新疆大学本科毕业论文(设计)

新疆大学
Xinjiang University

论文题目:	基于卷积神经网络预测流感病毒抗原关系
学生姓名:	宋文凯
学　　号:	20172501244
所属院系:	软件学院
专　　业:	软件工程
班　　级:	软件 2017-15
指导老师:	吕小毅
日　　期:	2021 年 5 月 16 日

摘　要

流感病毒的传播有着季节性的特征，每年造成全球数百万人感染以及数十万人死亡。目前防控流感最经济有效的手段是接种疫苗，其促使免疫系统产生对应抗体，从而保护宿主免受病毒感染。而疫苗株和流行株的抗原关系将直接决定是否需要更新疫苗。如果能通过计算机，实现快速预测两毒株抗原关系，则对实际流感防控工作有指导意义。

当前流感病毒抗原关系预测大多基于传统机器学习方法，需要依赖大量领域知识。本模型采用卷积神经网络，使用自 1968 年以来收集的共 22401 条 H3N2 病毒序列对数据，将长度为 329 的病毒 HA1 蛋白序列按照特定氨基酸理化性质和重要位点进行特征值映射处理，生成矩阵大小为 329*14 的输入矩阵，并进行模型训练。依托深度学习，本模型可以直接对原始数据进行特征提取，实现预测任务。相比于传统机器学习方法，本模型显著的提升了预测的准确度，有效的降低了过拟合的缺陷。

本模型结合不断增长的算力，能够有效的预测不同的病毒对之间的抗原关系，并且测试集中敏感性和特异性指标数值能达到 0.98，表现较为出色。因此本模型将对疫苗株的筛选工作有着重大的帮助。同时能够有效避免传统 HI 滴定实验的复杂、数据模糊和生化危险的问题，从而减少疫苗生产所需要的人物力消耗。

关键词：流感病毒；抗原关系预测；卷积神经网络；氨基酸理化性

ABSTRACT

The spread of influenza virus has seasonal characteristics, causing millions of infections and hundreds of thousands of deaths worldwide every year. At present, the most economical and effective way to prevent and control influenza is vaccination, which promotes the immune system to produce corresponding antibodies, thereby protecting the host from virus infection. The antigenic relationship between vaccine strains and epidemic strains will directly determine whether the vaccine needs to be updated. If a computer can be used to quickly predict the antigenic relationship between the two strains, it will be of guiding significance for the actual influenza prevention and control work.

Most of the current influenza virus antigen relationship predictions are based on traditional machine learning methods, which require a large amount of domain knowledge. This model uses a convolutional neural network and uses a total of 22,401 H3N2 virus sequence pair data collected since 1968. The virus HA1 protein sequence with a length of 329 is mapped according to specific amino acid physical and chemical properties and important sites to generate a matrix. The input matrix is 329*14, and the model is trained. Relying on deep learning, this model can directly extract features from raw data to achieve prediction tasks. Compared with traditional machine learning methods, this model significantly improves the accuracy of prediction and effectively reduces the defects of over-fitting.

This model combined with the ever-increasing computing power can effectively predict the antigenic relationship between different virus pairs, and the sensitivity and specificity index value of the test concentration can reach 0.98, which is relatively good. Therefore, this model will be of great help to the screening of vaccine strains. At the same time, it can effectively avoid the complexity of traditional HI titration experiments, obscure data and biochemical hazards, thereby reducing the human

effort required for vaccine production.

KEY WORDS：Influenza virus; antigen relationship prediction; convolution neural network; amino acid physical and chemical properties

目 录

1 绪论

二十世纪起，全球共发生了三次大流感，造成数千万人死亡，其中季节性流感每年便可造成超数十万人死亡[1]。如何防控季节性流感成为医学界的焦点问题，目前预防季节流感最经济有效的方法是接种对应的疫苗。世界卫生组织（World Health Organization，WHO）每年持续监测全球流感活动，同时收集病毒 HA 抗原的变化情况，以数据为基础分别在每年二月份和十月份对北半球和南半球给出下一个流感季使用的疫苗株。由于流感病毒属于 RNA 病毒，其具有高度的变异性[2]，因此如何快速的判定疫苗株与当前流行株是抗原相似还是抗原变异，将是评价疫苗株防护效果的重要指标。

1.1 选题背景

季节性流感病毒是一种能够引起人类急性呼吸道传染病的病原体，并对人类的健康产生持续的威胁。单股负链 RNA 是流感病毒的遗传物质，其结构大都为球形，自内而外依次为核心、衣壳和包膜。其中包膜镶嵌着两种表面抗原——血凝素（HA）和神经氨酸酶（NA）。病毒主要通过表面的 HA 蛋白与宿主细胞的唾液酸受体结合引起宿主感染，并进入宿主细胞进行繁殖扩增。而表面抗原是最容易引起人体免疫系统反应的抗原，免疫系统通过产生特定抗体与病毒的表面抗原结合，致使病毒失去感染能力[3]。

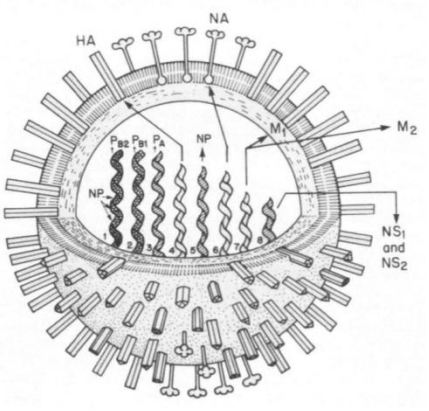

图 1-1 流感病毒结构图

其中 HA 和 NA 是季节性流感病毒的两个表面抗原，人们通常根据这两种抗原抗原性的不同，将其划分为不同的亚型。在感染人类的病毒中，HA 主要分为 H1、H2 和 H3，NA 主要分为 N1 和 N2。如图 1-1，流感病毒的 RNA 多聚酶 PB2、PB1、PA、血凝素（HA），核蛋白（NP）、神经氨酸酶（NA）均由第 1-6 基因节段编码。第 7、8 基因节段则分别编码包膜蛋白和非结构蛋白（NS）。可以看出流感病毒结构非常简单精巧，其增殖过程是，流感病毒复制各基因节段后，组装入子代病毒体中，完成增殖。但此过程极易发生基因重组或基因突变，进而导致新毒株的出现，这是流感病毒容易发生变异而出现大流行的原因[2]。

能刺激免疫系统产生特异抗体或致敏淋巴细胞的能力被称为抗原性，流感病毒的变异最终将会反应在抗原性的变化上，这将与其流行性密切相关，若抗原性变化小，免疫系统依旧能够对其产生大概率免疫，而某季节流感大流行，往往是由于抗原性的突变。抗原变异可分为抗原漂移（antigenic drift）和抗原性转变（antigenic shift）。抗原漂移，变异程度较小，一般是由核苷酸序列点的突变，进而导致 NA 和 HA 的抗原表位发生某些显著改变或由于人群体免疫选择性不同造成的。而抗原性转变，属于质变范畴，造成当前流行住与之前流行株完全失去联系。这种抗原性的转变会使人群原有的特异性免疫力失效，因此可引起世界性的流感流行。

1.1.1 血凝素 HA

血凝素（hemagglutinin，HA）是感染人类的病毒中最主要的中和抗原，空间结构表征为一条三聚体，如图 1-2 HA 蛋白结构所示。其由 3 条相同的糖基化多肽聚合而成，每条多肽又分为 HA1 和 HA2 两部分。在病毒感染的初始阶段其与神经氨酸受体结合，使病毒 RNA 及相关酶进入细胞。HA 抗体能够抑制病毒从感染细胞中释放出来，是重要的保护型抗体。但 HA 的抗原结构极易发生改变，导致抗原性变化。同时从基因层面，HA 的变异程度更快，病毒通过不断地位点突变，逃避人体免疫系统的攻击。抗原变化的本质是氨基酸位点变化和结构变化，导致抗原抗体不能结合。因此对流感病毒的探究主要集中在 HA 抗原的研究上。

由于 HA1 多肽链位于上方且含有更多的抗原表位，因此 HA1 是免疫系统识别以及抗体结合的重要区域。此次流感病毒抗原性的探究，将以 H3N2 病毒的 HA 蛋白中 HA1 部分为研究对象，探究其抗原性与其蛋白质序列间的关系。

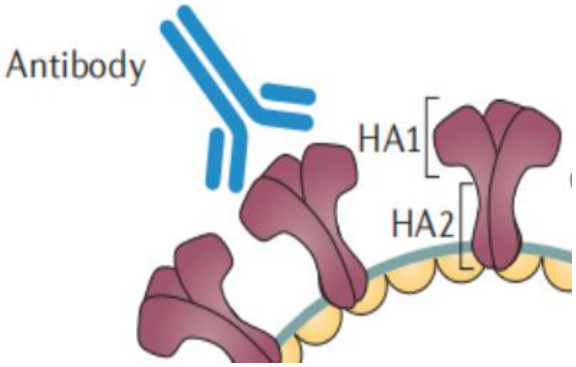

图 1-2 HA 蛋白结构图

1.1.2 表位

表位，位于抗原表面，决定抗原特异性的化学基因团又称抗原决定簇，是抗体与抗原的结合位点。单个抗原分子可以拥有多个不同的表位，每个表位都会因为氨基酸组合顺序，蛋白质空间结构以及氨基酸性质的影响而不同，具有抗原特异性。表位又包含线性表位和构象表位，其中线性表位是由抗原一级序列决定的，即蛋白质序列，构象表位是由多肽折叠形成的空间结构，可由空间内的氨基酸理化性质以及其他区域的空间属性影响。因此对 HA 抗原变化可以视为抗原表位的变化，包含氨基酸的变化，微环境的变化以及 3D 空间结构的变化[2]。

1.2 研究意义

流感的防控作为全球公共卫生的一项重大任务，当前最有效的防护措施是疫苗接种[4]。通过接种疫苗，使宿主体内产生具有保护作用的抗体。因为流感具有不断地进化现象，如果流行株和促使免疫系统产生特定保护抗体的疫苗株之间的抗原距离相近，产生的抗体便可与抗原结合，使宿主免受毒株感染。如果疫苗株和当前流行株的抗原距离远，疫苗的保护效力将会降低，不能产生很好的保护作用。因此实时监测流感病毒抗原的变化情况，并决定是否需要更新疫苗以及衡量疫苗效果[5]，都是十分重要的。

预测流感病毒对之间的抗原关系是疫苗株推荐过程中最重要的一步。我们可以根据推荐疫苗株与上一个流感流行季使用的疫苗株的抗原关系决定是否需要更新疫苗，从而减少在疫苗生产上所消耗的人力物力消耗。同时我们可以根据抗

原距离来评测当前疫苗的效果。目前的抗原距离D_{ij}的测算主要通过血凝抑制（hemagglutination inhibition，HI）实验进行。实验方法需要培养特定的病毒同时制备血清，因此 HI 滴定实验具有以下缺点[6]：

（1）准确度和灵敏度不高，将绝大部分病毒株都归为相似株。

（2）实验结果难以量化，由于需要大量的培养和制备，不同实验室得出的结果相差较大，甚至同一实验室的不同批次实验也有较大的差异。

（3）实验费时费力且有一定的生化危险，由于病毒传染性，因此在培养期间对实验环境要求较高，同时操作人员也有被感染的风险。

如今，随着第三代高通量测序技术的不断发展，同时流感病毒又是人类早期研究的病毒之一，病毒序列数据较多。其数据量已经初步满足大数据的要求。算力的增长又将深度学习带给生物信息界，病毒大数据已成为可能[7]。相对于传统的机器学习，深度学习能够通过特征学习和抽象特征，进行预测任务[8]。本论文提出的模型将有助于流感对康远距离的计算，帮助疫苗株的筛选和设计，预期能够降低流感疫苗开发的成本和风险，也能高效的评价疫苗效果[9]。

1.3 国内外研究现状

随着测序技术的进步，病毒序列的数据获取变得简单且准确。早期研究者如 Smith 在 2004 年从序列数据出发，构建抗原距离关系模型，成功将流感病毒按照抗原关系的远近划分为不同的簇。但其模型过于简单，单一使用序列数据构建模型，出现过拟合现象，只针对过去存在的数据有着较好的效果。而后出现了传统机器学习的计算方法，例如 Du[9]在五个抗原表位和传统的序列的基础上，又添加了氨基酸的理化性质作为特征，利用朴素贝叶斯模型进行构建。Wu 等设计了 H1N1 抗原距离计算模型，引入了空间距离这一概念，在空间结构上面获取了抗原的信息。目前基于序列扩充特征成为主流的抗原预测模型的出发点。

1.3.1 基于传统机器学习的抗原关系计算

传统的机器学习方法严重依赖大量的专业特征提取，这些工作将 HA 的氨基酸序列与抗原性有关的氨基酸属性相结合，将实验验证的抗原关系以及特征作为模型的训练数据，从而构建 HA 序列与抗原相似的关系模型。Wilson 等提出如果

抗原表位 HA 的氨基酸上，位点突变数目大于四，那么两病毒抗原属性就会发生变化[10]。Lee 等研究表明当表面抗原 HA 上的五个抗原表位，其中突变满足一定的数量关系时，病毒也会发生抗原性突变。Du 等开发了基于朴素贝叶斯的 H3N2 抗原关系预测模型，该模型整合四类导致抗原性改变的氨基酸特征[11]。每个抗原表位上的位点突变数目、氨基酸的物理化学性质、突变的氨基酸位点受空间影响异界糖基化位点的。Wu 等设计的 H1N1 抗原距离模型将抗原性和结构化特征变化结合，以量化抗原距离[12]。Kwoh 等提出了基于 AAIndex 数据库中的替换矩阵建立随机森林分类器[13]。

这类模型存在着三个主要的缺点：

（1）模型预测的准确性依赖于专业知识的积累而进行的特征提取，需要大量的先验特征，提取这些特征需要大量的专业知识。

（2）模型使用的是亚型特征，而亚型之间的特征是不同的，因此存在着适用性较小的情况，其只能针对某种特定的亚型进行相关预测分析，缺乏通用性。

（3）模型设计为线性模型，其计算模式过于简单，不能很好地表达序列到蛋白质这类复杂映射关系。

1.3.2 深度学习在抗原关系中的应用

深度学习在如今快速的发展，依托算力的提升，我们不仅能利用大量的特征数据进行复杂模型的构建，同时还能够进行模型迁移，使得模型能够运用到更多的相似场景中，能够有效的解决传统机器学习受领域知识限制以及较难跨亚型进行分析的两个缺点。同时由于流感病毒是人类长期以来一直关注的病毒，因此 H3N2 的相关性质和数据量较为充分，能够满足深度学习对数据量的要求。

目前抗原预测领域，深度学习的应用较少，是近些年的新趋势。例如 Yin 使用二维卷积神经网络，利用生物学的 ProtVec 编码方式对氨基酸序列进行编码后成功训练出跨不同亚型的 IAV-CNN 模型，同时 CNN 模块采用 SE 架构进行处理[14]。虽然模型的准确率较传统机器学习模型有了较大的准确度提高，但其依然是 90%左右的准确率。再例如周博，将 H3N2 抗原的氨基酸序列进行 one-hot 编码转换，作为神经网络的输入，使用 LSTM 和注意力机制的结合得到抗原性预测模型[15]。但其准确度仅有 80%左右。

上述两种模型存在着显著的缺点，仅仅考虑氨基酸序列作为输入，在序列上

增加结构方法，没有增添额外的理化性质。因此这也将成为本论文对其的改进点——增添额外的氨基酸理化性质以及 3D 结构特征。

1.4 本文工作

卷积神经网络能够对输入的矩阵继续训练，提取到相关的特征来进行预测。而输入矩阵质量的好坏将决定整个神经网络的质量（包括复杂程度和准确率），因此设计一个合理的，包含众多信息的输入矩阵将成为重要步骤。同时由于深度学习需要大量的数据，论文也收集了自 1968 年—2018 年的 H3N2 的 HA1 抗原蛋白序列数据作为数据来源。主要工作步骤如下：

（1）数据收集：分别从文献、世界卫生组织和各国疾病预防控制中心网站上收集 HI 实验数据和公共数据库收集抗原 HA 序列数据。

（2）数据清洗：将 HA 序列数据进行清理，处理包括同序列多名称、序列错误和序列异常等现象，同时统计序列对之间的序列差异 Seq-diff。

（3）数据规约：将清洗后的序列数据映射在 HI 滴定数据中，成为一一对应的数据对。通过 HI 滴定数据计算病毒毒株对的抗原距离D_{ij}，并且根据D_{ij}结果划分抗原相似或抗原变异，进行 0/1 化处理。

（4）序列特性信息计算：分别计算序列、利用空间结构计算的重要抗原位点、氢键供体、氢键受体、正电荷、负电荷和疏水性的特性值，并进行离散化和归一化处理。

（5）输入矩阵构建：根据序列和序列特性信息构建 329*14 的特征矩阵

（6）设计 CNN 模型：基于 SE 结构搭建 CNN 模型，据训练 CNN 模型并进行优化设计。

（7）使用不同工具进行对比实验，分析模型的优缺点。

2 甲型 H3N2 抗原特征表示

2.1 数据获取和预处理

2.1.1 数据获取与清洗

本项目需要使用两类 H3N2 数据：HI 实验数据和病毒抗原 HA1 序列数据。HI 实验数据来源于学术报告、科学文献、世界卫生组织流行病学周报、世界卫生组织流感参考和研究合作中心和各国疾病预防控制中心网站。病毒的抗原 HA1 序列数据来源于 Influenza Research Database、NCBI Influenza Virus Database 和 Global Initiative on Sharing All Influenza Data。

在文献中数据收集方面，主要在四篇论文的附录以及公开网站获取，如表 2-1 数据来源汇总所示。

表 2-1 数据来源汇总

数据来源	HI 滴定对数目	HA 序列数目	备注
CAO	3867	679	
LIAO	277	62	
PENG	791	621	不同簇为抗原变异，相同簇为抗原相似
WU	11 个抗原簇	195	
NCBI、PDB	4181	124	
数据合并	47633	791	
数据清洗	22401	791	序列差异<20

其中 Wu 的论文将大量的 H3N2 毒株按照抗原距离划分为不同的抗原簇，来探究不同簇间的进化位点。由于病毒的进化在生物学上呈现出，基因水平上的线性进化，抗原水平上的阶段进化呈现分簇现象[16]。因此不同簇间的毒株皆为抗原变异，相同簇间的毒株为抗原相似。将簇数据进行排列组合，能获得与 HI 滴定对效力相同的序列对。

由于数据来源较为广，通过数据分析，发现数据存在冗余和不匹配的问题，

包括：不同毒株名对应相同序列，造成的 HI 滴定数据重复、HI 滴定与 HA 序列不匹配和序列长度不统一等问题。因此本项目对数据的进行数据清洗，包含缺值处理、格式内容清洗、数据去重以及关联性验证。同时联系原文章作者，了解序列长度为 330 的原因，并将序列长度为 330 的进行经过数据去空位操作，还原为 329 长度的序列。通过初步梳理，共获得 22041 条含 HI 滴定数据的抗原对数据，以及 791 株毒株序列。

2.2.2 数据分析和预处理

对于 22041 条汇总数据两部分内容，第一部分为对应的两株病毒的序列，第二部分为毒株对间的抗原距离D_{ij}，其由 HI 滴定实验获取。

$$\text{Class}(\text{Pair1}, \text{Pair2}) = \begin{cases} 0 & D_{ij} \leq 4 \\ 1 & D_{ij} > 4 \end{cases} \quad (2\text{-}1)$$

本模型按照抗原距离常用的分类标准设定阈值如公式 2-1，对抗原相近或抗原相似进行分类。

表 2-2 序列差异分析

序列差异	Class=0	Class=1	序列差异	Class=0	Class=1
≤ 9	3047	913	≤ 15	4168	4467
>9	1314	16757	>15	193	13203
≤ 10	3370	1260	≤ 16	4218	5165
>10	991	16410	>16	143	12505
≤ 11	3657	1722	≤ 17	4244	5867
>11	704	15948	>17	117	11803
≤ 12	3858	2332	≤ 18	4276	6415
>12	503	15338	>18	85	11255
≤ 13	3994	3025	≤ 19	4298	6914
>13	367	14645	>19	63	10756
≤ 14	4094	3732	≤ 20	4313	7385
>14	267	113938	>20	48	10285

病毒序列对 Pair1 和 Pair2 长度均为 329，其中由 20 个氨基酸组成，由 20 个英文字母代替，缺损的氨基酸将使用‘-’表示。而序列差异（Seq-dff）指的是

逐位比较，依次遍历 329 个位点观察对应的两个位点是否相同，统计不同的个数。序列差异通常是传统机器学习的一个重要指标，特别是位于抗原表位上的差异。本模型中由于将两条序列均输入模型，因此序列差异和序列顺序将隐含在输入模型中。例如 IAV-CNN 中，使用 ProVec 编码其含义便是获取序列前后的信息，帮助模型更好的提取特征。而本模型选取构造长度为 329 的矩阵，也将位置信息保存在矩阵中，方便神经网络进行学习。

相关文献指出当两序列差异达到一定数值（例如大于 15）时，其抗原变异的可能性高达 90%[11]，因此为了保证训练样本的有效性，还需要对数据做进一步的分析，使数据正负样本数尽可能地保持 1：1，过滤出序列差异过大的序列。因此对序列差异进行统计如表 2-2 所示。

从上表可以看出，序列分布并不均匀，当序列差异小于 15 时，正负样本接近 1：1，同时序列差异大于 15 之后，只有 10%的数据属于抗原相近，因此序列差异是对数据正负样本分布的显著影响指标。在划分为训练集和验证集时，本模型将统计序列差异为定值时的数据，并进行 8-2 分类，使得同一序列差异下，80%的数据将作为训练集，20%的样本作为测试集。同时由于严格按照序列差异进行划分，因此训练集和测试集在统计学上具有相同的分布情况，能更好的模拟真实的数据情况，避免模型过拟合而出现错误。

2.2 模型特征选择

每种特定的模型在设计的初期，都将基于特定的方法进行相关的设计。目前为止大多数方法均基于序列的分析方法（Smith）[16]，利用序列间的潜在一维结构获得模型，根据氨基酸序列排列的差异预测抗原关系。该种在早期取得了一定的成果，但模型对新出现的流感序列预测能力较弱。而 HA 蛋白抗原的预测，归根结底属于蛋白层面的预测，不仅涉及一级结构，同时也涉及到二三级结构，才能决定一个蛋白质的性质。因此对于抗原的预测，还需要引入新的角度。

在现代生物学的视角下，如图 1-2，抗原与抗体的结合取决于表位和构象表位。表位是抗体结合的主要区域，抗原抗体能否结合，其本质上取决于此区域的理化性质。例如区域内的疏水性、极性、电荷量等，共同构成表位的微环境属性，影响抗原与抗体的结合。此类性质我们统称为非结构理化特征，抗原表位位点的

选取来自 Qiu 的文章[17]。

构象表位作为蛋白的 3D 结构特征也将影响抗原抗体的结合，例如在蛋白质的三级结构中，如果结构位置不能很好的匹配，影响抗体与抗原表位的接触。而在研究同一抗体是否能对不同抗原产生抑制效果时，也发现具有相似的 3D 结构区域，更容易发生免疫结合。因此将 3D 结构纳入抗原关系的特征选择，结合抗原表位筛选出重要位点。因此本模型将依据生物学知识，将单一的氨基酸序列增添额外的具有意义的生物学信息，作为特征输入。

2.3 非结构理化特征

20 种天然氨基酸作为蛋白质构成的基础，其自身的特性例如极性、电荷量等都将影响蛋白质的理化性质。Liao 通过分析氨基酸的 non-plar、polar、charged 等划分为不同的评分矩阵，对序列进行评分处理，之后进行回归分析，对于传统的单一序列分析，其显著的提高了模型的准确性[18]。同时其也探究了不同位点的氨基酸对抗原的影响关系，结果说明某些特定位置的氨基酸对抗原的差异有着重要影响。利用信息增益（IG）和熵（entropy）衡量 HA 上不同位点的得分，同样证明了特定位点对抗原变异的作用[19]。而且这类位点大都位于五个表位中，因此选择氨基酸理化性质和抗原表位上的位点作为非结构理化特征。

氨基酸的理化性质包含 200 多种，包括支链亲水性、分子量、等电点、羧基解离常数、酸碱滴定常数等。上述理化性质均属于氨基酸固有属性，因此可以通过查阅资料获取，本文采用 AAIndex（Amino Acid Index Database）作为数据来源。AAIndex 是有着氨基酸和氨基酸对的各种理化性质指数的数据库，其中又分别包含 AAIndex1、AAIndex2 和 AAIndex3 三种子库[20]，本次选用单氨基酸对应数据库 AAIndex1。其中的所有数据均来源于已出版的文献，AAIndex 可以通过 ftp 下载相关数据（ftp://ftp.genome.jp/pub/db/community/aaindex/）。

根据文献显示，氨基酸的 402 种指数，利用最小生成树构建后，性质具有明显的聚类特点[20]。因此无需选取过多的指数。选取以下五种性质，作为氨基酸的代表性质参与非结构理化性质的选择：氢键供体、氢键受体、正电荷、负电荷和疏水性。同时为了提高模型的收敛速度，收集数据后按照公式（2-2）进行归一化处理。

$$x_j^* = \frac{x_{ij} - x_i^{min}}{x_i^{\max} - x_i^{\min}} \quad (2\text{-}2)$$

氨基酸理化性质映射表所示。前四种属性，由于差距有明显的分类情况，因此在归一化的基础上，进行数据聚类处理，按照聚类情况划分为 0 和 1 做离散化处理，保证数据为无量纲表达式，便于不同理化性质间能够进行比较，同时也有利于神经网络的收敛速度，缩短训练过程。最后一种疏水性由于数据分布差距较小，因此仅使用归一化处理。最终结果如表 2-3 氨基酸理化性质映射表所示。同时上述选取的特征也围绕着蛋白质 3D 结构进行，例如疏水性在蛋白质二级结构预测中，发挥着重要作用[12]，不同的疏水性将影响蛋白质的二级和三级结构进一步的影响抗原与抗体的结合。而氢键供体和受体，则决定不同区域的吸引性，若区域内氢键供体数量较多，则能够很好的与氢键受体区域较多的氨基酸结合。因此在考虑众多能对蛋白质的结构产生影响的型之后，选取上述五种氨基酸理化性质，作为非结构化理化性质。

表 2-3 氨基酸理化性质映射表

Residue	hydrogen-bond Donor	hydrogen-bond Acceptor	Positive charge	Negative charge	Hydrophobicitya
A	0	0	0	0	0.616
R	1	0	1	0	0
N	1	1	0	0	0.236
D	1	1	0	1	0.028
C	0	0	0	0	0.68
Q	1	1	0	0	0.251
E	0	1	0	1	0.043
H	0	1	1	0	0.165
I	0	0	0	0	0.943
L	0	0	0	0	0.943
K	1	0	1	0	0.283
M	0	0	0	0	0.738
F	0	0	0	0	1
P	0	0	0	0	0.711
S	1	1	0	0	0.359
T	1	1	0	0	0.45
W	1	0	0	0	0.878
Y	1	1	0	0	0.88
V	0	0	0	0	0.825
G	0	0	0	0	0.501

2.4 空间结构特征

蛋白质的空间结构在蛋白质的功能上扮演着重要角色，并会直接或间接的影响抗原性。如图 2-1 HA 空间结构，由于构象表位本身就是由结构定义，因此位于抗原表位的氨基酸对空间造成的影响，会直接的改变构象表位的结构从而发生免疫逃离。而非表位的蛋白质区域中，某些蛋白质的改变可能会影响整体的结构，从而在结构上传递到抗原表位，间接的影响抗原的构象表位结构。因此探究空间结构特征能进一步的完善流感抗原预测模型。本模型将参考空间结构，初步筛选出结构位点，再结相关论文计算出的逃逸率等进一步筛选。

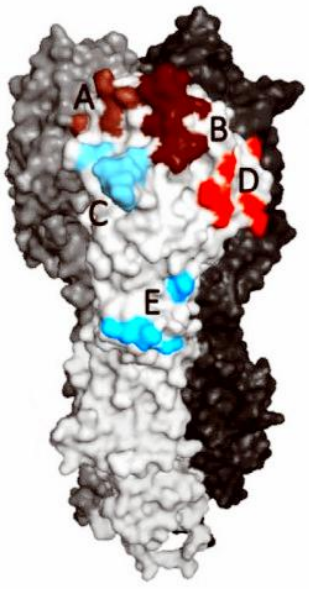

图 2-1 HA 空间结构

使用 PyMOL[23]进行 HA1 的空间结构建模，由于其为三聚体，因此图 2-1 仅展示其中一条上的抗原表位 A-E。将抗原表位映射在空间结构中时，能够观察到某些抗原空腔（epitope cavities），这些空腔将会显著影响抗原与抗体的结合，是重要的抗原备选位点[12]。依据结构信息计算相关的突出位点以及形成空腔的位点，再结合 Qiu 文章中计算的每个位点的突变率和逃离率，配合结构信息进一步筛选出潜在重要位点。其模型通过大量数据计算，测算空腔和突出备选位点的突变情况与位点位置之间的关系，得到逃逸率和突变率综合较高的位点[21]，作为模型重要位点的来源选择，放置在本模型中。最终得出如表 2-4 所示的 47 个抗原优势位点。

position 指代在序列长度为 329 的 HA1 序列种的位号，通过 position 进一步

映射对应的氨基酸信息。Cluster 表征此位点所在的表位，不同的表位在特定的情况下，突变效率不相同。Mutation rate 和 Escape rate 分别表示位点的突变概率，以及此处突变时发生的免疫逃逸概率（之前免疫系统的抗体不能识别突变后的抗原）。在五个抗原表位的所有候选位点中，最终筛选出 47 个位点，作为优势位点进行处理。

表 2-4 抗原优势位点

Position	Cluster	Mutation rate	Escape ratio	Position	Cluster	Mutation rate	Escape ratio
50	E	0.286	0.767	164	B	0.074	0.982
57	E	0.059	0.782	172	D	0.192	0.847
121	D	0.096	0.989	173	D	0.339	0.768
122	D	0.188	0.677	188	A	0.139	0.703
124	D	0.173	0.931	189	A	0.431	0.952
129	B	0.015	0.964	190	A	0.208	0.763
131	B	0.165	0.931	193	A	0.33	0.817
132	B	0.01	0.872	196	B	0.145	0.689
133	B	0.207	0.997	197	B	0.155	0.997
135	C	0.157	0.956	207	D	0.067	0.96
137	C	0.214	0.891	208	D	0.049	0.813
140	C	0.123	0.61	216	A	0.054	0.791
142	C	0.122	0.643	217	A	0.08	0.993
143	C	0.117	0.986	219	A	0.128	0.626
144	C	0.371	0.797	225	C	0.16	0.671
145	C	0.33	0.854	226	C	0.337	0.768
146	C	0.111	0.99	240	D	0.01	0.947
152	B	0.013	0.8	244	B	0.101	0.952
155	B	0.257	0.969	260	D	0.152	0.804
156	B	0.37	0.817	275	E	0.096	0.838
157	B	0.144	0.794	276	E	0.129	0.979
158	B	0.307	0.952	278	E	0.169	0.945
159	B	0.235	0.856	279	E	0.01	0.889
160	B	0.165	0.832				

3 基于 CNN 的抗原预测模型

3.1 卷积神经网络

卷积神经网络于二十世纪 80 至 90 年代提出，是一类包含卷积的前反馈神经网络，是深度学习领域内比较有代表性的算法之一。随着计算机性能的发展，深度学习再度繁荣发展，由于其能够按照阶层对信息进行平移，因此具有表征学习的能力，即能提取事物隐藏的信息之间的属性关系，被广泛应用在计算机视觉和自然语言处理领域[24]。

CNN 的结构包含输入层、隐藏层和输出层三部分。

输入层可以处理多维度的数据，并且需要预先设定估计维度和长度的矩阵，并且将数据进行标准化，把范围控制在固定的区间内将会有利于提升卷积神经网络的学习效率和表现。

$$Z^{l+1}(i,j) = \left[Z^{l} \otimes \omega^{l+1}\right](i,j) + b(i,j) \qquad (3\text{-}1)$$

隐藏层包含卷积层、池化层和全连接层三类常见的构架。如图 3-1 卷积示例，卷积层能够对输入的数据进行特征提取，每个元素对应一个权重系数和一个偏差量。计算公式如（3-1）所示

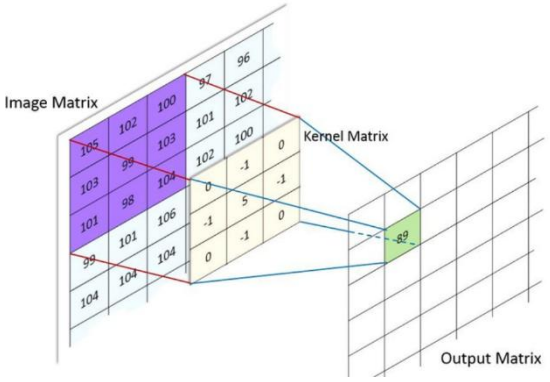

图 3-1 卷积示例

卷积的过程就是矩阵乘法的过程，通过不断地平移 Kernel 矩阵，将原始地输入矩阵计算成为不同的数值。期间需要利用激活函数和反馈神经网络调整 Kernel 的参数。池化层与卷积层类似，但它们采用的是降采样操作，加长滑动的

步长来降低特征的空间大小。

全连接神经网络是位于 CNN 隐藏层的最后部分，利用非线性组合来得到分类实现最终效果。不同于滑动卷积，其每一层都需要与上一层完全连接来实现高阶特征学习。

输出层为特定的结果，一般使用 Softmax 采用 0 或 1 表示。其中关于神经网络的训练和激活一般采用现成的方法，设定固定的 Learning-rate 进行调用，不作为本论文重点，故不做介绍。

3.2 模型设计与实现

CNN 网络的设计主要考虑三个部分

（1）输入矩阵的维度和大小，以及数据类型；

（2）神经网络构架结构；

（3）输出层大小设定；

本模块将完整介绍设计思路，代码实现将采用 Tensorflow1.7 版本进行[22]。

3.2.1 输入层和输出层设计

如图 3-2 所示，将 22041 对序列依次输入到不同的特征处理流程中。例如P_i序列对，当 329 位长度的氨基酸序列输入序列编码流中时，每一位都将获得自己独特的数值，最终组成P_{il}数组。多个编码流P_i序列对编码时，将会生成P_iM 矩阵，每个矩阵对应此序列对的特征属性，并且数值均在[0,1]之间。对于缺损的氨基酸，将对应的特征值设定为-1，保存为 float64 格式。同时根据深度学习模型训练的效果经验而言，稀疏矩阵的训练效果往往效果较差，因此在输入层编码时并未采用 one-hot 编码，而将 20 个氨基酸等分在[0,1]的区间内，避免周博模型中稀疏矩阵的问题[15]，从而加速模型的收敛，有效利用模型参数和数据。

在设计矩阵时，有着两种设计方案，feature_size=8 or 14

（1）采用传统模型的思路，以序列差作为矩阵输入，减少特征量，生成[22401,329,8]的矩阵,feature_size=8。

（2）将特征完整的映射进入矩阵，让模型自动提取矩阵信息，生成[22401,329,14]的矩阵,feature_size=14。

最终获得[22401,329,8]的矩阵，以 np.array 的形式保存为 npy 文件格式作为输入层读入数据。

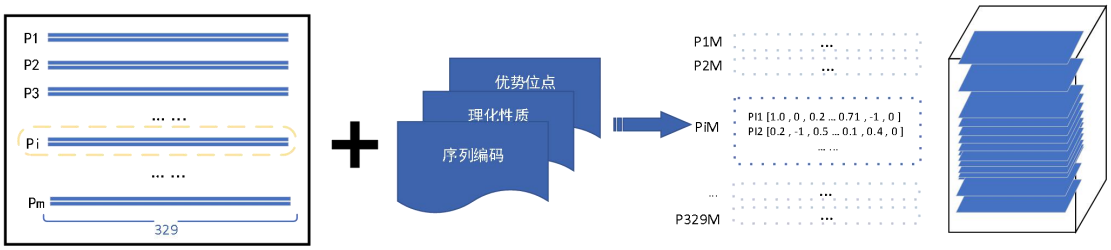

图 3-2 输入层编码处理

对于输出层的设计，由于样本标签为二分类问题，即判定相应的序列对是抗原相近还是抗原相似，因此我们将抗原相近定义为 0，抗原相近定义为 1，数据类型为 int32。定义输出层长度为 2，采用 Softmax 进行激活，则输出结果为 01 对应抗原相近，10 对应抗原相似。

3.2.2 网络结构设计

如图 3-3 神经网络设计示意，在第一个卷积层中，采用 SAME 模式，选取[3,feature_size]大小作为神经元大小，输出维度为 3*feature_size。以 feature_size=8 为例，每次神经网络向下滑动 1。对于单个神经元而言，输入[3,8]大小的数据，输出[1]，由于输出维度为 32 维度。每次滑动，都会产生 1*32 个输出值，因此隐藏层 1 设计为[3,8,32]。又因为使用 SAME 模式，最后两次滑动会复制最后一行结果进行输出。最终得到[329,32]大小的结果，进入隐藏层 2。卷积每次读取三行的策略是设计的重点，它能够保存序列之间的前后关系，方便提取特征。

隐藏层 2 的设计与隐藏层 1 类似，由于数据为[329,32]因此每个神经元大小取[3,32]，输出维度为 32，最终隐藏层设计为[3,32,32]。

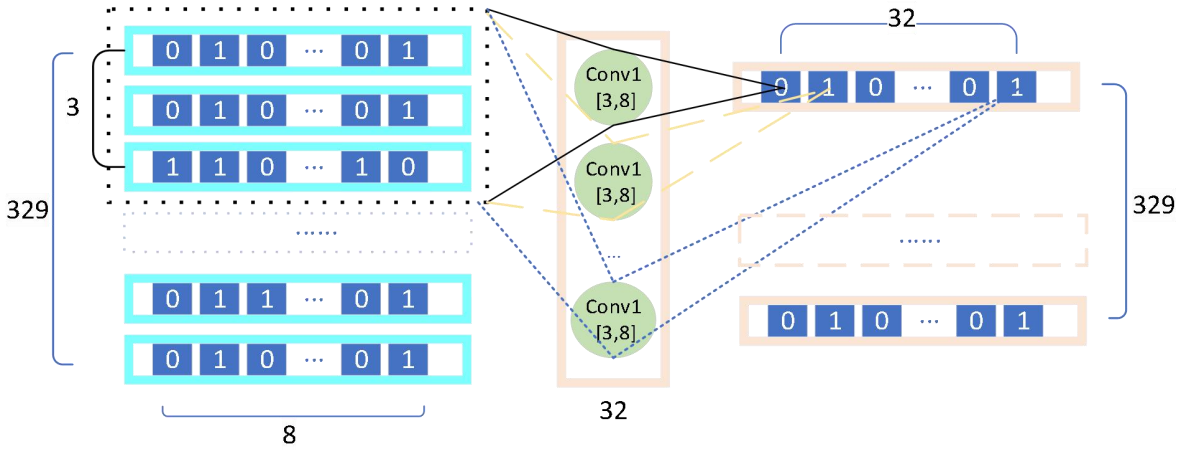

图 3-3 神经网络设计示例

随后进入一个池化层，由于池化层使得特征的行数减少一半，并且同样采用SAME 模式，因此池化层的输出结果为[165,32]。

池化层后，结果输入进全连接层，其本质上属于一维度神经网络，因此需要将二维数组降维为一维数据，因此第一层全连接层设计为[32*165,256]，第二层全连接层作为最后一层，[256,2]最终的结果将会以 one-hot 编码的形式输出。

最终代码如表 3-1 CNN 模型实现 所示下所示

表 3-1 CNN 模型实现

```
def model_function(data_batch, label_batch,num_neurons_fc1 ,learning_rate=0.001):
    #feature_size is a global data , be signed in function of dataset_input_fn()
    input_layer = tf.reshape(data_batch, [-1 , 329, feature_size])
    #the first convolutional layer
    W_conv1 = weight_variable([3,feature_size,32],name= 'W_conve1')
    b_conv1 = bias_variable([32],name = 'b_conve1')
    conv1_bn = tf.nn.conv1d(input_layer, W_conv1 ,stride=1 ,padding ='SAME')+b_conv1
    conv1 = tf.nn.relu(conv1_bn)
    # the pooling lay
    pool1 = tf.layers.max_pooling1d(conv1, pool_size = 2, strides= 2, padding='same')
    #the second convolutional layer
    W_conv2 = weight_variable([3,32,32],name= 'W_conve2')
    b_conv2 = bias_variable([32],name = 'b_conve2')
    conv2_bn = tf.nn.conv1d(pool1, W_conv2 ,stride=1 ,padding ='SAME')+b_conv2
    conv2 = tf.nn.relu(conv2_bn)
    # densely connected (fully connected) layer
    W_fc1 = weight_variable([32*(165), num_neurons_fc1], name = 'W_fc1')
    b_fc1 = bias_variable([num_neurons_fc1], name = 'b_fc1')
    pool4_flat = tf.reshape(conv2, [-1, 32*(165)])
    fc1_bn = tf.matmul(pool4_flat, W_fc1) + b_fc1
    fc1 = tf.nn.relu(fc1_bn)
    # readout layer or softmax regression
    W_fc2 = weight_variable([num_neurons_fc1, 2], name = 'W_fc2')
    b_fc2 = bias_variable([2], name = 'b_fc2')
    y_conv = tf.matmul(fc1, W_fc2) + b_fc2
    # calculate loss
    loss = tf.losses.softmax_cross_entropy(onehot_labels = label_batch, logits = y_conv)
    # calculate accuracy
    predictions = tf.argmax(y_conv, 1)
    actuals = tf.argmax(label_batch, 1)
    equality = tf.equal(predictions,actuals)
    accuracy = tf.reduce_mean(tf.cast(equality, tf.float32))
    train_op = tf.train.AdamOptimizer(learning_rate).minimize(loss)
    ones_like_actuals = tf.ones_like(actuals)
```

```
zeros_like_actuals = tf.zeros_like(actuals)
ones_like_predictions = tf.ones_like(predictions)
zeros_like_predictions = tf.zeros_like(predictions)
tp_op = tf.reduce_sum_tp(.)
tn_op = tf.reduce_sum_tn()
fp_op = tf.reduce_sum_fp()
fn_op = tf.reduce_sum_fn()
return accuracy, loss, train_op , tp_op, tn_op, fp_op, fn_op
```

表 3-1 中为本次模型的核心代码，采用 Tensorflow1.0 版本代码进行实现，代码中定义了神经网络的结构已经运算过程中所需要的评价函数，最终包含与测试和评价指标值，统一由 session 返回给主函数进行调用。同时 learning rate 和 batch_size 等可自由调节的参数，均以形式参数给出，方便调整模型参数。

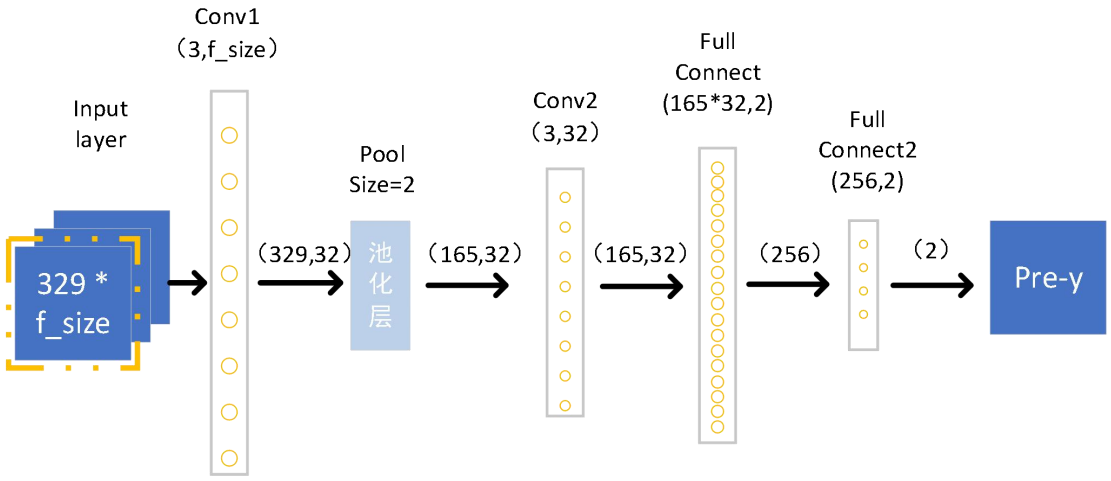

图 3-4 CNN 结构示意图

整体网络结构如图 3-4 CNN 结构示意图所示，模型支持 feature_size 的扩展，对于调整输入矩阵的大小有着重要的帮助。模型经过卷积层、池化层、卷积层和两个全连接层最终输入预测值 Pre-y。并且本模型在池化层后不同 feature_size 的结构相同，具有可对比性，能帮助我们更好的设计输入矩阵，搭配不同的特征模型。同时模型较为简单，有利于在当前数据量不够充分的情况下训练模型，理论上具有较好的收敛速度和训练效果。

4 实验结果与模型分析

4.1 模型参数和评价指标

经过预先的几次训练，目前采用如下参数：

模型对每组数据分组 batch_size 大小为 64，步长 rate 为 0.001，整体神经网络层次之间采用 padding 的 same 模式，epoch 轮数设定为 100 轮。

同时每次训练的结果进行相关保存，对模型的评价指标，本次采用五个指标，分别为 precision，accuracy，sensitivity，specificity 和 F1-score，其计算公式为

$$precision = \frac{TP}{TP+FP} \tag{4-1}$$

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{4-2}$$

$$sensitivity = \frac{TP}{TP+FN} \tag{4-3}$$

$$specificity = \frac{TN}{TN+FP} \tag{4-4}$$

$$F_1score = 2 * \frac{precision*sensitivity}{precision+sensitivity} \tag{4-5}$$

上述公式中 T 表示预测正确，F 表示预测错误，P 表示预测为 Postive，N 表示预测为 Negative。通过公式（4-1）、（4-2）、（4-3）、（4-4）和（4-5）我们能够很好的评价模型在不同情况下的效果，并且相比于单一的准确性，更为科学合理。依据上述指标我们对不同的 Seq-dff 分别进行一百轮训练，并保存结果。

4.2 模型结果分析

训练集和验证集的划分采用 80%和 20%的样本比例，并且为了保证数据相同的分布特性。根据上图 4-1 验证集结果可以看出，Feature8 的模型在 20 轮 epoch 后的各项评价指标均开始稳定，并且保证在 98%以上，说明我们的模型大大地提高了对于抗原对之间的关系，能够准确的预测抗原变异或相似，具有较强的准确性。而在 epoch58 时，模型出现了性能显著的下降，说明此时处于过拟合地状态。

因此后续使用时，在 Feature8 的情况下，尽量选取 40 轮的模型，作为最终的结果保存使用。

同时针对不同的输入矩阵，我们也使用 Feature14 作为输入，进行模型训练，训练结果如 4-2 Feature14 验证集结果所示。

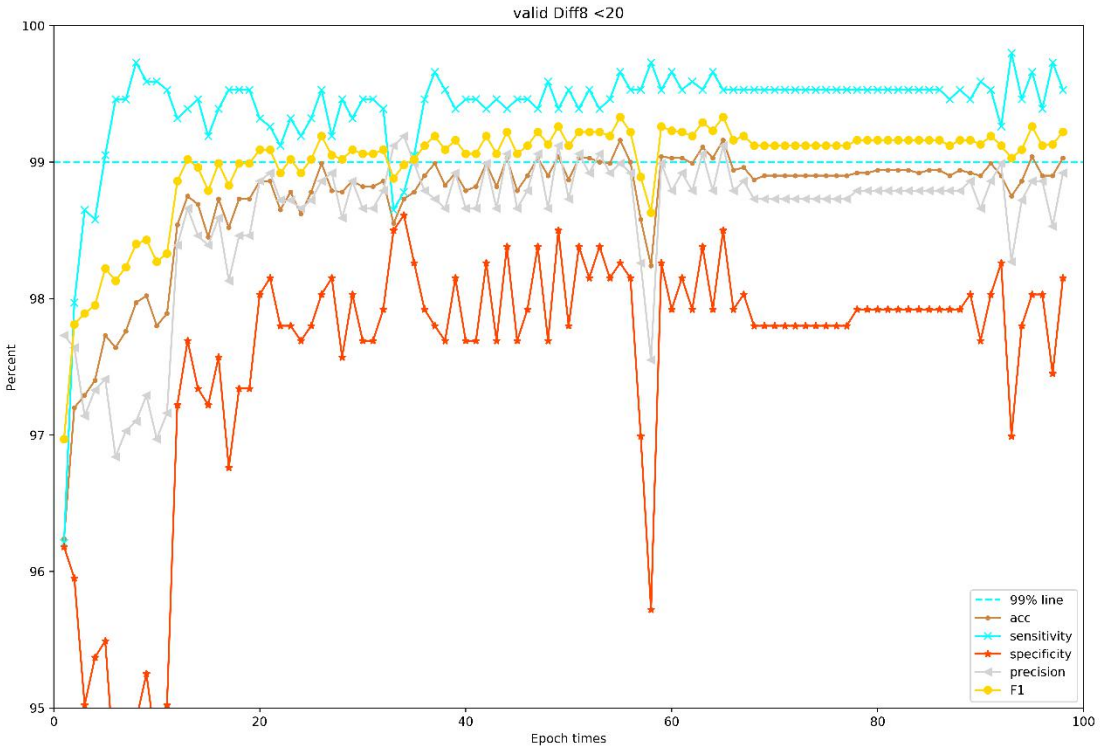

图 4-1 Feature8 验证集结果

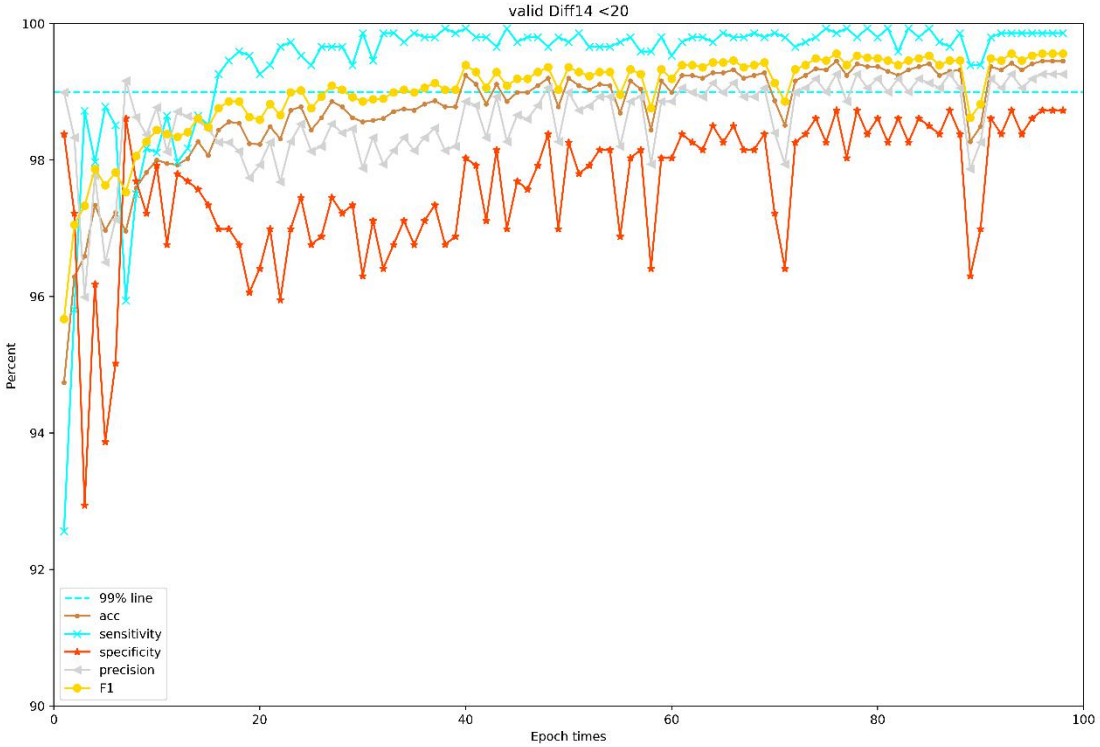

图 4-2 Feature14 验证集结果

可以看到结果表明，Feature14 经过 40 轮训练后各项指标均在 98%以上，具有比 Feature8 更好的效果，并且在 69 和 88 轮数据效果下降的情况下，数据波动不大，模型整体比 Feature8 拥有更好的稳定性和准确性。

由此可以见，我们的出发观点是正确的。在深度学习的模式下，通过增添抗原的理化性质，特别是 3D 结构特征筛选出的优势位点，增添更多的额外信息，从而使得深度学习能够超脱当前的知识领域进行更准确的判断和预测。

4.3 传统模型对比

为了更加展现模型的优缺点，我们收集了不同的该领域内比较有代表性的五种方法与之对比，分别是线性回归（LR）、临近算法（KNN）、支持向量机（SVM）、随机森林（RF）和 IAV-CNN。

表 4-1 模型对比结果

	LR	KNN	SVM	RF	IAV-CNN	Our Method
Accuracy	0.696	0.728	0.532	0.776	0.856	0.992
Precision	0.761	0.804	0.532	0.824	0.873	0.986
sensitivity	0.624	0.647	1.000	0.737	0.861	0.999
F1-score	0.686	0.717	0.695	0.778	0.867	0.993

可以看出，在表 4-1 中，相比于传统的机器学习方法，基于深度学习的 IAV-CNN 和我们的方法，比 LR，KNN，RF 等传统方法显著优秀，而 IAV-CNN 由于输入的是单一序列信息，但我们的方法增添了众多的特征，因此表现得更为优秀，说明模型结合众多的生物学信息，通过机器学习在更高维度的抽象特征是成功的，更趋近于预测抗原本质的能力，而非单纯的对序列进行回归预测分析。同时传统方法中对于序列差异大于 14 的序列对，往往判定为抗原变异，而在序列差异 14 到 20 的区间内，含有 10%的抗原相似的数据[11]，造成其模型敏感性和特异性评分较低。而本模型在此区间内，也有着很好的预测能力，因而本模型评分效果较好。

4.4 模型总结和展望

通过对以往模型的总结归纳，我们利用生物学的理化性质以及深度学习的特

点，成功的训练出抗原预测深度学习模型，该模型具有一定的实用性，能够预测两抗原对之间为抗原相似或抗原变异，进而对未来的流感疫苗研发做出一定的贡献，尤其是对于新兴的流感病毒，本模型更能利用生物学的特点给与准确的预测。

在候选模型中我们选取了重要表位、亲疏水性、极性和氢键结合位点等众多的特征，得益于生物信息学的快速发展，整合出本次更为科学的模型。但同时模型采用已有病毒的数据，先行计算出对突变有着重大影响的 47 个抗原优势位点，是基于一定的先验证知识得出。而病毒本身通过不断地进化，其结构和序列必然会改变，因此理论上本模型也仅仅针对近期的病毒具有很强的预测性，而面对未来新型结构的病毒，其准确度将会降低。而解决此问题的办法或许利用结构建模，对优势位点进行实时计算。例如计算 3D 结构使用的 MODELLER 同源建模，数据依赖于建模的准确程度，而 MODELLER 的准确程度一般在 60%—80%之间。最新的 AlphaGo 能够在预测蛋白结构上面拥着 90%以上的准确度。因此对于未来，希望能利用最新的模板预测技术，来进一步提高模型的能力。

通过查阅相关文献，对 Feature14 和 Feature8 模型效果的差异原因进行探究。认为是 Feature8 由于进行了序列特征差值计算，而序列对之间差异又较少，因此导致 Feature8 的输入矩阵更为稀疏。同时，认为的序列特征差值计算，使得矩阵原本的信息会遭受一定程度的破坏，使得深度学习模型不能很好的获取相关特征。而 Feature14 则会通过系统自动的进行差异或加算运算，即使矩阵本身比 Feature8 更大，训练需要更多的参数，但依旧可以训练出更为出色的模型。

同时也由于时间和精力的有限，本次模型仅针对于 H3N2 亚型，具有一定的局限性，未能预测更多例如 H1N1 等亚型。本模型设计初期，便考虑到不同亚型预测的问题，而深度学习可以通过迁移学习将相近的模型进行迁移训练，加快模型训练。而流感病毒不同亚型之间也有着共同的理化性质基础，无疑非常适合迁移学习。后期将继续进展相关工作，主动学习迁移学习的相关内容，将模型运用在更多的病毒亚型中。

参考文献

[1] Organization W H . Influenza (Seasonal): fact sheet. 2014.

[2] Trevor Bedford, Riley Steven, Barr Ian-G, et al. Global circulation patterns of seasonal influenza viruses vary with antigenic drift.[J]. Nature, 2015, 523(7559).

[3] M-Bouvier Nicole, Peter Palese. The biology of influenza viruses[J]. Vaccine, 2008, 26.

[4] J-L Virelizier. Host defenses against influenza virus: the role of anti-hemagglutinin antibody.[J]. Journal of immunology (Baltimore, Md. : 1950), 1975, 115(2).

[5] Kalus,Stöhr. Influenza—WHO cares[J]. Lancet Infectious Diseases, 2002.

[6] Claudia T , Daniele P , Stuart M , et al. Overview of Serological Techniques for Influenza Vaccine Evaluation: Past, Present and Future[J]. Vaccines, 2014, 2(4):707-734.

[7] Shu Y , Mccauley J . GISAID: Global initiative on sharing all influenza data – from vision to reality[J]. Eurosurveillance, 2017, 22(13).

[8] Seonwoo M , Byunghan L , Sungroh Y . Deep Learning in Bioinformatics[J]. Briefings in Bioinformatics, 2017(5):851.

[9] Nakaya H I , Wrammert J , Lee E K , et al. Systems biology of vaccination for seasonal influenza in humans.[J]. Nature Immunology.

[10] Wilson I A , Cox N J . Structural basis of immune recognition of influenza virus hemagglutinin.[J]. Annual Review of Immunology, 1990, 8(1):737.

[11] 杜向军. 甲型 H3N2 流感病毒基因组关联网络分析、抗原预测与疫苗株推荐[J]. 2010.

[12] Wu A , Peng Y , Du X , et al. Correlation of Influenza Virus Excess Mortality with Antigenic Variation: Application to Rapid Estimation of Influenza Mortality Burden[J]. Plos Computational Biology, 2010, 6(8).

[13] Zhou X , Yin R , Kwoh C K , et al. A context-free encoding scheme of protein sequences for predicting antigenicity of diverse influenza A viruses[J]. BMC genomics [electronic resource], 2018.

[14] RNA Viruses - Influenza A Virus; IAV-CNN: a 2D convolutional neural network model to

predict antigenic variants of influenza A virus[J]. Computers Networks & Communications, 2020.

[15] 周博. 基于深度神经网络的甲型 H3N2 流感病毒抗原性预测[D]. 2019.

[16] Smith D J , Lapedes A S , de Jong J C , et al. Mapping the Antigenic and Genetic Evolution of Influenza Virus[J]. Science, 2004, 305(5682):371-376.

[17] Qiu J , Qiu T , Yang Y , et al. Incorporating structure context of HA protein to improve antigenicity calculation for influenza virus A/H3N2[J]. Scientific Reports, 2016, 6:31156.

[18] Liao Y C , Lee M S , Ko C Y , et al. Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus[J]. Bioinformatics, 2008, 24(4):505-512.

[19] Huang J W , King C C , Yang J M . Co-evolution positions and rules for antigenic variants of human influenza A/H3N2 viruses[J]. BMC Bioinformatics, 2009, 10.

[20] Shuichi K , Piotr P , Maria P , et al. AAindex: amino acid index database, progress report 2008[J]. Nucleic Acids Research, 2008, 36(Database issue):D202-D205.

[21] Qiu Tianyi, Yiyan Yang, Jingxuan Qiu, et al. CE-BLAST makes it possible to compute antigenic similarity for newly emerging pathogens[J]. Nature Communications, 2018, 9(1).

[22] 华超. TensorFlow 与卷积神经网络[M]. 电子工业出版社, 2019.

[23] DeLano W L . The PyMol Molecular Graphics System[J]. Proteins Structure Function and Bioinformatics, 2002, 30:442-454.

[24] Angermueller C , P?Rnamaa T , Parts L , et al. Deep learning for computational biology[J]. Molecular Systems Biology, 2016, 12(7):878.

致　谢

经过几个月的艰苦努力，我终于完成了《基于卷积神经网络预测流感病毒抗原关系》的全部设计过程，并成功完成论文，取得了不错的结果。

本设计所涉及到的所有数据、模型、制图软件，均为开源或免费版本，并且已获得授权。所有代码和设计均是吕小毅老师和孟静博士的指导下独立完成的。论文中所有引用的文献也加入了文献应用中。撰写论文时，也保证理论的真实可靠，数据的真实可靠，本着严谨的科学精神完成本片论文。在学校期间感谢吕小毅教授和我爱我家实验室师兄师姐们的帮助，让我能够顺利的完成论文。同时感谢四川大学章乐教授对我的支持和中国医学科学院苏州系统医学研究所的蒋太交研究员，提供给我这样一个平台

天行健，君子自强不息！从高考结束，踏入大学的一刻起，我也不敢有丝毫放松，带着对高考的遗憾，认真追逐着大学脚步，脚踏实地的打好专业课的基础。从高中时的被动学习，到大学的主动求知，这一步我走的很久。也感谢我的高中——洛阳市第八中学三年的寄宿生活以及独特的上课模式，给予我独立思考的能力。此时此刻才明白，纵使是高考分数，也不过是人生中的过去风景，风雨如何？阳光又如何？回头来，依旧是一蓑烟雨任平生，也无风雨也无晴！

地势坤，君子以厚德载物！从大一的偏执，到如今的开朗豁达，一路走来免不了和同学朋友们磕磕碰碰，曾经的许许多多的行为如今也显得十分幼稚。但也就是那些幼稚，才有了如今的成熟。但无论怎样，陪伴着我一路走来的朋友，也让我明白，何为乐观、何为独立、何为友情、何为爱情。无论怎样我都相信，热爱、责任、真诚、善良、勇敢、自信，这些品质将成为做事的准则。感谢我的朋友和系统所师兄师姐的一路陪伴，特别感谢我孟静师姐这几个月来的辛勤付出！

由于我的学术水平有限，所写论文难免有不足之处，恳请各位老师和学友批评和指正。