写给物理学家的生成模型*

王 磊[1,†] 张 潘[2]
(1 中国科学院物理研究所 北京 100190)
(2 中国科学院理论物理研究所 北京 100190)

2024－05－16收到
† email：wanglei@iphy.ac.cn
DOI：10.7693/wl20240602

Generative models for physicists

WANG Lei† ZHANG Pan
(1 Institute of Physics，Chinese Academy of Sciences，Beijing 100190，China)
(2 Institute of Theoretical Physics，Chinese Academy of Sciences，Beijing 100190，China)

摘 要 科学研究的本质在于创造。生成式人工智能为更有创意的科学探索打开了无尽的想象空间。作为生成式人工智能的核心，生成模型学习数据样本背后的概率分布，并据此随机采样生成新的样本。生成模型和统计物理在本质上是同一枚硬币的两面。文章从物理的视角介绍扩散模型、自回归模型、流模型、变分自编码器等现代生成模型。生成模型在原子尺度物质结构的生成与设计中展现出巨大的潜力。不仅如此，基于和统计物理的内在联系，生成模型对于优化“大自然的损失函数”——变分自由能具有独特的优势，这为求解困难的统计物理和量子多体问题提供了新的可能。同时，物理学的洞察也在推动生成模型的发展和创新。通过借鉴物理学原理和方法，还可以设计出更加高效、更加统一的生成模型，以应对人工智能领域中的挑战。

关键词 生成模型，统计物理，相对熵，最大似然估计，变分自由能

Abstract The essence of scientific research is about creation. Generative artificial intelligence (AI) has opened up an endless space of imagination for more creative scientific research. As the core of generative AI, generative models learn the underlying probability distribution from data, and then randomly sample from it to generate new samples. Generative models and statistical physics are essentially two sides of the same coin. This article introduces modern generative models including diffusion models, autoregressive models, flow models, and variational autoencoders from a physics perspective. Generative models demonstrate tremendous potential in the generation and design of materials at the atomic scale. Moreover, based on their inherent connection with statistical physics, they have a unique advantage in optimizing Nature’s cost function, the variational free energy, thereby providing new possibilities for solving difficult problems in statistical physics and quantum many-body systems. At the same time, physical insights are also driving the development and innovation of generative models. By drawing inspiration from physical principles and methods, more efficient and unified generative models can be designed to address challenges in the field of AI.

Keywords generative models，statistical physics，relative entropy，maximum likelihood estimation，variational free energy

* 国家自然科学基金(批准号：T2225018；92270107；12325501)资助项目

1 引言

费曼在他的黑板上留下一句话：“what I can not create，I do not understand”[1]。三十年后，这句话被如今大红大紫的OpenAI当作信条挂在网站上[2]。确实，无论是在物理学还是人工智能的研究中，有能力创造往往才意味着最高层次的理解。

什么是创造？生成式人工智能(Generative AI)对此的回答是：学习数据样本背后的概率分布，并通过随机采样生成新的样本。这两年，人们用人工智能产品创造了无数令人惊艳的画作、引人入胜的故事、动人心弦的音乐。当然，还有广告文案、新闻报道、审稿报告、商业计划书、推荐信等等(例如这句话本身)。人工智能创造的源泉是数据本身，而人工智能创造的引擎则是“生成模型”：一种用于表达、学习和采样数据背后的概率分布的人工神经网络。生成模型和统计物理的关系非常紧密。一旦了解生成模型的物理学基因，就比较容易理解和改造它们，甚至发明新的生成模型。本文从物理学的角度介绍几类常见的生成模型，并举例说明它们在科学研究中的应用。

相对于性质预测之类的“判别式”任务，“生成式”人工智能更难、更基础、也更有用。用数学语言描述，性质预测的目的是拟合函数$y=f(\boldsymbol{x})$。这里$\boldsymbol{x}$是神经网络的输入，通常是代表微观结构的高维变量。y是输出，通常是代表宏观性质的低维变量。在性质预测之外，人们往往还更关心从宏观性质到微观结构的反向设计问题。由于从结构到性质的函数不可逆，简单地寻找它的反函数往往不能成功。概率建模提供了一个有用的视角。这时，性质预测就是要学习条件概率分布$p(y|\boldsymbol{x})$。而反向设计意味着给定宏观性质y，从条件概率$p(\boldsymbol{x}|y)$中采样生成新的微观构型$\boldsymbol{x}$。贝叶斯公式告诉我们$p(\boldsymbol{x}|y)\propto p(\boldsymbol{x})p(y|\boldsymbol{x})$。可见，把握微观构型的概率分布$p(\boldsymbol{x})$是“生成式”任务区别于“判别式”任务的关键。

图1形象地展示了生成模型在做什么。图1(a)中的蓝色点代表数据样本，生成模型需要学会数据背后的概率分布，并据此生成新的样本。生成模型能够处理的数据类型无所不包。如图1(b)所示，图像生成模型表达了像素取值的概率分布，而材料生成模型表达了原子类型和坐标的联合概率分布。可以想象，无论是所有像素取值所构成的图像空间，还是所有原子种类和排布所构成的材料空间，都巨大无比。而人们真正关注的自然图片和稳定材料，仅仅占据这些空间中一个小角落。生成模型需要尽量提取数据样本中的统计规律，才能生成浑然天成的新图片和新材料。

2 自然界中的概率分布

自玻尔兹曼以来，物理学对于自然界的描述就告别了决定论。微观世界充满了随机性，理解它自然需要掌握微观变量的联合概率分布。例如，水和冰由同样的水分子所组成，在基本组成单元上没有区别。但由于在不同温度下水分子构象$\boldsymbol{x}$的联合概率分布$p(\boldsymbol{x})$是不同的，因此在宏观上展现出不同的物理性质。另一个例子是统计物理中的经典伊辛模型，其中的自旋构型服从玻尔兹曼分布$p(\boldsymbol{x})=\frac{1}{Z}\mathrm{e}^{-\beta E(\boldsymbol{x})}$，其中$E(\boldsymbol{x})$是伊辛构型$\boldsymbol{x}$的能量，配分函数$Z$是概率分布的归一化因子。不同的温度$1/\beta$会导致不同的玻尔兹曼分布，甚至给出截然不同的磁化强度、比热等宏观物理量。

一般情况下，想要严格计算统计物理问题中

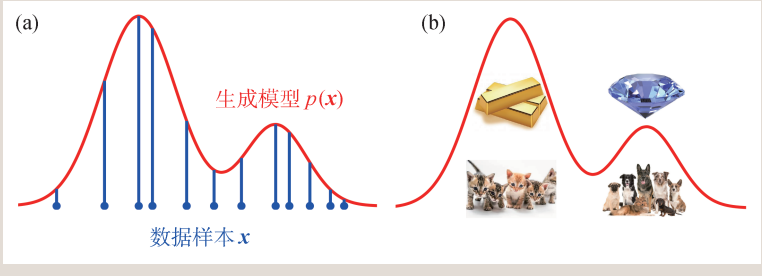

图1 (a)生成模型学习数据背后的概率分布，并据此采样生成新的样本；(b)在像素空间，不同类别的图片处于概率分布的不同模式。类似地，金属与绝缘体中的原子排布也处于晶体生成模型概率分布的不同模式

微观变量的概率分布是非常困难的，因为变量所处的概率空间太大了，归一化因子很难计算。以包含n个自旋的伊辛模型为例，直接计算配分函数需要做2^n次加法。哪怕在很少的变量数目n下，都需要很大的计算量才能完成：利用笔记本电脑我们可以计算得到n=30的配分函数，使用计算集群可以算到n=40个伊辛自旋，当n大于60时，即使用超级计算机也无法直接严格计算配分函数，因此也就不知道某个自旋构型的概率取值。在机器学习应用中，我们也面临类似的问题：一张图片、一段文本、一盘棋局都可以类比于伊辛模型的微观构型。它们背后都有各自关于像素、词和棋子位置的联合概率分布，而所有这些概率分布都存在于指数高维的空间中。

Box 1　相对熵的AB面

相对熵是信息论的一个基本量，也被称为Kullback—Leibler散度。它度量了概率分布之间的相似度。对于两个归一化的概率分布q和p，相对熵总是非负的[3]：

$$\mathrm{KL}(q\,||\,p)=\int \mathrm{d}\boldsymbol{x}\, q(\boldsymbol{x})\left[\ln q(\boldsymbol{x})-\ln p(\boldsymbol{x})\right]\geqslant 0\ , \qquad (1)$$

其中等号仅在两个概率分布相等时取到。因此，相对熵常常被用作训练生成模型的目标函数。通过最小化生成模型所表达的概率分布和目标概率分布之间的相对熵，可以让生成模型学会目标概率分布。

为什么要选择相对熵这个看起来奇怪的度量，而不直接使用类似于$\int \mathrm{d}\boldsymbol{x}\,|p(\boldsymbol{x})-q(\boldsymbol{x})|^2$这样的均方差度量？主要有两个原因：第一，概率分布$p(\boldsymbol{x})$和$q(\boldsymbol{x})$在数量级上可能差别巨大，均方差距离不能充分体现这点差别，而相对熵中比较概率函数的对数可以更好地处理这种数量级差异；第二，其实不值得在两个概率分布都取值极其小的区域比较它们的差异，因此，相对熵的定义中按照概率$q(\boldsymbol{x})$加权计算，即集中检查对于概率$q(\boldsymbol{x})$而言有意义的空间中两个概率分布对数的差别。

从定义(1)可见，相对熵对于概率分布q和p的互换并不是对称的。相对熵的AB面刚好对应了生成模型在数据建模和理论计算中的两种应用，见下表。首先，最小化数据集$\mathcal{D}$所代表的经验概率分布和模型分布之间的相对熵，等价于最小化以下的目标函数：

$$\mathcal{L}=-\sum_{x\in\mathcal{D}}\ln p(\boldsymbol{x})\,. \qquad (2)$$

因为$\ln p(\boldsymbol{x})$在统计学中被称为对数似然函数，最小化(2)式也被称为最大似然估计(maximum likelihood estimation)。注意，训练数据仅仅是目标概率分布中有代表性的样本，而不是目标分布本身。因此，过分地优化(2)式会导致过拟合现象。以图1(a)为例，如果模型学到的概率分布仅仅在蓝色数据点上非零，它就只会死记硬背训练数据，而不能再生成新的样本。

其次，在统计物理研究中人们往往知道体系的能量函数$E(\boldsymbol{x})$，而需要得到的是服从玻尔兹曼分布的样本$\boldsymbol{x}$以及配分函数Z。这种场景和数据驱动的最大似然估计恰恰相反。此时，可以将模型分布$p(\boldsymbol{x})$当作变分概率分布，并最小化它和物理系统的玻尔兹曼分布之间的相对熵。这等价于变分自由能：

$$F=\int \mathrm{d}\boldsymbol{x}\, p(\boldsymbol{x})\left[\frac{1}{\beta}\ln p(\boldsymbol{x})+E(\boldsymbol{x})\right]\geqslant -\frac{1}{\beta}\ln Z\ . \qquad (3)$$

其中不等号来自于相对熵的非负性$\mathrm{KL}\left(p||\dfrac{\mathrm{e}^{-\beta E(x)}}{Z}\right)\geqslant 0$。(3)式中两项的物理含义分别是变分概率分布的熵和能量期望值。当不等式取等号时，变分概率分布等于真实的玻尔兹曼分布，变分自由能计算也就严格地解决了问题。注意到变分计算并不依赖于事先准备好的训练样本，因为样本可以从生成模型概率分布$p(\boldsymbol{x})$中采样得来。此外，变分计算也不需要担心过拟合，变分自由能这个目标函数值越低越好。

生成模型的最大似然估计和变分自由能计算是同一枚硬币的两面

	最大似然估计((2)式)	变分自由能((3)式)
目标函数	$\mathrm{KL}(\mathcal{D}\,||\,p)$	$\mathrm{KL}\left(p||\dfrac{\mathrm{e}^{-\beta E(x)}}{Z}\right)$
已知量	数据样本	能量函数
未知量	概率分布	数据样本、配分函数

另外一个不太显然，但非常重要的联系是统计物理与贝叶斯推断的对应关系。在贝叶斯推断中，我们需要用一个模型描述一组观测数据$\mathcal{D}$。这里所谓“模型”是指给定参数$\boldsymbol{\theta}$预测数据的条件概率$p(\mathcal{D}|\boldsymbol{\theta})$，而所谓“推断”就是要确定如何根据观测数据选取模型参数。贝叶斯推断考虑对于给定的数据和参数的先验概率分布$p(\boldsymbol{\theta})$，从“后验概率”$p(\boldsymbol{\theta}|\mathcal{D})=\dfrac{1}{p(\mathcal{D})}p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$中选择参数。稍微改写一下后验概率的表达式，可以看到贝叶斯推断和统计物理之间的对应关系，见表1。其中，后验概率的归一化因子$p(\mathcal{D})$对应于配分函数Z，在贝叶斯推断中被称为边际似然，它描述了模型对数据的总体刻画能力。

表1　贝叶斯推断与统计物理之间的字典

	贝叶斯推断	统计物理
已知量	数据样本$\mathcal{D}$	能量函数$E(\boldsymbol{x})$
随机变量	模型参数$\boldsymbol{\theta}$	微观构型$\boldsymbol{x}$
采样分布	后验概率$p(\boldsymbol{\theta}\|\mathcal{D})$	玻尔兹曼分布$p(\boldsymbol{x})$
归一化因子	边际似然$p(\mathcal{D})$	配分函数Z

上面列举的对应关系也暗示了机器学习中涉及的不少问题都具有和统计物理问题相似的计算难度：计算高维概率分布函数的归一化因子涉及到理论计算机科学中所谓“#P难”的一类计数问题，一般被认为不存在快速(多项式时间复杂度)的求解算法。下文介绍的几种现代生成模型“各显神通”，巧妙地绕过了这个难题。结合Box 1中所介绍的训练目标函数，生成模型为高维变量的概率建模和变分计算打开了新的可能。

3　生成模型速览

我们从物理学的视角介绍图2所示的几类常见的生成模型。读者如果想要更加全面地了解生成模型，可以参考近期出版的教科书[4—6]。

3.1　玻尔兹曼机和扩散模型

玻尔兹曼机是一种经典的、由统计物理启发的生成模型[7]。它将模型分布表达为玻尔兹曼分布的形式(不妨令β=1)：

$$p(\boldsymbol{x})=\frac{\mathrm{e}^{-E(\boldsymbol{x})}}{Z}\ . \tag{4}$$

这种形式自然地将统计物理理论和计算方法引入概率建模。在统计物理中，反伊辛问题也是要从构型样本反推能量函数的形式[8]。例如，在低温下铁磁耦合的伊辛模型自旋全上和全下两个低能构型最容易被采样到，这就像是铁磁伊辛模型可以“记住”全黑和全白的图片。如果调节伊辛模型的耦合参数得到更多的低能构型，就会“记住”更多的图片，从而学会更加复杂的数据概率分布。

玻尔兹曼机虽然跟物理学关系紧密，但并不是一种好用的生成模型。首先，注意到(4)式分母的归一化因子Z，即配分函数，在一般情况下是难以计算的。这就给玻尔兹曼机的学习出了个难题。此外，就算我们能学到准确的能量函数，从所对应的玻尔兹曼分布中采样也往往是困难的。在这点上，生成模型采样和统计物理计算模拟的经验是一致的。例如，基于能量函数随机采样可能会面临弛豫时间长、遍历性差的问题，表现为样本长时间卡在图1(b)所示的某个分布模式中出不来。

扩散模型避免了玻尔兹曼机学习和采样中的两个难题。首先，考虑对数似然函数对于连续数据的梯度，可以绕开难以计算的配分函数：

$$\boldsymbol{f}(\boldsymbol{x})=\nabla\ln p(\boldsymbol{x})=-\nabla E(\boldsymbol{x})\ . \tag{5}$$

这个量在统计学中被称为“得分函数”(score function)，它的物理含义其实就是力。在仅仅给定数据样本的情况下，一种巧妙的估计得分函数的办法是对数据略加扰动，然后计算恢复力。因此，利用神经网络去学习式(5)被称为去噪得分匹配(denoising score matching)[9]。其次，为了缓解采样的困难，借鉴模拟退火算法的精神[10, 11]，使用噪声强度逐渐减小的退火朗之万动力学过程来采样。为此，在训练模型的过程中，相应地使用强度逐渐增加的噪声逐步扰动数据，并训练神经网络学习不同噪声水平所对应的得分函数。因为退火朗之万方程的采样过程相当于物理上的反向扩

散[12]，因此这类模型也被称为扩散模型。最近几年，扩散模型在图像、视频和分子结构的生成中大放异彩。从Sora到AlphaFold3中都有扩散模型的影子。

研究微观世界物质结构的标准手段之一是分子动力学模拟，它的出发点是物理体系的能量和力。而生成模型则绕过了基于物理原理的计算模拟，直接基于自然界已有数据生成新的微观物质结构。有意思的是，扩散模型的生成过程和分子动力学模拟神似，但又不一样。因为得分函数未必是真实的力，扩散模型在生成过程中允许抄近路，经过非物理的中间构型。反而基于真实力场的分子模拟往往受限于分子构象空间中崎岖的势能面，面临采样的困难。

3.2 自回归模型

自回归模型将高维随机变量$\boldsymbol{x}$的联合概率分布分解成一串低维变量的条件概率的连乘：

$$p(\boldsymbol{x})=p(x_1)p(x_2|x_1)p(x_3|x_1,x_2)\cdots\ . \quad (6)$$

这种分解的好处是每个条件概率因子都是低维的，更容易建模和计算。例如，一段文字$\boldsymbol{x}$的可能性是指数大的，而每个词x_i的可能性仅仅是词汇量那么大。只要每个条件概率因子是归一化的，就可以保证联合概率分布的归一化。此外，采样联合概率生成新的文本也是直接的：可以按照变量的顺序逐词生成，每步都基于上文信息预测下一个词的概率。因为预测条件概率$p(y|\boldsymbol{x})$被称作“回归”任务，逐词预测整个序列被称作“自”回归模型。原则上，形如(6)式的自回归概率分解对于任何变量顺序都成立。但由于在实践中通常使用神经网络来拟合条件概率，并通过共享不同条件概率因子的神经网络参数降低参数量，人们往往希望设计一种变量顺序使得分解后的条件概率契合数据的真实特征，从而更容易建模。

自回归模型天然适合语言数据序列性的特点，所以在自然语言概率建模中应用广泛。著名的GPT (Generative Pre-Trained Transformer)就是这样一类自回归生成模型。其中，Transformer[13]是一种对深度学习引擎(GPU和反向传播算法)都很友好的神经网络结构。当被用作自回归语言模型时，Transformer的输入是语言序列中每一个词的向量表示，而输出则是下一个词的概率表。并且，通过掩码操作使得输出相对于输入的雅可比矩阵是下三角形的，以保证模型概率的自回归性质。Transformer善于把握语言数据中的长程关联，是当下主流的语言模型架构。不仅如此，自回归模型的应用场景也完全可以是图像、棋局，或者是伊辛模型。只要对数据变量指定顺序，就可以用自回归模型对数据变量的联合概率建模，从而利用Transformer强大的学习能力逼近数据的联合概率分布，并生成新的数据样本。

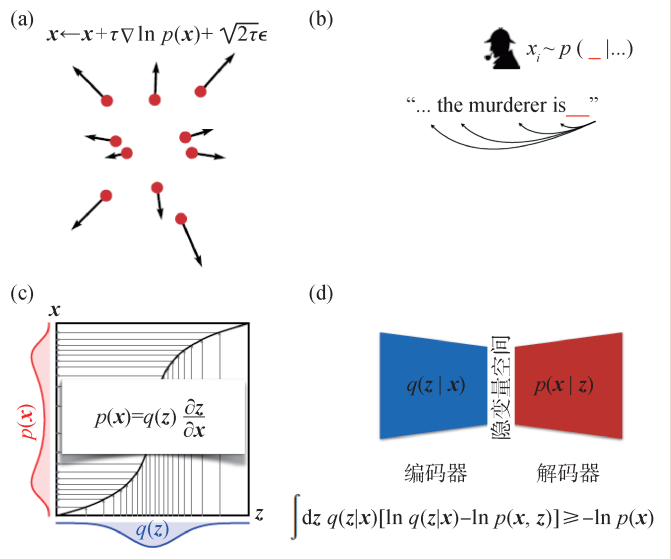

图2 四类生成模型和它们背后关键的数学公式 (a)扩散模型：朗之万方程；(b)自回归模型：条件概率分解；(c)流模型：变量替换；(d)变分自编码器：变分贝叶斯推断

自回归模型的程序实现并不复杂。仅仅只用60行的Python代码[14]就可以完全复现OpenAI开发的GPT2语言模型[15]的全部推断计算。Ondřej Čertík曾是美国洛斯阿拉莫斯国家实验室里一名从事计算物理研究的科学家，他看到了GPT2的“60行”代码实现忍不住动手试了一把，还通过把代码改写成Fortran语言实现了加速。完成这些实验之后，Čertík兴奋地在他的博客[16]中写道：“GPT2看起来就像一段典

型的计算物理代码，和我在整个职业生涯中开发和维护的许多代码类似”。“源码面前，了无秘密”[17]，请感兴趣的读者也去阅读、运行、调试一段自回归语言模型的源代码，一定会让你对语言模型产生不一样的体会。

当自回归语言模型越变越大之后，人们观察到两项有趣的现象：“能力涌现”和“标度率”。首先，模型似乎在某一规模突然获得某种能力[18]。这难免让人们产生一些浪漫的联想，例如将语言模型的能力涌现与相变现象联系起来，或者担心将来模型也会突然获得人类无法掌控的能力。此外，人们观察到当数据、模型和算力同步增加时，训练所得到的对数似然目标函数按照幂律降低。这种标度行为跨越了几个数量级，并且不依赖于自回归语言模型具体实现细节。引用论文[19]的原话：“如果能找到一个理论框架，可以推导出这些标度关系，那将是令人兴奋的！那就像是为我们所观察到的‘热力学’行为找到了‘统计力学’解释”。“涌现现象”和“标度率”这两个现象看起来相向而行，但它们都成为了追求“大”语言模型的核心驱动力。调和其中矛盾的关键似乎在于建立大语言模型能力和损失函数之间直接的联系[20, 21]。无论如何，大语言模型中所观察到的“能力涌现”和“标度率”是值得物理学家思考的问题。

3.3 流模型

流模型[22]通过可逆的神经网络实现数据到隐变量的映射$\boldsymbol{x}\leftrightarrow\boldsymbol{z}$，从而参数化概率分布：

$$p(\boldsymbol{x})=q(\boldsymbol{z})\left|\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}\right| . \tag{7}$$

这里源分布$q(\boldsymbol{z})$通常被选取为归一化因子并且便于采样的简单概率分布，例如正态分布。通过对源分布中的样本做变换，可以直接生成服从流模型概率分布的数据样本。神经网络变量替换的作用是在服从正态分布的样本中引入关联。(7)式中的雅可比行列式保证了流模型概率分布的归一化。

OpenAI开发过一个用于图像生成的流模型Glow[23]。除了直接生成自然的人脸图片之外，Glow模型还可以生成两张人脸图片之间的“平均人脸”。为此，先将两张图片分别映射到隐变量空间，再对隐变量插值，最后将插值结果映射回像素空间得到最后结果。可以想象，如果不利用隐变量分布简单便于插值的特点，而是直接将两张人脸图片的像素取平均，则不能得到一张自然的人脸图片。流模型可以表达数学和物理研究中常见的傅里叶变换、小波变换、正则变换等。不仅如此，流模型还将这些变换推广到了非线性、可学习的情况。例如，神经网络重正化群变换[24]和正则变换[25]通过学习流模型来识别多体物理系统中非线性的集体变量和慢自由度。

为了进一步揭示流模型背后的物理内涵，可以考虑公式(7)的连续时间极限。考虑用无穷小的变量替换$\boldsymbol{x}=\boldsymbol{z}+\tau\boldsymbol{v}$，再考虑$\tau\to 0$的连续时间极限，可以得到一个关于含时概率密度的连续性方程[26, 27]：

$$\frac{\partial p(\boldsymbol{x},t)}{\partial t}+\nabla\cdot[\,p(\boldsymbol{x},t)\boldsymbol{v}\,]=0 . \tag{8}$$

其中的含时概率密度在起始和终止时刻分别对应于源分布和模型概率分布。这给深层神经网络一个物理图像：它是初态概率分布流向末态概率分布的管道。

连续视角还更好地揭示了流模型和扩散模型的关系。前文提到，扩散模型的生成过程用到了退火朗之万方程，其概率密度演化遵循福克—普朗克方程$\frac{\partial p(\boldsymbol{x},t)}{\partial t}+\nabla\cdot[\,p(\boldsymbol{x},t)\boldsymbol{f}\,]-\nabla^2 p(\boldsymbol{x},t)=0$，其中$\boldsymbol{f}$是(5)式中介绍的得分函数。只要令(8)式中的$\boldsymbol{v}=\boldsymbol{f}-\nabla\ln p(\boldsymbol{x},t)$[28, 29]，就能发现两类模型具有统一的数学形式。这说明其实任何扩散模型背后都隐藏了一个流模型。因此，哪怕是在扩散模型如日中天的时候，也有一小波流模型的信徒坚信，只要找到正确的训练方式，流模型可以做得一样好，甚至更好。终于，2022年的秋天，几篇工作[30—32]几乎同时发现了一种新的训练连续时间流模型的方法：流匹配(flow matching)。流匹配方法受到了得分匹配方法的启发，但是它所训练得到的流模型可以比扩散模型更加简洁、灵活、高效，对于追求定量精确的科学问题尤为有用。

表2　变分自编码器与变分自由能之间的对应关系

	变分自编码器((9)式)	变分自由能((3)式)
积分变量	z	x
归一化因子	$p(x)$	Z
被积函数	$p(x,z)$	$e^{-\beta E(x)}$
目标函数	$\mathrm{KL}\left(q(z\|x)\|\frac{p(x,z)}{p(x)}\right)$	$\mathrm{KL}\left(p(x)\|\frac{e^{-\beta E(x)}}{Z}\right)$

例如，计算体系在不同构象下的自由能差在很多物理、化学和生物的应用中都非常关键。但是传统计算方法往往依赖于人为设计某种路径，使得在计算模拟过程中构象能够连接起来。而使用流匹配方法估计构象之间的自由能差[33]，仅仅需要不同构象下各自的数据样本，具有很强的通用性。

3.4　变分自编码器

自编码器的基本想法是把数据送进一个有瓶颈的神经网络并尽量重构它们。经过训练后，网络的前半部分是一个编码器，它将数据x转换成隐变量z。网络的后半部分是一个解码器，它将隐变量z变换为数据x。神经网络中存在瓶颈通常意味着隐变量的维度比原始数据更低，这会迫使神经网络不能够简单地把输入复制到输出。

为了计算自编码器所表达的概率，需要在数据和隐变量的联合概率分布中积掉隐变量：$p(x)=\int \mathrm{d}z\, p(x,z)$。可是，对于隐变量积分并不好算，这和计算表1中的配分函数和边际似然的难度类似。幸运的是，注意到与(3)式类似的变分自由能原理，我们有：

$$\mathcal{L}(x)=\int \mathrm{d}z\, q(z|x)\left[\ln q(z|x)-\ln p(x,z)\right] \geqslant -\ln p(x)\ . \tag{9}$$

因此，可以通过最小化式(9)来近似地最小化负对数似然函数。这里变分分布$q(z|x)$代表了从数据x推断隐变量z的编码器。仅当(9)式取到等号时，$q(z|x)$严格等于贝叶斯后验概率$p(z|x)=\frac{p(x,z)}{p(x)}$。另一方面，联合概率分布也可以写成$p(x,z)=p(x|z)p(z)$。这里$p(z)$是隐变量的先验分布，一般被选取为简单的概率分布，而$p(x|z)$是解码器，相当于一个从低维隐变量空间出发生成高维数据的生成模型。(9)式这样的不等式是变分自编码器名字中“变分”一词的来源。表2总结了变分自编码器和变分自由能(3)式的对应关系。

变分自编码器的低维隐变量空间给不少科学应用提供了便利。例如，将化学分子编码到化学隐变量空间[34]，再利用对于隐变量的梯度下降来优化分子性质，之后通过解码器直接生成分子构型。这个工作启发了大量关于物质微观结构的生成模型研究。

3.5　小结

生成模型远不止上面介绍的这些。例如，笔者还曾与合作者提出过“玻恩机”，一种量子波函数概率诠释所启发的生成模型。在生成模型不长的历史中，并不缺乏“惊艳逆袭”的传奇故事。也许下一个“爆款”的生成模型就来自本文的读者。除了提出新模型，走向统一也是生成模型发展的一大趋势。其实，上面介绍的扩散模型、流模型和变分自编码器可以统一地被看作是一类模型，其核心是概率密度的输运(transportation)[35]。因此，当下最成功的生成模型背后的机制只有两种：输运和自回归。这两种生成机制背后的道理都是“积跬步，行千里”，通过一步步的迭代积累生成样本。

哪种生成模型最厉害？这有点像在问谁在兵器谱上排名第一。在数据量不多时，一条首要的经验法则(有例外！)是看问题的特点：要处理的是离散还是连续变量？数据是否具有明显的序列化特征？一般来说，自回归模型适合于那些具有序列化表示的离散数据，而基于输运的流模型和扩散模型更适合于结构化的连续数据。对于更加复杂的情况，则可能需要考虑几种模型的组合。但是，假如数据量充分大，则进入了“手中无剑，心中有剑”的新境界：具体的模型架构已经不重要了，重要的是用来训练模型的数据本身[36]。

4　应用于物质科学的生成模型

训练生成模型所需要的数据量通常远大于模型的参数。因此，生成模型需要做好“有损压缩”：发现数据中的统计规律，从而生成符合规律的样本。本文介绍的四类现代生成模型都可以直接采样。它们生成样本的方式有点类似于经验直觉式的快速反应，生成速度和样本具体是什么无关。不仅如此，扩散模型、流模型和自回归模型所表达的概率分布都很容易归一化。这意味着它们都有“自知之明”：不仅可以生成样本，还可以用模型计算的似然函数给样本打分。

生成模型给科学研究打开了新的可能。但为了切实解决领域痛点问题，还是需要紧紧扣住生成模型的物理内涵，发挥出它们各自的特点，而不是仅仅搬运现有的程序框架。下面我们举例说明。

4.1　蛋白和材料生成

蛋白质是生命的物质基础，材料是人类文明的物质基础。在我们生活的地球上，经过亿万年的演化和数千年的文明，人类掌握了数十万种蛋白和无机晶体材料的结构。但是，它们只代表了非常少的可能性，蛋白和材料空间中还存在大量的结构有待人类探索与合成，其中也许就包含着新的灵丹妙药和先进材料。利用生成模型对原子级别的微观世界建模，可以绕过传统反向设计的思路，直接生成新的、有用的蛋白和材料结构。

例如，图3(a)展示了蛋白生成模型Chroma[37]根据对称群、子结构和肽链的整体形状生成的蛋白结构。Chroma的工作原理和图3(b)所展示的从文本生成图像的DaLLE-3模型一样，都是使用扩散模型从条件概率$p(\boldsymbol{x}|y)$中采样。除此以外，值得注意的是，蛋白其实也有显著的离散和序列化特征：它们是氨基酸序列在三维空间所形成的特定结构。可以使用自回归语言模型学习氨基酸序列的语言规律[39]，并通过氨基酸“造句”的方式生成新的序列组合。这类蛋白语言模型结合AlphaFold之类的工具，可以进一步从生成的氨基酸序列中预测蛋白的三维结构。

生成模型在材料的探索上也大有用武之地。生成模型压缩内化已知的晶体材料数据，把数据中蕴含的固体化学知识存储在神经网络权重中。一个好的晶体生成模型首先会是一个晶体“鉴赏家”，当用它看到新的晶体结构时，神经网络激活反映在材料空间中的联想能力，而模型计算的似然函数代表了关于晶体稳定性的化学直觉。不仅如此，它还是一个晶体“艺术家”。和那些生成文本、图片和视频的人工智能“艺术家”一样，晶体生成模型可以直接生成自然界中可以存在，但还没被发现的晶体材料。

扩散模型天然符合晶体生长的物理直觉：它似乎模拟了原子在三维空间中彼此相互作用，逐渐扩散到最终的位置的过程。因此，扩散模型是目前晶体生成的主流探索方向[40—42]。除此以外，

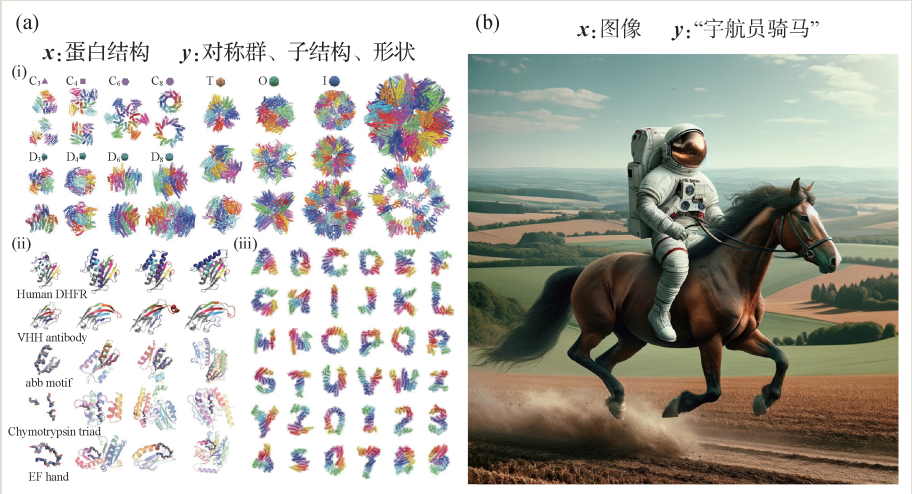

图3　(a) Chroma基于对称性(i)、子结构(ii)和形状(iii)生成的蛋白结构[37]；(b) DaLLE-3 基于文字提示生成图片[38]

也有一些别出心裁的工作使用自回归Transformer直接生成晶体的文本表示[43, 44]。这种做法抛弃了晶体的三维结构和对称性，但是居然也可以生成像模像样的晶体构型。也许，这又是深度学习给我们上的“苦涩的一课”[45]。无论如何，笔者认为如果考虑到晶体空间群的Wyckoff位置的自然排序，晶体材料可以有一种严谨的离散序列化表示[46]。在这种晶体语言中，“语序”由空间群Wyckoff占位的顺序决定，“语法”对应于固体化学规律，“同义词”代表可以互换的元素，“成语”对应常见的化学配位。使用晶体语言写出的一首首诗，正对应着自然界中那些漂亮的晶体。笔者相信，融合空间群对称性的晶体语言模型还有巨大的、有待挖掘的潜力。

另外，值得提一下Google DeepMind加速材料发现的工作[47]和本文主题“生成模型”的关系。抛开关于结果的一些争议不谈[48]，这个工作首先通过元素替换等手段产生大量备选结构，再利用机器学习势函数(详细讨论见[49])快速弛豫晶体结构并作出能量评估。其实，机器学习势函数也可以被看作是一种生成模型：它们是基于物理体系能量函数的玻尔兹曼机((4)式)。但不同于本文讨论的无监督学习方式，机器学习势函数通常需要第一性原理计算所得的能量和力作为训练标签。此外，正如前文对比扩散模型和分子动力学采样时所说的那样，基于真实能量函数的“生成模型”往往不容易在崎岖的势能空间走得很远。而本文介绍的几类现代生成模型避开了真实体系的能量函数和力的约束，可以直接生成多样且稳定的晶体结构。生成模型将是未来探索晶体材料空间的重要手段。

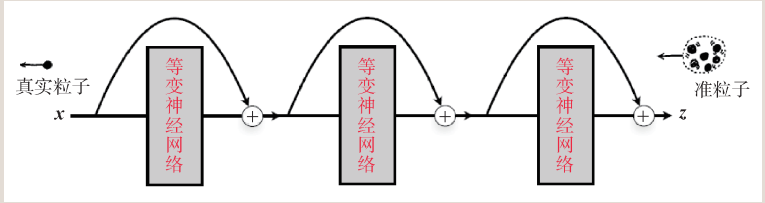

图4 使用深层的残差神经网络实现粒子和准粒子坐标之间的洄流变换。为了保证准粒子的统计性质，神经网络变换需要满足置换等变性质，即粒子置换操作和神经网络变换对易

4.2 变分自由能计算

公式(3)所示的变分自由能原理其实早就为人所知。在物理文献中，它常常被称为吉布斯—博戈留波夫—费曼(Gibbs—Bogoliubov—Feynman)不等式。它反映了大自然在有限温度下的一种倾向：平衡能量和熵从而使体系的自由能极小。然而，有点讽刺的是，试图将这一基本原理转化成实用的计算方法却是相当困难的。困难主要源于热力学熵的高昂计算代价——这背后的难点还是在于概率分布的归一化因子。在统计物理计算中，人们常常做出妥协，设计一些简单的概率来做变分计算，例如平均场方法或者贝特(Bethe)近似。这些妥协极大地限制了变分空间，并不能够精准地求解问题。而本文所介绍的自回归模型和流模型可以天然保证概率分布归一化，并且还能直接采样。使用这类现代的生成模型作为变分概率分布，终于可以发挥出吉布斯—博戈留波夫—费曼不等式中所蕴含的威力[24, 50, 51]。

除了经典的统计物理问题之外，基于生成模型的变分自由能计算也为解决量子多体问题提供了宝贵的机遇。对于哈密顿量为H的量子多体问题，最小化密度矩阵的相对熵可以给出和(3)式类似的目标函数：

$$F=\frac{1}{\beta}\mathrm{tr}\left(\rho\ln\rho\right)+\mathrm{tr}\left(\rho H\right)\,. \tag{10}$$

这里的难点是参数化多体密度矩阵，并且还要保证(10)式第一项的热力学熵易于计算。通过组合使用几种现代的生成模型可以解决这个难点。

下面我们以电子气低温状态方程研究为例展开具体讨论[52]。电子气是凝聚态物理研究的一个基础模型，它一方面是测试各种计算方法的试金石，另一方面，它也是费米液体和密度泛函理论中的基础模型。电子气密度矩阵的一般形式是$\rho=\sum_{k}p(\boldsymbol{k})\left|\psi_k\right\rangle\left\langle\psi_k\right|$，其中

$\boldsymbol{k}$是标记多电子能级的量子数，$p(\boldsymbol{k})$是能级占据的经典概率分布。可以使用自回归模型表达这个离散变量的联合概率分布，它的物理含义就是费米液体的朗道能量泛函。另一方面，$|\psi_k\rangle$是正交归一的多电子波函数，类比于前文(7)式，可以用“开根号”的流模型参数化这簇多体波函数：

$$\psi_k(\boldsymbol{x}) = \phi_k(\boldsymbol{z}) \left| \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} \right|^{\frac{1}{2}}, \qquad (11)$$

这里，$\boldsymbol{x}$和$\boldsymbol{z}$分别对应粒子坐标和准粒子坐标。(11)式中包含的神经网络变换的雅可比行列式保证了波函数的正交归一化。这类变换的一个相关的例子是费曼和他的学生研究液氦时引入的洄流(backflow)变换[53]：$z_i = x_i + \sum_{j \neq i} \eta(|x_i - x_j|)(x_j - x_i)$。洄流的物理含义是在相互作用体系中，粒子拖曳周围的粒子形成了准粒子。用准粒子坐标表达的简单波函数，如果写回到粒子坐标就是一个包含了关联效应的多体波函数。从深度学习的角度看，洄流变换对应于一个残差神经网络[54]。因此，其实可以把若干这样的变换串联起来形成图4所示的深层残差神经网络。可见，流模型抓住了费米液体理论中关键的绝热连续性(adiabatic continuity)的概念，而凝聚态物理学家讨论的准粒子就生活在流模型的隐变量空间。

在量子统计中，密度矩阵描述的是混合态。相对应地，我们使用自回归模型和流模型分别参数化了变分密度矩阵中的经典概率和量子波函数。这种参数化方式不仅符合物理约束，还蕴含了丰富的物理。一旦确立了参数化的变分空间，接下来就可以通过最小化变分自由能(10)式联合优化两个生成模型，从而近似解决电子气和更加一般的量子多体问题。基于生成模型的变分自由能方法将求解物理问题转化为对自然界的损失函数：自由能的随机梯度优化。这条研究路线值得关注，其实还有一个技术层面的理由，随机梯度优化简单朴实，却能充分发挥出深度学习引擎：微分编程框架[55]和专用加速硬件的强大计算能力，从而搭上人工智能飞速发展的便车，为解决困难的物理问题提供新机会。

表3　生成模型和统计物理课题之间的字典

生成模型	统计物理
对数似然函数$\ln p(\boldsymbol{x})$	能量函数
得分函数$\nabla \ln p(\boldsymbol{x})$	力
隐变量$\boldsymbol{z}$	集体变量
配分函数Z	自由能计算
自回归或输运采样	分子动力学模拟

5　结语

长期以来计算机最显著的特点是算得快、算得准。然而生成式人工智能赋予了计算机从经验中学习“直觉”的能力，甚至可以用来“创造”新的经验。定量描述这种“直觉”和“创造”的数学工具和统计物理一样，都是自然界中的概率分布。表3为有物理背景的读者提供了一个小小的字典。希望它和本文有助于揭开生成式人工智能的神秘面纱。

生成模型和物理研究血脉相连，它们都希望掌握数据背后的规律，并利用规律探索未知。当我们惊叹于生成式人工智能神奇的能力的同时，不妨也将目光投向我们赖以生存的宇宙，它也许就是一个更宏大的生成模型：从一个在餐巾纸上就写得下的作用量出发，产生了丰富多彩的世间万物。物理学研究的目的之一，就是基于实验观察所得的数据找到那个作用量。

致　谢　本文的写作受益于与张林峰、王涵、尤亦庄、吕健、李烁辉、谢浩、刘金国、吴典、董馨阳、欧仕刚等的合作与讨论。

参考文献

[1] 费曼的黑板图片. http://archives-dc.library.caltech.edu/islandora/object/ct1%3A483
[2] OpenAI官网关于生成模型的介绍. https://openai.com/research/generative-models
[3] Jensen不等式说明，对于凸函数f有：$f(x)$的平均值$\geq f(x$的平均值)。$-\ln x$是一个凸函数例子，据此可以证明相对熵的非负性。
[4] Tomczak J M. Deep Generative Modeling. Springer International Publishing，2022
[5] Murphy K P. Probabilistic Machine Learning: Advanced Topics. MIT Press，2023
[6] Bishop C M，Bishop H. Deep Learning：Foundations and Concepts. Springer International Publishing，2024

[7] Ackley D H, Hinton G E, Sejnowski T J. Cognitive Science, 1985, 9(1): 147
[8] Nguyen H C, Zecchina R, Berg J. Advances in Physics, 2017, 66(3): 197
[9] Vincent P. Neural Computation, 2011, 23(7): 1661
[10] Song Y, Ermon S. Generative Modeling by Estimating Gradients of the Data Distribution. In: Wallach H *et al* (Ed.). Advances in Neural Information Processing Systems 32, 2019
[11] Kirkpatrick S, Gelatt C D, Vecchi M P. Science, 1983, 220(4598): 671
[12] Sohl-Dickstein J, Weiss E A, Maheswaranathan N *et al*. International Conference on Machine Learning, 2015, 37: 2256
[13] Vaswani A, Shazeer N, Parmar N *et al*. Advances in Neural Information Processing Systems, 2017, 30: 6000
[14] 60行代码实现GPT2模型. https://jaykmody.com/blog/gpt-from-scratch/
[15] OpenAI开发的GPT2模型. https://openai.com/index/better-language-models/
[16] 300行FORTRAN代码实现GPT2模型. https://ondrejcertik.com/blog/2023/03/fastgpt-faster-than-pytorch-in-300-lines-of-fortran/
[17] 侯捷. STL源码剖析. 武汉: 华中科技大学出版社, 2002
[18] Wei J, Tay Y, Bommasani R *et al*. 2022, arXiv: 2206.07682
[19] Kaplan J *et al*. 2020, arXiv: 2001.08361
[20] Schaeffer R *et al*. 2023, arXiv: 2304.15004
[21] 关于大语言模型中涌现现象的讨论. https://www.jasonwei.net/blog/common-arguments-regarding-emergent-abilities
[22] Papamakarios G *et al*. Journal of Machine Learning Research, 2021, 22: 1
[23] Kingma D P, Dhariwal P. Advances in Neural Information Processing Systems, 2018, 32: 10236
[24] Li S H, Wang L. Phys. Rev. Lett., 2018, 121: 260601
[25] Li S H, Dong C X, Zhang L F *et al*. Phys. Rev. X, 2020, 10: 021020
[26] Chen R T Q, Rubanova Y, Bettencourt J *et al*. Advances in Neural Information Processing Systems, 2018, 32: 6572
[27] Zhang L F, Weinan E, Wang L. 2018, arXiv: 1809.10188
[28] Maoutsa D, Reich S, Opper M. Entropy, 2020, 22(8): 802
[29] Song Y, Sohl-Dickstein J, Kingma D P *et al*. 2020, arXiv: 2011.13456
[30] Lipman Y, Chen R T Q, Ben-Hamu H *et al*. 2022, arXiv: 2210.02747
[31] Liu X C, Gong C Y, Liu Q. 2022, arXiv: 2209.03003
[32] Albergo M S, Vanden-Eijnden E. 2022, arXiv: 2209.15571
[33] Zhao L, Wang L. Chin. Phys. Lett., 2023, 40: 120201
[34] Gómez-Bombarelli R, Wei J N, Duvenaud D *et al*. ACS Central Science, 2018, 4(2): 268
[35] Cédric V. Topics in Optimal Transportation, vol. 58. American Mathematical Soc., 2016
[36] 训练数据的重要性. 2023, https://nonint.com/2023/06/10/the-it-in-ai-models-is-the-dataset/
[37] Ingraham J B, Baranov M, Costello Z *et al*. Nature, 2023, 623: 1070
[38] DALL.E.3, https://openai.com/index/dall-e-3
[39] Madani A, Krause B, Greene E R *et al*. Nature Biotechnology, 2023, 41: 1099
[40] Xie T, Fu X, Ganea O *et al*. 2021, arXiv: 2110.06197
[41] Jiao R, Huang W B, Lin P J *et al*. Advances in Neural Information Processing Systems, 2024, 36: 11464
[42] Zeni C, Pinsler R, Zügner D *et al*. 2023, arXiv: 2312.03687
[43] Flam-Shepherd D, Aspuru-Guzik A. 2023, arXiv: 2305.05708
[44] Antunes L M, Butler K T, Grau-Crespo R. 2023, arXiv: 2307.04340
[45] The Bitter Lesson. http://www.incompleteideas.net/IncIdeas/BitterLesson.html
[46] Cao Z D, Luo X S, Lv J *et al*. 2024, arXiv: 2403.15734
[47] Merchant A, Batzner S, Schoenholz S S *et al*. Nature, 2023, 624: 80
[48] Cheetham A K, Seshadri R. Chemistry of Materials, 2024, 36(8): 3490
[49] 张林峰, 王涵. 模拟微观世界: 从薛定谔方程到大原子模型. 物理, 2024, 待发表
[50] Wu D, Wang L, Zhang P. Phys. Rev. Lett., 2019, 122(8): 080602
[51] Xie H, Li Z H, Wang H *et al*. Phys. Rev. Lett., 2023, 131(12): 126501
[52] Xie H, Zhang L F, Wang L. SciPost Physics, 2023, 14: 154
[53] Feynman R P, Cohen M. Physical Review, 1956, 102: 1189
[54] He K M, Zhang X Y, Ren S Q *et al*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. pp. 770—778
[55] 王磊, 刘金国. 物理, 2021, 50(2): 69

读者和编者

《物理》有奖征集封面素材

为充分体现物理科学的独特之美，本刊编辑部欢迎广大读者和作者踊跃投寄与物理学相关的封面素材。要求图片清晰，色泽饱满，富有较强的视觉冲击力和很好的物理科学内涵。

一经选用，均有稿酬并赠阅该年度《物理》杂志。

请将封面素材以附件形式发至：physics@iphy.ac.cn；联系电话：010-82649029。

《物理》编辑部