第 54 卷第 1 期
2026 年 1 月

硅 酸 盐 学 报
JOURNAL OF THE CHINESE CERAMIC SOCIETY

Vol. 54，No. 1
January，2026

http://jccsoc.cbpt.cnki.net

DOI：10.14062/j.issn.0454-5648.20250875

综 合 评 述

晶体结构生成式模型研究进展

魏 巍[1,2]，严 壮[1,3,4]，唐智勇[1,2]

(1. 中国科学院纳米系统与多级次制造重点实验室，国家纳米科学中心，北京 100190；2. 中国科学院大学，北京 100049；
3. 北京大学国际机器学习研究中心，北京 100871；4. 北京科学智能研究院，北京 100084)

摘 要：晶体材料的物理化学性质根本上取决于其三维空间中原子周期性排列所形成的晶体结构。晶体结构的高效预测与合理设计是新材料研发的关键环节，传统基于密度泛函理论(DFT)的高通量筛选和全局优化方法虽在简单体系中表现突出，但计算开销随原子数增加呈指数增长、对复杂多组分体系适应性差。近年来，人工智能的发展为晶体结构研究提供了新思路，其中，生成式模型(generative models)通过学习晶体数据库中结构与成分的分布规律，可在条件约束下高效探索庞大的化学空间，生成具备目标性能的候选结构。该类模型兼具低推理成本与强泛化能力，为突破传统方法瓶颈提供了关键路径。本文综述生成式模型在晶体结构研究中的最新进展，涵盖生成任务定义、结构表示方法、主要数据集、模型演进及典型应用，并展望其在材料智能设计和理论计算–实验验证闭环系统方面的未来发展方向。

关键词：晶体结构预测；生成式模型；科学智能；晶体材料设计
中图分类号：O742　　**文献标志码**：A　　**文章编号**：0454–5648(2026)01–0106–16
网络出版时间：2026–01–08

晶体作为原子或原子团簇在三维空间中呈周期性有序排列的固体材料，是支撑能源、电子、催化等领域技术进步的核心载体。其独特的物理化学性质并非完全由成分决定，而是源于原子层面的有序排列模式，即晶体结构。例如，同样由碳元素构成的金刚石与石墨，因碳原子分别采取 sp^3与 sp^2杂化的排列方式，前者表现出极高的硬度与绝缘性，后者则具备良好的导电性与润滑性[1]。晶体结构预测(CSP)旨在通过理论计算，从给定的化学组成出发，筛选出能量最低、稳定性最高或符合特定功能需求的晶体结构，为实验合成提供指导，但传统 CSP 方法如“穷举搜索法”“分子动力学模拟法”“进化算法”等，受限于技术原理，长期面临“被动搜索”与“高计算成本”的矛盾。

生成式模型(generative models)作为机器学习领域的重要分支，与传统的判别式模型聚焦分类或回归问题不同，其优势在于从数据中学习潜在规律并主动生成符合真实分布的新样本。在深度学习技术的推动下，生成式模型已形成以变分自编码器(VAE)、生成对抗网络(GAN)、图神经网络(GNN)、自注意力机制模型(Transformer)、基于马尔可夫链的扩散模型(Diffusion)及大语言模型(LLM)为代表的技术体系，在图像生成与自然语言处理等领域实现突破。这些技术特性使得生成式模型在晶体结构生成中展现出“主动生成”和“数据驱动”的优势，推动晶体结构研究从“被动搜索”向“主动设计”转型，为复杂体系晶体结构探索提供了全新思路。

1 晶体结构生成基础

1.1 核心挑战与生成式模型优势

晶体结构生成的目标是根据化学组成或性能需求，构建符合周期性规律、稳定且具备应用潜力的原子排列方案。传统探索方法存在 3 大挑战：一是组合空间爆炸，含 10 个原子的体系潜在结构数量可超 10^{11}种，穷举搜索难以覆盖亚稳态结构[2–3]；二是局部最优陷阱，分子动力学模拟等方法高度依赖初

收稿日期：2025–12–01。　修订日期：2025–12–25。
基金项目：北京市科学技术委员会中关村科技园区管理委员会资助(Z251100007525007)。
第一作者：魏 巍(1990—)，男，博士研究生。
通信作者：唐智勇(1971—)，男，博士，研究员；
严 壮(1998—)，男，博士，助理研究员。

Received date: 2025–12–01.　**Revised date:** 2025–12–25.
First author: WEI Wei (1990–), male, Doctoral candidate.
E-mail: weiw2025@nanoctr.cn
Correspondent author: TANG Zhiyong (1971–), male, Ph.D., Professor.
YAN Zhuang (1998–), male, Ph.D., Assistant Research Fellow.
E-mail: zytang@nanoctr.cn; yanz@pku.edu.cn

始结构先验知识，易陷入局部能量极小值[4–6]；三是计算成本高昂，进化算法等需反复通过 DFT 计算结构能量，单次预测往往耗时数周甚至数月[7–8]。

生成式模型通过数据驱动的主动生成模式，有效突破传统方法局限，展现出 3 大优势：一是效率显著提升，预训练后生成单个晶体结构仅需秒级时间，较传统 DFT 驱动方法提速数千倍[9]；二是条件生成能力突出，可通过融入成分、对称性、性能等约束，直接生成符合特定需求的结构[10–11]；三是亚稳态结构挖掘能力强，通过学习结构分布概率特征，能高效筛选出能量略高于最低值但具备特殊功能的亚稳态结构，拓展新材料边界[12]。

此外，生成式模型还可与实验合成深度融合形成“计算生成–筛选–实验”的闭环[13]，显著缩短新材料研发周期。这种模式已在金属有机框架材料(MOFs)设计中得到应用。例如，SMVAE 模型通过在隐空间中进行约束采样，能够以 61.5%的有效率生成符合几何与化学规则的 MOF 结构，显著提升了设计效率[14–15]。

1.2　主流生成式模型技术原理与晶体适配性

生成式模型通过学习晶体结构数据的潜在分布，实现从噪声或条件约束出发的结构生成。根据建模原理，可分为 6 类，这些方法在结构空间探索、生成质量与可控性方面各具优势，其数理基础与晶体适配性如下。

1.2.1　VAE　VAE 以变分推断与概率图模型为数理基础，其目标是通过近似后验分布学习数据的低维隐表示[16]。模型始于对联合概率分布[$P(X, Z)$]的建模，其中：$\boldsymbol{X}$ 表示晶体结构数据(如原子坐标矩阵)；Z 为隐变量(图 1a)。由于真实后验分布[$P(Z|X)$]难以直接计算，VAE 引入可学习的变分分布 $Q_\varphi(Z|X)$ 进行近似，并通过最大化证据下界(ELBO)间接最小化二者之间的 KL 散度，从而实现可行的变分推断优化。模型由编码器(encoder)与解码器(decoder)组成。编码器通过神经网络将输入 X 映射为变分分布的参数，通常为高斯分布的均值 $\mu(X)$与方差 $\sigma^2(X)$，即：

$$Q_\varphi(Z|X)=N\left(Z;\mu(X),\sigma^2(X)I\right) \tag{1}$$

解码器则学习生成分布：

$$P_\theta(X|Z)=N(X;f_\theta(Z),\sigma^2 I) \tag{2}$$

其中：$f_\theta(Z)$ 为解码网络输出函数：φ与 θ 分别表示编码器与解码器的参数。VAE 的训练目标是最大化 ELBO：

$$L(X;\theta,\varphi)=E_{\{Q_\varphi(Z|X)\}}[\ln P_\theta(X|Z)]-D_{\mathrm{KL}}(Q_\varphi(Z|X)\|P(Z)) \tag{3}$$

其中：第 1 项为重构损失(reconstruction loss)，用于衡量生成结构与原始样本的相似度；第 2 项为正则化项(regularization term)，用于约束隐空间分布逼近先验分布 $P(Z)$，该先验通常设为标准高斯分布。

在晶体结构生成任务中，可进一步引入条件信息 C 构建条件变分自编码器(CVAE)，可使隐变量与物理化学约束建立强关联，从而定向生成符合目标特性的晶体结构[17]，适用于需严格遵循晶体周期性、稳定性、成分结构等约束的材料设计场景，如 CDVAE 模型[18]。

1.2.2　GAN　GAN 以博弈论与生成模型理论为基础，通过生成器(generator)与判别器(discriminator)的 minimax 博弈实现数据分布拟合[19]。其数理基础是最小化生成分布 $P_G(X)$与真实数据分布 $P_{\mathrm{data}}(X)$之间的 JS 散度 $D_{\mathrm{JS}}[P_{\mathrm{data}}(X)\|P_G(X)]$ (图 1b)。生成器 $G(Z)$接收随机噪声 $Z\sim P(Z)$，通过非线性变换生成伪样本 $\hat{X}=G(Z)$；判别器 $D(X)$则输出样本来自真实数据集的概率 $D(X)\in[0,1]$。二者的博弈目标通过价值函数 $V(G, D)$定义：

$$\min_G \max_D V(G,D)=E_{X\sim P_{\mathrm{data}}}\left[\ln D(X)\right]+E_{Z\sim P(Z)}\left[\ln\left(1-D(G(Z))\right)\right] \tag{4}$$

当达到纳什均衡时，$P_G(X)=P_{\mathrm{data}}(X)$，生成器可输出符合真实分布的样本，判别器无法区分真伪 $[D(X)=0.5]$。条件生成对抗网络(CGAN)通过在生成器与判别器中同时引入条件变量 C(如晶格参数、空间群)，将目标函数扩展为 $V(G,D,C)$，使生成过程受约束引导，适配晶体周期性与组成要求，能生成细节丰富的新颖结构，尤其适用于探索非传统或亚稳态结构，如 CrystalGAN[8]。

1.2.3　GNN　GNN 的数理基础源于图论与消息传递机制，专为处理非欧几里得数据设计，例如晶体的原子邻接图(图 1c)。晶体结构可抽象为无向图 $G=(V, E)$，其中节点 V 代表原子，如特征含元素类型、电荷等；边 E 代表原子间相互作用，如特征含键长、键角等[20]。核心数学过程为邻域特征聚合，第 k 层的节点特征 H_v^k通过聚合其邻域节点 $N(v)$的第 k–1 层特征更新：

$$H_v^k=\sigma\left(W^k\cdot AGG\left(\{H_u^{k-1}\mid u\in N(v)\cup\{H_v^{k-1}\}\right)\right) \tag{5}$$

其中：σ为激活函数；W^k 为可学习参数；AGG 为聚合函数(如均值、求和或注意力加权)。

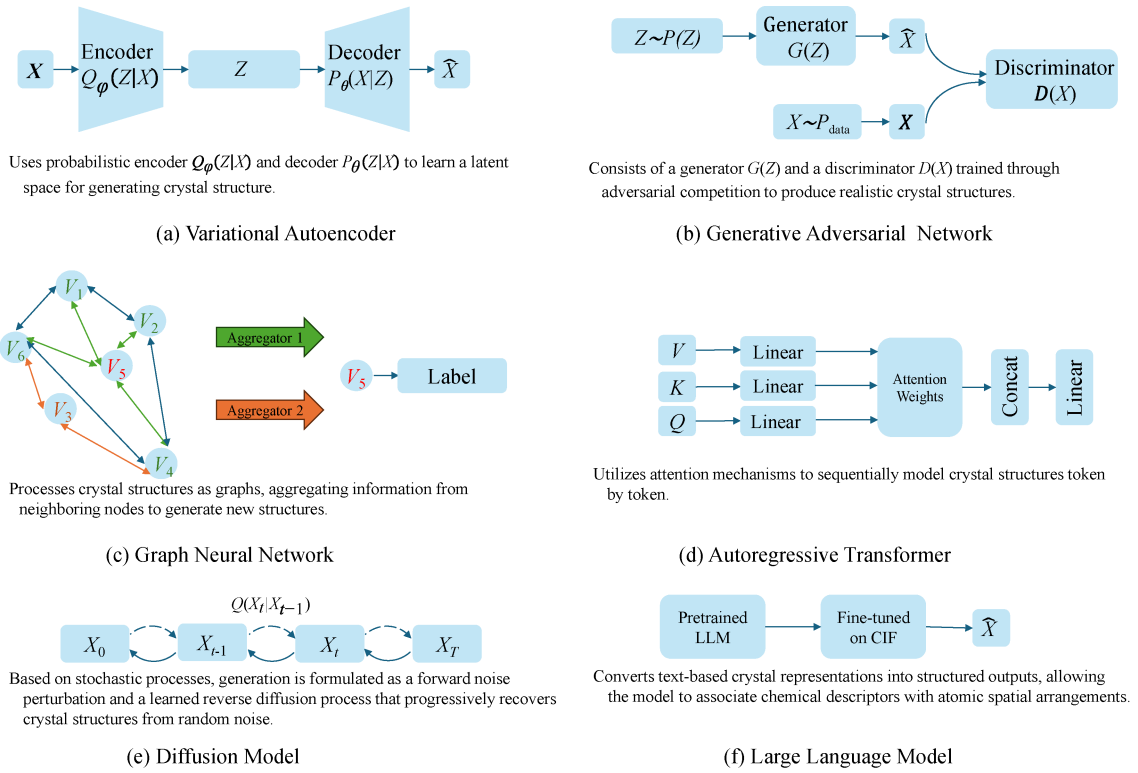


图 1 主要生成模型架构的示意图概述

Fig. 1 Schematic overviews of the main generative model architecture

在晶体生成任务中，GNN 通过递归邻域聚合有效编码原子的局部配位环境，因而被广泛用作结构的表征器；基于 GNN 构建的生成模型借助对抗训练机制鼓励跳出常见分布，适用于材料发现中的候选结构筛选阶段。例如，CGCNN 在晶体性质预测上的成功[21]。同时，图生成方法如 GraphVAE 在小分子/分子图生成中能够输出符合化学成键规则的拓扑结构，为将图生成技术推广至晶体网格提供了方法论基础[22]。

1.2.4 Transformer Transformer 的核心数理创新是自注意力机制，依托线性代数运算实现序列数据的全局依赖建模，突破了传统局部卷积或循环结构的局限(图 1d)。其数学逻辑围绕 Query 查询、Key 键、Value 值的交互展开。对于晶体结构的序列化表示，如按晶格顺序排列的原子特征序列 $X\in R^{n\times d}$，n 为原子数，d 为特征维度，首先通过线性变换生成 3 组向量：

$$Q=XW_Q,K=XW_K,V=XW_V \tag{6}$$

其中：$W_Q,W_K,W_V\in R^{d\times d_k}$ 为可学习矩阵。自注意力权重通过 Query 与 Key 的点积相似度计算，并经 Softmax 标准化：

$$\text{Attention}(Q,K,V)=\text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{7}$$

$\sqrt{d_k}$ 用于缓解点积过大导致的 Softmax 饱和问题[23]。多头注意力(MH)通过并行计算多个自注意力头并拼接结果，实现多视角特征融合：

$$\text{MH}(Q,K,V)=\text{Concat}(\text{Head}_1,\cdots,\text{Head}_h) \tag{8}$$

在晶体生成中，Transformer 可通过编码器学习已知结构的全局特征分布，解码器在接收成分/性能等条件约束后负责逐步生成原子/晶胞序列；其数理优势在于自注意力机制能够捕捉原子间的长程周期性关联与元素配比约束，从而天然适配需要建模长程晶格耦合与整体对称性依赖的生成任务[9, 24–25]。

1.2.5 Diffusion Diffusion 的数理框架建立在随机过程与概率密度反演理论之上，模拟“加噪–去噪”的渐进过程实现生成[26]。其包括前向加噪与反向去噪 2 个阶段，均可视为马尔可夫过程的离散近似(图 1e)。前向过程(有序加噪)从原始晶体结构 X_0 开始，在 T 个时间逐步添加高斯噪声，每个步骤的分布满足：

$$Q(X_t\mid X_{t-1})=N(X_t;\sqrt{1-\beta_t}X_{t-1},\beta_t I) \tag{9}$$

其中：$\beta_t\in(0,1)$ 为预设方差序列。利用马尔可夫性，

可直接由 X_0 计算任意 t 时刻的带噪样本：

$$q(X_t \mid X_0)=N(X_t;\sqrt{\alpha_t}X_0,(1-\alpha_t)I) \tag{10}$$

式中：$\alpha_t=\prod_{s=1}^{t}(1-\beta_s)$。当 $T\to\infty$ 时，X_{T} 逼近标准高斯噪声 $X_T\sim N(0,I)$。

而反向过程(学习去噪)则主要在训练神经网络 $\epsilon_\theta(X_t,t)$ 预测 t 时刻添加的噪声，通过最小化预测噪声与真实噪声的均方误差优化 $L(\theta)$：

$$L(\theta)=E_{X_0,Z,t}[\|\epsilon-\epsilon_\theta(\sqrt{\alpha_t}X_0+\sqrt{1-\alpha_t}Z,t)\|^2] \tag{11}$$

其中：$Z\sim N(0,I)$ 为随机噪声。生成时从 $X_T\sim N(0,I)$ 出发，通过迭代去噪重构 X_0：

$$X_{t-1}=\frac{1}{\sqrt{\alpha_t}}(X_t-\frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(X_t,t))+\sigma_t Z_t \tag{12}$$

在晶体生成中，可通过引入条件变量 C(如元素组成)构建条件去噪网络 $\epsilon_\theta(X,\ t,\ C)$，从而实现具备特定性能的目标结构定向生成。该方法的数理优势在于能兼顾生成样本的多样性与能量稳定性。

1.2.6　LLM　　LLM 在晶体生成中的数理框架是以 Transformer 编码器–解码器架构为基础，融合自回归生成理论与跨模态对齐机制，其关键能力体现在将晶体结构的文本化表示，如晶体学信息文件(CIF)格式、简化分子线性输入系统(SMILES)字符串转化为结构化输出，同时实现化学语义与空间结构的关联建模(图 1f)。

LLM 通过建模序列词元(token)的条件概率分布实现生成，对于晶体结构的文本序列 $U=(u_1,\cdots,u_n)$，如 CIF 文件的符号序列，生成过程遵循链式法则：

$$L(\theta;U)=\sum_{l}\ln P(u_i \mid u_{i-c},\cdots,u_{i-l};\theta) \tag{13}$$

其中：θ 为模型参数；下标 c 表示预测 u_i 时，模型只回看 c 个历史变量。每个 token 的条件概率通过 Transformer 解码器的 Softmax 层计算：

$$P(u_i \mid u_{i-c},\cdots,u_{i-1})=\mathrm{Softmax}(W_0\cdot\mathrm{Decoder}(u_{i-c},...,u_{i-1})) \tag{13}$$

其中：W_0 为输出层可学习权重，Decoder 通过多头注意力捕捉 CIF 序列中元素间的依赖关系，通过这种对晶体结构文本序列的逐 token 建模，LLM 能够自动学习到晶体学特征间的内在关联，例如特定空间群(如 *P*6/*mmm*)与典型晶格参数范围之间的对应关系，从而实现对晶体结构层级化描述的精准生成，例如先确定晶胞组成/空间群，再生成晶格参数，最后输出原子分数坐标[9, 27]。

不同晶体结构生成模型的数理基础与核心特性决定了其差异化适配场景：VAE 依托强条件约束能力与隐空间可解释性，适配稳定性优先、结构规则性强的定向生成任务，如 MOFs 材料设计、钙钛矿组分优化等对周期性与能量稳定性有明确要求的场景；GAN 通过生成器与判别器的博弈，擅长非传统/亚稳态结构探索与高多样性结构生成，适合纳米复合晶异质界面设计、高熵合金原子排布优化等材料发现的前沿探索阶段；GNN 凭借消息传递机制对原子局部关联的精准刻画，适配局部配位环境敏感、化学键约束严格的任务，如催化材料活性位点设计、分子筛拓扑结构生成，常作为结构表征器来保障化学成键合理性；Transformer 的自注意力机制能高效捕捉原子长程耦合与元素配比约束，适配长程周期性强、整体对称性依赖高的场景，如高温超导材料预测、半导体单晶缺陷设计等大尺寸晶胞或长程有序性突出的体系；Diffusion 通过“加噪–去噪”渐进式生成模式，平衡结构新颖性与能量稳定性，适配多样性与稳定性兼顾、多条件约束的复杂任务，如固态电池电解质设计等多组分体系的功能设计；LLM 支持晶体结构文本化表示与结构化输出的转化，适配需要跨模态驱动、具备快速响应需求的生成任务，如通过自然语言指令快速探索新型功能材料，方便实验人员将性能需求转化为结构设计方案。

1.3　条件约束机制与定向生成

条件生成式模型能够将化学组成、目标性能(如禁带宽度、催化活性)、晶体学约束(如空间群、晶格常数范围)等先验信息转化为模型可识别的条件信号，进而定向生成符合特定需求的晶体结构，成为连接“性能需求”与“结构生成”的关键技术载体(图 2)。这些条件信息涵盖晶体研发全流程的核心约束，其与晶体结构的关联建模是条件生成式模型的核心逻辑，具体理论基础是通过学习训练数据集中“晶体结构样本–条件信息”的联合概率分布($P(X|C)$)，实现对条件变量 C 的响应式结构生成。具体而言，模型首先通过编码器将高维的晶体结构数据 X 与条件信息 C 共同映射至低维隐空间，构建满足特定分布的隐变量 Z，此时隐空间中的每个样本 Z 均与对应的条件 C 形成强关联；在生成阶段，解码器则以隐变量 Z 和条件 C 为输入，通过多层非线性变换重构出符合条件约束的晶体结构数据 $\hat{X}$，并通过损失函数，最小化 X 与 $\hat{X}$ 的差异，优化模型对条件信息的拟合能力。

在实际的晶体结构生成实践中，这类模型的数理设计需特别适配晶体的周期性与稳定性约束，例如：在条件变量 C 的设计中，需将晶格常数、空间

群等晶体学特征转化为数值向量，确保模型能理解周期性排列规律；在隐空间映射过程中，需通过正则化约束原子坐标以保障其合理性，避免生成超出键长范围或违反空间对称性的结构；而在解码器输出阶段，部分模型还会融入能量预测模块，通过附加 DFT 能量或机器学习势能面的预测损失，确保生成结构满足能量稳定性条件。这种“条件约束–数理建模–晶体特性适配”的一体化设计，使条件生成式模型成为定向开发功能晶体中的主流技术路径之一。

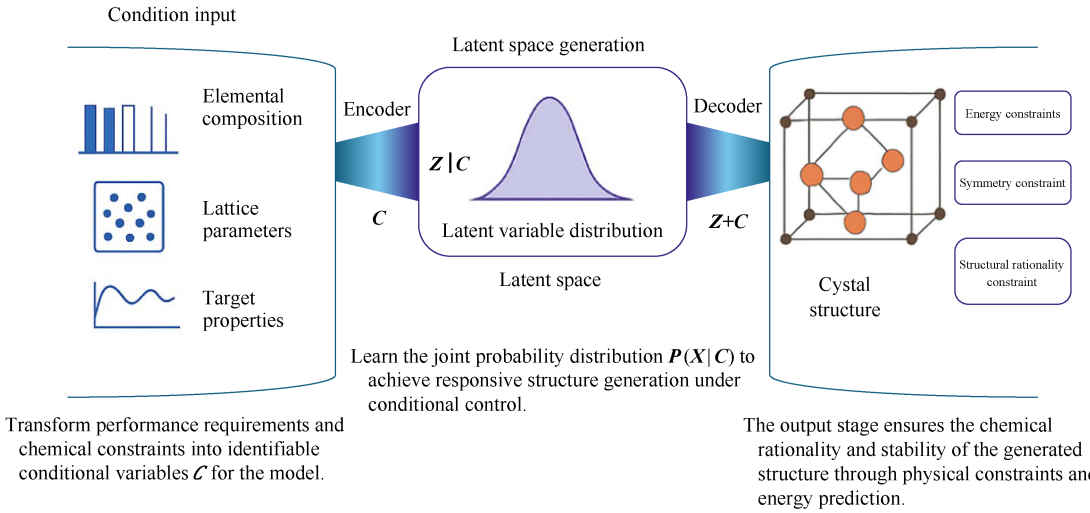


图 2 条件生成式模型流程

Fig. 2 Flow of conditional generative model

1.4 晶体结构描述符分类与适配场景

在生成式模型中，描述符是晶体结构与模型间的“信息桥梁”(表 1)，一方面将晶体的周期性、原子配位、元素组成等抽象结构特征转化为模型可解析的数值化、结构化数据，如晶格矩阵、图结构、晶体指纹等，避免关键特征丢失以确保模型学习到结构内在规律[21, 28]；另一方面通过定制化结构形式，如晶体指纹(GAN、GNN)，匹配不同模型为(如 VAE、GNN、Diffusion)的输入格式需求，降低特征提取难度，让模型聚焦“结构生成”任务[21, 29]。而生成式模型对描述符有 4 大主要要求：保真度上需完整保留晶体关键结构特征(如几何类描述符精准刻画原子排布、拓扑类描述符反映配位关系)，避免生成“非晶体”或不稳定结构[30–31]；适配性上需契合目标模型输入逻辑，如 3D-CNN 需要体素网格的规则三维数组、Transformer 需固定长度向量[29, 32]；有效性上需支持条件约束嵌入，如魏科夫位置关联空间群、超晶胞标记缺陷浓度，实现定向生成[10, 33]；高效性上需在保证精度的同时控制计算存储成本(如晶体指纹压缩维度，避免高分辨率体素网格的维度爆炸问题)，确保模型高效处理大规模数据[21]，以下将详细介绍各种晶体结构描述符。

表 1 晶体结构描述符分类比较表

Table 1 Comparison of crystal structure representations

Representation	Key features	Symmetry	Periodicity	Fine atomic	Long-range interaction	Addition boundary conditions
Reciprocal Space	Maps real-space periodicity into reciprocal space; represents diffraction features (*e.g.*, Miller indices, peak intensities) as fixed-length or frequency-domain vectors	☑	☑	☒	☑	☒
Wyckoff Positions	Encodes structure as “space group Wyckoff site atom mapping”; symmetry-informed and compact representation	☑	☑	☒	☑	☒
Voxel Grid	Discretizes the unit cell into 3D voxels; voxel values encode atomic occupancy/type; outputs 3D arrays or flattened tensors	☒	☒	☒	☒	☑
Point Cloud	Directly records atomic coordinates (x, y, z) and atomic numbers; supports dynamic number of atoms (N×4matrix)	☒	☒	☑	☒	☑
Graph	Represents the crystal as a multi-graph G=(V, E); nodes carry atomic features, edges encode distances/bond types; captures local chemistry/topology	☒	☒	☑	☑	☑

☑: Satisfied (This feature is clearly and excellently demonstrated, and the core functionality can be achieved without additional processing); ☒: Not Satisfied (This feature is weakly demonstrated, or requires additional operations to barely achieve the desired result, and cannot be naturally adapted).

1.4.1　倒易空间表示法(reciprocal space)　倒易空间表示法是将实空间晶体周期性结构映射至倒易空间，通过倒易点阵矩阵与衍射峰特征间接表征晶体结构，倒易点阵矩阵由实空间晶格矩阵转置逆运算得到，衍射峰特征则包含密勒指数、强度与半高宽等关键信息，可拼接为固定维度向量或转化为频域向量。从适配模型来看，其固定长度的特征向量能完美匹配 Transformer 与 Diffusion 对输入格式的要求，例如，在引入化学信息约束时，可将约束形式设定为“依据目标 X 射线衍射(XRD)谱生成对应原子结构”的条件生成任务，从而实现晶体衍射模式与结构构型的跨域对齐[34]。与其他方法相比，该方法独特优势在于能直接关联晶体光学、电学性能，如禁带宽度与倒易点阵参数的关联，这是晶格矩阵与原子坐标表示法等实空间方法难以实现的，但在刻画原子局部配位环境上，精度不及球谐函数表示法，更适合侧重宏观性能关联的晶体生成任务[35]。

1.4.2　魏科夫(Wyckoff)位置表示法　Wyckoff 位置表示法依托晶体空间群对称性，用 Wyckoff 符号与等效坐标集合描述原子对称排布，结构化形式为“空间群编号+Wyckoff 位置–原子映射表”，可转化为固定维度向量，如空间群独热(one-hot)向量与 Wyckoff 位置原子序数向量的拼接。在适配模型方面，其简洁的对称信息表示与明确的原子位置约束，使其适配 VAE 与晶体图神经网络(CGNN)，比如生成钙钛矿 ABX_3 结构时，可约束“空间群 221+Wyckoff a 位置为 A 原子”[10, 36]。相较于图结构表示法，该方法通过对称性减少冗余信息，计算效率更高，无须构建复杂的邻接矩阵；但灵活性不足，无法像超晶胞平铺表示法那样处理低对称性或含缺陷的晶体，仅适用于高对称性晶体生成，如立方相、四方相晶体。

1.4.3　体素网格表示法(voxel grid)　体素网格表示法将晶胞三维空间离散为均匀体素网格，以体素数值标识原子存在与否及类型，结构化输出为三维数组或扁平化一维向量，体素分辨率可根据需求调整，如 32×32×32 或 64×64×64[37]。从模型适配性看，其规则的三维网格结构天然适配 3D-CNN 与三维 Diffusion，在生成含异质界面的纳米复合晶时，可通过约束特定区域体素值控制原子分布。与点云表示法相比，体素网格表示法将连续空间离散化，虽会损失部分位置精度，但能更好适配依赖卷积运算的模型；而点云表示法保留原子精准坐标，却难以直接用于 3D-CNN。不过，体素网格表示法维度随分辨率提升急剧增加，计算成本高于晶体指纹表示法，更适合对空间局部结构调控有要求的复杂晶体生成[29]。

1.4.4　点云表示法(point cloud)　点云表示法用三维直角坐标记录单胞/超胞中每个原子的位矢并附加原子序数等属性，通常以 N×4 表示为(x, y, z, Z)矩阵存储；通过坐标中心化或尺度归一化可消除平移/尺度扰动，从而提高模型对平移与尺度扰动的鲁棒性。这种天然无序的集合结构与 PointNet[38]系列以及点云 Transformer[39–40]的处理范式高度契合，因此，被广泛用于点云分类、分割与近年的晶体生成任务。

与体素网格表示法相比，点云表示法无离散化损失，能精准保留原子间距与配位角度，更适合非周期性或准周期性晶体；但在处理高对称性晶体时，缺乏如 Wyckoff 位置表示法那样对对称信息的高效刻画，且无法像倒易空间表示法那样关联宏观衍射性能。此外，点云动态长度输入对模型批处理能力要求更高，这一点不如固定维度的晶体指纹表示法便捷。

1.4.5　图表示法(graph)　图表示法将晶体结构编码为一个多重图 $G=(V, E)$。其中，节点集合 V 中的每个节点 v_i 对应晶体中的一个原子；边集合 E 中的边 e_{ij} 则表示原子 i 与原子 j 之间的相互作用或键连接。节点通常附带元素级特征，如原子序数、电负性等；边特征可包含原子间距离(键长)、键类型或键角信息。此外，还可引入二阶甚至更高阶的图结构，用以描述更复杂的多体相互作用。

与点云和体素等几何表示不同，图结构本身并不保证几何信息的可逆性，通常无法从图直接恢复晶体的精确几何结构，因此，难以作为独立的生成式表示使用。在生成任务中，图表示更常作为辅助手段，与点云表示协同使用：点云负责更新原子空间位置，而图结构用于刻画局部拓扑关系，从而在迭代生成过程中同时保持几何合理性与化学可行性。

2　晶体结构数据集

晶体结构数据库是生成式模型发展与应用的“基石”(表 2)，其价值一是提供高质量训练数据基础，数据库通过标准化整理，如统一晶格参数单位、验证原子坐标合理性，为模型提供无噪声、高保真的学习样本，避免因数据质量问题导致模型学习偏差；二是支撑多维度约束生成，数据库中不仅包含晶体结构信息，还关联能量、热力学、光学等性能

数据，可帮助模型建立“结构–性能”映射关系，实现从“单纯生成结构”到“定向生成目标性能晶体”的跨越；三是拓展模型泛化边界，不同数据库覆盖的晶体类型、元素组合存在差异，多库融合使用可让模型接触更广泛的化学空间，避免对单一数据分布的过拟合，提升在未知晶体体系中的生成能力。

表 2 用于训练生成式模型的常见晶体数据集

Table 2 Common crystal datasets used for training generative models

Dataset	Size	Core feature	Key data contents	Typical applications
OMat24[50]	100 million	Largest open DFT dataset by Meta; integrates equilibrium and non-equilibrium structures	Atomic configurations, energies, stresses, electronic structures	High-throughput material screening
Materials Project[45]	Over 1 million	High degree of data standardization and complete attribute system	Crystal structures (lattice, coordinates, symmetry), formation energy, band gap, DOS, thermodynamic data	Semiconductor design, battery material discovery
Alexandria[48–49]	4.5 million	Links DFT structures with experimental synthesis and characterization data	Atomic configurations, synthesis parameters, XRD/SEM data, performance metrics	Experiment-driven structure generation and validation
ICSD[43–44]	327 000	Authoritative experimentally validated inorganic crystal structures	Atomic positions, space groups, lattice parameters	Structure prototype discovery, crystal structure prediction
OQMD[46–47]	1 317 811	Combines calculated formation energies and thermodynamic stabilities	DFT-calculated formation energy, stability, phase diagrams	Phase stability prediction, thermodynamic property modeling
CSD[41–42]	1 362 809	Experimental organic crystal repository	3D molecular structures, hydrogen bonding, symmetry data	Organic crystal design, pharmaceutical and coordination compound discovery
LLM4Mat-Bench[51]	1.97 million	Benchmark for evaluating LLMs in materials science reasoning	Multi-domain data (crystal info, text, property relations)	Evaluation of reasoning, prediction, and generation performance of LLMs in materials science
MGED	710 055	A national materials data infrastructure emphasizing standardized integration of heterogeneous experimental and computational data across the full R&D workflow	Crystal structures and compositions, synthesis and processing parameters, measured properties, characterization data, and first-principles and multiscale simulation results	Data-driven materials discovery, generative model training with experimental constraints, and experiment–computation closed-loop design
LAMBench	12 OOD force field datasets, 5 property calculation datasets; benchmarks 10 state-of-the-art LAMs	Comprehensive benchmark for LAMs evaluating generalizability, adaptability, and applicability; cross-domain coverage; automated high-throughput workflow; dimensionless metrics; open-source & interactive leaderboard	Standardized benchmark tasks built on multi-source DFT datasets; energy, force, and virial predictions; metrics for generalizability, adaptability, and applicability	Model comparison and validation, assessment of generalization and stability of atomistic foundation models, development of universal machine-learning interatomic potentials

2.1 CSD(剑桥结构数据库)

CSD 是由英国剑桥晶体数据中心(CCDC)维护的晶体结构数据库，起始于 1965 年，系统化汇编文献中已发表的小分子与金属有机晶体结构。CSD 收录了大量基于 X 射线、电子和中子衍射解析得到的精确三维坐标数据；截至 2025 年 1 月，数据库中记录的总结构数达到 1 362 809 个[41–42]。其主要聚焦有机体系，除原子三维坐标外，还记录了化学键信息、分子构型信息、附带化合物的合成方法、物理化学性质(如熔点、溶解度等)，是有机小分子和金属有机晶体领域数据最全面的数据库。在金属有机功能晶体与药物分子晶体生成等领域有重要应用价值。

2.2 ICSD(无机晶体结构数据库)

ICSD 最初由 Günter Bergerhoff 等在 1978 年建立，旨在系统性地汇总文献中已解析的无机晶体结构数据，目前由德国 FIZ Karlsruhe 与美国 NIST 协作维护，是国际上权威的无机晶体结构数据库之一。ICSD 收录了自 1913 年以来发表的完全解析的无机晶体结构条目，截至 2025 年 11 月，数据库条目数约为 327 000 条[43–44]，目前仍以每年一万条量级持续更新与扩充。数据库数据记录项经严格筛选与人工/自动化校验，典型条目包含晶体学基础信息，如晶格参数、原子分数坐标、空间群、完整的原子位置与占位因子、标准化后的晶胞与原子坐标、文献出处与实验条件注释等；部分条目还包含衍射相关信息或与理论计算的比对注记。由于 ICSD 对结构条目的高质量控制与统一的后处理标准，包括结构类型注释、标准化单元格及结构比对，其数据高度可比且被广泛用于相鉴定、晶体结构分析以及基于数据的材料筛选(如半导体、催化与电池材料的候选筛选与模型训练)中，被视为材料科学与晶体学研究的基础资源与行业准则。

2.3 Materials Project

Materials Project(MP)最初作为美国材料基因组计划(Materials Genome Initiative)框架下的重要实践项目，始于 2011 年 10 月，是由麻省理工学院和劳伦斯伯克利国家实验室联合开展的项目。该数据库包含百万级条目，包含结构数据与计算数据，整个数据体系完善[45]。数据库包含 80 多万条材料条目，以及超过 45 万条催化剂信息。其内容涵盖晶体基础

结构信息(晶格参数、原子分数坐标、空间群)、能量特性(形成能、结合能)、电子结构(禁带宽度、态密度)、热力学性质(比热容、热膨胀系数)，且所有数据均经过严格的质量验证与标准化处理，确保不同材料数据的可比性；因数据标准化程度高且属性丰富，其广泛应用于半导体材料与电池材料生成。

2.4　OQMD(开放量子材料数据库)

OQMD 是由 Northwestern University 的 Chris Wolverton 团队提出的。作为基于高通量第一性原理研发的计算平台，自 2010 年起逐步构建并对外开放，旨在通过系统性的 DFT 计算支持材料发现与相图构建[46–47]。截止至 2025 年 11 月，OQMD 包含 1 317 811 种材料的 DFT 计算数据，数据源头以 ICSD 的化合物修饰结构为主[46–47]。其数据字段包括晶体学信息、能量属性、电子结构信息，以及用于相图与稳定性分析的计算后处理结果。OQMD 提供完善的 API 与 qmpy 接口，数据以可复现的计算工作流与统一的计算参数记录，便于跨体系比较与大规模自动化分析。

2.5　Alexandria

Alexandria 是由 Ruhr University Bochum 的材料计算小组发布，以高通量第一性原理计算为主要内容的大规模材料数据库，现已发展为包含 450 万条 DFT 计算记录的开放数据集，不同子集的规模不同。例如，拥有 41.5 万条数据的 PBEsol/SCAN 子集和约 22.8 万条数据的立方钙钛矿子集。数据覆盖晶体结构、形成能、电子结构(带隙、态密度)及其他热力学性质。Alexandria 的重要特点是提供多泛函计算结果和丰富的假设/亚稳态构型，便于训练与评估机器学习势能、晶体生成模型及构建高通量筛选流程[48–49]。

2.6　OMat24

OMat24 是当前规模最大的开源 DFT 计算数据库，由 Meta 公司开源，包含超过 1.1 亿条计算记录[50]。数据覆盖范围极广，不仅收录晶体的平衡态原子构型如常规稳定相结构，还包含大量非平衡态构型如缺陷结构、过渡态结构，同时涵盖从二元化合物到多元复杂合金、氧化物的丰富的成分多样性数据，每条记录均附带 DFT 计算得到的能量、应力、电子结构等属性；凭借庞大数据规模与结构多样性，其主要用于高通量材料筛选与新型结构生成。

2.7　LLM4Mat-Bench

LLM4Mat-Bench 是由 Rubungo 等组成的研究团队发起，为评估与比较 LLM 在晶体材料性质预测上的能力而专门构建的大规模基准数据集与评测框架。该基准从 10 个公开的材料数据库收集并整合了约 197 万条晶体样本[51]，涵盖 45 种物理化学属性(如催化活性、磁学性质)，并同时提供化学组成、CIF 与文本描述，3 种模态输入。LLM4Mat-Bench 的目标是系统评测不同类型大语言模型在材料性质预测任务中的泛化能力与模态适应性，凭借其跨模态结构与标准化评测流程，LLM4Mat-Bench 已成为晶体属性预测与生成研究的重要基准之一。

2.8　MGED(材料基因工程数据库)

MGED 是由北京科技大学牵头建设的材料基因工程专用数据库[52]，是集数据采集、管理和应用软件于一体的综合平台，由中国的“十三五”重点研发计划“材料基因工程关键与支撑平台”支持。包含超过 71 万条材料相关数据，数据内容涵盖能源、催化、稀土功能材料等领域。主要为国家材料领域的未来研究和发展提供数据支撑，推动传统“试错法”研发模式转型，助力高端先进材料研发与产业升级。

2.9　LAMBench

LAMBench 是由北京科学智能研究院等机构提出的面向大原子模型的综合评测基准，旨在系统性评估不同原子级机器学习模型对第一性原理势能面的近似能力，自 2024 年起逐步构建并对外开放[53]。该基准并非传统意义上的材料数据库，而是基于多来源 DFT 计算数据与经典数据集构建的任务驱动型评测体系，用于统一比较模型在能量、原子力与应力张量预测中的精度、跨体系泛化能力及实际可用性。截至 2025 年，LAMBench 已集成来自无机材料、催化体系与分子体系等多个领域的标准化测试任务，并覆盖从小体系到大规模复杂结构的不同原子尺度。其评测指标在常规误差度量基础上，进一步引入模型守恒性、可微性、稳定性及推理效率等因素，以反映模型在分子动力学与材料模拟中的真实表现。通过模块化工具链与自动化工作流设计，LAMBench 为当前大原子模型在达成“近似通用势能面”这一核心目标上的能力差异提供了可复现、可比较的评估框架，已成为原子级基础模型方法学研究的重要参考基准。

3　目前的模型进展

自 2018 年以来，晶体结构生成式模型经历了从单一表征到多模态融合、从规则晶体生成到复杂异质体系设计的系统性演进(表 3)。当前模型发展

表 3 生成式模型及其描述符与数据库概况

Table 3 Overview of generative models, representations, and datasets

Acronym	Representations	Datasets	Architectures	Ref.	Year
CrystalGAN	Point Cloud	Pearson's Crystal Data[54]	GAN	[8]	2018
ED3DCS	Voxel	MP[45]	GAN	[55]	2019
iMatGen	Voxel Grid	MP[45]	VAE	[56]	2019
ICSG3D	Voxel	MP[45]	GAN	[29]	2020
CCCG	Point Cloud	MP[45]	GAN	[12]	2020
CubicGAN	Point Cloud	OQMD[46–47]	GAN	[57]	2021
FTCP	Reciprocal Space, Point Cloud	MP[45]	VAE	[86]	2022
CCDCGAN	Voxel Grid	MP[45]	VAE, GAN	[87]	2022
CDVAE	Point Cloud, Graph	Perov-5[58–59], Carbon-24[60], MP-20[45]	VAE, Diffusion	[18]	2022
PGCGM	Wyckoff	MP[45]	GAN	[61]	2023
XYZTransformer	Point Cloud	Perov-5[58–59], MP-20[45]	Transformer	[63]	2023
LCOMs	Point Cloud, Graph	OQMD[46–47], MatBench[90]	VAE, Diffusion	[64]	2023
SyMat	Point Cloud, Graph	Perov-5[58–59], Carbon-24[60], MP-20[45]	Diffusion	[65]	2023
GemsDiff	Point Cloud, Graph	MP[45]	Diffusion	[66]	2023
LCMGM	Reciprocal Space	OQMD[46–47], MP[45]	VAE, GAN	[91]	2024
DiffCSP++	Wyckoff	Perov-5[58–59], Carbon-24[60], MP-20[45], MPTS-52[45]	Diffusion	[33]	2024
CrystalFormer	Wyckoff	MP-20[45]	Transformer	[24]	2024
WyCryst	Wyckoff	MP[45]	VAE	[36]	2024
CGWGAN	Wyckoff	MP[45]	GAN	[68]	2024
CGMD	Point Cloud	AFLOW[70]	VAE, Diffusion	[69]	2024
DP-CDVAE	Point Cloud	Carbon-24[60]	Diffusion	[71]	2024
CrysTens	Point Cloud	Pearson's Crystal Database[54]	GAN, Diffusion	[72]	2024
CrystalTextLLM	Point Cloud	MP-20[45]	LLM	[92]	2024
DiffCSP	Point Cloud	Perov-5[58–59], Carbon-24[60], MP-20[45], MPTS-52[45]	Diffusion	[67]	2024
Con-CDVAE	Point Cloud, Graph	MP[45], OQMD[46, 47]	VAE, Diffusion	[17]	2024
UniMat	Point Cloud	Perov-5[58–59], Carbon-24[60], MP-20[45]	Diffusion	[93]	2024
FlowMM	Point Cloud	MP-20[45], MPTS-52[45]	Flow Matching	[76]	2024
GenMS	Point Cloud	MP[45], ICSD[43–44], OQMD[46–47], NOMAD[75]	LLM, Diffusion	[73]	2024
MatExpert	Point Cloud	NOMAD[75]	LLM	[74]	2024
FlowLLM	Point Cloud	MP-20[45]	LLM, Flow Matching	[94]	2024
TamGen	SMILES	PubChem[95]	GNN, LLM	[96]	2024
Cond-CDVAE	Point Cloud, Graph	MP[45]	VAE, Diffusion	[30]	2024
Matra-Genoa	Wyckoff	MP[45], Alexandria[48–49], WBM[78], ORB[97]	Transformer	[98]	2025
SymmCD	Wyckoff	MP-20[45], MPTS-52[45]	Diffusion	[99]	2025
WyckoffDiff	Wyckoff, GNN	WBM[78]	Diffusion	[77]	2025
Uni-3DAR	Voxel	Carbon-24[60], MP-20[45], MPTS-52[45]	Transformer	[32]	2025
EH-Diff	Point Cloud	Perov-5[58–59], Carbon-24[60], MP-20[45], MPTS-52[45]	Diffusion	[100]	2025
CrysBFN	Point Cloud	Perov-5[58–59], Carbon-24[60], MP-20[45], MPTS-52[45]	Bayesian	[81]	2025
TransVAE-CSP	Point Cloud, Graph	Perov-5[58–59], Carbon-24[60], MP-20[45]	VAE, Diffusion	[25]	2025
CrystalFlow	Point Cloud, Graph	MP-20[45], MPTS-52[45]	Flow Matching	[82]	2025
MatterGen	Point Cloud, Graph	MP[45], Alexandria[48–49]	Diffusion, Graph	[101]	2025
TGDMat	Point Cloud	Perov-5[58–59], Carbon-24[60], MP-20[45]	Diffusion	[102]	2025
NatureLM-Mat3D	Point Cloud	MP[45], NOMAD[75], OQMD[46–47]	LLM	[85]	2025
UniGenX	Point Cloud	MP-20[45], NOMAD[75], OQMD[46, 47], Carbon-24[60], MPTS-52[45], GEOM[103]	Transformer, Diffusion	[83]	2025
DAO	Point Cloud, Graph	CrysDB[79]	Diffusion	[79]	2025
Chemeleon	Point Cloud	MP-20[45]	Diffusion	[84]	2025

大体可分为 3 类：结构生成式模型、逆向设计与优化模型、结构解析与跨模态模型。以下将以技术演进脉络为主线，系统梳理代表性模型与其应用。

3.1　结构生成式模型

早期的晶体生成模型以单一表征与基础生成框架为核心。2018 年，CrystalGAN[8]开创性地使用非显式点云(point cloud)表征，以 Pearson's Crystal Database[54]为训练数据，生成对 GAN 框架，实现了三维晶体结构的自动生成，为后续点云表征的广泛应用奠定基础。在 2019 年，ED3DCS[55]与 iMatGen[56]引入体素/体素网络(voxel/voxel grid)表征，基于 MP 数据集构建 GAN 或 VAE 混合生成模型，能够在三维离散空间内生成完整的晶胞结构。2020 年，模型延续这一技术路径，ICSG3D[29]以来自 MP 数据库内的 CIFs 为数据来源，通过体素表征与 GAN 实现结构生成。CCCG[12]则基于 MP 数据集将点云表征与 GAN 结合，实现晶体结构生成。2021 年，CubicGAN[57]以 OQMD 为数据集，采用点云表征与 GAN 架构，聚焦立方晶系结构生成，同时推动了表征与数据的针对性适配。2022 年成为技术多元化转型的关键节点：CDVAE[18]创新性地融合点云与图(graph)双重表征，采用 VAE 与 Diffusion 混合架构，以 Perov-5[58–59]、Carbon-24[60]等数据集为支撑，通过自适应周期性 SE(3)等变图神经网络编码周期边界相互作用，实现稳定结构生成。

近三年模型发展呈现表征精细化与架构革新的双重特征，如 2023 年 PGCGM[61]将 Wyckoff 位置表征与 GAN 结合，基于 MP 数据集引入空间群对称变换的数据增强策略，使高对称性晶体结构生成的有效性较基线模型得到了提升；MOFDiff[62]将粗粒度(CG)Diffusion 与对比学习机制结合，先通过 GemNet-OC 编码器与对比损失实现 MOF 构建块(金属节点、有机连接体)的低维嵌入，再在"构建块身份–坐标–晶格"粗粒度空间内进行降噪扩散采样，同时采用等变图神经网络尊重旋转平移与置换对称性；创新设计退火组装算法(通过高斯密度重叠优化连接点匹配)与 UFF 力场弛豫，实现从粗粒度结构到全原子 MOF 的精准恢复；XYZTransformer[63]开创性地将 Transformer 与点云表征融合，以 Perov-5、MP-20[45]为训练数据，开启了自然语言与晶体结构的跨模态建模；同时，Diffusion 开始成为主流架构，LCOMs[64](点云–图表征、VAE/扩散)、SyMat(点云–图表征、扩散)[65]、GemsDiff(点云–图表征、扩散)[66]均基于 MP 或 OQMD 数据集，通过组合表征与 Diffusion 的结合，实现生成结构稳定性与多样性的提升。

2024 年模型数量与技术复杂度显著提升，形成多表征、多数据、多架构的融合格局：DiffCSP (Diffusion Model for Crystal Structure Prediction)[67]由清华大学与中国人民大学提出，在"晶格–分数坐标"空间内引入周期 E(3)等变去噪模型，兼顾晶格一致性与周期性约束，在 MP-20 数据集上 20 样本匹配率达 77.93%，生成速率较 DFT 方法提升 4500 倍。采用 Wyckoff 的表征应用也在持续深化，DiffCSP++[33]基于 Perov-5、MPTS-52[45]等多源数据集，以 Wyckoff 表征与 Diffusion 实现高精度结构生成；CrystalFormer[24]与 WyCryst[36]分别采用 Transformer 与 VAE 架构，基于 MP-20 或 MP 数据集，进一步深化了结构生成的对称性建模。CGWGAN[68]则将 Wyckoff 表征与 GAN 结合，进一步优化结构生成效率。同时，点云表征在晶体生成领域仍占重要地位。例如，CGMD[69]基于 AFLOW 数据集[70]，融合了 VAE 与扩散机制；DP-CDVAE[71]基于 Carbon-24 数据集，采用了扩散生成模型。这些工作持续推动着扩散模型在晶体生成中的工程化实现。CrysTens[72]则集成 GAN 与扩散双架构，提升生成结构的多样性。跨模态与多架构融合成为新方向，GenMS[73]结合 LLM 与 Diffusion，通过自然语言指令驱动结构生成；MatExpert[74]基于 NOMAD[75]数据集构建了一种由大语言模型驱动的多阶段晶体生成框架；FlowMM[76]引入 Flow Matching 技术，将晶体生成建模为连续归一化流过程，从而在保持生成质量与结构稳定性的同时，显著减少了推理阶段所需的积分步数。该方法以 Flow Matching 替代传统扩散模型，实现了更高效的稳定晶体结构生成。

2025 年模型进入深度集成与前沿探索阶段，呈现多维度技术突破：表征融合更为深入，WyckoffDiff[77]将 Wyckoff 位置表征与 GNN 结合，基于 WBM[78]数据集构建 Diffusion，通过对称编码实现快速稳定的结构生成；TransVAE-CSP[25]与 DAO[79]则采用点云–图组合表征，分别基于 VAE/扩散与扩散架构生成高精度结构；在架构创新方面，MOFFlow 模型[80]基于黎曼流匹配框架，利用 MOF 固有的模块化特性，将金属节点与有机连接体视为刚体，在 SE(3)空间(特殊欧几里得空间)中高效捕捉刚体的旋转 q、平移 τ 的动态，将预测维度从原子级的 3N 降至块级的 6M(6M≪3N)，大幅降低计算复杂度；同时设计分层架构，即原子级更新层通过图

神经网络编码构建块原子特征，块级更新层结合自定义的 MOFAttention 模块(融入晶格参数信息)迭代优化旋转、平移与晶格(ℓ)，实现对 MOF 结构的精准组装。CrysBFN[81]基于贝叶斯(Bayesian)架构，CrystalFlow[82]采用 Flow Matching，UniGenX[83]集成 Transformer 与 Diffusion，在 MP-20、NOMAD 等多源数据集上实现序列与结构的统一生成，性能较传统方法显著提升；跨模态能力进一步强化，Chemeleon[84]通过 Crystal CLIP 实现文本与结构嵌入的对齐，基于 MP-20 数据集构建文本引导的 Diffusion，突破了传统生成模型的输入限制；NatureLM-Mat3D[85]则基于 MP、NOMAD 等大规模数据集构建通用 LLM，推动晶体生成向通用人工智能方向发展。

3.2 逆向设计与优化模型

逆向设计与优化模型旨在以物性约束为引导，生成具有目标性能，如低形成能、高带隙或特定电导率等特性的新型晶体。在 2022 年，FTCP[86]引入 reciprocal space 与点云的组合表征，基于 MP 数据集构建 VAE 模型，突破了单一表征的局限，该模型以可逆晶体表征(invertible crystallographic representation)为核心逆向设计模型；CCDCGAN[87]则是一种基于体素表征的条件生成对抗网络，用于多组分体系的逆向晶体设计；由 DeepMind 开发的 GNoME[88]，采用强化学习与 GNN 相结合的主动学习策略，形成“模型筛选–DFT 验证–数据回流”的闭环优化体系。该模型在能量预测上达到 11 meV/atom 误差，稳定结构命中率超 80%，成功发现 220 万种稳定晶体，其中 736 种经实验验证，为电池电极与固态电解质的高通量筛选提供了数据支持。

3.3 结构解析与跨模态模型

结构解析与跨模态模型的目标是打通实验测量数据、晶体结构重建、性质预测的闭环，代表工作包括 XtalNet 模型[89]和 Chemeleon[84]。

XtalNet 由厦门大学、北京大学与深势科技联合开发，首次实现了粉末 X 射线衍射(PXRD)图谱到晶体结构的端到端预测。模型通过 BERT-based PXRD 特征提取器与等变图神经网络(EGNN)实现 PXRD–晶体嵌入空间的对齐，并以 Diffusion 条件生成晶格与分数坐标，在 hMOF-100 与 hMOF-400 数据集上 top-10 匹配率分别达 90.2%与 79%。其在真实 PXRD 实验数据上成功重建金属有机框架的金属节点结构，为 MOF 自动解析提供了新方案。

Chemeleon 模型是跨模态生成的代表，将 Crystal CLIP 与多通道扩散结合，能在文本输入(如“Zn–Ti–O 体系中可见光催化活性层状结构”)下生成对应晶体结构。模型在 MP-40 数据集上结构有效性 98%~99%，唯一性超 90%，成功预测多种固态电解质候选结构。其提出的“文本–结构语义对齐”范式，为人机协作式的功能材料生成奠定了基础。

4 模型的应用

近年来，生成式模型在材料科学中的应用逐渐从理论方法验证走向面向真实体系的结构设计与性能优化，其价值正在多个研究方向中得到系统体现。得益于模型在结构空间搜索、能量面近似及条件生成等方面的表达能力提升，这类方法已能够在化学约束、几何规则以及目标性能导向的多重条件下，高效提出具备可合成性与潜在应用价值的候选晶体结构。特别是在材料定向设计、药物固体形态预测、量子材料结构探索以及能源材料设计等典型领域，生成式模型通过与高通量计算、物性预测器及实验验证流程耦合，逐步形成“生成–筛选–验证”的闭环式研究范式。该范式不仅显著降低了结构搜索的计算与实验成本，也推动了材料研发从经验驱动向模型驱动的转变。本节将围绕材料筛选、药物研发、量子材料探索、能源材料设计这 4 大核心方向，总结不同生成式模型在实际材料体系中的应用路径。

4.1 材料筛选：定向设计稳定材料

MatterGen[101]由微软研究院与中国科学院深圳先进技术研究院提出，是以周期对称扩散机制为核心的生成类模型，它能够生成周期表范围内稳定且多样的无机晶体结构。模型通过包裹正态分布建模周期坐标，对晶格引入平均密度约束，并通过等变分数网络避免对称性重复学习。基于 60.8 万结构的 Alex-MP-20 数据集预训练，再以轻量适配器模块进行目标性质引导，实现低形成能或高模量晶体的定向生成。MatterGen 提出了 SUN(稳定–独特–新颖)评价体系。在该体系下，其生成结构与 DFT 弛豫结果的平均均方根偏差(RMSD)低至 0.076 Å。例如，其设计的 $TaCr_2O_6$，模量与目标值的偏差小于 20%。在“生成→筛选→验证”的闭环中，MatterGen 凭借其高效的候选结构生成能力和强大的属性条件化功能，显著提升了筛选阶段的效率与命中率，从而加速了整体材料设计进程。

4.2 药物研发：靶向药物设计开发

微软研究院与 GHDDI 合作开发的 TamGen 模型[96]，专为靶向药物设计打造，在结核病治疗药物

开发中取得显著成果。在结核病治疗药物领域，针对结核分枝杆菌(Mtb)关键靶点酪蛋白水解蛋白酶P(ClpP)，TamGen 通过“设计–优化–测试”3 步流程实现靶向分子生成：首先以 ClpP 的结合口袋结构(PDB ID 5DZK)为约束，生成 2612 个独特分子；经分子对接与 AI 活性预测筛选后，选取 4 个种子化合物与 3 个弱活性实验分子，进一步优化生成 8635 个分子；最终通过商业库相似性搜索与从头合成验证，发现 14 个具有显著抑制活性的 ClpP 抑制剂。其中，Analog-005 的半数抑制浓度(IC_{50})低至 1.88 μmol/L，自主合成的衍生物 Syn-A005-01、Syn-A005-02 的IC_{50}分别为 5.34 μmol/L、3.40 μmol/L，均展现出优于传统筛选的靶向活性。这些分子采用苯磺酰胺、二苯基脲等全新骨架，与现有肽模拟抑制剂(如硼替佐米)结构差异显著，有效解决了传统药物生物利用度低、稳定性不足的痛点，为耐药结核病治疗提供了新的药物候选方向。

在模型性能验证方面，TamGen 在 CrossDocked2020 基准数据集上，相较于 3D-AR、Pocket2Mol、TargetDiff 等主流生成式方法，在结合亲和力(对接分数)、类药性(QED)、合成可及性(SAS)等 6 项核心指标中 5 项位列前二，整体性能排名第一。其生成的分子融合环数量(平均 1.78 个)与 FDA 批准药物高度一致，合成可及性更优，且生成 100 个目标分子仅需 9 s，较最快的对比模型 ResGen 快 85 倍，兼顾了生成质量与效率。

4.3　量子材料探索：定向合成奇异磁性化合物

美国麻省理工学院团队提出了 SCIGEN 计算框架，通过将阿基米德晶格(Archimedean lattices)和 Lieb 晶格等几何对称性以结构约束形式嵌入生成模型，使生成的候选结构天然满足目标晶格的几何规则，从而显著提升了量子材料设计中的可控性与稳定性[104]。SCIGEN 能够面向量子自旋液体、高温超导、拓扑半金属、磁性材料与催化材料等场景生成多类结构，其中嵌入的晶格规则可确保模型在生成阶段即遵守周期性和平带特征等结构物理规律，有助于发现具有潜在量子特性的候选材料。

在生成效率方面，SCIGEN 在嵌入结构约束后能够一次性生成约 10^7个候选材料，并通过多阶段筛选将候选缩减至约 10^6个可行结构；进一步对约 2.6×10^4个候选进行高通量第一性原理计算，其中约 41%的结构表现出磁序相关特征，显著扩大可探索的量子材料空间[104]。SCIGEN 的约束模块无须重新训练即可切换晶格类型，从而支持按需生成多种结构族群，并已构建包含原始生成、预筛选与 DFT 验证结构在内的大规模材料数据库，可为量子计算、半导体与功能材料等产业提供基础研发资源。

在实验验证方面，研究团队从预测结果中选取了 2 种具有代表性的候选化合物，并成功合成了$TiPd_{0.22}Bi_{0.88}$与$Ti_{0.5}Pd_{1.5}Sb$ 2 种新材料。实验表征表明，这两种化合物分别呈现顺磁与抗磁行为，其磁性表现与 AI 模型的预测趋势保持一致[104]。这些结果证明，将几何约束融入生成模型能够有效提升预测材料的可实验性，从而加速量子材料的发现与验证过程。

4.4　能源材料设计：固态电池与催化材料精准设计

中国人民大学联合华为公司开发的 DAO“生成–预测”孪生框架[79]，在能源相关晶体材料开发中成果显著。在超导体领域，针对 CsV_3Sb_5家族材料，DAO-P 预测模型通过指数能量损失函数优化生成结构，对临界温度(T_c)的预测值(2.26 K)与实验值(2.30 K)偏差仅 1.7%，基于此设计的$CsV_3Sb_{4.8}Te_{0.2}$掺杂结构，经实验合成后 T_c 提升至 3.1 K，突破该体系本征上限。

该框架通过晶体结构生成与能量预测的协同作用，显著提高了生成效率与候选筛选精度。通过在标准晶体结构基准集(如 MP-20、MPTS-52)上的测试，DAO-G 能够生成高质量晶体结构，DAO-P 提供的能量引导有效避免陷入局部能量极小值，从而加快从大规模候选集合中发现实验可行结构的过程。该方法为复杂材料体系的计算发现和设计提供了可推广的基础模型架构，具备较高的跨体系泛化能力和加速材料研发的潜力。

5　结论与展望

本文系统梳理了生成式模型在晶体结构预测与设计中的研究进展。不难发现，随着生成式模型技术体系的日趋成熟，晶体结构预测已从传统的搜索式算法，逐步演进为基于概率分布学习的智能生成范式。这一范式转型不仅显著提升了结构探索的效率，拓展了可模拟体系的复杂度边界，更为功能材料的逆向设计提供了数据驱动的全新思路。

尽管该领域已取得一系列突破性进展，但晶体结构生成式模型仍面临 3 大核心挑战：其一，数据层面存在稀疏性问题，复杂多组分体系、亚稳态结构的高质量数据储备不足，且物理约束(如晶格周期性、群对称性、热力学稳定性)在模型中的融入不够充分，导致模型在小众体系中的泛化能力受限；其

二，模型生成结果的物理合理性与可解释性有待提升，现有方法多依赖数据驱动的黑箱建模，难以量化结构–性能关联的内在机制；其三，“模型生成–实验验证”的闭环体系尚未完全构建，多数生成结构仍停留在计算模拟层面，缺乏高效的实验转化路径与验证流程。基于上述挑战，未来的发展方向可聚焦以下 4 个维度：

1) 构建多模态融合的高质量综合数据库。未来需系统整合多尺度多模态实验表征数据、第一性原理计算数据及模型自生成的候选结构数据，形成覆盖“晶体结构–物理化学属性–实验合成可行性”的全链条数据库。同时，需在数据构建过程中嵌入群对称性、周期性边界条件、键合规则等物理先验知识，通过数据层面的约束强化，从源头提升模型对复杂体系的适配性与泛化能力。

2) 建立“模型–实验”闭环反馈机制。通过引入自动化实验平台与模型的动态反馈链路，构建“物理约束嵌入–模型智能推断–实验验证评估–数据回流优化”的持续迭代体系。模型可从实验结果(包括成功案例与失败数据)中提取有效信息，修正生成过程中的偏差，进而提升输出结构的物理合理性、可合成性与可解释性，实现理论建模与实验验证的深度耦合。

3) 加速无人化实验平台的集成与落地。推动无人实验室的技术攻关与工程化应用，实现样品制备、计算预测、表征测试、数据回流与模型优化的一体化整合。通过生成式模型与机器人实验平台、自动化表征设备的深度融合，构建从结构生成到实验验证的全流程自动化链路，有效填补当前“生成停留在计算层面”的核心短板，大幅缩短材料研发周期。

4) 促进生成式模型的产业化转化与工程化应用。依托标准化、可复用的晶体生成与筛选工程化流程，推动生成式模型在新能源电极材料、催化材料、光电功能材料等关键领域的规模化部署。通过构建“需求拆解–结构生成–性能筛选–工程适配”的产业化应用体系，为行业提供兼具创新性与落地性的材料解决方案，实现从理论创新到产业价值的转化。

总体而言，未来晶体材料研究将沿着 “智能化、自动化、产业化”3 条主线深度交织、协同发展。随着无人实验室的普及、多模态大模型的成熟以及物理约束增强型生成框架的完善，晶体结构生成有望构建“需求导向–智能生成–自动化验证–工程应用”的全流程闭环体系。这一发展趋势不仅标志着材料科学正从经验驱动迈向数据驱动与智能设计，更将推动该领域实现从“偶然发现”到“精准设计”的变革式创新。

参考文献:

[1] POPOV I V, GÖRNE A L, TCHOUGRÉEFF A L, et al. Relative stability of diamond and graphite as seen through bonds and hybridizations[J]. Phys Chem Chem Phys, 2019, 21(21): 10961–10969.

[2] GLASS C, OGANOV A, HANSEN N. Crystal structure prediction using evolutionary algorithms: principles and applications[J]. Comput Phys Commun, 2006, 175: 713.

[3] OGANOV A R. Crystal structure prediction: Reflections on present status and challenges[J]. Faraday Discuss, 2018, 211: 643–660.

[4] GARCIA-VILOCA M, POULSEN T D, TRUHLAR D G, et al. Sensitivity of molecular dynamics simulations to the choice of the X-ray structure used to model an enzymatic reaction[J]. Protein Sci, 2004, 13(9): 2341–2354.

[5] AMSLER M, GOEDECKER S. Crystal structure prediction using the minima hopping method[J]. J Chem Phys, 2010, 133(22): 224104.

[6] ORMEÑO F, GENERAL I J. Convergence and equilibrium in molecular dynamics simulations[J]. Commun Chem, 2024, 7(1): 26.

[7] LI C N, LIANG H P, ZHAO B Q, et al. Machine learning assisted crystal structure prediction made simple[J]. J Mater Inf, 2024, 4(3): 15.

[8] NOUIRA A, SOKOLOVSKA N, CRIVELLO J-C. CrystalGan: Learning to discover crystallographic structures with generative adversarial networks[J]. Arxiv, 2018, Doi:10.48550/arXiv.1810.11203.

[9] ANTUNES L M, BUTLER K T, GRAU-CRESPO R. Crystal structure generation with autoregressive large language modeling[J]. Nat Commun, 2024, 15(1): 10570.

[10] KAZEEV N, NONG W, ROMANOV I, et al. Wyckoff transformer: Generation of symmetric crystals[J]. Arxiv, 2025, Doi:10.48550/arXiv.2503.02407.

[11] NGUYEN T M, TAWFIK S A, TRAN T, et al. The search for superionic solid-state electrolytes using a physics-informed generative model[J]. Mater Horiz, 2025, 12(17): 6945–6955.

[12] KIM S, NOH J, GU G H, et al. Generative adversarial networks for crystal structure prediction[J]. ACS Cent Sci, 2020, 6(8): 1412–1420.

[13] CHEN Z A, MENG Z J, HE T, et al. Crystal structure prediction meets artificial intelligence[J]. J Phys Chem Lett, 2025, 16(10): 2581–2591.

[14] PARK J, KIM H, KANG Y, et al. From data to discovery: Recent trends of machine learning in metal–organic frameworks[J]. JACS Au, 2024, 4(10): 3727–3743.

[15] YAO Z P, SÁNCHEZ-LENGELING B, BOBBITT N S, et al. Inverse design of nanoporous crystalline reticular materials with deep generative models[J]. Nat Mach Intell, 2021, 3(1): 76–86.

[16] KINGMA D P, WELLING M. Auto-encoding variational bayes[J]. Arxiv, 2013, Doi:10.48550/arXiv.1312.6114.

[17] YE C Y, WENG H M, WU Q S. Con-CDVAE: A method for the conditional generation of crystal structures[J]. Comput Mater Today, 2024, 1: 100003.

[18] XIE T, FU X, GANEA O-E, et al. Crystal diffusion variational autoencoder for periodic material generation[J]. Arxiv, 2021, Doi:10.48550/arXiv.2110.06197.

[19] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Commun ACM, 2020, 63(11): 139–144.

[20] GASTEIGER J, GROß J, GüNNEMANN S. Directional message passing for molecular graphs[J]. Arxiv, 2020, Doi:10.48550/arXiv.2003.03123.

[21] XIE T, GROSSMAN J C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties[J]. Phys Rev Lett, 2018, 120(14): 145301.
[22] HAMILTON W, YING Z, LESKOVEC J. Inductive representation learning on large graphs[C]//NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, California, USA, 2017: 30: 1025–1035.
[23] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Arxiv, 2023, Doi: 10.48550/arXiv.1706.03762.
[24] CAO Z, LUO X, LV J, et al. Space group informed transformer for crystalline materials generation[J]. Arxiv, 2024, Doi:10.1016/ arXiv:2403.15734.
[25] CHEN Z Y, YUAN Y, ZHENG S M, et al. Transformer-enhanced variational autoencoder for crystal structure prediction[J]. arXiv E Prints, 2025: arXiv: 2502.09423.
[26] YANG L, ZHANG Z L, SONG Y, et al. Diffusion models: A comprehensive survey of methods and applications[J]. ACM Comput Surv, 2024, 56(4): 1–39.
[27] SONG Z L, LU S H, JU M G, et al. Accurate prediction of synthesizability and precursors of 3D crystal structures *via* large language models[J]. Nat Commun, 2025, 16(1): 6530.
[28] FABER F, LINDMAA A, VON LILIENFELD O A, et al. Crystal structure representations for machine learning models of formation energies[J]. Int J Quantum Chem, 2015, 115(16): 1094–1101.
[29] COURT C J, YILDIRIM B, JAIN A, et al. 3-D inorganic crystal structure generation and property prediction *via* representation learning[J]. J Chem Inf Model, 2020, 60(10): 4518–4535.
[30] LUO X S, WANG Z Y, GAO P Y, et al. Deep learning generative model for crystal structure prediction[J]. NPJ Comput Mater, 2024, 10: 254.
[31] OKABE R, CHENG M Y, CHOTRATTANAPITUK A, et al. Structural constraint integration in a generative model for the discovery of quantum materials[J]. Nat Mater, 2025: https://doi.org/10.1038/ s41563-025-02355-y.
[32] LU S Q, LIN H W, YAO L, et al. Unified cross-scale 3D generation and understanding *via* autoregressive modeling[J]. Arxiv, 2025, Doi:10.48550/arXiv.2503.16278.
[33] JIAO R, HUANG W, LIU Y, et al. Space group constrained crystal generation[J]. Arxiv, 2024, Doi:10.48550/arXiv.2402.03992.
[34] CHOUDHARY K. DiffractGPT: Atomic structure determination from X-ray diffraction patterns using a generative pretrained transformer[J]. J Phys Chem Lett, 2025, 16(8): 2110–2119.
[35] ZHANG Y, LI Y F, ZHANG X W, et al. Predicting crystal spin Hall conductivity using a multi-modal transformer with real-space and reciprocal-space fingerprints[J]. Phys Rev Materials, 2025, 9(9): 093804.
[36] ZHU R M, NONG W, YAMAZAKI S, et al. WyCryst: Wyckoff inorganic crystal generator framework[J]. Matter, 2024, 7(10): 3469–3488.
[37] RAO C P, LIU Y. Three-dimensional convolutional neural network (3D-CNN) for heterogeneous material homogenization[J]. Comput Mater Sci, 2020, 184: 109850.
[38] CHARLES R Q, HAO S, MO K C, et al. PointNet: Deep learning on point sets for 3D classification and segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. IEEE, 2017: 77–85.
[39] FUCHS F, WORRALL D, FISCHER V, et al. Se (3)-transformers: 3D roto-translation equivariant attention networks[J]. Arxiv, 2023, Doi: 10.48550/arXiv.2204.02394.
[40] GUO M-H, CAI J-X, LIU Z-N, et al. Pct: Point cloud transformer [J]. Comput Vis Media, 2021, 7: 187–199.
[41] ALLEN F H. The Cambridge Structural Database: A quarter of a million crystal structures and rising[J]. Acta Crystallogr B, 2002, 58(Pt 3 Pt 1): 380–388.
[42] BÜRGI H B. The Cambridge structural database and structural dynamics[J]. Struct Dyn, 2024, 11(2): 021302.
[43]HELLENBRANDT M. The inorganic crystal structure database (ICSD): Present and future[J]. Crystallogr Rev, 2004, 10(1): 17–22.
[44] ZAGORAC D, MÜLLER H, RUEHL S, et al. Recent developments in the Inorganic Crystal Structure Database: Theoretical crystal structure data and related features[J]. J Appl Crystallogr, 2019, 52(Pt 5): 918–925.
[45] JAIN A, ONG S P, HAUTIER G, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation[J]. APL Mater, 2013, 1: 011002.
[46] KIRKLIN S, SAAL J E, MEREDIG B, et al. The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies[J]. NPJ Comput Mater, 2015, 1: 15010.
[47] SAAL J E, KIRKLIN S, AYKOL M, et al. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD)[J]. JOM, 2013, 65(11): 1501–1509.
[48] WANG H C, SCHMIDT J, MARQUES M A L, et al. Symmetry-based computational search for novel binary and ternary 2D materials[J]. 2D Mater, 2023, 10(3): 035007.
[49] SCHMIDT J, HOFFMANN N, WANG H C, et al. Machine-learning-assisted determination of the global zero-temperature phase diagram of materials[J]. Adv Mater, 2023, 35(22): 2210788.
[50] BARROSO-LUQUE L, SHUAIBI M, FU X, et al. Open materials 2024 (omat24) inorganic materials dataset and models[J]. Arxiv, 2024, Doi:10.48550/arXiv.2410.12771.
[51]NIYONGABO RUBUNGO A, LI K M, HATTRICK-SIMPERS J, et al. LLM4Mat-bench: Benchmarking large language models for materials property prediction[J]. Mach Learn: Sci Technol, 2025, 6(2): 020501.
[52] XIE J X, SU Y J, ZHANG D W, et al. A vision of materials genome engineering in China[J]. Engineering, 2022, 10: 10–12.
[53] PENG A, CAI C, GUO M, et al. LAMBench: A benchmark for large atomic models[J]. arXiv preprint arXiv:250419578, 2025,
[54] VILLARS P, CENZUAL K. Pearson's Crystal Data: Crystal Structure Database for Inorganic Compounds[M]. Ohio: ASM International, 2023.
[55] HOFFMANN J, MAESTRATI L, SAWADA Y, et al. Data-driven approach to encoding and decoding 3-d crystal structures[J]. Arxiv, 2019, Doi:10.48550/arXiv.1909.00949.
[56] NOH J, KIM J, STEIN H S, et al. Inverse design of solid-state materials *via* a continuous representation[J]. Matter, 2019, 1(5): 1370–1384.
[57] ZHAO Y, AL-FAHDI M, HU M, et al. High-throughput discovery of novel cubic crystal materials using deep generative neural networks[J]. Adv Sci, 2021, 8(20): 2100566.
[58] CASTELLI I E, OLSEN T, DATTA S, et al. Computational screening of perovskite metal oxides for optimal solar light capture[J]. Energy Environ Sci, 2012, 5(2): 5814–5819.
[59] CASTELLI I E, LANDIS D D, THYGESEN K S, et al. New cubic perovskites for one- and two-photon water splitting using the computational materials repository[J]. Energy Environ Sci, 2012, 5(10): 9034–9043.
[60] PICKARD C J, NEEDS R J. Ab initio random structure searching[J]. J Phys Condens Matter, 2011, 23(5): 053201.
[61] ZHAO Y, SIRIWARDANE E M D, WU Z Y, et al. Physics guided deep learning for generative design of crystal materials with symmetry constraints[J]. NPJ Comput Mater, 2023, 9: 38.
[62] FU X, XIE T, ROSEN A S, et al. Mofdiff: Coarse-grained diffusion for metal-organic framework design [J]. Arxiv, 2023, Doi:10.48550/ arXiv.2310.10732.
[63] FLAM-SHEPHERD D, ASPURU-GUZIK A. Language models can

generate molecules, materials, and protein binding sites directly in three dimensions as XYZ, CIF, and PDB files[J]. Arxiv, 2023, Doi:10.48550/arXiv.2305.05708.

[64] QI H, GENG X, RANDO S, et al. Latent conservative objective models for data-driven crystal structure prediction[J]. Arxiv, 2023, Doi:10.48550/arXiv.2310.10056.

[65] LUO Y, LIU C, JI S. Towards symmetry-aware generation of periodic materials[J]. Arxiv, 2023, Doi: 10.48550/arXiv.2307.02707.

[66] KLIPFEL A, FREGIER Y, SAYEDE A, et al. Vector field oriented diffusion model for crystal material generation[J]. Proc AAAI Conf Artif Intell, 2024, 38(20): 22193–22201.

[67] JIAO R, HUANG W, LIN P, et al. Crystal structure prediction by joint equivariant diffusion[C]//NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023: 17464–17497.

[68] SU T H, CAO B, HU S B, et al. CGWGAN: Crystal generative framework based on Wyckoff generative adversarial network[J]. J Mater Inf, 2024, 4(4): 20.

[69] NOVITSKIY L, LAZAREV V, TIUTIULNIKOV M, et al. Unleashing the power of novel conditional generative approaches for new materials discovery[J]. Arxiv, 2024, Doi:10.48550/arXiv.2411.03156.

[70] CURTAROLO S, SETYAWAN W, WANG S D, et al. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput *ab initio* calculations[J]. Comput Mater Sci, 2012, 58: 227–235.

[71] PAKORNCHOTE T, CHOOMPHON-ANOMAKHUN N, ARRERUT S, et al. Diffusion probabilistic models enhance variational autoencoder for crystal structure generative modeling[J]. Sci Rep, 2024, 14(1): 1275.

[72] ALVERSON M, BAIRD S G, MURDOCK R, et al. Generative adversarial networks and diffusion models in material discovery[J]. Digit Discov, 2024, 3(1): 62–80.

[73] YANG S, BATZNER S, GAO R, et al. Generative hierarchical materials search[J]. Arxiv, 2024, Doi:doi.org/10.48550/arXiv.2409.06762

[74] DING Q, MIRET S, LIU B. Matexpert: Decomposing materials discovery by mimicking human experts[J]. Arxiv, 2024, Doi:10.48550/arXiv.2410.21317.

[75] SCHEIDGEN M, HIMANEN L, LADINES A N, et al. NOMAD: A distributed web-based platform for managingmaterials science research data[J]. J Open Source Softw, 2023, 8(90): 5388.

[76] MILLER B K, CHEN R T, SRIRAM A, et al. Flowmm: Generating materials with riemannian flow matching[J]. Arxiv, 2024, Doi:10.48550/arXiv.2406.04713.

[77] EKSTRöM KELVINIUS F, ANDERSSON O, PARACKAL A S, et al. WyckoffDiff–A Generative Diffusion Model for Crystal Symmetry[C]// ICML, 2025: 15130–15147.

[78] WANG H C, BOTTI S, MARQUES M A L. Predicting stable crystalline compounds using chemical similarity[J]. NPJ Comput Mater, 2021, 7: 12.

[79] WU L, HUANG W, JIAO R, et al. Siamese foundation models for crystal structure prediction[J]. Arxiv, 2025, Doi:10.48550/arXiv.2503.10471.

[80] KIM N, KIM S, KIM M, et al. Mofflow: Flow matching for structure prediction of metal-organic frameworks[J]. Arxiv, 2024, Doi:10.48550/arXiv.2410.17270.

[81] WU H, SONG Y, GONG J, et al. A periodic bayesian flow for material generation[J]. Arxiv, 2025, Doi:10.48550/arXiv.2502.02016.

[82] LUO X S, WANG Z Y, WANG Q C, et al. CrystalFlow: A flow-based generative model for crystalline materials[J]. Nat Commun, 2025, 16(1): 9267.

[83] ZHANG G, LI Y, LUO R, et al. Unigenx: Unified generation of sequence and structure with autoregressive diffusion[J]. Arxiv, 2025, Doi:10.48550/arXiv.2503.06687.

[84] PARK H, ONWULI A, WALSH A. Exploration of crystal chemical space using text-guided generative artificial intelligence[J]. Nat Commun, 2025, 16(1): 4379.

[85] XIA Y, JIN P, XIE S, et al. Nature Language Model: Deciphering the language of nature for scientific discovery[J]. Arxiv, 2025, Doi:10.48550/arXiv.2502.07527.

[86] REN Z K, TIAN S I P, NOH J, et al. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties[J]. Matter, 2022, 5(1): 314–335.

[87] LONG T, ZHANG Y X, FORTUNATO N M, et al. Inverse design of crystal structures for multicomponent systems[J]. Acta Mater, 2022, 231: 117898.

[88] MERCHANT A, BATZNER S, SCHOENHOLZ S S, et al. Scaling deep learning for materials discovery[J]. Nature, 2023, 624(7990): 80–85.

[89] LAI Q S, XU F J, YAO L, et al. End-to-end crystal structure prediction from powder X-ray diffraction[J]. Adv Sci, 2025, 12(8): 2410722.

[90] DUNN A, WANG Q, GANOSE A, et al. Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm[J]. NPJ Comput Mater, 2020, 6: 138.

[91] CHENEBUAH E T, NGANBE M, TCHAGANG A B. A deep generative modeling architecture for designing lattice-constrained perovskite materials[J]. NPJ Comput Mater, 2024, 10: 198.

[92] GRUVER N, SRIRAM A, MADOTTO A, et al. Fine-tuned language models generate stable inorganic materials as text[J]. Arxiv 2024, Doi:10.48550/arXiv.2402.04379.

[93] YANG S, CHO K, MERCHANT A, et al. Scalable diffusion for materials generation[J]. Arxiv, 2023, Doi:10.48550/arXiv.2311.09235.

[94] CHEN R, MILLER B, SRIRAM A, et al. FlowLLM: Flow matching for material generation with large language models as base distributions[C]//Advances in Neural Information Processing Systems 37. Vancouver, BC, Canada. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 46025–46046.

[95] KIM S, CHEN J, CHENG T J, et al. PubChem 2025 update[J]. Nucleic Acids Res, 2025, 53(D1): D1516–D1525.

[96] WU K H, XIA Y C, DENG P, et al. TamGen: Drug design with target-aware molecule generation through a chemical language model[J]. Nat Commun, 2024, 15: 9360.

[97] NEUMANN M, GIN J, RHODES B, et al. Orb: A fast, scalable neural network potential[J]. Arxiv, 2024, Doi:10.48550/arXiv.2410.22570.

[98] DE BREUCK P-P, PIRACHA H A, RIGNANESE G-M, et al. A generative material transformer using Wyckoff representation[J]. Arxiv, 2025, Doi:10.48550/arXiv.2501.16051.

[99] LEVY D, PANIGRAHI S S, KABA S-O, et al. SymmCD: Symmetry-Preserving crystal generation with diffusion models[J]. Arxiv, 2025, Doi:10.48550/arXiv.2502.03638.

[100] LIU Y, ZHOU C, ZHANG S, et al. Equivariant hypergraph diffusion for crystal structure prediction[J]. Arxiv, 2025, Doi:10.48550/arXiv.2501.18850.

[101] ZENI C, PINSLER R, ZÜGNER D, et al. A generative model for inorganic materials design[J]. Nature, 2025, 639(8055): 624–632.

[102] DAS K, KHASTAGIR S, GOYAL P, et al. Periodic materials generation using text-guided joint diffusion model[J]. Arxiv, 2025, Doi:10.48550/arXiv.2503.00522.

[103] AXELROD S, GÓMEZ-BOMBARELLI R. GEOM, energy-annotated molecular conformations for property prediction and molecular generation[J]. Sci Data, 2022, 9(1): 185.

Review on Crystal Structure Generative Models

WEI Wei[1,2], *YAN Zhuang*[1,3,4], *TANG Zhiyong*[1,2]
(1. Cas Key Laboratory of Nano System and Hierarchical Fabrication, National Center for Nanoscience and Technology, Beijing 100190; 2. University of Chinese Academy of Sciences, Beijing 100049; 3. Center for Machine Learning Research, Peking University, Beijing 100871; 4. AI for Science Institute, Beijing 100084)

Extended Abstract

Generative models for crystalline materials have emerged as a transformative paradigm in computational materials science, offering new possibilities for crystal structure prediction (CSP), inverse structural design, and functional materials discovery. Compared with traditional search-based CSP frameworks that depend on iterative global optimization and handcrafted physical constraints, generative approaches learn the statistical distribution of atomic arrangements directly from data, thereby enabling efficient sampling of chemically valid and symmetry-consistent structures. Recent advances in geometric deep learning, equivariant neural networks, variational autoencoders, and diffusion-based modeling have substantially enhanced both the fidelity and scalability of generative crystal models. However, despite the rapid development, a comprehensive and systematic review that integrates representations, generative mechanisms, and practical applications of these models is still sparse.

This review first analyzes the major classes of crystal representations. According to the level of geometric completeness and symmetry preservation, representations can be categorized into reciprocal-space encodings (*e.g.*, diffraction-based descriptors), symmetry-aware Wyckoff-position formalisms, voxelized three-dimensional grids, point-cloud coordinate systems, and graph-based multirelational structures. The specific characteristics of each representation are summarized, including their capabilities for handling periodicity, long-range interactions, spatial boundary conditions, and compatibility with equivariant neural networks. Hybrid schemes—such as the point-cloud/graph combined formulations used in DiffCSP, MatterGen, and GNOME—provide balanced expressiveness and geometric accuracy, enabling efficient modeling of complex crystalline systems.

The generative mechanisms of crystal models are more intricate than those of molecular generative models. According to their learning paradigm, recent models may be divided into four categories. Multimodal-aligned models (*e.g.*, Chemeleon) employ contrastive learning to align textual descriptions with crystal structures, enabling text-conditioned generation across complex chemical spaces. Experiment-driven models (*e.g.*, XtalNet) directly map PXRD spectra to crystal structures via integrated pretraining and conditional generation, circumventing traditional database matching. Twin-model collaborative frameworks (*e.g.*, DAO) combine generative modules with high-precision property predictors, achieving concurrent optimization of structural stability and targeted properties. Dynamics-informed models incorporate nonlinear kinetic effects such as impurity-induced frustration and non-equilibrium crystallization into sampling strategies. These mechanisms collectively provide diverse pathways for addressing the unique challenges posed by crystalline systems.

The applications of generative crystal models are further examined across pharmaceuticals, quantum materials, and energy-related functional materials. In pharmaceutical solid-state chemistry, pocket-conditioned molecular generative models provide insights into integrating solid-form considerations with molecular design. In quantum materials research, constraint-embedded generators such as SCIGEN enable large-scale exploration of lattice-specific structures—including Archimedean and Lieb lattices—yielding candidates for quantum spin liquids, flat-band materials, and exotic magnetic phases. In energy materials, the DAO framework demonstrates the potential of “generation–prediction” coupling by successfully designing superconductors and high-capacity layered cathodes, showing strong coherence between computational predictions and experimental verification.

Summary and prospects Although notable progress has been made in generative crystal models, significant challenges remain in both scientific understanding and practical deployment. Existing datasets are limited in structural diversity, physical fidelity, and multimodal coverage, making it necessary to construct large-scale, high-quality datasets integrating experimental observations, first-principles calculations, and model-generated structures. The physical interpretability and reliability of generated structures, particularly in multicomponent and nonequilibrium systems, require systematic investigation. Establishing closed-loop platforms that integrate model generation, autonomous synthesis, high-throughput characterization, and iterative retraining remains a key step toward practical materials discovery. Physics-aware constraints, uncertainty quantification, and robust generation under sparse data conditions are essential for improving model reliability. Furthermore, industrial applications demand models capable of handling large unit cells, defect chemistry, compositional tunability, and synthesis feasibility. With the development of large materials models, automated laboratories, and data-centric infrastructures, generative crystalline modeling is expected to become a core driver in building an intelligent, automated, and industrialized materials innovation ecosystem.

Keywords crystal structure prediction; generative models; artificial intelligence for science; crystalline material design