**基于深度学习的蛋白质残基接触位点预测系统设计**

**摘 要**

伴随着人类基因组计划的顺利实施以及后基因组时代的到来，每天都会产生大量的基因和蛋白序列等数据。这些生物数据量随着社会的发展爆炸式的增长，最终产生了数以万计的生物分子序列数据。

蛋白质与配体物质相互作用是完成特定功能的关键，因此对蛋白质相互作用进行研究不仅有助于揭示生命活动的本质,而且对疾病发生机制的了解及有效药物的开发均起到推动性的作用。而生物信息和计算机技术的交差发展为深入了解蛋白质相互作用机理提供了基础条件。

本次课程实验尝试借助模式识别和深度学习的研究方法来预测金属离子锌-蛋白质残基相互作用位点，分别构建了支持向量机模型和密集连接网络来推理蛋白质残基-金属离子锌的相互作用位点，并且尝试将卷积神经网络模型与支持向量机模型进行堆叠组合，有效的提升了模型推理精度。

**关键词** 蛋白质残基相互作用；金属离子锌；接触位点预测；深度学习；卷积神经网络

**Prediction system of protein residue contact sites based on deep learning**

**Abstract**

With the smooth implementation of the Human Genome Project and the advent of the post-genomic era, a large amount of gene and protein sequence data are generated every day. The amount of these biological data has exploded with the development of society, and finally generated tens of thousands of biomolecule sequence data.

The interaction between proteins and ligand substances is the key to complete specific functions. Therefore, the study of protein interactions not only helps to reveal the essence of life activities, but also promotes the understanding of disease occurrence mechanism and the development of effective drugs. The cross development of biological information and computer technology provides the basic conditions for a deep understanding of protein interaction mechanisms.

In this course, we tried to use pattern recognition and deep learning research methods to predict the metal ion zinc-protein residue interaction site, and constructed a support vector machine model and a dense connection network to reason about the protein residue-metal ion zinc interaction. Function site, and try to stack and combine the convolutional neural network model and the support vector machine model, which effectively improves the model inference accuracy.

**Keywords** protein residue interaction; metal ion zinc; interaction site prediction; deep learning; convolutional neural network

目 录

[第1章 绪论 1](#_Toc40822832)

[1.1 课题研究背景、目的以及意义 2](#_Toc40822833)

[3.3.1 研究背景 2](#_Toc40822834)

[3.3.2 研究目的及意义 2](#_Toc40822835)

[1.2 国内外研究现状 2](#_Toc40822836)

[1.2.1 蛋白质-蛋白质相互作用位点预测研究现状 2](#_Toc40822837)

[1.2.2 锌离子结合蛋白质作用位点预测研究现状 2](#_Toc40822838)

[1.3 论文主要研究工作 2](#_Toc40822839)

[1.2.1 论文主要研究工作内容 2](#_Toc40822840)

[第2章 蛋白质功能位点预测相关知识 1](#_Toc40822841)

[2.1 蛋白质 1](#_Toc40822842)

[2.1.1 蛋白质的定义 1](#_Toc40822843)

[2.1.2 蛋白质的结构 1](#_Toc40822844)

[2.1.3 蛋白质功能位点 2](#_Toc40822845)

[2.1.4 蛋白质作用相关数据库 2](#_Toc40822846)

[2.2 蛋白质功能位点预测流程 3](#_Toc40822847)

[2.3 主要预测方法 4](#_Toc40822848)

[2.3.1 人工神经网络 4](#_Toc40822849)

[2.3.2 卷积神经网络 6](#_Toc40822850)

[2.3.3 密集连接网络 7](#_Toc40822851)

[2.3.4 支持向量机 8](#_Toc40822852)

[2.3.5 深度学习中的关键技术 8](#_Toc40822853)

[2.4 预测算法的评价指标 9](#_Toc40822854)

[第3章 数据集的构造 10](#_Toc40822855)

[3.1 数据集的处理 10](#_Toc40822856)

[3.1.1 数据集的选取 11](#_Toc40822857)

[3.2 特征的选择 10](#_Toc40822858)

[3.3.3 氨基酸的离散型编码 11](#_Toc40822859)

[3.3.4 蛋白质位置特异性得分矩阵 11](#_Toc40822860)

[3.3.5 蛋白质二级结构 11](#_Toc40822861)

[3.3 数据平衡化处理 10](#_Toc40822862)

[3.3.1 欠采样 11](#_Toc40822863)

[3.3.2 过采样 11](#_Toc40822864)

[第4章 基于密集连接网络的金属离子锌结合蛋白质作用位点预测方法 10](#_Toc40822865)

[4.1 数据集预处理 10](#_Toc40822866)

[4.2 网络模型结构 10](#_Toc40822867)

[4.3 性能评估 10](#_Toc40822868)

[4.4 本章小结 10](#_Toc40822869)

[第5章 基于支持向量机的金属离子锌结合蛋白质作用位点预测方法 10](#_Toc40822870)

[第6章 基于密集连接网络与支持向量机模型融合的金属离子锌结合蛋白质作用位点预测方法 10](#_Toc40822871)

[结 论 11](#_Toc40822872)

[致 谢 12](#_Toc40822873)

[参 考 文 献 13](#_Toc40822874)

# 绪论

## 课题研究背景、目的以及意义

### 研究背景

伴随着人类基因组计划的顺利实施以及后基因组时代的到来，每天都会产生大量的基因和蛋白序列等数据。这些生物数据量随着社会的发展爆炸式的增长，最终产生了数以万计的生物分子序列数据。从GOLD数据库（网址：<https://gold.jgi.doe.gov/statistics> 需要用Google Chrome浏览器打开）查询可知，截止 2019 年 12 月，已有近2万个基因组完成测序，图 1.1 显示了 1997年到 2019年被完整测序的基因组数目。

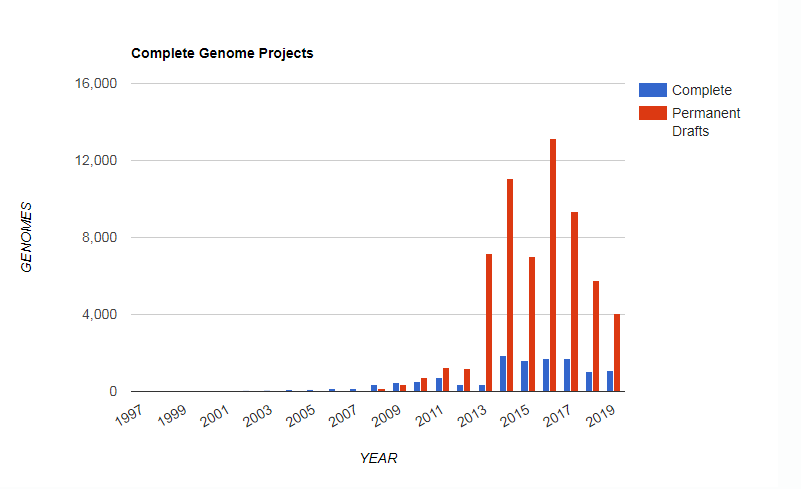


图 1.1 1997 年到 2019 年被完整测序的基因组数目

蛋白质是生命活动的体现者，它是一切生物藉以表现生命的最重要的基本单元，生物体系的运作与蛋白质间的相互作用密不可分。蛋白质相互作用是生命活动的基础，对蛋白质相互作用的研究能够让我们从更深层面上认识生命活动本身，同时也能够用于蛋白质功能的预测[13]。蛋白质结构预测不仅可以有效降低可以使用的蛋白质序列数目与蛋白质结构数目之间的鸿沟，从而辅助结构基因组学的研究，也是人们认识和了解蛋白质折叠机理的一种重要途径(Dill and MacCallum, 2012)[14]。

在生物信息学中，蛋白质结合位点的预测一直是主要的研究课题。目前传统蛋白质结合位点的检测方法主要还是借助物理化学实验[17]，一般有传统的小规模生化试验方法和高通量试验方法。传统的生化实验方法可以检测出很少的蛋白质与配体之间的相互作用，高通量实验方法则可以大规模的鉴定多对相互作用对[15]。对于蛋白质与配体间相互作用的研究，无论是小规模的生化试验方法还高通量实验方法，都在这方面的研究中取得了一定的研究成果并且进行了一些应用，但是还是存在一些问题[15]：

1. 化学实验的完成需要一定的环境限制，对实验环境要求比较严格，造成实验结果复现难的问题。
2. 化学实验的基础是大量的生物数据，实验代价相对比较昂贵，实验耗时较长。
3. 现代化的测序工艺进步，蛋白质与配体作用对预测速度和蛋白质序列测序速度不匹配产生了矛盾。

随着生物信息学的发展，国内外越来越多的研究学者在尝试借助模式识别的研究方法来对预测蛋白质和配体的相互作用。现有的生物信息学的研究方法有基于统计分析的方法与基于机器学习的方法。相较于统计分析方法，机器学习的方法可捕获更加复杂的序列特征，因此在该问题上的预测效果更佳。尽管目前已经产生了一系列基于机器学习的作用位点预测算法，但预测精度仍有待提高[16]。目前现有的机器学习方法大多数依赖人为干预的蛋白质一级结构序列特征提取过程，提取到的数据特征并不能很充分的表达蛋白质序列的特性。

### 研究目的及意义

传统的蛋白质与配体相互作用检测方法或多或少存在问题，生物信息学发展到目前已经有不少学者将模式识别的研究思想应用到了蛋白质相互作用位点的研究。目前用来研究蛋白质相互作用位点的传统的机器学习算法主要有人工神经网络、支持向量机[8]、决策树、随机森林以及聚类算法等一系列的方法。近几年来国内外学者开始将目光转向了机器学习中的一个独特的领域——深度学习。目前深度学习在自然语言处理、图像处理和机器翻译等方面以及取得了巨大的成果，深度学习具有学习能力强、覆盖范围广、适应能力强、可移植性好等特点。而卷积神经网络（CNN）和长短时期神经网络（LSTM）在深度学习领域更是表现力优秀的神经网络模型，它们在图像处理和序列分析场景下常常能取得相当不错的效果。蛋白质一级结构是氨基酸序列，在深度学习的角度属于序列预测问题（seq2seq问题），国内外一大批学者尝试以LSTM为基础网络模型进行改进，将LSTM深度学习神经网络模型成功运用到了蛋白质与配体结合位点预测问题上[20]，让神经网络自己去提取序列特征，并且取得了不错的效果。而卷积神经网络CNN作为图像处理领域的表现优异的网络，也逐渐被研究学者运用到了蛋白质结合位点预测场景。卷积神经网络在可以通过不同维度的卷积核进行不同层次的特征提取，从而丰富蛋白质序列特征，更好的对蛋白质结构和理化特性进行表达[18]。卷积神经网络CNN[19]，不仅具备强大的自主学习能力，而且可以从样本数据中挖掘出隐含特征，简化了人工表征特征的过程。本课题重点研究将卷积神经网络应用于蛋白质与金属离子锌的相互作用位点预测，使用经典的CNN结构构建模型，通过在不同的数据集上的训练准确度来评价模型。

## 国内外研究现状

### 蛋白质-蛋白质相互作用位点预测研究现状

蛋白质通过与其他蛋白质，化合物，RNA和DNA相互作用，在几乎所有生物事件中都起着至关重要的作用。了解相互作用位点的特征是了解分子识别过程的基础。蛋白质很少单独发挥作用，经常通过成为大型分子网络的一部分来发挥其功能，并通过复杂的蛋白质-蛋白质相互作用（PPI）调控网络来协调作用。因此，蛋白质复合物而不是单个组分将决定生物系统的行为。PPI对于细胞功能的几乎所有方面都是至关重要的，包括信号和代谢途径的调控，蛋白质合成，DNA复制和基因翻译以及免疫学识别[1]。特别地，鉴定两个相互作用的蛋白质之间的结合位点将为理解和确定蛋白质复合物的功能和结构，促进药理学靶标的鉴定以及最终用于药物设计提供有价值的线索。因此预测蛋白质相互作用位点具有重要意义。

迄今为止，国外学者已经提出了几种预测PPI位点的方法，并且根据所使用的特征可以将它们大致分为三类。第一类方法仅基于序列信息[2]。Ofran与Rost在研究PPI位点过程中使用了相同的数据集，该数据集包括333个复合物中的1,134条链和59,559个接触残基[3]。在他们的2003年研究[中](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3429425/" \l "pone.0043927-Ofran1)预测精度达到了70%，正确地预测了20％（66/333）的复合物中至少一个相互作用位点。第二类方法整合了二级结构信息和序列信息[4]。Wang的研究工作[4]基于由69条蛋白质链组成的异源二聚体的非冗余数据集，其灵敏度为66.3％，特异性为49.7％，准确度为0.654，相关系数为0.297。Zhou及其同事[5]构建了615对非同源复合物形成蛋白进行训练的预测因子，并在一组不同的129对非同源复合物形成蛋白上进行了测试。该方法，可以正确预测了11004个界面残基中的70％。第三类方法使用3D结构信息或具有序列信息的集成3D结构进行预测[6]。Aytuna等人[6]提出了一种算法，该算法运行在67个接口的模板数据集和6170个蛋白质结构的顺序非冗余数据集上。预测的62,616种潜在相互作用的大多数已在公共数据库（包括相互作用蛋白数据库，生物分子相互作用网络数据库和PDB）中得到验证。在Sikic等人的工作中，他们使用了与Ofran和Rost的研究相同的数据集[3]。在这项研究中，基于序列的预测实现了84％的准确率和26％的召回率。与结构信息结合后，预测性能将提高到76％的精度和38％的召回率[7]。基于不同的特征，已经提出了几种机器学习方法来预测蛋白质相互作用的位点，例如神经网络[3]，支持向量机（SVM）[8]，贝叶斯网络[9]和隐马尔可夫模型（HMM）[10]。Li等人[11]在一组50种随机选择的蛋白质上测试了训练的最佳SVM模型。预测核心界面残基的敏感性，特异性和MCC分别为60.6％，53.4％和0.243。Bradford 等人[9]使用贝叶斯网络预测180个蛋白质基准数据集上的蛋白质-蛋白质结合位点，成功率为82％。在Li及其同事的研究中[11]，当使用1276条非冗余的异质复合蛋白链作为训练和测试集时，最佳精度，召回率，准确性和MCC分别为0.536、0.595 0.692和0.328。

### 锌离子结合蛋白质作用位点预测研究现状

很大一部分蛋白质需要金属来稳定其结构或发挥其功能。锌是这些金属中最常见的一种，它与具有多种生物学作用的蛋白质有关。最近的一项研究估计，人类蛋白质组中40％的锌结合蛋白是转录因子。其余的60％主要是参与离子迁移的酶和蛋白质（[Andreini等，2006](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2144590/#B04)）。金属离子锌具有许多化学性质，因此具有多种生物学功能（[Vallee和Auld 1990](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2144590/#B55)）。由于其D壳已充满，因此不会发生氧化或还原反应。这在氧化还原电势可能波动的生物环境中提供了一定程度的稳定性。它具有不同的配位数和几种类型的连接残基的能力可能使蛋白质调节其结合位点的亲和力和功能[12]。

## 论文主要研究工作

### 论文主要研究工作内容

本课题围绕深度学习在蛋白质残基接触位点预测上的应用，主要利用卷积神经网络对金属离子锌结合蛋白质相互作用位点进行预测。蛋白质结合位点预测问题存在数据非平衡问题，即样本中正负样本数量差距过大，正类数量远远小于负类数量。因此本课题尝试对样本数据进行欠采样与过采样，防止样本数据特征丢失过多和特征错误补偿。

第一章主要介绍了蛋白质结合位点的研究现状以及背景资料

第二章主要介绍了蛋白质功能位点预测的相关知识，包括蛋白质、蛋白质作用位点的定义，常用的蛋白质作用数据库；本次课题利用的预测算法模型。

第三章主要介绍了本次课题的数据集来源以及以及数据特征的选取

第四章主要介绍

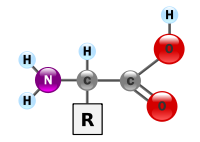
# **蛋白质功能位点预测相关知识**

## 蛋白质

蛋白质作为生命的物质基础，承担着关键的生命活动。蛋白质是由氨基酸脱水缩合而成的大分子化合物，是构成生物细胞的基本物质之一。

### 蛋白质的定义

蛋白质由一个或多个α-[氨基酸](https://zh.wikipedia.org/wiki/%E6%B0%A8%E5%9F%BA%E9%85%B8)[残基](https://zh.wikipedia.org/wiki/%E6%AE%98%E5%9F%BA)组成的长链条组成。 α-[氨基酸](https://zh.wikipedia.org/wiki/%E6%B0%A8%E5%9F%BA%E9%85%B8)分子呈线性排列，相邻α-氨基酸[残基](https://zh.wikipedia.org/wiki/%E6%AE%98%E5%9F%BA" \o "残基)的[羧基](https://zh.wikipedia.org/wiki/%E7%BE%A7%E5%9F%BA)和[氨基](https://zh.wikipedia.org/wiki/%E6%B0%A8%E5%9F%BA)通过[肽键](https://zh.wikipedia.org/wiki/%E8%82%BD%E9%94%AE)连接在一起。下面是蛋白质的结构式如图2.1.1所示：



### 蛋白质的结构

大多数的蛋白质都自然[折叠](https://zh.wikipedia.org/wiki/%E8%9B%8B%E7%99%BD%E8%B4%A8%E6%8A%98%E5%8F%A0)为一个特定的三维结构，这一特定结构被称为[天然状态](https://zh.wikipedia.org/w/index.php?title=%E5%A4%A9%E7%84%B6%E7%8A%B6%E6%80%81&action=edit&redlink=1)。虽然多数蛋白可以通过本身氨基酸序列的性质进行自我折叠，但还是有许多蛋白质需要[分子伴侣](https://zh.wikipedia.org/wiki/%E5%88%86%E5%AD%90%E4%BC%B4%E4%BE%A3)的帮助来进行正确的折叠。在高温或极端pH等条件下，蛋白质会失去其天然结构和活性，这一现象就称为[变性](https://zh.wikipedia.org/wiki/%E5%8F%98%E6%80%A7_(%E7%94%9F%E7%89%A9%E5%8C%96%E5%AD%A6)" \o "变性 (生物化学))。生物化学家常常用以下四个方面来表示蛋白质的结构：

[一级结构](https://zh.wikipedia.org/wiki/%E8%9B%8B%E7%99%BD%E8%B3%AA%E4%B8%80%E7%B4%9A%E7%B5%90%E6%A7%8B)（Primary structure）：组成蛋白质多肽链的线性氨基酸序列。一个蛋白质是一个[聚酰胺](https://zh.wikipedia.org/wiki/%E8%81%9A%E9%85%B0%E8%83%BA" \o "聚酰胺)。

[二级结构](https://zh.wikipedia.org/wiki/%E8%9B%8B%E7%99%BD%E8%B4%A8%E4%BA%8C%E7%BA%A7%E7%BB%93%E6%9E%84)（Secondary structure）：依靠不同氨基酸之间的C=O和N-H基团间的[氢键](https://zh.wikipedia.org/wiki/%E6%B0%A2%E9%94%AE)形成的稳定结构，主要为[α螺旋](https://zh.wikipedia.org/wiki/%CE%91%E8%9E%BA%E6%97%8B)和[β折叠](https://zh.wikipedia.org/wiki/%CE%92%E6%8A%98%E5%8F%A0)。因为二级结构是局部的，不同的二级结构的许多区域可存在于相同的蛋白质分子。

[三级结构](https://zh.wikipedia.org/wiki/%E8%9B%8B%E7%99%BD%E8%B3%AA%E4%B8%89%E7%B4%9A%E7%B5%90%E6%A7%8B)（Tertiary structure）：通过多个二级结构元素在三维空间的排列所形成的一个蛋白质分子的三维结构，是单个蛋白质分子的整体形状。蛋白质的三级结构大都有一个[疏水核心](https://zh.wikipedia.org/wiki/%E7%96%8F%E6%B0%B4%E6%A0%B8%E5%BF%83" \o "疏水核心)来稳定结构，同时具有稳定作用的还有[盐桥 (蛋白质)](https://zh.wikipedia.org/w/index.php?title=%E7%9B%90%E6%A1%A5_(%E8%9B%8B%E7%99%BD%E8%B4%A8)&action=edit&redlink=1)、氢键和[二硫键](https://zh.wikipedia.org/wiki/%E4%BA%8C%E7%A1%AB%E9%8D%B5" \o "二硫键)，甚至[翻译后修饰](https://zh.wikipedia.org/wiki/%E7%BF%BB%E8%AF%91%E5%90%8E%E4%BF%AE%E9%A5%B0)。“三级结构”常常可以用“折叠”一词来表示。三级结构控制蛋白质的基本功能。

[四级结构](https://zh.wikipedia.org/wiki/%E8%9B%8B%E7%99%BD%E8%B4%A8%E5%9B%9B%E7%BA%A7%E7%BB%93%E6%9E%84)（Quaternary structure）：由几个蛋白质分子（多肽链），通常称为[蛋白质亚基](https://zh.wikipedia.org/wiki/%E8%9B%8B%E7%99%BD%E8%B4%A8%E4%BA%9A%E5%9F%BA" \o "蛋白质亚基)所形成的结构，在功能上作为一个[蛋白质复合体](https://zh.wikipedia.org/wiki/%E8%9B%8B%E7%99%BD%E8%B4%A8%E5%A4%8D%E5%90%88%E4%BD%93)。

### 蛋白质功能位点

化学家 Anfisen 等人提出蛋白质分子的一级结构决定其三维结构，而三维结构及其

动态变化是蛋白质发挥其功能的基础。蛋白质的功能往往由其与不同配体相互作用决定的，在相互作用过程中，蛋白质分子中有一类残基对其行使功能起着关键作用，这类残基称为蛋白质的功能位点[15]。常用的蛋白质功能位点类型有蛋白质-蛋白质相互作用位点、蛋白质

与金属离子相互作用位点、蛋白质翻译后修饰位点如磷酸化、糖基化和泛素化位点等、蛋白

质与核酸相互作用位点等。一般来讲，蛋白质功能位点的识别对蛋白质生物学功能预测具有重要意义。比如金属结合蛋白质作用位点，金属离子采取不同的方式结合蛋白质，蛋白质具有不同的功能，如结构维护、分子识别与催化，信号传递与存储、基因转录与表达调节等。本文主要研究金属结合蛋白质作用位点的预测方法。

### 蛋白质作用相关数据库

| **数据库** | **物种** | **描述** |
| --- | --- | --- |
| Bind | Many | 生物分子相互作用数据库 |
| BioGrid | Many | 蛋白质-蛋白质以及基因相互作用的数据库 |
| DIP | Many | 相互作用的蛋白质数据库 |
| HPID | Human | 人类蛋白质相互作用数据库 |
| HPRD | Human | 人类蛋白质参考数据库 |
| IntAct | Many | EBI 的蛋白质相互作用数据库 |
| MINT | Many | 分子相互作用数据库 |
| MIPS | Mammals | 哺乳动物蛋白质相互作用数据库 |
| OPHID | Human | 在线预测人类相互作用数据库 |
| STRING | Many | 实验和生物信息学方法预测的蛋白质网络数据库 |
| PDB | Many | 国际上唯一的生物大分子结构数据库，主要包括蛋白质、核酸等实验测定的结构信息 |
| ProSite | Many | 收集了生物学有显著意义的蛋白质位点和序列模式，并能快速鉴别蛋白质属于相应的蛋白质家族 |

## 蛋白质功能位点预测流程

本次课题主要利用深度学习算法来预测蛋白质功能位点，主要原理是通过对实验测定的蛋白质功能位点数据进行分析，找到存在的内在规律，提取特征属性，利用深度学习方法建立预测模型，预测蛋白质残基是否是作用的功能位点。

蛋白质功能位点预测识别是一个典型的二类分类问题。金属离子锌结合蛋白质作用位点主要是针对四种氨基酸残基，半胱氨酸（CYS）、组氨酸（HIS)、天冬氦酸（ASP)、谷氨酸（GLU)，是否与蛋白质发生结合作用，则金属结合位点的预测问题转变成了判断这四种残基是否是结合位点和非结合位点。

首先，对选取蛋白质属性特征进行提取，并向量化作为模型的输入然后选择算法构建分类模型，并进行模型训练直至满足精度要求；最后，利用该训练的模型对未知测试样本进行预测。预测流程如图2.2所示：

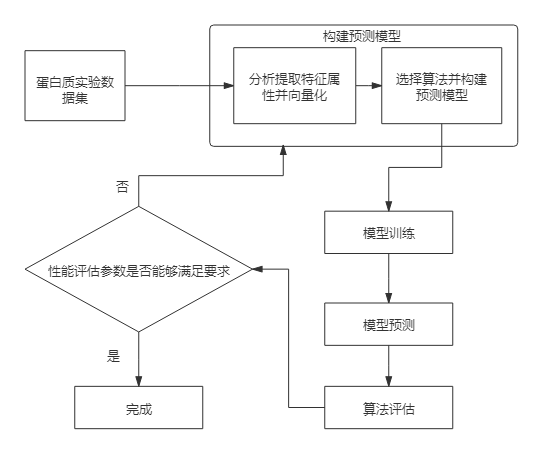


图 2.3 蛋白质残基接触位点预测流程

## 主要预测方法

### 人工神经网络

人工神经网络(Artificial Neural Network，ANN)[21]自上世纪80年代以来就是人工智能领域的研究热点，人脑以计算机无法理解的方式来解释现实事物。人工神经网络是仿真人类大脑的神经元网络，希望能够是计算机以类似人类的学习逻辑来学习事物并且做出推理决策。

人工神经网络是计算机领域用于处理机器学习问题的强大工具，其广泛应用于回归与分类等问题中，它模拟了生物体神经细胞的运作原理，将一个个具有层次关系，连接关系的人工神经元组成网络结构，通过数学表达的方式模拟神经元之间的信号传递，从而可建立一个具有输入与输出关系、并可通过网络方式可视化的非线性方程，我们称之为人工神经网络。一般而言，ANN可通过合理的网络结构配置拟合任意非线性函数，因此它也用于处理内部表达较为复杂的非线性系统或黑箱模型。人工神经网络内部的结构是通过学习不断地进行改变的，它是一种自适应系统。典型的神经网络主要分为两部分：

1. 神经元

神经元模型是神经网络中最基本的组成成分。我们先忽略激活函数，神经元包含输入和输出。通常一个人工神经网络模型由*L*层神经元组成，其中：第一层称为输入层，最后一层（第*L*层）被称为输出层，其它各层均被称为隐含层（第2层~第*L*-1层）。典型的人工神经网络如图2.3.1.1所示：

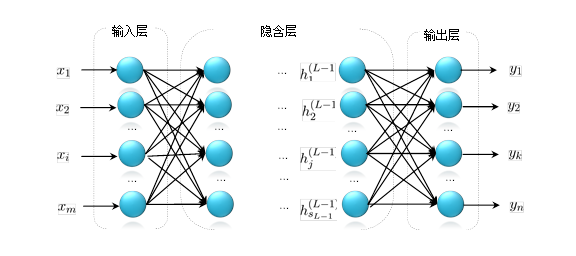


图 2.3.1.1 典型的人工神经网络结构

输入向量为：

输出向量为：

第隐含层各神经元的输出为：

其中，为第层神经元的个数。

设 为从层第个神经元与层第 个神经元之间的连接权重；为第层第个神经元的偏置，那么：

其中，为层第个神经元的输入，为神经元的激活函数。通常在多层神经网络中采用非线性激活函数，而不是用线性激活函数，因为采用基于线性激活函数的多层神经网络本质上还是多个线性函数的叠加，其结果仍然为一个线性函数。

1. 激活函数

BP神经网络通常使用下面两种非线性激活函数：

### 卷积神经网络

卷积神经网络（Convolutional Neural Network, CNN）是一种前馈神经网络，它的卷积层（Convolutional layer）可以作用在部分覆盖范围内的周围单元，在图像处理领域具有出色的表现。与传统的人工神经网络结构不同，它包含有非常特殊的卷积层和降采样层（又称之为池化层、汇合层），其中卷积层和前一层采用局部连接和权值共享的方式进行连接，从而大大降低了参数数量。降采样层可以大幅降低输入维度，从而降低网络复杂度，使网络具有更高的鲁棒性，同时能够有效的防止过拟合。由于以上设计，卷积网络主要用来识别缩放，位移以及其他形式扭曲不变的二维图形，并且可以直接以原始图片作为输入，而无需进行复杂的预处理工作。典型的卷积神经网络结构图如图2.3.2所示：

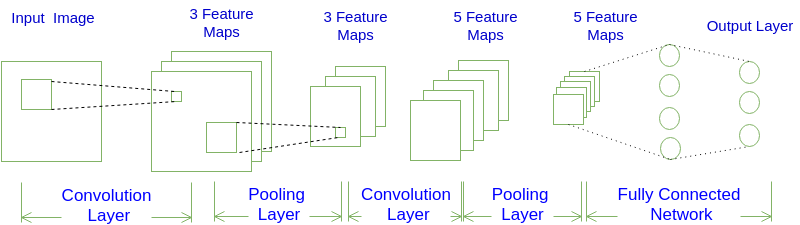


图 2.3.2 典型的卷积神经网络结构

卷积神经网络通常包含以下几种层：

1. 卷积层（Convolutional layer）

卷积神经网路中每层卷积层由若干卷积单元组成，每个卷积单元的参数都是通过反向传播算法优化得到的。卷积运算的目的是提取输入的不同特征，第一层卷积层可能只能提取一些低级的特征如边缘、线条和角等层级，更多层的网络能从低级特征中迭代提取更复杂的特征。

1. 线性整流层（Rectified Linear Units layer, ReLU layer）

这一层神经的活性化函数（Activation function）使用线性整流（Rectified Linear Units, ReLU）。

1. 池化层（Pooling layer）

通常在卷积层之后会得到维度很大的特征，将特征切成几个区域，取其最大值或平均值，得到新的、维度较小的特征。

1. 全连接层（Fully-Connected layer）

把所有局部特征结合变成全局特征，用来计算最后每一类的得分。

### 密集连接网络

密集连接网络（DenseNet）是一种结构特殊的卷积神经网络，它的特点是在神经网络的前向传递中，每一层都和前面的所有层直接连接，每层的输入来自于之前所有层的输出。一个普通的有层的神经网络会产生个层与层之间的连接，而同样是层的DenseNet则会因为它层与层相互连接的特点，产生多达个连接。正因为如此特殊的网络结构，Densenet网络可以减轻梯度弥散问题，加强特征传播，鼓励特征重用，大大减少参数数量。Densenet网络结构示意图如图2.3.3所示：

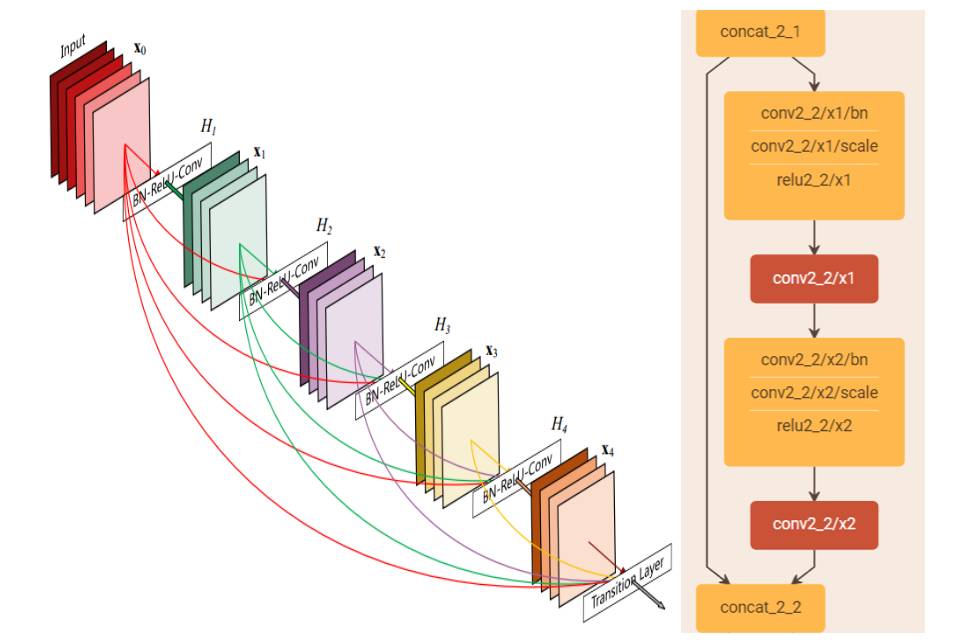


图 2.3.2 密集连接（Densenet）网络结构

### 支持向量机

支持向量机（Support Vector Machines, SVM）是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；SVM还包括核技巧，这使它成为实质上的非线性分类器。SVM的的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。SVM的的学习算法就是求解凸二次规划的最优化算法。

### 深度学习中的关键技术

1. 激活函数（Activation Function）

在计算网络中，一个节点的激活函数定义了该节点在给定的输入或输入的集合下的输出。标准的计算机芯片电路可以看作是根据输入得到"开"(1)或"关"(0)输出的数字网络激活函数。这与神经网络中的线性感知机的行为类似。 一种函数（例如 ReLU 或 S 型函数），用于对上一层的所有输入求加权和，然后生成一个输出值（通常为非线性值），并将其传递给下一层。

1. Dropout机制

[Dropout](https://mp.weixin.qq.com/cgi-bin/appmsg?t=media/appmsg_edit&action=edit&type=10&appmsgid=503278593&isMul=1&token=805468140&lang=zh_CN) 机制是一种避免前馈神经网络中出现的过拟合现象的简单技术。在每轮迭代中，网络中的每个神经元以 p 的概率被丢弃。当训练完成后，尽管神经元的输出要乘以该神经元被丢弃的概率 p，整体的网络架构还是会被使用。这样做抵消了没有神经元被丢弃的神经网络尺寸过大的影响，并且可以被解释为在训练时可能出现的网络上取平均。每一层的 dropout 概率可能不尽相同，原始的 Dropout 论文建议输入层的 p=0.2，而隐藏层的 p=0.5。输出层中的神经元不会被丢弃。该方法如图 2.3.5.2 所示。

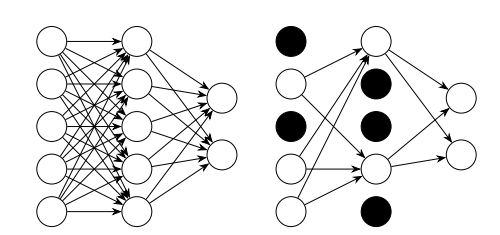


图 2.3.5.2 标准 Dropout 的示例

1. 损失函数（Loss Function）

损失函数用来评价模型的预测值和真实值不一样的程度，损失函数越好，通常模型的性能越好。不同的模型用的损失函数一般也不一样。损失函数分为经验风险损失函数和结构风险损失函数。经验风险损失函数指预测结果和实际结果的差别，结构风险损失函数是指经验风险损失函数加上正则项。

1. 注意力机制

--------------------------------------------------------------------------------------------------------------

## 预测算法的评价指标

衡量一个分类器预测性能的好坏，选择合适的验证方法和评价指标是非常关键的。下面介绍常用的评估验证方法。为了更好地、更公正地评价预测方法的性能，采用下面的评价指标参数对其进行评价。

1. 召回率
2. 精确度
3. 特异度
4. 准确度
5. MCC尺度
6. F1得分

True Positive（真正， TP）被模型预测为正的正样本；True Negative（真负 ， TN）被模型预测为负的负样本； False Positive （假正， FP）被模型预测为正的负样本； False Negative（假负， FN）被模型预测为负的正样本。

# 数据集的构造

## 数据集的处理

### 数据集的选取

选择 Zhao 等人[10]研究中的数据集（Zhao\_dataset）作为本次课题的训练集，Passerini 等人[11]用的数据集作为测试集。在 Zhao 数据集中去掉与 Passerini 重复的蛋白质序列，最后得到 391条蛋白质链。根据上文锌结合位点的定义，在 Zhao 数据集上得到锌结合位点 2023 个，非锌结合位点 14493 个。

## 特征的选择

### 氨基酸的离散型编码

深度学习框架可以以长度为 33 的蛋白质片段作为输入如图3.3.3.1所示。但原始的蛋白质片段数据是由氨基酸字母表示的类别数据，需要对其进行数值转化才能变成可计算的数据。这里参考普遍用于自然语言处理问题中的编码方式，既 one-of-K 编码。在one-of-K 编码中，K 代表字典的大小，对于氨基酸的编码 K 可以等于 20 代表 20 种常见的氨基酸，每一个氨基酸都用一个维度为 20 的向量表示，只有对应这个氨基酸的位置是 1 其他位置都是 0。然而对于那些蛋白质片段长度不足 33 的，需要用额外的特殊符号 “-”表示，因此实际的 K=21。对于其他非常见氨基酸或者不确定的氨基酸，例如一般蛋白质数据库中以 X 表示的氨基酸，其编码向量所有位置的值都为 0.05，代表 20 种氨基酸的均匀分布。最终，一个长度为 33 的蛋白质序列片段可以用 33×21 的矩阵表示，其具体的 one-of-21 编码方式如图3.3.3.1所示。

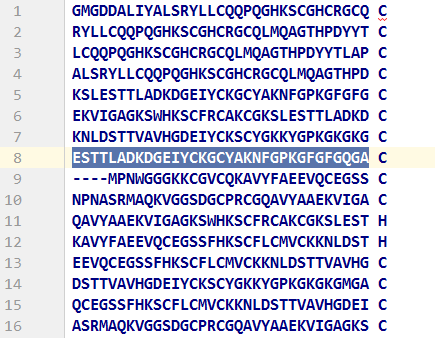


图 3.3.3.1 长度为 33 的蛋白质片段

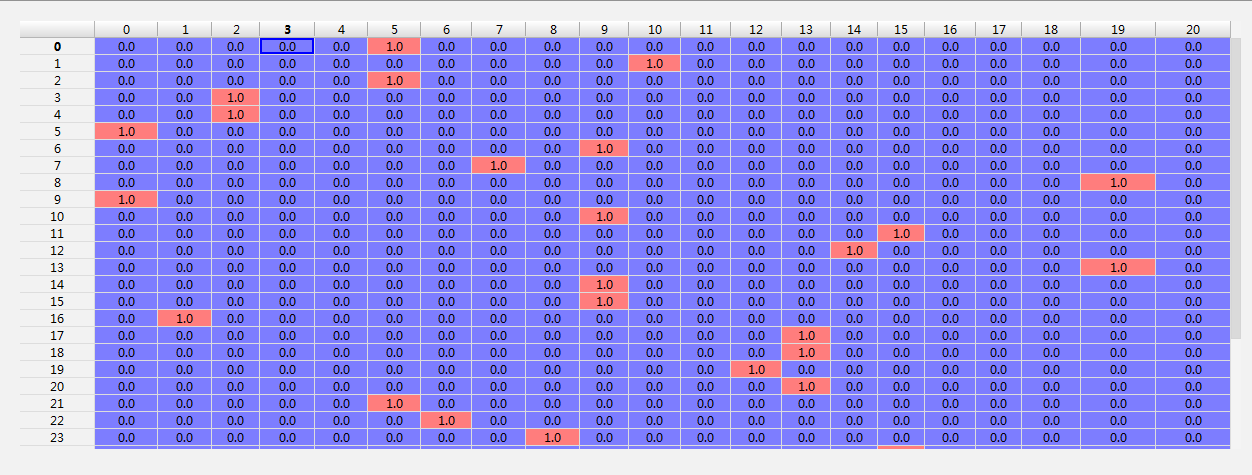


图 3.3.3.2 长度为 33 的蛋白质片段one-of-K编码部分数据

### 蛋白质位置特异性得分矩阵

位置特异性打分矩阵（PSSM）由蛋白质序列在数据库中迭代进行多序列比对得到。生物进化理论认为，如果两种生物体是由同一个祖先进化而来，则它们的遗传物质会倾向于具有一定的相似性。也就是说同源物种的遗传物质在整个演化进程中不断地通过局部的序列缺失和重组以及部分碱基的替换等方式进行演变，从而造成了物种之间的显性差异。而多序列比对的方法能够提取出蛋白质序列中所包含的相似性信息，并由此寻找序列中能够决定生物体性状的序列模式。以 Lichtarge 和 Thornton 等人的研究为代表的目前已有的一些结果表明演化信息的重要度评分越高，则该位置的残基是结合位点的可能性越大。因此，使用位置特异性打分矩阵作为预测蛋白质结合位点的特征，具有合理性。在蛋白质功能以及蛋白质相互作用预测，蛋白质二级结构预测问题中 PSSM 得分矩阵都是不可或缺的一种特征。并且在前人对于蛋白质-ATP 结合位点预测问题的研究中已经证明了PSSM 位置特异性得分矩阵对于这个问题的重要性

### 蛋白质二级结构

蛋白质的二级结构特征主要反映的是局部空间构象，这种构象是重复的且有规则的。蛋白质序列主链原子的三维空间排列就是蛋白质的空间构象，这种排布在一定程度上可以决定蛋白质的功能作用。由于蛋白质序列上的各个氨基酸连接起来的肽链的不断运动，使得蛋白质链的构象有很多，这就形成了蛋白质的二级结构，包括：α-螺旋，β-转角以及β-折叠和无规卷曲。蛋白质二级结构中，α-螺旋，β-折叠，分别用H和E来描述，其余的无规则卷曲部分用C来描述。这样，一个蛋白质二级结构序列由3个抽象字符构成蛋白质主要二级结构 蛋白质二级结构是氨基 酸残基在蛋白质多肽链中的局部空间构象，其具有 8 种类型， 分别是 α-螺旋（H）、β-桥（B）、折叠（E）、螺旋-3（G）、 螺旋-5（I）、转角（T）、卷曲（S）和环（L）。 把八类二级结构粗略地归类为螺旋、折叠和卷曲三类。

## 数据平衡化处理

金属锌离子与蛋白质结合位点预测是一个二分类问题，实际上真正结合的作用位点相对于非结合位点数量非常少。锌离子结合蛋白质作用位点预测和蛋白质-蛋白质作用位点预测相似，都是一个典型的非平衡分类问题。经过清洗处理后的数据集非结合位点数量13976个，结合位点数量1958个。对数据进行合理的采样处理能够避免传统机器学习方法对非平衡数据集分类的偏向性，更好地提高非平衡数据分类的预测准确性。

目前采样方法（SamplingMethods）分为两类：欠采样（UnderSampling）和过采样（OverSampling）。欠采样是从多数类样本中按照某种规则选择少量样本，再合并原有少数类样本作为新的训练数据集。过采样与欠采样相反，过采样按照某种规则对少数类的样本进行差值来增加新的样本，再合并原有多数类样本作为新的训练数据集。采样方法分类如图

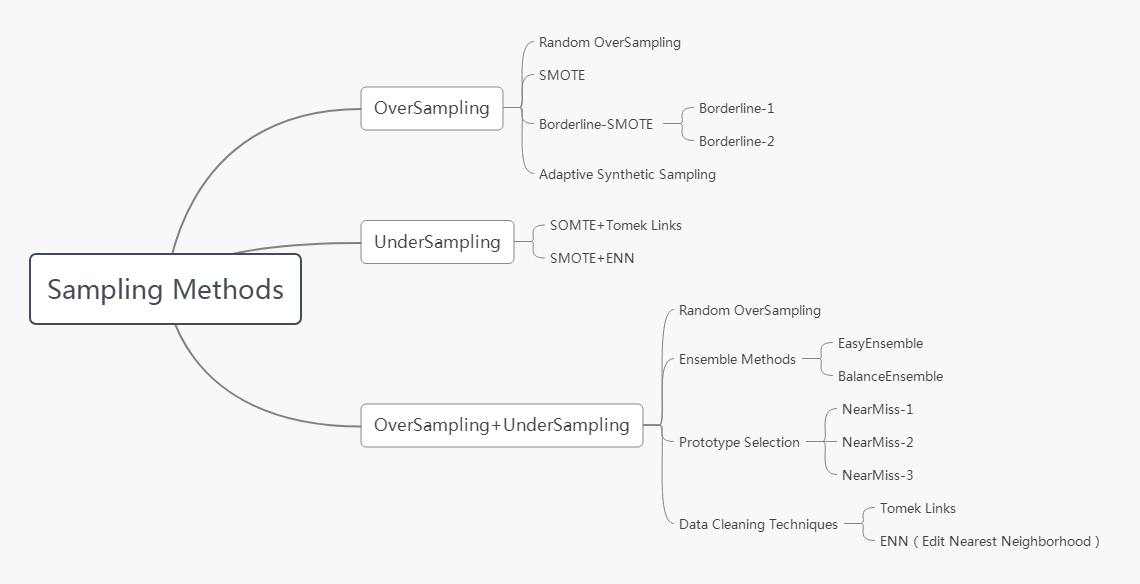


图 3.3 采样方法分类

### 欠采样

欠采样是从多数类样本中按照某种规则选择少量样本，再合并原有少数类样本作为新的训练数据集。本次课题实验中采用的是随机欠采样（RandomUnderSampling）来减少多数类的样本数量。如果随机删除的数量过大，那么合成的训练集丢失了某些数据信息，最终会使得模型只学习到总体模型的一部分特征。

### 过采样

过采样按照某种规则对少数类的样本进行差值来增加新的样本，再合并原有多数类样本作为新的训练数据集。本次课题实验中采用的自适应综合过采样（ADASYN）来增加少数类的样本数量。自适应综合过采样在于它为“难以学习”的实例创建更多数据的自适应性，并允许您对机器学习模型进行更多负样本采样，最终可以综合平衡数据。

# **基于密集连接网络的金属离子锌结合蛋白质作用位点预测方法**

## 数据集预处理

目前经过清洗处理后的锌离子与蛋白质作用数据集非结合位点数量13976个，结合位点数量1958个，数据集存在严重的正负样本数据分布不平衡问题。正负样本分布如图4.1.1所示。实验使用随机欠采样算法（RandomUnderSampling）对正样本即非结合位点样本进行随机删除50%的样本数据，负样本数据不做任何处理，将正负样本的数量调整到6988：1958；然后使用自适应综合过采样算法（ADASYN）对负样本即结合位点样本进行自适应插值扩充，正样本数据不做任何处理，将正负样本的数量调整到6988: 6871。正负样本分布如图4.1.2所示。

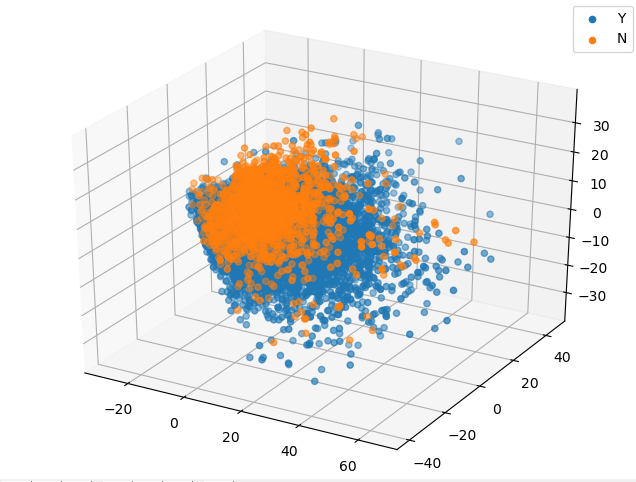


图 4.1.1 预处理前正负样本数据分布

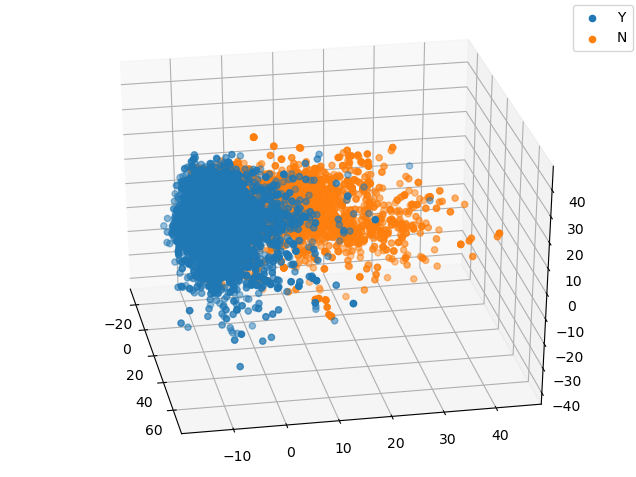
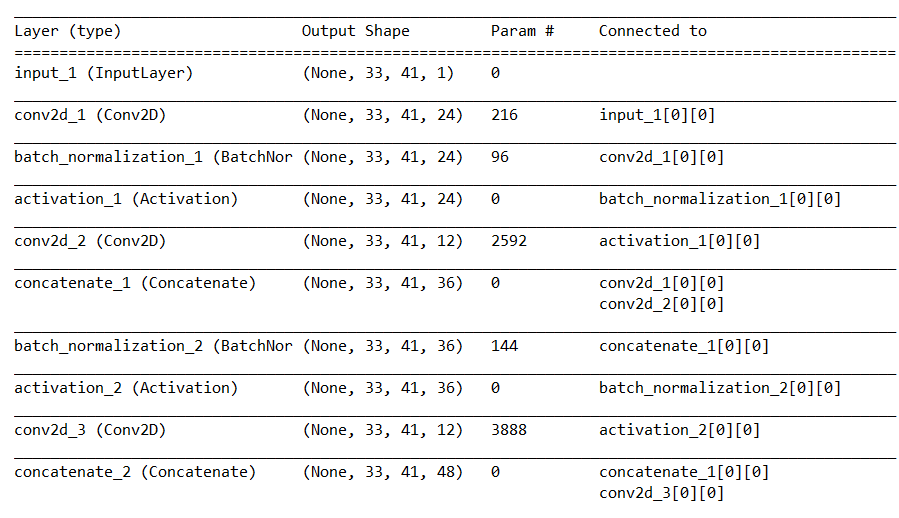
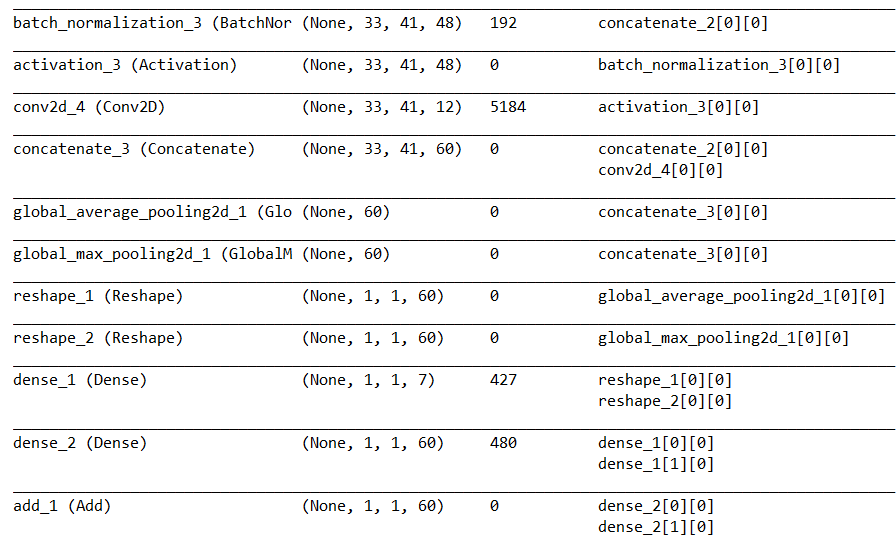


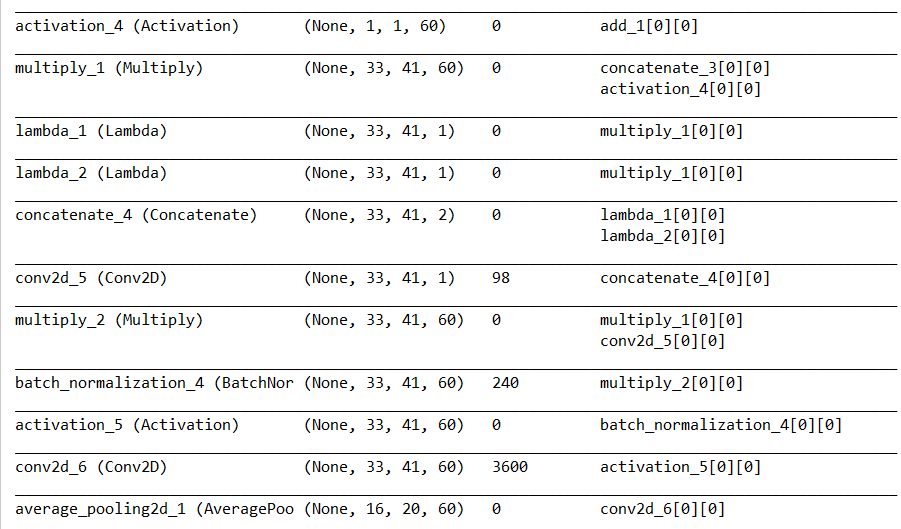
图 4.1.2 预处理后正负样本数据分布

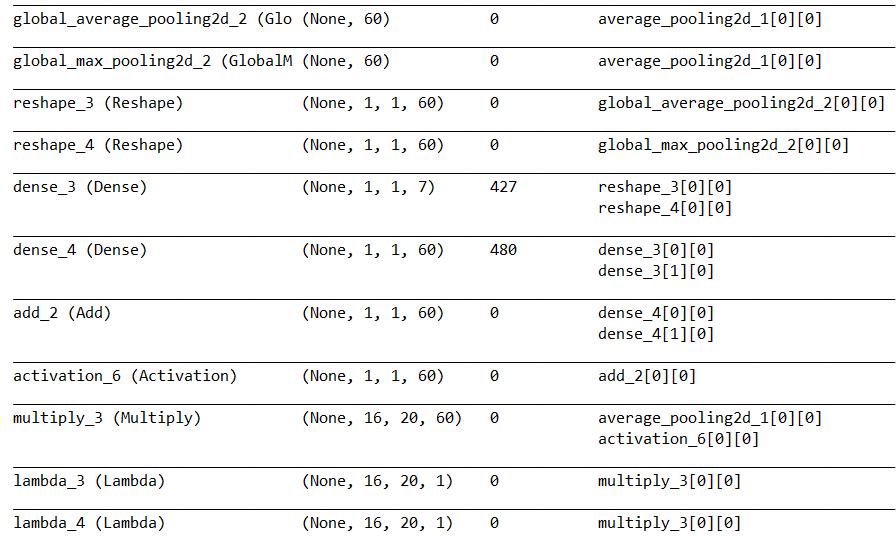
## 网络模型结构

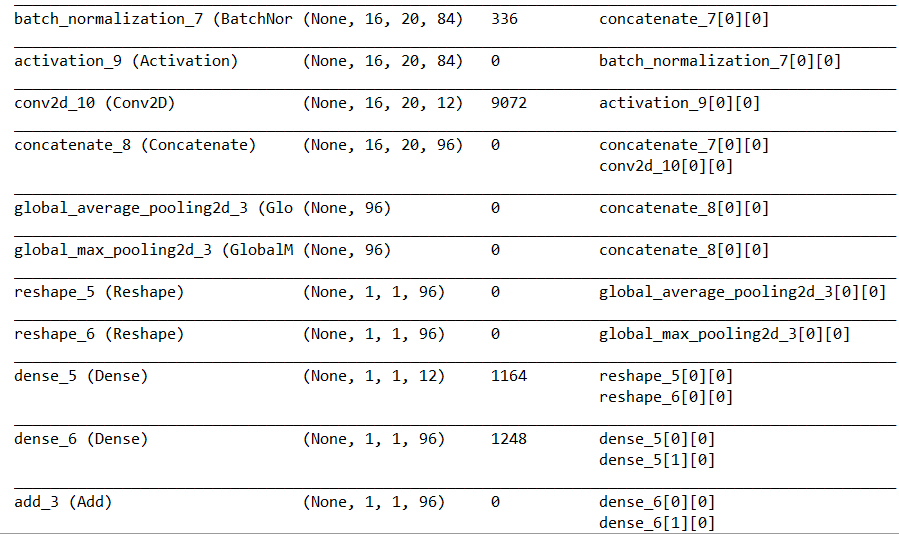
本次课题实验选用的是密集连接网络（Densenet），密集连接网络脱离了加深网络层数(ResNet)和加宽网络结构(Inception)来提升网络性能的定式思维,从特征的角度考虑,通过特征重用和旁路(Bypass)设置,既大幅度减少了网络的参数量,又在一定程度上缓解了gradient vanishing问题的产生。网络模型结构如图4.2.1-所示：

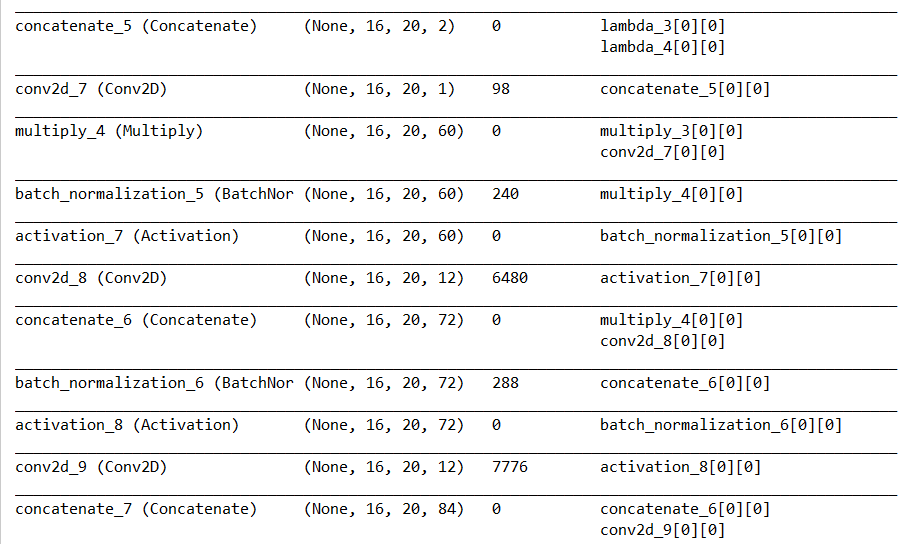


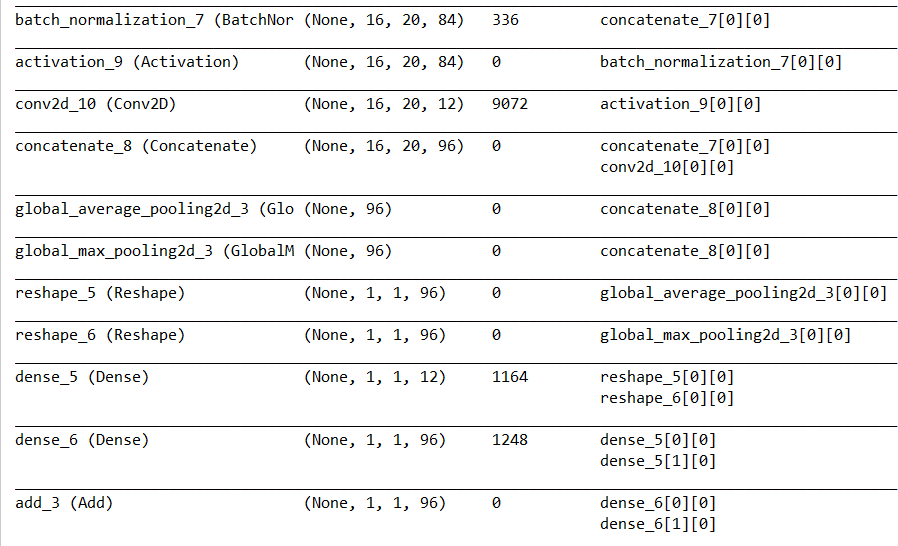


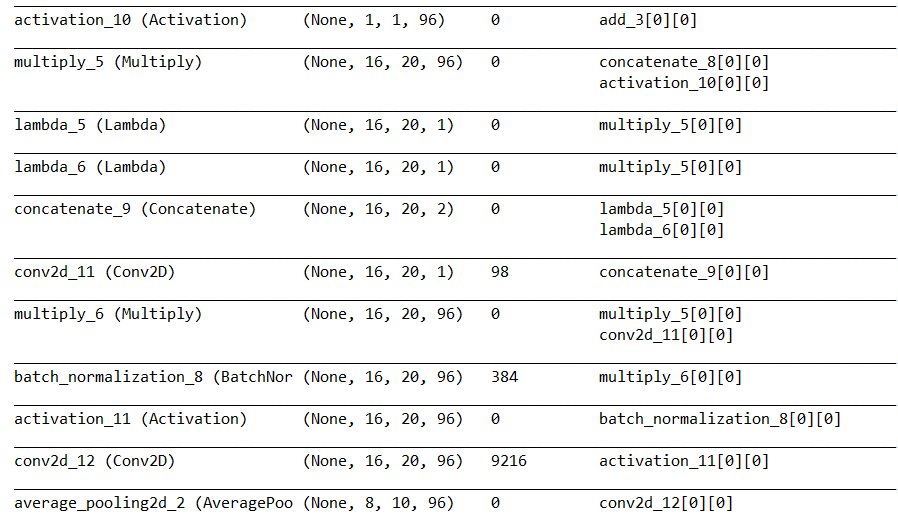


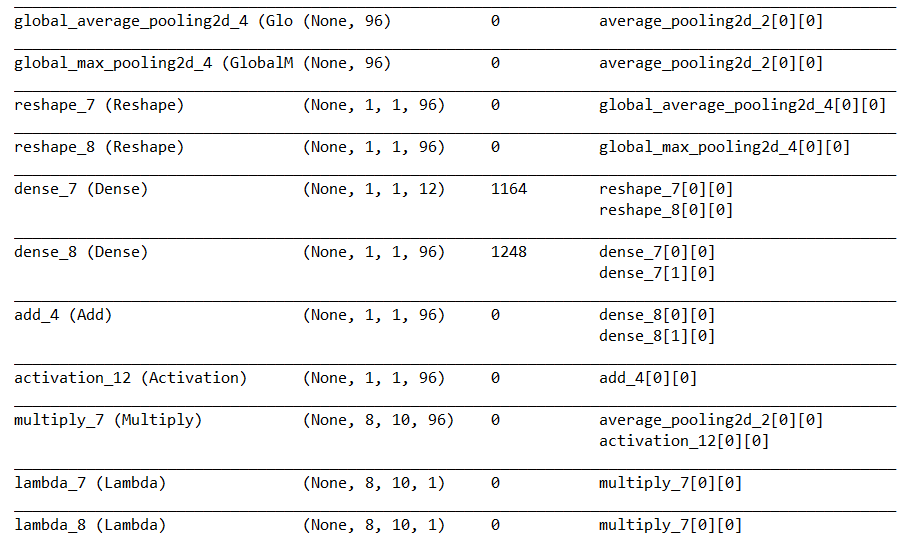


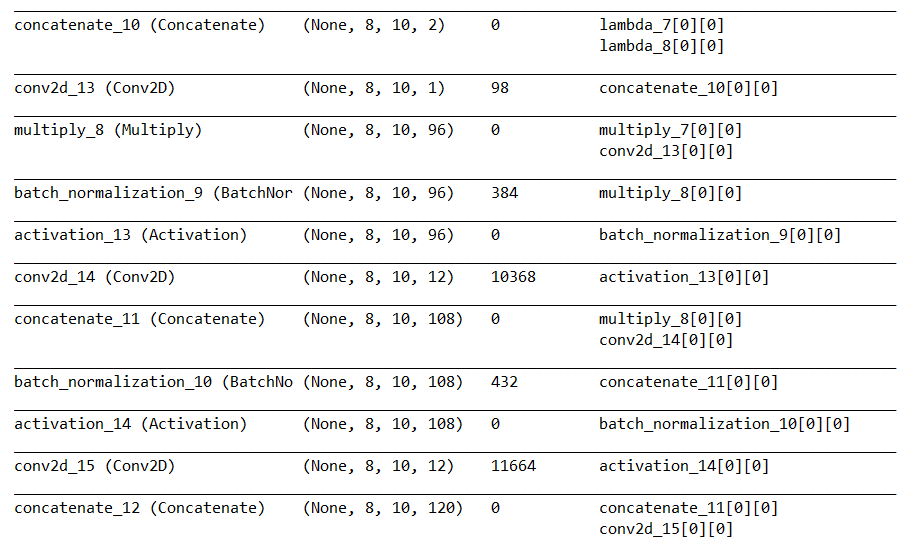


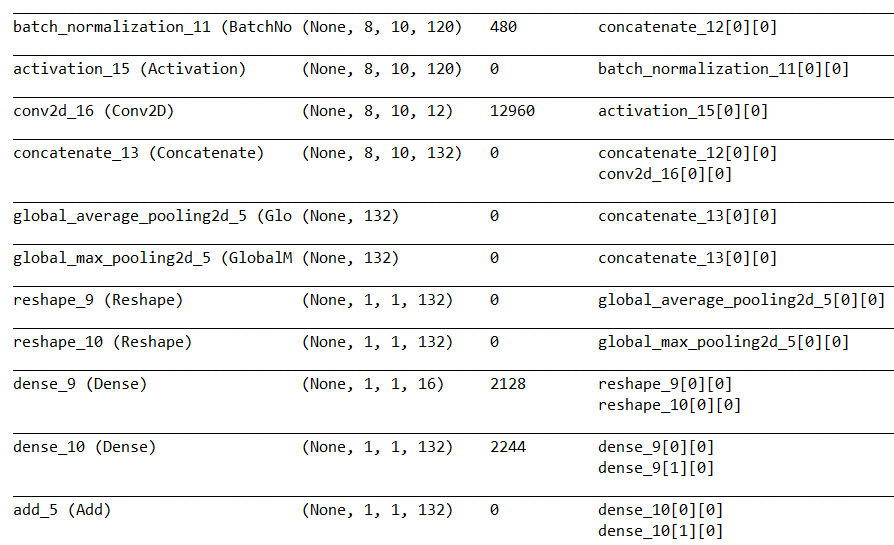


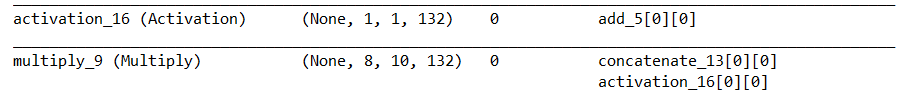


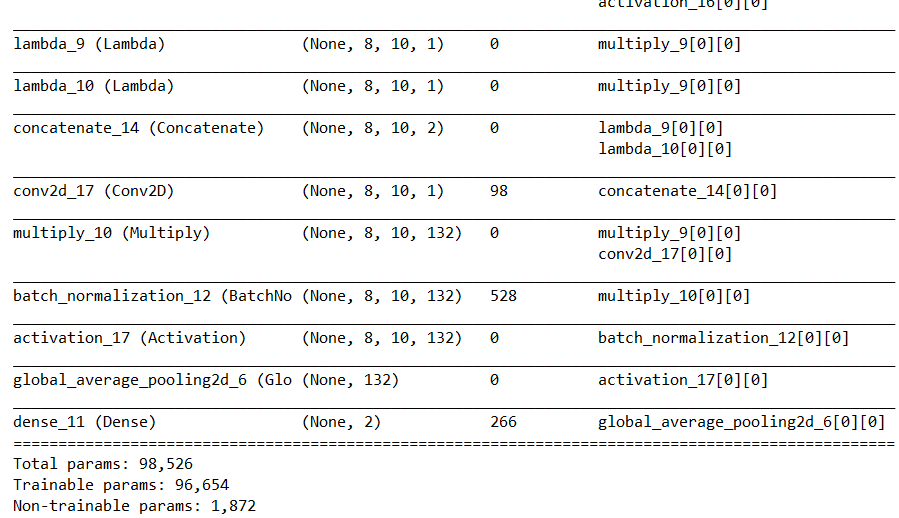












## 性能评估

## 本章小结

# 基于支持向量机的金属离子锌结合蛋白质作用位点预测方法

数据平衡化处理

基于

# 基于密集连接网络与支持向量机模型融合的金属离子锌结合蛋白质作用位点预测方法

# 结 论

×××××××××××（小4号宋体，1.5倍行距）×××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××…………

# 致 谢

×××××××××（小4号宋体，1.5倍行距）×××××××××××××××××××××…………

# **参 考 文 献**

1. Alberts B (1989) Molecular biology of the cell: Garland Pub.
2. Gallet X, Charloteaux B, Thomas A, Brasseur R (2000) A fast method to predict protein interaction sites from sequences. J Mol Biol 302: 917–926. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/10993732)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=J+Mol+Biol&title=A+fast+method+to+predict+protein+interaction+sites+from+sequences&author=X+Gallet&author=B+Charloteaux&author=A+Thomas&author=R+Brasseur&volume=302&publication_year=2000&pages=917-926&pmid=10993732&)]
3. Ofran Y, Rost B (2003) Predicted protein-protein interaction sites from local sequence information. FEBS Lett 544: 236–239. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/12782323)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=FEBS+Lett&title=Predicted+protein-protein+interaction+sites+from+local+sequence+information&author=Y+Ofran&author=B+Rost&volume=544&publication_year=2003&pages=236-239&pmid=12782323&)]
4. Wang B, Chen P, Huang DS, Li JJ, Lok TM, et al. (2006) Predicting protein interaction sites from residue spatial sequence profile and evolution rate. FEBS Lett 580: 380–384. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/16376878)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=FEBS+Lett&title=Predicting+protein+interaction+sites+from+residue+spatial+sequence+profile+and+evolution+rate&author=B+Wang&author=P+Chen&author=DS+Huang&author=JJ+Li&author=TM+Lok&volume=580&publication_year=2006&pages=380-384&pmid=16376878&)]
5. Zhou HX, Shan Y (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 44: 336–343. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/11455607)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Proteins&title=Prediction+of+protein+interaction+sites+from+sequence+profile+and+residue+neighbor+list&author=HX+Zhou&author=Y+Shan&volume=44&publication_year=2001&pages=336-343&pmid=11455607&)]
6. Aytuna AS, Gursoy A, Keskin O (2005) Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. Bioinformatics 21: 2850–2855. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/15855251)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Bioinformatics&title=Prediction+of+protein-protein+interactions+by+combining+structure+and+sequence+conservation+in+protein+interfaces&author=AS+Aytuna&author=A+Gursoy&author=O+Keskin&volume=21&publication_year=2005&pages=2850-2855&pmid=15855251&)]
7. Sikic M, Tomic S, Vlahovicek K (2009) Prediction of protein-protein interaction sites in sequences and 3D structures by random forests. PLoS Comput Biol 5: e1000278. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2621338/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/19180183)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=PLoS+Comput+Biol&title=Prediction+of+protein-protein+interaction+sites+in+sequences+and+3D+structures+by+random+forests&author=M+Sikic&author=S+Tomic&author=K+Vlahovicek&volume=5&publication_year=2009&pages=e1000278&pmid=19180183&)]
8. Bradford JR, Westhead DR (2005) Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics 21: 1487–1494. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/15613384)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Bioinformatics&title=Improved+prediction+of+protein-protein+binding+sites+using+a+support+vector+machines+approach&author=JR+Bradford&author=DR+Westhead&volume=21&publication_year=2005&pages=1487-1494&pmid=15613384&)]
9. Bradford JR, Needham CJ, Bulpitt AJ, Westhead DR (2006) Insights into protein-protein interfaces using a Bayesian network prediction method. J Mol Biol 362: 365–386. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/16919296)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=J+Mol+Biol&title=Insights+into+protein-protein+interfaces+using+a+Bayesian+network+prediction+method&author=JR+Bradford&author=CJ+Needham&author=AJ+Bulpitt&author=DR+Westhead&volume=362&publication_year=2006&pages=365-386&pmid=16919296&)]
10. Bernardes JS, Fernandez JH, Vasconcelos AT (2008) Structural descriptor database: a new tool for sequence-based functional site prediction. BMC Bioinformatics 9: 492. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2612011/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/19032768)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=BMC+Bioinformatics&title=Structural+descriptor+database:+a+new+tool+for+sequence-based+functional+site+prediction&author=JS+Bernardes&author=JH+Fernandez&author=AT+Vasconcelos&volume=9&publication_year=2008&pages=492&pmid=19032768&)]
11. Li N, Sun Z, Jiang F (2008) Prediction of protein-protein binding site by using core interface residue and support vector machine. BMC Bioinformatics 9: 553. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2627892/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/19102736)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=BMC+Bioinformatics&title=Prediction+of+protein-protein+binding+site+by+using+core+interface+residue+and+support+vector+machine&author=N+Li&author=Z+Sun&author=F+Jiang&volume=9&publication_year=2008&pages=553&pmid=19102736&)]

1. [Jessica C. Ebert](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ebert%20JC%5BAuthor%5D&cauthor=true&cauthor_uid=18042678) and [Russ B. Altman](https://www.ncbi.nlm.nih.gov/pubmed/?term=Altman%20RB%5BAuthor%5D&cauthor=true&cauthor_uid=18042678) (2008) Robust recognition of zinc binding sites in proteins
2. 王菲露.基于机器学习方法的蛋白质相互作用位点及二级结构预测出[D].安徽大学. 2009.5
3. 熊大鹏.基于深度学习架构的蛋白质远程残基接触预测研究[D].生命科学学院. 2017.5
4. 李慧.蛋白质功能位点预测方法研究[D] .南京航天航空大学. 2018.10
5. 王多林.基于深度学习的蛋白质翻译后修饰位点预测研究[D].吉林大学. 2018.6.5
6. 陈震.基于序列信息的蛋白质功能位点预测的算法开发[D].中国农业大学. 2014.6
7. Haoyang Zeng, Matthew D. Edwards, Ge Liu and David K. Gifford.Convolutional neural network architectures for predicting DNAprotein binding[D].MIT.June.2016
8. Yifeng Cui, Qiwen Dong, Daocheng Hong & Xikun Wang (2019) Predicting protein-ligand binding residues with deep convolutional neural networks
9. Shweta Yadav, Asif Ekbal, Sriparna Saha, Ankit Kumar, Pushpak Bhattacharyya (2018) Feature assisted stacked attentive shortest dependency path based Bi-LSTM model for protein–protein interaction.

1. [Martin T. Hagan](https://www.semanticscholar.org/author/Martin-T.-Hagan/144250204), [Howard B. Demuth](https://www.semanticscholar.org/author/Howard-B.-Demuth/51078936), [Mark Beale](https://www.semanticscholar.org/author/Mark-Beale/92665081)（1995）Neural network design