http://dx.doi.org/10.52810/FAAI.2024.003

REVIEW

基于机器学习和深度学习的蛋白质结构预测研究进展

崔佳轩[1,*]

[1] 贵州大学，大数据与信息工程学院，贵州 550025

学术编辑： 金学波；**收稿日期：** 2024-02-09；**录用日期：** 2024-05-11；**发布日期：** 2024-05-20

***通讯作者：** ✉崔佳轩，1363049441@qq.com

文章引用

崔佳轩. 基于机器学习和深度学习的蛋白质结构预测研究进展. 人工智能前沿与应用, 2024, **1**(1): 32–44.

Citation

Cui, J. (2024). Research Progress in Protein Structure Prediction Using Machine Learning and Deep Learning. Frontiers and Applications of Artificial Intelligence, 1(1), 32–44.

摘　要

蛋白质结构预测是生物信息学领域的一个核心问题，对于理解蛋白质功能、药物设计以及疾病研究具有重要意义。传统的蛋白质结构预测方法受限于计算复杂度和预测精度。近年来，随着机器学习和深度学习技术的快速发展，这些先进的方法被广泛应用于蛋白质结构预测中，显著提高了预测的准确性和效率。本文首先介绍了蛋白质结构预测的背景和重要性，然后详细阐述了机器学习和深度学习在蛋白质结构预测中的应用，包括常用的算法、模型架构以及优化策略。最后，本文展望了基于机器学习和深度学习的蛋白质结构预测在未来的发展方向和潜在挑战，为相关领域的研究者提供了有价值的参考。

关键词: 蛋白质结构预测；深度学习；机器学习；卷积神经网络；Transformer 模型；生成式对抗网络

Research Progress in Protein Structure Prediction Using Machine Learning and Deep Learning

Jiaxuan Cui[1,*]

[1]School of Big Data and Information Engineering, Guizhou University, Guizhou 550025, China

Academic Editor: Xuebo Jin; **Submitted:** 2024-02-09; **Accepted:** 2024-05-11; **Published:** 2024-05-20

***Correspondence Author:** ✉Jiaxuan Cui, 1363049441@qq.com

Abstract

Protein structure prediction is a core problem in the field of bioinformatics, which is significant for understanding protein functions, drug design, and disease research. Traditional protein structure prediction methods are limited by computational complexity and prediction accuracy. In recent years, with the rapid development of machine learning and deep learning techniques, these advanced methods have been widely used in protein structure prediction, significantly improving prediction accuracy and efficiency. This article first introduces the background and importance of protein structure prediction, and then elaborates on the application of machine learning and deep learning in protein structure prediction, including commonly used algorithms, model architectures, and optimization strategies. Finally, this article looks ahead to the future development directions and potential challenges of protein structure prediction based on machine learning and deep learning, providing valuable references for researchers in related fields.

Keywords: Computer vision, feature tracking, optical flow method, visual features, visual tracking

1 序言

在生物学领域，蛋白质作为生命活动的主要承担者，扮演着至关重要的角色。其三维结构对于理解蛋白质的生物功能、蛋白质与蛋白质之间以及蛋白质与小分子之间的相互作用具有决定性意义。然而，通过实验手段测定蛋白质的三维结构既耗时又昂贵。因此，开发高效的计算方法来预测蛋白质结构成为了生物信息学和计算生物学领域的迫切需求。近年来，随着深度学习技术的异军突起，基于深度学习的蛋白质结构预测方法取得了令人瞩目的进展。

深度学习，作为一种强大的机器学习技术，通过模拟人脑神经网络的运作原理，能够自动从大规模数据中提取有用的特征，进而实现复杂的模式识别和预测任务。在蛋白质结构预测中，深度学习模型能够从蛋白质的氨基酸序列出发，自动捕捉序列中隐含的高级结构和功能信息，为蛋白质的三维结构预测提供了全新的视角和解决方案。

传统的蛋白质结构预测方法往往受限于物理、化学和统计学的原理，通过模拟蛋白质折叠过程或搜索已知结构数据库来预测蛋白质的三维结构。然而，这些方法在计算资源和模拟精度方面存在局限性，难以处理大规模的蛋白质序列，并且在处理复杂蛋白质结构时表现不佳。相比之下，基于深度学习的蛋白质结构预测方法能够处理更大规模的蛋白质序列，并且通过学习数据中的复杂模式，能够更准确地预测蛋白质的三维结构。

近年来，基于深度学习的蛋白质结构预测算法如 AlphaFold、RoseTTAFold 等在国际蛋白质结构预测竞赛（如 CASP）[1] 中屡获佳绩，充分展示了深度学习在蛋白质结构预测领域的巨大潜力。本综述旨在系统梳理和分析近年来基于深度学习的蛋白质结构预测方法的研究进展，探讨不同深度学习模型在蛋白质结构预测中的应用和性能，并展望未来的发展方向。通过回顾相关文献和研究成果，我们期望为相关领域的研究者提供一个全面的视角，推动基于深度学习在工业领域 [23–31] 的进一步发展和应用。

2 基于传统机器学习预测模型

从机器学习 [32–43] 的角度来看，蛋白质结构预测中的接触预测问题可以类比为计算机视觉中的图像分割问题。在图像分割中，输入是一个具有特定维度的图像（H×W×Z），其中 H、W 和 Z 分别代表图像的高度、宽度和通道数，而输出则是一个二维矩阵（H×W），用于标记每个像素是否属于特定对象。类似地，在蛋白质接触预测中，输入是一个表征蛋白质特征的矩阵（L×L×N），其中 L 是蛋白质序列的长度，N 是通道数，而输出则是

一个二维矩阵（L×L），表示不同氨基酸残基之间接触的概率。

支持向量机（SVM）是一种广泛使用的分类器，它通过训练已知类别的蛋白质数据来构建分类模型。在蛋白质结构预测中，SVM 可以用于预测蛋白质的二级结构，如 α-螺旋、β-折叠等。输入特征可以包括氨基酸序列、物理化学性质等，而输出则是预测的二级结构类别。Wang[2] 提出一种基于并行多类支持向量机（SVM）的蛋白质结构预测方法。该方法在 SVM-pro-file 二类支持向量机的基础上，采用加权一对多的多类分类方法对蛋白质序列作出唯一的类别判断，提高了基于支持向量机的蛋白质结构预测算法的应用范围，同时利用主从二级模型对算法进行并行处理，降低了算法复杂性，提高了蛋白质结构预测的效率。

传统的人工神经网络（ANN）也在早期的研究中得到了应用 [44–55]。传统的人工神经网络（ANN）作为一种经典的机器学习模型，ANN 以其强大的自学习和非线性映射能力，为从蛋白质序列预测其结构提供了一种有效途径，在蛋白质结构预测领域曾发挥过重要作用。ANN 的特点主要体现在其通用性、自学习性和鲁棒性上。作为一种通用的函数逼近器，ANN 能够模拟复杂的非线性关系，从而学习从蛋白质的一维序列信息到其复杂三维结构的映射。同时，通过训练过程中的权重调整，ANN 能够自我学习并不断优化其预测性能。此外，ANN 还表现出对输入数据噪声和变化的鲁棒性，这在处理具有一定变异性的蛋白质序列时尤为重要。在结构上，ANN 是由神经元组成的网络结构，通常由输入层、隐藏层和输出层构成（如图 1 所示）。输入层负责接收经过编码的蛋白质序列特征，如氨基酸的理化性质、进化保守性等。隐藏层则通过捕捉这些特征之间的复杂关系，进一步提炼和抽象信息。最终，输出层根据具体的预测任务，给出蛋白质的二级结构、溶剂可及性或其他结构相关属性的预测结果。

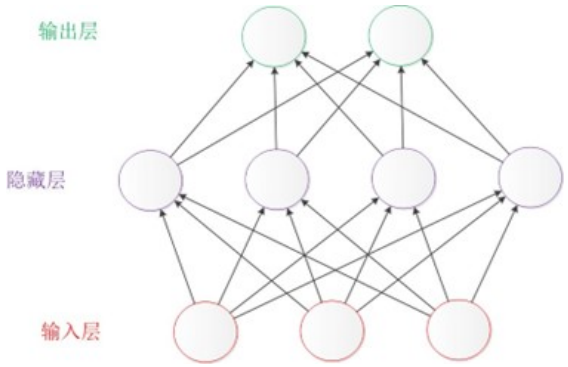

图 1. ANN 网络结构示意图

在蛋白质结构预测方面，ANN 曾被广泛应用于从序列出发预测蛋白质的二级结构。例如，通过训练 ANN 模型来识别 α-螺旋和 β-折叠等局部结构元素。此外，ANN 还被用于预测蛋白质的稳定性、折叠速率以及与其他分子的相互作用等。

人工神经网络（ANN）在蛋白质结构预测领域优于其他机器学习方法，凭借其卓越的表征学习能力和对复杂非线性关系的精准捕捉，使得它能够较好地从原始数据中自动提取关键特征，为预测提供有力支持。此外，ANN 的高度灵活性和可扩展性也为其在应对各种预测任务时提供了极大的便利。然而，这种强大的学习能力也伴随着一些挑战。首先，ANN 在数据量不足时容易陷入过拟合的困境，这可能会影响其在未知数据上的表现。其次，训练 ANN 需要大量的计算资源和时间，这对于一些资源有限的环境来说是一个不小的负担。最后，尽管 ANN 能够做出准确的预测，但其内部的决策逻辑往往难以直观解释，这在某些需要明确决策依据的场景中可能构成一定的障碍。因此，在选择使用 ANN 进行蛋白质结构预测时，需要全面考虑其优势和劣势，以做出明智的决策。

在蛋白质二级结构预测领域，误差反传前向网络（BP）、径向基函数网络（RBF）、广义回归神经网络（GRNN）、

串并联叠层网络（CF）以及 Elman 网络（ELM）是五种常用的神经网络模型。BP 神经网络 [3] 是一种单向传播的多层前馈神经网络。BP 学习算法是 Rumelhart 等在 1986 年提出的。它是具有三层或三层以上的神经网络，包括输入层、中间层（隐层）和输出层。上下层之间实现全连接，而每层神经元之间无连接。当学习样本提供给网络后，神经元的激活值从输入层经各中间层向输出层传播，在输出层的各神经元获得网络的输出响应。接下来，按照减少目标输出与实际误差的方向，从输出层经过各个中间层逐层修正各个连接权值，最后回到输入层，这种算法称为“误差反向传播算法”，即 BP 算法。随着这种误差的传播修正不断进行，网络对输入模式响应的正确率也不断上升。BP（Error Back Propagation）神经网络通过其多层前馈结构和误差反向传播算法，能够有效地学习蛋白质序列与其二级结构之间的复杂映射关系。通过将蛋白质序列编码为数值向量并输入到 BP 神经网络中，网络能够通过迭代训练不断调整其内部参数，从而准确预测新的蛋白质序列的二级结构。

GRNN[4] 是一种基于概率密度的回归神经网络。它不需要迭代训练过程，而是通过计算输入样本与训练样本之间的相似度，直接进行回归预测。GRNN 由四层构成：输入层、模式层、求和层和输出层。输入层接收样本数据，模式层计算输入与训练样本之间的相似度，求和层对模式层的输出进行加权求和，最终输出层给出预测结果。GRNN 在处理小数据集和非线性回归问题上表现出色。GRNN（广义回归神经网络）利用概率密度进行回归预测，它特别适合处理小数据集和非线性问题。在蛋白质结构预测中，GRNN 能够快速计算输入蛋白质序列与训练样本之间的相似度，并直接给出预测结果，这对于快速筛选和预测蛋白质结构具有重要意义。

RBF 神经网络 [5] 是一种三层前馈神经网络，具有强大的局部逼近能力。它的特点是使用径向基函数作为隐藏层的激活函数。RBF 网络包括输入层、隐藏层和输出层。输入层接收数据，隐藏层通过径向基函数对输入进行非线性变换，然后输出层对隐藏层的输出进行线性组合以产生最终结果。RBF 网络在模式识别、函数逼近等领域有广泛应用。RBF（径向基函数）神经网络以其强大的局部逼近能力和使用径向基函数作为激活函数的特点，在蛋白质结构预测中展现出优异的性能。它能够捕捉蛋白质序列中的局部特征，并通过线性组合产生最终的预测结果。

3 基于深度学习的预测模型

3.1 卷积神经网络（CNN）

随着深度学习技术的兴起，越来越多的研究者开始将深度学习应用于蛋白质结构预测中。深度学习模型能够从大量的数据中自动提取有用的特征，并学习复杂的非线性映射关系，从而更准确地预测蛋白质的三维结构。

卷积神经网络（CNN，Convolutional Neural Networks）是一种深度学习模型，特别适合处理图像相关问题。CNN（卷积神经网络）的结构主要包括数据输入层、卷积层、激活函数层、池化层和全连接层（如图 2 所示）。CNN 的特点在于其能够自动提取输入数据的特征，通过卷积层对局部区域进行感知并提取出有用的特征，然后通过池化层进行下采样以减少数据维度，最后通过全连接层进行分类或回归。CNN 的另一个显著特点是权值共享，这大大减少了网络参数的数量，使得模型更加高效且易于训练。CNN 在图像识别、目标检测、语音识别等领域有着广泛的应用，其强大的特征提取能力使得它在处理复杂模式识别问题时表现出色。与其他机器学习方法相比，如支持向量机（SVM）或决策树等，CNN 能够自动学习特征表示，而无需手动设计和选择特征，这使得它在处理原始数据时具有更大的灵活性和适应性。此外，CNN 通过逐层卷积和池化操作，能够捕捉到图像中的层次化特征，从而在处理具有复杂结构和纹理的图像时具有显著优势。

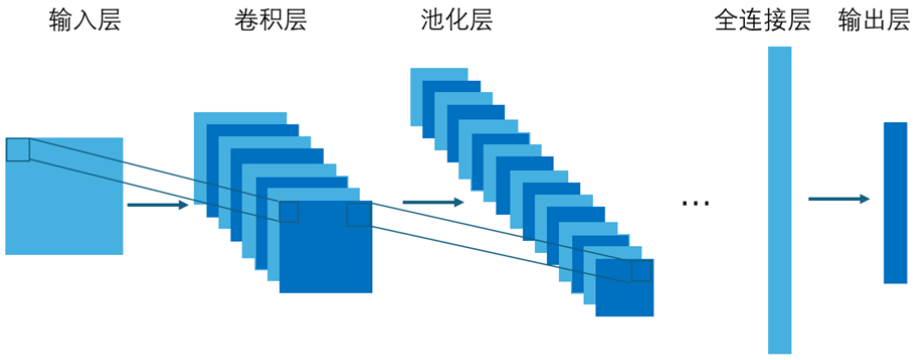

图 2. CNN 网络结构示意图

图 2 CNN 网络结构示意图 RaptorX-Contact[6] 是 Xu 等人开发的一种基于深度学习的接触图预测工具。该方法通过组合两个深度残差神经网络来构建一个超深度神经网络，将进化耦合和序列守恒信息结合起来进行接触预测。第一个残差网络对序列特征进行一系列一维卷积变换；第二个残差网络则对第一个网络的输出以及其他信息进行二维卷积变换。通过使用非常深的残差网络，RaptorX-Contact 能够更准确地模拟接触发生模式和复杂的序列结构关系，从而获得更高质量的接触预测结果。

Zhang 的团队开发了一种名为 ResPre[7] 的方法，该方法基于深度残差卷积神经网络来预测残差-残差接触。ResPre 在从 CASP 实验收集的蛋白质上进行了测试，并展现出了较高的预测准确率，优于其他先进的方法。ResPre 的主要优点是利用精度矩阵来减少接触图的噪声，并通过残差网络提高了深度学习模型的训练效果。

此外, Yang[8] 的团队开发了 MapPred 方法, 该方法在残差神经网络框架中使用了宏基因组序列数据。MapPred 由 DeepMSA 和 DeepMeta 两种方法组成，都是基于残差神经网络进行训练的。DeepMSA 利用协方差特征矩阵进行训练，并通过考虑接触映射的对称性来提高训练效率。而 DeepMeta 则结合预测的接触点和其他序列剖面特征进行工作。实验结果表明，宏基因组序列数据的贡献是显著的，MapPred 在接触图预测方面取得了优异的表现。

Fukuda[9] 等人开发了 DeepECA 方法，这是一种基于卷积神经网络（CNN）的进化耦合分析方法，用于直接从多序列比对（MSA）预测接触图。DeepECA 方法能够使用来自深层或浅层 MSA 的信息，并通过多任务模型同时预测二级结构和接触图。在基准数据集上的实验结果表明，DeepECA 在接触图预测方面具有一定的改进效果。

Zhang 的团队则开发了一种名为 TripletRes[10] 的方法，该方法同样用于预测蛋白质接触图，但基于深度残差神经网络的端到端训练。其独特之处在于能够从离散的距离剖面中推断蛋白质接触图，并直接融合从全基因组和宏基因组数据库中提取的三组共同进化矩阵。这种融合最大限度地减少了接触模型训练过程中的信息损失。在 CASP 11&12 和 CAMEO 实验的 245 个非同源蛋白上的测试中，TripletRes 在远程接触精度方面表现卓越，相较于 CASP12 中的其他顶级方法，其提升幅度至少达到了 58.4% 和 44.4%。值得一提的是，在 CASP13 实验中，TripletRes 在 top-L/5 远程接触预测方面的精度达到了最高的 71.6%。

Kihara 的团队还开发了一种名为 AttentiveDist[11] 的方法，专门用于蛋白质残基间距离的预测。该方法基于残差网络（ResNets）并融入了注意力机制，以确定与每个残差对最相关的 MSA。网络中包含 45 个残差块，前 5 个用于特征编码。通过使用具有不同 e 值的 4 个 MSA 生成不同的输入进行网络训练。在最后一个残差块结束时，模型会分支成 5 个不同的路径来预测 5 个不同的输出：Cβ 距离预测、每个氨基酸残基对的 3 个侧链取

向角和主二面角。AttentiveDist 的成功证明了在不同的基于 MSA 的特征上使用基于注意的方法与 MSA 中的共同进化信息之间的紧密关联。

AlphaFold[12] 是由 DeepMind 开发的一种革命性的蛋白质结构预测方法，在 CASP13 中脱颖而出。该方法的核心在于其图预测组件，该组件利用经过 PDB 结构训练的卷积神经网络（CNN）来预测任意残基对之间的 Cβ–Cβ 距离。结合查询序列的氨基酸表示和从多序列比对（MSA）生成的特征，CNN 网络能够预测每对残基的离散概率分布。这个分布与真实距离高度相似，为后续的蛋白质模型构建提供了坚实基础。然而，在实验条件下，它只能准确预测一小部分没有模板的蛋白质。尽管如此，学术界很快复制了 AlphaFold1 的性能并推动了其进一步发展。在蛋白质结构预测领域的应用展现出了显著的优势和劣势。其优势主要在于强大的特征提取能力，能够自动学习和识别蛋白质序列中的关键模式，这对于理解蛋白质的功能和预测其结构至关重要。同时，CNN 对高维数据的出色处理能力，使得它能够更好地捕捉蛋白质的三维结构信息，为科研人员提供了更深入的分析视角。然而，CNN 在蛋白质结构预测中也存在一些劣势。由于 CNN 的卷积操作本质上是局部的，它在处理长距离依赖关系时可能表现出不足。这意味着，当需要考虑蛋白质序列中远距离氨基酸之间的相互作用时，CNN 的性能可能会受到限制。此外，CNN 的感受野大小固定，可能限制了其对全局结构信息的捕捉能力，尤其是对于较大的蛋白质分子而言。

3.2 基于 Transformer 模型

Transformer 是一种基于自注意力机制的深度学习模型，具有强大的序列建模能力。Transformer 的结构由编码器和解码器两大部分构成 Transformer（如图 3 所示）。编码器通常由 6 个相同的层堆叠而成，每一层都包含多头自注意力机制和全连接前馈神经网络，用于将输入序列转换为高级别的特征表示。解码器也由类似的层堆叠而成，但额外插入了一个对编码器输出执行多头注意力的子层，用于根据编码器的输出和之前的输出来生成序列的下一个元素。此外，Transformer 还包含位置编码来提供序列中单词的位置信息，以及残差连接和层归一化来辅助模型训练和提高性能。输入序列首先通过词嵌入转换为向量表示，并加上位置编码，解码器的输出则通过线性层和 softmax 层产生下一个可能的单词或标记的概率分布。这种结构设计使 Transformer 能够并行处理输入序列中的所有元素，并有效捕获序列中的长期依赖关 Transformer 其特点在于完全依赖于自注意力来计算输入序列中各个位置之间的相关性，从而捕捉长距离依赖关系。

Transformer 通过多头自注意力机制，能够同时关注输入序列的不同部分，提取出丰富的上下文信息。此外，Transformer 采用残差连接和层归一化技术，有效地缓解了深度神经网络中的梯度消失和表示瓶颈问题。它的作用主要体现在自然语言处理任务中，如机器翻译、文本摘要和问答系统等，能够显著提高模型的性能。与其他机器学习方法相比，如循环神经网络（RNN）或卷积神经网络（CNN），Transformer 并行计算能力更强，训练速度更快，同时由于其自注意力机制，能够更好地处理长序列数据中的依赖关系，避免了 RNN 中的梯度消失和长距离依赖问题。因此，Transformer 在自然语言处理领域取得了显著的成果，并逐渐成为许多先进模型的基础架构。

在 CASP14 上，DeepMind 提出了令人瞩目的 AlphaFold2[13, 14]。该方法以出色的预测精度赢得了第 14 届结构预测的关键评估（CASP14）大赛的冠军。在单个域的平均 GDTts 接近 90 的出色表现下，AlphaFold2 展示了其在蛋白质结构预测领域的卓越能力。运用了一种创新的神经网络模型——Evoformer，该模型融合了蛋白质的进化原理、物理特性和几何构型规则，显著提升了蛋白质结构的预测精准度。Evoformer 的设计理念受到 MSA-Transformer 的启发。Transformer 模型，作为一种新兴的自注意力网络，能够通过自注意力机制深入挖掘序列数据中的核心特征，并在多个 AI 领域展现了其应用价值。由 Transformer 演化而来的 MSA-Transformer 则采用 MSA 表征作为数据输入，借助注意力机制来解析蛋白质序列数据。Evoformer 采纳了与 MSA-Transformer 相似的双结构设计，一方面捕捉氨基酸残基之间的多序列比对信息，另一方面挖掘结构上的约束特征，这样

的双重机制大幅优化了预测效果。与 AlphaFold1 相比 AlphaFold2 的输入是“原始”的多个序列比对这使得深度学习网络能够直接从 MSA 中提取协同进化类型的信息。在随后的 CASP15 中参赛者广泛采用了整合了 AlphaFold2 的预测方法各模型在整体折叠和界面接触预测方面均表现出色。相较于 CASP14 期间 31% 的成功率 CASP15 实现了令人印象深刻的 90% 的成功率。AlphaFold2 的成功得益于其自注意力机制、自蒸馏的训练方式以及高效的搜索算法这些技术使得它能够从大量的蛋白质序列和结构数据中深度学习蛋白质的特征和规律从而高效地预测出未知的蛋白质结构并达到了接近实验水平的精度。

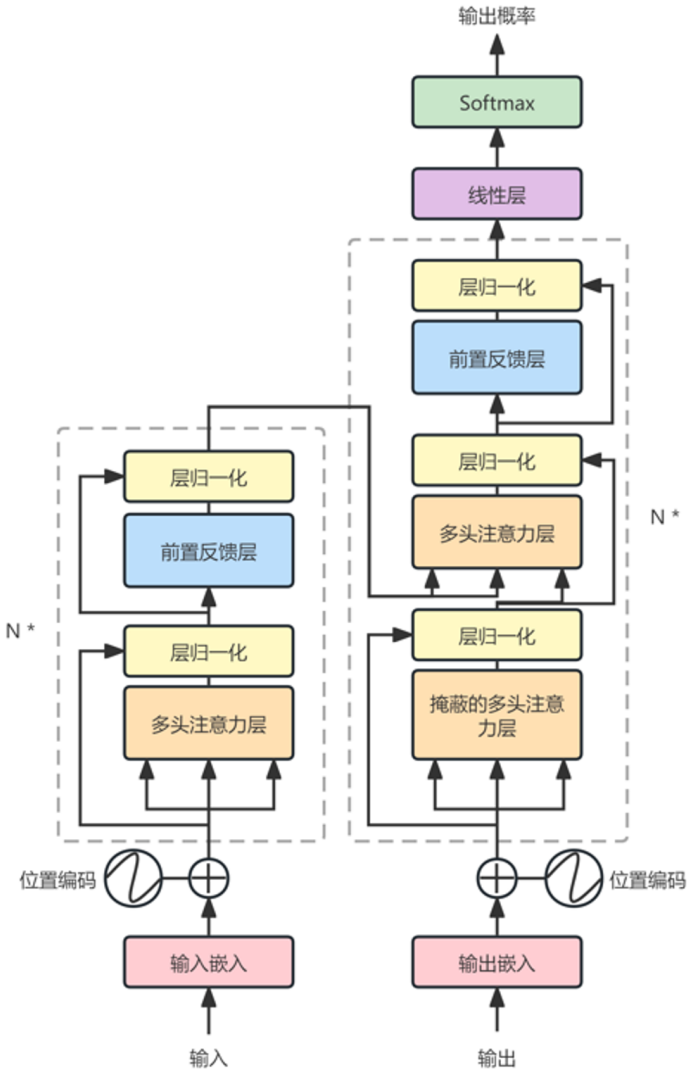

图 3. tranformer 网络结构示意图

ColabFold[15] 则是一种快速且易于使用的软件用于预测蛋白质结构和同聚体复合物。它通过替换 AlphaFold2 的同源性搜索为速度更快 40-60 倍的 MMseqs2（多对多序列搜索）来加速单个预测过程。此外通过避免重新编译和添加提前停止标准 ColabFold 将批量预测速度提高了约 90 倍。研究表明在 CASP14 目标上 ColabFold 的表现优于 AlphaFold-Colab 并在 ClusPro 数据集上与 AlphaFold-multimer 在预测质量方面相媲美。

RoseTTAFold[16, 17] 作为与 AlphaFold 同时发布的一种方法也展示了其独特的优势。该方法支持 X 射线晶体学和低温 EM 建模问题的解决方案在没有实验确定的结构的情况下提供了对蛋白质功能的深入洞察并能够快速生成蛋白质-蛋白质复合物的准确模型。对蛋白质-蛋白质复杂数据集的进一步训练有望进一步提高多蛋白质组装结构的建模能力。此外 RoseTTAFold 可以轻松地与现有的小分子和蛋白质粘结剂设计方法相结合为新药

发现等领域提供有力支持。RoseTTAFold 的创新之处在于其独特采用的三轨神经网络设计该架构能够将残基间的距离与方向、序列信息以及原子坐标紧密结合从而显著提升预测的精确度和效率。同时 RoseTTAFold 还能够充分利用多序列比对与共进化信息这些数据为其提供了对蛋白质结构的深入理解使其建模能力更为强大。因此 RoseTTAFold 在快速预测蛋白质-蛋白质复合物结构方面展现出了卓越的性能为蛋白质科学领域的研究提供了有力工具。

另一方面，ESMFold[16, 18] 是一种基于预训练语言模型的蛋白质结构预测方法，它建立在 Evolutionary Scale Modeling（ESM）的基础之上。尽管其主体结构与 AlphaFold2 有诸多相似之处，包括数据解析、编码器、解码器以及循环部分，但 ESMFold 在推理的神经网络结构上进行了简化。它消除了对明确同源序列（以多序列比对（MSA）形式）输入的需求，并且无需进行 Jax 图编译，从而节省了大量时间。这种优化使得 ESMFold 的推理速度比 AlphaFold2 快了一个数量级，甚至在实际应用中，其速度优势更为显著。

trRosettaX-Single[16, 19] 是 trRosettaX 系列中的一个重要成员。它在孤儿蛋白上的表现优于 AlphaFold2 和 RoseTTAFold，同时在人类设计的蛋白上也取得了良好成绩（平均模板建模得分（TM-score）为 0.79）。实验测试表明，完整的 trRosettaX-Single 管道比 AlphaFold2 快两倍，且使用的计算资源显著减少（<10%）。此外，与 trRosettaX 一样，trRosettaX-Single 也采用了端到端的训练方式，这种设计使得模型能够更深入地学习蛋白质序列与结构之间的关系，从而进一步提高了预测的准确性和效率。

Transformer 模型在蛋白质结构预测方面展现出显著的优劣。其优势在于能够捕捉蛋白质序列中的长程依赖关系，通过自注意力机制有效地处理序列数据，从而在预测蛋白质结构时提供更高的准确性。此外，Transformer 模型的并行计算能力使其能够高效处理大规模数据集，进一步提升了预测的效率。然而，Transformer 模型也存在一些劣势，特别是在处理超长序列时可能会遇到计算复杂度和内存消耗较高的问题。同时，模型的可解释性相对较弱，难以直观理解模型内部的决策过程。总体而言，Transformer 在蛋白质结构预测中表现出色，但仍需权衡其计算资源和可解释性的挑战。

3.3 GAN

GAN（生成对抗网络）作为一种前沿的深度学习技术，在蛋白质结构预测领域的应用虽然尚处于初级阶段，但其潜力已引起广泛关注。GAN 的核心机制在于其独特的生成器与判别器的对抗训练，通过这种训练方式，生成器能够产生高度逼真的数据样本，而判别器则负责鉴别这些样本的真实性（如图 4 所示）。在蛋白质结构预测中，GAN 的引入为这一领域带来了新的视角和方法。传统的蛋白质结构预测方法往往依赖于复杂的物理模型和大量的实验数据，而 GAN 则能够通过学习蛋白质序列与结构之间的复杂映射关系，生成与真实蛋白质结构高度相似的数据样本。这一特点使得 GAN 在蛋白质结构预测中具有独特的优势，尤其是在处理复杂和未知的蛋白质结构时。

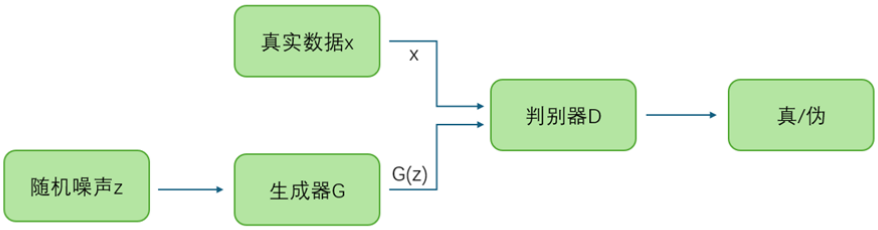

图 4. GAN 结构图

Yang[20] 在提出了一种创新的深度学习模型 SA-GAN，该模型在卷积神经网络（CNN）的基础上进行了重要

的改进与完善。他们巧妙地融合了生成对抗网络（GAN）和自注意力（Self-Attention）机制，同时将蕴含丰富进化信息的位置特异性得分矩阵（PSSM）作为模型输入的一部分。这样的设计使得网络能够更有效地捕获长程交互信息。作者证明了 GAN 在提取蛋白质数据特征方面的积极作用，同时也突显了自注意力模型在捕获数据和特征内部相关性方面的关键能力。将这两者优势结合，形成的自注意力生成对抗网络（SA-GAN）模型在三类蛋白质二级结构预测的准确率上实现了显著提升，并展示了良好的可扩展性。

Li[21] 提出了一种新型的蛋白质 8 态二级结构预测方法 WG-ICRN，将 WGAN 和 ICRN 相结合。在 WG-ICRN 中，WG-ICRN 方法采用 WGAN 提取氨基酸序列的蛋白质特征，结合 PSSM 构建特征矩阵 WG-data，丰富了数据基础。进而，该方法引入改进的卷积残差网络（ICRN）处理 WG-data，通过融入改进的 Inception 模块的 ResNet 提升网络性能，实现二级结构预测。实验表明，WG-ICRN 在六个数据集上表现优于其他四种方法，显示了 WGAN 与 ICRN 结合的优越性。但 WGAN 训练需平衡生成器与鉴别器，且方法对全局残基影响考虑不足。

Kihara[22] 的团队推出了一种名为 ContactGAN 的蛋白质接触图预测方法，该方法基于生成对抗网络（GAN）。通过 GAN 的精细化处理，ContactGAN 能够生成更为精确的蛋白质接触图。在包括 CASP13 蛋白结构建模靶点在内的数据集上进行测试时，该方法相较于最近的接触预测方法展现出了显著且一致的改进。无疑，ContactGAN 将成为结构预测流程中一个有价值的补充，进一步提升接触预测的精确度。

GAN（生成对抗网络）在蛋白质结构预测方面的优势在于其强大的生成能力。GAN 能够生成与真实蛋白质结构高度相似的数据，为预测提供多样化的候选结构。此外，GAN 通过对抗训练的方式，不断优化生成器和判别器，从而提高生成结构的准确性和逼真度。然而，GAN 也存在一些劣势。首先，GAN 模型相对复杂，需要大量的数据和计算资源进行训练，且训练过程可能不稳定。其次，GAN 生成的结构有时可能缺乏生物学上的合理性，需要进一步验证和修正。最后，GAN 在生成高分辨率、精细的蛋白质结构方面仍面临挑战。总体而言，GAN 为蛋白质结构预测提供了新的视角和方法，但其实用性和可靠性仍需进一步研究和改进。

4 结论

随着生物信息学和计算技术的飞速发展，机器学习和深度学习已经成为蛋白质结构预测领域的重要工具。本文深入探讨了这些先进技术在该领域的应用，展示了它们在提高预测准确性和效率方面的巨大潜力。然而，尽管取得了显著的进展，相关方法仍然面临着许多挑战和未来的发展方向。首先，数据的质量和多样性是蛋白质结构预测中亟待解决的问题。目前可用的数据集仍然有限，且存在着不平衡和噪声等问题。为了进一步提高预测的准确性，深度模型需要构建更大规模、更高质量的数据集，并探索有效的数据增强和预处理技术。其次，模型的可解释性和泛化能力也是未来研究的重要方向。当前的深度学习模型虽然取得了令人瞩目的成果，但往往缺乏可解释性，使得我们难以理解其内部的工作机制。此外，模型在面对未见过的蛋白质序列时，其泛化能力也受到一定的限制。

因此，未来的研究需要致力于开发更具可解释性的模型，并探索提高模型泛化能力的方法。最后，蛋白质结构的动态性和复杂性也为预测带来了挑战。蛋白质的结构并非静态不变，而是受到环境、配体和其他分子的影响而发生变化。因此，未来的研究需要更多地关注蛋白质的动态结构预测，以及考虑蛋白质与其他分子的相互作用。

随着技术的不断进步和创新，我们可以预见未来的蛋白质结构预测将更加精准、高效和智能化。新的算法、模型架构和优化策略将不断涌现，为我们提供更强大的工具来揭示蛋白质的神秘面纱。同时，随着数据的不断积累和计算能力的增强，我们也有望解决当前面临的一些挑战，推动蛋白质结构预测领域的持续发展。

创新说明

介绍了蛋白质结构预测的背景和重要性，然后详细阐述了机器学习和深度学习在蛋白质结构预测中的应用，包括常用的算法、模型架构以及优化策略。

展望了基于机器学习和深度学习的蛋白质结构预测在未来的发展方向和潜在挑战，为相关领域的研究者提供了有价值的参考。

参考文献

[1] Prediction Center. (n.d.). CASP: Critical Assessment of protein Structure Prediction. Retrieved from https://predictioncenter.org/ on March 15, 2023

[2] 王栋, 孙济洲, 李福超, 等. 基于并行多类支持向量机的蛋白质结构预测 [J]. 计算机应用研究,2011,28(02):465-468.

[3] 王菲露, 宋杰, 宋杨.BP 神经网络在蛋白质二级结构预测中的应用 [J]. 计算机技术与发展,2009,19(05):217-219+223.

[4] 王菲露, 宋杨. 基于广义回归神经网络的蛋白质二级结构预测 [J]. 计算机仿真,2012,29(02):184-187.

[5] 张斌, 尹京苑, 薛丹. 基于 RBF 神经网络的蛋白质二级结构预测 [J]. 生物信息学,2011,9(03):224-228+234.

[6] WANG S, SUN S Q, LI Z, et al. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model [J]. Plos Computational Biology, 2017, 13(1).

[7] XU J B. Distance-based protein folding powered by deep learning [J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116(34): 16856-65.

[8] WU Q, PENG Z L, ANISHCHENKO I, et al. Protein contact prediction using metagenome sequence data and residual neural networks [J]. Bioinformatics, 2020, 36(1): 41-8.

[9] FUKUDA H, TOMII K. DeepECA: an end-to-end learning framework for protein contact prediction from a multiple sequence alignment [J]. Bmc Bioinformatics, 2020, 21(1).

[10] LI Y, ZHANG C X, BELL E W, et al. Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks [J]. Plos Computational Biology, 2021, 17(3).

[11] JAIN A, TERASHI G, KAGAYA Y, et al. AttentiveDist: Protein Inter-Residue Distance Prediction Using Deep Learning with Attention on Quadruple Multiple Sequence Alignments [J]. bioRxiv, 2020.

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

[13] 张弘, 王慧洁, 鲁睿捷, 等. 蛋白质结构预测模型 AlphaFold2 的应用进展 [J/OL]. 生物工程学报:1-14[2024-04-22].https://doi.org/10.13345/j.cjb.230677.

[14] 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-+.

[15] MIRDITA M, SCHüTZE K, MORIWAKI Y, et al. ColabFold: making protein folding accessible to all [J]. Nature Methods, 2022, 19(6): 679-+.

[16] Meng, Q. Z., et al. (2023). "Improved structure-related prediction for insufficient homologous proteins using MSA enhancement and pre-trained language model." Briefings in Bioinformatics 24(4).

[17] LIU S, WU K, CHEN C. Obtaining protein foldability information from computational models of AlphaFold2 and RoseTTAFold [J]. Computational and Structural Biotechnology Journal, 2022, 20: 4481-9.

[18] NGUYEN P T, HARRIS B J, MATEOS D L, et al. Structural modeling of ion channels using AlphaFold2, RoseTTAFold2, and ESMFold [J]. Channels, 2024, 18(1).

[19] WANG W, PENG Z, YANG J. Single-sequence protein structure prediction using supervised transformer protein language models [J]. Nature computational science, 2022, 2(12): 804-14.

[20] 杨璐, 董洪伟. 基于自注意力机制和 GAN 的蛋白质二级结构预测 [J]. 中国科技论文在线精品论文,2023,16(02):148-159.

[21] LI S, YUAN L, MA Y M, et al. WG-ICRN: Protein 8-state secondary structure prediction based on Wasserstein generative adversarial networks and residual networks with Inception modules [J]. Mathematical Biosciences and Engineering, 2023, 20(5): 7721-37.

[22] MADDHURI VENKATA SUBRAMANIYA S R, TERASHI G, JAIN A, et al. Protein Contact Map Denoising Using Generative Adversarial Networks [J]. bioRxiv, 2020.

[23] Fang, F. A. N. G., Tan, W., & Liu, J. Z. (2005). Tuning of coordinated controllers for boiler-turbine units. Acta Automatica Sinica, 31(2), 291-296.

[24] Lv, Y., Fang, F. A. N. G., Yang, T., & Romero, C. E. (2020). An early fault detection method for induced draft fans based on MSET with informative memory matrix selection. ISA transactions, 102, 325-334.

[25] Zhang, X., Fang, F., & Liu, J. (2019). Weather-classification-MARS-based photovoltaic power forecasting for energy imbalance market. IEEE Transactions on Industrial Electronics, 66(11), 8692-8702.

[26] Wei, L., & Fang, F. (2016). H_∞-LQR-Based Coordinated Control for Large Coal-Fired Boiler–Turbine Generation Units. IEEE Transactions on Industrial Electronics, 64(6), 5212-5221.

[27] Liu, J., Song, D., Li, Q., Yang, J., Hu, Y., Fang, F., & Joo, Y. H. (2023). Life cycle cost modelling and economic analysis of wind power: A state of art review. Energy Conversion and Management, 277, 116628.

[28] Fang, F., Zhu, Z., Jin, S., & Hu, S. (2020). Two-layer game theoretic microgrid capacity optimization considering uncertainty of renewable energy. IEEE Systems Journal, 15(3), 4260-4271.

[29] Fang, F., & Xiong, Y. (2014). Event-driven-based water level control for nuclear steam generators. IEEE Transactions on Industrial electronics, 61(10), 5480-5489.

[30] Liu, J., Zeng, D., Tian, L., Gao, M., Wang, W., Niu, Y., & Fang, F. (2015). Control strategy for operating flexibility of coal-fired power plants in alternate electrical power systems. Proceedings of the CSEE, 35(21), 5385-5394.

[31] Fang, F., & Wu, X. (2020). A win–win mode: The complementary and coexistence of 5G networks and edge computing. IEEE Internet of Things Journal, 8(6), 3983-4003.

[32] Wang, N., Fang, F., & Feng, M. (2014, May). Multi-objective optimal analysis of comfort and energy management for intelligent buildings. In The 26th Chinese control and decision conference (2014 CCDC) (pp. 2783-2788). IEEE.

[33] Wang, W., Liu, J., Zeng, D., Fang, F., & Niu, Y. (2020). Modeling and flexible load control of combined heat and power units. Applied Thermal Engineering, 166, 114624.

[34] Lv, Y., Lv, X., Fang, F., Yang, T., & Romero, C. E. (2020). Adaptive selective catalytic reduction model development using typical operating data in coal-fired power plants. Energy, 192, 116589.

[35] Fang, F., Jizhen, L., & Wen, T. (2004). Nonlinear internal model control for the boiler-turbine coordinate systems of power unit. PROCEEDINGS-CHINESE SOCIETY OF ELECTRICAL ENGINEERING, 24(4), 195-199.

[36] Chang, K., Wang, Y., Ren, H., Wang, M., Liang, S., Han, Y., ... & Li, X. (2023). Chipgpt: How far are we from natural language hardware design. arXiv preprint arXiv:2305.14019.

[37] Wang, Y., Han, Y., Zhang, L., Li, H., & Li, X. (2015, June). ProPRAM: Exploiting the transparent logic resources in non-volatile memory for near data computing. In Proceedings of the 52nd Annual Design Automation Conference (pp. 1-6).

[38] Chen, W., Wang, Y., Yang, S., Liu, C., & Zhang, L. (2020, March). You only search once: A fast automation framework for single-stage dnn/accelerator co-design. In 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE) (pp. 1283-1286). IEEE.

[39] Hamdioui, S., Pouyan, P., Li, H., Wang, Y., Raychowdhur, A., & Yoon, I. (2017, November). Test and reliability of

emerging non-volatile memories. In 2017 IEEE 26th Asian Test Symposium (ATS) (pp. 175-183). IEEE.

[40] Ma, X., Wang, Y., Wang, Y., Cai, X., & Han, Y. (2022). Survey on chiplets: interface, interconnect and integration methodology. CCF Transactions on High Performance Computing, 4(1), 43-52.

[41] Wu, B., Wang, C., Wang, Z., Wang, Y., Zhang, D., Liu, D., ... & Hu, X. S. (2020). Field-free 3T2SOT MRAM for non-volatile cache memories. IEEE Transactions on Circuits and Systems I: Regular Papers, 67(12), 4660-4669.

[42] Zhao, X., Wang, Y., Liu, C., Shi, C., Tu, K., & Zhang, L. (2020, July). BitPruner: Network pruning for bit-serial accelerators. In 2020 57th ACM/IEEE Design Automation Conference (DAC) (pp. 1-6). IEEE.

[43] Han, Y., Wang, Y., Li, H., & Li, X. (2014, November). Data-aware DRAM refresh to squeeze the margin of retention time in hybrid memory cube. In 2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD) (pp. 295-300). IEEE.

[44] Wang, Y., Li, H., & Li, X. (2017). A case of on-chip memory subsystem design for low-power CNN accelerators. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 37(10), 1971-1984.

[45] Liu, C., Chu, C., Xu, D., Wang, Y., Wang, Q., Li, H., ... & Cheng, K. T. (2021). HyCA: A hybrid computing architecture for fault-tolerant deep learning. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 41(10), 3400-3413.

[46] Xu, D., Chu, C., Wang, Q., Liu, C., Wang, Y., Zhang, L., ... & Cheng, K. T. (2020, October). A hybrid computing architecture for fault-tolerant deep learning accelerators. In 2020 IEEE 38th International Conference on Computer Design (ICCD) (pp. 478-485). IEEE.

[47] Wang, C., Wang, Y., Han, Y., Song, L., Quan, Z., Li, J., & Li, X. (2017, January). CNN-based object detection solutions for embedded heterogeneous multicore SoCs. In 2017 22nd Asia and South Pacific design automation conference (ASP-DAC) (pp. 105-110). IEEE.

[48] Liu, B., Chen, X., Wang, Y., Han, Y., Li, J., Xu, H., & Li, X. (2019, January). Addressing the issue of processing element under-utilization in general-purpose systolic deep learning accelerators. In Proceedings of the 24th Asia and South Pacific Design Automation Conference (pp. 733-738).

[49] Li, C., Wang, Y., Liu, C., Liang, S., Li, H., & Li, X. (2021). GLIST: Towards in-storage graph learning. In 2021 USENIX Annual Technical Conference (USENIX ATC 21) (pp. 225-238).

[50] Qu, S., Li, B., Wang, Y., Xu, D., Zhao, X., & Zhang, L. (2020, July). RaQu: An automatic high-utilization CNN quantization and mapping framework for general-purpose RRAM accelerator. In 2020 57th ACM/IEEE Design Automation Conference (DAC) (pp. 1-6). IEEE.

[51] Wang, Y., Deng, J., Fang, Y., Li, H., & Li, X. (2017). Resilience-aware frequency tuning for neural-network-based approximate computing chips. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 25(10), 2736-2748.

[52] Li, W., Wang, Y., Li, H., & Li, X. (2019, January). P3M: a PIM-based neural network model protection scheme for deep learning accelerator. In Proceedings of the 24th Asia and South Pacific Design Automation Conference (pp. 633-638).

[53] Xu, D., Zhu, Z., Liu, C., Wang, Y., Zhao, S., Zhang, L., ... & Cheng, K. T. (2021). Reliability evaluation and analysis of FPGA-based neural network acceleration system. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 29(3), 472-484.

[54] Li, J., Chen, Z., Cheng, L., & Liu, X. (2022). Energy data generation with wasserstein deep convolutional generative adversarial networks. Energy, 257, 124694.

[55] Liu, Q., Cheng, L., Alves, R., Ozcelebi, T., Kuipers, F., Xu, G., ... & Chen, S. (2021). Cluster-based flow control in hybrid software-defined wireless sensor networks. Computer Networks, 187, 107788.

崔佳轩 2023 年入学贵州大学电子信息类专业。研究方向为信号处理与通信技术、嵌入式系统开发以及人工智能应用等。

Jiaxuan Cui enrolled in the Electronic Information major at Guizhou University in 2023, focusing on research directions such as signal processing and communication technology, embedded system development, and artificial intelligence applications.