特约评述

DOI: 10.12211/2096-8280.2023-008

人工智能蛋白质结构设计算法研究进展

陈志航，季梦麟，戚逸飞
（复旦大学药学院，上海 201203）

摘要：蛋白质是各类生命活动不可缺少的承担者，其序列决定了折叠后的三维结构和功能。这些具有特定功能的蛋白质在生物医学等多个领域具有重要的应用价值。计算蛋白质设计可以根据所需的蛋白功能和结构设计氨基酸序列，生成自然界中不存在的蛋白质。传统计算蛋白质设计通常采用能量函数和特定的搜索优化算法获得设计的序列。近年来，随着先进算法的发展、大数据的积累和计算机硬件算力的增长，人工智能技术得到了蓬勃发展，并逐渐应用于蛋白质设计领域。本文综述了近年人工智能在蛋白质结构设计中的进展，侧重于各类算法的介绍，从固定骨架设计、可变骨架设计和序列结构生成三个方面回顾了最新的蛋白质结构设计算法，并阐明了其相对于传统计算方法的新颖性和创新性。在人工智能技术的赋能下，蛋白质设计的成功率和合理性获得大幅提高，按需功能蛋白设计的时代即将到来。

关键词：蛋白质设计；蛋白质工程；人工智能；深度学习；蛋白质序列与结构

中图分类号：Q816　　**文献标志码：**A

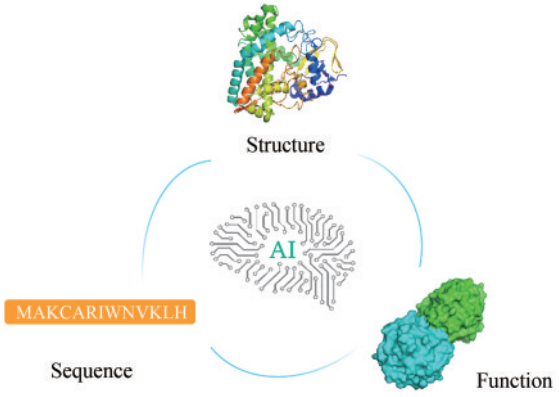
Research progress of artificial intelligence in desiging protein structures

CHEN Zhihang，JI Menglin，QI Yifei
（*School of Pharmacy*，*Fudan University*，*Shanghai 201203*，*China*）

Abstract: Proteins are essential to life as they carry out a great variety of biological functions. Protein sequences determine their three-dimensional structures, and therefore physiological functions. Proteins with specific functions have important applications in many fields such as biomedicine, where they are utilized in drug design and delivery. In the past, protein engineering and directed evolution are commonly used to improve the activity and stability of proteins. These methods, however, are both complex and expensive, as they require a large number of biological experiments for validation. Computational protein design (CPD) allows the design of amino acid sequences based on desired protein functions and structures, and more intriguingly, generation of proteins even not found in nature. Conventional CPD uses energy function and optimization algorithm to design protein sequences. In recent years, with the rapid development of artificial intelligence (AI) technique, the accumulation of big data and the development of high speed computing, AI has made great progresses in learning, and been successfully applied in CPD. In this review, based on

收稿日期：2023-01-13 修回日期：2023-03-15
基金项目：国家自然科学基金（22033001）
引用本文：陈志航，季梦麟，戚逸飞. 人工智能蛋白质结构设计算法研究进展[J]. 合成生物学，2023，4(3)：464-487
Citation：CHEN Zhihang，JI Menglin，QI Yifei. Research progress of artificial intelligence in desiging protein structures[J]. Synthetic Biology Journal，2023，4(3)：464-487

the input constraints and sampling space size, we present a systematic overview of recent applications of AI in protein design from three aspects: fixed-backbone design, flexible-backbone design, and sequence structure generation. We focus on algorithms and protein feature encoding, present the effect of dataset size and architectural improvements on model performance in prediction, and showcase several enzymes, antibodies, and binding proteins that were successfully designed using these models. The advantages of AI compared with traditional CPD methods are also discussed. Finally, we highlight challenges in AI-aided protein design, and propose some strategies for solutions.



Keywords: protein design; protein engineering; artificial intelligence; deep learning; protein sequence and structure

蛋白质是生物体内的“生命机器”，在转录、翻译、信号传导和细胞周期调控等几乎所有的生命活动过程中发挥着至关重要的作用。天然蛋白质以一种极端经济且严谨的方式对其氨基酸序列进行编码，并在体内自发折叠成特定三维结构来实现其生物活性。探寻蛋白质结构和功能的关系在过去几十年内一直是基础医学和生物学研究的焦点。随着对蛋白质功能研究的深入和实际应用的展开，天然蛋白质已无法完成人类日益增长的需求。对蛋白质的改造和设计也从依赖天然蛋白的随机突变和定向进化，向理性设计甚至是从头设计（*de novo* design）全新的具有特定功能的蛋白质转变。

蛋白质的氨基酸序列排布方式决定了其折叠后结构和活性功能。对于一个链长为 100 个氨基酸的蛋白质，其可能的氨基酸序列组合有 20^{100} 种。在如此广大的序列空间内进行氨基酸序列的优化搜索面临着巨大的困难[1]。蛋白计算设计避免了相对随机的突变策略，并提供了基于蛋白质的生物物理和生物化学原理的指导性设计蓝图。计算蛋白质设计的目标是设计一个能够折叠成预定义的结构且具有所需功能的氨基酸序列，通常会从一个已知的结构出发，保留活性位点，并修改部分序列以提高所设计蛋白质的稳定性或实现新的功能[2-3]。

依据 Anfinsen 的折叠热力学假说[4]，蛋白质折叠到最低自由能状态，其 3D 结构由氨基酸序列决定。然而，在折叠过程中最重要的不是折叠态的绝对能量，而是折叠态与最低的备选态之间的能量差。这种计算不仅涉及到所有可能的氨基酸序列，而且涉及到所有可能的结构，因此多数现有的方法都集中在寻找所需结构的最低能量氨基酸序列这个更容易处理的问题上。目前广泛使用的方法仍然是基于能量函数和启发式采样方法的算法[5]。RosettaDesign[6]、FoldX[7]、EvoDesign/EvoEF2[8] 等设计方法使用使用蛋白质结构参数化的打分项来量化氨基酸序列和特定三维结构之间的匹配度，其中 RosettaDesign 是目前使用最为广泛的方法之一。RosettaDesign 采用能量函数[9] 来捕捉序列-结构关系，对结构中每个残基侧链的氨基酸种类和构象进行采样，并使用蒙特卡洛模拟退火等方法进行优化以获得低能序列和构象。在

过去的三十年中，基于能量函数的蛋白计算设计取得了巨大的进展，包括设计新颖的3D折叠[10]、酶[11]和复合物[11]，更包括免疫信号[12-13]、靶向治疗蛋白[14-15]、蛋白质开关[16-17]、自组装蛋白[18-19]等。尽管取得了这些成功，但是基于能量函数的蛋白质设计方法准确度仍然较低，在没有多轮实验试错的情况下无法可靠使用，导致蛋白设计实验成功率难以提升[20]。

以深度学习为代表的人工智能技术，随着算法和算力的发展以及大数据的积累，近期在多个领域取得了重要进展。在生物学和化学领域中，深度神经网络的优势在于可以从蛋白质结构的原子坐标、氨基酸种类、二级结构等简单的输入数据中学习高阶特征。深度学习模型一旦学会了蛋白质特征间的关系，就可以用来为结构生物学和生物分子的设计提供新的见解和指导。海量具备真实性和可用性的数据[21-24]使得深度学习表现出比经典物理模型或其他机器学习方法更好的性能[25]。目前，深度学习已被应用于蛋白质-配体打分[26-29]、蛋白质-蛋白质相互作用预测[30-32]、化合物性质预测[33]、分子结构生成[34-36]等诸多领域[37]，近期更是在蛋白质结构预测方面取得了引人注目的进展。以AlphaFold[38]和RoseTTAFold[39]为代表的结构预测算法通过多序列比对（multiple sequence alignment，MSA）、基于注意力机制的序列分析和蛋白三维结构生成等模块，以端到端的方法大幅提高了蛋白三维结构预测的准确率。

在蛋白质设计领域，近年来设计方法也逐渐从基于物理化学原理的打分函数，转变到利用深度学习进行设计的策略。本文将回顾近年深度学习在蛋白设计方向的研究进展，按照模型的采样方式、搜索空间大小和蛋白设计任务的难易程度分成三个方面：①固定主链构象的蛋白质设计；②可变骨架的序列设计；③结构和序列生成模型。在固定骨架设计任务中，模型已知蛋白骨架的走向和残基位置，仅需对骨架上的序列进行设计；可变骨架设计中则允许一定程度的蛋白骨架结构柔性，模型搜索空间增大，设计的自由度提高；生成模型可从头生成全新的蛋白序列和骨架，或根据局部功能位点进行结构补全，解决了前两类设计方法中初始骨架来源的问题。

1 固定主链构象的蛋白质设计

固定骨架蛋白质设计的目标是找到一个最能折叠成目标结构的氨基酸序列，也可以看作是找到一个折叠成目标结构的概率比其他所有序列都大的序列[40-41]。

1.1 早期工作

SPIN使用一个基于片段局部特征和能量非局部轮廓的神经网络来解决基于固定骨架结构的蛋白序列设计问题[42]，其输入特征包括目标蛋白质主链的φ、ψ二面角，通过比较相邻5个残基的结构片段得到局部片段衍生序列图谱[43]，并采用DFIRE统计势[44]计算全局能量。SPIN在500个蛋白质的测试集上平均序列恢复率约为30%。

Qi团队[45]开发了用于蛋白计算设计的神经网络模型，使用目标残基及其相邻残基的距离、主链二面角和二级结构等几何特征，以约3倍于SPIN的训练集对神经网络进行训练，将序列恢复率提高至33%。同期，SPIN2[46]使用一个具有三个隐藏层的神经网络，在蛋白骨架特征中添加另外两个骨架二面角θ和$\imath$，并改用正弦和余弦表示作为特征输入，将序列恢复率提高至34%。

SPIN2仅使用一维结构特征，不足以表征具有复杂三维结构的蛋白质。SPROF[47]则通过两两残基距离的二维距离矩阵来表示蛋白质中残基之间距离（图1）。SPROF使用双向长短时记忆网络与自注意力二维卷积神经网络来预测蛋白质序列。SPROF方法在独立测试集上取得了39.8%的序列恢复率，明显高于仅从一维结构特征训练的SPIN2方法取得的34.6%。

1.2 卷积神经网络

卷积神经网络（convolutional neural network，CNN）[48]是最成功的神经网络架构之一，主要包括卷积和池化两种基本操作。在蛋白质设计中，卷积层用于对蛋白质残基间距离图或蛋白质在三维空间网格中的密度距离分布进行变换并提取特征，更深的卷积网络能从输入特征中迭代提取更复杂的特征。池化层通过连续降采样的方式逐渐降低

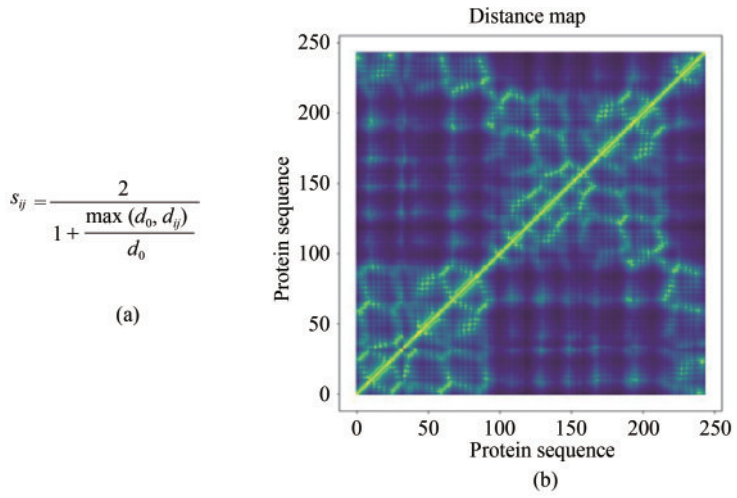


图 1　SPROF 中残基距离计算方法

（a）d_{ij} 为残基 i 和 j 的 C_α 原子之间的距离，d_0=0.4 nm；（b）蛋白质残基-残基距离矩阵

Fig. 1　Calculating the distance of residues in SPROF

(a) d_{ij} is the distance between the Cα atoms of residues i and j, d_0=0.4 nm, and (b) matrix for residue-residue distance of a protein structure.

数据的空间尺寸，以减少网络中的参数数量，使得计算资源耗费变少，也有效控制过拟合。另外，卷积使得模型能够处理大小可变的输入数据。

ProDCoNN [49]、Anand 等 [50] 发展的方法和 DenseCPD [51] 均使用三维卷积网络从目标残基周围的三维结构环境特征来预测残基类型（图 2）。模型以残基周围的原子密度和原子类型网格作为输入，使用 DenseNet [52] 等多层卷积网络对密度分布数据进行学习，捕获不同尺度下的结构信息。网络中的卷积层提取蛋白质共价键信息、键角、二面角和二级结构的特征图，池化层精简特征图数量，最后输出目标残基为 20 种天然氨基酸的概率大小。其中，ProDCoNN 和 Anand 模型分别在相同的 T500 和 TS50 上达到约 40% 的序列恢复率，DenseCPD 则达到 51%。

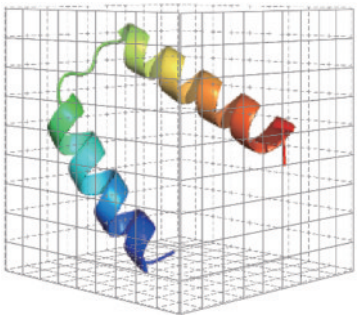


图 2　三维卷积神经网络提取网格中的蛋白质空间结构信息

Fig. 2　Extracting spatial information of a protein structure from 3D convolutional neural network

MutCompute [53] 使用残基原子（C，H，O，N，S）坐标、部分电荷（partialcharge）和溶剂可及表面积（solvent-accessible surface area，SASA）作为结构特征输入 3D-CNN 网络。MutCompute 以蛋白质中心目标残基的 C_α 为中心，掩蔽 2 nm 立方体内的所有肽原子，构造为该残基的局部化学微环境（microenvironment）样本，以这种方式从 19 300 个蛋白质结构中构造 170 万个微环境作为训练集。训练后的模型能够识别稳定的突变，根据残基局部化学微环境预测蛋白质中不稳定的位点。Lu 等 [54] 使用 MutCompute 模型设计了一种聚对苯二甲酸乙二醇酯（PET）水解酶，指导野生型水解酶 PETase 组合 N233K/R224Q/S121E 和骨架的 D186H/R280A 五个位点的突变，得到的突变体 FAST-PETase 具有优异的催化活性和热稳定性。FAST-PETase 在 30～50 ℃和一系列 pH 水平之间显示出优越的 PET 水解活性，适用于至少 51 种未经处理的 PET 降解，工业上可广泛用于塑料的回收与循环。

TrDesign [55] 使用基于卷积神经网络的结构预测模型 trRosetta 进行反向序列设计。首先将随机氨基酸序列输入到蛋白质结构预测模型 trRosetta [56] 中，输出残基之间距离、角度和二面角的分布（图 3）。其次计算预测分布与目标蛋白结构分布之间的差异，使用梯度反向传播来更新氨基酸序列，重复该过程直到收敛。TrDesign 通过 trRosetta 遍历全局构象势能面，和 RosettaDesign 单点能量

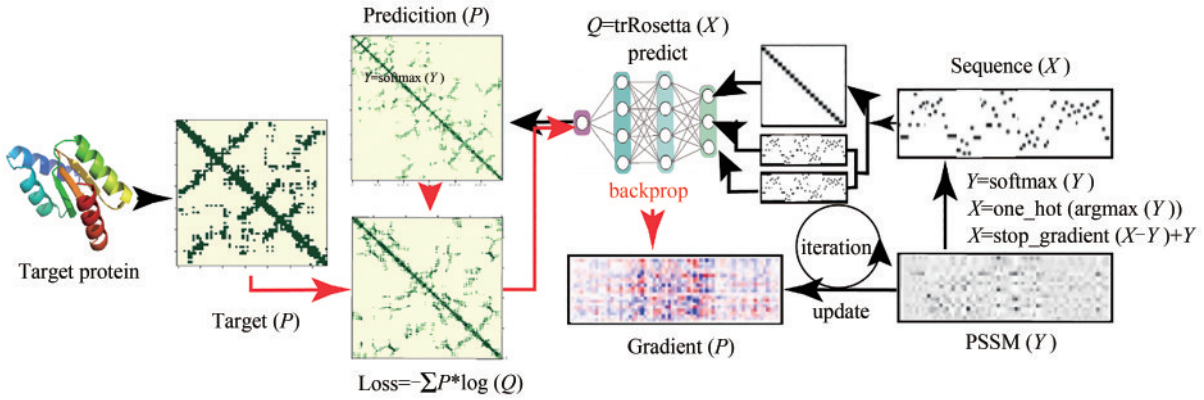


图3 trDesign模型架构图

Fig. 3 Architecture for the trDesign model

计算方法相比，能够多方面捕获序列折叠势能，保证设计蛋白质的可折叠性和稳定性。高分辨率的Rosetta模型用于创建目标结构的深度能量极小值，而低分辨率的trRosetta模型用于减少在能量极小值点备选序列的数量。将两种方法结合，能够在遍历势能面的同时减少候选序列的数量。然而使用trRosetta进行反向序列设计需要反复运行trRosetta模型，计算效率不高并且容易陷入势能面上次优解。

1.3 图神经网络

图神经网络（graph neural network，GNN）运行在图（graph）这种非欧氏数据结构上，已被广泛应用于知识图谱、社交网络、药物发现和蛋白质生物信息学等领域[57-58]。蛋白质结构可用图进行编码，残基信息编码在节点特征中，空间中相邻残基之间的关系可编码为边特征。

在蛋白质序列中距离较远的一对残基在折叠后的三维结构中可能存在近距离相互作用。在网络中引入注意力机制使图网络能够识别残基在三维空间中的紧密/稀疏关系，在考虑全局构象的同时又聚焦局部关键特征。此外，图结构在表示蛋白质结构时，可同时描述主链柔性拓扑结构的全局整体特征和精确原子位置的局部细节特征。使用图结构表征蛋白质具有更高的灵活性和较高的计算效率。

GraphTrans[59]使用图$G=\{V, E\}$表示蛋白质结构，节点特征$V=\{v_1, v_2\cdots v_N\}$描述每个残基的氨基酸类型，边特征$E=\{e_{ij}\}_{i\neq j}$捕捉它们之间的关系（图4）。模型通过三维结构的自回归解码Transformer[60]以捕获序列残基之间稀疏的成对依赖关系信息。GraphTrans模型可以有效地捕获序列和结构之间的高阶依赖关系，序列恢复率在Ollikainen 40测试集上达到39.2%，高于RosettaDesign的33.1%；在CATH测试集上残基困惑度（per-residue perplexities）为6.85，精度比以往基于神经网络（LSTM：17.13；SPIN2：12.61）的模型显著提高。

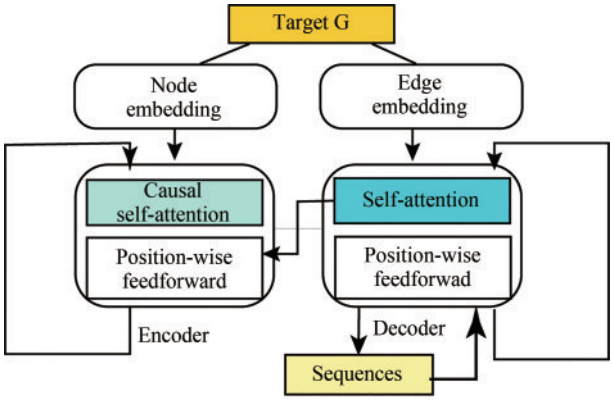


图4 GraphTrans编码器与解码器示意图

Fig. 4 Architecture for the GraphTrans encoder and decoder

一个给定的蛋白质结构，对应于单一的距离矩阵，可以由许多不同的满足距离矩阵约束的同源序列折叠而成。ProteinSolver[61]是一个预训练的图卷积神经网络，将使用氨基酸序列填充特定目标结构表述为一个约束满足问题（constraint satisfaction problem），其目标是在兼顾长程和短程的约束的同时，为链中的残基分配氨基酸标签，使得残基之间的作用力是有利的。训练好的ProteinSolver网络能够以很高的准确度快速生成数

千个匹配特定蛋白质拓扑结构的序列。

为同时将蛋白质残基的几何结构和关系特征纳入统一网络架构，Jing 等[62]提出使用几何向量感知器（geometric vector perceptron，GVP）（图 5）来代替多层感知器（multi-layer perceptron，MLP）。给定一个标量和向量输入特征（$\mathbf{s}$，$\mathbf{V}$）的元组，GVP 将残基原子三维坐标转化为残基距离特征，并将其与标量特征组合，输出一个更新的元组（$\mathbf{s}'$，$\mathbf{V}'$）。GVP 模型在标量特征进行转换之前，会将其与转换后向量特征的范数进行拼接，这允许模型从输入向量中提取旋转不变信息，以便图中节点的信息传播。GVP-GNN[62]使用 GVP 层来增强 GNN 对于几何结构特征的感知，并能够在欧氏向量特征上执行和表达，在蛋白质结构的质量评估和序列设计方面具有独特的优势。

Orellana 等[63]对上述 GVP 的结构提出了改进，使用图卷积神经网络（graph convolutional neural network，GCN）同时对节点和结构信息进行端到端的学习。模型添加每个氨基酸骨架中所有原子之间的归一化距离作为节点特征；将每个氨基酸的 C_α 与其 k 个最近邻氨基酸的 C_α 之间的标准化距离（k 值邻近，k=35）作为边特征，然后将节点和边特征嵌入空间进行编码，并将其引入到 GCN 模型中，输出为序列中每个位置的氨基酸种类，可用于指导基于能量函数的蛋白设计方法。该模型的序列恢复率从以往模型的 40.2% 提高到 44.7%。

TERMinator[64]使用三级 motifs（TERM）捕获序列-结构关系[65]，融合了残基原子坐标信息作为特征。TERMinator 提取目标蛋白中与 TERM 结构匹配的信息来构建节点和边，嵌入空间编码后输入图神经网络中，输出序列空间中拟合了能量函数的 Potts 模型。GNN Potts 模型编码器接受 TERM 数据并提取特征，使用使用马尔科夫链蒙特卡洛（Markov chain Monte Carlo，MCMC）模拟退火算法生成最优序列，输出位置氨基酸标签。作者还进行消融了实验，完整的 TERMinator 模型（恢复率 41.73%）性能强于消融 TERM 信息输入的模型（恢复率 40.29%），表明联合使用 TERM 和空间坐标作为特征有利于蛋白质设计。

ESM-IF1[66]使用 GVP 来学习向量特征的等变转换和标量特征的不变变换。该工作尝试以下三种架构：①GVP-GNN；②更宽和更深的 GVP-GNN-large；③由 GVP-GNN 结构编码器和 Transformer 组成的混合模型。ESM-IF1 使用 AlphaFold2 预测的 1200 万个结构，将训练数据增加了近 3 个数据级，克服了实验数据的限制，最终在 CATH 4.3 测试集上进行评估并根据残基困惑度（perplexity，越低越好）和序列恢复率进行比较。GVP-GNN-large 和 GVP-Transformer 模型均在序列恢复率上比简单 GVP-GNN 提高约 9%，达到与 DenseCPD 相当的 51%，且困惑度由 6 降低至 4。在突变效应的 zero-shot 多项预测测试中（包括复合物稳定性、结合亲

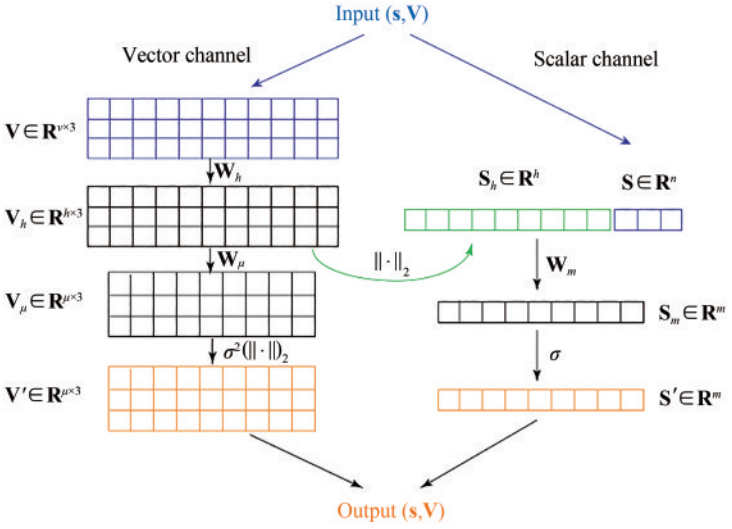


图 5　GVP 原理示意图

Fig. 5　Architecture for GVP

合力和插入效应），ESM-IF1均取得优异的性能表现。

McPartion[67]引入了一种深度SE(3)-等变图Transformer架构，直接对从蛋白质主链结构衍生的特征进行操作，实现了同时预测每个残基的氨基酸类型和侧链构象。局部感知图（locality aware graph）Transformer利用蛋白质主链的几何形状来优化单个残基和残基对的特征表示，并将注意力限制在空间上相邻的残基对上。该模块的输出和蛋白质主链坐标一起被传递到张量融合网络（tensor fusion network，TFN）[68]输出一个标量和残基位置，然后由TFN-Transformer为每个输入残基产生侧链构象和氨基酸类型。作者评估了5种不同的残基掩蔽方法并分别进行了损失函数、网络架构和模型超参数的消融实验，发现从损失函数中移除侧链坐标均方根偏差（root mean squared deviation，RMSD）和预测的侧链原子之间的成对距离两个特征显著降低了测试蛋白上的天然序列恢复率。除此之外，移除模型中的TFN-Transformer层对恢复率的影响最大。与几种现有的序列设计方法对比而言，该模型在4个测试集上展现了更高的序列恢复率。

ABACUS-R[69-71]使用一个多任务学习的编码器-解码器网络，根据固定骨架上局部环境预测中心位置的残基类型（图6）。网络的输入是目标残基与最邻近k个残基联合形成的局部特征，包含空间层面的相对位置与取向信息、序列层面的相对位置信息以及邻近残基的残基类型。ABACUS-R模型不需要显式地模拟侧链，从而避免优化的过程。模型拟合了给定结构下侧链类型的能量函数，通过在目标骨架上残基的迭代，逐轮降低随机残基数目，使得设计结果逐渐收敛，产生自洽的整体序列。ABACUS-R在单个残基平均序列恢复率达到53%，多个湿实验结果（包括X射线晶体学解析的晶体结构）表明，ABACUS-R在设计精度和成功率方面都优于基于能量函数的从头序列设计方法。

Roney等[72]认为AlphaFold从蛋白质的共进化

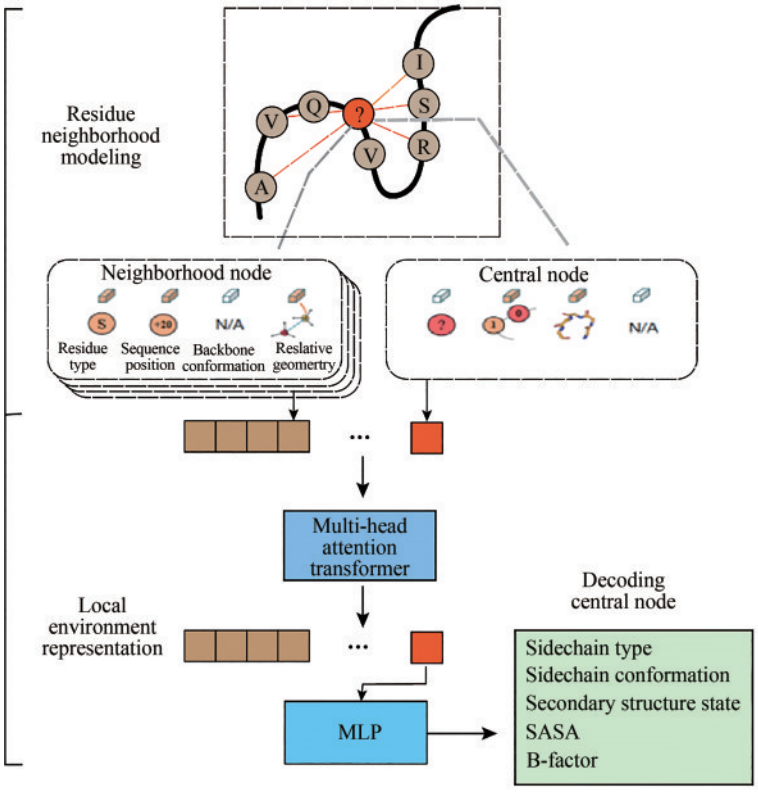


图6 ABACUS-R模型架构示意图

Fig. 6 Architecture for the ABACUS-R model

数据中学习了一个高精度的能量函数，可以在不使用任何共进化数据的情况下，确定蛋白质 3D 结构和序列之间的关系，从而用于蛋白质设计问题中。该流程类似于 TrDesign，将目标蛋白骨架结构提供给 AlphaFold 作为模板，最小化目标结构和预测结构之间的差异，并优化关于输入序列的复合置信度评分（composite confidence score）。该设计方法的序列恢复率达到约 30%。

ProteinMPNN[73] 参考 GraphTrans，使用具有 3 个编码器和 3 个解码器层以及每层宽度为 128 的消息传递网络（message passing network，MPNN）。作者认为相较于残基主链二面角和旋转走向，残基 N、C_α、C、O 和 C_β 原子之间的距离提供了更好的归纳偏置来捕获残基之间的相互作用。将上述特征输入 MPNN 网络（图 7），使模型预测序列恢复从 41.2% 增加到 49.0%。

虽然不少蛋白设计模型都致力于提升设计序列的恢复率，但在实际的蛋白设计应用中，恢复率最高的序列并不一定是最优解。因此，ProteinMPNN 在设计时使用了采样温度来获取更多的差异序列。PDB 数据库在收集蛋白质晶体结构数据时会根据序列对原子坐标进行修正，ProteinMPNN 训练时在骨架上添加高斯噪声来避免模型学到这种修正带来的误差，以提高模型稳定性并增强模型的泛化能力。噪声的添加在大部分情况下降低了 ProteinMPNN 的序列恢复率，并使 AlphaFold 对设计序列进行结构预测时更具有鲁棒性。

ProteinMPNN 还使用一种 order-agnostic 方法使得模型能在结构一部分固定的情况下设计其他部分，这使得 ProteinMPNN 适用于更复杂的结构，例如蛋白-蛋白复合物、环状蛋白、蛋白质纳米颗粒等。除了计算实验，研究人员使用 ProteinMPNN 进行了蛋白质单体、蛋白质纳米笼和蛋白质功能设计，并对先前使用 RosettaDesign 设计失败的蛋白进行了重新设计。这些设计蛋白能在大肠杆菌体系中可溶表达，并在生化实验中验证了其结构和活性，证明了 ProteinMPNN 设计蛋白的可靠性和合理性。

如果一个设计氨基酸序列的每个残基都与其局部环境很好地吻合，那么它就有望折叠成一个与目标结构相似的结构，ProDESIGN-LE[74] 便采用该思路。ProDESIGN-LE 以每个邻近残基的残基类型和相对于中心残基的 3×3 变换矩阵 $\boldsymbol{R}$ 和三维平移向量 $\boldsymbol{t}$ 来表示中心残基的局部环境，将特征输入一个 3 层的 Transformer 来学习残基对其局部环境的依赖性，并输出其嵌入图，后进一步使用全连接层将嵌入图转化为 20 种氨基酸类型的分布。训练好的 Transformer 模型在目标结构的序列上迭代地选择合适的残基，并相应地更新相邻残基的局部环境，最终获得所有残基都与自身局部环境匹配良好的设计序列。ProDESIGN-LE 模型在计算指标评估和实验验证上均取得不错的结果，在设计的 5 个 CAT Ⅲ 蛋白中，有 3 个具有良好的溶解性。

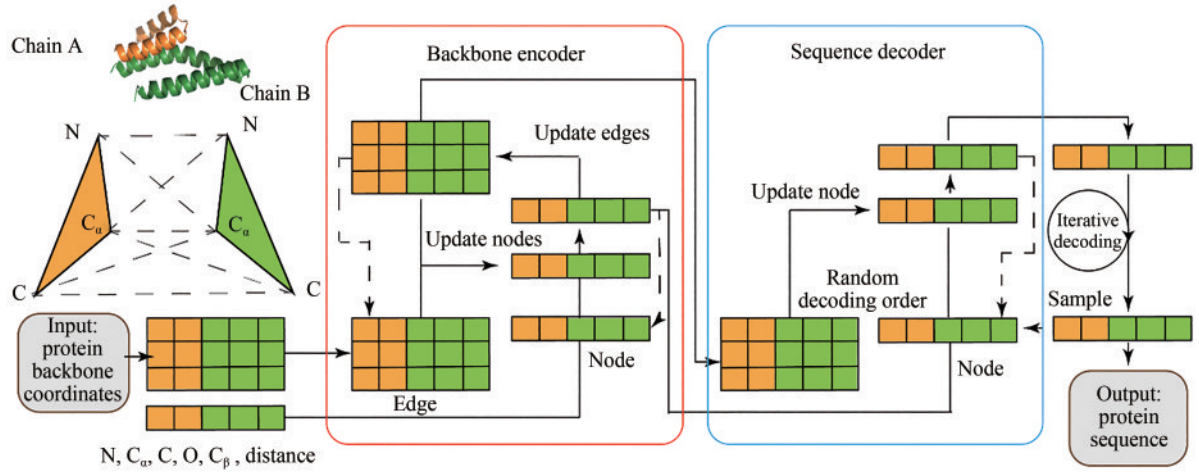


图 7　ProteinMPNN 模型的整体结构

Fig. 7　Architecture for the ProteinMPNN model

与CNN方法相比，图模型不需要像CNN那样单独处理每个残基及其周边结构，从而减小了编码的数据规模并提高了训练效率。GNN能够充分挖掘结构信息并获得不错的序列恢复率，能够正确处理序列中残基对的长、短程相互作用关系，可以在效率和精度之间取得较好的平衡。

随着固定骨架蛋白质序列设计模型的不断发展，其预测性能和精度大幅度提升，序列恢复率逐步提升，预测困惑度逐步降低（表1，表2）。

表1 固定骨架序列设计模型在CATH 4.2测试集上的序列恢复率和困惑度

Table 1 Sequence recovery rate and perplexity of the fixed-backbone sequence design model on CATH 4.2 test set

模型 Models	恢复率/%(↑) Recovery/% (↑)	困惑度(↓) Perplexity (↓)
GraphTrans	35.82	6.63
StructGNN[76]	37.1	6.49
GVP-GNN-large	39.20	6.17
GVP-GNN-Transformer	38.30	6.44
GVP-GNN-Transformer+AF2	51.60	4.01
ProteinMPNN	45.96	4.61
ProDesign	50.22	4.69
PiFold[77]	50.22	4.62
LM-DESIGN[75]（PiFold）	55.65	4.52

2 可变骨架的序列设计

与固定骨架设计问题不同，在可变骨架设计问题中，蛋白质确切的骨架结构通常都是未知的，因此在设计过程中需要同时考虑优化序列和结构。

2.1 幻想设计

深度学习神经网络能够从蛋白质结构或节点关系中识别和提取特征并将这些特征显著增强后输出。若反其道行之，对神经元输入一些抽象的特征，让每个神经元模拟出最可能具有这些特征的蛋白结构，再将结构信息反传回网络，经过多轮迭代优化即能生成最合适的蛋白序列或结构。2015年Google发布的DeepDream便是能够以此原理在图片中产生不存在的物品，生成的图片如同梦境中的画面一样。

前文提到trRosetta能够快速预测一个蛋白质序列的空间约束，Anishchenko等[80]重新训练了一个背景网络，将输入trRosetta的序列在自身的输出结构上不断迭代，使预测结构的空间约束逐渐具有清晰的分布，这种方法被称为幻想（hallucination）设计。首先将一个随机序列转换为折叠蛋白序列

表2 固定骨架序列设计模型在TS50 &TS500测试集上的序列恢复率和困惑度

Table 2 Sequence recovery rate and perplexity of the fixed-backbone sequence design model on TS50 &TS500 test sets

模型类别 Group	模型 Models	TS50		TS500	
		恢复率/%(↑) Recovery/%(↑)	困惑度(↓) Perplexity(↓)	恢复率/%(↑) Recovery/%(↑)	困惑度(↓) Perplexity (↓)
MLP	SPIN	30.00	—	—	—
	SPIN2	34.00	—	—	—
	Wang's model	33.00	—	—	—
CNN	SPROF	39.80	—	—	—
	ProDCoNN	46.50	—	—	—
	DenseCPD	50.71	—	55.53	—
GNN	StructGNN	43.89	5.40	45.69	4.98
	GraphTrans	42.20	5.60	44.66	5.16
	GVP-GNN	44.14	4.71	49.14	4.20
	GCA[78]	47.02	5.09	47.74	4.72
	ADesign[79]	48.36	5.25	49.23	4.93
	ProteinMPNN	54.43	3.93	58.08	3.53
	PiFold	58.72	3.86	60.42	3.44
	LM-DESIGN（PiFold）	57.89	3.50	67.78	3.19

的编码，同时输入随机噪声得到背景的空间约束。使用马尔科夫链蒙特卡洛（MCMC）算法对序列进行随机突变，再将其输入 trRosetta 模型中逐轮预测空间约束，以 Kullback-Leibler（KL）散度对序列约束和背景约束的分布差异进行优化，使得到的空间约束逐渐逼近真实蛋白质，并借此折叠蛋白 3D 结构（图 8）。

TrDesign-motif[81] 将 trRosetta 和 hallucination 有机结合起来用于蛋白质结合 motif 的设计。对于活性位点，初始输入骨架的 2D 特征作为目标分布，让 motif 功能部分预测序列与原结构尽可能地相似；而在自由幻想部分，将随机噪声的 2D 特征分布作为背景，让生成的序列尽可能远离其分布。使用混合的损失函数来优化结构和序列，创造出一个携带功能 motif 片段的新蛋白结构。

RFDesign 使用 constrained hallucination[82] 对幻想算法进行约束，优化序列，在保证预测结构的功能基序（motif）与目标结构接近的同时，自由幻想生成其非功能位点（图 9）。inpainting[82] 进行蛋白结构补全（即 RFjoint2[82]），将 trRosetta 换成 RoseTTAFold 框架，并尝试不同的结构掩蔽方式训练一个蛋白结构和序列模型，从功能位点出发填充非功能区的序列和结构，创建一个可行的蛋白质主链。inpainting 可以同时进行结构和序列生成，不依赖于 trRosetta 或反向传播的更新，可以通过输入主链走向来提高性能。

研究人员使用以上三种幻想方法设计了金属蛋白、酶活性位点和蛋白结合蛋白等，并都进行了计算机模拟和实验测试相结合的验证[82]。模型中的 inpainting 和 hallucinate 模块能够实现大肠杆菌铁蛋白（*E. coli* bacterioferritin）双铁结合位点的重新构建，在设计的 96 个铁蛋白结构中有 76 个可溶性表达，8 个具有金属结合的特征光谱位移，3 个具有与 AlphaFold 折叠结构一致的二级结构（圆二

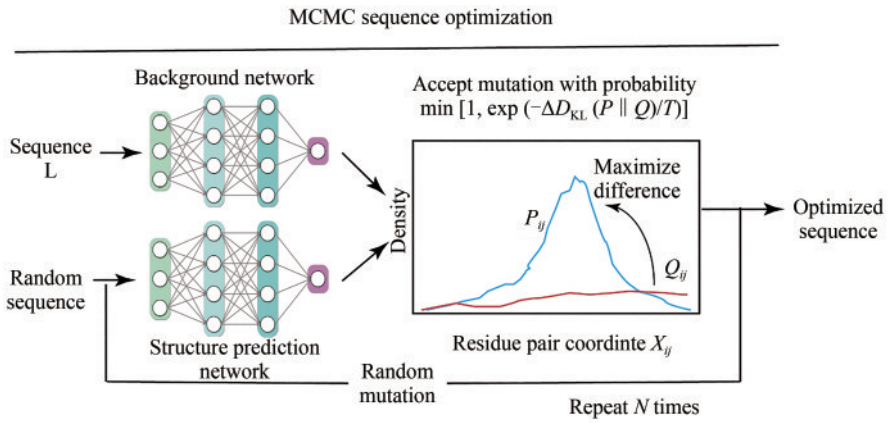


图 8　hallucination 模型原理示意图

Fig. 8　Architecture for the hallucination model

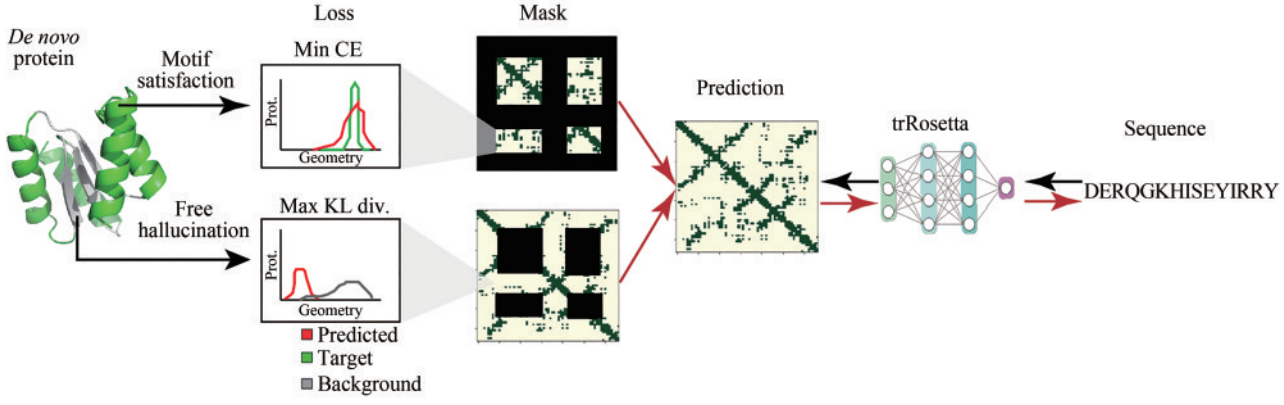


图 9　Constrained hallucination 模型原理示意图

Fig. 9　Architecture for the Constrained hallucination model

色光谱鉴定），并且能够稳定地与金属络合。幻想设计能够产生碳酸酐酶Ⅱ上三个Zn^{2+}配位组氨酸和环上苏氨酸组成的基序，并正确放置Zn^{2+}配位；幻想模型还构建了参与甾体激素生物合成的D5-3-酮甾体异构酶（KSI）的催化侧链，两种酶的活性位点与天然晶体结构几乎完全匹配。文章中还展示了幻想设计通过固定靶点蛋白和结合蛋白部分位点，修复缺失位点（inpainting）或自由幻想（hallucinate）全新的骨架结构来设计蛋白质结合蛋白的过程。其中设计的结合蛋白pdl1_inp_1与PD-L1结合能力（K_d=326 nmol/L）相较于野生型PD-1（K_d=3.9 mmol/L）增强；设计的TrkA在配体结合时呈现与天然结构相同的二聚化现象；多种设计的Mdm2癌基因结合蛋白与抑癌蛋白p53的天然N端螺旋结合紧密。

然而，RFDesign在使用RoseTTAFold生成时，由于采用单次运行预测缺失结构的方式，生成的序列长度和结构质量都受到一定限制。

Zhang等[83]基于上文提到的hallucinate方法，提出一种从头设计蛋白质折叠的自动自适应优化工具包AutoFoldFinder，通过序列优化的方式产生具有新蛋白元件排列方式的氨基酸序列与结构，使用同余系数图对齐（congruence coefficient map alignment，CM-Align）替换hallucinate方法中的KL散度，无需对整个接触图的全局比较，能够更精细地反映接触图在局部二级结构上的特征差异。AutoFoldFinder通过序列优化将生成一千条蛋白质序列中低相似度序列比例从22%提升至30.9%，加入CM-Align方法后，超过50%的结构与已知结构有显著差异。

最近Baker团队[84]发布了首个使用深度学习工具从头设计荧光酶结构的工作。研究人员选择合成荧光素酶底物二苯基特拉嗪（diphenylterrazine，DTZ）作为目标酶的作用底物，作者首先构建了DTZ阴离子构象系综，随后围绕每个构象，使用RIFGen方法[85-86]枚举了与DTZ相互作用的氨基酸侧链旋转异构体相互作用场（RIF），最后使用RIFDock将每个DTZ构象和RIF在约4000个天然蛋白骨架的中心腔中进行对接，以最大化蛋白-DTZ相互作用。此方法发现与DTZ结构互补的结合口袋中大多为核转运因子2（nuclear transport factor 2，NTF2）家族蛋白，将对接获得的骨架和口袋使用family-wide hallucination方法进行优化设计。

family-wide hallucination集成了无限制幻想设计[80，82]与Rosetta序列设计方法[55]，对环（loop）和可变区域（variable regions）的序列和结构进行从头设计，并对核心区域的结构进行序列优化。该方法从2000个天然NTF2序列出发，在序列空间中进行蒙特卡洛搜索，每一步都进行一次序列变化，并使用trRosetta进行结构预测。模型的损失函数由两部分构成：结构保守区域基于与NTF2-like蛋白实验结构的输入残基距离和方向分布的一致性进行评估；而可变区域基于网络预测与背景分布之间的KL散度计算的预测残基间几何结构的置信度进行评估。氢键网络也被纳入设计的结构中，以增加结构特异性。实验数据显示family-wide hallucination生成的1615个骨架在原生结构的空间内采样更多，并且比原生骨架或非深度学习能量优化生成的骨架具有更强的序列结构关系。

研究人员运用以上方法生成的蛋白骨架设计了人工荧光素酶，能够以高选择性催化DTZ的氧化化学发光。其中活性最强的酶LuxSit-i在保持与天然荧光酶催化效率相当的同时大大提高了对底物的特异性和热稳定性（变性温度>95 ℃）。

2.2 能量模型

可变骨架的蛋白质设计可以分解成骨架结构的生成和固定骨架设计两个独立的子任务。中国科学技术大学刘海燕组[87]提出了一种全新的、使用神经网络形式能量项的统计模型——SCUBA，使基于连续采样和优化主链中心能量面来设计新主链的方法成为可能。SCUBA模型将主链的可设计性分解为几个关键因素的作用，包括局部构象倾向性、肽主链氢键几何构象以及手性附着和紧密排列的侧链所需的骨架空间。研究者使用统计能量项来表示各种相互作用，用一种名为邻接计数神经网络（neighbor counting-neural network，NC-NN）的通用方法训练。NC-NN包含两步过程，首先通过基于核的密度估计（即邻接计数）从原始结构数据估计统计能量值，然后训练神经

网络（三层全连接感知机）表示势。得到的统计能量项，除了可以提供易于计算的函数值和导数用于结构采样和优化外，还可以高保真地表示复杂的、高维且高度相关的真实结构数据分布。

在模板未知条件下，使用神经网络形式的能量项模型 SCUBA 驱动的随机动力学（stochastic dynamics）和模拟退火算法（simulated annealing）来生成可设计的新蛋白质主链骨架，再使用前文中提到的 ABACUS2 [69] 对主链骨架序列进行序列优化和骨架松弛 [10] 设计的迭代，从而完成对蛋白质的可变骨架从头设计任务。在 9 种用 SCUBA 设计的高精度骨架蛋白结构中，其中有 4 种具有新颖的非天然结构。这一结果充分展示了 SCUBA 在蛋白设计中的实用性，特别是在设计功能蛋白时，能量函数驱动的骨架采样和优化可以很容易地进行定制，以促进对结构空间的广泛探索。另外，SCUBA+ABACUS2 [87] 策略所设计的蛋白质具有高于天然蛋白质骨架的热稳定性，设计成功率约为 42%（38 个经实验验证的蛋白质中有 16 个成功折叠，14 个 H2E4 蛋白质和 4 个 H4 蛋白质），设计的骨架与实验获得的结构一致，达到原子精度，同时设计的 H2E4 和 H4 蛋白与具有相似结构的已知天然蛋白质具有低序列同一性（平均同一性 14%）。

Liang 等 [88] 随后发展了一个基于级数展开的能量函数模型 OSCAR-Design。在四个独立的阶段中优化目标函数 $E_{total}=E_{side}+E_{bb}+E_{ref}$ 的各项参数，最大化原结构和其他旋转异构体之间的能量差；最小化天然环结构中选择环诱饵之间的 RMSD，最大化氨基酸组成与天然序列的相似性；惩罚埋藏的非氢键极性原子。作者使用 Monte Carlo 模拟退火算法对 OSCAR-Design 进行测试。OSCAR-Design 在侧链和 loop 预测任务中与 OSCAR [89-90] 和 LEAP [91] 一样准确。在从头设计任务中，OSCAR-Design 在测试集达到 38%～43% 天然序列恢复率，成功还原了 75% 的亲疏水性残基，氨基酸组成的整体相似性达到 90%。

3　结构和序列生成模型

在第一部分介绍的蛋白质设计工作中，设计过程往往从设计蛋白的主链结构开始，该结构可以源自天然蛋白质，蛋白结构预测模型的输出，根据对天然蛋白的观察、比较等方式手工搭建的大致三维构象等。近年来机器学习领域生成模型的巨大进展为生成全新的蛋白质结构和序列奠定了基础。深度生成模型在快速发现新颖、合理的蛋白质结构方面有着巨大的潜力。

3.1　生成对抗网络（GAN）与变分自编码器（VAE）

Huang 团队 [92] 提出了一种基于生成对抗网络（generative adversarial network， GAN）的生成模型，策略具体细节如图 10 所示。蛋白质的结构使用蛋白质主链上成对 C_α 之间的距离（以 Å 为单位）来表示。GAN 模型中的生成器通过输入一个正态分布随机变量 $z\sim N(0,I)$，输出一个成对距离图，判别器判断生成器输出的结果是真实的（数据样本）或是虚假的（生成器输出），而后生成器对生

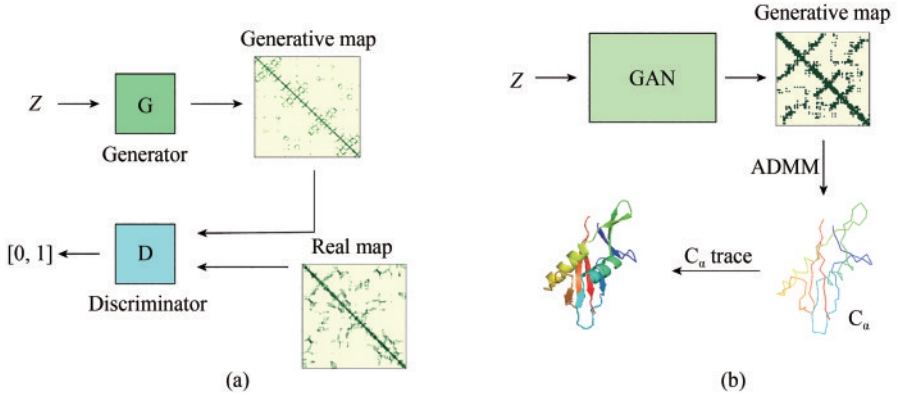


图 10　生成对抗模型用于蛋白二维接触图和三维骨架的生成

Fig. 10　Generative adversarial network for generating contact map and 3D backbone structure.

成的结果不断迭代优化用以欺骗判别器，整个模型最终输出得到合理的成对距离图。得到的距离图随后通过交替方向乘子法（alternating direction multiplier method，ADMM）折叠成3D结构从而得到C_α的坐标，最后使用一个快速追踪脚本将C_α原子的坐标匹配到一个合理的蛋白质骨架。研究者将此方案应用于补全蛋白质结构中缺失残基的任务，同时还扩展生成建模程序来解决端到端的结构恢复问题，并减少当前模型在精细局部结构中出错的问题。在后续研究中，Huang等[93]进一步优化了他们的方案，通过所有主链原子之间的成对距离来表示蛋白质结构，并提出了一种以可微分的方式直接恢复和细化相应主链坐标的方法（图11）。具体来说，在GAN生成骨架原子距离矩阵之后，采用卷积神经网络，通过自编码器损失从成对距离矩阵中恢复蛋白质骨架坐标。相较于ADMM恢复方法，这种新提出的方案是一种快速、完全可微分的方法，即生成的3D骨架坐标的错误可以反向传播到生成器网络。

以上提到的GAN方法在结构生成领域表现出了较好的性能，但也存在一定的弊端，例如生成的距离约束不能保证是欧氏有效的，因此不能恢复完全满足生成约束的3D坐标[94]。2020年Huang等[94]提出了一种构建蛋白质骨架的新方法Ig-VAE，使用变分自编码器（variational autoencoder，VAE）直接生成免疫球蛋白的三维坐标。模型的架构如图12所示。首先通过输入蛋白的原子坐标计算出主链残基二面角和距离矩阵，其次将距离矩阵输入编码器压缩特征得到低维的潜在空间表征，潜在空间表征传递给解码器，解码器直接生成蛋

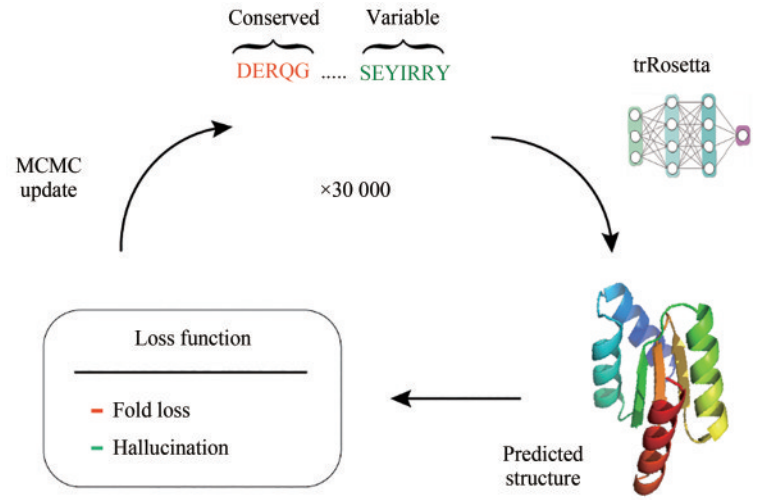


图11 Famliy-wide幻想蛋白质结构生成模型架构图

Fig. 11 Architectuer for the famliy-wide hallucination protein structure generation model

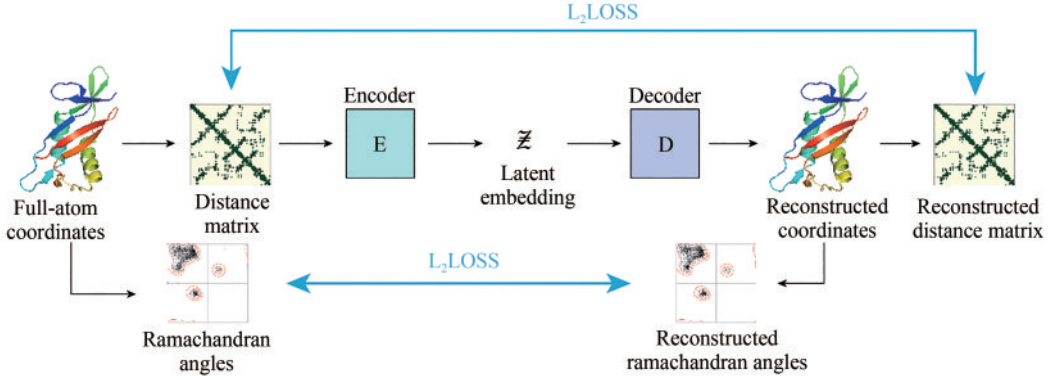


图12 Ig-VAE模型架构

Fig. 12 Framwork for Ig-VAE

白 3D 空间中的坐标（图 12）。通过重构出的坐标重新计算主链残基二面角和距离矩阵，角度和距离矩阵的误差都通过 3D 坐标反向传播进网络中。训练完成后，Ig-VAE 在结构嵌入及重构、隐空间插值以及生成能力方面表现良好，是一种构建单结构域抗体的有效工具。

2022 年许锦波组 [95] 提出了一种直接在三维坐标空间中对蛋白质结构进行建模的、基于 VAE 的模型，相比于先前提出的直接坐标生成模型 [3]，其应用仅限于固定长度的蛋白质，新提出的模型通过提取关于蛋白质几何形状的不变表征（invariant representations），并使用局部对齐的坐标损失函数直接在坐标空间上执行梯度优化，解决了输入和输出空间中的旋转和平移等方差，因此可以直接、灵活地对三维结构进行建模。

此外基于 VAE 的模型还有 Guo 等 [96] 提出的 DECO-VAE 模型。在该模型中，训练数据集中的 3D 结构首先表示为二维接触图，而后经由图神经网络提取节点和边特征输入编码器，解码器的输出以既定的方法还原为蛋白质 3D 结构。Harteveld 等 [97] 提出的 GENESIS 模型通过优化蛋白质拓扑晶格模型在距离和角度特征图中的 2D 表示来去噪蛋白质拓扑晶格模型草图。GENESIS 结合 trRosetta [80] 设计框架，为不同的蛋白质折叠生成了大量的不同序列。

3.2　扩散模型

现有的蛋白质 3D 结构生成方法仅限于在高度约束的环境中生成蛋白的拓扑结构 [94]。去噪扩散概率模型（denoising diffusion probabilistic models，DDPM）是一类从复杂数据分布中采样的生成模型。DDPM 定义了一个正向扩散过程，将数据扰动为噪声，学习反向过程中每一步的噪声为何，再逐步从数据分布中将随机高斯噪声去噪最终产生样本。近年来 DDPM 已被训练用来重建不同形式的被噪声破坏的数据（例如图像或文本）。DDPM 应用于蛋白设计领域则是将加噪后的蛋白质结构多步迭代后还原为真实结构用以训练；使用训练好的模型对输入随机的高斯噪声逐步“去噪”来生成折叠性质完好的蛋白结构，实现蛋白设计或结构生成。

DDPM 模型 [98-99] 输入的随机性使得去噪轨迹和输出的结构具备高度多样性，模型不需要起始的三维拓扑结构信息，但可以通过提供额外初始结构信息或施加外部约束条件，引导结构生成过程中每个步骤的迭代，直至特定的设计目标（图 13）。

Trippe 等 [101] 开发了 ProtDiff（一种蛋白骨架扩散概率模型）以及 SMCDiff（一种以模体为条件的骨架生成方法）。ProtDiff 模型采用分子 E(3) 等变扩散模型用于蛋白质结构生成。SMCDiff 是一种基于顺序蒙特卡洛的模体-骨架问题解决模型，将无条件训练的扩散概率模型用于条件采样。模体-骨架生成整体框架包含两个步骤，首先训练 ProtDiff 来学习蛋白质骨架上的分布，然后使用 SMCDiff 和 ProtDiff 来修补给定模体。评估结果表明，该框架能够生成多样化的超过 20 个氨基酸骨架结构，计算时间缩短至数分钟甚至更短。2022 年 Wu 等 [102] 提出了 FoldingDiff，一种使用 Transformer 作为主要架构训练的去噪扩散概率模型（图 14）。对于蛋白质

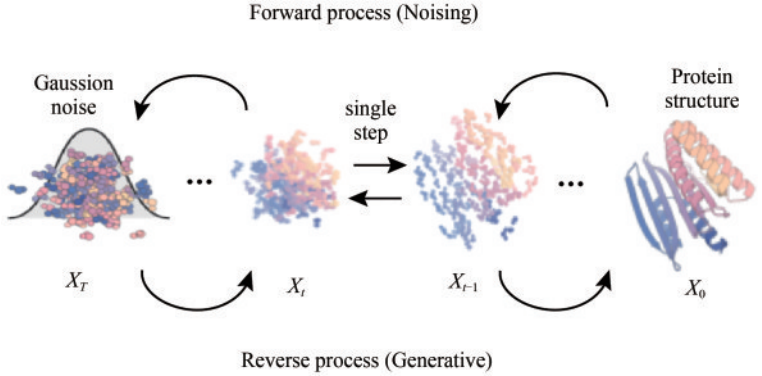


图 13　蛋白质结构生成扩散模型的原理示意图 [100]

Fig. 13　Schematic diagram of the diffusion model for protein structure generation[100]

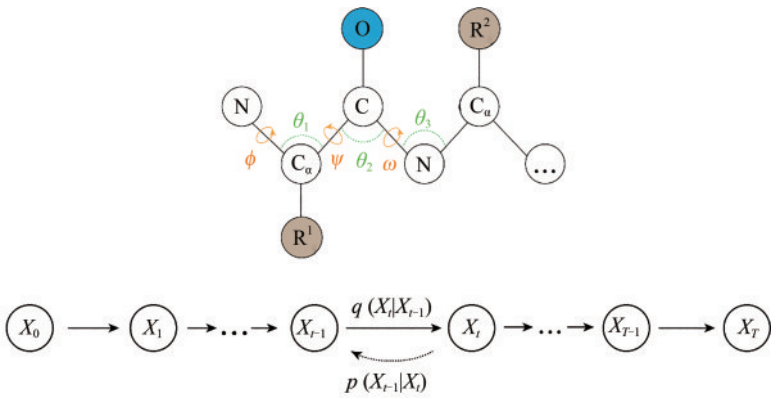


图14 FoldingDiff训练流程

Fig. 14 Training flow of the FoldingDiff model

的3D结构，研究者们使用氨基酸残基间的角度（ψ、ω、φ、θ_1、θ_2、θ_3）来表示，其中3个角为二面角，另外3个角为键角。训练天然蛋白骨架X_0开始，通过正向过程向其中迭代添加高斯噪声，直到X_t时刻角度无法辨识。反向过程中，研究者们采用了一个双向的Transformer架构，在正向过程中得到的实例上学习反向去噪过程。经过训练得到的扩散模型可以生成高质量的、多样化的、在生物学上合理的蛋白质结构。生成的结构可带有手性，同时表现出高度的可设计性。

除了上述的仅能生成蛋白主链骨架的模型外，DDPM模型还能够联合生成蛋白质的结构和序列，完成蛋白质的从头设计任务。

ProteinSGM[103]模型可以从头产生真实的蛋白质，并且可以将输入的蛋白骨架和功能位点修复为预定义长度的完整蛋白结构。ProteinSGM将两个残基之间的6D坐标特征作为输入特征，将其转化为2D的蛋白质残基接触矩阵（图15）。扩散模型在2D接触矩阵上逐渐添加噪声并迭代进行学习正向扩散的进程，训练完成的模型再对噪声反向逐步去噪，从噪声中生成真实的残基接触矩阵样本，后转化为蛋白质6D坐标。使用模型的输出残基约束指导Rosetta Design[104]和Relax生成与6D坐标约束相对应的蛋白质结构。因为连续时间扩散模型的采样需要大量正向传播的得分网络来求解反向梯度，而RosettaDesign依赖于昂贵的蒙特卡洛算法来遍历结构势能面找到局部最小值对应的低能量结构，因此模型在高通量设计任务中选择外接结构预测算法（如AlphaFold2等）来减小计算量。

Ingraham等[105]提出的Chroma模型，能够直接对新的蛋白质结构和序列进行采样，并调节生成过程，使其达到所需的特性和功能，同时实现完整蛋白复合物的3D结构和序列的联合建模且计

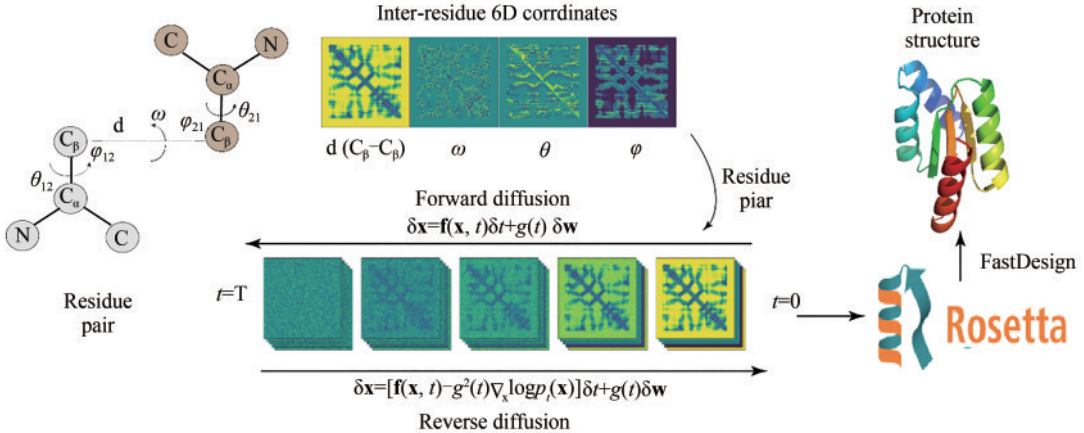


图15 ProteinSGM蛋白编码和模型架构图

Fig. 15 Protein structure encoding and model architecture of ProteinSGM

算效率十分可观。模型可以在不同线索下实现条件采样，而无需重新训练。Chroma 实现了一种可编程蛋白质设计的新模式，这种模式为生成特定和量身定制的蛋白质提供了可行性。

Anand [106] 模型通过定义二级结构和残基接触矩阵约束嵌入到高维空间，再使用 IPA 模块降维到三维空间中表征蛋白结构。作者使用 AlphaFold 网络架构 [38] 中的不变点注意力（invariant point attention，IPA）模块替换 Transformer 中的标准注意力模块保证模型的平移旋转不变性，使用类似于 BERT [107] 的扩散方法在骨架上生成序列。与其他 DDPM 模型不同，该模型不使用随机产生的高斯噪声，而是通过随机掩盖部分残基，在［0，1］中作为 t 的函数进行线性插值来训练模型；在生成时，模型在 $t = T$ 时掩蔽所有的残基来进行反向过程，从 $t = T$ 到 $t = 0$ 的时间步进行迭代采样。模型还允许人为给定条件信息编码蛋白结构。该模型完全从真实蛋白结构数据中学习，并生成蛋白质拓扑结构的条件约束，以产生全原子骨架构型以及序列和侧链预测。作者用了 3 个独立训练的模型分别生成蛋白结构、序列和旋转异构体，并将模型应用于无序列从头生成、蛋白补全、序列设计、侧链旋转异构体重排等任务中，结果表明其具有作为端到端的蛋白质从头设计工具的潜力。

Baker 组 [100] 随后推出基于 RoseTTAFold（RF）的扩散模型 RFdiffusion。将扩散模型建模为预训练后微调的 RoseTTAFold 模型（图 16）。在使用 RoseTTAFold 进行经典结构预测时，模型的结构输入来自同源模板结构，每个模板结构都有相关的每个残基的“置信度”值。在 RFdiffusion 中，结构输入来自于部分（去）噪声的结构，置信度特征被重新参数化以表示当前的去噪时间步，模型在该时间步的条件上进行结构预测，然后计算当前输入结构到预测的最终结构的噪声插值，生成去噪的结构并输入到下一个时间步。RFdiffusion 有着 RF 的序列信息通道，类似于前文中提到的 RFjoint，能够在扩散生成时逐渐地恢复被遮蔽的序列，通过输入部分遮蔽的序列和完整结构模板来预测未知位置的氨基酸分布，实现部分序列设计。为了生成用于训练或推断的加噪蛋白质结构，作者用 N-C_α-C 骨架对残基编码并进行正向扩散。对于平移，用 3D 高斯噪声对残基 C_α 坐标进行局部扰动；对于旋转，使用等变的 SO(3)-Transformer [108] 在旋转矩阵上模拟布朗运动生成噪声 [109]，使得模型具有全局的旋转不变性和高维的表征能力。在后续无条件约束策略设计和限制拓扑结构设计两种策略下，RFdiffusion 设计了包括蛋白质单体、蛋白质-肽复合物、对称寡聚体、酶和金属结合蛋白等多种类型的蛋白，证明了 RFdiffusion 在蛋白设计任务中的有效性和通用性。

2022 年刘海燕组 [110] 提出的 SCUBA-D，可以从包含不同类型或数量噪声的原始骨架中生成高质量的骨架。整个模型包含三个主要部分：一个低分辨率去噪模块，用于从初始骨架结构生成先验骨架结构；一个语言模型辅助的结构扩散模块，用于生成高分辨率的输出结构；一个判别器网络，用于辅助训练去噪扩散模块。在此框架中，初始结构可以是完全随机的也可以带有若干约束，低分辨率去噪模块经过训练可以处理不同类型的初始结构。对不同的初始结构，该模块的目标是生

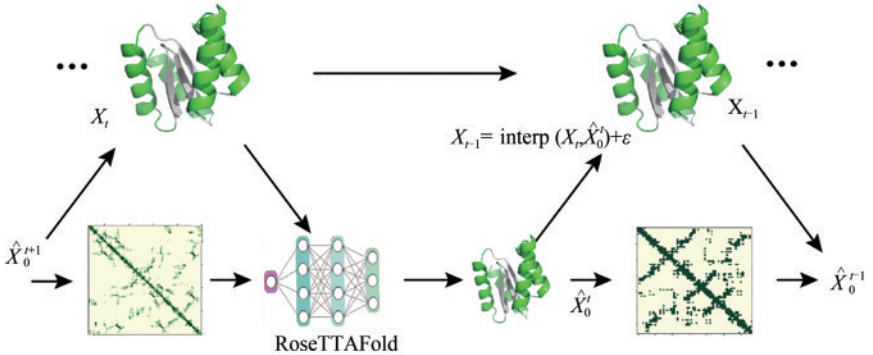


图 16　RFdiffusion 模型原理示意图

Fig. 16　Schematic diagram of the RFdiffusion model

成一个经过优化的粗糙的骨架结构，并保留所有初始结构中包含的拓扑信息。而后语言模型辅助的结构扩散模块获取低分辨率去噪模块的输出先验骨架结构，使用一系列去噪步骤对其进行细化，最终得到高分辨率的输出结构，其中使用氨基酸序列语言模型（ESM-1b模型[111]）辅助结构扩散过程。为了保证生成结构的高物理可信度，在架构中还使用了两个GAN风格的判别器，在训练中提供额外的损失。而后研究者将结构预测用于在生成骨架上设计的序列，来评估模型生成骨架的质量。结果表明，模型可以始终生成高质量的骨架结构，具有十分广阔的应用前景。

目前，扩散模型在抗体设计中的应用已有报道的工作。2022年Luo等[112]提出了DiffAb模型，该模型基于扩散概率模型以及等变神经网络对抗原抗体互补决定区（complementarity-determining regions）进行联合建模，可以生成针对特定抗原结构的抗体。研究者们同时对蛋白序列、坐标以及每个氨基酸的方向都进行了建模，使得模型可以实现原子级别分辨率的抗体设计且对旋转和平移等变。模型训练完成后，研究者将模型应用于序列结构协同设计、基于主链的抗体序列设计以及抗体优化任务中，结果表明模型在3个任务上均有出色的表现。

基于自注意力架构的蛋白质结构预测模型能够很好地捕获序列和结构之间的关系并高度准确地预测蛋白3D结构，但在生成能力上较弱；而基于序列空间反向传播迭代的蛋白幻想（hallucination）模型的性能高度依赖于输入的序列条件和生成标准。扩散模型使用的基于结构预测模型的3D噪声迭代方法，能够通过外部条件保留特定功能片段进行设计，也能在更广阔的序列和结构空间中进行探索，同时保证生成蛋白的合理性与多样性。

3.3 蛋白质序列生成

在蛋白质巨大的序列空间中，想要得到特定的序列以匹配到已知三维结构中执行特定的生物功能，无疑是一个巨大的挑战。近年来发展的人工智能方法不依赖于盲目搜索，而是基于推理的过程，直接从训练样本中学习序列与结构功能的关系，充分探索蛋白质序列空间，得到新颖的蛋白质序列。以下将简要介绍近年来发表的蛋白质序列的生成模型。

蛋白质序列生成模型的发展主要受到自然语言处理领域出色模型的启发。Repecka等[113]提出了一种基于生成对抗网络的蛋白质序列生成模型——ProteinGAN（图17）。ProteinGAN模型使用生成对抗网络架构，训练数据为苹果脱氢酶家族的16 706个蛋白序列。模型输入长为128的随机向量（均值为0，方差为0.5），由生成器生成蛋白质序列并将其呈递给判别器。在与自然蛋白质序列比较后，判别器对得到的序列进行打分，判断其为自然序列或是生成序列。生成器学习生成与自然序列近似的氨基酸序列用以欺骗判别器。经

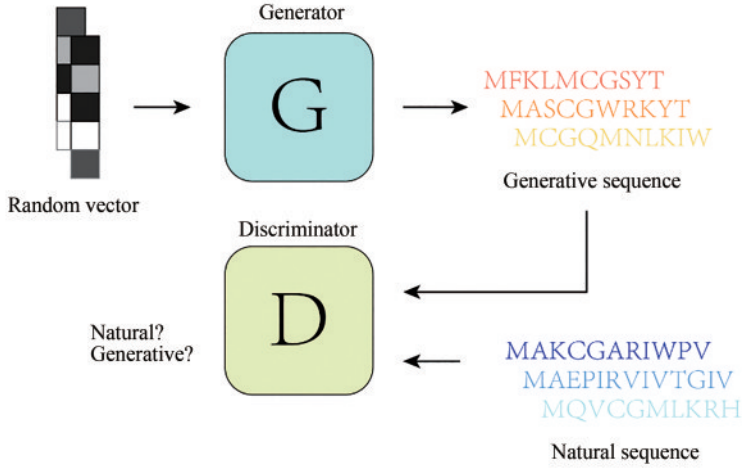


图17 ProteinGAN基本架构

Fig. 17 Architecture for ProteinGAN

过2.5M步训练之后，98%的生成序列包含苹果酸脱氢酶的全部主要结构域，同时序列聚类中的不同氨基酸序列之间相似度不超过10%，这表明模型已极大程度上探索了苹果酸脱氢酶家族的序列空间。

随着Transformer模型[60]在自然语言处理领域大放异彩，越来越多的研究者将Transformer架构应用到蛋白质序列生成领域，由此产生了许多基于Transformer的序列生成模型。2020年Madami等[114]提出了ProGen模型。ProGen是一种条件Transformer语言模型。该模型使用带有一系列蛋白性质标签的氨基酸序列进行训练，实现可控生成。ProGen生成的蛋白质在能量上与天然蛋白质相近，具有理想的生物功能。由Elnaggar等[115]提出的ProtTrans模型，使用4种不同的语言模型（两种自回归语言模型Tranformer-XL、XLNet以及两种自编码模型Bert、Albert）在蛋白质数据集上进行预训练，从序列中学习提取有用的特征，而后引入下游监督任务，以实现单个残基和单个蛋白性质的预测。这些模型原则上具有序列生成能力。2021年Gligorijević等[116]提出了一种序列去噪自编码器，该模型与一个功能预测器相结合，可以从大量未标记的蛋白质数据中学习蛋白质序列的多样性，而功能预测器可对序列采样的方向进行指导。在测试阶段，研究者进一步探究了模型在设计带有金属结合位点的序列以及重新设计功能增强的角质酶的能力。

2022年Moffat等[117]提出了DARK架构，用于在不断迭代扩展的合成蛋白质序列上有效地训练生成模型，该模型使用了标准的Transformer解码器架构，可生成具有不同有序结构的新序列。随后，Ferruz等人提出了ProtGPT2模型[118]，该模型是一个自回归Transformer模型，拥有7.38亿参数。模型的训练在Uniref-50数据集上进行。训练完成后生成的序列显示出与自然序列相似的预测稳定性与动态特性，同时在进化上与当前的蛋白质序列空间相距甚远。Hesslow等[119]提出RITA模型是一个拥有12亿参数的自回归生成模型。该模型在UniRef-100数据集超过2.8亿个蛋白质序列上进行训练。研究者们探究了模型大小对自回归模型性能的影响，结果表明随着模型规模的增大，模型的表现有了显著的提升。而后Nijkamp等[120]提出的ProGen2自回归Transformer模型具有更大的规模，模型参数最多可达64亿，模型的训练在从基因组、宏基因组和免疫库数据库中提取的超过10亿种蛋白质的不同序列组成的数据集上进行。为了评估ProGen2生成序列的能力，研究者选择在以下三种情境对模型进行评估，即：预训练后一般序列的生成，微调后的可以折叠成特殊结构的序列生成，以及在抗体序列数据集上进行预训练后的抗体序列生成。结果表明，截至ProGen2模型的提出，ProGen2在生成合理序列方面的表现为当前最佳。

4 总结与展望

在过去的数年中，人工智能技术在蛋白质设计上取得了巨大的成功。先进的人工智能模型凭借其强大的特征提取、数据统计和函数拟合能力，从现有蛋白质结构和序列数据中学习基本的特征和相互作用关系，拟合出具有泛化能力的函数模型，以应用于各类蛋白设计任务中。部分深度学习蛋白设计模型设计的蛋白已经被实验验证具有所需的结构和功能。

深度学习模型的性能高度依赖于标注准确的多样性数据。蛋白结构数据库需要从昂贵的生物实验结果中收录蛋白质功能和性质相关的数据。通常，这些不断积累的数据需要加以筛选和整理后才能作为深度学习模型的训练集和测试集。为保证深度学习神经网络能够充分捕获输入蛋白质结构和序列中的一般性质和潜在的依赖关系，一个具备合理性和可及性的蛋白质特征表示方式颇为重要。从最简单直接的独热编码、二级结构类型和组成原子在三维空间中的位置坐标，到高维空间中的嵌入图，再到依据邻近氨基酸残基的环境表示方式，为同时兼顾关键部位的贡献和全局构象的完整表征，研究人员提出了多种蛋白质结构和序列的特征提取和编码表示方法。对特定的蛋白质设计任务，如何选择合适的蛋白序列结构表征方式和人工智能模型，是研究者面临的最主要问题。

目前，深度学习模型在蛋白质设计任务上的

普及和应用依然存在着诸多问题和挑战。

其一，和海量的蛋白序列相比，蛋白结构数据库中收录数据的规模远远不足。在数据缺乏的情况下，构思再精妙的模型也难以展现其高准确和强泛化能力。另外，在深度学习模型的训练数据中进行合理的数据增强或运用掩蔽策略进行训练也会使模型的性能有所提升。

其二，目前对于蛋白设计模型的性能评估大多为天然序列恢复率和预测结构与原结构之间的差异，然而这两个指标仅能够衡量设计序列或结构与原蛋白的全局相似程度，并不能很好地量化设计蛋白的物理化学性质。Dauparas 等[73] 在ProteinMPNN文章中也指出天然序列恢复率对结构分辨率敏感，并且与局部残基距离误差相关性不高（R_{pearson}约为0.5），并不是一个能够很好地评价蛋白序列预测模型性能的指标。单个关键残基预测的错误对整体天然序列恢复率影响不大，但对序列折叠能力是毁灭性的。未来的方向可能是引入更多的评价指标，局部指标包括二级结构恢复率、溶剂可及表面、设计序列中无序残基比例等[121]。设计结构的全局评估可以使用结构预测模型折叠的结构并计算与目标结构的差异；长时间分子动力学模拟能够衡量序列折叠后结构的稳定性、展现结合蛋白与靶点之间的相互作用构象。将深度学习方法与传统的基于能量函数的蛋白质设计方法联用或前后相接，将深度学习模型生成的大量候选序列或结构输入基于物理化学的能量函数模型中进行验证和筛选，挑选出最优序列进行实验验证。充分发挥深度学习模型的高通量序列生成能力和物理化学模型对于蛋白的可表达性、可溶性以及聚集效应等物理化学性质的把握能力。

其三，蛋白质生理功能的实现大多是一个动态的过程，并且酶的活性位点具有一定的柔性。目前蛋白设计模型着重于对单一蛋白质功能构象结构的模仿或满足，力求设计蛋白的可折叠性、可溶性和稳定性，然而在功能位点和结合界面缺乏足够的关注。因此设计蛋白质的结合和变构现象，依然是当前研究中的难点。

最后，绝大多数模型难以同时考虑设计蛋白的性质，如可表达性、可溶性、稳定性、免疫原性等，只是拟合了天然蛋白从结构到序列的映射关系。从头设计具有强活性但低免疫原性和毒性的蛋白质药物，并佐以大量的生物实验结果，是人工智能蛋白质设计方法展现自己广阔应用前景的最有力方式。

传统蛋白质设计方法中使用的人工推导的能量函数能够遍历势能面，指导着蛋白序列结构生成优化的方向，并且具备生物物理和生物化学上的可解释性。深度神经网络学习到的能量函数比传统的更精确，但其神经网络模型内部的特征表示和数据传输可能缺乏一定的可解释性。希望未来的探索能够逐步打开深度学习模型内部的“黑盒子”，在模型输出结果的精确性和计算过程的可解释上有所改善。

近年在深度学习的赋能下，蛋白质设计的成功率和合理性得到了大幅提高。未来人工智能技术将更多地应用于抗体、酶、多肽药物等各类功能蛋白的设计中。可以预见的是，按需设计功能蛋白质的时代即将到来。

参 考 文 献

[1] HUANG P S, BOYKEN S E, BAKER D. The coming of age of *de novo* protein design[J]. Nature, 2016, 537(7620): 320-327.

[2] KHERSONSKY O, LIPSH R, AVIZEMER Z, et al. Automated design of efficient and functionally diverse enzyme repertoires[J]. Molecular Cell, 2018, 72(1): 178-186.e5.

[3] GLASGOW A A, HUANG Y M, MANDELL D J, et al. Computational design of a modular protein sense-response system[J]. Science, 2019, 366(6468): 1024-1028.

[4] ANFINSEN C B. Principles that govern the folding of protein chains[J]. Science, 1973, 181(4096): 223-230.

[5] LEAVER-FAY A, O'MEARA M J, TYKA M, et al. Scientific benchmarks for guiding macromolecular energy function improvement[J]. Methods in Enzymology, 2013, 523: 109-143.

[6] LEMAN J K, WEITZNER B D, LEWIS S M, et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks[J]. Nature Methods, 2020, 17(7): 665-680.

[7] NADRA A D, SERRANO L, ALIBÉS A. Chapter one-DNA-binding specificity prediction with FoldX[M]//Methods in enzymology. New York: Academic Press. 2011, 498: 3-18.

[8] HUANG X Q, PEARCE R, ZHANG Y. EvoEF2: accurate and fast energy function for computational protein design[J]. Bioinformatics, 2020, 36(4): 1135-1142.

[9] ALFORD R F, LEAVER-FAY A, JELIAZKOV J R, et al. The Rosetta all-atom energy function for macromolecular modeling and design[J]. Journal of Chemical Theory and Computation, 2017, 13(6): 3031-3048.

[10] KUHLMAN B, DANTAS G, IRETON G C, et al. Design of a novel globular protein fold with atomic-level accuracy[J]. Science, 2003, 302(5649): 1364-1368.

[11] SIEGEL J B, ZANGHELLINI A, LOVICK H M, et al. Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction[J]. Science, 2010, 329(5989): 309-313.

[12] SILVA D A, YU S, ULGE U Y, et al. *De novo* design of potent and selective mimics of IL-2 and IL-15[J]. Nature, 2019, 565 (7738): 186-191.

[13] MOHAN K, UEDA G, KIM A R, et al. Topological control of cytokine receptor signaling induces differential effects in hematopoiesis[J]. Science, 2019, 364(6442): eaav7532.

[14] CHEVALIER A, SILVA D A, ROCKLIN G J, et al. Massively parallel *de novo* protein design for targeted therapeutics[J]. Nature, 2017, 550(7674): 74-79.

[15] CAO L X, GORESHNIK I, COVENTRY B, et al. *De novo* design of picomolar SARS-CoV-2 miniprotein inhibitors[J]. Science, 2020, 370(6515): 426-431.

[16] LANGAN R A, BOYKEN S E, NG A H, et al. *De novo* design of bioactive protein switches[J]. Nature, 2019, 572(7768): 205-210.

[17] DAWSON W M, LANG E J M, RHYS G G, et al. Structural resolution of switchable states of a *de novo* peptide assembly[J]. Nature Communications, 2021, 12: 1530.

[18] SHEN H, FALLAS J A, LYNCH E, et al. *De novo* design of self-assembling helical protein filaments[J]. Science, 2018, 362 (6415): 705-709.

[19] HSIA Y, BALE J B, GONEN S, et al. Design of a hyperstable 60-subunit protein icosahedron[J]. Nature, 2016, 535(7610): 136-139.

[20] ROCKLIN G J, CHIDYAUSIKU T M, GORESHNIK I, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing[J]. Science, 2017, 357(6347): 168-175.

[21] BERMAN H M, WESTBROOK J, FENG Z K, et al. The protein data bank[J]. Nucleic Acids Research, 2000, 28(1): 235-242.

[22] FOX N K, BRENNER S E, CHANDONIA J M. SCOPe: structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures[J]. Nucleic Acids Research, 2014, 42(D1): D304-D309.

[23] CONSORTIUM T U, BATEMAN A, MARTIN M J, et al. UniProt: the universal protein knowledgebase[J]. Nucleic Acids Research, 2017, 45(D1): D158-D169.

[24] MISTRY J, CHUGURANSKY S, WILLIAMS L, et al. Pfam: the protein families database in 2021[J]. Nucleic Acids Research, 2021, 49(D1): D412-D419.

[25] FRAPPIER V, KEATING A E. Data-driven computational protein design[J]. Current Opinion in Structural Biology, 2021, 69: 63-69.

[26] KWON Y, SHIN W H, KO J, et al. AK-score: accurate protein-ligand binding affinity prediction using an ensemble of 3D-convolutional neural networks[J]. International Journal of Molecular Sciences, 2020, 21(22): 8424.

[27] JIANG D J, HSIEH C Y, WU Z X, et al. InteractionGraphNet: a novel and efficient deep graph representation learning framework for accurate protein-ligand interaction predictions[J]. Journal of Medicinal Chemistry, 2021, 64(24): 18209-18232.

[28] JONES D, KIM H, ZHANG X H, et al. Improved protein-ligand binding affinity prediction with structure-based deep fusion inference[J]. Journal of Chemical Information and Modeling, 2021, 61(4): 1583-1592.

[29] JIMÉNEZ J, ŠKALIČ M, MARTÍNEZ-ROSELL G, et al. K_{DEEP}: protein-ligand absolute binding affinity prediction *via* 3D-convolutional neural networks[J]. Journal of Chemical Information and Modeling, 2018, 58(2): 287-296.

[30] SLEDZIESKI S, SINGH R, COWEN L, et al. D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions[J]. Cell Systems, 2021, 12(10): 969-982.e6.

[31] BARANWAL M, MAGNER A, SALDINGER J, et al. Struct2Graph: a graph attention network for structure based predictions of protein-protein interactions[J]. BMC Bioinformatics, 2022, 23(1): 370.

[32] WANG S, CHEN W Q, HAN P F, et al. RGN: residue-based graph attention and convolutional network for protein-protein interaction site prediction[J]. Journal of Chemical Information and Modeling, 2022, 62(23): 5961-5974.

[33] SHEN W X, ZENG X, ZHU F, et al. Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations[J]. Nature Machine Intelligence, 2021, 3(4): 334-343.

[34] BUTTON A, MERK D, HISS J A, et al. Automated *de novo* molecular design by hybrid machine intelligence and rule-driven chemical synthesis[J]. Nature Machine Intelligence, 2019, 1 (7): 307-315.

[35] DE CAO N, KIPF T. MolGAN: an implicit generative model

for small molecular graphs[EB/OL]. arXiv, 2018: 1805.11973 [2023-10-01]. https://arxiv.org/abs/1805.11973

[36] WINTER R, MONTANARI F, STEFFEN A, et al. Efficient multi-objective molecular optimization in a continuous latent space[J]. Chemical Science, 2019, 10(34): 8016-8024.

[37] DING W Z, NAKAI K T, GONG H P. Protein design *via* deep learning[J]. Briefings in Bioinformatics, 2022, 23(3): bbac102.

[38] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596 (7873): 583-589.

[39] BAEK M, DIMAIO F, ANISHCHENKO I, et al. Accurate prediction of protein structures and interactions using a three-track neural network[J]. Science, 2021, 373(6557): 871-876.

[40] DAHIYAT B I, MAYO S L. Protein design automation[J]. Protein Science, 1996, 5(5): 895-903.

[41] DAHIYAT B I, MAYO S L. *De novo* protein design: fully automated sequence selection[J]. Science, 1997, 278(5335): 82-87.

[42] LI Z X, YANG Y D, FARAGGI E, et al. Direct prediction of profiles of sequences compatible with a protein structure by neural networks with fragment-based local and energy-based nonlocal profiles[J]. Proteins: Structure, Function, and Bioinformatics, 2014, 82(10): 2565-2573.

[43] DAI L, YANG Y D, KIM H R, et al. Improving computational protein design by using structure-derived sequence profile[J]. Proteins: Structure, Function, and Bioinformatics, 2010, 78 (10): 2338-2348.

[44] YANG Y D, ZHOU Y Q. *Ab initio* folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions[J]. Protein Science, 2008, 17(7): 1212-1219.

[45] WANG J X, CAO H L, ZHANG J Z H, et al. Computational protein design with deep learning neural networks[J]. Scientific Reports, 2018, 8: 6349.

[46] O'CONNELL J, LI Z X, HANSON J, et al. SPIN2: predicting sequence profiles from protein structures using deep neural networks[J]. Proteins: Structure, Function, and Bioinformatics, 2018, 86(6): 629-633.

[47] CHEN S, SUN Z, LIN L H, et al. To improve protein sequence profile prediction through image captioning on pairwise residue distance map[J]. Journal of Chemical Information and Modeling, 2020, 60(1): 391-399.

[48] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[49] ZHANG Y, CHEN Y, WANG C R, et al. ProDCoNN: protein design using a convolutional neural network[J]. Proteins: Structure, Function, and Bioinformatics, 2020, 88(7): 819-829.

[50] ANAND N, EGUCHI R, MATHEWS I I, et al. Protein sequence design with a learned potential[J]. Nature Communications, 2022, 13: 746.

[51] QI Y F, ZHANG J Z H. DenseCPD: improving the accuracy of neural-network-based computational protein sequence design with DenseNet[J]. Journal of Chemical Information and Modeling, 2020, 60(3): 1245-1252.

[52] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). July 21-26, 2017, Honolulu, HI, USA. IEEE, 2017: 2261-2269.

[53] SHROFF R, COLE A W, DIAZ D J, et al. Discovery of novel gain-of-function mutations guided by structure-based deep learning[J]. ACS Synthetic Biology, 2020, 9(11): 2927-2935.

[54] LU H Y, DIAZ D J, CZARNECKI N J, et al. Machine learning-aided engineering of hydrolases for PET depolymerization[J]. Nature, 2022, 604(7907): 662-667.

[55] NORN C, WICKY B I M, JUERGENS D, et al. Protein sequence design by conformational landscape optimization[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(11): e2017228118.

[56] YANG J Y, ANISHCHENKO I, PARK H, et al. Improved protein structure prediction using predicted interresidue orientations[J]. Proceedings of the National Academy of Sciences of the United States of America, 2020, 117(3): 1496-1503.

[57] WANG X, FLANNERY S T, KIHARA D. Protein docking model evaluation by graph neural networks[J]. Frontiers in Molecular Biosciences, 2021, 8: 647915.

[58] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. arXiv, 2016: 1609.02907 [2023-01-10]. https://arxiv.org/abs/1609.02907

[59] INGRAHAM J, GARG V K, BARZILAY R, et al. Generative models for graph-based protein design[C/OL]// Advances in Neural Information Processing Systems 32 (NeurIPS 2019), December 2019, Vancouver, Canada, Neural Information Processing Systems Foundation, 2019[2023-01-10]. https://dspace.mit. edu/bitstream/handle/1721.1/129731/NeurIPS-2019-generative-models-for-graph-based-protein-design-Paper.pdf?sequence=2&isAllowed=y.

[60] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. December 4-9, 2017, Long Beach, California, USA. New York: ACM, 2017: 6000-6010.

[61] STROKACH A, BECERRA D, CORBI-VERGE C, et al. Fast

and flexible protein design using deep graph neural networks[J]. Cell Systems, 2020, 11(4): 402-411.e4.

[62] JING B, EISMANN S, SURIANA P, et al. Learning from protein structure with geometric vector perceptrons[EB/OL]. arXiv, 2020: 2009.01411[2023-01-10]. https://arxiv.org/abs/2009.01411.

[63] ORELLANA G A, CACERES-DELPIANO J, IBAÑEZ R, et al. Protein sequence sampling and prediction from structural data[EB/OL]. bioRxiv, 2021[2023-01-10] https://www.biorxiv.org/content/10.1101/2021.09.06.459171v3.

[64] LI A J, LU M, DESTA I, et al. Neural network-derived Potts models for structure-based protein design using backbone atomic coordinates and tertiary motifs[J]. Protein Science, 2023, 32(2): e4554.

[65] ZHENG F, ZHANG J, GRIGORYAN G. Tertiary structural propensities reveal fundamental sequence/structure relationships[J]. Structure, 2015, 23(5): 961-971.

[66] HSU C, VERKUIL R, LIU J, et al. Learning inverse folding from millions of predicted structures[C/OL]//Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, PMLR. 2022: 8946-8970[2023-01-10]. https://www.biorxiv.org/content/10.1101/2022.04.10.487779v2.

[67] MCPARTLON M, LAI B, XU J B. A deep SE(3)-equivariant model for learning inverse protein folding[EB/OL]. bioXiv, 202[2023-01-10]. https://www.biorxiv.org/content/10.1101/2022.04.15.488492v1.

[68] ZADEH A, CHEN M, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis[EB/OL]. arXiv, 2017: 1707.07250 [2023-01-10]. https://arxiv.org/abs/1707.07250.

[69] XIONG P, HU X H, HUANG B, et al. Increasing the efficiency and accuracy of the ABACUS protein sequence design method[J]. Bioinformatics, 2020, 36(1): 136-144.

[70] LIU Y F, ZHANG L, WANG W L, et al. Rotamer-free protein sequence design based on deep learning and self-consistency[J]. Nature Computational Science, 2022, 2(7): 451-462.

[71] XIONG P, WANG M, ZHOU X Q, et al. Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability[J]. Nature Communications, 2014, 5: 5330.

[72] RONEY J P, OVCHINNIKOV S. State-of-the-art estimation of protein model accuracy using AlphaFold[J]. Physical Review Letters, 2022, 129(23): 238101.

[73] DAUPARAS J, ANISHCHENKO I, BENNETT N, et al. Robust deep learning-based protein sequence design using ProteinMPNN[J]. Science, 2022, 378(6615): 49-56.

[74] HUANG B, FAN T W, WANG K Y, et al. Accurate and efficient protein sequence design through learning concise local environment of residues[J]. Bioinformatics, 2023: btad122.

[75] ZHENG Z, DENG Y, XUE D, et al. Structure-informed language models are protein designers[EB/OL]. arXiv, 2023: 2302.01649[2023-02-10]. https://arxiv.org/abs/2302.01649.

[76] INGRAHAM J, GARG V K, BARZILAY R, et al. Generative models for graph-based protein design[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems, 8-14 December 2019, Vancouver, Canada, Curran Associates Inc, 2019: 1417[2023-01-10]. https://proceedings.neurips.cc/paper/2019/hash/f3a4ff4839c56a5f460c88cce3666a2b-Abstract.html.

[77] GAO Z Y, TAN C, LI S Z. ProDesign: toward effective and efficient protein design[EB/OL]. arXiv, 2022[2023-01-10]. https://arxiv.org/abs/2209.12643v1.

[78] TAN C, GAO Z Y, XIA J, et al. Generative *de novo* protein design with global context[EB/OL]. arXiv, 2022[2023-01-10]. https://arxiv.org/abs/2204.10673.

[79] GAO Z Y, TAN C, LI S Z. AlphaDesign: a graph protein design method and benchmark on AlphaFoldDB[EB/OL]. arXiv, 2022[2023-01-10]. https://arxiv.org/abs/2202.01079v2.

[80] ANISHCHENKO I, PELLOCK S J, CHIDYAUSIKU T M, et al. *De novo* protein design by deep network hallucination[J]. Nature, 2021, 600(7889): 547-552.

[81] TISCHER D, LISANZA S, WANG J, et al. Design of proteins presenting discontinuous functional sites using deep learning [EB/OL]. bioXiv, 2020[2023-01-10]. https://www.biorxiv.org/content/10.1101/2020.11.29.402743v1.

[82] WANG J, LISANZA S, JUERGENS D, et al. Scaffolding protein functional sites using deep learning[J]. Science, 2022, 377 (6604): 387-394.

[83] ZHANG S H, XU Y J, PEI J F, et al. AutoFoldFinder: an automated adaptive optimization toolkit for *de novo* protein fold design[EB/OL]. 2021[2023-01-10]. https://www.mlsb.io/papers_2021/MLSB2021_AutoFoldFinder.pdf.

[84] YEH A H W, NORN C, KIPNIS Y, et al. *De novo* design of luciferases using deep learning[J]. Nature, 2023, 614(7949): 774-780.

[85] DOU J Y, VOROBIEVA A A, SHEFFLER W, et al. *De novo* design of a fluorescence-activating β-barrel[J]. Nature, 2018, 561(7724): 485-491.

[86] CAO L X, COVENTRY B, GORESHNIK I, et al. Design of protein-binding proteins from the target structure alone[J]. Nature, 2022, 605(7910): 551-560.

[87] HUANG B, XU Y, HU X H, et al. A backbone-centred energy function of neural networks for protein design[J]. Nature, 2022, 602(7897): 523-528.

[88] LIANG S D, LI Z X, ZHAN J, et al. *De novo* protein design by an energy function based on series expansion in distance and orientation dependence[J]. Bioinformatics, 2021, 38(1): 86-93.

[89] LIANG S D, ZHENG D D, ZHANG C, et al. Fast and accurate prediction of protein side-chain conformations[J]. Bioinformatics, 2011, 27(20): 2913-2914.

[90] LIANG S D, ZHOU Y Q, GRISHIN N, et al. Protein side chain modeling with orientation-dependent atomic force fields derived by series expansions[J]. Journal of Computational Chemistry, 2011, 32(8): 1680-1686.

[91] LIANG S D, ZHANG C, ZHOU Y Q. LEAP: highly accurate prediction of protein loop conformations by integrating coarse-grained sampling and optimized energy scores with all-atom refinement of backbone and side chains[J]. Journal of Computational Chemistry, 2014, 35(4): 335-341.

[92] ANAND N, HUANG P S. Generative modeling for protein structures[C/OL]// 6th International Conference on Learning Representations, Vancouver, BC, Canada, April 30-May 3, 2018[2023-01-10]. https://openreview.net/forum?id=HJFXnYJvG.

[93] ANAND N, EGUCHI R, HUANG P S. Fully differentiable full-atom protein backbone generation[EB/OL]. ICLR 2019 Workshop on Deep Generative Models for Highly Structured Data, 2019[2023-01-10]. https://openreview. net/group? id= ICLR.cc/2019/Workshop/DeepGenStruct.

[94] EGUCHI R R, CHOE C A, HUANG P S. Ig-VAE: generative modeling of protein structure by direct 3D coordinate generation[J]. PLoS Computational Biology, 2022, 18(6): e1010271.

[95] LAI B Q, MCPARTLON M, XU J B. End-to-End deep structure generative model for protein design[EB/OL]. bioRxiv, 2022[2023-01-10]. https://www. biorxiv. org/content/10.1101/2022.07.09.499440v1.

[96] GUO X J, DU Y Q, TADEPALLI S, et al. Generating tertiary protein structures *via* interpretable graph variational autoencoders[J]. Bioinformatics Advances, 2021, 1(1): vbab036.

[97] HARTEVELD Z, SOUTHERN J, LOUKAS A, et al. Deep sharpening of topological features for *de novo* protein design[EB/OL]. ICLR 2022 Machine Learning for Drug Discovery, 2022 [2023-01-10]. https://openreview.net/forum?id=DwN81YIXGQP.

[98] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[EB/OL]. arXiv, 2020: 2006.11239. https://arxiv.org/abs/2006.11239.

[99] SOHL-DICKSTEIN J, WEISS E A, MAHESWARANATHAN N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37. July 6-11, 2015, Lille, France. New York: ACM, 2015: 2256-2265.

[100] WATSON J L, JUERGENS D, BENNETT N R, et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models[EB/OL]. bioXiv, 2022[2023-01-10]. https://www.biorxiv.org/content/10.1101/2022.12.09.519842v2.

[101] TRIPPE B L, YIM J, TISCHER D, et al. Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem[EB/OL]. arXiv, 2022: 2206.04119[2023-01-10]. https://arxiv.org/abs/2206.04119.

[102] WU K E, YANG K K, BERG R V D, et al. Protein structure generation *via* folding diffusion[EB/OL]. arXiv, 2022: 2209.15611 [2023-01-10]. https://arxiv.org/abs/2209.15611.

[103] LEE J S, KIM P. ProteinSGM: score-based generative modeling for *de novo* protein design[EB/OL]. 2022[2023-01-10]. https://doi.org/10.21203/rs.3.rs-1855828/v1.

[104] LEAVER-FAY A, TYKA M, LEWIS S M, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules[J]. Methods in Enzymology, 2011, 487: 545-574.

[105] INGRAHAM J, BARANOV M, COSTELLO Z, et al. Illuminating protein space with a programmable generative model [EB/OL]. bioXiv, 2022[2023-01-10]. https://www. biorxiv. org/content/10.1101/2022.12.01.518682v1.

[106] ANAND N, ACHIM T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models [EB/OL]. arXiv, 2022: 2205.15019[2023-01-10]. https://arxiv.org/abs/2205.15019.

[107] DEVLIN J, CHANG M, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. arXiv, 2018: 1810.04805[2023-01-10]. https://arxiv.org/abs/1810.04805.

[108] DE BORTOLI V, MATHIEU E, HUTCHINSON M, et al. Riemannian score-based generative modelling[EB/OL]. arXiv, 2022: 2202.02763[2023-01-10]. https://arxiv.org/abs/2202.02763.

[109] LEACH A, SCHMON S M, DEGIACOMI M T, et al. Denoising diffusion probabilistic models on SO(3) for rotational alignment[EB/OL]. ICLR 2022 Workshop on Geometrical and Topological Representation Learning, 2022[2023-01-10]. https://openreview.net/forum?id=BY88eBbkpe5.

[110] LIU Y F, CHEN L H, LIU H Y. *De novo* protein backbone generation based on diffusion with structured priors and adversarial training[EB/OL]. bioRxiv, 2022[2023-01-10]. https://www.biorxiv.org/content/10.1101/2022.12.17.520847v1.

[111] RIVES A, MEIER J, SERCU T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 mil-

lion protein sequences[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(15): e2016239118.

[112] LUO S T, SU Y F, PENG X G, et al. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures[EB/OL]. bioXiv, 2022[2023-01-10]. https://www.biorxiv.org/content/10.1101/2022.07.10.499510v5.

[113] REPECKA D, JAUNISKIS V, KARPUS L, et al. Expanding functional protein sequence spaces using generative adversarial networks[J]. Nature Machine Intelligence, 2021, 3(4): 324-333.

[114] MADANI A, MCCANN B, NAIK N, et al. ProGen: Language modeling for protein generation[EB/OL]. arXiv, 2020: 2004.03497[2023-01-10]. https://arxiv.org/abs/2004.03497.

[115] ELNAGGAR A, HEINZINGER M, DALLAGO C, et al. ProtTrans: towards cracking the language of lifes code through self-supervised deep learning and high performance computing [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(10), 7112-7127.

[116] GLIGORIJEVIĆ V, BERENBERG D, RA S, et al. Function-guided protein design by deep manifold sampling[EB/OL]. bioRxiv, 2021[2023-01-10]. https://www.biorxiv.org/content/10.1101/2021.12.22.473759v1.

[117] MOFFAT L, KANDATHIL S M, JONES D T. Design in the DARK: learning deep generative models for *de novo* protein design[EB/OL]. bioRxiv, 2022[2023-01-10]. https://www.biorxiv.org/content/10.1101/2022.01.27.478087v1.

[118] FERRUZ N, SCHMIDT S, HÖCKER B. ProtGPT2 is a deep unsurprised language model for protein design[J]. Nature Communications, 2022,13(1): 4348.

[119] HESSLOW D, ZANICHELLI N, NOTIN P, et al. RITA: a study on scaling up generative protein sequence models[EB/OL]. arXiv, 2022: 2205.05789[2023-01-10]. https://arxiv. org/abs/2205.05789.

[120] NIJKAMP E, RUFFOLO J, WEINSTEIN E N, et al. ProGen2: exploring the boundaries of protein language models[EB/OL]. arXiv, 2022[2023-01-10]. https://arxiv.org/abs/2206.13517.

[121] LI Z X, YANG Y D, ZHAN J, et al. Energy functions in *de novo* protein design: current challenges and future prospects[J]. Annual Review of Biophysics, 2013, 42: 315-335.

通讯作者：戚逸飞（1983—），男，副研究员，硕士生导师。研究方向为生物大分子结构和功能模拟以及人工智能药物设计。

E-mail：yfqi@fudan.edu.cn

第一作者：陈志航（1998—），男，硕士研究生。研究方向为人工智能蛋白质设计。

E-mail：zhihangchen21@m.fudan.edu.cn

第一作者：季梦麟（2000—），男，硕士研究生。研究方向为人工智能蛋白质设计。

E-mail：22211030067@m.fudan.edu.cn