2025年臺灣國際科學展覽會
優勝作品專輯

作品編號　190025

參展科別　電腦科學與資訊工程

作品名稱　分子結構語言與熔沸點性質的人工智慧預測

得獎獎項　四等獎

就讀學校　國立臺灣師範大學附屬高級中學

指導教師　蔡明剛

　　李啟龍

作者姓名　吳泰澄

關鍵詞　機器學習、圖像神經網路、分子性質預測

作者簡介

我是師大附中數理資優班1612班的學生吳泰澄，目前主要專注於人工智慧有關的研究，閒餘時間也會寫網頁。我之所以能完成這篇研究，要特別感謝我的指導教授、學長、同學以及家長的協助與支持，最後，我認為比起當一個每天按照教科書開處方的醫生，我願意做更有創造性的事情，如:研究。

附件三

研究報告封面

2025 年臺灣國際科學展覽會

研究報告

區別：

科別：電腦科學與資訊工程

作品名稱：分子結構語言與熔沸點性質的人工智慧預測

關鍵詞：機器學習、圖像神經網路、分子性質預測（最多三個）

編號：
（編號由國立臺灣科學教育館統一填列）

目錄

中文摘要

背景：預測分子性質如溶解度、毒性及熔沸點對於基礎科學至關重要。然而，實驗測量這些性質耗時且昂貴，因此本研究使用多種機器學習模型藉由調整變相來準確預測熔、沸點。

方法：本研究使用超過一萬筆數據及兩種類型的機器學習方法：淺度與深度學習。淺度學習由 PyCaret 實現，並以 Mordred 作為分子描述器；深度學習使用圖神經網路，包括（CMPNN 和 GCN），並調整隱藏層參數。

結果：CMPNN 在目前嘗試的模型中表現最佳。發現影響沸點預測的關鍵特徵是 piPC1，與鍵級相關；熔點則是 AATS0d，與 σ 電子的 Moreau-Broto 自相關有關。

結論：CMPNN 模型在沸點與熔點預測中均表現最佳。沸點中深度學習模型優於淺度學習模型（$p<0.05$）。此外，使用 SHAP 成功找出 piPC1 和 AATS0d 對最關鍵。本研究不僅得出了高準確性的模型，還發現了影響分子性質的關鍵特徵，且可擴展至其他預測。

英文摘要

Background: Predicting molecular properties such as solubility, toxicity, melting, and boiling points is crucial for fundamental science research. However, experimental measurements are often time-consuming and cost-intensive, so we use machine learning (ML) as an approach to improve prediction accuracy.

Methods: A dataset containing over 10k compounds was used for training shallow and deep ML models. Shallow machine learning models were implemented via PyCaret and Mordred as feature extraction. For deep machine learning models, graph neural networks (GNNs), specifically CMPNN(Communicative Message Passing Neural Network) and GCN(Graph Convolutional Network), were trained, and tuned by adjusting the number of hidden layers and sizes (neurons) in each layer.

Results: The CMPNN model outperforms the GCN and shallow ML model for boiling point prediction(best: $R^2 = 0.76$, MAE = 23.89K for b.p.; best: $R^2 = 0.87$; MAE = 23.73K for m.p.). The top molecular descriptor of the b.p. prediction is piPC1, which is related to bond order, and that of m.p. is AATS0d, which is related to σ electron Moreau-Broto autocorrelation.

Conclusions: The prediction of molecular properties was improved by a comprehensive research of shallow and deep learning approaches, showcasing CMPNN model can reach the highest performance in the prediction of m.p. and b.p.($R^2 = 0.87$ in m.p.; $R^2 = 0.76$ in b.p.). In this study, we found that the deep learning model works better than shallow ML in predicting m.p.($p<0.05$). This study uses SHAP analysis to successfully identify piPC1 and AATS0d as the key prediction factors of b.p. and m.p. respectively. Moreover, this approach can be applied to predict other molecular properties. To conclude, this study not only shows a highly accurate model but also identifies the key factors of m.p. and b.p.

壹、研究動機

在某次製作實驗預報時要查出各個化合物的熔點、沸點及其他性質，人工一一查詢十分費時而且很多時候因為化合物比較不常用到而實驗室或藥商未必會去檢測，又或是其性質不穩定，不會被直接販售而使需要研究者在實驗室自行合成，此時就不會有所謂的性質表可以參考，而某些網站就會提供由現有模型依靠演算法計算出來的數據，但此類數據又極為不準確，有些熔點差距甚至可以到 100 度卡爾文的差距，因此我就想到了在資訊課所學的大數據分析，將化合物的立體結構以大數據分析試圖找出其熔點、沸點、閃點、氧化性、及危險性等物理及化學性質。

貳、研究目的及研究問題

一、研究目的

本研究旨在：

1. 探討分子性質與立體結構關聯性，分析其機制。
2. 設計淺度與深度模型，並進行效果差異的比較。
3. 利用變數間的交互作用，尋找表現最佳的模型。
4. 分析結構因素對化學性質的影響，並進行比較。

二、研究問題

化合物有成千上萬種，其基本性質及危險性對於研究者而言至關重要，尤其在於預防實驗意外的發生及加速藥物開發，因此本實驗將提出兩種不同方法將給定的化學結構或其 IUPAC 名轉換為其基本性質，未來運用在不同性質上可以更準確且有依據。

本研究將實驗分為兩個部份：淺度機器學習、深度機器學習，此兩者的目的均在於將化合物的結構轉換成性質。

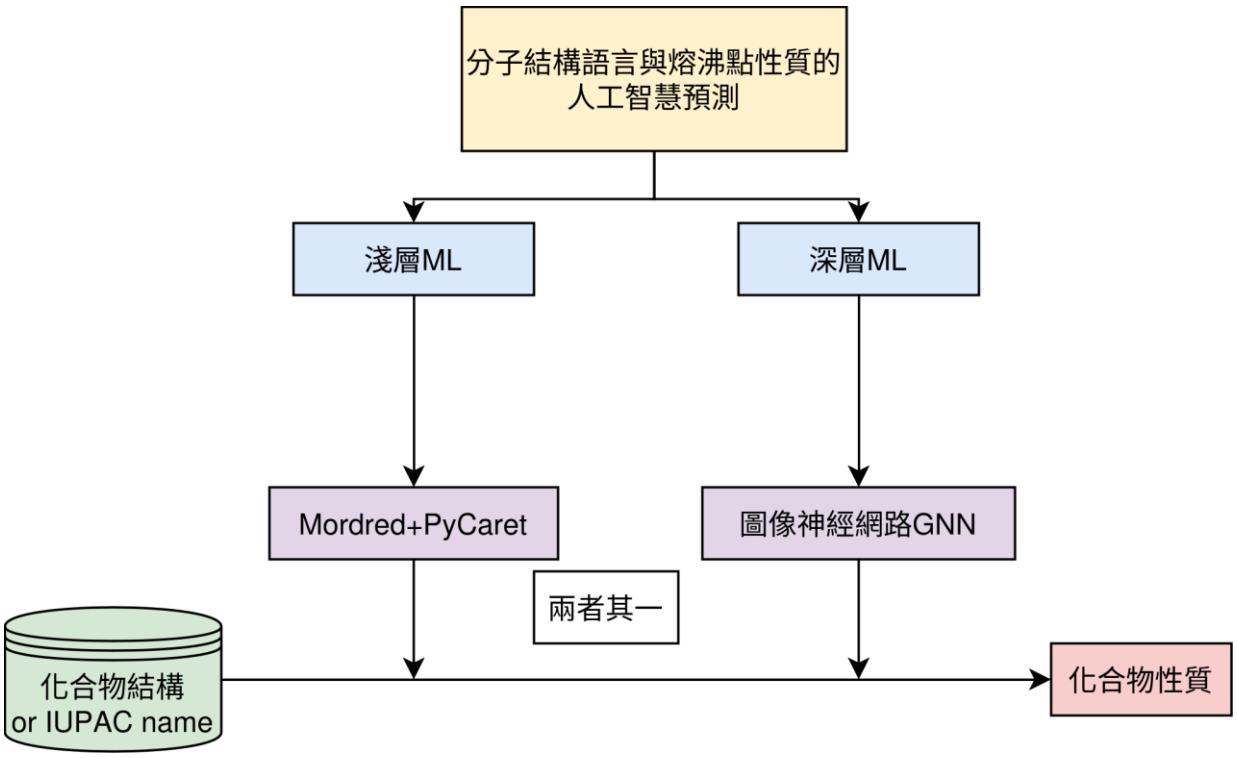


圖 1: 研究架構圖（作者自行繪製）

（一）、淺度機器學習

淺度機器學習是指較簡單的學習方式，並沒有多層的堆疊，故通常所需的資源及時間也較少，曾有諸多研究利用此一性質預測熔、沸點，其中以人工挑選出極化作用力、分散力、分子構型對稱性作為主要特徵，並以 XGBoost 訓練熔點，R^2 分數可以得到 0.83(葉宗融 &蔡明剛, 2020)。此方法通常會先透過分子描述器將分子特徵萃取（feature extract）出來，再以萃取出的特徵進行挑選、特徵工程等，最後套入模型並擬合。

（二）、深度機器學習

深度機器學習（Deep Learning，DL），是一種比淺度機器學習更複雜、耗費更多資源的學習方式，基本上涵蓋多層模型，常見的類型包括但不限於卷積神經網絡（Convolutional Neural Network，CNN）、循環神經網路（Recurrent neural network，RNN）、及本研究所使用的圖像神經網路（Graph neural network，GNN）等，有研究認為使用 GNN 可以更貼近化合物的立體結構 (Reiser et al., 2022)，因此本研究將以 GNN 進行嘗試，而所採用的神經網路類型之一便是圖卷積神經網路（Graph Convolutional Networks，GCN），其概念和一般的 CNN 類似，只是數據改成非歐基理得數據，因此 GCN 卷積的對象會變成是鄰近的點，此網路在過去的研究中被認為是表現最佳的 GNN 之一 (Zhang et al., 2019)，可以擷取成功的結構特徵 (Kipf & Welling, 2017)。

而 GNN 類網路係透過將點的特徵萃取出來，部份模型亦支援萃取邊的特徵，常見的特

徵整理請參照肆之三之七 ，並將這些點（不論有無實際鍵結，亦可設定有實際鍵結）進行聚集，如：卷積、交互等運算。

（三）、淺度與深度機器學習之比較

在Qu et al., 2022中提到，在使用NIST Thermodynamics Research Center (TRC) SOURCE Data Archival System 資料集的情況下，使用傳統的官能基貢獻模型（即Stein and Brown, 1994的方法）對分子沸點的模型預測效果只能達到 MAE=11.84K，但 MEGNet 模型可以達到 MAE=5.77K，且該研究相信其模型表現已很接近極限，另外該研究亦發現該模型可以找出原資料集中錯誤的分子沸點數據。

參、研究設備及器材

（一）、硬體

本研究採用自有之伺服器，規格如下：

- 處理器：Intel Xeon Gold 6414U (64 cores)
- 隨機存取記憶體：512GB

惟研究過程中從未用盡所有運算資源，本實驗亦不須此類高規格之伺服器重現。
另外亦視需求在 Google Colab 及自備筆電上進行較簡易的運算。

（二）、軟體

本研究採用 Python 作為主要程式語言，同時搭配現有的自動機器學習模型（AutoML）達成淺度機器學習。
本研究中使用的淺度機器學習軟體如下：

- PyCaret 3.3.2

深度學習模型如下：

- OpenChem 的主要 github 版本（github master branch@f42707）
- CMPNN 的主要 github 版本（github master branch@b647df2）
- DeepChem 2.8.0

同時採用 RDkit 分析化學結構，Matplotlib 繪圖，Mordred 作為分子描述器。

肆、研究過程或方法

一、背景

（一）、評分方式

本研究主要採用 R^2 分數及 MAE 作為評估標準，並以 RMSE 作為損失函數，以下的所有參數以 x_{real} 代表真實數值、x_{pred} 代表模型計算的數值。

R^2 分數，決定係數（Coefficient of determination）定義如下：

$$R^2 = 1 - \frac{\sum_{k=0}^{n}(x_{k,real} - \bar{x}_{k,real})^2}{\sum_{k=0}^{n}(x_{k,pred} - x_{k,real})^2}$$

RMSE，Root Mean Square Error，方均根定義如下：

$$RMSE = \frac{\sqrt{\sum_{k=0}^{n}(x_{k,real} - x_{k,pred})^2}}{n}$$

MSE，Mean Square Error，均方誤差定義如下：

$$MSE = \frac{\sum_{k=0}^{n}(x_{k,real} - x_{k,pred})^2}{n}$$

MAE，Mean Absolute Error，均絕對值誤差定義如下：

$$MAE = \frac{\sum_{k=0}^{n}|x_{k,real} - x_{k,pred}|}{n}$$

本研究之所以採計 RMSE 而非 MSE 或 MAE 係因 RMSE 可以減少 MSE 及 MAE 之間評估的損失，故採取 RMSE 作為損失函數。

（二）、預測項目

本研究主要針藥典沸點及熔點，如未另外註明，溫度的單位均為絕對溫標。

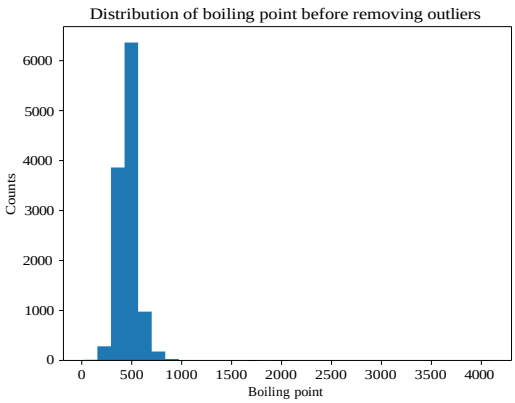
沸點：係指物質之飽和蒸氣壓與外界（10^5 帕）達到平衡時之溫度，亦可理解為物質劇烈的從液相轉為氣相時的溫度，大部份物質具有此性質，但少數物質可能達一定溫度後會分解成其他物質，故本研究不納入此類化合物。目前學界認為和沸點相關的主要是分子間的作用力，包含氫鍵、偶極力等，而分子間的作用力又會和接觸面積、電荷分佈、凡得瓦體積等有關，因此在預測上具有一定難度。

熔點：係指固相之飽和蒸氣壓與外界（10^5 帕）達到平衡時之溫度，也可被思考成物質（晶體）從固相轉為液相時的溫度。根據Hughes et al., 2008的研究，沸點在分子性質的預測

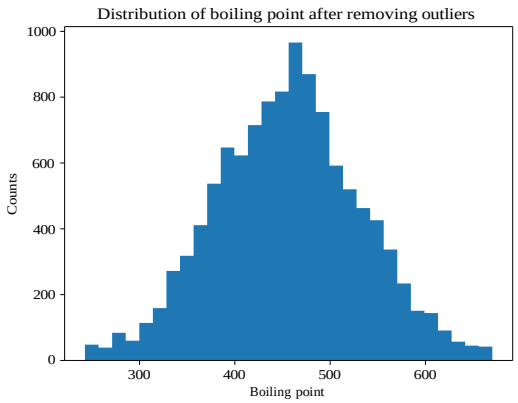
上不如其他性質來的容易，該研究使用了 MOE 2D/3D 的描述符（目前已被包含在 Mordred 中）並針對 LogS、LogP、熔點等性質使用不同模型進行預測，發現熔點為三者中最難預測的，其次依序為 LogS、LogP，同時也表明，缺乏固相晶體資訊為其中的描述符應為熔點較難預測的原因。除了晶體資訊外，熔點還和分子的對稱性、凡得瓦體積、分子間作用力等有關。

（三）、資料來源

本研究的沸點資料集採用ChEDL 資料集的沸點資料(Bell et al., 2016-2024) 中 Yaws's boiling point 數據資料集，其紀錄了各分子的 SMILES 及沸點，Yaws 沸點資料集原始資料共有 11719 筆，沸點分佈如圖2a，第一四分位數（Q_1）為 403K，平均為 451K，第三四分位數（Q_3）為 510K 本研究移除大於 Q_3 加 1.5 個 IQR（四分位距）及小於 Q_1-1.5IQR 之離群值後，計有 11324 筆沸點資料，分佈如圖2b。



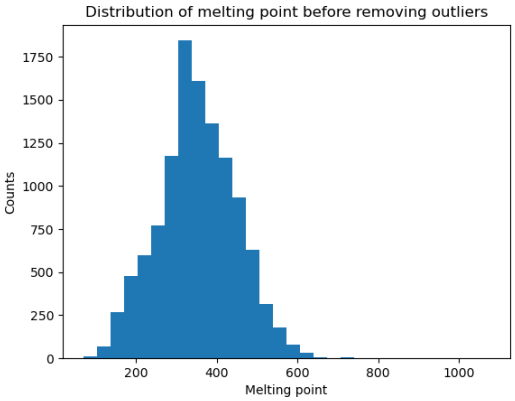
(a) 沸點去除離群值前的分佈（作者自行繪製）



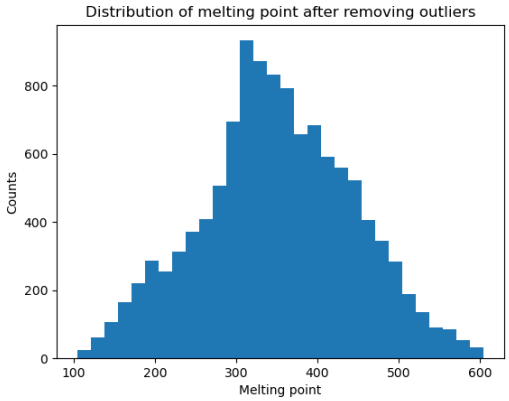
(b) 沸點去除離群值後的分佈（作者自行繪製）

圖 2: 沸點去除離群值前後的分佈比對圖（作者自行繪製）

本研究的熔點資料集採用同樣來自 ChEDL 資料集的沸點資料中Andrew Lang, 2011所編之 Open Notebook 資料集，共計 11549 筆資料，其中 9 筆資料 CAS 編號有誤，$Q_1 = 292K$，平均 $= 349K$，$Q_3 = 418K$，比照沸點的方式移除離群值後共計 11477 筆資料，移除前後分佈如圖3。



(a) 熔點去除離群值前的分佈（作者自行繪製）



(b) 熔點去除離群值後的分佈（作者自行繪製）

圖 3: 熔點去除離群值前後的分佈比對圖（作者自行繪製）

（四）、資料結構

本研究中原始資料係採用 SMILES（簡化分子線性輸入規範，Simplified molecular input line entry specification）儲存，其可以代表一個分子的唯一結構。舉例來說，阿斯匹靈，又稱 2-乙醯氧基苯甲酸，結構如下圖所示：

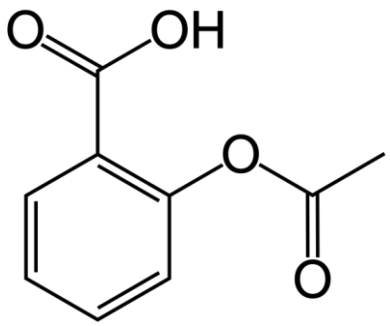


圖 4: 阿斯匹靈結構式（Fvasconcellos 繪製，公有領域）

該結構可以用 SMILES 代表成：「O=C(C)Oc1ccccc1C(=O)O」。

二、淺度機器學習

（一）、特徵

由於原始數據中僅用一串文字（SMILES）去描述一個分子，模型無法解讀此串文字，因此需要特徵萃取，將此文字轉換成一系列數值，又稱為描述符或特徵，在淺度機器學習中，常用的便是分子描述器，其中又以 Mordred 和 PaDEL 提供的特徵數最多且最常被使用。因此本研究採用由Moriwaki et al., 2018提出的 Mordred 分子描述器，該分子描述器共可產生 1614 個特徵，其中包含 763 個數字特徵，本研究捨棄所有包含物件 (即非數值) 的特徵，僅

保留每分子 763 個特徵。

本研究先移除與分子性質呈現低相關的特徵（共變異數小於 0.1）、特徵數據集中彼此之間呈現高度相關性的特徵（共變異數大於 0.7），並且以變異數分析（ANOVA，Analysis of variance）選出相關性最高的前 k 個特徵（即 SelectKBest）。因為模型複雜度會隨選取特徵數增加而增加，故表現會在特定一處最佳，呈現之曲線應為一開口向下之曲線，故設計出實驗一，參見肆之四之一。

（二）、機器學習模型

本研究中採用了自動機器學習軟體 PyCaret(Ali, 2020)，其特點為可以將所有主流回歸模型全部跑一次並自動調整其超參數，將繁重的工作流程簡化。

另外，本研究使用 10-fold 交叉驗證，即將訓練資料集化分成 10 份，並以 9 份做訓練、1 份做驗證，每一份小資料集均會輪流當驗證資料集，如此可以減少模型對資料的敏感性。參見實驗二：肆之五之二。

（三）、模型解釋

淺度機器學習的好處之一便是可以精確的得出模型特徵和預測數值之間的因果關係，其中一個方法便是透過 SHAP（SHapley Additive exPlanations），透過此方法可得到每個特徵的 Shapley values，將其取絕對值並依大到小排序即可獲得特徵影響模型之排名。

三、深度機器學習

本研究使用的圖像神經網路，會先將各節點（原子）及邊（鍵）依據本節第七段轉換成數值特徵以描述此原子或鍵，之後由各模型將這些特徵進行卷積或回歸。

常見的 GNN 模型共有以下幾種，本研究中挑選了幾個常見的模型並測試，其中包含自行架設之 GCN+MLP（Multi-layer Perceptrons，多層傳感器）網路。

（一）、**MEGNet** 模型

MEGNet 的工作原理如圖5，其主要係在每次訓練一個分子時依序更新邊、原子（節點）及全域特徵的資訊，此三層可另由數層神經網路架構堆疊而成。

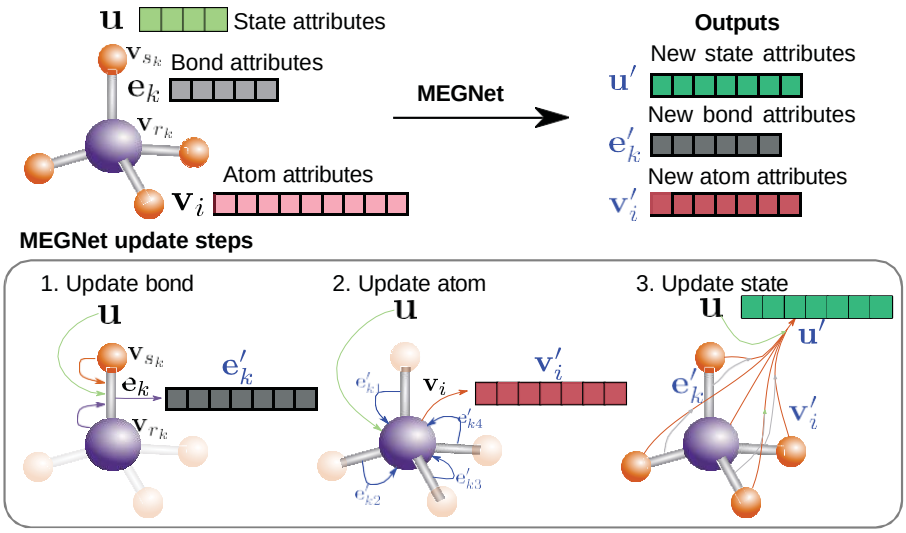


圖 5: MEGNet 的工作原理（Chen et al., 2019繪製）

後來又有基於此研究提出更新的模型 M3GNet(Chen & Ong, 2022)，但該模型注重應用於晶體系統而非分子系統。此二模型皆可由 MatGL 函式庫提供實做 (Ko et al., 2021)。

（二）、**SchNet** 模型

SchNet 在分子能量、原子間力場量的計算與預測上有極大優勢（且此二項目的研究為 SchNet 模型的設計目標），透過加入 Continuous Filter Convolution Layer，以原子在空間中的位置萃取出特徵 (Schütt et al., 2017)，同時也有研究者將 SchNet 應用於沸點的預測實驗中，雖其研究目的並非模型本身，但其計算沸點之 MAE 可小到約 14K。（使用特殊 train/test 割分方法）(Li & Rangarajan, 2022)。

（三）、**OpenChem**

此函式庫中包含了幾種常用的模型並以 Pytorch 作為後端，其中之一便是 GCN 及 MLP 搭配在一起的神經網路。本研究中使用該函式庫中的此二網路並加以堆疊、排序，其中 GCN 層的隱藏維度為 n，一共堆疊五層，MLP 的部份輸出維度則分別是 n、n/2、n/4 直到等於 128 再接上 1。(Korshunova et al., 2021)

n 的切確數值涉及模型的複雜度，詳見下章實驗設計肆之四之三。

（四）、自行架設之神經網路

本研究嘗試過許多類型的網路後，亦試著自行組建 MLP 神經網路，依次更新邊、節點並將兩者混雜在一起，每層以線性回歸實做，降低複雜性，使得總參數量在 5282。

由於此模型不涉及卷積，基本上只有各點的特徵之集合作為輸入，並以一個特定的數值作為輸出，因此此類模型在結構上可預期會較其他標準方法差，經過簡單的測試後，實驗亦

證實此理論，其表現低於基準值（$R^2 < 0$，若以數據的平均數作為預測值，則 R^2 應為零），故後續實驗中不以本模型採計。

（五）、訊息傳遞網路

訊息傳遞網路（Message Passing Neural Network，MPNN）會通過在邊之間傳遞消息並逐步更新節點狀態來學習分子性質，從而捕捉分子的結構訊息 (Gilmer et al., 2017)，另有基於此模型延伸的 DMPNN（Directed Message Passing Neural Network），此模型將分子視為有相圖，透過考慮分子圖中鍵（邊）的方向性來改進訊息傳遞過程。另外還有另一種神經網路–CMPNN（Communicative Message Passing Neural Network），進一步擴展了 MPNN 模型，其特點是允許節點和邊之間的訊息進行交互連結。在 CMPNN 中，節點和邊會彼此影響，增加了訊息傳遞過程中交換訊息的多元性。(Song et al., 2020) 此三模型對於特徵萃取的方法比較如下圖：

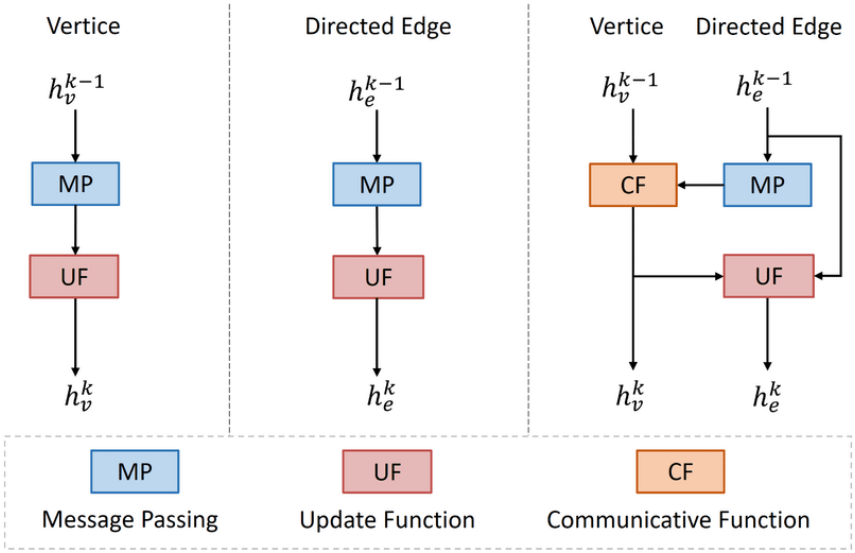


圖 6: 三種訊息傳遞網路比較圖，左側為 MPNN，中間為 DMPNN，右側為 CMPNN（由 Song et al., 2020繪製）

此類網路主要由兩部份構成，訊息傳遞神經網路層（Message Passsing Neural Network，MPN）和前饋神經網路層（Feedforward Neural Network，FNN），其中本研究使用的 MPN 中主要的有 1 層激活函數（ReLU）、五層線性回歸（其中包含節點、邊各一層，隱藏層兩層、輸出層一層）、1 層 BatchGRU 及最後再一層線性回歸。FNN 的部份則較為簡單，每層僅由正則化（Dropout）、線性回歸、激活函數（ReLU）構成。各層之間的參數數量詳見下章實驗設計。(Song et al., 2020)

（六）、量子力學方法（非 **GNN** 類）

雖然採用量子力學的模型的在 QM9 資料集（限制總原子數量小於 14） 中表現較佳（MAE：1.21 kcal/mol），而 GNN 類表現最好的模型（SchNet），表現最好到 1.23kcal/mol，但此模型每次跑都需要幾分鐘（原文：few mins），因此可能會對後面的應用有顯著影響。(Tsubaki & Mizoguchi, 2020)

（七）、萃取特徵

本章節旨在探討應該如何依據前人的研究結果選出所需要的特徵，依據Chen et al., 2019、Xiong et al., 2020、Pocha et al., 2021等研究整理出常見的特徵如下：

方式	名稱	特徵儲存方式	簡介
原子	原子種類/數	one-hot	表達原子的種類或原子數
原子	掌性	one-hot/空	R/S
原子	環的大小	整數/空	環的尺寸/若不在環中則為空
原子	電負度	one-hot	電負度以 0.5-4.0，分成十類
原子	共價半徑	one-hot	25-250pm，分成十類
原子	電子親和力	one-hot	-2.5-3.7，分成十類
原子	混成軌域	one-hot/空	sp,sp2,sp3,... 或更多，若無混成則為空
原子	電子接收	布林值	是否為電子接收者
原子	電子給予	布林值	是否為電子給予者
原子	芳香環	布林值	是否在芳香環中
原子	自由基	整數	自由基數量
原子	電荷	整數	形式電荷
鍵	鍵級	one-hot/空	1、2、3 或在芳香環中
鍵	同環	布林值	和出發原子是否在同一個環中
鍵	拓樸圖距離	整數	和出發原子最短路徑上有多少原子
鍵	對稱形式	one-hot	無、任何、Z form、E form
全域	平均原子重量	浮點數	分子量除以原子數量
全域	平均鍵數	浮點數	總鍵數除以原子數量

表 1: GNN 模型常用的特徵總表（作者自行整理）

值得一提的是，Pocha et al., 2021曾以其中數種特徵的選擇[1]做過實驗，其實驗結果顯示不選擇芳香性相關、在環中、形式電荷有關的特徵結果表現較好，而鄰居為氫原子的數量與鄰居為非氫原子的數量則對結果有正面影響。

[1]其所選之特徵均為原子特徵

	1	2	3	4	5	6	7	8	9	10	11	12
原子種類	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
非氫原子鄰居	✓		✓						✓	✓	✓	✓
氫原子鄰居	✓			✓				✓		✓	✓	✓
形式電荷	✓				✓			✓	✓		✓	✓
在環中	✓					✓		✓	✓	✓		✓
芳香性	✓						✓	✓	✓	✓	✓	

表 2: Pocha et al., 2021研究所選之特徵（Pocha et al., 2021繪製，作者自行翻譯）

唯此工作研究之實驗資料集係和量子力學、溶解度、代謝穩定性有關，故對於本研究欲預測之項目而言不一定適用。

四、實驗設計

（一）、特徵數量對淺度機器學習表現之影響

本實驗中將透過梯度下降的概念，先從 20 個特徵向上提昇特徵數量直到模型表現開始下降為止，之後縮小範圍，找出表現最好時特徵數量有多少個。

（二）、不同模型及超參數對淺度機器學習表現之影響

本實驗中我們將數據以 8:2 的比例區分成訓練數據組及測試數據組。訓練數據組中以 10-fold 交叉驗證法進行驗證並以 PyCaret 自動機器學習進行回歸預測實作並紀錄不同模型最佳化之後的學習表現。

（三）、不同模型的超參數對深度機器學習表現之影響

本研究中以 10-fold 交叉驗證法分別測試了前章（肆之三一至六小節）所述之數種模型，並挑選兩類較具代表性的 GCN+MLP、MPNN，分別由 OpenChem 和 CMPNN 實做，另外為了和現有研究比較，同時使用了 DeepChem（另一 GCN+MLP 模型，引自Feng et al., 2024）。

OpenChem 的部份主要由 5 層 GCN 網路接上 3 層 MLP 構成，本實驗中將透過改變 GCN 隱藏層的參數數量（Hidden Size）來找出需要適當描述此資料集需要多少參數，以達成此類模型較佳的表現。

CMPNN 的部份則由數層特徵提取層連接而成，又可稱為訊息傳遞網路，在此之後會接上前饋神經網路，詳見肆之三之五，本實驗分為兩部份：

其一為改變訊息傳遞網路的隱藏層參數數量，從 150 上升到 1500，每次上升 150，藉此觀察需要適當描述此資料集需要多少參數，以達成此類模型較佳的表現。

其二為調整前饋神經網路的層數，從 4 層減少到 1 層，並找出適當的層數使模型表現更佳。

伍、研究結果

一、特徵數量對淺度機器學習表現之影響

本研究以 Xgboost 模型對沸點數據從 k=20 個特徵向上增加所使用的特徵數量訓練，會發現模型對測試數據集的預測表現隨所使用的特徵數量增加而改變，而在約 k=50 的地方模型對測試數據集的預測表現能力開始下降，R^2 分數會從 0.63 降至 0.50，之後本研究採用左右逼近法找出最佳特徵，如圖7，發現 k=35 附近有最好的表現。

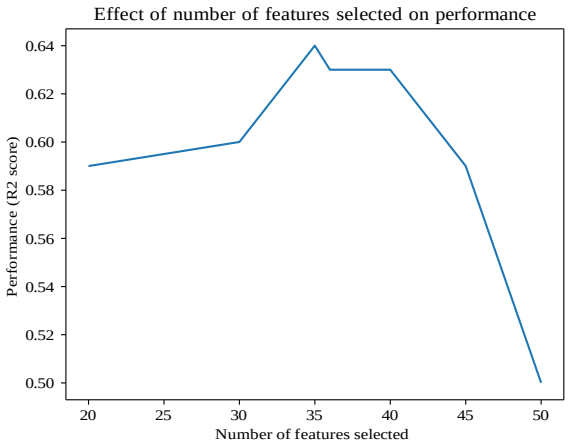


圖 7: 所選的特徵數量對表現作圖分析（作者自行繪製）

本研究發現在選取過多特徵的時候，模型表現會受到相關性較低的特徵干擾，而表現有明顯下降，且訓練時間也更著被拉長。

二、不同模型及超參數對淺度機器學習表現之影響

各模型對沸點的表現其實十分接近，並以 Extra Tree Regressor 為首，Random Forest Regressor 為次。不同模型之間的表現如表3：

模型 (b.p.)	R^2 分數	RMSE	MAE
Extra Trees Regressor	0.6821	42.84	25.67
Random Forest Regressor	0.6636	44.07	27.10
LightGBM (Light Gradient Boosting Machine)	0.6583	44.42	28.09
XGBoost (Extreme Gradient Boosting)	0.6357	45.85	28.04
GBR (Gradient Boosting Regressor)	0.6044	47.79	32.42

表 3: 沸點表現最好的前五個模型之測試分數（作者自行繪製）

從上表可以看出前四名的差異並不大，但到第五名的梯度提昇迴歸器（GBR）時則有明顯差異。之後對 Extra Tree Regressor 進行殘差分析結果如下：

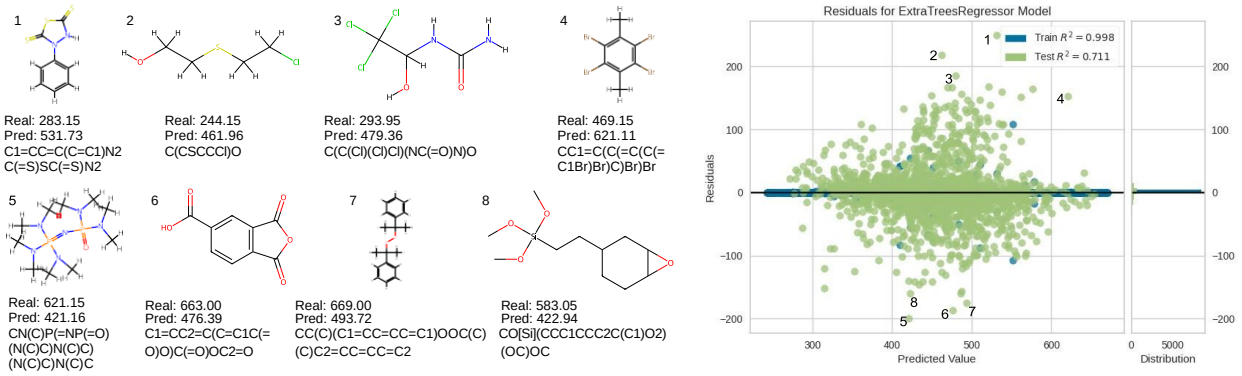
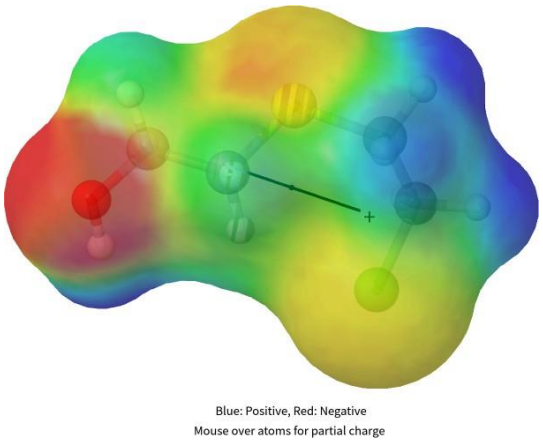
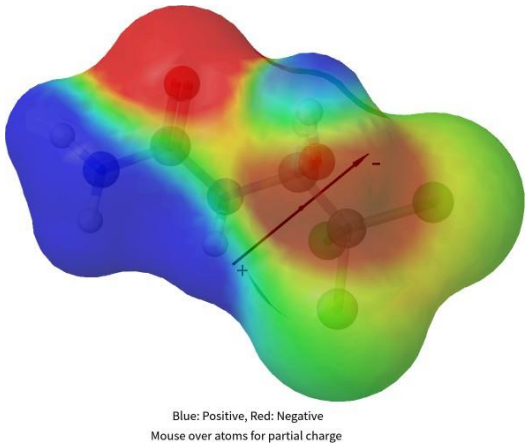


圖 8: 沸點 Extre Tree Regressor 殘差分析（作者自行繪製）

我們挑選殘差最大的前三個分子（1-3）來做觀察，由於分子 1 有超過 10 個重原子，故我們挑選分子 2、3 以molcalc.org（以 GAMESS，General Atomic and Molecular Electronic Structure System 作為後端）進行量子力學的模擬運算，算出他們的極性與表面電荷分佈，他們的極性分別為 2.58 及 3.43 德拜（Debye），但同時此二分子的表面積也不算小，均在 300 多平方埃左右，和正己烷差不多。



(a) 分子 2 的極性及表面積電荷圖（molcalc 繪製）



(b) 分子 3 的極性及表面積電荷圖（molcalc 繪製）

圖 9: 分子 2、3 的極性及表面積電荷圖（molcalc 繪製）

另外可以從此圖看出此模型變異數一致（Homoscedasticity），殘差大致成常態分佈，唯右上角存在一個較大的殘差（分子 4）。

對於沸點而言情況亦大致相同，但值得注意的是前四名的差距比沸點更小，且整體而言 R^2、RMSE 分數表現均較沸點高。

模型 (m.p.)	R^2 分數	RMSE	MAE
Extra Trees Regressor	0.8039	41.58	30.16
LightGBM (Light Gradient Boosting Machine)	0.7996	42.04	31.25
Random Forest Regressor	0.7939	42.63	31.33
XGBoost (Extreme Gradient Boosting)	0.7908	42.95	31.67
GBR (Gradient Boosting Regressor)	0.7586	46.14	34.97

表 4: 熔點表現最好的前五個模型之測試分數（作者自行繪製）

接著本研究一樣對其進行殘差分析，結果如下圖：

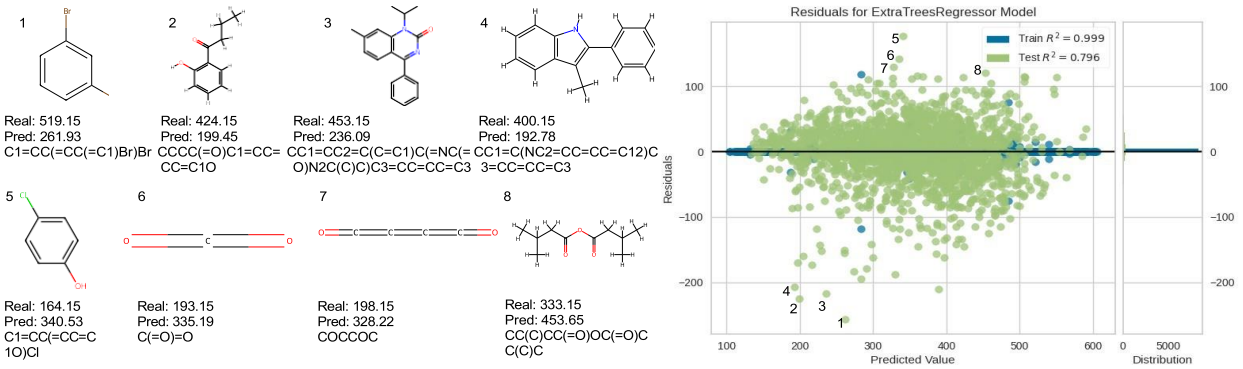


圖 10: 熔點 Extre Tree Regressor 殘差分析（作者自行繪製）

從這張圖可以看出分子 6 到 8 具有很高的對稱性，而 6、7 更是直線形分子，尤其是分子 6，二氧化碳，其沸點被遠遠的高估約 100 多度，殘差僅次於分子 5，而根據高中教科書及學界目前的說法，對稱性越高，沸點越高，惟此模型顯然高估了這個理論。至於最被低估的分子 1-4 則均存在至少一個極性原子（N、O、鹵素等），均非純烴類化合物，且 3、4 中更有雜環。

同樣的可以看出此模型變異數一致，唯左下角的散布了較多的離群值。

三、不同模型的超參數對深度機器學習表現之影響

本章節均先以熔點為例，並將比較後最好的模型套用到沸點上。

（一）、**GCN+MLP** 模型參數數量對學習表現之影響

結果： OpenChem 卷積層參數數量對表現的影響，如下表：

Hidden Size	128	256	512
R^2	0.7009	0.7073	0.6713
MAE	20.30	20.09	22.07

表 5: OpenChem 之卷積層隱藏層參數數量對表現之影響（作者自行繪製）

發現：

圖 11: OpenChem 之卷積層隱藏層參數數量對表現之影響（作者自行繪製）

1. 從表中可看出過多參數數量會略微的影響表現（R^2 降低約 0.03）

2. 每層 128 個參數足以描述分子結構，提昇至 256 僅有些微增加

思考：綜觀表現的變化可發現在 512 時有顯著下降，應該是因為模型過於複雜而低度擬合（或欠擬合，Underfitting），過度提昇參數數量不會再有任何顯著增加。

（二）、**MPNN** 類模型卷積層隱藏層參數數量對學習表現之影響

結果： CMPNN 卷積層隱藏層參數數量對表現的影響如下表：

	150	300	450	600	750
參數數量	618001	2079151	4440301	7701451	11862601
R^2	0.7553	0.7906	0.8176	0.8281	0.8361
MAE	35.07	32.1	29.65	28.63	27.63
	900	1050	1200	1350	1500
參數數量	16923751	22884901	29746051	37507201	46168351
R^2	0.838	0.8464	0.8465	0.8464	0.8456
MAE	27.52	25.98	26.12	25.82	25.90

表 6: CMPNN 之卷積層隱藏層參數數量對表現之影響（作者自行繪製）

發現：

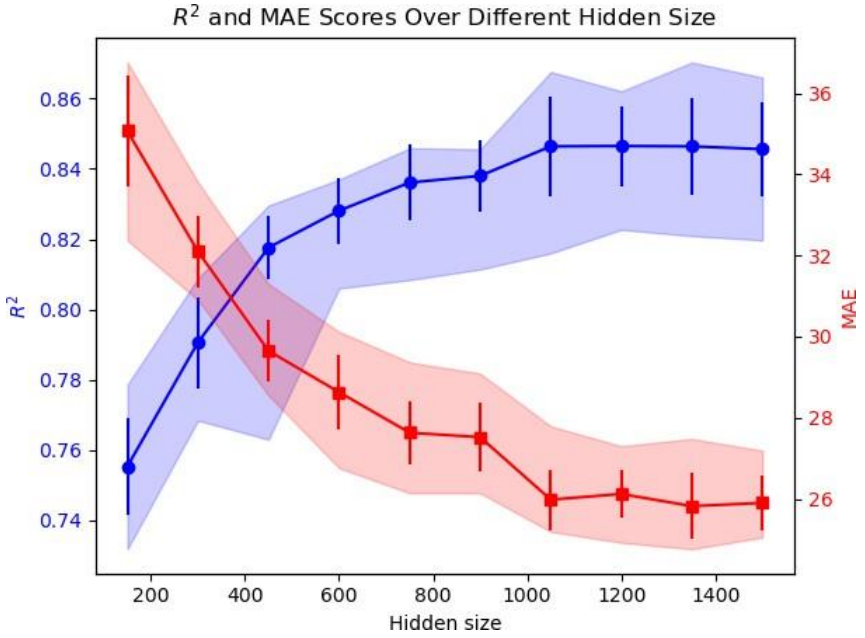


圖 12: CMPNN 之卷積層隱藏層參數數量對表現之影響（作者自行繪製）

1. 每層參數數量在 1000 附近（1050），整體模型參數 22,884,901，已經可以充分描述此資料集的性質

2. 每層參數數量超過 1050 並不會顯著增加表現，反而會拉長訓練時間

3. 本實驗最多可以改善約 0.1 個 R^2 及將 MAE 降低 10K

思考：表現應該會在某部份達到最低值，且作圖結果也支持此假說，因此模型實際上不需要過多的變數即可在較短的時間內擬合此資料集。而在特徵萃取的時候，尤其是連接節點和邊的時候，適當的參數數量會有顯著的影響，可被思考為需要適當的描述資訊方可捕捉此分子的特性。

（三）、**MPNN** 類模型卷積層隱藏層參數數量對學習表現之影響

結果： CMPNN 之 FNN 隱藏層參數數量對表現的影響如下表：

層數	4	3	2	1
R^2	0.8380	0.8453	0.846	0.8521
MAE	27.52	26.18	26.27	24.94

表 7: CMPNN 之 FNN 層數對表現之影響 (作者自行繪製)

發現：

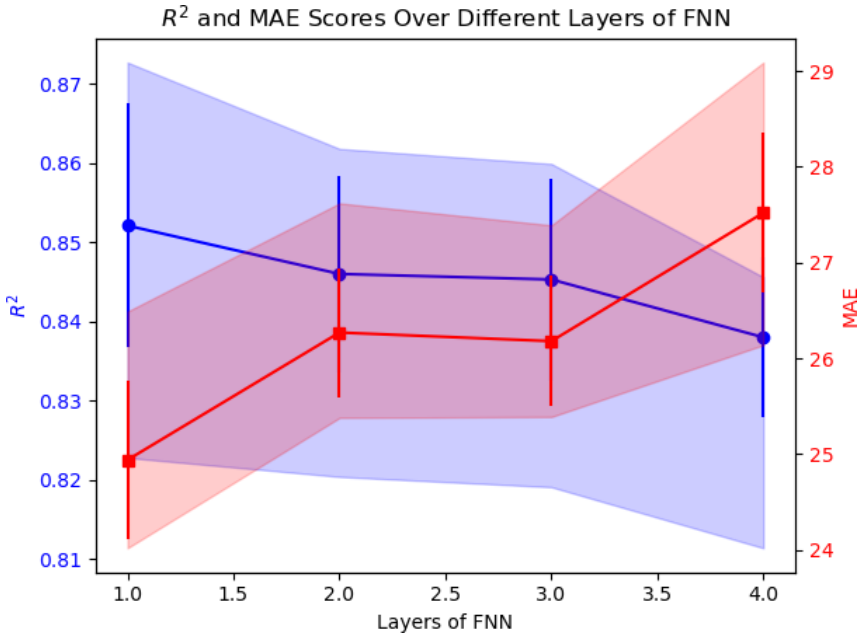


圖 13: CMPNN 之 FNN 隱藏層參數數量對表現之影響（作者自行繪製）

1. 整體差異並不大，R^2 在 0.87 及 0.81 間浮動，MAE 也均落在 24 到 29 之間差異並不大

2. 由僅堆疊一層表現稍微勝出，R^2 平均稍微高於 0.85，MAE 則在 25 附近

思考：影響模型表現的關鍵不在此層可能是因為此模型並沒有涉及特徵之間的相互運算，而僅是把上一層運算完的結果進行轉換，所以模型的表現其實已經在訊息傳遞層就決定了，和此層較無關係，而層數越少表現有些微增加的原因應為中間少掉數層正則化，因此損失較少，且在 4 層以下的時候都沒有發生過擬合的現象，因此正則化在此模型及數據集中並沒有其存在的必要性。

陸、討論

一、淺度機器學習模型表現貢獻分析

根據沸點模型分析出的主要分子描述符（參圖14）中以 piPC1 最為重要，此描述符是在 PathCount 家族中的其中一員，其計算方式為涉及鍵的數量及鍵級，大致代表鍵級對共價鍵加權的總和。FilterItLogS 則是 Filter-it™ LogS 描述器，係為一套模擬該化合物在水中溶解度之公式，而 VSA_EState3 則是屬於 MOE 型的描述符，和分子的靜電勢及表面積分佈有關，WPath 則是維納指數為分子結構圖（不含氫）之所有最短路徑和，此結果可以呼應Wiener, 1947的研究，本研究將其適用範圍從原研究的烷類到更多樣的化合物上。AATSCov 是以凡得瓦體積加權後以 moreau-broto 自相關計算出來的。AATSCov 是以凡得瓦體積加權後以 moreau-broto 自相關計算出來的。

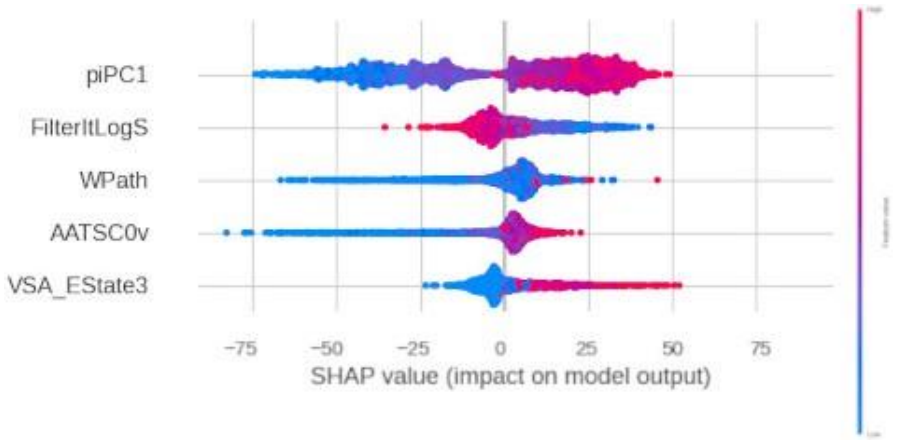


圖 14: 沸點之 Shapley values 分析（前五重要）（作者自行繪製）

另外根據熔點模型分析（參圖15），主要影響的分子描述符是 AATSod，此描述符為 σ 電子加權後以 moreau-broto 自相關計算出來的。第二高的 TopoPSA(NO) 則是僅使用氮、氧等極性原子計算出的表面積（Polar Surface Area，PSA），亦即在一個分子的表面上氮、氧的表面積為多少。nHBDon 則是代表氫鍵給予者（通常是 N、O、F 等電負度大的原子）的數量，piPC2 則和 piPC1 類似，只是將他的級數上升到了 2 級，最後 SlogP_VSA1 是經過 Wildman-Crippen 方法計算出的 LogP 和凡得瓦表面積有關。

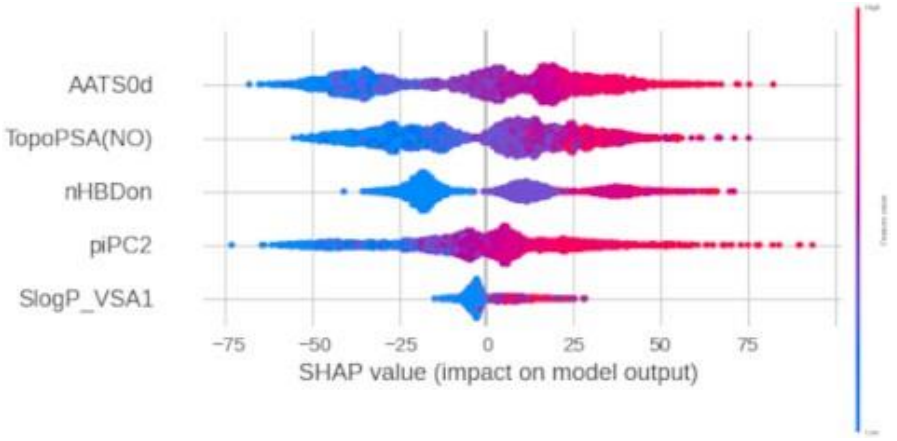


圖 15: 熔點之 Shapley values 分析（前五重要）（作者自行繪製）

二、淺度機器學習與深度機器學習比較

本章節將比較深度和淺度機器學習表現的差異，深度機器學習以表現較好的 CMPNN 作為代表。

首先對於沸點而言，我們分別對於兩個指標（R^2 及 MAE）做出下列虛無假設（H_0）：

1. $R^2_{CMPNN} = R^2_{shallow}$
2. $MAE_{CMPNN} = MAE_{shallow}$

3. 兩種模型一樣好

並透過學生 t 檢定（student's t test）計算，MAE 結果如下：

- T 值為：1.4964，小於 T 臨界值 1.7341
- p 值為：0.1519

故在 95% 的信心區間內接受 H_0，並無顯著差異。

另一方面，R^2 結果如下：

- T 值為：-1.8024，小於 T 臨界值-1.7341
- p 值為：0.0441

故在 95% 的信心區間內拒絕 H_0，淺度機器學習較深度機器學習好。

雖然從結果上來看 R^2 中淺度機器學習較深度機器學習確實會比深度機器學習好，但是因為 R_2 較 MAE 而言比較誤差的方式並不是完全準確的，因此本研究認為結果仍應為無顯著差異，即接受虛無假設。值得注意的是深度機器學習模型的表現標準差很大，是沸點的 10 倍左右，亦即很不穩定。

其次對於熔點而言，因為淺度機器學習的平均值加上 3 個標準差後仍然小於深度機器學習，所以較為容易判斷，且經計算後 T 值（8.9907）確實遠大於 T 臨界值（1.7341），p 值約為 2e-8，而 R^2 也有類似情況。故在熔點之中，深度機器學習模型的表現顯著高於淺度（$p<0.05$）。這可能係因為深度機器學習可以捕捉非線性特徵，且複雜度較高，較可以捕捉分子的特徵。

關於淺度是否會比深度模型好的比較可能會因為資料量、模型複雜度及資料特性有關，故深度模型在某些情況下如本研究中的沸點就不會顯著高於淺度。

三、本研究表現和以往研究對比

本章節統整了各種嘗試過的模型並將之和其他研究進行對比，整理如下：

對於沸點而言，經過嘗試不同模型結構及特徵萃取方式，發現在使用相同資料集的情況下表現最好的模型是 CMPNN，其 R^2 分數最高可達到 0.76，比較結果可參考表8。

模型 (b.p.)	R^2	MAE	備註
淺度機器學習平均	0.68	25.67	Mordred+Extra Tree Regressor
淺度機器學習最佳	0.71	24.04	Mordred+Extra Tree Regressor
OpenChem	0.71	20.09	
CMPNN(10 folds mean)	0.60	26.67	
CMPNN(10 folds best)	**0.76**	**23.89**	
Qu et al., 2022	X	5.8	非本研究資料集
Jocab method*	X	11.84	非本研究資料集

表 8: 深度機器學習模型之比較（沸點）（作者自行繪製）

* 注：Jocab method 是指Stein and Brown, 1994所提出的一種方式，其計算方式簡單，主要是計算每個官能基會對性質有多少影響，但同時根據以往研究，此方式較不能精準的捕捉訊息。

和Qu et al., 2022的研究相比，雖然在 MAE 上有差距，但此極有可能係因資料集的數量和多樣性有關，該研究僅使用了約三千筆資料，而本研究將近是該研究的四倍，因此本研究之模型在使用相同資料集的情況下未必會比較差，然而該研究並未公佈原始模型的程式碼或使用的資料集內容。

而針對熔點的部份，本研究最高可以達到 $R^2 = 0.87$，MAE=23.73K，使用的模型一樣是 CMPNN，各模型的表現比較如下表：

模型 (m.p.)	R^2	MAE	備註
淺度機器學習平均	0.80	30.16	Mordred+Extra Tree Regressor
淺度機器學習最佳	0.83	28.89	Mordred+Extra Tree Regressor
OpenChem	0.72	38.84	
CMPNN 平均	0.85	24.81	
CMPNN 最佳	**0.87**	**23.73**	
Feng et al., 2024（GCN）	0.77	32.32	
Feng et al., 2024（GCN）	0.76	28.79	非本研究資料集
CMPNN 平均	0.77	27.45	非本研究資料集
CMPNN 最佳	0.81	24.93	非本研究資料集

表 9: 深度機器學習模型之比較（熔點）（作者自行繪製）

和Feng et al., 2024的研究相比，本研究的 R^2 分數高出了 0.1，且同時在 MAE 的部份也相進步了 10K，比起過去研究不論使用本研究或該研究的資料集均有顯著提昇，尤其在本研究的資料集上。

四、未來展望

本研究未來可以朝著更準確、更能解釋的方向前進。

準確性的部份可以由直接修改模型結構，如加入全域變數（如分子量等可能有關性質）或是以其他類似模型代替 BatchGRU，亦可依據伍之二的殘差分析適當的加入其他描述符以捕捉和性質較相關但現有分子描述器沒有的特徵。

解釋性的部份則可以進一步以神經網路的特徵去對照淺度機器學習的 SHAP 值分析結果，並以此觀察是否具有一致性，增加神經網路的可信任性（Trustworthy）。

柒、結論

從製作實驗預報開始發想，本研究使用了兩種不同的方式（淺度及深度機器學習）成功的預測並解釋了熔、沸點的預測，其中表現最佳的 CMPNN 模型，預測表現可以達到沸點： $R^2 = 0.76$，$MAE = 23.89K$，熔點：$R^2 = 0.87$，$MAE = 23.73K$，均將 MAE 下降到 23K 左右，並且找出 piPC1 和沸點的預測最有相關，而熔點則是 AATSod，未來此研究可以應用在更多性質如藥性、溶解度、反應性等的預測上。

捌、參考文獻

Ali, M. (2020, April). *Pycaret: An open source, low-code machine learning library in python* [PyCaret version 1.0]. https://www.pycaret.org

Andrew Lang, J.-C. B. (2011). Open Notebook Science Melting Point Data — lulu.com [[Accessed 21-09-2024]].

Bell, C., Cortes-Pena, Y. R., & Contributors. (2016-2024). Chemicals: Chemical properties component of chemical engineering design library (chedl) [Accessed: 2024-09-10]. https://github.com/CalebBell/chemicals

Chen, C., & Ong, S. P. (2022). A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, *2*(11), 718–728. https://doi.org/10.1038/s43588-022-00349-3

Chen, C., Ye, W., Zuo, Y., Zheng, C., & Ong, S. P. (2019). Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, *31*(9), 3564–3572. https://doi.org/10.1021/acs.chemmater.9b01294

Feng, H., Qin, L., Zhang, B., & Zhou, J. (2024). Prediction and interpretability of melting points of ionic liquids using graph neural networks. *ACS Omega*, *9*(14), 16016–16025. https://doi.org/10.1021/acsomega.3c09543

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 1263–1272.

Hughes, L. D., Palmer, D. S., Nigsch, F., & Mitchell, J. B. O. (2008). Why are some properties more difficult to predict than others? a study of qspr models of solubility, melting point, and log p [PMID: 18186622]. *Journal of Chemical Information and Modeling*, *48*(1), 220–232. https://doi.org/10.1021/ci700307p

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks.

Ko, T. W., Nassar, M., Miret, S., Liu, E., Qi, J., & Ong, S. P. (2021, June). *Materials Graph Library* (Version 0.5.3). https://doi.org/10.5281/zenodo.8025189

Korshunova, M., Ginsburg, B., Tropsha, A., & Isayev, O. (2021). Openchem: A deep learning toolkit for computational chemistry and drug design [PMID: 33393291]. *Journal of Chemical Information and Modeling*, *61*(1), 7–13. https://doi.org/10.1021/acs.jcim.0c00971

Li, B., & Rangarajan, S. (2022). A diversity maximizing active learning strategy for graph neural network models of chemical properties. *Mol. Syst. Des. Eng.*, *7*, 1697–1706. https://doi.org/10.1039/D2ME00073C

Moriwaki, H., Tian, Y.-S., Kawashita, N., & Takagi, T. (2018). Mordred: A molecular descriptor calculator. *Journal of Cheminformatics*, *10*(1), 4. https://doi.org/10.1186/s13321-018-0258-y

Pocha, A., Danel, T., Podlewska, S., Tabor, J., & Maziarka, Ł. (2021). Comparison of atom representations in graph neural networks for molecular property prediction. *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN52387.2021.9533698

Qu, C., Kearsley, A. J., Schneider, B. I., Keyrouz, W., & Allison, T. C. (2022). Graph convolutional neural network applied to the prediction of normal boiling point. *Journal of Molecular Graphics and Modelling*, *112*, 108149. https://doi.org/10.1016/j.jmgm.2022.108149

Reiser, P., Neubert, M., Eberhard, A., Torresi, L., Zhou, C., Shao, C., Metni, H., van Hoesel, C., Schopmans, H., Sommer, T., & Friederich, P. (2022). Graph neural networks for materials science and chemistry. *Communications Materials*, *3*(1), 93. https://doi.org/10.1038/s43246-022-00315-6

Schütt, K. T., Kindermans, P.-J., Sauceda, H. E., Chmiela, S., Tkatchenko, A., & Müller, K.-R. (2017). Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. https://arxiv.org/abs/1706.08566

Song, Y., Zheng, S., Niu, Z., Fu, Z.-h., Lu, Y., & Yang, Y. (2020, July). Communicative representation learning on attributed molecular graphs [Main track]. In C. Bessiere (Ed.), *Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI-20* (pp. 2831–2838). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2020/392

Stein, S. E., & Brown, R. L. (1994). Estimation of normal boiling points from group contributions. *Journal of Chemical Information and Computer Sciences*, *34*(3), 581–587. https://doi.org/10.1021/ci00019a016

Tsubaki, M., & Mizoguchi, T. (2020). On the equivalence of molecular graph convolution and molecular wave function with poor basis set. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 1982–1993, Vol. 33). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1534b76d325a8f591b52d302e7181331-Paper.pdf

Wiener, H. (1947). Structural determination of paraffin boiling points [PMID: 20291038]. *Journal of the American Chemical Society*, *69*(1), 17–20. https : / / doi . org / 10 . 1021 / ja01193a005

Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., & Zheng, M. (2020). Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism [PMID: 31408336]. *Journal of Medicinal Chemistry*, *63*(16), 8749–8760. https://doi.org/10.1021/acs.jmedchem.9b00959

Zhang, S., Tong, H., Xu, J., & Maciejewski, R. (2019). Graph convolutional networks: A comprehensive review. *Computational Social Networks*, *6*(1), 11. https://doi.org/10.1186/s40649-019-0069-y

葉宗融 & 蔡明剛. (2020). 以機器學習方法預測分子熔點的應用範例. 化學, *78*(2), 119–125. https://doi.org/10.6623/chem.202006_78(2).008

【評語】190025

建議本研究設法取得化合物熔點、沸點的真實資料，以供機器學習來訓練預測的模型。

另外建議本研究應朝全面性的化合物做熔點、沸點做預測，以增加本系統的可用性。新特徵的導入也是需要考慮的方向。