國立臺灣大學生物資源暨農學院生物機電工程學系

碩士論文

Department of Biomechatronics Engineering
College of Bioresources and Agriculture
National Taiwan University

Master Thesis

利用深度學習預測T細胞受體與抗原結合的特異性

Using deep learning to predict antigen binding specificity of T-cell receptors

劉又瑋

YU-WEI LIU

指導教授：陳倩瑜 博士

Advisor : Chien-Yu Chen, Ph.D.

中華民國 111 年 7 月

July 2022

doi:10.6342/NTU202203779

致謝

兩年時光飛逝，感謝當初同意讓我加入實驗室的倩瑜老師，讓我在實驗室裡學到了很多東西，雖然大學所學與生物資訊無關，但老師與各位學長的耐心指導下，讓我在實驗室裡可以很快地進入狀況；最要感謝的是弘曄，加入實驗室之後總是可以提供給我很有建設性的意見，讓我能夠順利的解決各種問題；接者感謝在實驗室幫助過我的學長們東祈、文策、昀翔，一起度過在實驗室的時光；再來感謝同屆的淘喻以及如秀，一起修課討論課業。很開心再可以再 c4Lab 遇到很多好朋友。最後要感謝我的爸媽，讓我再要重考的時候支持我，最後才能順利進入台大就讀並加入 c4Lab。

doi:10.6342/NTU202203779

摘要

預測 T 細胞受體 （T cell receptor，TCR ） 與主要組織相容性複合物（Major histocompatibility complex，MHC） 和胜肽（Peptide） 結合的相互作用，仍然是極具挑戰性的計算問題。這一挑戰主要源於三個主要因素：實驗數據準確性、稀缺性和問題本身的高複雜性。一般而言，關於新生抗原（Neoantigen）和抗原生物學中未解決的基本問題之一是：為什麼並非所有新生抗原或抗原都會引發 T 細胞反應，對此，如果能準確預測新生抗原/抗原和 TCR 之間相互作用，將對於了解癌症進展、預後和對免疫治療的反應之相關研究至關重要。另一方面，近期許多自然語言處理（Natural Language Processing，NLP）相關研究顯示，可將蛋白質序列視為句子，而將胺基酸視為單詞，因此，許多相關研究開始嘗試使用類似自然語言處理的技術，從蛋白質序列數據庫中提取有用的生物信息。日前，有一些可公開使用的蛋白質語言預訓練模型被釋出，而且已被證明有助於各種下游預測任務。因此，本研究旨於建立了一個以蛋白質語言模型 ProtBert 為編碼基礎的預測模型，預測由 I 類主要組織相容性複合物呈現的新生抗原和一般 T 細胞抗原的 TCR 結合特異性。本研究針對兩個預測問題，一個是預測 MHC-I 和 peptide 的結合問題，一個是 TCR 和 peptide-MHC（pMHC）的結合問題，比較不同編碼方式，結果顯示蛋白質語言模型在兩個問題上都可以提升預測準確率。最終，本研究提出搭配集成學習，進一步提升以 ProtBert 為基礎的預測模型之準確性，期望能強化預測 T 細胞受體與抗原結合特異性之後續應用。

doi:10.6342/NTU202203779

Abstract

Predicting the interaction of T cell receptors (TCR) with complexes of major histocompatibility and peptide (pMHC) remains challenging. This challenge involves three main issues: accuracy of data, sparse and problem complexity. One of the fundamental and unanswered question about neoantigen and antigen is why not all antigen elicits T cell responses although the peptide might have been present on the MHC cell surface. Accurate and comprehensive characterization of the interactions between neoantigen/antigen and TCR is critical for understanding cancer progressions, prognosis, and the response of immunotherapy. On the other hand, many recent NLP studies have shown that protein sequences can be regarded as sentences and amino acids as words. In this regard, researchers can use natural language processing to extract biological information from protein sequence databases. Recently, there are some successful pre-training protein language models publicly available. This study then developed a prediction model based on protein language model ProtBert to predict TCR binding specificity of neoantigen/antigen presented by major histocompatibility complex class I. The results demonstrated that using protein language model can improve the accuracy of prediction on both problems: predicting MHC-peptide binding and TCR-pMHC binding. Moreover, this study integrated ensemble learning to further improve the prediction accuracy. The ProtBert-based ensemble model is expected to facilitate the immunogenomics studies related to TCR binding in the near future.

doi:10.6342/NTU202203779

目錄

doi:10.6342/NTU202203779

doi:10.6342/NTU202203779

doi:10.6342/NTU202203779

圖目錄

doi:10.6342/NTU202203779

doi:10.6342/NTU202203779

表目錄

doi:10.6342/NTU202203779

第一章 研究目的

T 細胞受體是 T 細胞表面的特異性受體 （T cell receptor，TCR ），負責識別由主要組織相容性複合物 （Major histocompatibility complex，MHC） 呈現於細胞表面上的抗原，TCR 由兩種蛋白質鏈（α 和 β）所組成，TCR 透過將 V-D-J 三種基因片段進行基因重組，來製作大量不同的 α 和 β 序列，進而和不同的 peptide-MHC（pMHC）結合。

主要組織相容性複合物與感染所產生外來抗原或是癌症所產生的新生抗原（Neoantigen）結合，並呈現在細胞表面上，從而使 T 細胞能夠識別，TCR β 與 pMHC 相互作用由三個環定義，這些環決定了 TCR 的特異性，並表示為互補決定區（Complementarity-determining regions，CDR），目前已知 CDR1 和 CDR2 與 MHC 作用，而 CDR3 主要與胜肽（peptide） 相互作用，也有一些研究指出 TCR α 中的 V 基因也會一定程度影響 TCR-pMHC 結合。近年來，一些更複雜且有效的深度學習方法顯示出在預測上的突破，例如: pMTnet(Lu, Zhang et al. 2021)、NetTCR(Montemurro, Schuster et al. 2021)與 ERGO-II(Springer, Tickotsky et al. 2021)。

另一方面，蛋白質語言模型（Protein Language Model，pLM）將蛋白質序列解釋為一個句子，並將其組成元素（胺基酸）解釋為單詞，因此可以使用針對自然語言處理（Natural Language Processing，NLP）的技術從蛋白質序列中提取有用的生物資訊，蛋白質序列受限於特定的功能而產生特定的 3D 結構，這些約束反映了 NLP 中的語法和詞意的規則，目前已經有一些成功的預訓練蛋白質語言模型像是 ProtBert(Elnaggar, Heinzinger et al. 2020)和 ESM1b(Rao, Meier et al. 2020)，被證明有助於各種預測任務的微調。

doi:10.6342/NTU202203779

遷移學習（Transfer learning）是一種深度學習的方法，針對相似問題的模型，將預訓練好的模型一部分或是全部用於訓練所需的模型上，遷移學習分成兩種，第一種是將感興趣問題的相關數據微調於預訓練模型中，利用預訓練模型的權重當作新模型的初始權重。第二種是特徵擷取，將所需任務中的數據輸入到預訓練模型中，然後提取與預訓練模型的相關信息，並將其做為新模型的特徵。在本研究中，將利用遷移學習來做特徵擷取，利用預訓練好的蛋白質語言模型 ProBert 為編碼基礎建立一個預測模型，專注於 TCR 以及 MHC-I 類 peptide 結合進行預測，使用的標準指標以及與 pMTnet 這篇論文相同的訓練以及測試資料來評估新模型的性能，來展示本研究提出的方法優於 pMTnet。

doi:10.6342/NTU202203779

第二章　文獻探討

2.1　TCR 與 peptide-MHC 複合物結合

四種不同的 T 細胞抗原受體多 peptide（α、β、γ、δ）形成兩種不同的異二聚體（α:β 和 γ:δ），它們在一級序列、基因組織和重排（recombination）方式上與免疫球蛋白非常相似。T 細胞受體僅存在於細胞表面，並且當它們嵌入主要組織相容性複合物 （MHC） 分子中時才能識別抗原片段。而帶有 αβ 的 T 細胞受體的 T 細胞在大多數脊椎動物的免疫反應(Krogsgaard and Davis 2005)中至關重要，這些細胞的特異性受到 αβ T 細胞受體控制。與 B 細胞以及 γδ T 細胞相比，αβ T 淋巴細胞面臨更複雜的配體（ligand），這個配體由主要組織相容性複合體的抗原片段所組成。

TCR-pMHC 交互作用完全在兩個細胞表面上進行，這是一個非常特殊的環境，在這種環境下，高親和力可能產生更多的障礙，T 細胞必須避免對任何表達 peptide-MHC 複合物產生強烈反應，但又必須要對特異性結合對特定子集。互補決定區中包含 pMHC 結合的位點，而抗原跟 CDR3 有密切關係。TCR 的變異性在結合當中佔了很大的比重。

doi:10.6342/NTU202203779

2.2 自然語言處理

自然語言處理是人工智慧和語言學領域的分支，透過複雜的數學模型及演算法來讓機器去認知了解我們的語言系統，機器翻譯就是 NLP 的應用的一種，將需要被翻譯的文本輸入進 NLP 模型中，模型就會辨識理解再生成目標語句。

目前在 NLP 中最廣為研究人員使用的演算法模型即是 BERT (Devlin, Chang et al. 2018)，BERT 是一個無監督式的模型，全名為轉譯器的雙向編碼表述（Bidirectional Encoder Representations from Transformers，BERT），是 Google 基於 Transformer 架構上的一個開源模型，BERT 訓練資料來自維基百科以及 BooksCorpus 總計 33 億個字。

BERT 是傳統語言模型的變形，而語言模型（Language Model，LM）就是在給定一些詞彙下去評估下一個詞彙的出線的機率，在訓練 BERT 時讓它進行兩個任務，第一個是克漏字填空，第二個是判斷兩個句子是否相連。

訓練一個無監督式的 LM 的好處，數據是無限大的，由於不需要標記標籤，所以網路上的所有文本都可以當作是資料集。訓練一個 BERT 這種無監視 LM 的模型成本分非常高，因此微調 BERT 來實現各種任務成為常態。LM 可以學會語言結構，透過特徵擷取或是微調（fine-tuning） 都可以提升表現以及減少各式 NLP 模型的成本。

doi:10.6342/NTU202203779

2.3 蛋白質語言模型

蛋白質由 20 種不同的胺基酸所組成，像珠子一樣，這些胺基酸以一維序列的方式串起來，這些一維序列透過摺疊成獨特的三維結構來執行特定的任務。蛋白質胺基酸序列的資料比起蛋白質結構的資料多了好幾個數量級，通過基於機器學習的預測方式來縮小序列註釋是生物學以及生物資訊面臨的挑戰之一，近年來透過遷移學習利用大量未標記的數據來彌補這個差距。

蛋白質語言模型將蛋白質序列視為句子，並將其組成成分胺基酸當成單詞，ProtTrans (Elnaggar, Heinzinger et al. 2020) 訓練蛋白質語言模型並基於序列提取特徵的方法，資料利用 3930 億個胺基酸來進行訓練，是 BERT 的資料幾百倍以上，利用無監督式的方式預訓練在目前已發表的 NLP 模型當中，像是 BERT、Albert(Lan, Chen et al. 2019)、T5(Raffel, Shazeer et al. 2019)等，實現兩個目標，第一個比較不同的 NLP 模型訓練成蛋白質語言模型後的要能，第二個比較了後續微調在其他任務上的效能。

doi:10.6342/NTU202203779

2.4 TCR-pMHC 資料庫

2.4.1 VDJdb

VDJdb (Shugay, Bagaev et al. 2017)是抗原特異性 T 細胞受體的綜合資料庫，通過手動處理已經發表過的研究，這些研究定義了 T 細胞對配體的受異性。VDJdb 由已知的 MHC-I 和 MHC-II 呈遞已知配體的能力。

在 VDJdb 中，每個配對都包含了 TCRα、TCRβ、同源配體（包含 peptide 序列）以及 MHC-I、MHC-II。VDJdb 的出現與免疫組庫高通量測序（Immune repertoire sequencing technology 或 RepSeq）的出現密切相關，該技術能夠從單個樣本中快速獲取數百萬個不同的 TCR 序列。圖 1 為資料整理的流程，通過將 TCR 特異性與高通量篩選技術配對的方式，有助於發現一系列免疫相關疾病。

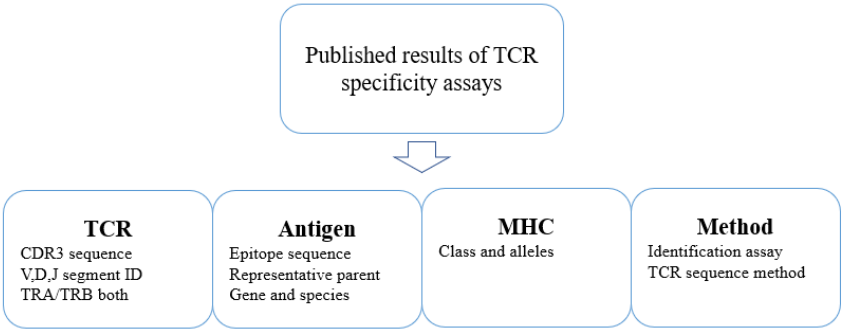

圖 1 VDJdb 資料整理流程，取自文獻(Shugay, Bagaev et al. 2017)

doi:10.6342/NTU202203779

2.4.2 McPAS-TCR

McPAS-TCR (Tickotsky, Sagiv et al. 2017)建立在與人類與鼠的各種疾病以及疾病相關的 T 細胞，在 PubMed 中查詢 CDR3 序列、T 細胞受體，用手動的方式搜索整理，將 TCR 序列與抗原與相關的疾病以及器官連接。該數據庫可以通過疾病狀況、T 細胞類型、抗原、物種、MHC、檢測類型和其他標准進行查詢。

2.5 TCR-pMHC 結合預測工具

2.5.1 NetTCR

NetTCR (Montemurro, Schuster et al. 2021)利用淺層的捲積神經網路來解決 TCR-pMHC 的問題，固定 HLA 為 A*02:01 ，探討 TCR 中 CDR3α、CDR3β 及 peptide 的結合，圖 2 為模型架構。比較從最簡單的序列相似性模型到捲積神經網路（Convolutional Neural Network，CNN），對這些模型進行交叉驗證訓練，並用 Levenshtein 相似度演算法，對 TCR 的序列進行嚴格的冗餘序列刪除。結論表示比起只有 CDR3β 與 peptide 下，加入 CDR3α 的資訊有助於提升預測準確率，以及對序列進行冗餘序列刪除也可以提升準確率。

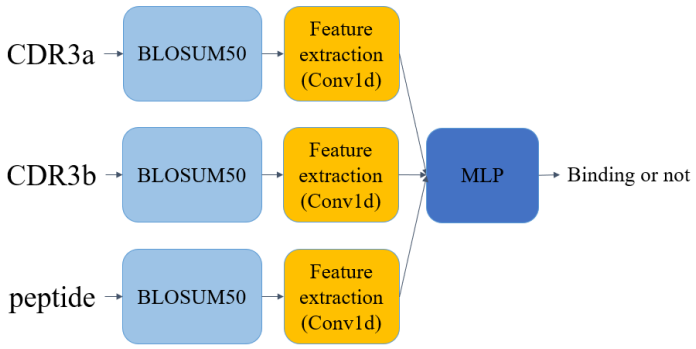

圖 2 NetTCR 模型架構圖，本圖參照(Montemurro, Schuster et al. 2021)之原圖之概念重繪

doi:10.6342/NTU202203779

2.5.2 ERGO-II

ERGO-II (Springer, Tickotsky et al. 2021)基於 TCRβ 和 peptide 的雙重編碼，包含其他額外的特徵 α、V 和 J 基因、MHC 分型。所有特徵都經過編碼，編碼用作多層感知器 MLP（Multilayer Perceptron，MLP） 的輸入。

圖 3 為 ERGO-II 模型架構圖，對於 TCRα 以及 TCRβ 編碼測試了兩種編碼方式，一種是利用長短期記憶（Long Short-Term Memory，LSTM）編碼，一種是用自編碼器（AutoEncoder，AE）額外利用外部的 TCR 數據進行預訓練，再微調到 ERGO-II 的分類器上，並將 TCR 序列長度固定在 28 個胺基酸。peptide 的部分利用 LSTM 編碼。其他特徵都使用嵌入矩陣進行編碼，每個特徵都擴增成 50 維的特徵向量，其他特徵包括 Vα、Vβ、Jα、Jβ 及 MHC，最後預測方面分為 MLP-I 以及 MLP-II，差別在有沒有 Vα 及 Jα 兩種特徵。

ERGO-II 結合了 TCR-pMHC 兩個大型數據集（McPAS-TCR 以及 VDJdb），由於數據集中只有 TCR-pMHC positive 資料，透過隨機抽取的方式擴增 5 倍的負樣本資料，從數據集中隨機抽取兩個樣本，TCRβ 序列、TCRα 序列、V、J 基因和 T 細胞類型取自第一個樣品。peptide 和 MHC 取自第二個樣本。此論文將這些特徵值合併到一個新的負樣本資料。

此論文最終結果發現對於大多數的 peptide，V 和 J 基因對於預測準確率有顯著的提升，隨後才是 CDR3α，而 MHC 在不同的 peptide 中貢獻不同。

doi:10.6342/NTU202203779

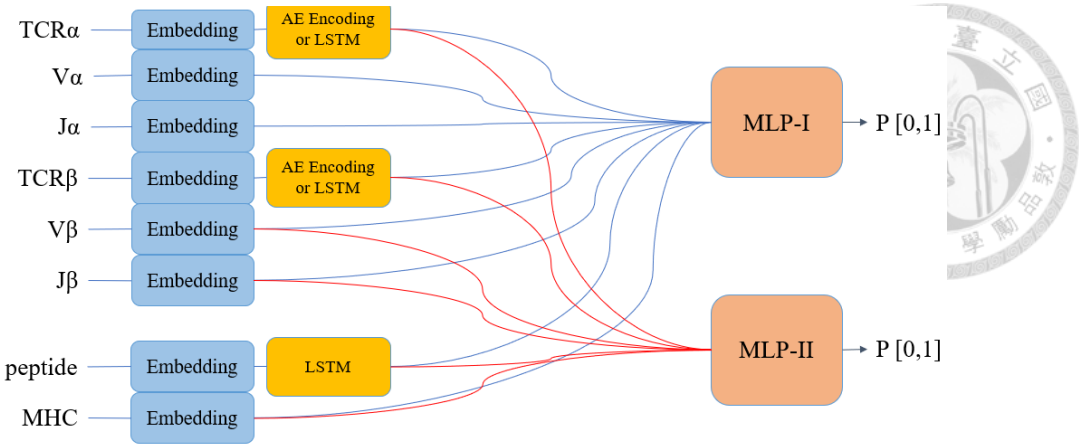

圖 3 ERGO-II 模型架構圖，本圖參照(Springer, Tickotsky et al. 2021) 之原圖之概念重繪

2.5.3 pMTnet

pMTnet (Lu, Zhang et al. 2021)基於遷移學習的模型，以預測 MHC-I 複合物呈現的新抗原和 TCR 結合的特異性，使用 TCR （CDR3β）、peptide、MHC-I 實現配對的準確預測，模型分為三個部分，分為 TCRβ 嵌入模型、pMHC 嵌入模型、MLP 預測模型。

TCRβ 嵌入模型的部分將 TCRβ 序列利用 Atchley factor 編碼，用五個數值代表每個胺基酸的生化特性，在無監督的情況下建立一個 Stacked AutoEncoder，通過比較輸出以及輸入的相似度來訓練，最後用 Bottleneck 來當作 TCR 的嵌入，圖 4 為 TCR 編碼模型圖。

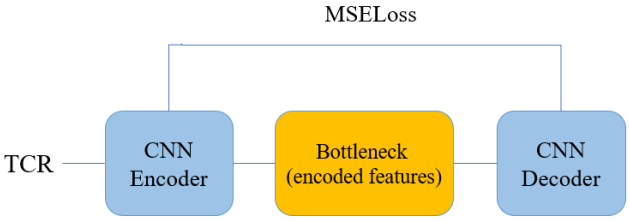

圖 4 pMTnet TCR 編碼器，本圖取自(Lu, Zhang et al. 2021) 之原圖之概念重繪

doi:10.6342/NTU202203779

pMHC 嵌入模型的部分，重建了 NetMHCpan(Reynisson, Alvarez et al. 2020)，NetMHCpan 是一個預測 peptide 與 MHC 結合的模型，利用偽序列對 MHC 編碼，偽序列由與 peptide 接出的胺基酸所組成，僅包含 34 個殘基，然後使用 BLOSUM50 矩陣對這 34 個殘基進行編碼，另一方面對 peptide 也用 BLOSUM50 進行編碼。這裡使用 MHC 序列而不是類型作為輸入，比起 NetMHCpan 原始的模型新增一個輸出大小為 16 的 LSTM 層加速收斂，最後取單神經元前一層也就是輸出的前一層當作 pMHC 的嵌入，圖 5 為 pMHC 編碼器模型圖。

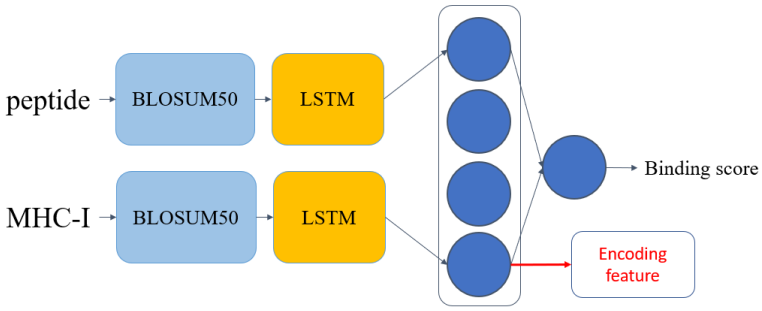

圖 5 pMTnet pMHC 編碼器，本圖取自(Lu, Zhang et al. 2021) 之原圖之概念重繪

最後將 TCR 以及 pMHC 的嵌入輸入進全連接層進行預測，在訓練過程中，已知 pMHC 和 TCR 之間的正面數據，本研究隨機將這些 TCR 和 pMHC 不匹配以創建十倍以上的負面數據，圖 6 為 pMTnet 最終模型架構。

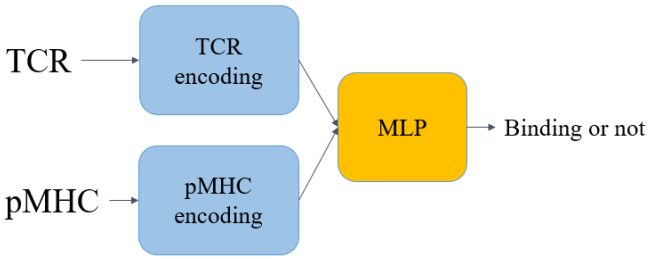

圖 6 pMTnet 模型架構圖，本圖取自(Lu, Zhang et al. 2021) 之原圖之概念重繪

doi:10.6342/NTU202203779

第三章 研究方法

3.1 資料介紹

3.1.1 NetMHCpan 資料收集

NetMHCpan 是一篇基於 peptide 及 MHC-I 兩個序列作為輸入的模型，該數據是從 IEDB(Vita, Mahajan et al. 2019) 用於重新訓練 I 類結合預測工具的數據集，包含 170,470 筆 peptide 跟 MHC-I 結合親和力測量值，涵蓋人類 109 種 I 類 MHC，包含 42 個 HLA-A、56 個 HLA-B、11 個 HLA-C 等位基因，peptide 長度從 8 個胺基酸到 14 個胺基酸，而超過 90%的長度落在 9、10 個胺基酸，圖 7 為 peptide 長度統計圖。

對於每個 HLA，選擇了結合親和力大於 50nM 的 peptide，標籤部分為了要標準化到 0~1 之間，採用了以下公式$1 - \log(\text{aff})/\log(50000)$，其中 aff 為以 nM 當單位的親和力值。將 80%數據用於訓練、20%用於驗證。

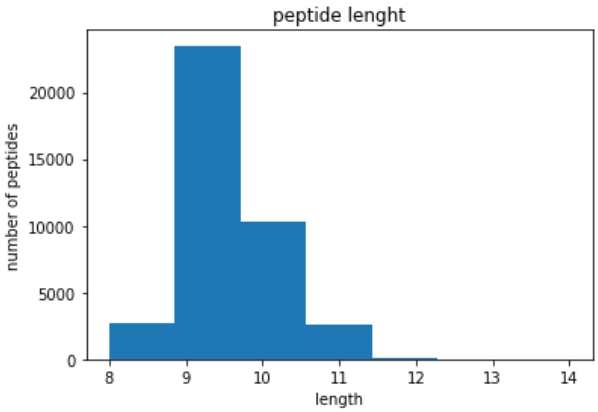

圖 7 NetMHCpan peptide 資料長度統計

doi:10.6342/NTU202203779

3.1.2 pMTnet 資料收集

pMTnet 是一篇基於 TCRβ、peptide、MHC-I 三個序列輸入的模型，資料來源為現在已經發表的論文，重複的資料刪除，這些資料僅保留高信心的資料，例如：VDJdb 只保留包含 VDJ 三種都有的資料並且分數大於零。

在訓練資料上有 32607 筆 TCR-pMHC 結合正資料集，其中包括 28,604 筆不重複的 TCRβ 中 CDR3 的序列長度從最短 5 個胺基酸，最長 26 個胺基酸，428 種 peptide，最短 8 個胺基酸，最長 20 個胺基酸，HLA 部分包含 31 種 HLA-A、30 種 HLA-B、2 種 HLA-C 等位基因，pMHC 有 514 筆 peptide 與 MHC-I 結合的序列，圖 8 為訓練資料中 TCRβ 以及 peptide 長度統計圖，表 1 為不重複序列統計。

為了要了解 TCR 多樣性與 pMHC 結合影響，透過隨機匹配 TCR 和 pMHC 創建十倍的負資料集，每一個 TCR 序列配上 10 個隨機的 pMHC 這些配對不會跟正資料集交集。在隨機匹配時我們會固定 pMHC 部分，確保 pMHC 中 peptide 與 MHC-I 一定結合，例如：隨機挑選兩個樣本 A 和 B，TCR 部分選取 A 樣本，pMHC 選取 B 樣本。並保證訓練資料以及測試資料中，不會有相同的 TCR-pMHC。

a b

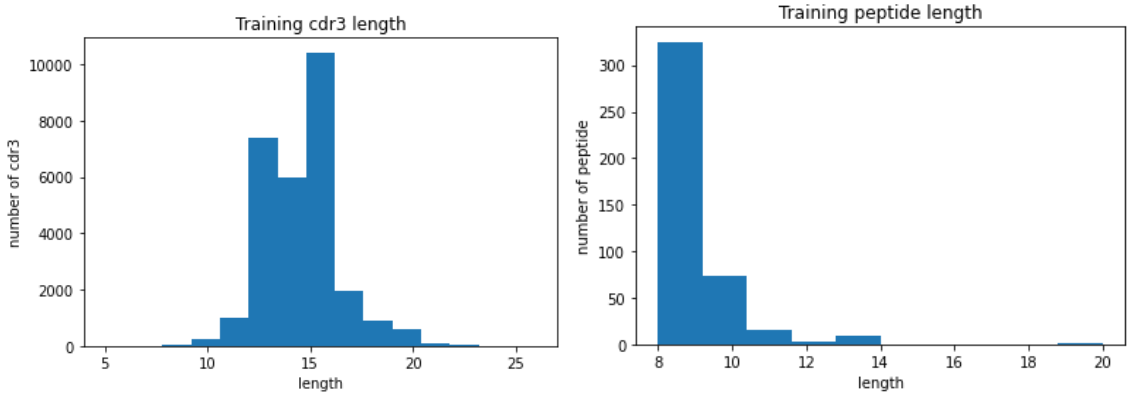

圖 8 a. pMTnet 訓練資料 TCRβ 長度統計圖 b. pMTnet 訓練資料 peptide 長度統計

doi:10.6342/NTU202203779

表 1 訓練資料不重複序列統計

	Number of unique sequences
TCRβ	28,604
peptide	428
MHC-I	63
pMHC	514

在測試資料上主要是來自高通量實驗的數據，總共有 619 筆 TCR-pMHC 結合的資料，其中包括 271 筆獨立的 TCRβ 中 CDR3 的序列長度從最短 9 個胺基酸，最長 22 個胺基酸，224 種 peptide，最短 8 個胺基酸，最長 13 個胺基酸，HLA 部分包含 6 種 HLA-A、16 種 HLA-B、2 種 HLA-C 等位基因，pMHC 有 236 筆 peptide 與 MHC-I 結合的序列，圖 9 為驗證資料中 TCRβ 以及 peptide 長度統計圖，表 2 為不重複序列統計。

整理過程中刪掉訓練資料出現過的，因此測試集在 TCR-pMHC 上是完全獨立的資料，在透過隨機匹配方式創建十倍的負資料集。

a b

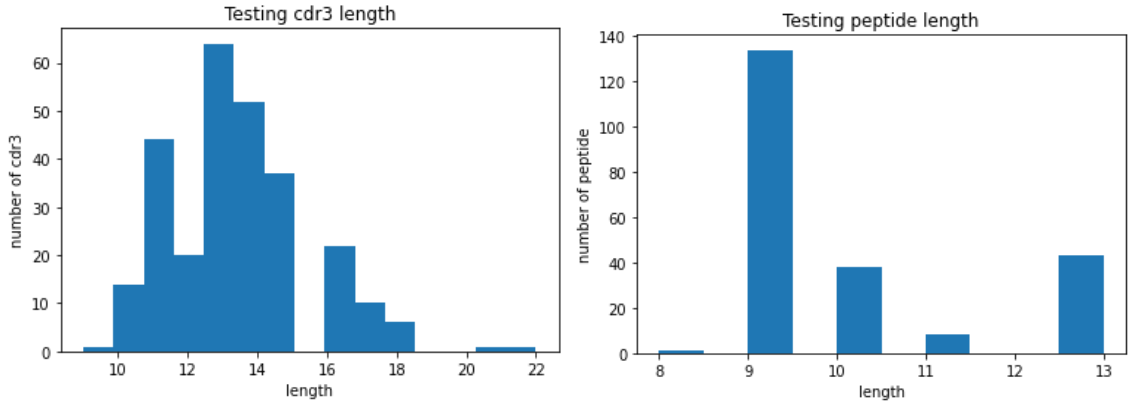

圖 9 a. pMTnet 測試資料 TCRβ 長度統計 b. pMTnet 測試資料 peptide 長度統計

doi:10.6342/NTU202203779

表 2 測試資料不重複序列統計

	Number of unique sequences
TCRβ	271
peptide	224
MHC-I	24
pMHC	236

3.2 實驗模型

圖 10 為模型架構圖，模型類似於 NetTCR，先將三種序列利用蛋白質編碼方法（Protein encoding method），再通過一維捲積神經網路做特徵擷取，通過不同的步數來提取特徵，通過一維捲積層、池化層、激活函數、最後將三個不同步數捲積網路的輸出合併，圖 11 為特徵擷取，再輸入到多層感知器做分類圖 12 分類。

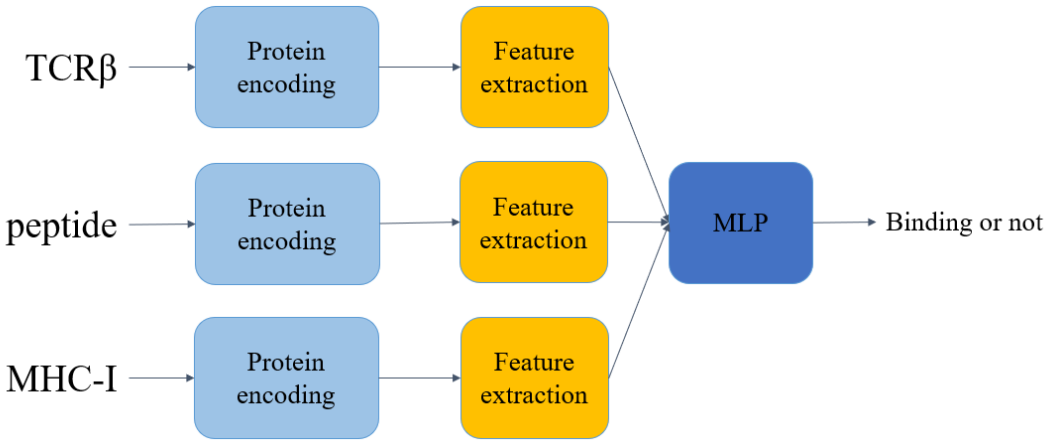

圖 10 預測 TCR-pMHC 結合模型架構圖

doi:10.6342/NTU202203779

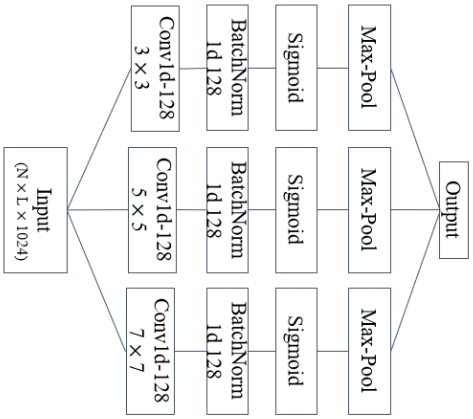

圖 11 ProtBert 為編碼基礎模型特徵擷取（feature extraction）

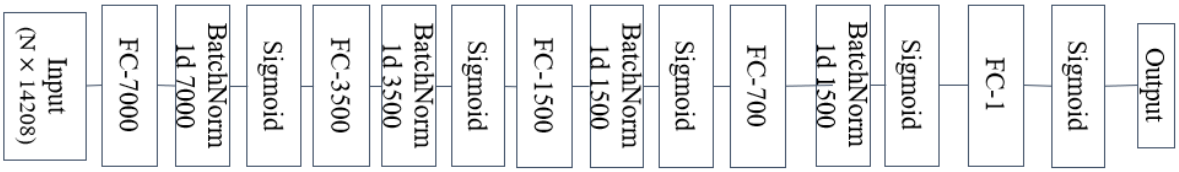

圖 12 ProtBert 為編碼基礎模型 MLP

集成學習，在同一資料集上透過訓練多個模型，這些模型在學習上有好有壞，透過將這些模型結合來讓模型提升效能降低過擬和，在這邊我們利用集成平均分類（Ensemble Averaging Classifier）中的 Simple Averaging，其方法是先將資料訓練出好幾個模型，再將模型輸出加起來平均。

圖 13 為 Ensemble 模型架構圖，在這裡本研究總共訓練了 20 個模型，並對輸出做 Simple Averaging，模型部分輸入採用 TCR、peptide、以及 34 個偽序列的 MHC-I，三個序列透過特徵擷取後合併再用 MLP 輸出。將平均過的輸出去計算 AUC，跟利用遷移學習的 pMTnet 做比較。

doi:10.6342/NTU202203779

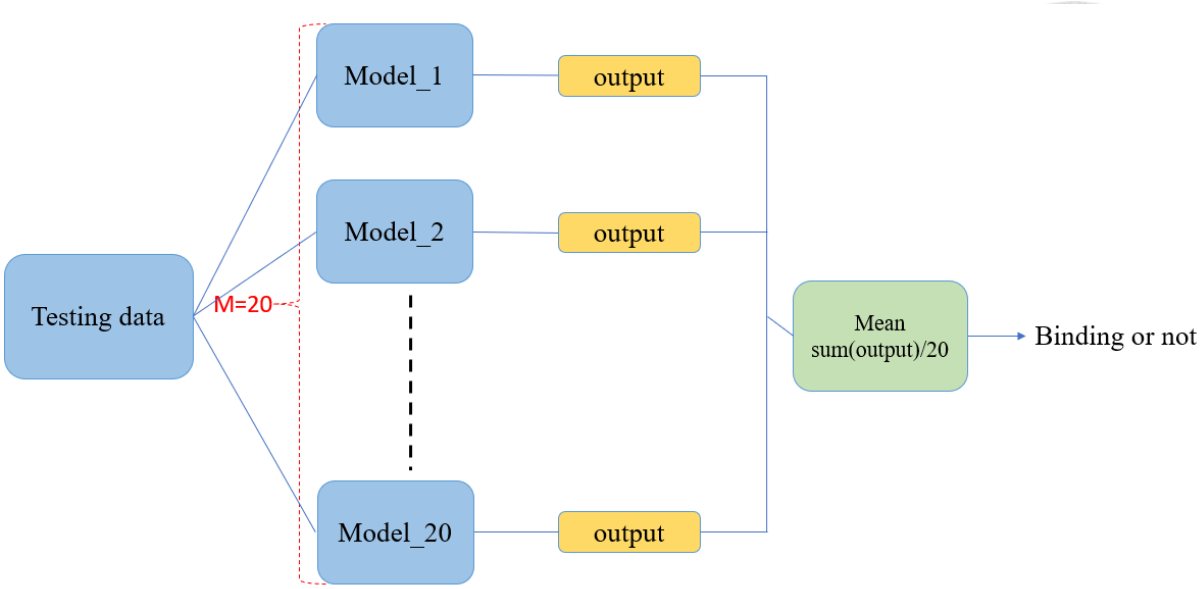

圖 13 Ensemble 模型架構圖

3.3 實驗流程

3.3.1 比較不同蛋白質編碼工具在 NetMHCpan 資料預測上影響

為了要探討 ProtBert 在 peptide 與 MHC-I 結合的效能，本研究比較不同的蛋白質編碼工具來探討對於模型的影響，分別是 BLOSUM50、One-Hot encoding 以及 ProtBert。NetMHCpan 的資料標籤為 0 或 1 的值，所以模型訓練會採用迴歸分析。

輸入資料

在 BLOSUM50(Henikoff and Henikoff 1992) 以及獨熱編碼（One-Hot encoding）上，peptide 以及 MHC-I 擴增成 21 維，為了要使用批量訓練，需要把序列長度統一，因此會使用 peptide 部分填充到長度為 15，MHC-I 部分使用偽序列(Nielsen,

doi:10.6342/NTU202203779

Lundegaard et al. 2007)，MHC-I 部分選定 182 個胺基酸，資料來源從 IPD-IMGT/HLA 當作參考，並利用 peptide 結合的 182 胺基酸當作基準，從 182 胺基酸中選定 34 個胺基酸，這 34 個胺基酸是 HLA 序列與 peptide 接觸的胺基酸所組成，接觸的定義是 HLA-A、B 結構中與 peptide 距離 4Å 以內，選定偽序列可以在不失特徵下降低序列長度，能夠維持訓練成果而且降低時間，這 34 個胺基酸位置分別是 7, 9, 24, 45, 59, 62, 63, 66, 67, 79, 70, 73, 74, 76, 77, 80, 81, 84, 95, 97, 99, 114, 116, 118, 143, 147, 150, 152, 156, 158, 159, 163, 167, 171。

ProtBert 部分 peptide 以及 MHC-I 擴增成 1024 維，由於 Bert 的輸入方式，序列在輸入前需要在起頭的部分加入[BOS]代表句子開頭，在結尾部分加入[EOS]代表句子結尾，例如：原始序列 CASSSYEQYF 改成[BOS] CASSSYEQYF [EOS]。peptide 部分填充成 20 加 2 總長為 22，MHC-I 部分一樣使用偽序列 34 加 2 總長 36。

模型配置

模型部分使用跟 NetMHCpan 類似的前饋神經網路，圖 14 為模型架構圖，N 為批次，L 為序列長度，我們在前饋神經網路前對 peptide 與 MHC-I 序列增加兩層一維的捲積神經網路來提取特徵圖 15 以及圖 16 為 BLOSUM50 及 One-Hot encoding 特徵擷取以及分類器，圖 17 以及圖 18 為 ProtBert 特徵擷取以及分類器，為了要加速損失收斂，損失函數的部分使用均方誤差（Mean Square Error，MSE），優化器的部分使用 Adam，學習率初始為 1e-3 分別在週期為 60, 100 時下降十倍，並訓練 250 個週期。

doi:10.6342/NTU202203779

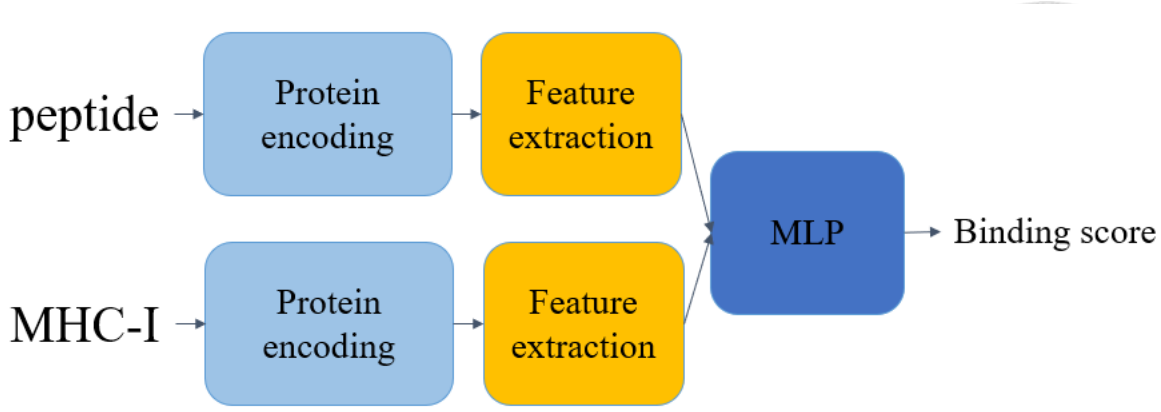

圖 14 預測 peptide 及 MHC-I 結合模型架構圖

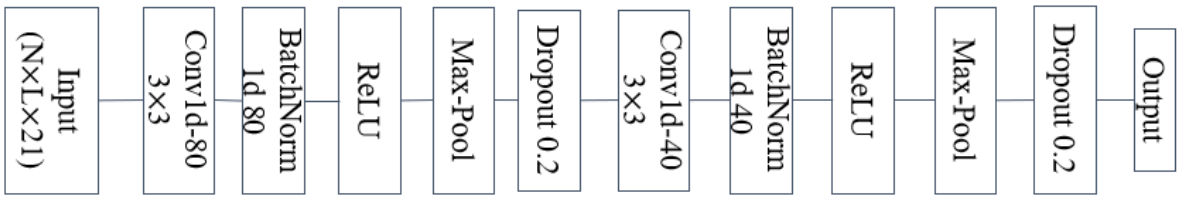

圖 15 BLOSUM50 和 One-Hot encoding 兩者方式的特徵擷取

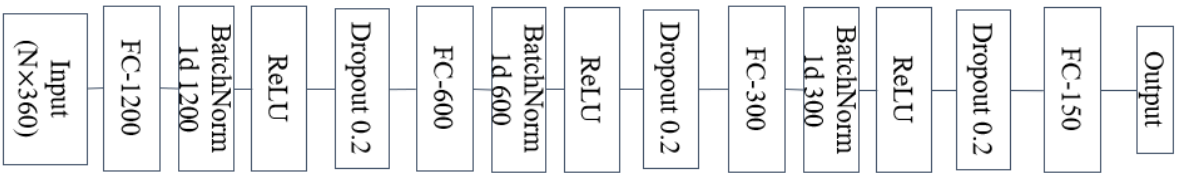

圖 16 BLOSUM50 和 One-Hot encoding 兩者方式的 MLP

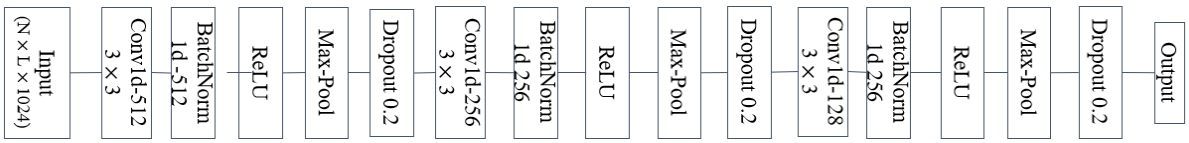

圖 17 ProtBert 和 nn.Embedding 兩者方式的特徵擷取

doi:10.6342/NTU202203779

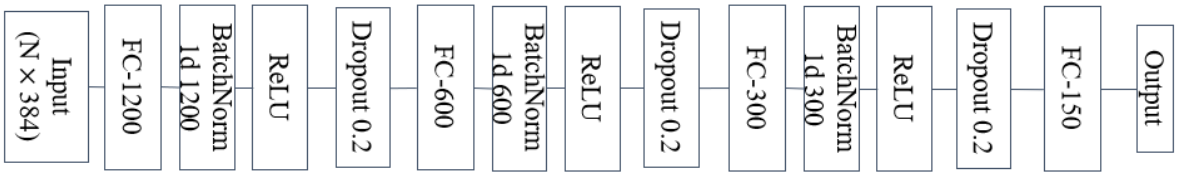

圖 18 ProtBert 和 nn.Embedding 兩者方式的 MLP

3.3.2 比較不同蛋白質編碼工具在 pMTnet 資料預測上影響

探討 ProtBert 在 TCR 與 pMHC 結合效能的影響，本研究比較不同的蛋白質編碼工具對模型的影響，分別是 BLOSUM50、One-Hot encoding、Pytorch 中的 nn.Embedding 以及 ProtBert，pMTnet 只提供正資料集標籤設為 1，負資料集利用隨機匹配的方式產生十倍標籤設為 0，所以模型會採用分類模型。

輸入資料

輸入資料部分 BLOSUM50 以及 One-Hot encoding 擴增成 21 維，nn.Embedding 以及 ProtBert 擴增成 1024 維，序列長度上，peptide 填充到長度為 20，TCR 填充到長度為 26，MHC-I 一樣使用偽序列選定 34 個胺基酸。

模型配置

模型配置與圖 10 相同，圖 19 為 BLOSUM50 以及 One-Hot encoding 特徵擷取，圖 20 為分類器，損失函數的部分採用二分類交叉熵損失函數（Binary Cross Entropy，BCE），並將 weight 設定為 10，由於負樣本為正樣本的十倍，因此損失函數的權重分別為正樣本的 10/11 和負樣本的 1/11。優化器的部分使用 Adam，學習率初始為 1e-3 分別在週期為 80, 140 時下降十倍，並訓練 200 個週期。

doi:10.6342/NTU202203779

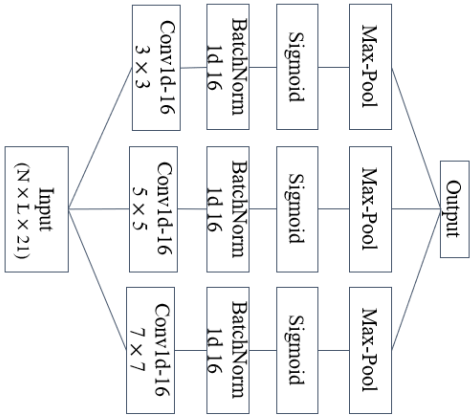

圖 19 BLOSUM50 和 One-Hot encoding 兩者方式的特徵擷取

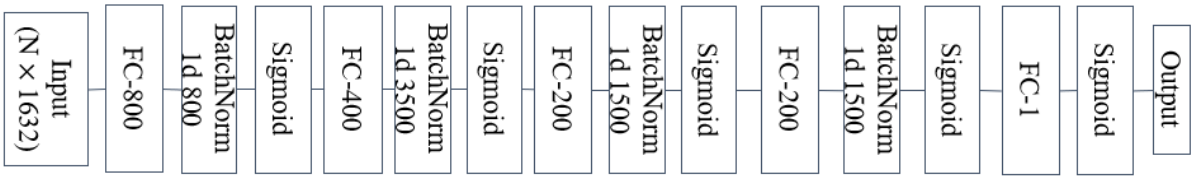

圖 20 BLOSUM50 和 One-Hot encoding 兩者方式的 MLP

3.3.3 比較不同 MHC-I 長度對 pMTnet 資料預測上影響

探討有無 MHC-I 這個特徵以及不同的 MHC-I 長度對於預測上的影響，在此比較中都會以 ProtBert 為底的模型來比較，圖 21 為模型架構比較圖，除了全長 MHC、無 MHC 以及長度為 34 的偽序列以外，MHCfovea (Lee, Chang et al. 2021)是一篇預測 MHC-I 與 peptide 結合，此篇也使用了偽序列，從 MHC-I 類 182 個胺基酸中選取了 42 個胺基酸，這 42 個胺基酸使用了 46 個 HLA-A、85 個 HLA-B、19 個 HLA-C 總計 150 個等位基因來選擇重要位置，首先刪除了沒有多態性也就是所有基因位置都有相同

doi:10.6342/NTU202203779

的胺基酸，再來計算每個位置的重要性，最後選定 42 個胺基酸位置分別是 1, 9, 11, 12, 24, 31, 32, 43, 44, 45, 62, 63, 65, 66, 67, 69, 70, 71, 73, 74, 76, 77, 79, 80, 94, 95, 97, 98, 109, 114, 116, 127, 131, 138, 142, 143, 144, 145, 152, 156, 163, 180。

資料部分一樣使用擴增過的 pMTnet 資料，模型部分，TCR 以及 peptide 部分長度固定，在沒有 MHC 的模型，先將 TCR 以及 peptide 編碼後分別輸入進一維捲積神經網路做特徵擷取，輸出合併利用 MLP 分類，有 MHC 的模型更改 MHC 長度，並在 MLP 中調整神經元數量。

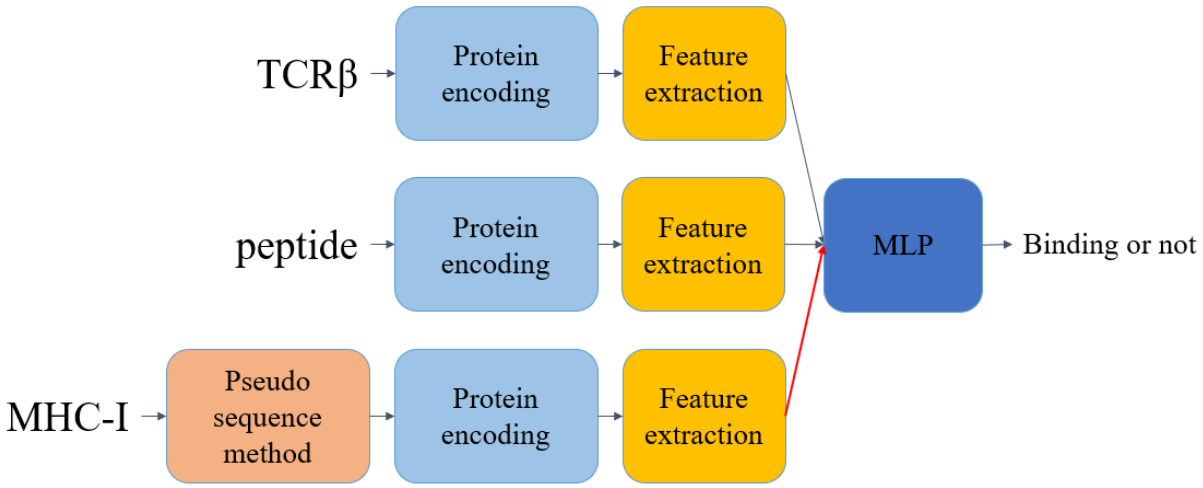

圖 21 比較 MHC 序列長度模型圖，如果是沒有 MHC 版本，則紅線部分不會輸入進 MLP

doi:10.6342/NTU202203779

3.3.4 比較不同填充（padding）方式對 pMTnet 資料預測上影響

本研究擬探討填充（padding）到序列前與序列後對預測上的影響，訓練過程中，MHC-I 一樣選定 34 個位置的胺基酸，比較 TCR 與 peptide 序列擴充前與後的影響，模型架構與圖 10 相同，資料部分一樣使用擴增過的 pMTnet 資料。

分為三種方式，一種是原本的序列後填充，peptide 填充到長度為 20，TCR 填充到長度為 26，MHC-I 一樣使用偽序列選定 34 個胺基酸。一種是序列前填充，peptide 填充到長度為 20，TCR 填充到長度為 26，MHC-I 一樣使用偽序列選定 34 個胺基酸。最後一種也是序列前填充，不一樣的地方是 peptide 的長度只選定前八個的胺基酸。

3.3.5 探討在訓練時刪掉特定的等位基因群對測試集的影響

資料部分使用 pMTnet 的資料，模型架構與圖 10 相同，選定 HLA-A*11 以及 HLA-B*57 兩個等位基因群做兩次訓練，第一次訓練把訓練資料裡 A*11 的資料刪除，並挑出測試集中 A*11 的資料來測試。第二次訓練把訓練資料裡 B*57 的資料刪除，並挑出測試集中 B*57 的資料來測試。

3.3.6 訓練結果評估

在回歸分析上，由於標籤是 0-1 的連續值，所以訓練結果會以皮爾森積動差相關係數來評估模型的好壞，輸出與標籤越相似模型學習的越好。

在分類模型上，訓練結果會以 ROC Curve （Receiver Operating Characteristic

doi:10.6342/NTU202203779

Curve）、PR Curve （Precision Recall Curve）、AUC （Area Under Curve）來評估模型的好壞，圖 22 為混淆矩陣。

ROC Curve 為各種決策門檻下比較假陽率（False Positive Rate，FPR）以及真陽率（True Positive Rate，TPR）間的變化，圖中 FPR 為 X 軸 TPR 為 Y 軸，其中 FPR 代表分類器的預測為正（Positive），但實際是錯的，即預測錯誤；TPR 代表分類器的預測為正（Positive），而實際也是正的，即預測正確。ROC cure 呈現 FPR 以及 TPR 之間的相對關係，圖中越靠近（ 0 , 1 ）代表完美分類，所以 ROC Curve 越靠近左上越好。

PR Curve 為各種決策門檻下比較 Recall 以及 Precision，Recall 為 X 軸，Precision 為 Y 軸，Recall 事實為真的樣本中有幾個是預測正確的，Precision 為陽性的樣本中有幾個是預測正確的，在算法中沒有考慮真陰性（True Negative），只專注於正樣本中，因此就算負資料集遠大於正資料集 Recall 以及 Precision 仍然是有效的指標，Precision, Recall = 1 也就是完美預測，因此我們的 PR 曲線越往右上角凸起則代表更好的模型表現，反之越平則代表越差。

AUC 代表曲線下的面積，為分類器預測能力的常用數值，前面提到 ROC Curve 越靠近左上方越好，下方面積越大則代表模型效能更好，PR Curve 越靠近右上方，下方面積越大則代表模型效能越好。

		Predict Class	
		N	P
Actual Class	N	TN	FP
	P	FN	TP

$$\text{TPR} = \frac{TP}{TP+FN}$$
$$\text{FPR} = \frac{FP}{TN+FP}$$
$$\text{Precision} = \frac{TP}{TP+FP}$$
$$\text{Recall} = \frac{TP}{TP+FN}$$

圖 22 混淆矩陣

doi:10.6342/NTU202203779

第四章 結果與討論

4.1 NetMHCpan 資料分析

在評估預測 peptide 與 MHC-I 相互作用結合中，預測模型是從 NetMHCpan 資料所建構，在這裡研究了兩種類型的模型，一種是基於預訓練的蛋白質語言模型編碼（ProtBert），一種基於統計以及二進制的方法（BLOSUM50、One-Hot encoding），我們計算了不同的 Protein encoding 方法的 correlation，我們以驗證集的資料作為模型評估標準，表 3 顯示，利用 ProtBert 為編碼基礎的模型有者最好的表現，而且相比於另外兩個模型有者大幅度的進步，圖 23 為訓練過程的驗證集的 loss 以及 correlation 學習曲線。這一小節驗證蛋白質語言模型 ProtBert 比起過往的 BLOSUM50 或是 One-Hot encoding 兩種編碼方式有者更好的效能。

表 3 peptide 及 MHC-I 模型預測結果 correlation 比較

Protein encoding method	**Pearson correlation**
BLOSUM50	0.68
One-Hot	0.73
ProtBert （MHC34）	0.83

doi:10.6342/NTU202203779

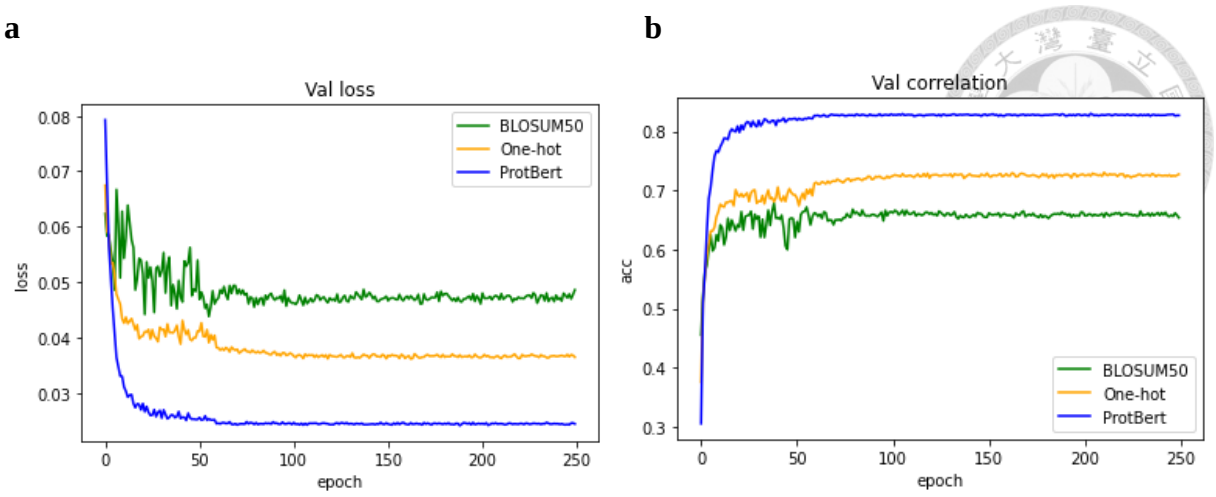

圖 23 三種不同的蛋白質編碼工具在驗證資料上的學習曲線 a.損失學習曲線 b. correlation 學習曲線

4.2 pMTnet 資料分析

在評估 TCR 以及 pMHC 結合中，預測模型是使用 pMTnet 擴增十倍所產生的負資料集，在這裡一樣研究了四種不同的蛋白質編碼方式，ProtBert、nn.Embedding、BLOSUM50 及 One-Hot encoding，本研究計算了四種不同模型的 AUC 以及 ROC Curve，表 4 為 AUC 比較，圖 24 為 ROC curve 比較圖，以測試集為評估標準，其中又以 ProtBert 的效能最好，在 AUC 上達到 0.78。

doi:10.6342/NTU202203779

表 4 TCR-pMHC 模型預測結果 ROC curve AUC 比較

Protein encoding method	**ROC curve AUC**
BLOSUM50	0.66
One-Hot	0.71
nn.Embedding	0.74
ProtBert	0.78

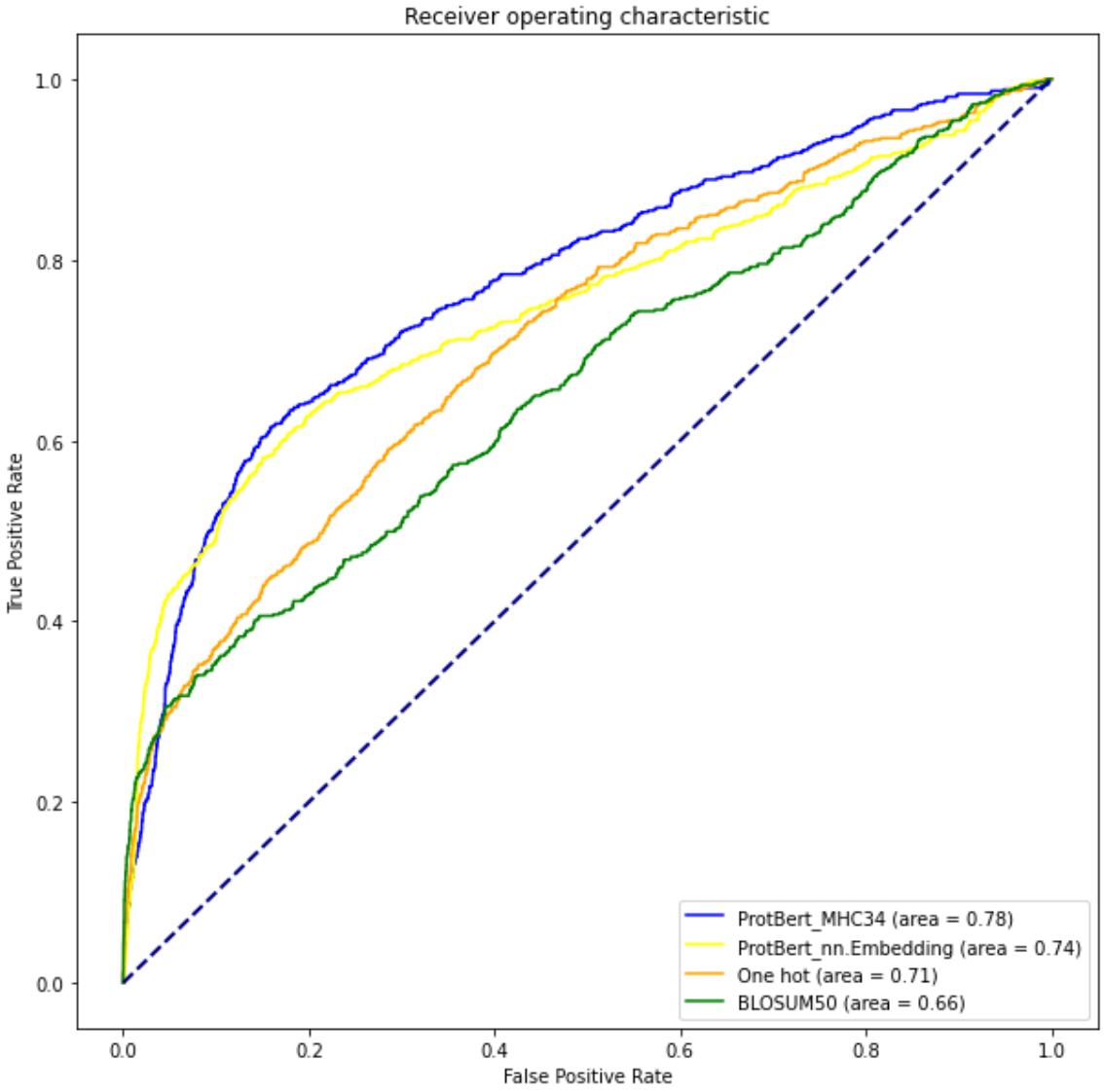

圖 24 TCR-pMHC 模型預測結果 ROC curve 比較

doi:10.6342/NTU202203779

4.3 探討 MHC 長度對 AUC 的影響

在沒有 MHC 這特徵，AUC 降到了跟使用 One-Hot encoding 差不多降低了 0.06 以上，由此可知 MHC 這個特徵的重要性。ERGO-II 這篇論文中有提到 MHC 的特徵提升效果不大，這篇論文當中 MHC 並非使用序列，而是使用 50 維的嵌入向量來表示，不同的 MHC 有者不同的向量，但在我們的研究當中可以發現 MHC 的特徵非常重要，可以推斷在特徵擷取上使用序列比起使用嵌入向量上更能夠提高效能。

而在偽序列 34 個胺基酸以及 42 個胺基酸的比較上，AUC 相差 0.03，並沒有相差太多，在 TCR-peptide 預測模型中，加入 MHC 序列能夠提升預測效能，但是在 MHC 全長部分推論是因為包含了太多的特徵，導致 AUC 只有 0.61 而已。表 5 為四種不同 MHC 長度的 AUC 比較，圖 25 四種不同 MHC 長度的 ROC curve。

表 5 不同的 MHC-I 長度對 TCR-pMHC 模型預測結果 ROC curve AUC 比較

Protein encoding method	**ROC curve AUC**
No MHC	0.69
MHC34	0.78
MHC42	0.75
MHC182 （total length）	0.61

doi:10.6342/NTU202203779

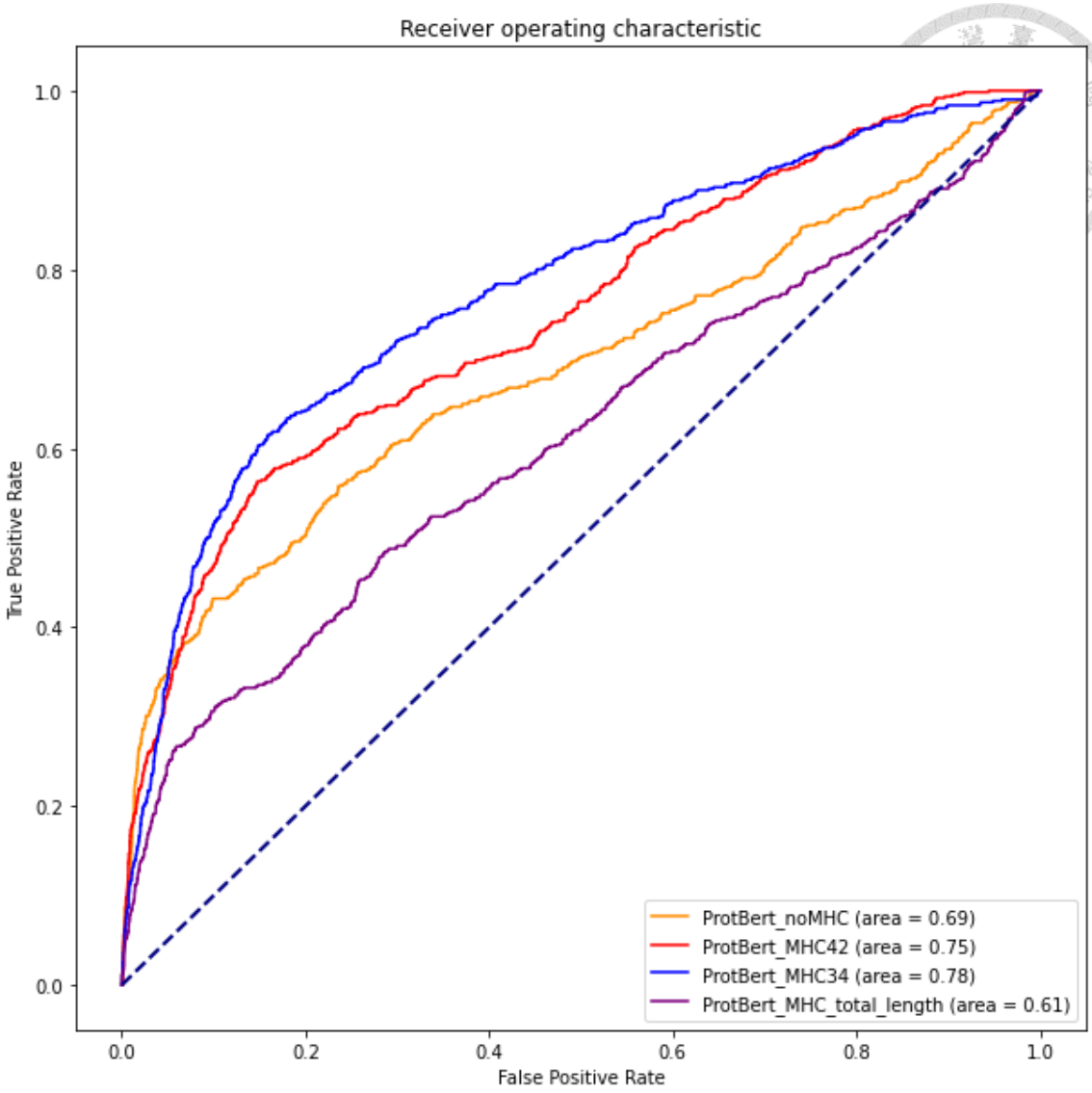

圖 25 不同的 MHC-I 長度對 TCR-pMHC 模型預測結果 ROC curve 比較

doi:10.6342/NTU202203779

4.4 探討不同填充方式對 AUC 的影響

結果上序列後填充對於預測結果上較好，而在序列前填充，只取 8 個胺基酸的 peptide 序列在預測上會比取整段填充 peptide 好。比起序列後填充的 AUC，序列前填充大概會差 0.08-0.14 的 AUC，圖 26 為三種不同填充方式比較圖。

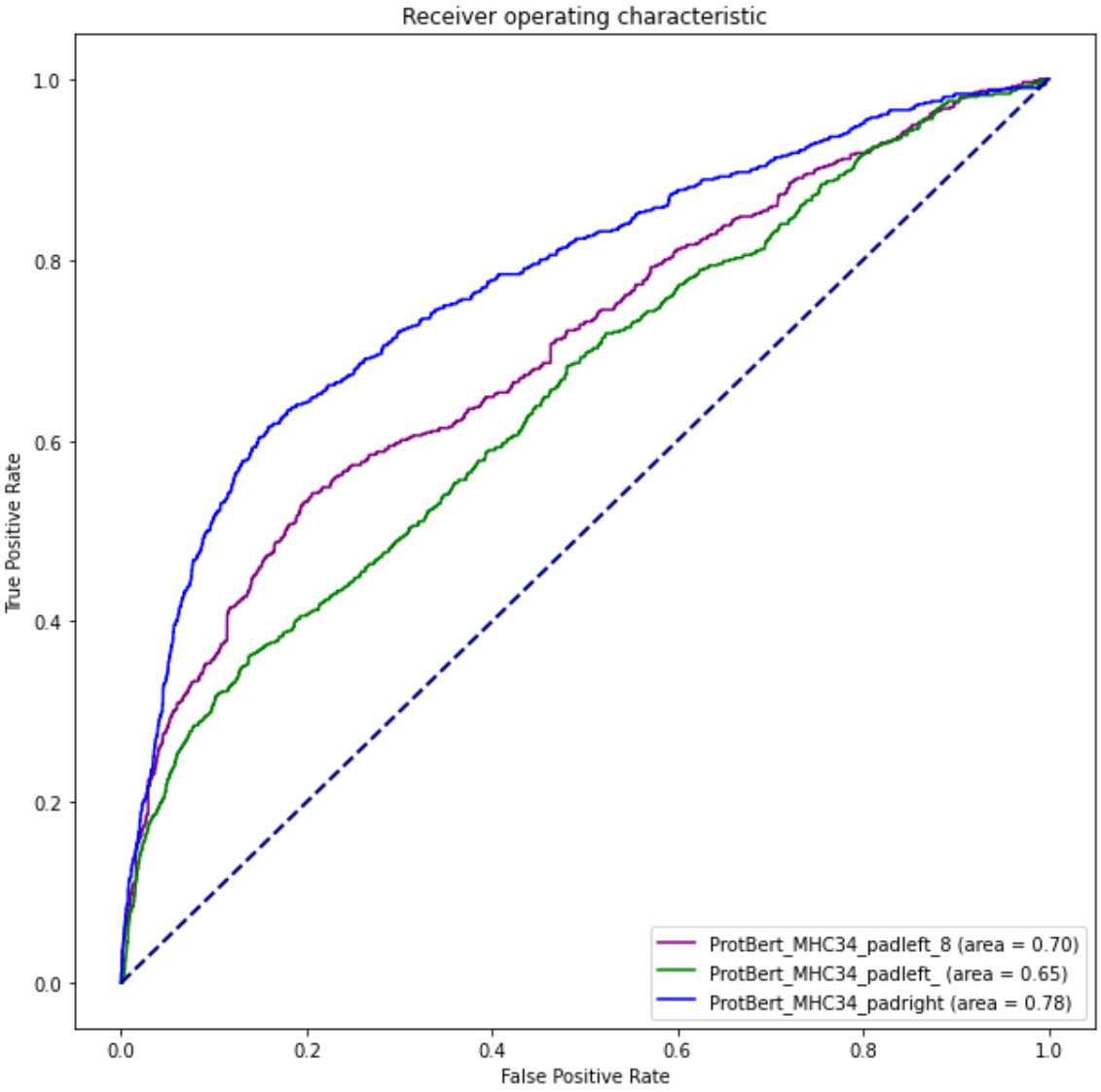

圖 26 不同的填充方式對 TCR-pMHC 模型預測結果 ROC curve 比較，padleft_8 為只選取 8 個胺基酸，padleft_為整段序列前填充，padright 為整段序列後填充

doi:10.6342/NTU202203779

4.5 Ensemble ProtBert 為編碼基礎的模型對 AUC 的影響

由於模型每一次訓練的差異性非常大，因此透過集成（Ensemble） 的方式平均每一次訓練的結果，以此達到更好的效能。本研究與最近發布的模型 pMTnet 做比較，pMTnet 為兩階段的模型，分為 TCR 和 pMHC 兩序列編碼以及 MLP 分類模型，表 6 為兩個模型 ROC curve 的 AUC 比較。

我們將整理過後的資料用 pMTnet 的原始碼重新進行預測，比起原始 pMTnet 論文的數值在 ROC curve AUC 降低了 0.015 在 PR curve AUC 上降低了 0.07。

圖 27 以及圖 28 為 Ensemble 過後與 pMTnet 比較的 ROC curve 以及 PR curve，我們發現在 Ensemble 後的模型比起 pMTnet ，在 ROC Curve 的 AUC 提高 0.035，表 7 為 PR curve 的 AUC 比較，在 PR Curve 的 AUC 提高 0.09。比起 pMTnet 需要預訓練 TCR 編碼器以及 pMHC 編碼器，我們的模型透過 ProtBert 已經預訓練好的蛋白質語言模型進行特徵擷取，能夠達到更好的預測效能。

表 6 Ensemble ProtBert 為編碼基礎的模型與 pMTnet 模型預測結果 ROC curve AUC 比較

Model	**ROC curve AUC**
pMTnet(Lu, Zhang et al. 2021)	0.827
pMTnet （本研究重新實作）	0.812
ProtBert-base ensemble	0.847

doi:10.6342/NTU202203779

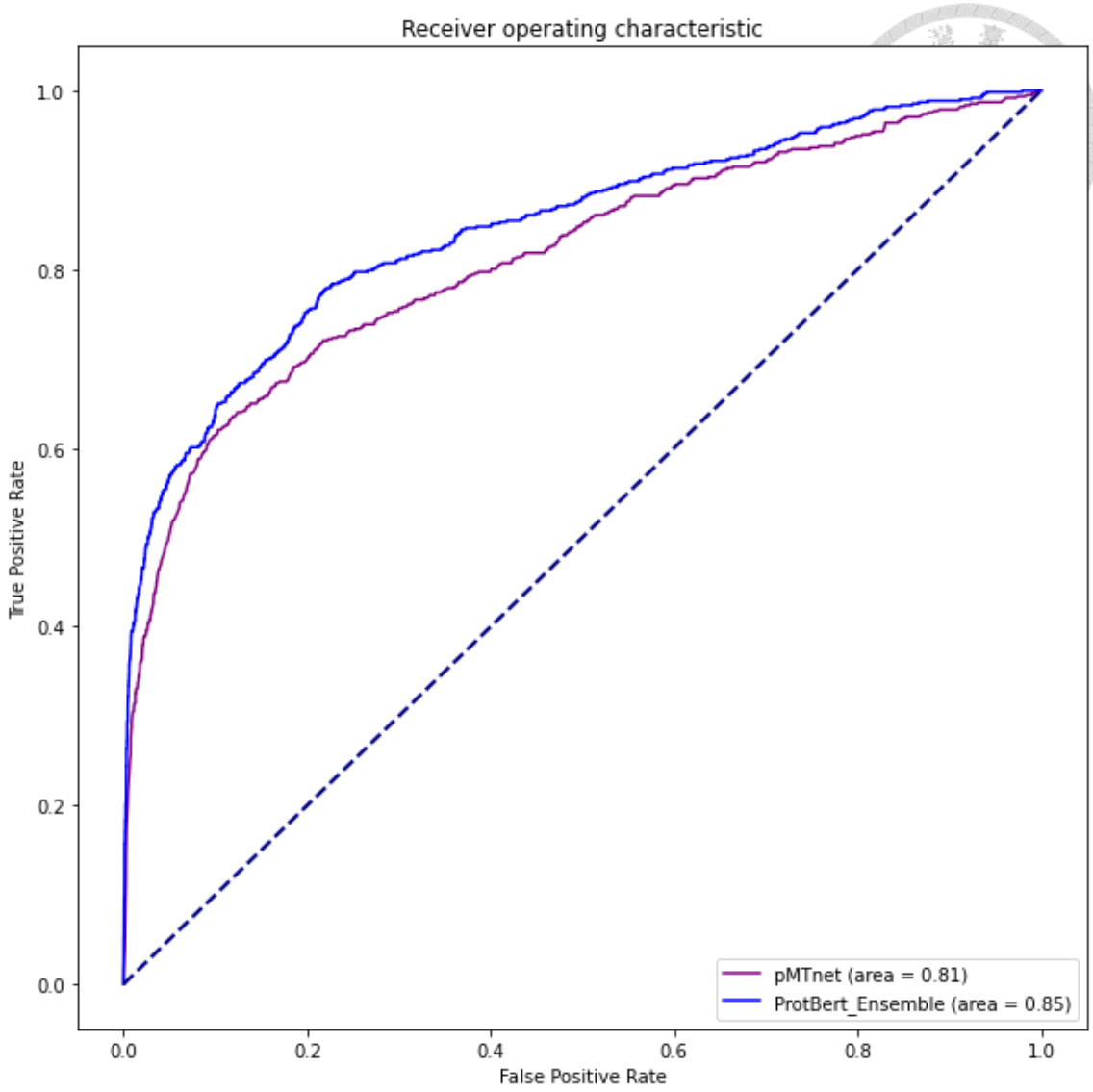

圖 27 Ensemble ProtBert 為編碼基礎的模型 ROC curve

表 7 Ensemble ProtBert 與 pMTnet 模型預測結果 PR curve AUC 比較

Model	**PR Curve AUC**
pMTnet(Lu, Zhang et al. 2021)	0.566
pMTnet （本研究重新實作）	0.494
ProtBert-base ensemble	0.585

doi:10.6342/NTU202203779

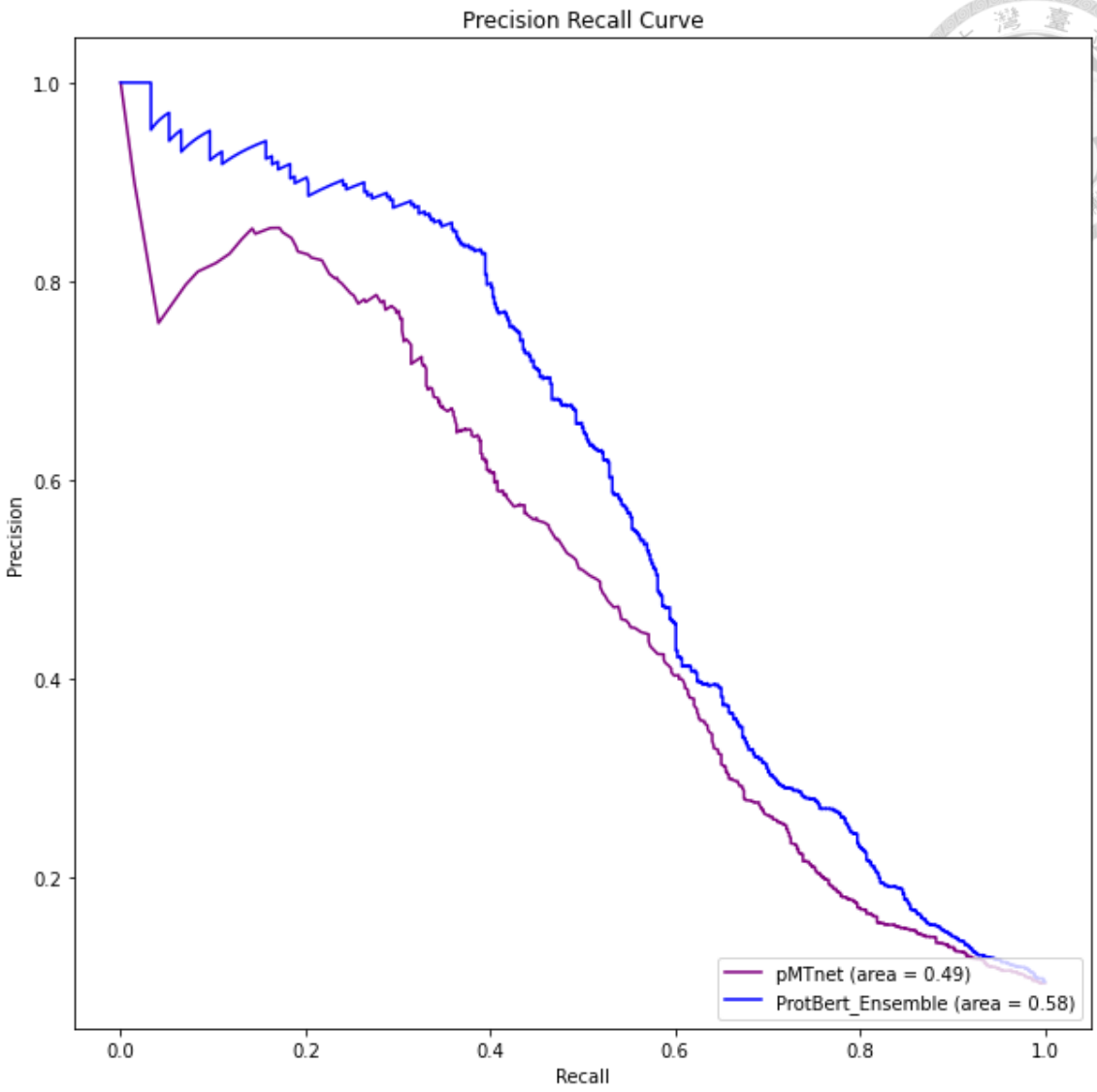

圖 28 Ensemble ProtBert 為編碼基礎的模型 PR curve

4.6 訓練時刪除特定的等位基因群對預測的影響

用原先訓練好的模型與刪除特定基因群的模型來比較，測試集分別是 HLA-A*11 以及 HLA-B*57， 圖 29 為刪除 HLA-A*11 等位基因群對 ROC curve AUC 的比較圖，黃色線使用的模型是在訓練資料中有包含 A*11 的資料，藍色線沒有，最後在測試集中 A*11 的資料中進行驗證。圖 30 為刪除 HLA-B*57 等位基因群對 ROC curve AUC 的比較圖，黃色線使用的模型是在訓練資料中有包含 B*57 的資料，藍色線沒有，最後在測

doi:10.6342/NTU202203779

試集中 B*57 的資料中進行驗證。從結果來看刪除特定基因群後會降低準確率，數據越多下降的越多，未來可以嘗試更小的族群像是專注在 HLA 蛋白質上，或許會有更好的結果。

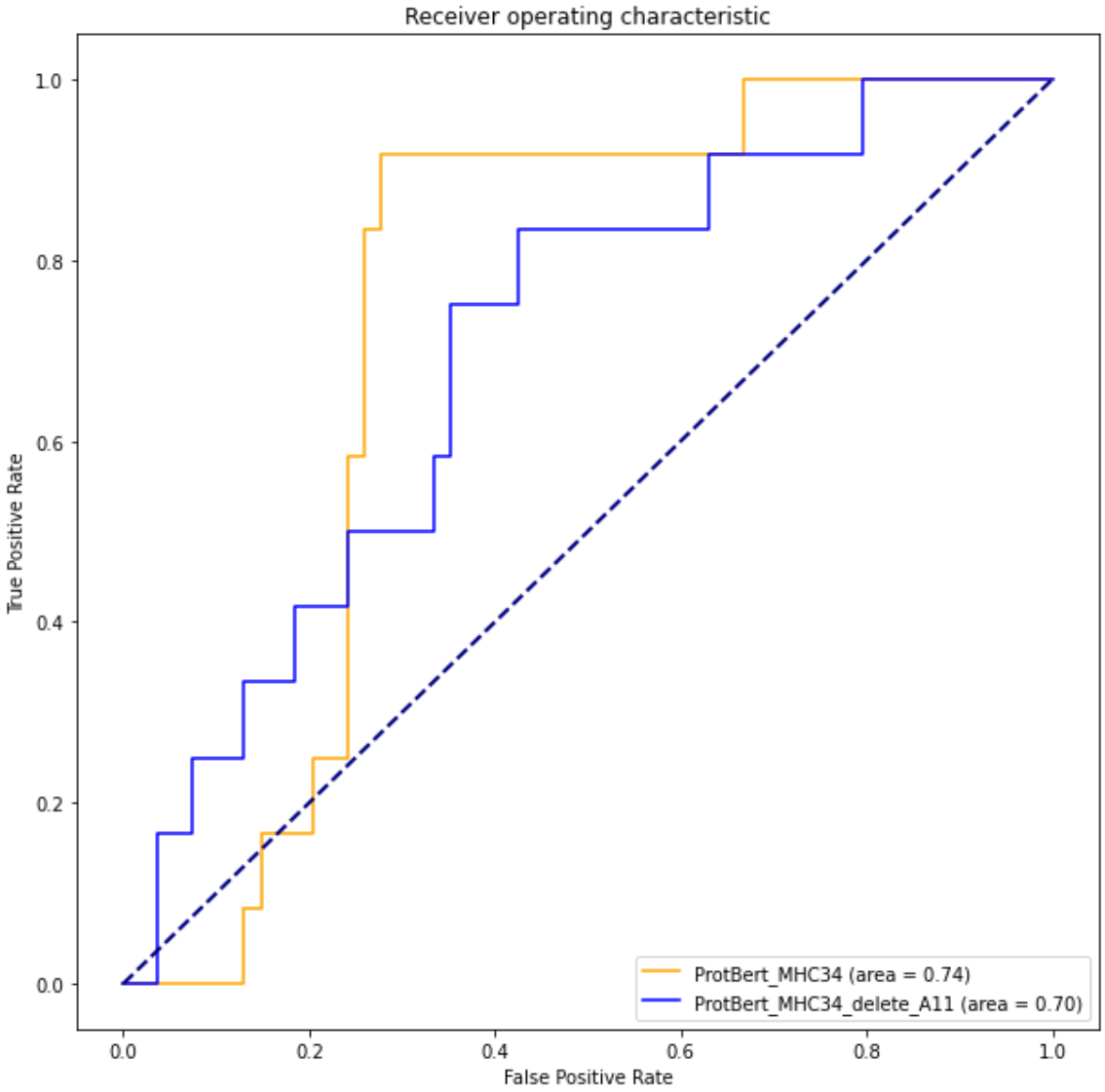

圖 29 刪除 HLA-A*11 對於 ROC curve AUC 影響

doi:10.6342/NTU202203779

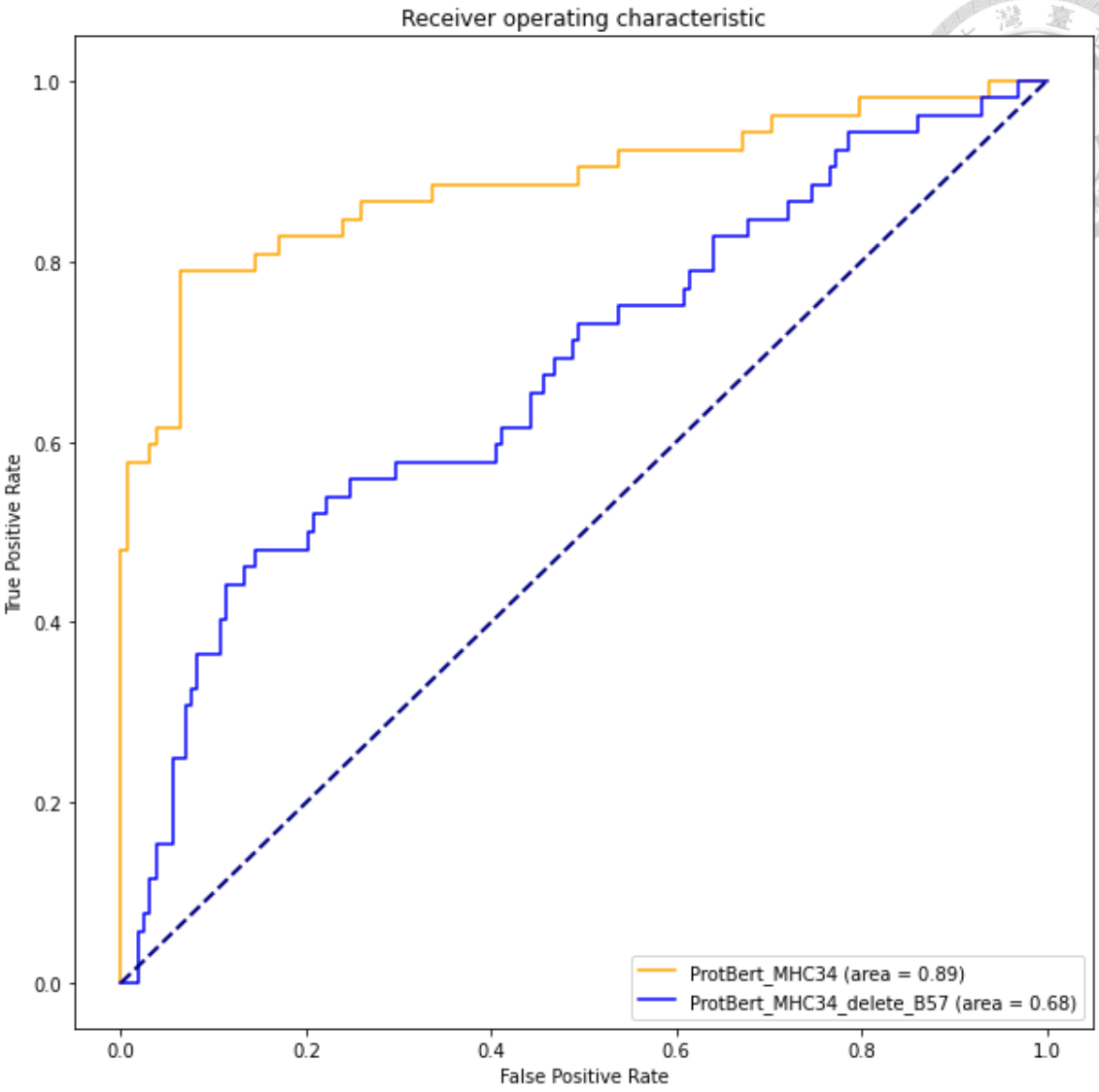

圖 30 刪除 HLA-B*11 對於 ROC curve AUC 影響

doi:10.6342/NTU202203779

4.7 探討加入 TCRα的資訊對 AUC 的影響

蒐集了 VDJdb 以及 McPAS-TCR 兩個資料庫，選定有同時包含 TCR α、TCR β、peptide 以及 MHC-I 的資料，而這些資料為正資料集，負資料集通過隨機匹配的方式擴增 5 倍。

模型方面將 TCR α、TCR β、peptide、MHC-I 序列輸入進一維捲積網路的特徵擷取，合併之後在輸入進分類器進行分類，從結果中可以發現，有加入 TCR α這個資訊可以提升約 0.1 的 AUC，圖 31 為加入 TCR α這個特徵的比較圖。

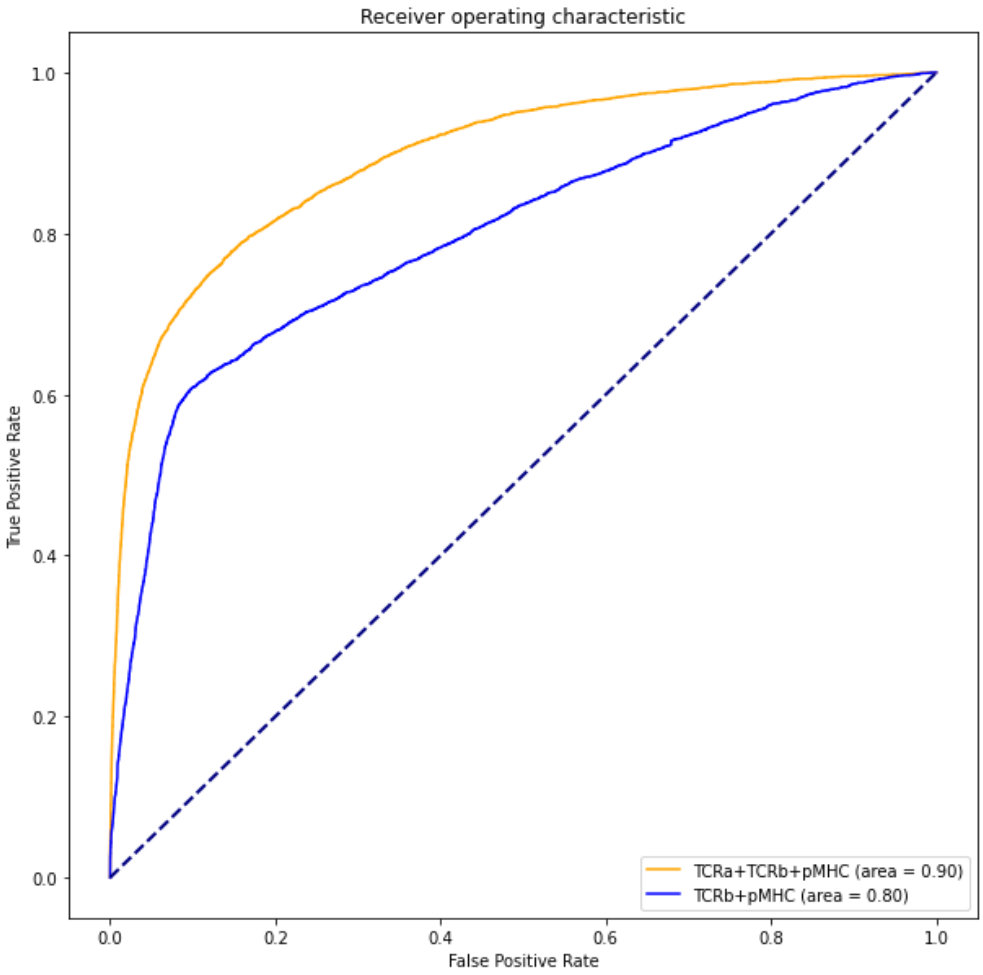

圖 31 加入 TCR α 後對於 ROC curve AUC 比較圖

doi:10.6342/NTU202203779

第五章 結論

預測 TCR 與 pMHC 結合是研究免疫系統反應的一個重要問題，並且已經開發了許多方法來解決這個問題，最近的一些工作利用蛋白質語言模型來解決各種蛋白質結合的問題，目前更多的研究釋出了更多更大信息量的模型像是 ProtTrans 接受了超過 2.5 億個蛋白質序列。本研究能夠預測 I 類 peptide-MHC 與 TCR 結合的特異性，只要給定 TCR 序列、peptide 序列與 MHC-I 類型。本研究開發創新的算法設計模型，透過微調 ProtBert 這個蛋白質語言模型用做特徵擷取，與 BLOSUM50 以及 One-Hot encoding 進行比較，先利用 NetMHCpan 的資料來驗證利用 ProtBert 能夠進一步提升預測 peptide 與 MHC-I 的預測。進一步探討了在 TCR 與 pMHC 結合預測上，得到一樣的結果，因此更加肯定 ProtBert 優於另外兩種蛋白質編碼工具。本研究也探討了在 TCR 與 peptide 的預測結合中，加入 MHC-I 的資訊能夠有效的提升預測準確率。最後利用集成學習的方式，對 ProtBert 為基礎的模型進行集成訓練，利用此方法可以在目前的數據分析上優於其他的機器學習模型 pMTnet ，在 ROC curve 的 AUC 高過 pMTnet 0.035，並且在 PR curve 的 AUC 高過 0.09。

雖然在結果上優於過去的研究，本研究也發現加入 TCR α的資訊能夠提升準確率，但在 TCR-pMHC 預測上有個關鍵的問題是準確負數據的可用性以及正數據集的數量，因此未來需要增加更多的正數據集，以及產生真正的負數據，是這項任務必須解決的問題，才能讓 TCR-pMHC 結合預測上以及在遇到新的 MHC-I 時能夠更加準確。

doi:10.6342/NTU202203779

第六章 參考文獻

Devlin, J., M.-W. Chang, K. Lee and K. Toutanova (2018) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv:1810.04805.

Elnaggar, A., M. Heinzinger, C. Dallago, G. Rihawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik and B. Rost (2020) "ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing." arXiv:2007.06225.

Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." Proc Natl Acad Sci U S A **89**(22): 10915-10919.

Krogsgaard, M. and M. M. Davis (2005). "How T cells 'see' antigen." Nature Immunology **6**(3): 239-245.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut (2019) "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." arXiv:1909.11942.

Lee, K.-H., Y.-C. Chang, T.-F. Chen, H.-F. Juan, H.-K. Tsai and C.-Y. Chen (2021). "Connecting MHC-I-binding motifs with HLA alleles via deep learning." Communications Biology **4**(1): 1194.

Lu, T., Z. Zhang, J. Zhu, Y. Wang, P. Jiang, X. Xiao, C. Bernatchez, J. V. Heymach, D. L. Gibbons, J. Wang, L. Xu, A. Reuben and T. Wang (2021). "Deep learning-based prediction of the T cell receptor–antigen binding specificity." Nature Machine Intelligence **3**(10): 864-875.

Montemurro, A., V. Schuster, H. R. Povlsen, A. K. Bentzen, V. Jurtz, W. D. Chronister, A. Crinklaw, S. R. Hadrup, O. Winther, B. Peters, L. E. Jessen and M. Nielsen (2021). "NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data." Communications Biology **4**(1): 1060.

Nielsen, M., C. Lundegaard, T. Blicher, K. Lamberth, M. Harndahl, S. Justesen, G. Røder, B. Peters, A. Sette, O. Lund and S. Buus (2007). "NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence." PLOS ONE **2**(8): e796.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu (2019) "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." arXiv:1910.10683.

Rao, R., J. Meier, T. Sercu, S. Ovchinnikov and A. Rives (2020). "Transformer protein language models are unsupervised structure learners." bioRxiv: 2020.2012.2015.422761.

Reynisson, B., B. Alvarez, S. Paul, B. Peters and M. Nielsen (2020). "NetMHCpan-4.1 and

doi:10.6342/NTU202203779

NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data." Nucleic Acids Res **48**(W1): W449-w454.

Shugay, M., D. V. Bagaev, I. V. Zvyagin, R. M. Vroomans, J. C. Crawford, G. Dolton, E. A. Komech, A. L. Sycheva, A. E. Koneva, E. S. Egorov, A. V. Eliseev, E. Van Dyk, P. Dash, M. Attaf, C. Rius, K. Ladell, J. E. McLaren, K. K. Matthews, E B. Clemens, D. C. Douek, F. Luciani, D. van Baarle, K. Kedzierska, C. Kesmir, P. G. Thomas, D. A. Price, A. K. Sewell and D. M. Chudakov (2017). "VDJdb: a curated database of T-cell receptor sequences with known antigen specificity." Nucleic Acids Research **46**(D1): D419-D427.

Springer, I., N. Tickotsky and Y. Louzoun (2021). "Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction." Frontiers in Immunology **12**.

Tickotsky, N., T. Sagiv, J. Prilusky, E. Shifrut and N. Friedman (2017). "McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences." Bioinformatics **33**(18): 2924-2929.

Vita, R., S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler, A. Sette and B. Peters (2019). "The Immune Epitope Database (IEDB): 2018 update." Nucleic Acids Res **47**(D1): D339-d343.

doi:10.6342/NTU202203779