以深度學習做蛋白質與配體結合之快速篩檢預測

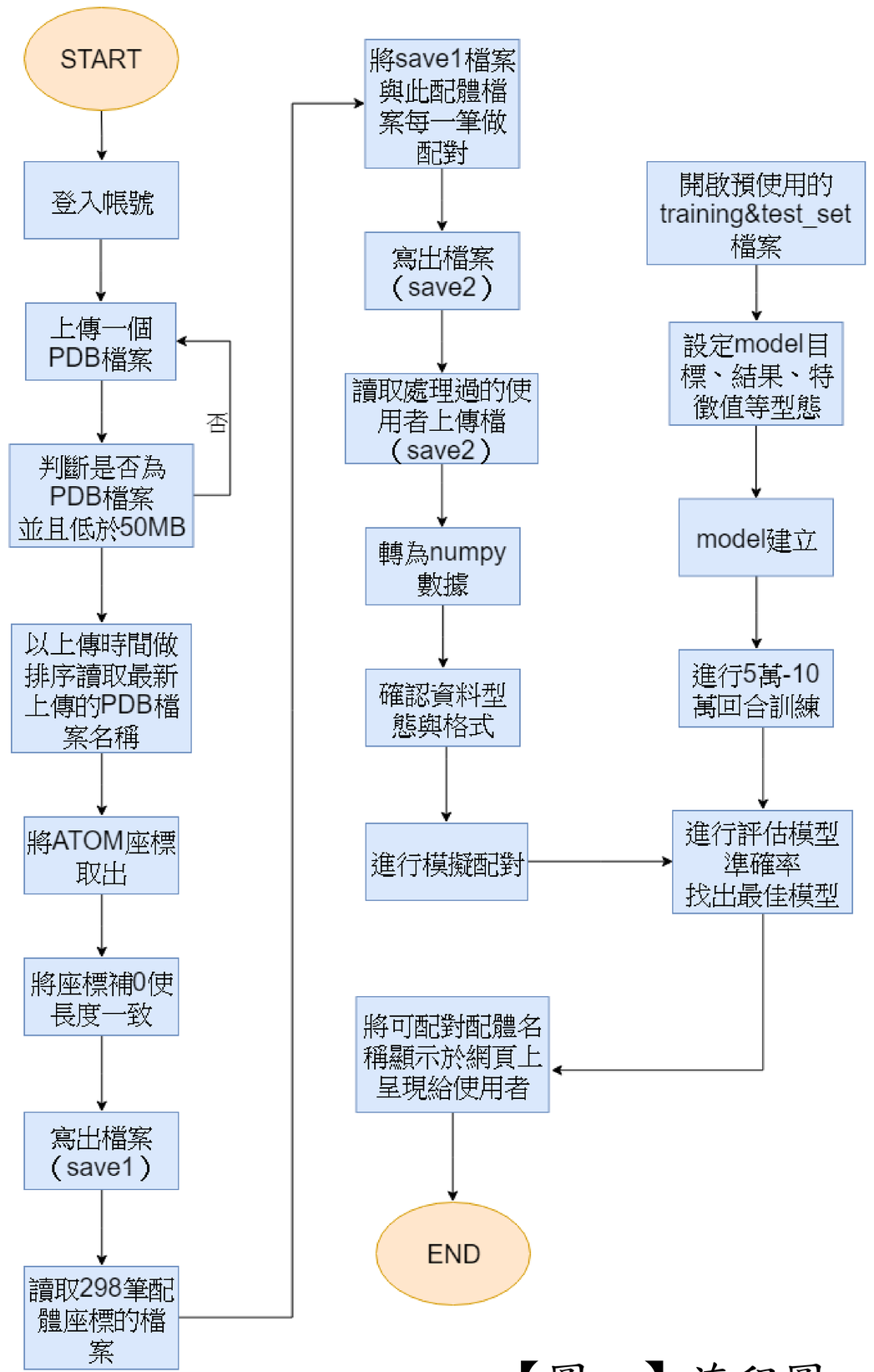
慈濟大學

指導老師：許弘駿 教授　組員：林欣霓、吳欣純、李孟庭

摘要

本專題實驗使用深度學習的技術進行預測，選用DNN（Deep Neural Network）網路來進行蛋白質與配體結合預測，我們是針對單一特定物種與抗生素相關之蛋白質，從整理出的配體中，在實驗時篩選出可能接合之配體，而選出的配體是與該蛋白質結合的可能性較高的。在實驗與研究時，可以利用篩選出的配體進行操作，使實驗者節省時間與成本。蛋白質的上傳與可結合之配體於網頁上執行以及呈現。

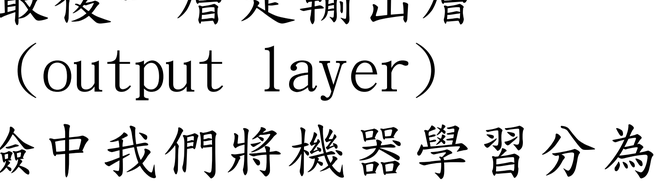
流程圖



【圖一】流程圖

DNN概念與應用

DNN（DeepNeural Networks）



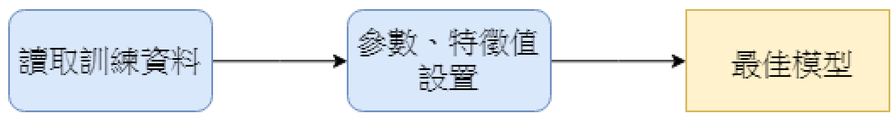
【圖二】DNN示意圖

DNN神經網路層可以分為：

- 第一層是輸入層（input layer）
- 中間的層數都是隱藏層（hidden layer）
- 最後一層是輸出層（output layer）

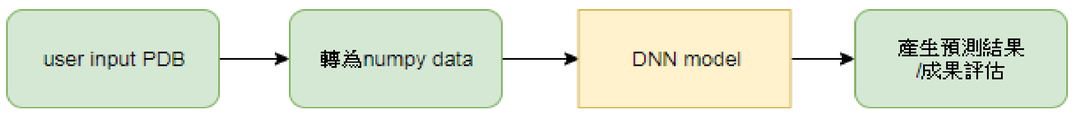
在實驗中我們將機器學習分為二個階段

→訓練階段：讀取訓練資料（CSV檔第一列放此檔案有幾筆資料及特徵值長度），進行模型訓練調整（模型準確率判斷），找出實驗最佳模型。



【圖三】訓練階段

→預測階段（使用者模擬）：將使用者上傳的檔案讀取轉為numpy檔，利用最佳模型，進行配體預測（最後將預測的結果與正確答案做正確率、召回率、F1評估）。



【圖四】預測階段（使用者模擬）

研究工具

硬體設備　電腦環境 Windows 10

軟體

Python　Tensorflow

xampp　php

drupal

以深度學習做蛋白質與配體結合之快速篩檢預測

慈濟大學

指導老師：許弘駿 教授　組員：林欣霓、吳欣純、李孟庭

資料處理→訓練資料

進行DNN訓練的資料需為CSV檔。目前訓練資料為412筆，檔案正確與錯誤資料數量各一半。而資料來源為大腸桿菌且為抗生素類型的蛋白質和配體座標，將長度不一致的座標後面補0，使每一筆蛋白質資料長度一致，再加上資料正確=1或錯誤=0〔蛋白質29670, 配體732, 結果1/0〕，訓練資料特徵值總長度：30402（不包含結果1或0）。

資料處理→測試資料

取出每種大腸桿菌且為抗生素類型的配體座標→共298筆，依照訓練資料時的配體長度補0，長度為732。

我們模擬使用者上傳的蛋白質PDB檔案，取其XYZ座標與整理好的298筆配體前後接合，測試資料特徵值長度為30402。

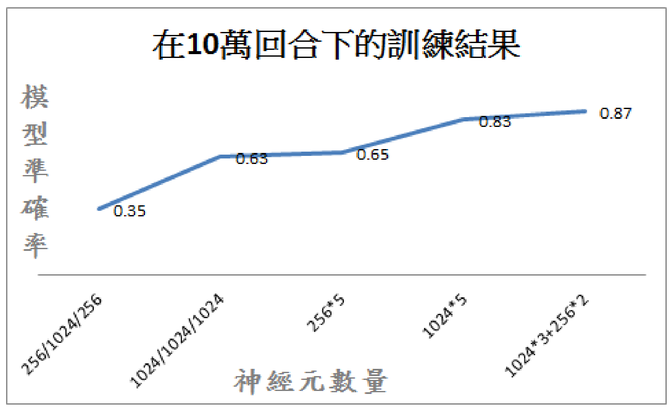
網頁呈現

【圖五】首頁

【圖六】結果呈現

【圖七】上傳PDB檔案介面

實驗結果與討論



【圖八】訓練結果

我們最終實驗結果，選用5層神經層，其每層神經元為1024，1024，1024，256，256的模型進行10萬回合的運算，模型準確率0.87，預測配體結果的評估為正確率90.6%、召回率44%、F1=6.6%。

結果不理想的原因可能為：

（1）蛋白質與配體結合專一性太高

（2）資料量太少→大腸桿菌且為抗生素類型的蛋白質資料蒐集不易，在深度學習訓練上，資料量的龐大有助於準確率上升

未來可以嘗試不同神經網路與嘗試蒐集更多資料。