

2019 年臺灣國際科學展覽會 優勝作品專輯

作品編號 190013

參展科別 電腦科學與資訊工程

作品名稱 利用深度學習預測中草藥的藥性功能與毒性

得獎獎項 大會獎：二等獎
出國正選代表

就讀學校 南投縣私立普台高級中學

指導教師 李紹銘

作者姓名 李致翰

關鍵詞 中草藥、深度學習、化學結構

作者簡介



我是李致翰，目前就讀普台高中二年級。自從去過幾次程式設計與人工智慧的暑期營隊後，我就開始喜歡上這個領域，並在空閒時間自學寫程式和閱讀人工智能相關文章。對於機器學習，我特別想要了解它對其他領域有什麼幫助，也就是如何使用其他領域的數據，並建構確實能夠分析那些資料的機器學習模型。雖然過程中總是被大量的名詞和知識弄得頭昏腦脹，但最後我對能建立模型並弄懂其背後原理感到很有成就。

摘要

近年來中醫在慢性病的治療上已獲得很大的進展，許多中草藥的功能與特性皆是利用人體實驗來找出其效用，中藥內含的成分極為複雜，其功能與毒性測試常依賴於過去經驗醫學，許多中藥仍有待探討與實證，其對人體細胞和基因的影響仍不如西醫。因此目前國內外紛紛開始建置中草藥、成分、化合物相關數據庫供查詢，然而如何讓中醫的功能與毒性科學化變得是很重要的議題。本研究利用化學結構的數據分析來探討中藥成分與人體臟器與毒性的關係性。我們採用深度學習模型以中草藥的化學性質作為輸入，透過化學結構的圖像傳播，來預測中草藥相對應之臟器有效功能與毒性預測，希望透過本研究可以提供中藥對健康影響的依據並作為未來輔助中醫的工具，讓人們可更加了解食用中藥對人體可能有的正面與負面影響。

Abstract

With the advancement of drug research, there was significant achievement in treating chronic diseases with chinese medicine. Chinese medicine has higher variety and more complex composition and it is difficult to quantitatively analyze its functions and toxicity. From now on, there is still a large number of Chinese medicine waiting to excavated. Therefore, many countries had in-depth study on the functional pharmacology of Chinese medicine, and drug safety. They built up databases of Chinese herbal medicines, ingredients and compounds for enquiry. This study aims to develop predict graph convolutional neural network method to predict the function and the toxic effect of Chinese herbal medicine on human body parts. We provided the basis for evaluating the health effects of Chinese herbal medicines for further research and development.

一、研究動機與目的

全球普遍接受西方科學醫學，其講求以科學精準的方法來確定有效性和安全性；近年來，中醫藥日益受到全球矚目與重視，然而其功能與毒性測試常依賴於過去經驗醫學，許多中藥仍有待探討與實證，其對人體細胞和基因的影響仍不如西醫。因此以科學觀點易為人詬病且較不為大眾接受，中藥的現代化也是時勢所趨。另一方面，中藥治療逐漸普及並被廣泛的利用，很多人誤以為中草藥來自天然必定安全無毒，其實大部分的中藥都各有其毒性。例如：含馬兜鈴酸及相關化合物的有毒草藥，很可能是導致亞洲尤其是中國內地和台灣地區肝癌的重要原因之一。前全球已有多種草藥被挖掘出成分及其化合物結構，配合中藥化學的相關知識，或由分子生物學相關的實驗而得的資料，利用現代化學分析和生物技術，探索中藥的有效成分以及毒性測試，不僅有利於發揮中醫藥治療疾病的優勢，同時成為現代醫學對於中醫藥的研究提供廣闊的思路。中國及西方國家致力發展建置中藥化學結構資料庫，將中藥物質、藥理作用、指紋圖譜及用藥安全的研究紀錄於中藥庫網站如 TCMID、Herding 等(Choi et al., 2016; Xue et al., 2012)，提供草藥成分、化合物成分等供大眾查詢。

我們發現以前曾有使用大數據來完成「老藥新用」的目的。過去曾經有醫學界人士發現舊藥除了原本的療效以外，其實還另有別的效用，進而發現「老藥的新用」比起直接研發新藥的來得容易，且在藥品通過政府許可時，比起新開發的藥物會快上許多。但是測試過程一樣需要相當多的資金及時間來做實驗驗證，非常不划算，因此為了解決這個問題，科技公司或電腦科學家開發出了大數據分析工具，例如本研究使用的 Deepchem 是一款史丹佛大學領導的開放原始碼庫，專門用在對藥物研發、量子化學及生物的深度學習上。這些工具都能被用在建立能夠透過規律推測真實可能會發生的情況的模型，節省作實驗所需的時間。不只西藥有用所謂的人工智慧做過研究，中醫也有，舉例來說，美國國家生物技術信息中心在去年發表過一篇論文，其使用卷積神經網路和人臉圖像預測一個人在中醫中分的九大體質：平和質、氣虛質、陽虛質、陰虛質、痰濕質、濕熱質、血瘀質、氣鬱質、和特稟質，並且最後準確率到達 65.29%。但在過往的藥品研究中，對於西藥的研究相對較為充足，在中藥材的方面著墨較少，因此本研究也效仿這些過去的實驗方法，應用在中藥上，嘗試進行藥效的預測，並試圖達到更好的效果。

民眾對於中草藥的功能不瞭解，加上無研究驗證，使得民眾認為中醫食補養生觀念並不科學。中草藥的有效功能稱為歸經是中醫千年來的用藥理論，是指食物或藥物作用在人體的臟腑經絡的治療作用或保健功效。人體經絡分布在人體，如同幾條高速公路一般，構成身體的「經絡」系統，每條經絡有其對應的臟器，中醫認為透過五臟六腑彼此聯繫的網絡，是將人的血氣運行、及營養和廢物傳達到身體各個部位，主要用於治療和保健的作用。經脈包含心經、胃經、脾經、腎經、肝經、膽經、肺經、小腸經、大腸經、膀胱經、心包經、三焦經共十二經絡

(臟器)。本研究旨在以我們已知中草藥的化合物結構，進行化學分子描述後，將其輸入圖卷積神經網路進行建模，進而預測臟器與毒性之關聯性，瞭解化合物結構對人體哪些臟器是有幫助的或是具有毒性，以做為未來中草藥成分對健康影響評估之依據。

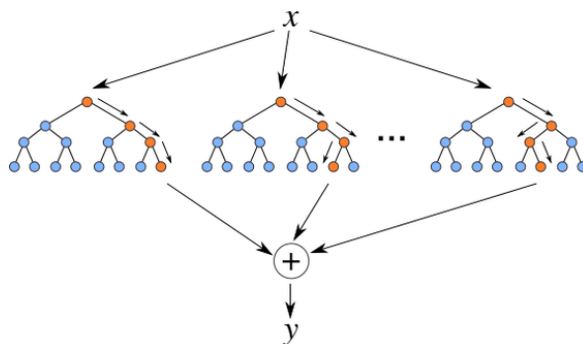
二、研究相關理論

(一)分子科學

分子科學是一個跨學科的知識領域，主要研究分子的結構與功能、相互作用、結構和各種物理化學性質。要了解這些分子的性質，化學家需進行不斷的實驗，以證實成效且傳統的數值模擬方法對一些極為複雜的問題是無能為力的。此時，機器學習憑藉高維空間中的解析能力，進一步對複雜問題進行抽象、簡化與估算。傳統的化學家們為了瞭解分子的性質必須不斷實驗以驗證，不僅費時費力，效率也非常低。但是近年來由於人共智慧及機器學習的普及和發展，化學家已經可以透過電腦模型來做一般實驗無法做的預測分析，現在更有相關的『工具』（演算法，分子特徵化系統等等）供使用。透過機器學習的方式來研究實驗對象，幫助研究者從已知的生化反應進而構建清晰直觀的模型，來評估化合物之風險高低，可能有助於開發出代替傳統用動物實驗測量化合物性質的方法。預測分子的性質是一個尤其重要的問題，需要先進的機器學習技術以及圖形表示學習算法。近年來機器學習的研究從應用於圖片或是語音方面，轉而協助科學研究，今年的研究指出利用神經信息傳遞（MPNN：Message Passing Neural Network）模型實踐在化學領域的搜索問題上，也就是說利用分子數據的自然表徵為一個圖（graph），其中原子是節點（node）、鍵結是邊（edge），而鍵結的強弱及代表邊的權重。這樣的想法就如同在 ImageNet 中的所提出的卷積神經網路(CNN)，是利用在圖像數據的區域特徵以及不變性的知識，往往能更好地進行歸納與推理來達到優良的效果(Altae-Tran, Ramsundar, Pappu, & Pande, 2017)。在自然語言的處理上，也有利用句子所表達的依存結構樹(dependency tree)來有效的預測語言上的情緒。(Wu et al., 2018)MoleculeNet 專題提供多個公共生化數據集並開發之前提出的多個分子特徵化(molecular featurization)和學習演算法的開源程式碼實現(DeepChem)。在 MoleculeNet 專題中已經實現的基於圖形模型的核心結構(如:卷積模型、有向無環圖模型、消息傳遞、深度張量(Tensor)等方法(Kearnes, McCloskey, Berndl, Pande, & Riley, 2016)。

(二)機器學習

首先，我們先介紹以下分類方法：隨機森林、支持向量機、邏輯式回歸、類神經網路與卷積神經網路。決策樹是隨機森林(Random Forest)的基本決策模型，隨機森林生成很多決策樹，每棵樹依賴於從原本的資料 x 來進行隨機取樣資料或是特徵變數，得到不同的訓練資料後，根據訓練資料中某些屬性或資料進行分類判別時，得到一群決策樹分類器。隨機森林中的每一棵決策樹都會給出自己的分類選擇，整體分類的結果採平均的分數作為輸出的決策 y ，如圖一。



圖一 隨機森林概念(取自 <https://lz5z.com/机器学习常用算法—随机森林/>)

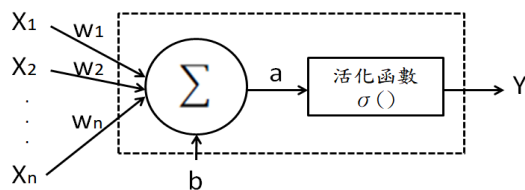
有利於事件發生的勝算定義為事件發生的機率除以事件未發生的機率。變數的勝算比和相對應的迴歸係數之間存在一獨特的關係。邏輯式迴歸(Logistic Regression)，是對一個或多個自變數 x 與應變數 y 之間關係做建模的迴歸分析，每個自變數在羅吉斯迴歸方程式中能表示勝算比等於勝算比和自變數的係數之間的關係使我們於進行模行參數估計時更易計算勝算比的估計值。基本上與線性迴歸(Linear Regression)的概念相同，但對於連續性的資料用線性迴歸，二元的資料適合用邏輯式迴歸。

深度學習(Deep Learning)是機器學習(Machine Learning)中一種基於對資料進行特徵學習的方式，透過訓練資料逐層特徵提取進行分類與預測，能夠用於資訊檢索、分類和迴歸等特定問題中。現在深度學習中在許多領域上都有突破性的發展，例如語音及音頻識別、影像分類與識別、影像檢索等。深度學習的概念源於類神經網路，其一種仿照生物神經網路結構和功能的計算模型，在結構上如圖二，我們把一個神經網路劃分為輸入層、輸出層和隱藏層。輸入層的每個結點對應一個可預測的變數，而輸出層的結點則對應目標變數。在輸入層和輸出層之間是隱藏層，而隱藏層的層數和每層內部的結點(神經元)個數則決定了神經網路的複雜度。其中， x 為輸入， W 為權重， b 為偏移量，單一神經元使用線性方程式將輸入總和成 a ，如下面之(1)方程式：

$$a = \sum_{i=1}^n w_i x_i + b \quad (1)$$

再將 a 值經由活化函數後，得到單一神經元的輸出 y ，其較易地處理解釋變數的非線性和互動效應。如下面之(2)方程式：

$$y = \sigma(a) = \frac{1}{1 + e^{-a}} \quad (2)$$



圖二 類神經網路(單一神經元為例)

實際上神經元數與隱藏層數目則依不同的問題個別定出適當的數目，用以表現輸入人工神經元間的交互作用，類神經網路必須透過訓練的方式反覆學習，讓輸入都能正確對應到其輸出類別，越來越接近目標值。隨著訓練次數(epoch)的增加，我們將計算所得輸出值與期望目標輸出值之差異，透過隨機梯度下降法(Stochastic Gradient Descent, SGD)得到類神經網路建構中的權重值的修正量逐步調整，使得目標值與神經網路之輸出值間的誤差越來越小，透過來回計算來產生最佳的權重。卷積神經網路 (Convolutional Neural Network，簡稱 CNN) 是早期最廣泛的深度學習模型，主要用以處理影像結構的演算法，CNN 的作法是用移動視窗(Sliding Windows)的概念，利用固定大小的移動視窗，逐一掃描整張圖片，將每次框出來的圖像變成特徵向量維新的輸出以判斷類別歸類。傳統的卷積神經網路從低維度且規則的網格(Grid)來學習局部固有結構 (Local Stationary Structures) 並將它們組合成多種尺度分層模式來處理圖像、影片及聲音辨識的工作。儘管卷積神經網路被證明可成功解決大量機器學習問題，但是在很多真實問題中，數據所產生的關聯圖像並不規則。對於我們所要達成的目標：如何能從中草藥的化合物結構對應於臟器的有效性與毒性，我們意識到以下需解決的問題：

1、決定一個分子的特性的並不是分子的整體，而是其中的各個部分特徵，例如碳酸基就是一個特徵。就如同圖像中的特徵，也不全是整張照片而是部分。

2、圖像的特徵是像素，對於化合物化學結構我們可以採取分子指紋(fingerprint)的方式來表達

3、對於圖像的關聯性是屬於網格式，化學結構的部分我們可以採取圖廣播的方式來表達彼此之間的關聯性。

因此若能有效地解決上述的問題，建立起一個有效的模型有別於傳統圖像而適用於化學結構的圖形分析，必定是很重要的關鍵。

三、研究流程

研究將匯入中草藥化合物資料，並用分子指紋來對各個化學分子作描述，接著採用 Deepchem 的開放原始碼庫，卷積神經網路來進行分析建模，以下為研究流程及研究方法做說明。

(一)導入資料源

本研究採用國外 TCM-Mesh 網站資料，此網站特別收集中醫藥學相關資訊記錄於數據庫中，提供中藥製劑的網絡藥理學分析服務。目前該數據庫有包含草

藥 6,235 種，化合物 383,840 種，基因 14,298 種，病害 6,204 種，基因與疾病關聯 144,723 種，基因相互作用 3,440,231 對，副作用記錄 163,221 種，毒素記錄 71 種，也使用了 1,293 種 FDA 准核的藥物的化合物用於評估該數據庫(Zhang, Yu, Bai, & Ning, 2017)，該平台草藥來源取自 TCMID 數據庫。本研究整合網站提供的草藥及人體經絡對應關係資料，分別是「草藥詳細資訊」(herb-info.txt，herb-ingredients.txt)與「草藥對應化學藥品編號」(toxicity.txt)，其資料欄位如圖三，將草藥-成分-化合物結構做有系統性的連結，製出整合資料，以利預測草藥與經絡關聯性。整合後的欄位有：

- 1.草藥學名 (pinyin_name)
- 2.草藥英文名 (English name)
- 3.拉丁文 (Latin name)
- 4.性味 (properties)：根據味覺經絡的辨別及臨床治療反應效果而定的。
- 5.歸經 (Meridians)：分別是心經、胃經、脾經、腎經、肝經、膽經、肺經、小腸經、大腸經、膀胱經、心包經、三焦經共十二經絡。
- 6.草藥類別 (use part)：有分樹脂類、果實種子類及根莖類
- 7.療效 (effect)
- 8.應用的症狀 (indication)
- 9.記錄成分 (chemical)
- 10.化學藥品編號 (CID)

	herb	name	chemical	English.name	Latin.name	properties	meridians	use.part	effect
1	A WEI	kamolanol	CID046883037	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Spleen	balsam	To transform concretion and disperse
2	A WEI	kamolanol	CID046883037	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Stomach	balsam	To transform concretion and disperse
3	A WEI	kamolanol	CID046883037	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Liver	balsam	To transform concretion and disperse
4	A WEI	camphene	CID000006616	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Spleen	balsam	To transform concretion and disperse
5	A WEI	camphene	CID000006616	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Stomach	balsam	To transform concretion and disperse
6	A WEI	camphene	CID000006616	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Liver	balsam	To transform concretion and disperse
7	A WEI	conferol	CID011892267	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Spleen	balsam	To transform concretion and disperse
8	A WEI	conferol	CID011892267	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Stomach	balsam	To transform concretion and disperse
9	A WEI	conferol	CID011892267	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Liver	balsam	To transform concretion and disperse
10	A WEI	confertifolin	CID000442187	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Spleen	balsam	To transform concretion and disperse
11	A WEI	confertifolin	CID000442187	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Stomach	balsam	To transform concretion and disperse
12	A WEI	confertifolin	CID000442187	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Liver	balsam	To transform concretion and disperse
13	A WEI	assafoetidol b	CID000636584	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Spleen	balsam	To transform concretion and disperse
14	A WEI	assafoetidol b	CID000636584	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Stomach	balsam	To transform concretion and disperse
15	A WEI	assafoetidol b	CID000636584	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Liver	balsam	To transform concretion and disperse
16	A WEI	6-o-e-feruloylajugol	CID006325178	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Spleen	balsam	To transform concretion and disperse
17	A WEI	6-o-e-feruloylajugol	CID006325178	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Stomach	balsam	To transform concretion and disperse
18	A WEI	6-o-e-feruloylajugol	CID006325178	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Liver	balsam	To transform concretion and disperse
19	A WEI	beta-pinene	CID000014896	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Spleen	balsam	To transform concretion and disperse
20	A WEI	beta-pinene	CID000014896	Asafetida Giantfennel Resin	Ferula assafoetida	Warm,Pungent,Bitter	Stomach	balsam	To transform concretion and disperse

圖三 草藥-成分-化合物關聯

以人參（學名 REN SHEN）為例，性味田，微苦，歸心經、脾經及肺經，療效為補充原氣，恢復脈搏和乾涸，補脾健肺，產生液體，安養精神，降低血糖等。它的主要成分是 aposcopolamine，三磷酸腺苷 (adenosine triphosphate)，精氨酸雙果糖，葡萄糖 (argininyl-fructosyl-glucose)，腺嘌呤 (adenine)，木皂甙 A (araloside a)，生物素 (biotin)，菜油固醇(campesterol)，這些成分所屬的化學藥品編號(CID)分別為 6419932、5957、11968784、190、197091、171548、173183。(資

料取自 TCM-MESH 網站)

我們將預測化學子結構與中草藥有效功能與毒性的關聯性視為機器學習領域中的分類問題，由於一個中草藥可能對於多種臟器皆有其功效，所以又被歸類為多標記分類問題(multi-label classification)，如圖四所表示。

	herb	bladder	cardiovascular	gallbladder	heart	kidney	large intestine	liver	lung	small intestine	spleen	stomach	three end
0	A WEI	0	0	0	0	0	0	1	0	0	1	1	0
1	AI DI CHA	0	0	0	0	0	0	1	1	0	0	0	0
2	AI YE	0	0	0	0	1	0	1	0	0	1	0	0
3	AN XI XIANG	0	0	0	1	0	0	0	0	0	1	0	0
4	BA DOU	0	0	0	0	0	1	0	0	0	0	1	0

圖四 多標記分類問題

(二)資料前處理

1、剔除對應不到歸經的草藥

並不是每個草藥都有對應的歸經，因此剔除歸經對應不到歸經的資料，共計剔除了 700 多筆資料。

2、剔除未知化學成分的草藥

要做中藥科學研究的前提是要了解草藥中的化學成分，以化學的觀點研究中醫，部分草藥並沒有化學成分，因此將其去除，剔除未知化學成分的草藥共計 864 筆。

3、轉置化學結構(SMILES)

化學分子描述的方法，本研究採用最典型的「簡化分子線性輸入規範」(Simplified molecular input line entry specification, 簡稱 SMILES, 又稱化合物線性標記方法)。首先根據化學分子結構建置樹狀結構圖，須去掉氫(H)及打開環，被拆掉的鍵端原子用數字標記，針對樹狀結構做深度搜尋(DFS)，來轉化分子結構以轉化為簡單的線性向量字串，以利解讀化學分子結構。下圖三為以 Python 介面使用 RDKit 轉換 SMILES 的例圖五。(RDKit 是一款開源化學訊息學與機器學習工具包，提供 C++ 和 Python 的 API 接口。)

```

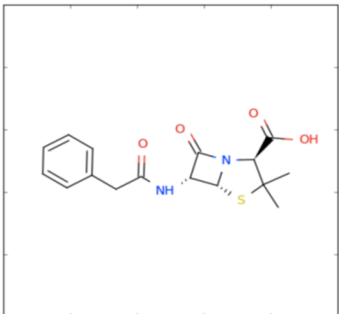
In [1]: # import libraries
from rdkit import Chem
from rdkit.Chem import Draw
import matplotlib.pyplot as plt
import matplotlib inline

# define molecule
penicillin_g_smiles = 'CC1([C@@H](N2[C@H](S1)[C@@H](C2=O)NC(=O)Cc3ccccc3)C(=O)O)C'
penicillin_g = Chem.MolFromSmiles(penicillin_g_smiles)

# draw molecule
Draw.MolToMPL(penicillin_g, size=(200, 200))

Out[1]:

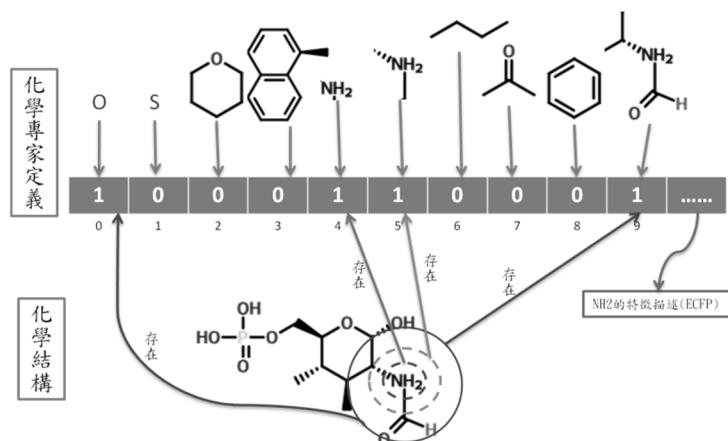
```



圖五-RDKit 轉換 SMILES 的例圖

4、化學結構轉成分子指紋(fingerprint)

化學資訊的描述分成靜態與動態兩種：靜態的描述包含目前廣泛應用於化學結構之特徵描述(fingerprint)，利用專家已知的化學小分子與已知的化學子結構如原子數、氫鍵等資訊進行描述，而有些是從化合物的化學結構來進一步計算其他的化學特性，如：重量、凡得瓦力等，若此化合物具有此結構或特性則標明 1；反之，則標明為 0。根據化學人定義好的化學問題做標記，例如：有沒有包含一些子結構在內，有則 1，反之無則 0，或是會不會融解、會不會分裂等問題。化學資訊的描述有很多種，其中一種是化學結構之特徵描述或稱指紋(Extended Connectivity Fingerprint or ECFP)。我們針對化合物元素做特徵描述，利用專家已知的化學子結構描述，其特徵維度的大小為固定（長度固定為 1024），透過 PubChem 開放式系統取得化學結構，化學結構裡有原子和原子之間的連結鍵，我們以非氫原子這些化學原子為節點（node）、原子之間的鍵則為邊（edge），找出以每個節點為圓心向外擴張至直徑距離為 2 的子結構，針對這個化學子結構計算其化學特性，最後得到化學分數，判斷是否包含專家認定的子結構特徵，以產出 0 或 1 的 ASCII 碼序列值，以此類推，造訪每個節點以產出該節點距離為 2 的子結構的特徵描述。分子指紋(ECFP)的產生過程如下圖六：

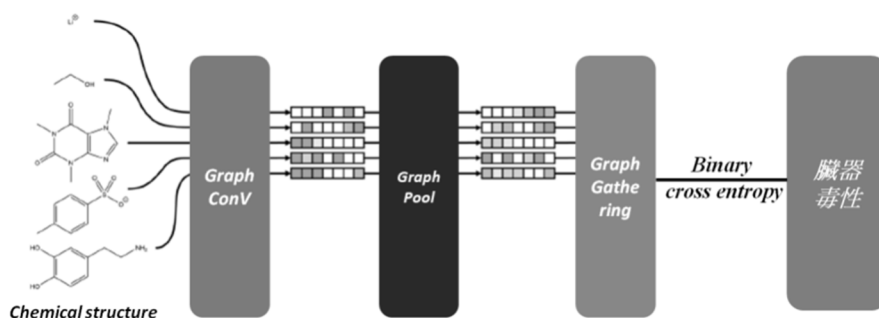


圖六-分子指紋(ECFP)的產生過程

分子 NH2 的特徵值為該分子以一串二進制位數值，代表分子中特定的子結構存在或不存在。以 NH2 為圓心，當距離為 0 時，第 4 個定義值為 1，表示第 4 個定義的子結構存在於 NH2 中；再增加為距離為 1 時，第 5 個定義的化學子結構為 1，表示該子結構存在；同理，當增加到距離為 2 時，第 9 個化學子結構為 1，以此類推，最後產出 NH2 二進位序列的特徵描述，這就是 ECFP 的過程。

(三)圖卷積網路(Graph CNN)

由於化學結構不同於影像的地方在於化學結構是部份相接的，並非像圖像一樣每個像素都有一樣多個相接的像素。我們使用圖卷積神經網路的過程如圖七，其功能包含圖特徵卷積、特徵最大池化以及圖結構累積化。本研究神經元隱藏層作法是化學子結構圖特徵卷積(GraphConv)→ 圖特徵最大池化(GraphPool)→ 圖特徵卷積 (GraphConv)→ 圖特徵最大池化 (GraphPool)→ 圖結構累積化 (GraphGather)。



圖七-圖卷積網路流程架構

1、輸入層

把化學子結構想像成是由數個像素組成，每個化學元素皆經由 ECFP 產出的化學元素 1024 個的化學指紋當作輸入層，在輸入資料模型之前，這些值可能會造成資料不準確或是偏頗於數值較高的特徵值，因此我們對

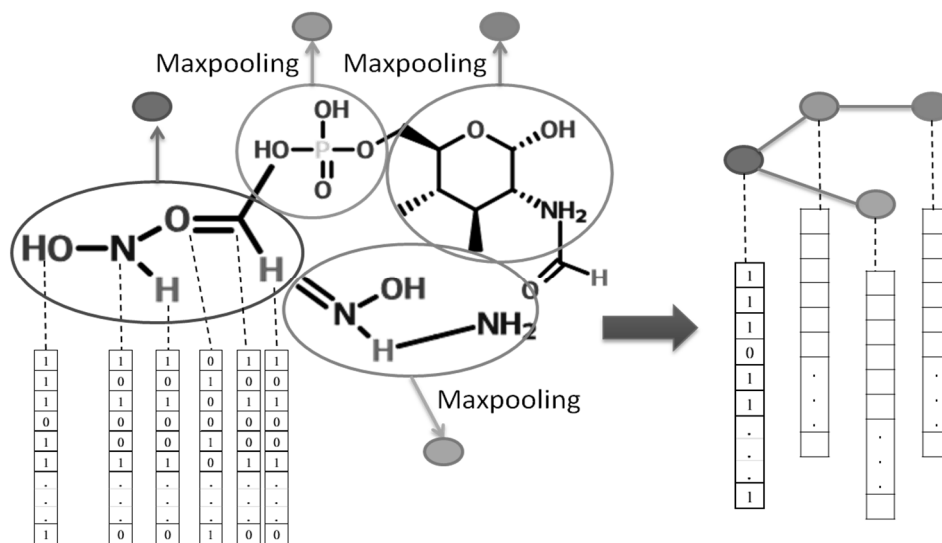
一維特徵值做正規化，以使資料不偏於極端值。

2、化學子結構圖特徵卷積

圖卷積網路的概念是用特徵的移動視窗(sliding window)掃描過所有的輸入層，產生出降為的特徵子結構，並且電腦學習將具有緊密性及關聯性較高的化學元素視為特徵群。

3、化學子結構圖特徵最大池化

將上述所產生出來的特徵群中最大關鍵特徵值為此一群的主特徵之代表，即為化學結構特徵的最大池(Graph Pool)，就像影像辨識 CNN 方法中的 max pooling 一樣，因為化學結構的每個元素及相鄰元素都有距離，可能為單鍵或雙鍵，所以為了找出輸入的特徵描述值與化學結構的關係，必須計算權重(weight)以取得加權值。要計算加權值之前，需要利用 KNN 的演算法來選取每一個節點的卷積的鄰域(neighbor vertex connectivity)，將一個節點視為中心點，把其鄰域的節點集結，最後加權為一個值，表達一個鄰域的核心結構。將緊密性及關聯性較高的化學元素視為一個移動視窗，特徵值的大小取決於化合物關聯程度，越是緊密關聯性的化合物，即歸在同一視窗中，最後從中取得最大的關鍵特徵值為此一群特徵之代表，即為化學子結構特徵最大池化。如圖八所表示。



圖八-圖特徵最大池化

4、整體化學圖結構累積化

前面我們將化學結構的簡化為幾個核心子結構，那麼現在我們將這些特徵再進一步加權為整個化學分子的整體特徵，這個累積將可以利用作為後續預測。

5、輸出層

最後一個步驟要探討擁有某些化合物的中草藥與人體臟器、毒性的關聯係，我們將其視為多個二元分類的問題，使用 sigmoid 函式使數字介於 0 到 1 之間，一個中草藥最後輸出的結果就是十三個值，對應到十二個經

絡以及毒性預測，而只有其有對應的功效與毒性值是 1，其他都是 0。

在訓練前，我們將資料分成十等分，八等份作為訓練集，剩下的兩等份作為測試集，導入訓練集進行模型訓練，圖卷積網路會透過輸入的數據及給定結果進行訓練，重複調整權重及偏差值等參數，此目的是讓神經網路自己學習去預測和識別。精準度的測試需比較模型的輸出和事實。然而，數據下的有毒與沒有毒性的樣本筆數比例處於非常不平衡的現象，數據分析下常關注的類型是少數類別樣本(minority)，對於資料導向分析的機器學習模型訓練時應用於此領域時，含有少量少數類型事例的資料集訓練後的模型進行預測時，通常對多數類的預測精度很高，而少數類的預測精確性卻很差。因此我們再切分訓練資料集時，著重於資料不平衡的特徵來設計模型，利用分層抽樣 (stratified sampling) 將資料劃分為不同的層(有毒性與沒有毒性)，然後從不同的層中隨機地抽取樣本，從而保證有毒的樣本在每一次抽樣中都具有資料，不會造成不平衡資料所造成抽樣的偏頗。

本研究以本研究使用了三種非線性及一種線性模型、卷積神經網路 (GCN)、支持向量機(SVM)、隨機森林(Random Forest)及邏輯斯回歸(Logistic Regression)，四種模型做中草藥臟器功能與毒性之預測，並針對各個訓練及做準確率比較。我們可以透過混淆矩陣上為本研究判斷模型好壞的工具。定義以下的評估方式：
positive(TP)：實際有關係，預測結果為 1(有關係)，則為判斷正確。
true negative(TN)：實際沒關係，預測結果為 0(沒關係)，則為判斷正確。
false positive(FP)：實際沒關係，預測結果為 1(有關係)，則為判斷錯誤。
false negative(FN)：實際有關係，預測結果為 0(沒關係)，則為判斷錯誤。
ROC 曲線是利用真陽性率(true positive rate; TPR)，又稱為敏感度(sensitivity)，其公式如 $TP/TP+FN$ ；以及偽陽性率(false-positiive rate; FPR)，以 1-特異度(specificity)表示，特異度的公式如 $TN/FP+TN$ 。我們利用曲線下的面積(Area Under Curve; AUC)來判別 ROC 曲線的鑑別力，AUC 數值的範圍從 0 到 1，數值愈大愈好。我們將中草藥的經絡以及毒性的 AUC 以及整體評估的效果呈現於表一與二。

表一 測試組經絡個別預測評估(AUC)

方法	膀胱經	心經	膽經	心包經	腎經	大腸經	肝經	肺經	小腸經	脾經	胃經	三焦經	毒性
隨機森林模型	0.68	0.74	0.65	0.60	0.65	0.64	0.65	0.64	0.75	0.63	0.68	0.86	0.98
邏輯式回歸	0.63	0.83	0.68	0.62	0.64	0.68	0.65	0.62	0.74	0.64	0.61	0.76	0.95
圖卷積神經網路	0.66	0.89	0.74	0.63	0.68	0.67	0.66	0.67	0.68	0.66	0.64	0.87	0.98

表二 測試組整體預測評估(AUC)

方法	訓練資料	測試資料
隨機森林模型	0.92	0.83
邏輯式回歸	0.87	0.81
圖卷積神經網路	0.95	0.85

由表二我們發現對於心經與三焦經的分類比較起其他經絡分類效果較好，圖卷積神經網路的預測類別效果基本上都優於其他兩種方式，唯一小腸經的部分，隨機森林與邏輯式回歸模型的分類的表現比較好。總體而言，表二表示圖卷積網路確實比其他的模型準確約 2-4%，可能是由於其特別針對圖形的結構來進行分類，使得效果有所提升。

四、結論

我們利用中草藥數據建構了一個草藥藥性功能與毒性的預測模型，以此了解中草藥對十二臟器與毒性之關聯性。在比較各種模型的時候，圖卷積網路準確度最高，這也證實了化學結構是有規律的，每個節點彼此間都有關連，比較像圖而非隨機的規律，透過圖卷積網路能夠用找尋分子的特徵，透過圖結構最大池化與累積化，將所有特徵集合，這個過程能夠確保所有特徵是一並用來當作判斷中草藥藥物標的的依據，我們的方法有助於評估新的中草藥或是其藥效功能與毒性還很模糊的草藥，供後續研發探討，已達成舊藥新用的概念。我們貢獻除了提供一個比以往更科學的方法來判定中草藥的正面與負面的影響外，這不僅有助於中醫的發展，更可對整體的醫學界解釋傳統中藥的效果，進而供後續研發。

五、應用與未來展望

本研究透過化合物的分子特徵，找出類似的分子結構，若能將草藥成分解析萃取，有助於研發草藥或化學家朝向正確的方向以解析分子結構對疾病的臨床實驗。另外，相對於中藥，目前的西藥多以合成或從天然產物中提煉而成，若能將本研究的結果應用到西藥上，找出特定的化學結構對於身體的影響，以及其交互作用，進而能夠針對病症合成出適用之藥物，將更能達到對症下藥的目的。雖然此模型確實可以根據輸入的單一草藥的化學性質去推測其對應的藥性功能與毒性，但實際在食用中藥時絕對不止食用一種，而且可能也會以煎或熬的方式調配藥湯，這些都會改變草藥的化學性質。因此我們未來能夠進一步的測試不同中藥之間的交互作用，或是不同的劑量，可能會增強或削弱彼此的作用，以及不同的食用方式是否也會對藥性產生不同的影響。另外，由於一些不同的中藥之化學結構也會對人體有相同的影響，而人又有不同的體質，若是能取得同一藥材對於不同體質的影響程度，進而和其他有相同效用的藥材相比較，並結合前述中對於人體體質的預測，在實際運作上就更能針對不同的狀況提供最佳的中藥材組合。

在研究途中，我們發現我們需要的工具或程式碼都有多方提供，例如大學、科技公司，而且很多是開源的，免費提供給任何有興趣的人使用，而現在的趨勢也是大家互相用別人寫的程式或製作的套件再進行調整，很少有完全自己寫的案

例，這樣更能促精進交流及新突破，因此做機器學習的研究應著重於參數的決定，選擇效果最好的演算法等調整，並多使用較知名的工具包含開源程式碼，以建構更創新、更準確的預測模型。

六、參考文獻

- 杜健鵬, & 史大卓. (2010). 中醫藥學治療慢性病的優勢和對現代醫學診療模式的啟示. 北京中醫藥, 29(4), 271-273.
- 郝征, 吳學會, 祁建春, & 馬曉峰. (2017). 大健康背景下的中醫養生科普軟件開發. 中醫藥管理雜誌, 23, 066.
- TCM-MESH 網站平台, 查詢介面, <http://mesh.tcm.microbioinformatics.org/>
- TCMID 網站平台, <http://www.megabionet.org/tcmid/>
- PubChem 開放式平台, <https://pubchem.ncbi.nlm.nih.gov/>
- Tox21 挑戰, <https://tripod.nih.gov/tox21/challenge/>
- 卷積神經網路, <http://cs231n.github.io/convolutional-networks/>
- Altae-Tran, H., Ramsundar, B., Pappu, A. S., & Pande, V. (2017). Low data drug discovery with one-shot learning. ACS central science, 3(4), 283-293.
- Choi, W., Choi, C.-H., Kim, Y. R., Kim, S.-J., Na, C.-S., & Lee, H. (2016). HerDing: herb recommendation system to treat diseases using genes and chemicals. Database, 2016.
- Cong, W.-h., & Chen, K.-j. (2007). FDA acknowledges TCM belonging to whole medical systems. Chinese journal of integrative medicine, 13(2), 83-84.
- Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. Paper presented at the Advances in neural information processing systems.
- Kearnes, S., McCloskey, K., Berndl, M., Pande, V., & Riley, P. (2016). Molecular graph convolutions: moving beyond fingerprints. Journal of computer-aided molecular design, 30(8), 595-608.
- Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. (2016). DeepTox: toxicity prediction using deep learning. Frontiers in Environmental Science, 3, 80.
- Shao, L., & ZHANG, B. (2013). Traditional Chinese medicine network pharmacology: theory, methodology and application. Chinese journal of natural medicines, 11(2), 110-120.
- Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., . . . Pande, V. (2018). MoleculeNet: a benchmark for molecular machine learning. Chemical science, 9(2), 513-530.
- Xue, R., Fang, Z., Zhang, M., Yi, Z., Wen, C., & Shi, T. (2012). TCMID: traditional Chinese medicine integrative database for herb molecular mechanism analysis.

Nucleic acids research, 41(D1), D1089-D1095.

Zhang, R.-z., Yu, S.-j., Bai, H., & Ning, K. (2017). TCM-Mesh: The database and analytical system for network pharmacology analysis for TCM preparations.

Scientific reports, 7(1), 2821.

【評語】 190013

本研究利用深度學習模型，讓中醫的功能與毒性科學化，以中草藥的化學性質作為輸入，透過化學結構的圖像傳播，來預測中草藥相對應之臟器有效功能與毒性預測。本研究利用分層抽樣保證有毒的樣本在每次抽樣中都具有資料，並找到化學結構與圖像的共同與不同之處，建立起適用於化學結構的模型，提供了一個比以往更科學的方法來判定中草藥的正面與負面影響。建議加強說明參數如何設定，如不同模型的初始參數如何決定及其值為何、訓練的步數/次數、訓練終止的設定值等。