网络或反卷积网络的卷积核大小、填充和步幅。

对于种群中的每个有序对（M_X， M_Y）， M_X 被视为底物， M_Y被视为催化剂。将 M_Y 中编码的催化转化应用于 M_X的状态，生成候选产物分子 M_Z。这一过程产生了大量潜在的反应产物。

从这一产物池中，通过随机采样新分子来构建下一代分子种群。具体而言，下一代中百分之九十的分子从新生成的反应产物中采样，其余百分之十则从当前种群中复制，以维持连续性并防止多样性丧失。种群规模保持不变。

由于每个分子都编码其催化转化的参数，对分子的任何修改都会直接改变其在后续反应中施加的转化。因此，通过反复应用反应和扰动，分子状态与催化功能都在持续重塑。这种递归耦合产生了一种自组织动力学，其中分子层面的表征与反应算子协同进化。

仿真环境由一万五千个分子组成，其中 M_{13}、 M_7和 M_3每种类型各有五千个分子。所有分子均以随机值初始化，且未施加空间结构。每个反应周期结束后，都会向所有分子添加小幅随机扰动，以促进变异。通过这些反应、采样和扰动循环，该模拟考察分层自编码结构能否在分子种群中出现并趋于稳定。

4.2 模拟结果

基于上述反应网络，通过反复进行反应周期，考察了不同长度分子的分布变化。结果如图 5 所示。

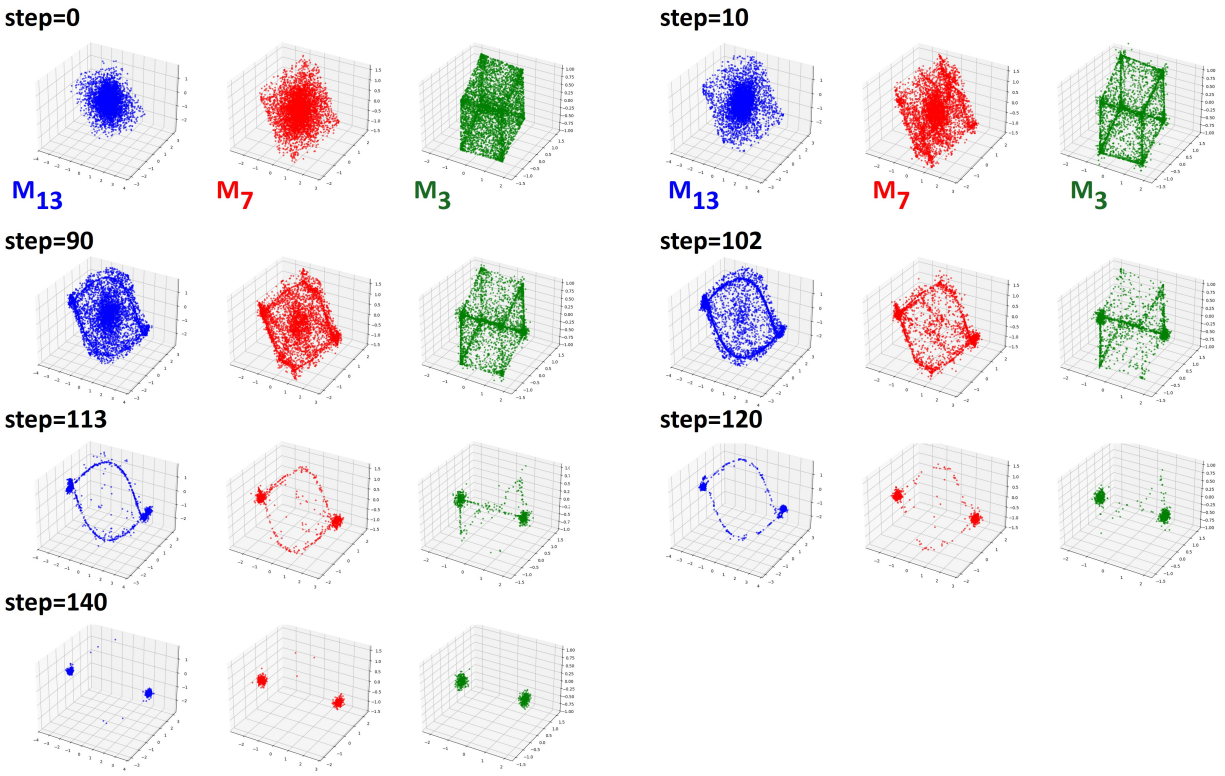


图 5：每个步骤中 M_{13}、 M_7 和 M_3 分子状态的分布。通过应用主成分分析（PCA）进行降维，将这些分布以三维形式可视化。

在初始配置中，所有分子值均在 -1 至 1 的范围内随机初始化。经过一次变换步骤后，没有观察到具有特征性的结构模式（如图中的步骤 $= 0$ 所示）。然而，随着反应的进行，在 M_3 分子种群中可以观察到一种独特的结构。图中可见的立方体状分布反映了双曲正切函数的特征行为，该函数倾向于将数值推向其输出范围的极端边界。尽管在大约 90 个步骤之前， M_7 分子种群中偶尔会出现暂时性的结构模式，这些模式似乎受到 M_3 分子结构的影响，但它们并未持续存在。从长期来看，在M_7 或 M_{13} 种群中均未观察到明显的结构组织。

有趣的是，在某个时刻（图 5 中大约第 102 步或第 113 步）， M_3 种群开始收敛为两种截然不同的状态。随后， M_7 和 M_{13} 种群也逐渐发生变化，在遵循相似趋势的同时呈现出略有不同的结构，最终所有分子种群都收敛为两个截然不同的簇。

为了评估反应网络是否已具备跨分子层编码和解码信息的能力，我们基于均方误差（MSE）测量了重构误差。该指标旨在评估原始分子中的信息经过一系列催化变换后保留得如何。重构误差较低表明，中间层保留了足够的信息，能够可靠地重构原始分子，这与堆叠自编码器的行为类似。

一般而言，评估过程包括：（i）选取一种类型的分子；（ii）使其经过一个或多个中间分子层的转化；（iii）将最终重构出的分子与原始同类型种群中的成员进行比较。这使我们能够量化信息在整个编码–解码路径中保留的程度。例如， M_{13} 分子的重构准确率按如下方式评估。取两个

M_{13}分子，分别记为 $M_{(i)\,13}$和 $M_{13}^{(j)}$，通过反应生成一个 M_7分子，记为$M_{(k)\,7}$。随后，两个这样的 M_7分子（ $M_7^{(k)}$和 $M_{(l)\,7}$）通过反应生成一个新的M_{13} 分子，记为 $M_{13}^{(new)}$ 。对于每个重构出的分子，计算其与原始 M_{13} 种群中最相近成员之间的均方误差，从而得到重构误差

$$RE_{13} = \frac{1}{5000} \sum_{i=1}^{5000} \min_j \mathrm{MSE}(M_{13}^{(new,i)}, M_{13}^{(orig,j)}). \tag{1}$$

该过程通过路径 $M_{13} \to M_7 \to M_{13}$量化重构。我们将相同的方法应用于其他路径，包括 $M_7 \to M_3 \to M_7$ 和 $M_{13} \to M_7 \to M_3 \to M_7 \to M_{13}$，从而评估多个层级转换中的重构性能。

图6显示了重建误差随时间的变化。起初，重构误差非常大，但随着系统趋近两态收敛，误差迅速减小。在最终的两态构型中，有理由认为所有层都保留了相同的信息，这一点可从重构误差的降低得到印证。

为了考察 M_3 分子是否保留了关于 M_7 和 M_{13} 分子的信息，我们使用第140步的状态，根据 M_3 分子重构了 M_7 和 M_{13} 。结果如图7所示。如图所示，成功地根据 M_3 分子重构出了 M_7 和 M_{13} 分子状态分布。这表明，指定较低层状态所需的信息已嵌入 M_3 分子的高层表示中。

这些结果表明，当自催化反应网络受到促进自我复制的选择压力时，它可以自发形成本文提出的层级信息编码与解码结构。

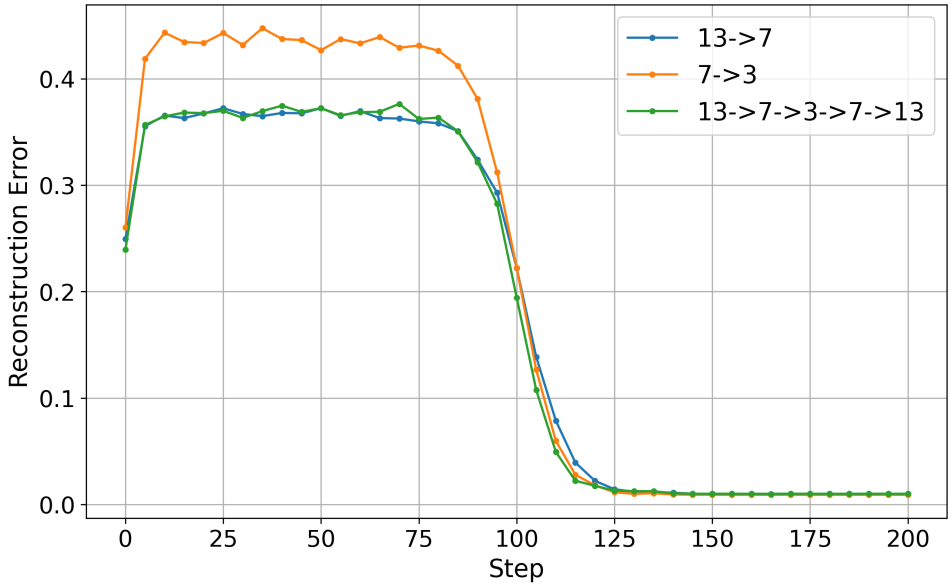


图6：每一步的重建误差。详情请参阅正文。

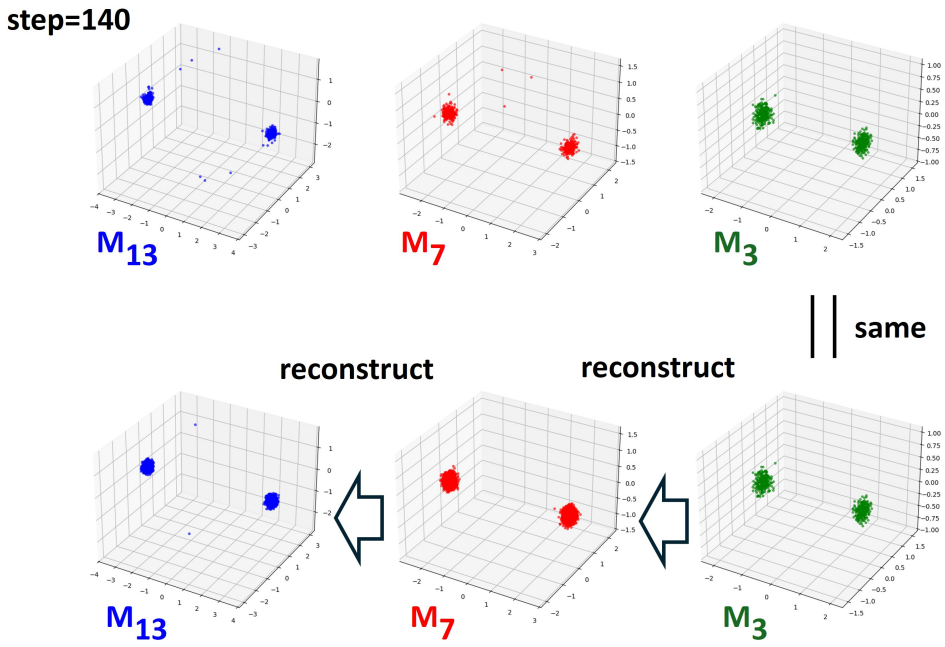


图7：从 M_3 个分子中解码 M_7 和 M_{13} 个分子的结果。上方图展示了第129步的原始分布，下方图展示了从相同的 M_3 个分子解码得到的分布。

4.3 模型的局限性

在本次模拟中，选择压力极为简单，只关注数值序列复制的难易程度。因此，没有任何驱动力促使序列复杂度增加。结果，所有层最终都收敛到二元状态。在这种简化条件下，序列本身并不需要变得更加复杂，层与层之间也没有进行信息压缩的必要。这一结果反映出本模型所定义的选择压力过于简单。

在更真实的进化情境中，情况则截然不同。尽管复制以及由此产生的有限空间或物质资源竞争仍是选择的主要来源，但这些基本压力会间接表现为各种与生存和稳健性相关的约束。例如，生物体必须维持结构完整性、抵御环境波动，并协调多种功能过程，才能继续占据生态空间。潜在竞争所带来的这些间接后果，可能产生有利于分子或发育复杂度增加的压力。在这些条件下，一旦序列变得足够复杂，本研究提出的分层信息压缩和冗余多层表征就可能更加显著，也更容易被观察到。

需要强调的是，本模型旨在作概念性演示。它略去了生化约束、物理具身以及种群层面进化动力学。此外，该假说目前缺乏直接的生物学证据，而要为所提出的多层编码架构找到分子层面的实现方式，仍是一项尚待解决的挑战。尽管如此，正如理论人工生命研究中常见的情况，这一框架的价值在于展示一种构造性可能：层级编码与解码可以在自复制系统中涌现，并且原则上可能支持不连续进化转变。通过阐明这一机制的生成潜力

及其局限性，该假说为未来针对进化中多阶段信息处理的实证、计算和综合研究奠定了基础。

5 与其他“进化即学习”假说的讨论

进化与学习之间的关系在进化生物学中长期以来一直受到探讨，并提出了各种概念框架。其中，鲍德温效应仍然是一个尤其具有影响力的概念（Baldwin，1896）。鲍德温效应描述了一种学习与遗传进化相互作用的机制，将学习本身置于自然选择的目标之下。能够通过学习进行适应的生物体获得生存优势，而在进化时间尺度上，支持这种学习的神经和发育架构会得到选择性稳定。此外，有人提出，最初通过学习获得的行为或性状最终可能会被遗传同化，使其能够通过遗传固定而无需学习便得以出现。

这一概念机制通过数学建模和模拟得以形式化。Hinton and Nowlan 展示了学习能够显著加速进化过程（Hinton & Nowlan, 1987）。在此基础上，Watson and Szathmáry 提出，进化本身可以被视为一种学习过程，在这一过程中，进化景观会逐渐被“记忆”，并由进化历史塑造（Watson & Szathmáry, 2016）。在他们看来，这种记忆不仅编码于基因之间的相关性和模块化发育机制之中，也编码于生物体通过相互作用构建并改变的生态关系之中。这一视角将进化系统视为本质上能够自我改造的系统，既能重组自身的内部发育结构，也能重组选择环境的关键方面。他们的框架假定存在一个适应度函数，用来评估各种状态的“优劣”，

从而使他们能够将进化描述为一种类似于学习的优化过程，尽管这一适应度函数未必是固定或绝对的。

Vanchurin 等人后来拓展了这一思路，提出进化不仅通过适应度优化推进，也通过最小化相对于环境的预测误差推进（Vanchurin 等人，2022）。从这一观点来看，选择有利于能够改善预测或预期行为的机制。Watson and Szathmáry（2016）和 Vanchurin 等人（2022）都强调，进化系统在多个组织层级上运作，从而形成层级适应动力学，这种动力学类似于不同尺度上的学习。

当前假说所隐含的学习过程与这些方法的侧重点不同。堆叠自编码器假说并未将进化解释为种群层面的优化或预测过程，而是提出类似学习的动力学可能已嵌入自我复制过程本身。这里，自我复制并非遗传信息的直接复制，而是一系列跨越多个表征层、对这些信息进行压缩和重构的转换过程，类似于深度学习中的层级自编码器。该过程被假设会形成多层潜在空间，对遗传信息中日益抽象（或语义化）的特征进行编码（Bengio 等人，2013；Hinton 与 Salakhutdinov，2006）。不同层级上的突变会扰动不同的抽象层次，使得连续和不连续的进化变化都能通过解码器过程发生。

这一框架并不取代基于适应度的选择，也不取代 Hinton、Watson 或 Vanchurin 提出的类学习动态，与较为传统的进化理论也并不冲突。相反，它补充了这些观点，提出自我复制本身的信息架构可能赋予进化系统某种层级表征学习能力。由此，该假说提供了一种生成机制，可将潜在表征中的微观扰动与宏观

进化转变联系起来，从而为进化如何在统一的信息过程中同时呈现连续性与非连续性，提供一种可能的途径。

6 结论

本研究提出了一种关于进化的新颖理论框架——堆叠自编码器进化假说。该假说认为，进化并非仅仅是自我复制的过程，而是一个动态的层级信息转换过程，其中遗传编码在多个抽象层级上被压缩和重构。因此，进化可能并非仅通过表层的序列层面突变发挥作用，还会通过抽象语义空间中的扰动，促成快速且适应性强的进化转变。

通过将自编码器的计算原理引入进化理论，这一假说为进化过程提供了新的计算视角。它指出，有望将传统进化理论与自监督学习、表征学习等先进计算框架相结合；这些框架近年来在机器学习领域得到了积极发展。例如，这一假说提出了一种可能性：Transformer模型等先进神经架构——它们在自然语言处理领域取得了显著成功——或许也内在地嵌入于进化过程中。正如人工智能系统能够生成多样且有意义的语言输出，进化本身或许也已具备在语义空间中运作的能力，而这一空间如同人类语言一样丰富且结构严谨。

与间断平衡或跳跃式进化等用以刻画快速变化模式的描述性解释不同，本框架明确了一种具体机制，借此可以产生协调一致的、全系统范围的新性状。具体而言，较深潜在层中的突变会扰动压缩表征；这些表征一经解码，便会同时在许多性状上引发结构化变化，从而将微观信息

扰动与单一生成过程中的宏观进化转变联系起来。这一假说将进化重新理解为植根于层级信息变换的涌现计算过程。通过超越随机突变与选择的局限，它为大规模且具有语义结构的进化转变如何从遗传编码的信息架构中产生提供了一种生成性解释。

代码可用性

本研究中用于模拟和分析的代码已公开，网址为：
`https://github.com/HiroIizuka/SAEvo`。

参考文献

Asquith, P. J. (2007). 今西锦司关于社会性和进化结果观点的来源。《生物科学杂志》，*32*(4), 635–641.

Baldwin, J. M. (1896). 进化中的一个新因素。《美国博物学家》，*30*, 441–451.

Bengio, Y., Courville, A., & Vincent, P. (2013). 表征学习：综述与新视角。*IEEE*《模式分析与机器智能汇刊》，*35*(8), 1798–1828.

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2006年). 深度网络的贪心逐层训练。神经信息处理系统进展，*19*。

Carroll,S. B. (2005). 《无尽的美丽形态》：进化发育生物学的新科学与动物界的形成. WW Norton&Company.

Darwin, C. (1859). 《物种起源》，*1859*年。伦敦：John Murray。

Eldredge, N., Gould, S. J., 等（1972）。间断平衡：系统渐变论的另一种解释。古生物学模型，*82*，115。

Erhan, D., Courville, A., Bengio, Y., & Vincent, P.（2010）。为什么无监督预训练有助于深度学习？第十三届人工智能与统计国际会议论文集，201–208。Gilbert, W.（1986）。生命起源：RNA世界。《自然》，*319*（6055），618–618。Goldschmidt, R.（1940）。《进化的物质基础》。（由耶鲁大学出版社于1982年重印。Hickinbotham, S., Clark, E., Nellis, A., Stepney, S., Clarke, T., & Young, P.（2016）。在自动机化学中最大化相邻可能。*Artificial Life*，*22*（1），49–75。Hinton, G. E., & Nowlan, S.（1987）。学习如何引导进化。复杂系统，*1*（3），495–502。Hinton, G. E., & Salakhutdinov, R. R.（2006）。利用神经网络降低数据的维度。《科学》，*313*（5786），504–507。Ikegami, T., & Hashimoto, T.（1995）。机器与磁带自我复制网络中的主动突变。*Artificial Life*，*2*（3），305–318。Kauffman, S. A.（2000）。《探索》。牛津大学出版社。Kimura, M.（1979）。中性理论。《科学美国人》。Krizhevsky, A., Sutskever, I., & Hinton, G. E.（2012）。使用深度卷积神经网络进行ImageNet分类。神经信息处理系统进展，*25*。Langton, C. G.（1984）。元胞自动机中的自我复制。*Physica D*：非线性现象，*10*（1-2），135–144。LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P.（1998）。应用于文档识别的基于梯度的学习。《*IEEE*会刊》，*86*（11），2278–2324。Masci, J., Meier, U., Cireşan, D., & Schmidhuber, J.（2011）。用于层次化特征提取的堆叠式卷积自编码器。人工神经网络与机器学习–*ICANN 2011*：第*21*届人工神经网络国际会议，芬兰埃斯波，*2011*年*6*月*14—17*日，会议论文集，第*I*部分 *21*，52–59。

McMenamin, M. A. (2013). 寒武纪大爆发：动物生物多样性的构建。*BioScience*, *63*(10), 834–835. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A. Y., et al. (2011). 多模态深度学习。*ICML*, *11*, 689–696. Ofria, C., & Wilke, C. O. (2004). Avida：用于计算进化生物学研究的软件平台。 *Artificial life*, *10*(2), 191–229. Ray, T. S. (1992). 一种生命合成方法。 *Proceedings of Artifcial Life*, *2*, 371– 408. Salzberg, C. (2003). 自我复制元胞自动机的涌现进化动力学。硕士论文，阿姆斯特丹大学，阿姆斯特丹，荷兰。 Sayama, H. (1999). 在简单的元胞自动机空间中演化出一种新型、结构上可解体的自我复制环。 *Artificial Life*, *5*(4), 343–365. Schluter, D. (2000). 适应辐射的生态学。牛津大学出版社。 Tria, F., Loreto, V., Servedio, V. D. P., & Strogatz, S. H. (2014). 相关新颖性的动力学。 *Scientific reports*, *4*(1), 5890. Valentine, J. W. (2004). 门的起源。芝加哥大学出版社。 Vanchurin, V., Wolf, Y. I., Katsnelson, M. I., & Koonin, E. V. (2022). 探索将进化视为多层次学习的理论。 美国国家科学院院刊,*119*(6), e2120037119. Varela, F. J. (1979). 生物自主性原理。Elsevier North Holland，纽约，纽约州。 Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., & Bottou, L. (2010). 堆叠去噪自编码器：运用局部去噪准则，在深度网络中学习有用的表征。 机器学习研究期刊, *11*(12). Watson, R. A., & Szathmáry, E. (2016). 进化如何学习？ 生态学与进化趋势, *31*(2), 147–157.