國立臺灣師範大學理學院資訊工程學系

碩士論文

Department of Computer Science and Information Engineering

College of Science

National Taiwan Normal University

Master's Thesis

基於 Transformer 的化合物-蛋白質交互作用預測方法之改進

An Improved Transformer-based Approach for Compound-Protein Interaction Prediction

陳威宇

Wei-Yu Chen

指導教授：陳柏琳 博士

Advisor: Berlin Chen, PhD.

中華民國 114 年 8 月

August 2025

摘要

近年來，化合物-蛋白質交互作用 (Compound-Protein Interaction, CPI) 預測已經成為計算化學領域的研究熱點之一。隨著深度學習技術的興起，越來越多的基於神經網路的 CPI 預測方法得到了開發和應用。其中，Transformer 模型是採用自注意力機制 (Self-attention) 的深度學習模型，具有強大的建模能力，因此有越來越多模型使用了此方法。不過，基於此方法的模型在預測 CPI 的任務上存在著一些問題，例如訓練的成本太大、對於3D 空間相互作用的捕捉能力較弱等，而這些問題也影響到預測的準確率。為了找到比傳統 Transformer 還更能準確預測的方法，我們從模型架構、輸入特徵的選擇以及損失函數等面向尋找改進的方法，期望能找出可以提升準確率，甚至降低運算成本的方法。本論文以 CAT-CPI (Ying et al., 2022) 的模型架構為基礎，結合 TransformerCPI (Chen et al., 2020) 對於化合物特徵的提取方式，提出了基於 Transformer 的 CPI 預測之改進方法。TransformerCPI 針對一維的 SMILES 序列產生了對應的原子特徵，而 CAT-CPI 則是使用二維的化合物圖像作為輸入，利用 CNN 學習化合物圖像的局部細節特徵，並且取得了優秀的結果。因此本模型結合兩者的特色，同時以一維的原子特徵和二維的分子圖像作為輸入，利用不同的化學結構資訊互補來提高模型的預測能力。此外我們也嘗試以 Performer、Conformer 等不同的架構取代傳統的 Transformer 來提升預測的準確率與運算的速度，並觀察不同的損失函數 (Loss Functions) 對於訓練結果的影響。我們使用 Human、Celegans 以及 Davis 資料集對所有改進方法進行實驗，發現與只使用分子圖像的方法相比，原子特徵與分子圖像結合的輸入能有效提升預測的準確率，且以 Performer 和 Conformer 等模型取代 Transformer 也可些微提升預測的能力。

關鍵字：深度學習、CPI、DTI、分子圖像、Transformer

Abstract

In recent years, Compound-Protein Interaction (CPI) prediction has emerged as one of the major research focuses within the field of computational chemistry. With the rapid development of deep learning technologies, an increasing number of neural network-based CPI prediction methods have been proposed and adopted. Among them, Transformer models—which leverage the self-attention mechanism—offer strong modeling capabilities and have thus become widely utilized. However, Transformer-based approaches still face several challenges in CPI prediction tasks, such as high computational costs and limited ability to capture three-dimensional spatial interactions, which in turn affect predictive accuracy. To explore more effective methods beyond the conventional Transformer architecture, we investigate possible improvements in model design, input representation, and loss function strategy, aiming to enhance prediction performance while potentially reducing computational burden. This thesis presents an improved Transformer-based approach for CPI prediction by building upon the architecture of CAT-CPI (Ying et al., 2022) and incorporating the compound feature extraction technique used in TransformerCPI (Chen et al., 2020). TransformerCPI generates atomic-level features from one-dimensional SMILES sequences, while CAT-CPI utilizes two-dimensional compound images as input and applies convolutional neural networks (CNNs) to learn local structural details, achieving promising results. Our proposed model integrates both approaches by simultaneously using one-dimensional atomic features and two-dimensional molecular images as inputs, allowing complementary chemical structure information to enhance prediction capability. Additionally, we explore alternative model architectures such as Performer and Conformer to replace the standard Transformer, aiming to improve prediction accuracy and computational efficiency. We also examine the impact of different loss functions on training outcomes. Experiments conducted on the Human, Celegans, and Davis datasets demonstrate that incorporating both atomic features and molecular images yields better predictive performance compared to using molecular images alone. Moreover, replacing the Transformer with Performer or Conformer models results in moderate improvements in accuracy.

Keywords: Deep Learning, CPI, DTI, Molecular Images, Transformer

目錄

圖目錄

表目錄

第一章 緒論

1.1 研究背景

Compound-Protein Interaction (CPI) 是指化合物與蛋白質之間的相互作用。化合物通常是潛在的藥物分子，而蛋白質則是可能被這些化合物調節的靶標分子 (Target Molecule) 。

CPI 預測是一種計算化學 (Computational Chemistry) 任務，旨在預測化合物與蛋白質之間的相互作用。這是一項重要的任務，因為它可以為藥物發現 (Drug Discovery) 提供有價值的引導訊息。在藥物發現過程中，科學家需要識別分子的潛在藥物性質，包括其與蛋白質的相互作用。CPI 預測可以透過計算模擬 (Computational Simulation) 來實現，使用各種計算方法，包括機器學習和深度學習等方法。這些方法使用大量的化合物和蛋白質結構數據，透過學習這些數據的模式和規律，預測新的化合物與蛋白質之間的相互作用。這些預測結果可以為藥物發現提供指引，縮短藥物研發的時間和成本，同時也可以為生命科學研究提供有價值的資訊，如分子結構分析、蛋白質功能預測等。

近年來，CPI 預測已經成為計算化學領域的研究熱點之一。在過去的幾十年中，隨著計算技術和生物科技的快速發展，累積了大量的化合物和蛋白質結構數據。這些數據為 CPI 預測提供了重要的基礎，並促進了各種計算方法的發展。

早期的 CPI 預測方法主要使用分子對接技術 (Molecular Docking) [1]，該技術可以模擬分子之間的相互作用。但是，分子對接方法在計算效率和準確性方面存在一定的局限性。隨著深度學習技術的興起，越來越多的基於神經網路的 CPI 預測方法得到了開發和應用。這些方法利用深度神經網路學習蛋白質和化合物的特徵表示 (Feature Representations) ，從而實現對 CPI 的準確預測。

使用深度學習網絡進行 CPI 預測的常用方法包括 Graph Convolutional Networks (GCN) [2]、Recurrent Neural Networks (RNN) [3]、Siamese Networks [4]

以及 Transformer Networks [5] 等。GCN 利用化合物和蛋白質的圖結構訊息，透過卷積操作 (Convolution) 提取特徵，構建 CPI 預測模型。RNN 透過建立序列模型 (Sequence Models)，將蛋白質和化合物的序列訊息輸入模型，學習其之間的相互作用。Siamese Networks 透過將化合物和蛋白質的特徵映射到相同的空間中，建立相似度模型 (Similarity Models)，預測它們之間的相互作用。Transformer Networks 則是透過對蛋白質和化合物的序列訊息進行編碼，利用多頭注意力機制 (Multi-head Attention) 學習其之間的相互作用。近年來有越來越多模型使用了基於 Transformer Networks 的方法，原因在於該方法主要有以下幾個優勢：

1. 強大的建模能力：Transformer Networks 是一種基於自注意力機制 (Self-attention) 的深度學習模型，具有強大的建模能力，能夠學習到複雜的非線性關係，適用於多種不同的數據類型，如序列、圖等。
2. 能夠利用序列訊息：CPI 預測任務中，蛋白質和化合物的序列訊息往往是非常重要的，Transformer Networks 能夠有效地捕捉這些資訊，並用於預測相互作用。
3. 處理不定長的輸入：蛋白質和化合物的序列長度可能不同，Transformer Networks 能夠處理不定長的輸入，能夠更好地適應不同的數據特點。
4. 穩健性 (Robustness) 強：相對於其他深度學習模型，如 RNN 和 CNN，Transformer Networks 對輸入序列中的隨機噪聲和錯誤具有更強的穩健性。
5. 可解釋性強：相對於其他深度學習模型，Transformer Networks 具有更好的可解釋性，能夠可視化注意力權重，直觀地顯示哪些位置在預測過程中得到了更多的關注。

由此可見，使用 Transformer Networks 進行 CPI 預測有許多優點。然而，該方法也存在一些需要留意的問題。首先，Transformer Networks 的模型複雜度相對較高，需要大量的參數和計算資源，因此在實際應用中需要考慮到硬體設備和運算速度等因素。此外，此方法需要大量的蛋白質和化合物資料，並且資料的質量和數量也會對預測精度產生重要影響，因此需要保證資料的質量和數量。由於

CPI 預測往往會受到資料集本身的偏差和不平衡性等因素的影響，因此在使用 Transformer Networks 進行 CPI 預測時，可能需要進行相應的數據處理和模型優化等措施，以提高預測精度和穩定性。

1.2 研究動機

在藥物開發與再定位等生醫應用中，準確預測化合物與蛋白質之間的交互作用具有關鍵價值。相較於實驗方法（如高通量篩選、高解析結構分析）所需的高昂成本與長期週期，基於深度學習的計算預測工具提供了更具效率與擴展性的解決方案。然而，現有 CPI 預測方法仍存在多方面的挑戰與限制，導致其在實務應用中難以全面發揮效益。

首先，許多模型仍仰賴單一輸入模態，例如僅以化合物的 SMILES 表示或蛋白質序列作為輸入，缺乏對分子結構特性與序列間互補訊息的整合，限制了模型對交互行為的理解深度。即使有部分研究嘗試加入分子圖、圖像或結構資訊，但模態融合策略仍多為早期串接（Early Concatenation），缺乏系統性分析與自適應權重設計，導致模態間可能產生訊號衝突或冗餘。

其次，在模型架構層面，雖然 Transformer 因其強大的序列建模能力而被廣泛應用於生醫序列任務，但其注意力機制在長序列或高維輸入下的計算複雜度高，使得模型在面對實際資料時經常遭遇效率瓶頸。此外，原始架構並未考慮結構資訊的區域依賴關係，對於建模局部結構互動仍有其侷限。

最後，CPI 資料本身往往呈現高度不平衡，正樣本數遠低於負樣本，使得模型在訓練過程中易於偏向預測多數類別，造成召回率下降，進而影響整體應用價值。傳統使用的交叉熵損失函數並未設計用於處理此類不平衡情境，導致分類邊界難以有效對齊。

1.3 研究貢獻

在本研究中，我們將使用基於 Transformer Networks 的方法進行 CPI 預測。以現

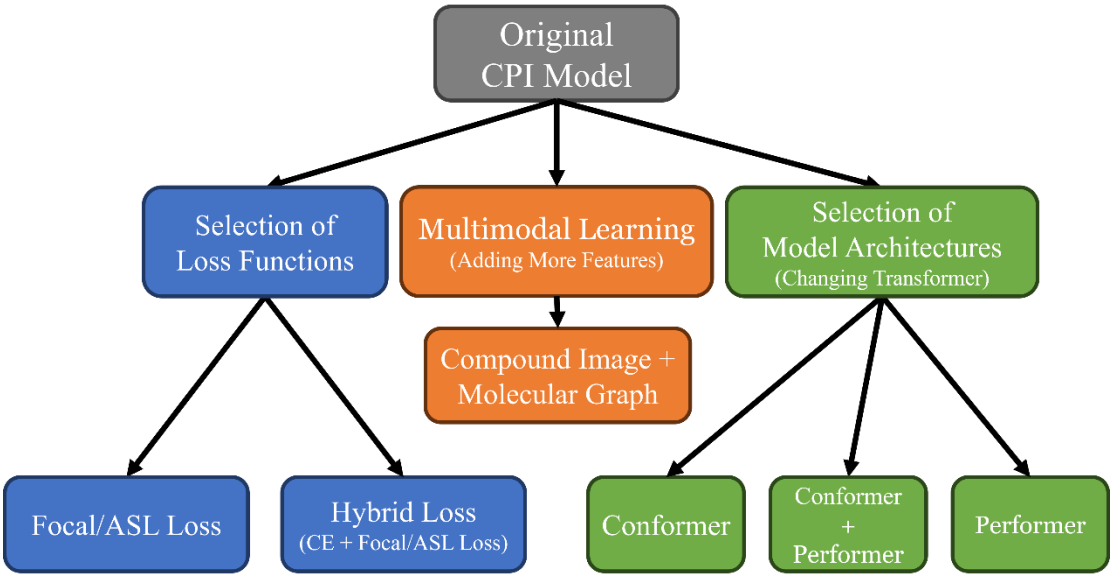


圖 1 本研究的 CPI 模型改進方向，由左而右分別是更改損失函數、多模態學習（增加更多特徵）以及模型架構的選擇

有表現優良的模型為基礎，提出一個可彈性調整輸入模態、模型架構與損失函數的多模態預測框架。我們以各種不同面向改進該模型，最後比較各種改進方法的優劣，找出最好的方式。改進的方法主要包括增加訓練的特徵、改良神經網路的架構，以及修改損失函數等，如圖 1 所示。

在訓練特徵方面，我們將採用多模態學習 (Multimodal Learning)。對於化合物的資料部分，我們利用 1D 的 SMILES 序列產生兩種訓練資料作為輸入：分子圖 (Molecular Graphs) 和分子圖像 (Molecular Images)。分子圖包含原子特徵和共價鍵的資訊，而分子圖像能清楚的顯示分子的原子、結構和化學鍵等資訊。藉由多種訓練資料，期望能利用不同的化學結構資訊互補，提高模型的預測能力。

在神經網路部分，一般的 Transformer 存在著計算複雜度高、難以建模局部結構、長距離交互效果有限等問題，因此在本研究中也將嘗試使用其他的 Transformer 來強化建模能力。例如引入局部卷積 (CNN) 來建模局部結構，然後用 Transformer 捕捉全局依賴關係的 Conformer [6]、能夠使計算複雜度從 $O(N^2)$ 降至 $O(N)$ 的 Performer [7] 和 Linformer [8] 、透過連續狀態空間模型 (State Space Model, SSM) 來替代自注意力的 Mamba [9] 等，期望能提升模型訓練的效

率、降低計算成本。

在損失函數方面，雖然一般對於分類問題都是採用交叉熵 (Cross Entropy) [10] 作為損失函數，但是對於不平衡的數據則可能降低分類的準確率。Focal Loss [11] 便是為了解決不平衡數據的問題所提出的方法，透過調整損失的加權方式減少了容易樣本的損失貢獻，提高模型在難分類樣本上的準確率。此外還有 Asymmetric Loss [12]，他針對 False Positive (FP) 和 False Negative (FN) 設定不同的權重，進一步優化 Precision 和 Recall，是 Focal Loss 的改良版本。本研究除了將交叉熵直接改用其他損失函數以外，還會將交叉熵和 Focal Loss 或 Asymmetric Loss 混合，根據 epoch 來調整兩個損失函數的權重，如此一來可以結合 Cross Entropy 的穩定收斂 和 Focal Loss 的少數類別增強效果。

本研究的整體目標為：一、建立一個結構具擴展性與模組化的 CPI 預測模型；二、驗證多模態特徵輸入與高效架構對模型表現之影響；三、提出具穩定性與泛化能力的訓練策略，以實際提升預測準確性與實用性。透過本研究，我們期望為 CPI 任務提供更具彈性、效率與準確度的解決方案，並對未來相關領域如虛擬篩選、藥物再定位等應用產生實質貢獻。

第二章 文獻探討

2.1 理論基礎

在一篇整理 CPI 預測方法的綜述文章中 [13]，將 CPI 預測的整體流程大致上分為「資料準備」、「模型訓練」以及「預測」三個階段。

2.1.1 資料準備

在資料準備階段，首先要從特定的資料庫中取得化合物與蛋白質的資料。這些資料庫可以細分為三類：以化學為中心 (Chemistry-Centric) 、以蛋白質為中心 (Protein-Centric) 以及綜合的 (Integrated) 資料庫。以化學為中心的資料庫主要側重於整合來自化學實驗的訊息。它們包含 Simplified Molecular-Input LineEntry System (SMILES)、InChI 密鑰等資料及其具有相應親和力的相互作用/靶向蛋白質。例如 PubChem, ChEMBL, DUD-E 等。蛋白質資料庫一般提供序列訊息，很少包含與化合物相關的訊息。例如 UniProt, Protein Data Bank 等。其他資料庫包括除化合物或蛋白質之外的綜合訊息，例如與基因、疾病或表型 (Phenotypes) 的關聯。例如 BindingDB, Davis 等。

取得化合物與蛋白質的資料後，接著根據資料的格式選擇合適的編碼方式 (Encoding)。在化合物方面，常見的資料格式包含字串 (Strings)、指紋 (Fingerprints)、圖形 (Graphs)。基於字串的方法將化合物轉換為上下文感知字串 (Context-Aware Strings)，例如 SMILES、SMARTS 與 SELFIES。SMILES 可以編碼為 one-hot 和 multi-hot 向量的混合。此外也可以使用 word2vec [14] 進行編碼，它通過將字符映射到實數向量來構建詞嵌入 (Word Embeddings)。word2vec 還可以結合 RNN 等序列模型，透過將固定長度的字符視為「一個詞」來生成整個化學句子的強大嵌入。基於 SMILES 的 SMARTS 專注於局部子結構 (Localized Substructures)，SELFIES 則能生成保證有效分子結構的字串。化學指紋 (Chemical Fingerprints) 透過將化合物與預定義的子結構集進行比較，將化合物編

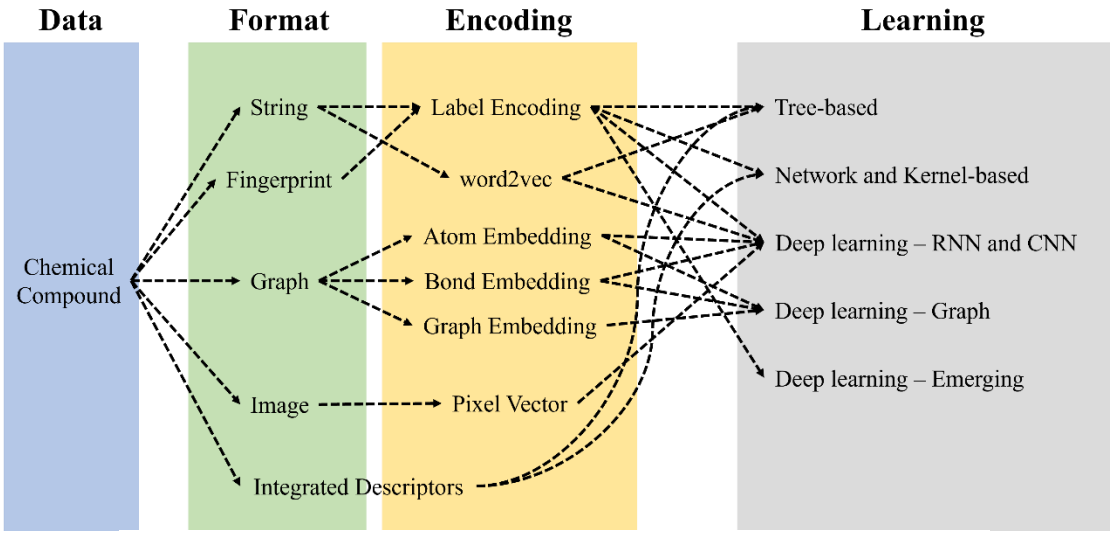


圖 2 化合物的資料處理流向（從格式到模型），引用自[13]

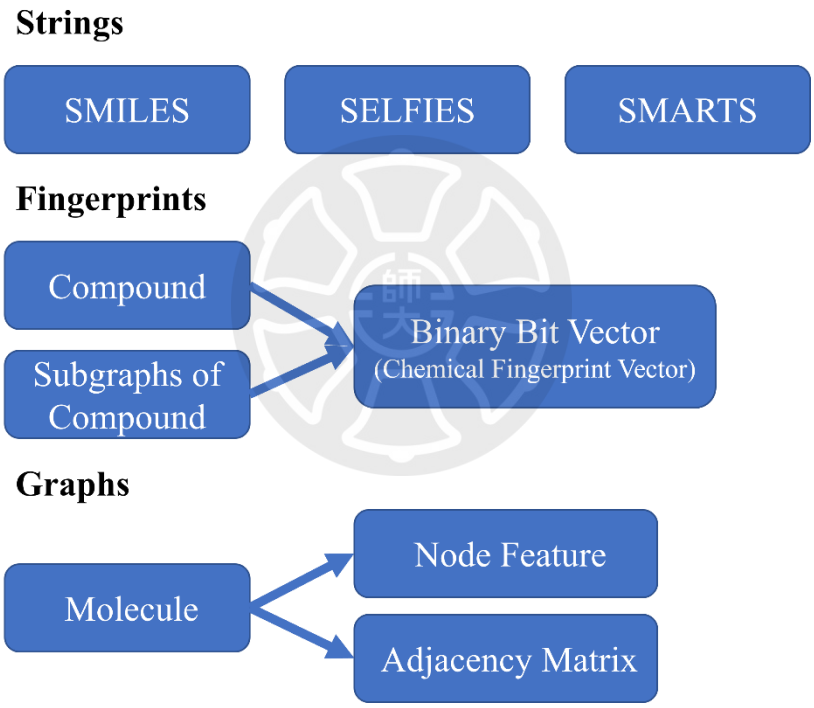


圖 3 化合物的資料格式與編碼機制，引用自[13]

碼為二進制位向量。化合物的圖形表示則是將輸入的分子轉化為一組矩陣——通常表示為具有原子/鍵訊息的圖形的鄰接矩陣 (Adjacency Matrices)。此外還有使用圖像 (Image) 來表示分子的方式，因為圖像可以很清楚的表示物體的位置、幾何形狀和空間結構。有了分子的圖像，我們就能清楚看到原子、結構、化學鍵等訊息，且比起字串、圖形更能包含相當複雜的訊息。圖 2 整理了化合物的資料格式、對應的編碼方式以及模型，圖 3 整理了化合物的編碼機制。

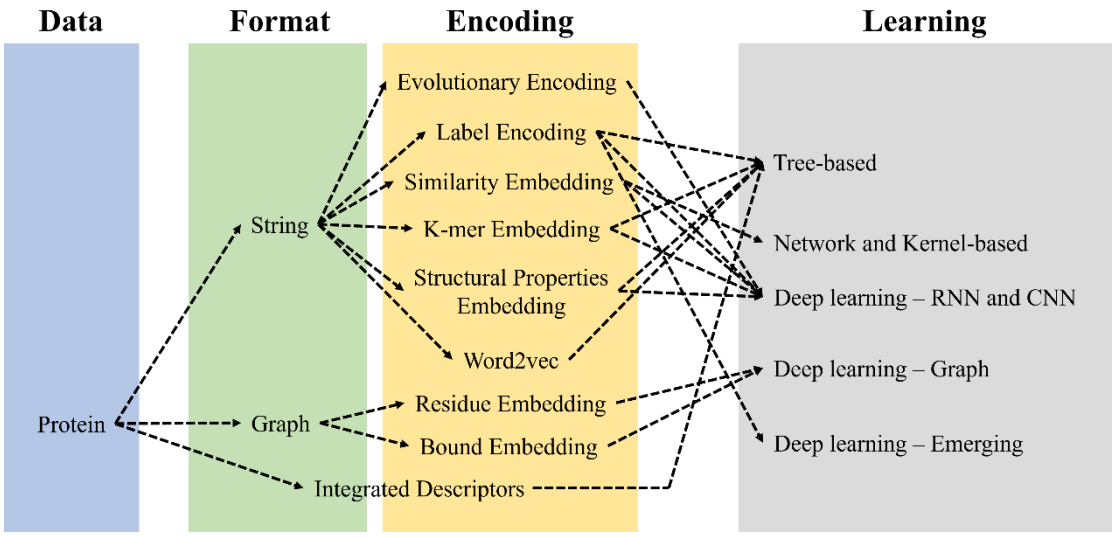


圖 4 蛋白質的資料處理流向（從格式到模型），引用自[13]

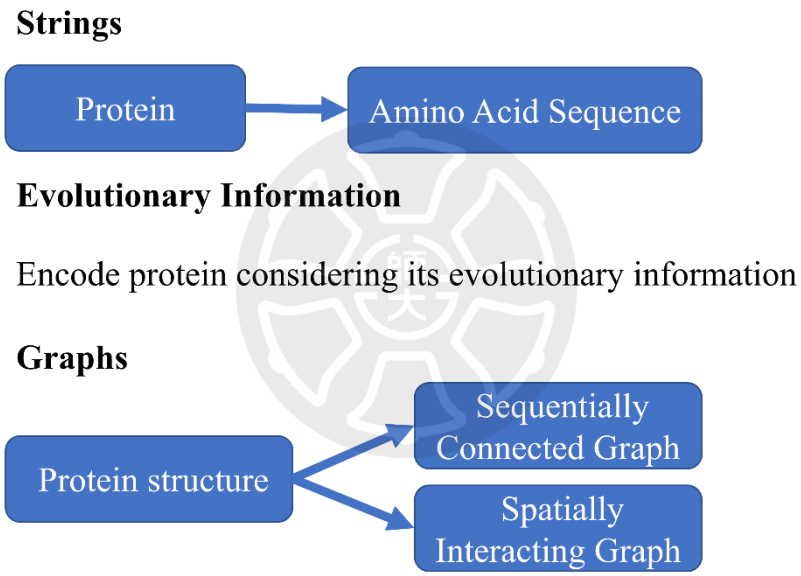


圖 5 蛋白質的資料格式與編碼機制，引用自[13]

蛋白質基本上是一個氨基酸殘基的序列，包含了高度保守的進化訊息(Evolutionary Information)。考慮到其結構特性，通常會用 one-hot, word2vec 或 *k*-mer-based，以序列的方式進行編碼，例如 TransformerCPI [15] 即是使用 word2vec 進行編碼。另外也可以將蛋白質結構轉換為化學屬性空間圖 (Spatial Graph)，其中殘基節點和預設距離內兩個殘基之間的邊。圖 4 整理了蛋白質的資料格式、對應的編碼方式以及模型，圖 5 整理了蛋白質的編碼機制。

AlphaFold [16] 和 AlphaFold2 [17] 表明，將蛋白質進化訊息與蛋白質結構

訊息結合使用，對於從蛋白質序列預測蛋白質結構非常有效。其中，AlphaFold2 是由 CNN 和自注意力機制構成的深度學習網路。網路架構包含兩個主要模塊，一個是用於預測蛋白質序列的二級和三級結構的序列模型，另一個是用於預測蛋白質的原子細節的結構模型。在訓練過程中，AlphaFold2 利用了大量的蛋白質序列和結構數據進行訓練，並且採用了一種稱為知識蒸餾 (Knowledge Distillation) 的技術，將多個網路的預測結果融合在一起，從而提高了預測的準確度。AlphaFold2 在第 14 屆的蛋白質結構預測比賽 (CASP14) 中的平均相似度分數 (GDT_TS) 達到了約 92.4，比第二名高出約 20 個百分點，被稱作結構生物學「革命性」的突破、蛋白質研究領域的里程碑。

最新版本的 AlphaFold3 [18] 則進一步拓展了預測能力，不僅能處理蛋白質本身，還可解析蛋白質－配體、蛋白質－核酸等複合分子結構。其架構在 AlphaFold2 基礎上整合了擴散 (Diffusion-based) 生成模型與更強化的 Evoformer（或稱 Pairformer），能同時預測多分子交互作用，並在蛋白質－配體交互準確率上提升 50%，在核酸交互方面也有明顯進步。此外，AlphaFold3 於 2024 年 11 月正式開放學術用途之模型權重與程式碼，使研究者能更自由地應用於藥物設計與複合體結構預測。

2.1.2 模型訓練

資料經過編碼處理後，將會作為一個或多個模型的組合的輸入，以學習資料的模式 (Patterns)。這些模型可以分為「機器學習」與「深度學習」兩大類。機器學習的方法包含了使用基於決策樹 (Decision Tree) 的 Tree-based Methods，以及使用基於支持向量機 (Support Vector Machine, SVM) 的 Network-based and Kernel-based Methods。深度學習的方法主要包含了利用 RNN 與 CNN 來處理序列訊息的 Sequence-based Methods，以及使用 GCN 處理圖形資料的 Graph-based Methods。其他還有使用生成對抗網路 (Generative Adversarial Network, GAN) 或是自編碼器 (Autoencoder) 來做特徵提取的方法。

2.1.3 預測

在使用資料訓練模型後，可以使用不同的指標來評估模型，可分為回歸方法 (Regression) 與分類方法 (Classification)。回歸方法是藉由預測親和力值 (Affinity Value) 來做預測 CPI，分類方法中則是藉由預測交互標籤 (Interaction Label) 來預測 CPI。CPI 預測的整體流程如圖 6。

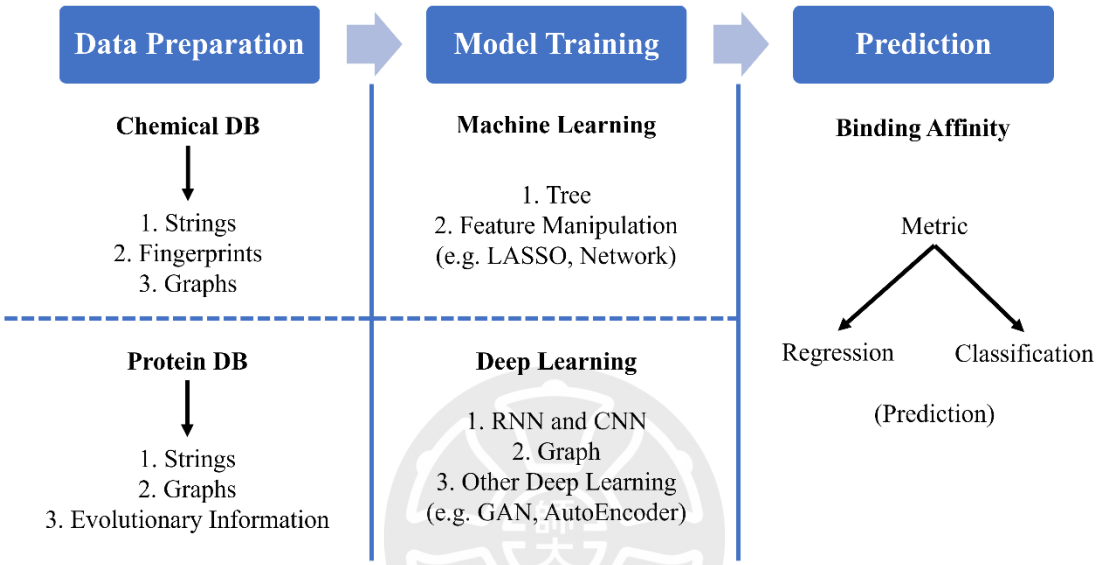


圖 6 CPI 預測的過程概覽，引用自[13]

2.2 Transformer 模型

2.2.1 簡介

Transformer 模型 [5] 是一種採用自注意力機制的深度學習模型，在 CPI 預測中屬於 Sequence-based Methods。由 Google 在 2017 年提出，在自然語言處理領域取得了極好的效果，近年來也被廣泛應用於其他領域，如圖像處理、推薦系統以及音訊處理等。

Transformer 的主要結構包含編碼器 (Encoder) 和解碼器 (Decoder) 兩個部分，每個部分都由多層子層組成。編碼器的任務是將輸入序列進行編碼，解碼器的任務是根據編碼器的輸出和上一個時刻的輸出，預測下一個時刻的輸出。每個子層包含多個注意力機制和前饋神經網路 (Feedforward Neural Network)，其中注

意力機制可以將輸入序列中的重要訊息與上下文相關性相結合，並將其轉換為輸出序列。

在編碼器中，輸入序列先經過一個嵌入層轉換為向量表示，然後經過多層 Self-Attention 和前饋神經網路進行特徵提取和編碼。在解碼器中，輸出序列的前一個時刻的輸出經過嵌入層和多層 Masked Self-Attention 進行特徵提取和編碼，然後再和編碼器的輸出進行多層 Attention 進行特徵提取和解碼。

Transformer 模型的基本構建單元是縮放點積注意力 (Scaled Dot-Product Attention) 單元。對於每個注意力單元，Transformer 模型取三個輸入，分別為查詢 (query) Q、鍵 (key) K以及值 (value) V，學習一組權重矩陣 (W_Q, W_K, W_V)。對所有標記的注意力計算可以表示為使用 softmax 函式的一個大型矩陣計算，如式 (2.1)。可分成四個步驟：

1. 計算矩陣Q和K的點積S
2. 為求梯度穩定性將S標準化，即除以向量維度的平方根$\sqrt{d_k}$
3. 使用 softmax 函數轉換為機率
4. 乘以 V獲得加權值矩陣。

$$Attention\,(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

一組權重矩陣 (W_Q, W_K, W_V) 稱為一個注意力頭 (Attention Head)，Transformer 模型中每一層都包含多個注意力頭。每個注意力頭都表示不同標記相互之間的注意力，而多個注意力頭則可以針對不同的「相關性」計算不同的注意力權重。

相較於傳統的 RNN 與 CNN，Transformer 沒有序列上的循環或卷積，因此可以進行平行計算，提高了訓練速度。使用自注意力機制來計算輸入序列中各個位置之間的相對重要性，進而在不同層次上對不同位置進行加權編碼，使模型可以更好地捕捉長距離依賴關係。此外也引入了殘差連接 (Residual Connection/Skip Connections) 和層規範化機制 (Layer normalization)，使模型更

易於訓練和穩定。在 CPI 預測中，利用 Transformer 模型對蛋白質序列和化合物分子進行建模，可以充分利用其強大的建模能力和平行計算的優勢，提高 CPI 預測的準確度和效率。

傳統的 Transformer 雖然在自然語言處理 (Natural Language Processing, NLP)、電腦視覺 (Computer Vision, CV)、藥物發現等領域取得巨大成功，但仍然存在一些問題。首先是計算效率的問題，Transformer 的 Multi-Head Self-Attention 需要計算整個序列的 QK^T 內積，導致時間複雜度為 $O(N^2d)$、記憶體需求為 $O(N^2)$，因此訓練成本極高。其次，Transformer 雖然可以捕捉長距離關係，但實際上長序列訊息可能衰減，且無法建模空間關係（例如 3D 分子結構），因此 Transformer 對長距離相互作用捕捉的能力較弱。再來，傳統 Transformer 只使用一維的序列訊息，但是分子是 Graph 結構、蛋白質有 3D 摺疊結構，因此使用傳統 Transformer 可能會錯過 3D 空間中的相互作用。最後，Transformer 模型通常需要大規模數據來訓練，但是 CPI 資料集通常有限，這可能導致 Transformer 容易過擬合。基於以上問題，後續有許多研究提出了解決方案，成功利用改進的模型提升性能。以下介紹一些方法，包括 Conformer [6] 、Performer [7] 、Linformer [8] 還有 Mamba [9]。相關的實作細節則會在第三章討論。

2.2.2 Conformer

在 2020 年由 Google 發表的論文中提出了將 Transformer 與 CNN 架構結合而成的新架構 Conformer，隨後成功應用於自動語音辨識 (Automatic Speech Recognition, ASR) 的技術上。Transformer 的 Self-Attention Layer 在大範圍內提取特徵資訊有較好的效果，但缺乏提取局部細微特徵；反之，CNN 的 Convolution Layer 可以很好的提取局部細微的特徵，但針對全域特徵則需要大量的參數與深度。Conformer 的架構便是希望能結合 Self-Attention Layer 與 Convolution Layer 各自的優點。將 Conformer 應用於分子化學和生物訊息學領域時，則可以用來處理分子的三維立體結構，預測分子的生物活性和與蛋白質的相互作用，且有助於

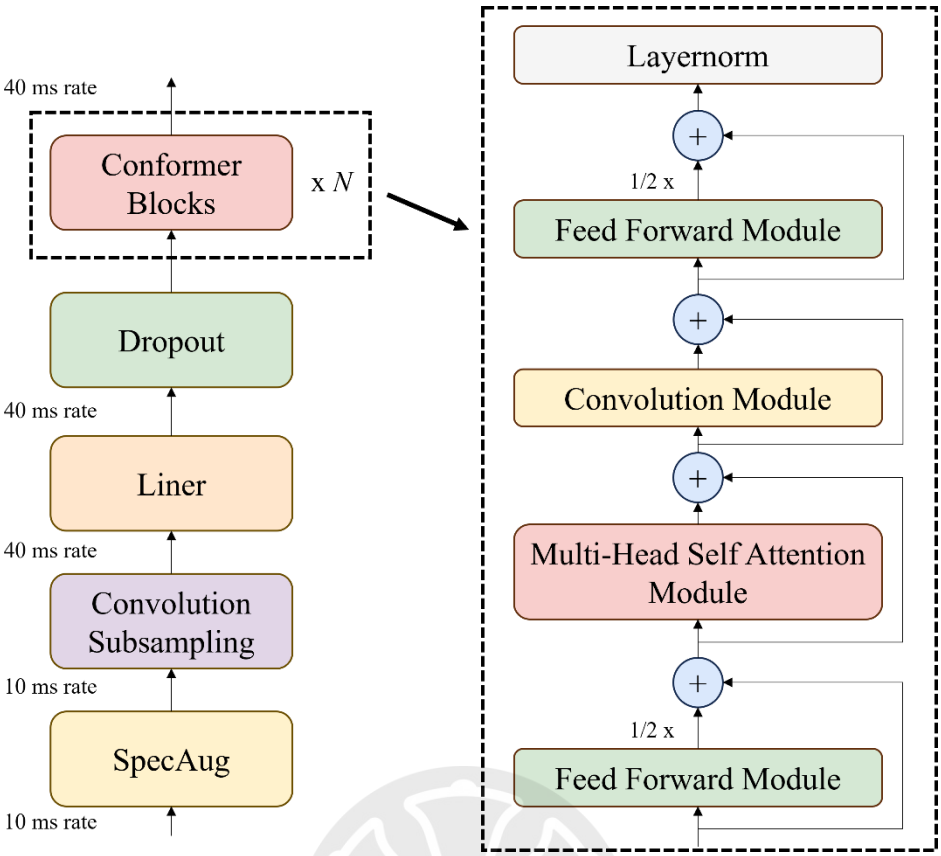


圖 7 Conformer 的網路架構，引用自[6]

處理分子結構的圖形表示，甚至允許在同一模型中結合不同類型的數據，例如分子結構和蛋白質序列。Conformer 的網路架構如圖 7 所示，其將 FFN 分為兩層放置於 block 首尾，並使用 GLU (Gated Linear Unit) 作為卷積模組中的非線性激活的這種架構又被稱為 Macaron-style 結構。

2.2.3 Performer

Performer 是 2021 年由 Krzysztof Choromanski 等人所提出的一種 Transformer 的架構變體，其核心貢獻在於將傳統 Transformer 中的多頭自注意力 (Multi-Head Self-Attention, MHSA) 模組替換為具備線性時間與空間複雜度的核化注意力 (Kernelized Attention) 機制。傳統 Transformer 的 Self-Attention 計算複雜度為 $O(N^2)$，其中N為輸入序列的長度，這對於長序列資料（如蛋白質序列、圖分子表示或文字）造成了巨大的記憶體與計算負擔。而 Performer 所提出的 Fast Attention Via positive Orthogonal Random features (FAVOR+) 技術，則透過正交隨

機特徵映射 (Positive Random Features) 將原本不可拆解的 Softmax Attention 核 (Kernel) 進行近似，成功地將複雜度降為 $O(N)$ 。

具體而言，Performer 將傳統的 Attention 計算方式近似如下：

$$Attention\,(Q,K,V) \approx \phi(Q)\,(\phi(K)^T V) \tag{2.2}$$

其中 $\phi(\cdot)$ 表示特徵映射函數，將查詢 (Q) 與鍵 (K) 映射至高維隨機特徵空間。該映射保留了 Softmax 核的特性，且能有效分解注意力機制中的非線性成分，實現更高效的計算。

此外，Performer 允許透過調整特徵映射維度 (nb_features) 以控制近似精度與運算資源間的權衡。當 nb_features 趨近於每個 Attention Head 的維度時，Performer 的表示能力可逼近甚至匹敵標準 Transformer 的表現，而其計算效率則顯著提高。此特性特別適用於處理序列長度極長或模型需要高效擴展的任務場景。

2.2.4 Linformer

這是 Sinong Wang 等人在 2020 年所提出的方法。Linformer 透過對自注意力機制中的注意力矩陣進行低秩分解 (Low-rank Matrix Decomposition)，將計算複雜度從平方級降低到線性級。這種方法減少了計算資源的需求，使得 Transformer 能夠在資源受限的環境中高效運行。

2.2.5 Mamba

Performer 和 Linformer 都是基於 Transformer 改進的方法，但是這個由 Albert Gu 和 Tri Dao 在 2023 年底提出的全新方法卻是一種基於狀態空間模型 (SSM) 的序列建模架構。該模型透過讓 SSM 參數成為輸入的函數，替代 Transformer 的自注意力機制，以線性時間複雜度處理長序列數據，解決了離散模態下的不足。而 Mamba 也在語言、語音和基因組學等領域上達到了最先進的性能。

2.3 CPI 模型回顧

以下介紹近年一些採用 Transformer 架構的 CPI 預測模型，它們都在 CPI 預測任務上取得相當優異的成果。

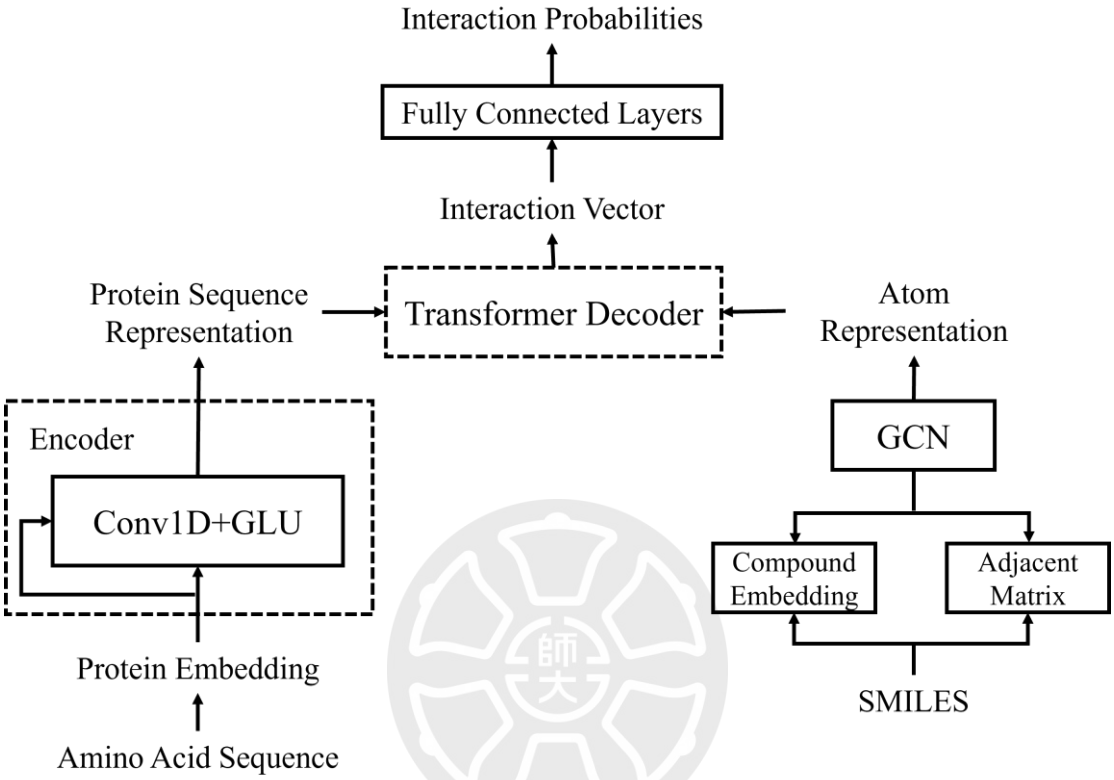


圖 8 TransformerCPI 的模型架構，引用自[15]

TransformerCPI [15] 使用 Transformer 的自注意力機制來學習 SMILES 序列中的語義關係。該模型簡化了 Transformer 的架構，包含將編碼器的部分替換成 Gated Convolutional Network (Gated CNN)，以及將解碼器的 mask operation 做修改以確保模型可用於整個序列。此外還設計了較嚴格的標籤反轉實驗 (label reversal experiment) 來測試模型是否學習到真正的交互特徵。圖 8 為其模型架構。

DISAE [19] 全名為蒸餾序列比對嵌入 (Distilled Sequence Alignment Embedding)，是一個透過將進化訊息納入未標記蛋白質序列的自監督式學習而設計出的蛋白質序列的表示方法。DISAE 可以利用所有蛋白質序列及其多序列比對 (MSA) 來捕獲蛋白質之間的功能關係，而無需了解它們的結構和功能。該研究將蛋白質利用 DISAE 做嵌入，使用 ALBERT [20] 進行預訓練後，再透過

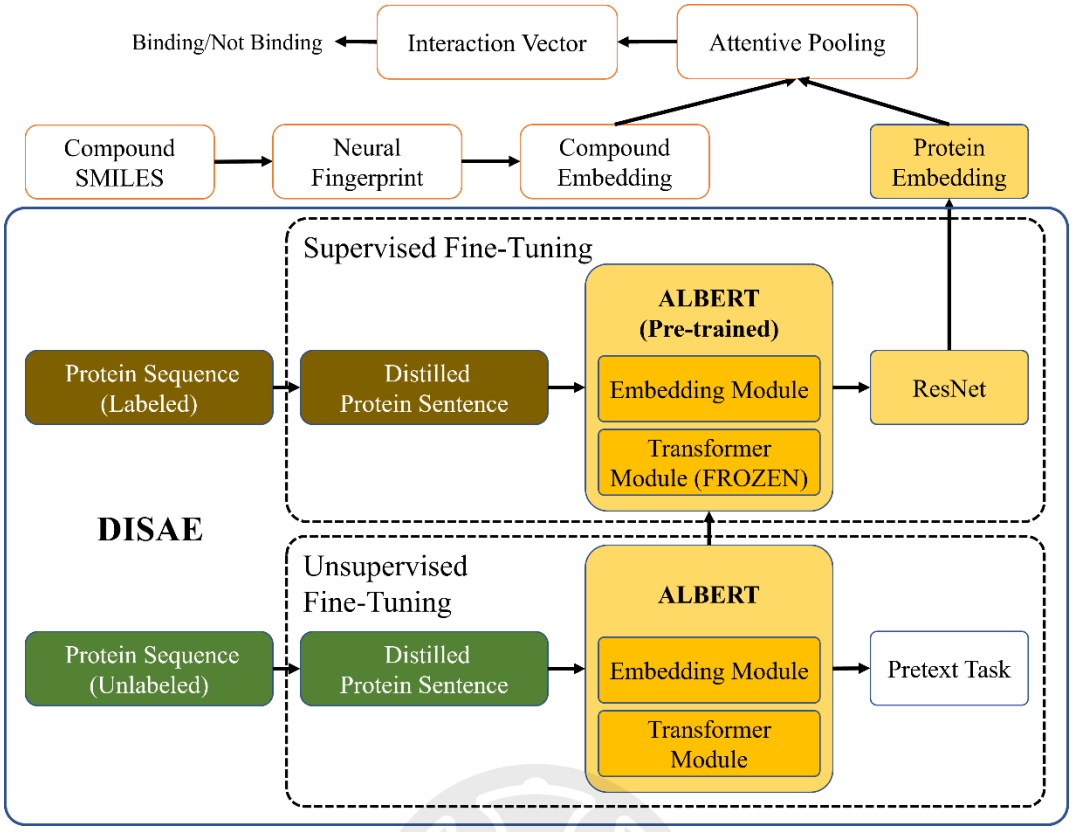


圖 9 MSA-Regularized Protein Sequence Transformer 的模型架構，引用自[19]

Module-Based Fine-Tuning 進行全基因組 CPI 預測 (Whole-Genome CPI Prediction)，以預測與進化分歧的未標註蛋白質的化學結合 (Chemical Binding) 。圖 9 為其模型架構。

CAT-CPI [21] 結合了 CNN 和 Transformer 來預測 CPI。它使用 CNN 學習化合物圖像的局部細節特徵，再使用 Transformer 的編碼器做語義學習 (Semantic Learning)。蛋白質的特徵則是使用 *k*-gram 的滑動窗口 (Sliding Window Division) 方法來學習。在分別取得化合物與蛋白質的特徵圖 (Feature Map) 後，最後使用 Feature Relearning 模組了解化合物和蛋白質特徵的相互作用特徵，藉此得到預測結果。圖 10 為其模型架構。

MCL-DTI [22] 利用了藥物的多模態資訊 (Multimodal Information)，分別是分子影像和化學特徵，以及靶點的 FASTA 作為輸入，以更全面的學習藥物和靶點的特徵。此外引入了雙向交叉注意力機制 (Bidirectional Cross-Attention Mechanism) 來提高 DTI 的性能，其中 Multi-head Self Attention 用來捕捉特徵本

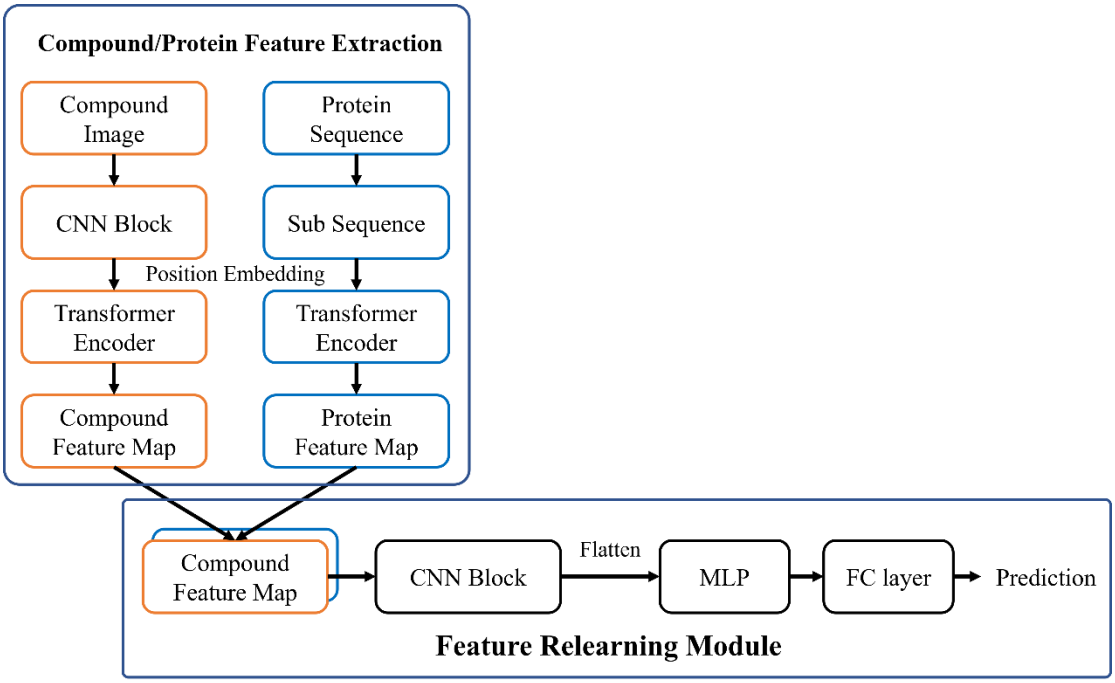


圖 10 CAT-CPI 的模型架構，引用自[21]

身的內部關係，而 Multi-head Cross Attention 則是用於捕捉藥物與靶點之間的相互作用訊息，兩者相輔相成。圖 11 為其模型架構。

MDL-CPI [23] 是一個 Multi-View 的深度學習方法。它分別使用 BERT-CNN 和 GNN 提取蛋白質和化合物的資訊並映射到低維特徵空間，並同時使用 AE2 (Autoencoder in Autoencoder Networks) 來獲取蛋白質和化合物高集中相關的特徵資訊，稱為統一特徵資訊。Autoencoder 是一種在編碼器和解碼器之間具有內部隱藏層的神經網路。透過 BERT-CNN、GNN 以及 AE2 的輸出連接起來作為最終的關聯表示，避免蛋白質和化合物之間重要的交互作用資訊遺失的問題。圖 12 為其模型架構。

PerceiverCPI [24] 將 CPI 視為回歸 (Regression) 問題，輸入的化合物以 Extended Connectivity Fingerprint (ECFP) 和分子圖表示，蛋白質則使用 Tasks Assessing Protein Embeddings (TAPE) 編碼，輸出為 Binding Score。模型的設計源自於 Perceiver IO [25] 和 Directed Message-Passing Neural Network (D-MPNN)，利用非對稱注意力機制提取化合物的資訊，然後用額外的交叉注意力 (Cross-Attention) 模組與蛋白質的資訊結合，藉此捕捉 CPI 的資訊。圖 13 為其模型架

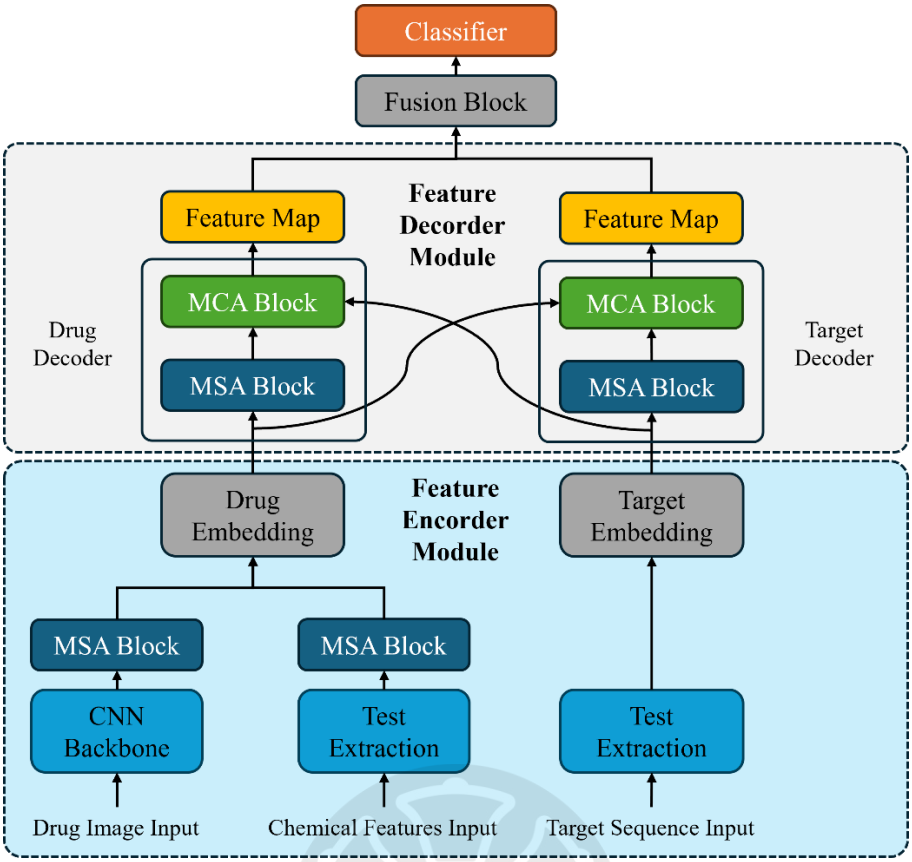


圖 12 MCL-DTI 的模型架構，引用自[22]

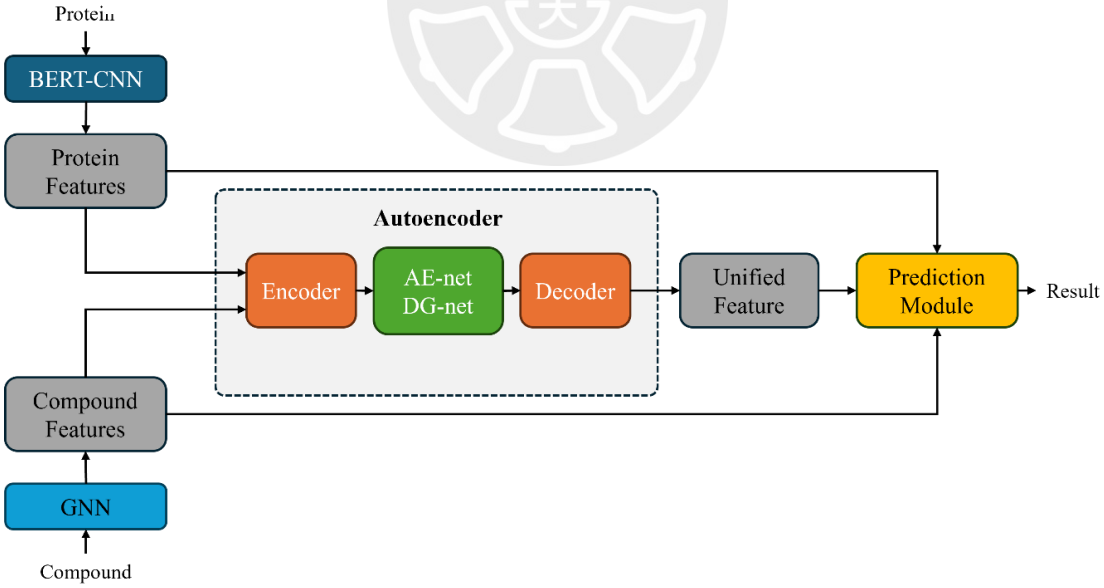


圖 11 MDL-CPI 的模型架構，引用自[23]

構。

CAT-DTI [26] 是一種結合 Cross-Attention 機制與 Transformer 架構的預測模型，旨在提升模型對於異質資料分佈的適應能力。該方法首先透過 GCN 與

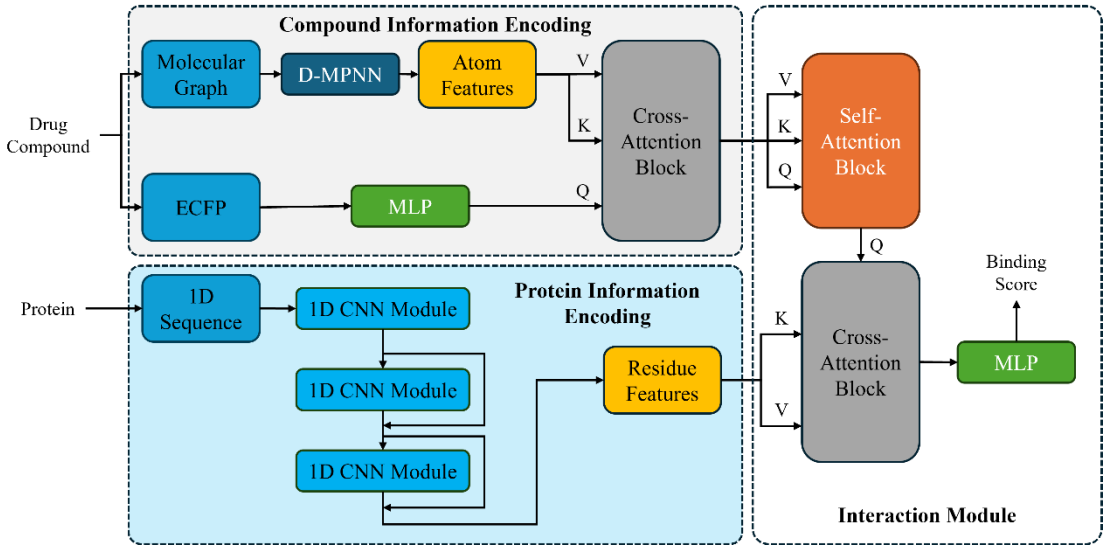


圖 13 PerceiverCPI 的模型架構，引用自[24]

CNN-Transformer 混合編碼器，分別提取藥物分子圖與蛋白質序列的結構與語意特徵。隨後，模型引入 Cross-Attention 模組進行藥物與蛋白質特徵融合，有效捕捉交互關係。為強化模型在跨領域預測任務中的泛化能力，CAT-DTI 採用 Conditional Domain Adversarial Network (CDAN) 進行特徵對齊。

SP-DTI [27] 是一種結合蛋白質結構資訊與 Transformer 架構的預測模型，特別引入 Subpocket 分析以提升交互建模的精細度。該方法利用 AlphaFold 預測蛋白質 3D 結構，並透過 Cavity Identification and Analysis Routine (CAVIAR) 演算法識別並分解 Binding Pocket 為多個 Subpockets。模型融合 ESM-2 與 ChemBERTa 預訓練嵌入，並以 Subpocket-Informed Transformer 建構藥物與蛋白質的交互表示。SP-DTI 在 Unseen Protein 與 Cross-Domain 測試中表現優異，展現其在結構導向預測上的強大泛化能力。

第三章 方法與步驟

3.1 模型架構簡介

本模型採用基於 Transformer [5] 的方法，以 CAT-CPI [21] 為基底，結合 TransformerCPI [15] 對於 SMILES 序列的編碼方式，並且針對模型架構做了改良。本模型的架構如圖 14 所示，主要包括以下幾個步驟：

1. 化合物特徵提取：首先使用 SMILES 序列生成由原子特徵和共價鍵組成的分子圖，以及 2D 的分子圖像。接著使用 GCN 對分子圖，CNN 和 Transformer Encoder 對分子圖像進行特徵提取，得到兩個固定大小的向量表示。
2. 蛋白質特徵提取：首先使用 *k*-gram 方法將蛋白質序列分割成子序列並編碼，再使用 Transformer Encoder 對編碼後的子序列進行特徵提取，得到一個固定大小的化合物與蛋白質向量表示。
3. 預測：將化合物向量、蛋白質向量整合成一個新的特徵向量，送進全連接層以預測化合物和蛋白質之間是否發生相互作用。

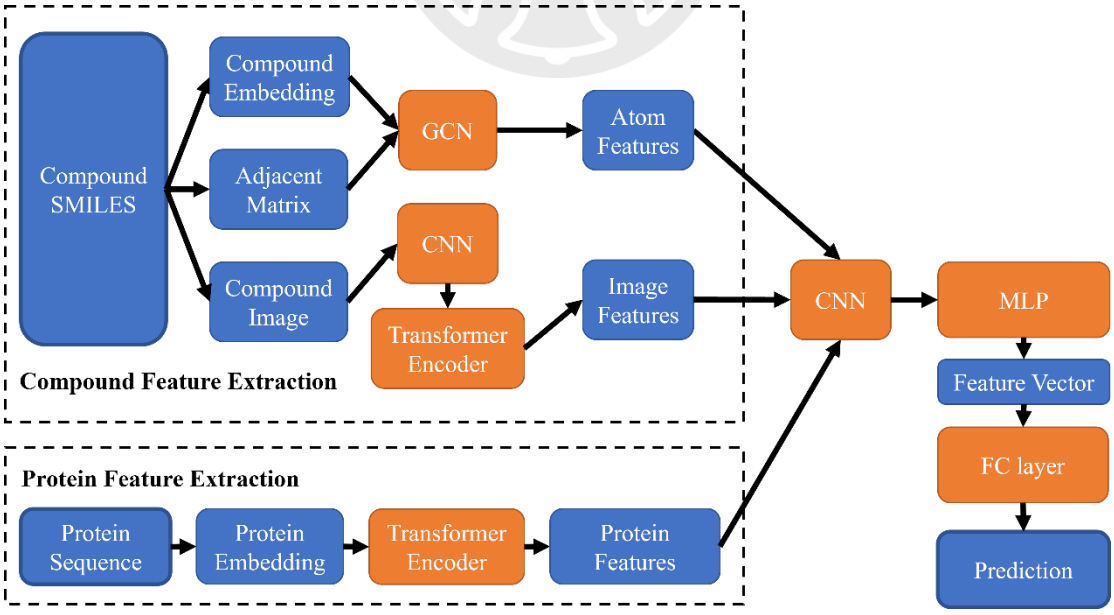


圖 14 本研究的 CPI 預測模型架構

3.1.1 化合物特徵提取

化合物的 SMILES 序列可以提供化學結構的序列訊息，例如它們所包含的原子、化學鍵和官能基等，這些訊息通常與化合物的生物活性和蛋白質交互性有關。我們以化合物的 SMILES 序列作為模型的輸入，並且利用 RDKit 的開源工具包將 SMILES 序列轉換為兩種資料，如圖 15 所示：一種是分子圖 (Molecular Graphs)，由原子特徵 (Atom Features) 和共價鍵 (Covalent Bonds) 組成，另一種則是 2D 的分子圖像 (Molecular Images)。其中，原子特徵是一個大小為 34 的向量，包含原子類型、原子度數、形式電荷、自由基電子數、雜化類型、是否為芳香性原子、連接的氫原子數、手性以及構型等資訊。我們在這裡將化合物的分子圖表示為 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$，其中 $V \in \mathbb{R}^{a \times f}$ 是分子中的a原子集，每個原子表示為 f維特徵向量，$\mathcal{E}$是分子中的共價鍵集，以鄰接矩陣 (Adjacency Matrix) $A \in \mathbb{R}^{a \times a}$表示。另一方面，化合物的圖像可以提供圖像特徵，例如圖像的形狀、顏色和紋理等，這些特徵通常與化合物的物理、化學特性有關，例如極性、溶解度和分子量等。

將這兩種不同的化合物資料格式結合起來，透過多模態學習 (Multimodal Learning) 來進行 CPI 預測，能夠同時利用圖結構與影像特徵，進一步提升模型表現。具體來說，將 Graph 所提供結構訊息和 Image 所提供空間與視覺訊息分別提取出來，然後進行特徵融合，可以得到更全面和準確的化合物表示。這樣的化合物表示可以更好地反映化學分子的結構和特性，從而提高模型的預測準確性。在分子圖的部分，我們使用 GCN，透過整合其相鄰原子特徵來學習每個原子的表示。在分子圖像的部分，我們採用和 CAT-CPI 一樣的 CNN Block 架構來初步提取特徵。這個 CNN Block 是由 Convolution Layer, Batch Normalization Layer, Activation Layer 和 Pooling Layer 組成的。接著使用一般的 Transformer 編碼器（架構已在章節 2.2 說明）進一步提取特徵。最終，我們得到化合物的特徵圖 (Feature Map)。

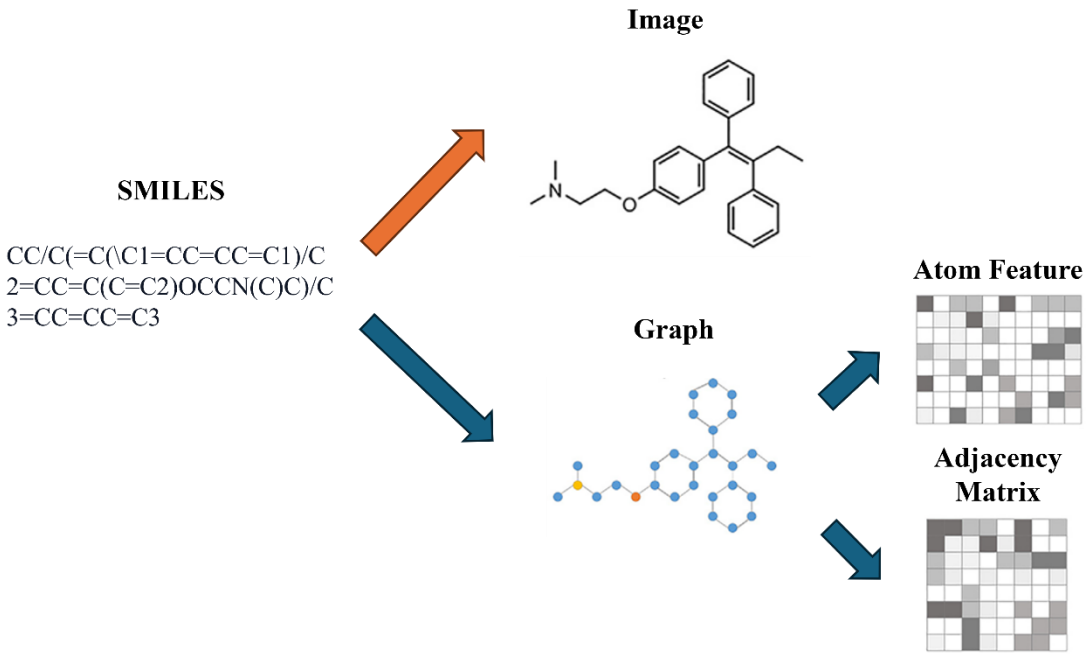


圖 15 本研究使用的化合物資料特徵示意圖。我們利用 RDKit 將 SMILE 序列轉換成 Image 和 Graph 後進行多模態學習

3.1.2 蛋白質特徵提取

由於氨基酸種類較少，蛋白質的表示方式較為簡單，導致深度學習模型難以有效學習其特徵。為了解決這個問題，我們採用了 *k*-gram 分割方法，透過滑動窗口將蛋白質序列劃分為 k 個氨基酸為一組的子序列，並以此建立蛋白質子序列的語料庫（Corpus），作為後續語意建模之用。在建立語料庫後，蛋白質序列可以透過語料庫中的編號進行編碼。這些子序列依據氨基酸類別數量進行嵌入，形成數值表示。然後根據蛋白質的長度特徵，從編碼後的子序列中抽取 N 個字符串，得到該蛋白質的最終表示。接著就可以使用與化合物特徵提取時相同的 Transformer 編碼器來處理這些子序列，最後得到蛋白質的特徵圖。

3.1.3 預測

有了編碼過的化合物與蛋白質的特徵圖後，接著利用多層感知器 (Multilayer Perceptron, MLP) 透過多層神經元的計算得到一個新的向量。這種方法的好處是可以保留化合物向量和蛋白質向量的所有訊息，同時也可以解決維度災難的問題。

然後再使用 CNN 提取一次特徵，因為 CNN 有非常強大的特徵聚合能力，透過對化合物和蛋白質的堆疊特徵圖進行卷積運算，可以有效地提取兩者的相互作用。為了取得最後的分類結果，我們在模型的最後一層加入一個全連接層 (Fully Connected Layer) ，以融合後的特徵向量作為輸入，輸出兩個對應於正負樣本的 Logits，並透過 softmax 函數轉換為機率分佈，表示化合物與蛋白質之間發生或不發生作用的預測機率，其中機率較高者所對應的類別即為預測結果。我們使用 Adam Optimizer 來調整與優化模型的超參數。

3.2 Transformer 架構的實踐細節

為了評估不同 Transformer 架構在 CPI 預測任務中的效果，本研究實作並比較了四種注意力機制架構，分別為 CAT-CPI、Performer、Conformer 以及 PerformerConformer。這些架構皆建立於 Transformer Block 的變形基礎上，並整合不同來源的輸入特徵，包括 Compound Image、Molecular Graph（透過 GCN 提取原子特徵）與 Protein Sequence。

上述各種架構均套用相同的輸入設定與融合模組，包括 Compound Image 通過 CNN、Protein Sequence 通過 Embedding，原子特徵通過多層 GCN 取得後與其他通道特徵融合。所有模型在訓練與測試階段皆維持一致的超參數設計，確保公平比較。

3.2.1 CAT-CPI 之多模態版本

原始的 CAT-CPI 只處理來自 Compound Image 與 Protein Sequence 的特徵。其模型架構包括使用 CNN 模組擷取 Compound Image 特徵，以及使用 embedding + Transformer 處理蛋白質序列，再將兩者提取出的多模態特徵進行融合與預測。在本研究中，我們保留 CAT-CPI 的主要結構作為 Baseline，並進一步擴充其輸入來源，加入第三個模態：Molecular Graph（分子圖）資訊，以增強模型對原子層級結構的理解。

有關分子圖的處理方式，我們採用 TransformerCPI 的方法，將化合物的 SMILES 描述轉換為圖結構，節點為原子、邊為共價鍵，並使用 GCN (Graph Convolutional Network) 提取原子層級的結構特徵。為與 CAT 原始架構對齊，我們對 GCN 的輸出特徵進行 Reshape 與插值處理，使其轉換為類似影像的張量格式（大小為 B×1×256×256），再與 Compound image 和 Protein Sequence 所得到的表示在通道維度上進行張量拼接 (Tensor Concatenation) 。拼接後的三模態融合特徵經過共用的 CNN 模組與 1D 卷積進一步壓縮，最終輸入全連接層進行交互預測。

為保持一致性，本研究沿用 CAT 架構的整體流程，並統一設定 Transformer Block 的深度與維度（例如 Embedding Dimension = 256，Block 數量 = 4），作為 Baseline 架構。此外，我們保留 CAT 原始的 Transformer Block 作為特徵擴展模組，並在必要時替換為其他注意力架構（如 Performer 或 Conformer）以進行比較。所有模態在輸入模型前皆進行對齊處理，並統一設定 Embedding Dimension、Block 數量與 MLP 比例，以確保實驗條件的一致性與公平性。

3.2.2 Performer

本研究使用 performer-pytorch 套件替換 CAT-CPI 的注意力機制來實現 Performer 架構。Performer 透過 FAVOR+ 技術將原始 softmax 注意力近似為兩步矩陣運算，如式(2.2)所示。這個近似過程的關鍵在於 nb_features 參數，即映射後的特徵維度，其大小會影響近似 softmax 的精確程度與計算資源需求。根據 performer-pytorch 的實作邏輯，若未明確指定，nb_features 預設為：

$$\text{nb\_features} = \lfloor 0.5 \times \text{dim\_head} \rfloor \tag{3.1}$$

其中 $\text{dim\_head} = \text{embedding dimension}/\text{num heads}$。以本研究的設定為例，當 $\text{embedding dimension} = 256$ 且 $\text{num heads} = 8$ 時，$\text{dim\_head} = 32$，因此預設 $\text{nb\_features} = 16$。

然而，根據我們的觀察，過小的 nb_features 會導致 Attention 表示過於粗

糙，進而降低模型表現。因此，本研究進一步將 nb_features 作為可調參數，嘗試不同設定（如 32、64 等）以分析其對預測效能的影響。

3.2.3 Conformer

如圖 7 所示，原始的 Conformer block 包含四個主要模組，依序為：FeedForward (FFN) → Multi-Head Self-Attention (MHSA) → Convolution Module (Conv) → FeedForward。每個子模組皆包含 LayerNorm、殘差連接與 dropout 操作。然而在本研究的實作上，為了確保與其他 Block 類型的一致性與模組統一，我們的 Conformer Block 採用的是 Self-Attention 加上輕量卷積模組組合的簡化架構，未包含 Macaron 結構與 GLU。卷積模組使用逐通道 Depthwise Convolution、GELU 激活與 BatchNorm 組成，有效補強局部依賴建模能力。

3.2.4 Performer-Conformer

為結合 Performer 的計算效率與 Conformer 的結構建模能力，本研究提出一個新架構：Performer-Conformer，也就是將 Conformer 中的 Self-Attention 模組替換為 Performer attention，保留其他結構模組不變。我們使用與 Conformer 相同的卷積模組，並將 nb_features 設計為可調參數來確保 Performer Attention 的有效性。

3.3 損失函數

我們將 CPI 預測的任務視為一個二元分類的問題。對於這樣的問題，交叉熵損失 (Cross-Entropy Loss) [10] 是最常見的目標函數。然而在實際應用中，我們發現使用交叉熵損失可能會帶來以下問題：

- **類別不平衡**

在某些 CPI 數據集中，正樣本（確定有交互作用的化合物-蛋白質對）與負樣本（無交互作用的化合物-蛋白質對）可能會有極度不均衡的情況。由於交叉熵損失假設所有類別的重要性相等，這會導致模型更傾向於預測多數類別，從而影響

少數類別的識別能力。

- **難分類樣本的影響力不足**

在 CPI 領域，某些蛋白質-化合物對的交互模式較為複雜，難以分類。交叉熵對這些「難分類樣本」的關注度不足，可能會導致模型過度關注易分類樣本，而忽略關鍵的複雜模式。

為了解決這些問題，我們研究並實驗了多種改進的損失函數，包括傳統的 Cross-Entropy Loss、針對類別不平衡問題而設計的 Focal Loss [11]、Asymmetric Loss [12] 以及損失函數混合的策略，期望能提升模型在類別不平衡情況下的表現。

3.3.1 各種損失函數介紹

以下介紹部分適用於二元分類 CPI 任務的損失函數：

- **Cross-Entropy Loss (Claude Shannon, 1948)**

這是分類問題最常用的損失函數，對於所有樣本給予相同的關注度，學習速度快且應用廣泛，但是也可能導致類別不平衡的問題。Cross-Entropy Loss 的公式如下所示。

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{\mathrm{i}} y_i \log \hat{y}_i \tag{3.2}$$

其中，y_i是真實標籤，$\hat{y}_i$是模型預測的機率分佈。

- **Focal Loss (Lin et al., 2017)**

為了解決類別不平衡問題，Focal Loss 透過調整易分類樣本的權重，使模型更關注難分類樣本，提高少數類別的學習效果。Focal Loss 的公式如下所示。

$$\mathcal{L}_{\mathrm{FL}} = -\sum_{\mathrm{i}} \alpha(1-\hat{y}_i)^{\gamma} y_i \log \hat{y}_i \tag{3.3}$$

其中，α是調整正負樣本權重的超參數，γ是聚焦參數。當$\gamma > 0$時，模型會減少對易分類樣本的損失權重，增加對難分類樣本的學習關注度。

- **Asymmetric Loss (Ridnik et al., 2021)**

針對極端類別不平衡的情況，Asymmetric Loss 透過不同的懲罰機制來降低負樣本的影響，並強化對正樣本的學習，使模型在不同錯誤類型上的控制更為精細。Asymmetric Loss 的公式如下所示。

$$\mathcal{L}_{\text{ASL}} = \begin{cases} -(1-\hat{y}_i)^{\gamma^+} y_i \log \hat{y}_i, & y_i = 1 \\ -(\hat{y}_i)^{\gamma^-} (1-y_i) \log(1-\hat{y}_i), & y_i = 0 \end{cases} \tag{3.4}$$

其中，γ^+和γ^-分別控制正樣本與負樣本的損失權重。

3.3.2 混合損失函數

雖然相較於 Cross-Entropy Loss，Focal Loss 和 Asymmetric Loss 理論上對於不平衡的資料會有較好的學習效果，但如果是相對平衡的資料，使用後者則可能會降低整體的準確率。且更改損失函數也會影響到收斂的速度。為了同時利用 Cross-Entropy Loss 的穩定性與 Focal Loss 的強化學習能力，我們提出一個混合損失函數（Hybrid Loss）的方法，透過權重分配的方式結合兩者，提升模型對不平衡數據與難分類樣本的適應能力。以 Cross-Entropy Loss 和 Focal Loss 的混合為例，我們可以用以下的式子來表示：

$$\mathcal{L}_{\text{Hybrid}} = \lambda \mathcal{L}_{\text{CE}} + (1-\lambda)\mathcal{L}_{\text{FL}} \tag{3.5}$$

其中λ是超參數，控制 Cross-Entropy Loss 和 Focal Loss 之間的權重平衡。λ 會隨著訓練時的 epoch 由大變小，因此一開始 Cross-Entropy Loss 的影響力最大，但是會隨著時間越來越小，Focal Loss 則與之相反。另外也可以把 Focal Loss 替換成 Asymmetric Loss，也就是把式(3.5)的$\mathcal{L}_{\text{FL}}$更改成$\mathcal{L}_{\text{ASL}}$。透過兩種方式的結合，我們期望模型能夠兼具泛化能力以及對難分類樣本的強化學習能力。

第四章 實驗與結果

4.1 實驗資料

在 CPI 預測模型的評估實驗中，通常會使用已經公開發布的 CPI 資料集。這些資料集的每一筆資料都是由一個化合物、一個蛋白質以及它們之間是否有交互作用的分類標籤（0 或 1）來表示。

在本研究中，我們使用 Human [28], Celegans [28], 以及 Davis [29] 三種資料集來進行實驗。其中，Human 和 Celegans 所包含的化合物和蛋白質的數量相近，且正負樣本的比例趨近於 1:1，是屬於平衡的資料集。Davis 所涵蓋的蛋白質數量遠多於化合物，且負樣本數遠高於正樣本數，因此是極度不平衡的資料集。我們比照 CAT-CPI 的處理方式，將每個資料集依比例分割成訓練集、驗證集和測試集。三種資料集的組成如表 1 所示。

表 1 本研究所使用的資料集之組成

Dataset	Compounds	Proteins	Samples	Positive Samples
Human	1,709	2,043	6,212	3,364
Celegans	1,723	1,708	7,511	3,893
Davis	68	379	11,103	1,506

4.2 t-SNE 視覺化分析

為了針對模型所提取的特徵進行視覺化分析，本研究採用 *t*-distributed Stochastic Neighbor Embedding（*t*-SNE） [30] 作為降維方法。*t*-SNE 是一種非線性降維技術，能夠將高維度資料映射至二維或三維空間，並在低維空間中保留原始資料的局部結構。其主要原理是透過比較高維空間與低維空間中樣本間的相似度分布，使相似樣本在視覺化結果中聚集、而不相似者分離，進而揭示潛在的分群結構。由於本研究所處理的特徵嵌入維度較高，*t*-SNE 能有效協助觀察正負樣本間在嵌

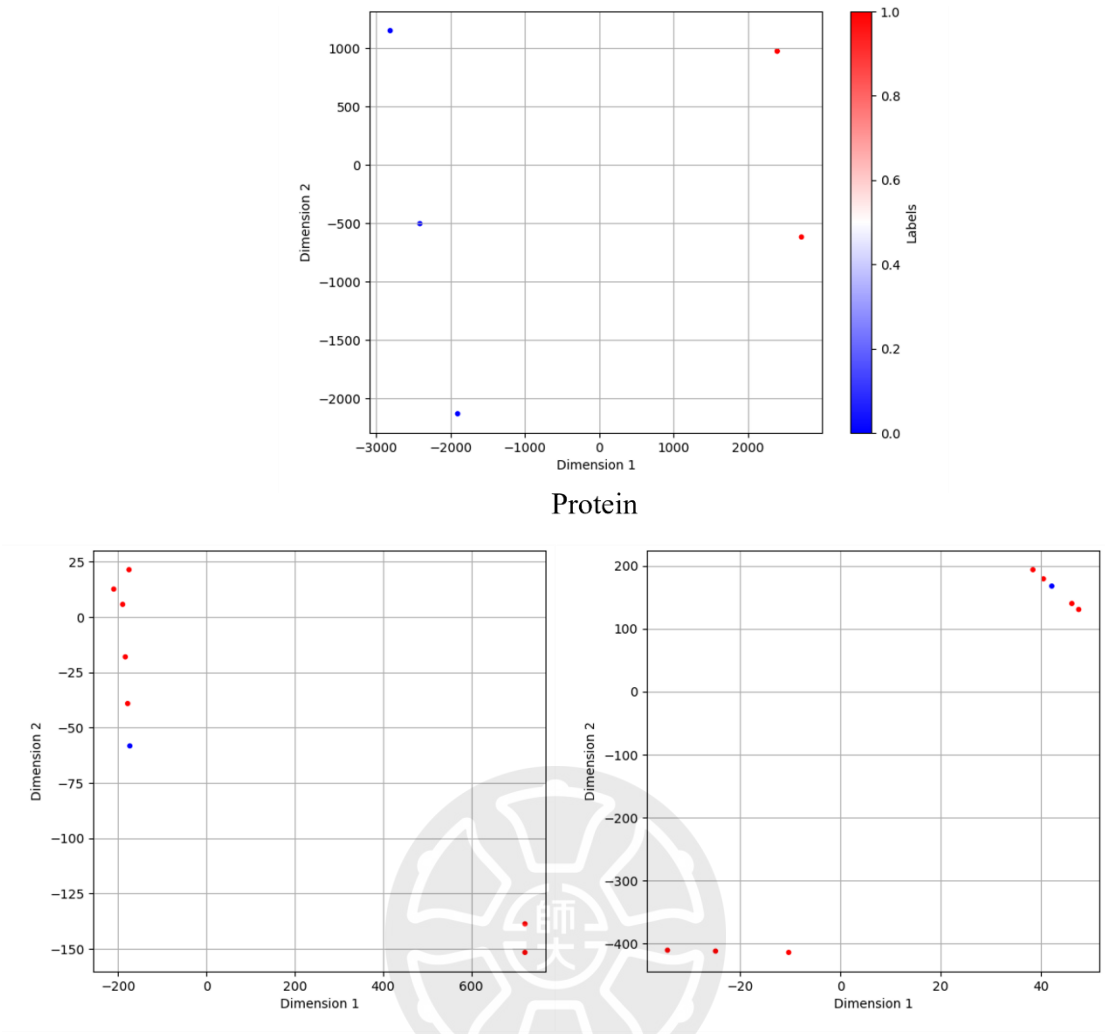


圖 16 Human 資料集的 *t*-SNE 視覺化。我們挑選出 6 個相同化合物的樣本以及 9 個相同蛋白質的樣本進行視覺化分析

入空間中的分布差異，作為判斷模態表示能力的輔助依據。

為了初步檢視模型所提取的多模態特徵是否具有區分CPI正負樣本的能力，我們以 *t*-SNE 對三種特徵（Compound、Adjacency、Protein）在三個資料集（Celegans、Human、Davis）上進行視覺化。其中，Compound 和 Adjacency 對應到分子圖 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$。

在 Human 和 Celegans 資料集中（圖 15 和 16），Compound 特徵嵌入顯示出最明顯的分群現象，正負樣本分布在圖中呈現出相對分離的區域，特別是 Human 資料的 Compound 嵌入，紅藍點聚落清晰、邊界分明，說明以 SMILES 序列為

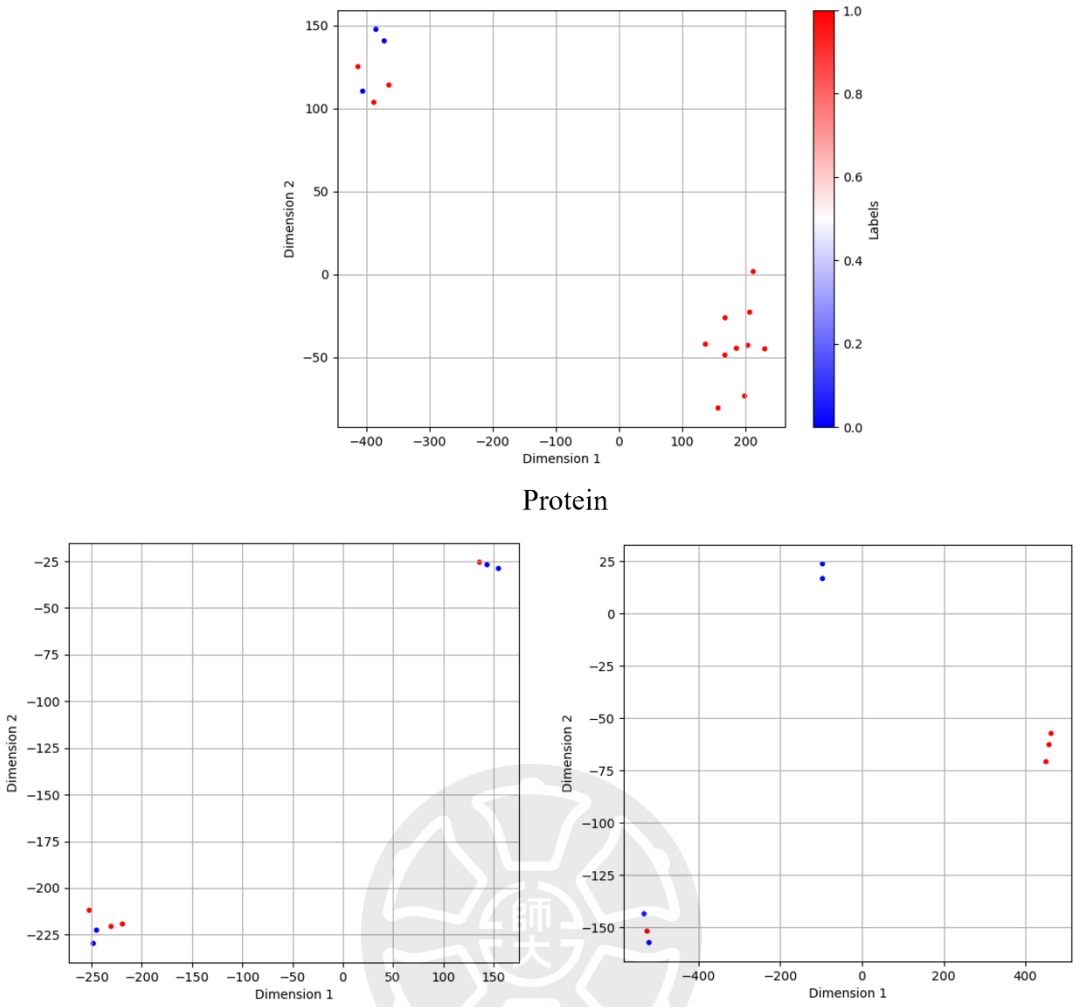
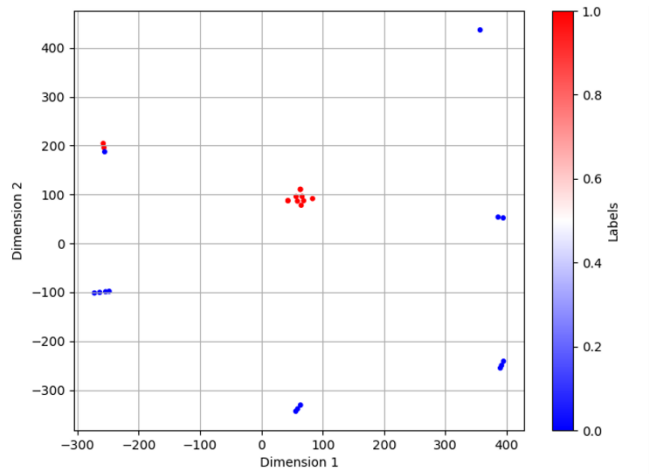


圖 17 Celegans 資料集的 *t*-SNE 視覺化。我們挑選出 16 個相同化合物的樣本以及 8 個相同蛋白質的樣本進行視覺化分析

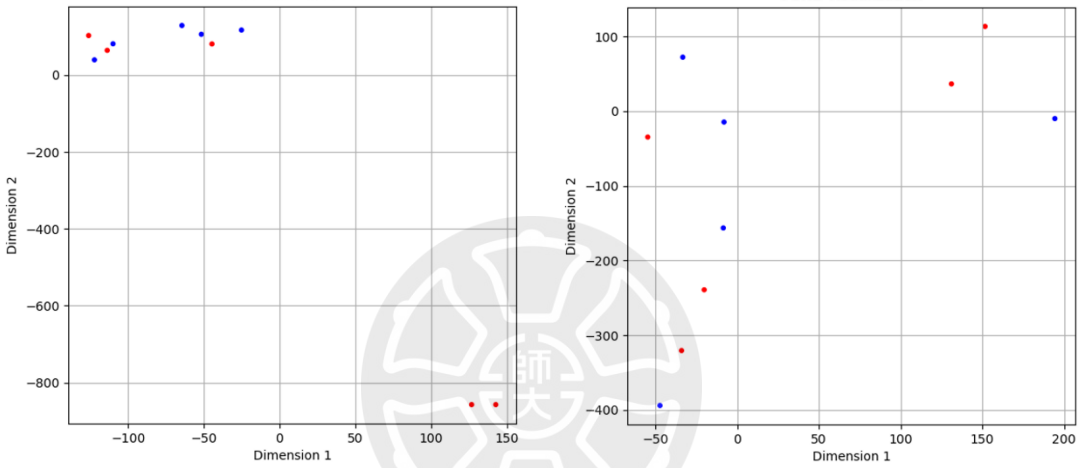
基礎的原子特徵具備良好的判別能力。Protein 特徵則表現次之，雖然群聚程度不如 Compound，但在 Human 資料中也能看出某些區域具備正負樣本的分離趨勢。Adjacency 特徵則在各資料集中分群效果相對較弱，紅藍點分布混雜，未呈現明顯的交互作用判別性。

至於 Davis 資料集（圖 17），由於負樣本數遠多於正樣本，各模態的分群現象相對模糊。Compound 嵌入仍保有部分聚落，但紅藍點界線不明確；Protein 和 Adjacency 則大多呈混合分布，可見資料不平衡對特徵學習造成顯著影響。

整體而言，本次視覺化結果指出：Compound 特徵在三個資料集中皆具備較強的表徵能力與可分性，而 Protein 特徵則視資料品質而定，表現有所差異；純



Protein



Adjacency　　Compound

圖 18 Davis 資料集的 *t*-SNE 視覺化。我們挑選出 26 個相同化合物的樣本以及 11 個相同蛋白質的樣本進行視覺化分析

粹的共價鍵資訊（Adjacency）在表示交互作用方面效果有限，未能清楚區分正負樣本。

4.3 評估方法

對於 CPI 預測模型的評估，除了損失函數的 Loss 值以外，通常會使用一些評估指標來衡量其性能。在本實驗中，我們使用以下的評估指標：

1. AUC (Area Under the ROC Curve) ：ROC 曲線下的面積，可以用來評估模型的分類能力。
2. Precision/Recall：Precision 是指預測為正樣本的樣本中，真正為正樣本的比

例；Recall 是指所有真正為正樣本的樣本中，被正確預測為正樣本的比例。

3. F1-score：綜合考慮 Precision 和 Recall，常用於評價二分類模型的性能。

4. Time：指的是從訓練開始到測試結束所花費的時間，單位為秒。

在實驗的設計部份，由於這是一個由多個面向進行方法改進的研究，我們會先針對各個面向分別進行修改，用前述的三個資料集分別訓練 20 個 epoch 後，與作為基準的 CAT-CPI 做性能的比較，並觀察各評估指標的變化。分別找出最好的改進方法後，再將其整合並評估最終版本的性能。以損失函數的改進為例，我們會將原本只用 Cross Entropy Loss 的方法與 Focal Loss、Asymmetric Loss，連同兩種損失函數混合的方法一起比較。我們也會探討相同方法，不同資料集對於各項評估指標的影響。

4.4 實驗結果

我們針對作為基底的 CAT-CPI 模型做了三個方面改進的實驗。在特徵方面，除了原有的分子圖像外，加上了由原子和共價鍵特徵組成的分子圖。在模型架構方面，我們將 Transformer 更改為 Performer 和 Conformer，以及將 Conformer 的 Self-Attention 替換為 Performer 的核化注意力的 Performer-Conformer。在損失函數方面，我們不僅嘗試單純使用 Focal Loss、Asymmetric Loss，也嘗試了混合的損失函數。模型皆以 Python 3.10.12 運行在 i9-9900KF CPU, 64GB RAM, NVIDIA TITAN RTX 的機器上。參數方面基本上與 CAT-CPI 的設定相同，例如 Learning Rate = 0.001、Batch Size = 128。以下針對三個面向的實驗結果進行分析。

4.4.1 多模態學習

為探討多模態特徵是否能有效提升 CPI 預測的準確性，本研究在原始 CAT 架構的基礎上，新增第三模態：Molecular Graph（分子圖）資訊，作為 Compound Image 的補充，並和 Protein Sequence 特徵共同輸入模型進行預測。我們首先在 CAT-CPI 架構下進行實驗，結果如表 2。

表 2 在 CAT-CPI 架構下，有無使用多模態學習的實驗結果，分別是只使用分子圖像 (CI) 以及同時使用分子圖像和分子圖 (CI+MG)

Methods	**Loss**	**ROC-AUC**	**PR-AUC**	**Precision**	**Recall**	**F1**	**Time**
Human							
CI	**0.192**	0.986	0.987	0.934	**0.955**	**0.944**	**187.175**
CI+MG	0.195	**0.988**	**0.989**	**0.937**	0.952	**0.944**	268.552
Celegans							
CI	0.157	0.991	0.993	0.962	0.952	0.957	**223.161**
CI+MG	**0.150**	**0.992**	**0.993**	**0.979**	**0.954**	**0.967**	371.709
Davis							
CI	0.562	0.912	0.442	0.214	**0.853**	0.343	**202.457**
CI+MG	**0.504**	**0.916**	**0.465**	**0.221**	0.842	**0.350**	293.519

實驗結果顯示，在 Human 與 Celegans 資料集上，加入 Molecular Graphs 模態（CI+MG）後，多數指標上略優於單一模態。以 Human 為例，ROC-AUC 從 0.986 提升至 0.988，PR-AUC 從 0.987 提升至 0.989，F1 分數維持 0.944，整體效能持平略升；Celegans 的改善更為明顯，F1 分數由 0.957 提升至 0.967，顯示 Atomic Features 在結構明確、樣本平衡的資料集中可有效補強 Compound Images 特徵，提升預測精度。

但在 Davis 資料集中，結果則出現不同趨勢。CI 模式下 ROC-AUC 為 0.912，而 CI+MG 模式略升至 0.916，PR-AUC 由 0.442 升至 0.465，F1 分數也由 0.343 微幅提升至 0.350。雖整體表現有小幅進步，但其訓練時間也大幅增加（由 202.457 秒升至 293.519 秒），整體成本效益需考量。

接下來，我們將相同的多模態設計應用於 Performer 架構，以檢驗此趨勢是否具有一致性。Performer 架構使用 FAVOR+ 技術將注意力計算從二次複雜度降為線性，有效減少長序列計算資源。實驗結果（表 4.3）顯示類似現象仍成立。

表 3 在 Performer 架構下，有無使用多模態學習的實驗結果，分別是只使用分子圖像 (CI) 以及同時使用分子圖像和分子圖 (CI+MG)

Methods	**Loss**	**ROC-AUC**	**PR-AUC**	**Precision**	**Recall**	**F1**	**Time**
Human							
CI	0.217	0.984	0.986	0.939	0.933	0.936	**190.913**
CI+MG	**0.199**	**0.986**	**0.988**	**0.939**	**0.942**	**0.941**	281.931
Celegans							
CI	0.198	0.988	0.991	0.959	0.939	0.949	**235.84**
CI+MG	**0.137**	**0.994**	**0.995**	**0.972**	**0.957**	**0.964**	384.737
Davis							
CI	0.577	**0.914**	**0.462**	0.204	**0.846**	0.329	**220.715**
CI+MG	**0.515**	0.913	0.458	**0.221**	0.842	**0.350**	297.292

在 Human 資料集上，ROC-AUC 提升至 0.986（CI 為 0.984）、PR-AUC 提升至 0.988、F1 分數由 0.936 上升至 0.941；Celegans 資料集中，CI+MG 模式的 F1 高達 0.964，為所有條件下最佳。而在 Davis 資料集中，CI 模式 F1 為 0.329，加入 MG 後上升至 0.350，ROC-AUC 為 0.914（CI）與 0.913（CI+MG），幾乎持平。

在樣本數與類別分佈相對均衡的資料集（如 Human 與 Celegans）中，Molecular Graphs 可補充 Compound Images 所未涵蓋之結構資訊，穩定提升模型性能。然而在如 Davis 這類樣本高度不平衡且圖結構較雜的情境中，MG 模態的效果則相對有限，即便在 Performer 架構中也難以產生顯著效益。未來可進一步透過模態注意力加權、特徵選擇機制或融合方式優化，以提升多模態學習的穩定性與適應性。

4.4.2 模型架構

在「模型架構」的實驗中，我們評估多種 Transformer 架構於 CPI 預測任務的表

表 4 以 Human 資料集進行模型架構實驗之結果

Methods	Loss	ROC-AUC	PR-AUC	Precision	Recall	F1	Time
CAT	**0.192**	**0.986**	**0.987**	0.934	**0.955**	0.944	187.175
Performer (nb=16)	0.207	0.985	**0.987**	0.949	0.952	**0.950**	192.917
Performer (nb=32)	0.203	0.985	**0.987**	**0.951**	0.936	0.943	**177.569**
Performer (nb=64)	0.196	0.985	**0.987**	0.942	0.942	0.942	188.406
Conformer	0.217	0.983	0.985	**0.951**	0.929	0.940	190.317
Performer-Conformer	0.224	0.980	0.981	0.939	0.946	0.942	200.211

現差異，包含原始 CAT-CPI 架構（CAT）、高效注意力版本 Performer（nb_features = 16、32、64）、強化結構建模能力的 Conformer，以及本研究設計之融合架構 Performer-Conformer（nb_features = 16）。為確保公平比較，所有模型均使用相同資料處理與訓練流程，並在三個資料集上進行評估：Human、Celegans 與 Davis，實驗結果如表 4.4～4.6 所示。

在 Human 資料集上，各架構整體表現皆非常接近，ROC-AUC 均落在 0.980–0.986 間，PR-AUC 亦在 0.981–0.989 間。Performer (nb_features=16) 取得最高 F1 分數（0.950），而 CAT 則擁有最高 Recall（0.955），顯示原始 CAT 架構在召回敏感任務中仍具優勢。Performer-Conformer 架構則展現穩定的整體性能，F1 值與其他架構相當，但訓練時間也最長。整體而言，在資料平衡且訊號明確的情境下，簡化架構如 CAT 或低 nb_features 的 Performer 已能達到極佳效能。

表 5 以 Celegans 資料集進行模型架構實驗之結果

Methods	Loss	ROC-AUC	PR-AUC	Precision	Recall	F1	Time
CAT	0.157	**0.991**	**0.993**	0.962	**0.952**	0.957	223.161
Performer (nb=16)	0.190	0.988	0.991	0.961	0.939	0.950	**214.805**
Performer (nb=32)	**0.155**	**0.991**	**0.993**	0.959	0.944	0.951	222.047
Performer (nb=64)	0.184	0.989	0.991	**0.974**	0.949	**0.961**	227.909
Conformer	0.186	0.990	0.991	0.971	0.942	0.956	233.411
Performer-Conformer	0.178	0.989	0.990	0.966	0.947	0.956	245.631

在 Celegans 資料集上，模型差異更為明顯。Performer (nb_features=64) 架構達成全體最高的 F1 分數(0.961)與 Precision(0.974)，表現優於 CAT(F1＝0.957)。Conformer 與 Performer-Conformer 亦表現良好，F1 分數皆達 0.956，顯示在結構穩定且樣本充分的資料中，增加卷積模組或提升 nb_features 皆能增強模型判別能力。不過，Conformer 類架構的訓練時間仍顯著增加。

Davis 資料集為極度不平衡且樣本複雜性較高的實例（正負樣本比例約 1:6），模型表現差異最為顯著。Performer (nb_features=16) 取得最高 ROC-AUC（0.919）與 F1 分數（0.352），顯示線性注意力對於處理長序列與複雜特徵有明顯優勢。若以 Recall 為觀察重點，Performer (nb_features=64) 的表現最為突出，Recall 達到 0.912，顯示此架構傾向召回更多正樣本。然而，其 Precision 與 F1 均較低（分別為 0.154 與 0.264），顯示過度偏向召回可能犧牲預測準確性。相對地，CAT 架構在 Recall 表現最差(0.853)，儘管在其他指標如 PR-AUC(0.442)與 F1（0.343）方面表現尚可，整體呈現預測保守但穩定的趨勢。Conformer 與

表 6 以 Davis 資料集進行模型架構實驗之結果

Methods	Loss	ROC-AUC	PR-AUC	Precision	Recall	F1	Time
CAT	0.562	0.912	0.442	0.214	0.853	0.343	202.457
Performer (nb=16)	**0.521**	**0.919**	**0.485**	**0.221**	0.863	**0.352**	**196.493**
Performer (nb=32)	0.581	0.915	0.475	0.205	0.863	0.331	203.696
Performer (nb=64)	0.858	0.911	0.454	0.154	**0.912**	0.264	204.944
Conformer	0.580	0.917	0.454	0.214	0.888	0.345	206.088
Performer-Conformer	0.664	0.916	0.477	0.192	0.891	0.316	223.256

Performer-Conformer 在 PR-AUC 上略有提升，但在 F1 分數方面無明顯優勢，顯示複合式架構在此資料集下仍需更細緻的調整以避免過度擬合。

綜合以上的實驗結果，在 Human 與 Celegans 資料集中，各模型皆達到高水準準確度，其中 Performer 在 nb_features = 16 或 64 時通常可取得略高的 F1 分數與 Precision，而 CAT 在 Human 中具有最高的 Recall。Conformer 類模型亦展現穩定性能，但需較長訓練時間。在 Davis 資料集中，Performer (nb_features = 16) 顯著優於其他架構，在 ROC-AUC 與 F1 分數皆為最高，適合處理不平衡且複雜資料；相對地，CAT 的 Recall 最低，顯示其保守預測策略在此場景效果有限。整體而言，模型效能受資料特性影響明顯，Performer 架構在維持效能與效率間具良好折衷。

值得留意的是，Performer-Conformer 結合架構在本研究中採用的設定預設為 nb_features = 16，與單獨使用的 Performer (nb_features =16) 架構相同。然而，兩者在效能上仍呈現差異：以 Human 資料集為例，雖然 Performer (nb_features

=16) 在 F1 分數上略高（0.950），但 Performer-Conformer 的結果更為穩定（F1 = 0.942），且在 Precision 與 Recall 之間達到良好平衡；在 Celegans 資料集中，兩者效能亦相近（F1 ≈ 0.956～0.961），顯示 Conformer 的加入可在不增加 nb_features 設定的情況下，補足 Performer 對局部結構捕捉的不足。

此外，單獨使用 Performer 模型時，nb_features 設定對效能有顯著影響：以 nb_features = 32 可在三個資料集中取得穩定且具泛化性的表現，尤其在 Celegans 中達到 F1 = 0.951，而 nb_features = 64 雖在 Celegans 表現最佳（F1 = 0.961），但在 Davis 資料中出現明顯下滑（F1 = 0.264），顯示高維度可能導致過擬合或模型不穩定。綜合而言，若採用單一架構，Performer (nb_features = 32) 為最具通用性之選擇；若可接受稍高之時間成本，Performer-Conformer（預設 nb_features = 16）則具備更穩定的泛化能力與結構適應性。

4.4.3 損失函數

為進一步提升 CPI 預測模型的分類能力與對不平衡樣本的適應性，本研究比較了四種損失函數在三個資料集中的表現，包括傳統的 Cross Entropy（CE）、針對少數類優化的 Focal Loss、強調不對稱懲罰的 Asymmetric Loss，以及兩者加權融合的 Hybrid Loss（CE + Focal）。實驗結果如表 4.7～4.9。請留意由於各損失函數的計算方式不同，不能直接拿 Loss 值的大小進行比較。

在 Human 資料集上，CE 損失函數表現整體最佳，達到最高的 ROC-AUC（0.986）、PR-AUC（0.987）與 F1 分數（0.944），其中 Recall 亦為四者最高（0.955），顯示標準交叉熵在樣本分佈相對平衡的情境下仍具良好表現。Hybrid Loss 雖在 F1（0.937）與 Recall（0.929）略低於 CE，但 Precision 為所有方法中為第二高（0.945），僅略差於 Focal Loss 的 0.947，顯示其在穩定性與召回間取得良好折衷。Focal 與 Asymmetric Loss 整體性能略遜，特別是 Asymmetric 在 Precision 上下降顯著（0.872）。

表 7 以 Human 資料集進行損失函數實驗之結果

Methods	Loss	ROC-AUC	PR-AUC	Precision	Recall	F1	Time
CE	0.192	**0.986**	**0.987**	0.934	0.955	**0.944**	187.175
Focal	0.032	0.981	0.982	**0.947**	0.917	0.932	**183.568**
Asymmetric	0.047	0.985	0.986	0.872	**0.984**	0.925	190.345
Hybrid	0.034	0.984	0.986	0.945	0.929	0.937	185.052

表 8 以 Celegans 資料集進行損失函數實驗之結果

Methods	Loss	ROC-AUC	PR-AUC	Precision	Recall	F1	Time
CE	0.157	**0.991**	**0.993**	0.962	0.952	0.957	**223.161**
Focal	0.022	0.988	0.990	**0.974**	0.949	**0.961**	228.864
Asymmetric	0.033	0.987	0.990	0.912	**0.975**	0.942	228.363
Hybrid	0.021	**0.991**	**0.993**	0.966	0.947	0.956	230.044

表 9 以 Davis 資料集進行損失函數實驗之結果

Methods	Loss	ROC-AUC	PR-AUC	Precision	Recall	F1	Time
CE	0.562	0.912	0.442	0.214	0.853	0.343	202.457
Focal	0.162	0.898	0.414	0.160	**0.891**	0.271	**201.029**
Asymmetric	0.043	0.878	0.419	**0.239**	0.747	**0.363**	202.801
Hybrid	0.123	**0.914**	**0.461**	0.214	0.87	0.343	203.104

在 Celegans 資料集中，各損失函數差異相對縮小。Hybrid Loss 與 CE 皆達成最高 ROC-AUC（0.991）與 PR-AUC（0.993），但 Hybrid 的 Precision 較高（0.966），Focal Loss 的 F1 則略高於其他方法（0.961）。Asymmetric Loss 的 Recall 雖最高（0.975），但 Precision 明顯偏低（0.912），反映出其在平衡性資料中過度懲罰負類可能導致分類界線失衡。

Davis 資料集則顯示出不同趨勢。在此極度不平衡資料中，CE 與 Hybrid Loss 取得相同的 F1 分數(0.343)，但 Hybrid Loss 在 PR-AUC(0.461)與 ROC-AUC（0.914）上皆略優於 CE（0.442 與 0.912），顯示其在兼顧召回與準確性方面具優勢。Focal Loss 在 Recall 上表現不錯（0.891），但 Precision 僅 0.16，導致 F1 僅為 0.271。Asymmetric Loss 在 Precision（0.239）與 F1（0.363）表現略優於 Focal，顯示在高度不平衡下調整懲罰策略確有助益。

綜合三組資料集的結果顯示，在樣本分佈平衡或接近平衡的情境下（如 Human、Celegans），Cross Entropy 表現穩定且全面，無需過度調整；而在不平衡資料情境下（如 Davis），Hybrid Loss 能夠提供更佳的整體分類效果，尤其在 AUC 與 Precision 指標上更具優勢。雖然 Focal 與 Asymmetric Loss 設計上針對難分類或類別不平衡情境，但實驗結果顯示它們未必能穩定提升所有指標，可能仍需與資料特性更精密匹配。未來可考慮自適應加權或動態調整損失策略，以進一步提升模型在不同資料條件下的表現穩定性與泛化能力。

4.4.4 整合與比較

綜合前面的實驗結果，我們整合出兩個改良後的 CPI 預測模型作為本研究最佳的模型候補：Performer 和 Performer-Conformer，並和原始的 CAT-CPI 做比較。三組模型的差異如下：

- CAT-CPI：原始論文的模型，沒有使用分子圖作為輸入，損失函數為 Cross-Entropy Loss。
- Performer：nb_features = 32，有使用分子圖，損失函數為 Hybrid Loss。
- Performer-Conformer：nb_features = 32，有使用分子圖，損失函數為 Hybrid Loss。

我們針對這三種架構進行了 50 個 epoch 的實驗，結果如圖 18～20。

我們可以清楚觀察到不同模型架構在 CPI 預測任務中的優劣與適應性差異。首先，Performer 在 Human 與 Celegans 資料集中皆取得最高的 F1 分數與 PR-

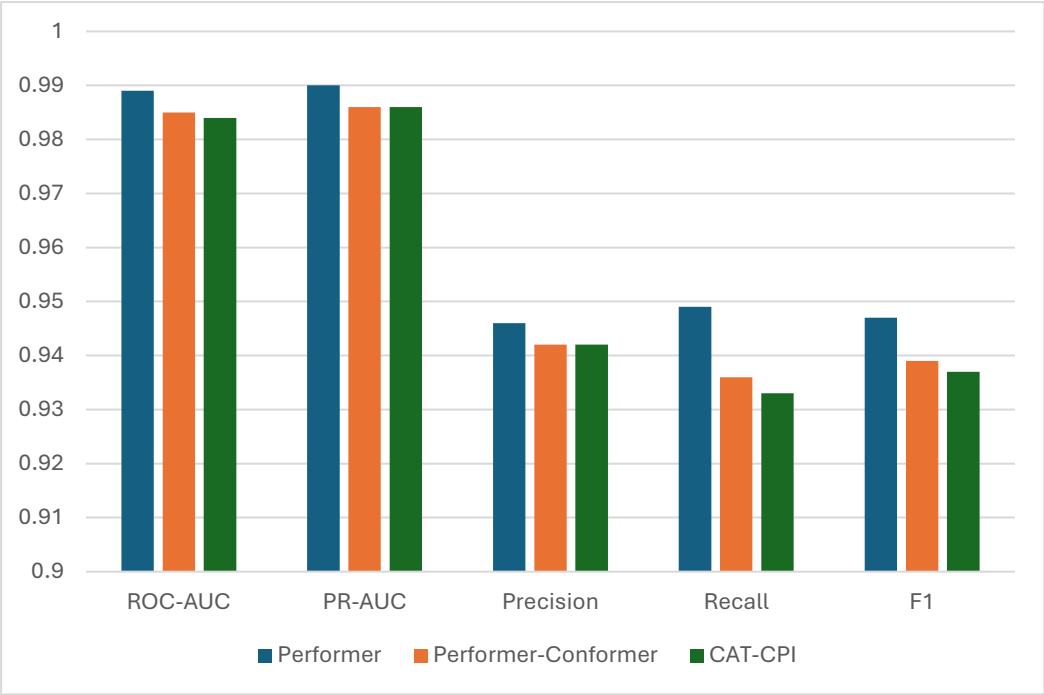


圖 19 以 Human 資料集進行三種不同架構的實驗結果

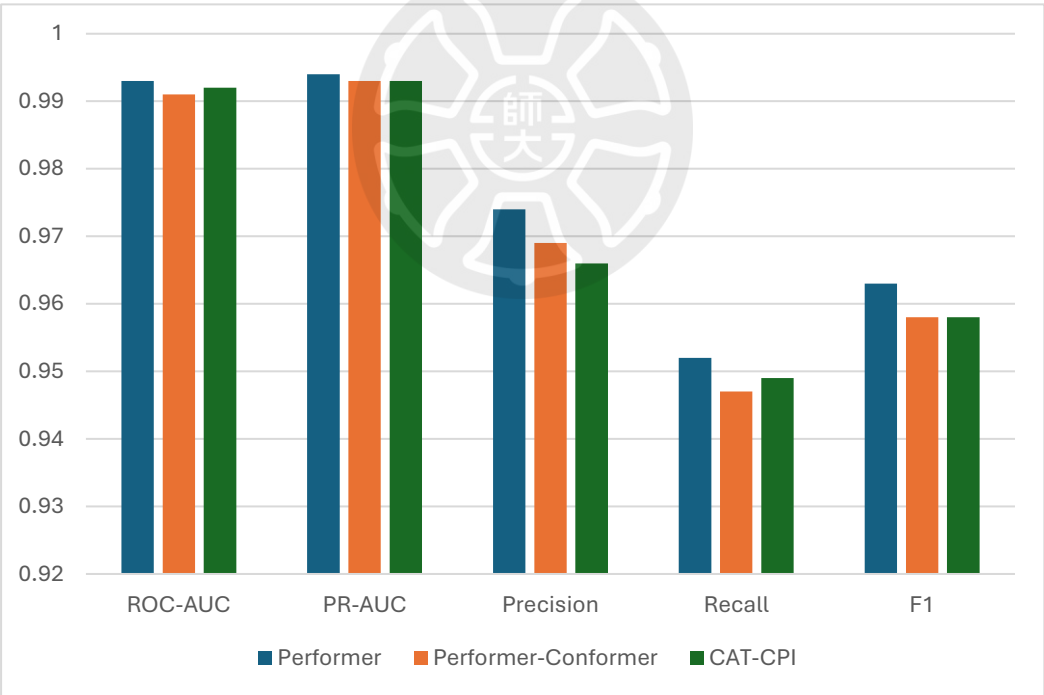


圖 20 以 Celegans 資料集進行三種不同架構的實驗結果

AUC，顯示其在樣本均衡或結構較清晰的資料中，具備極佳的表示力與分類能力。此架構能穩定地捕捉到多模態輸入中的關鍵資訊，特別是在分子圖像與原子特徵共同提供結構訊息的情境下表現尤為優異。

Performer-Conformer 則展現出更強的穩定性與泛化能力。雖然在 Human 與

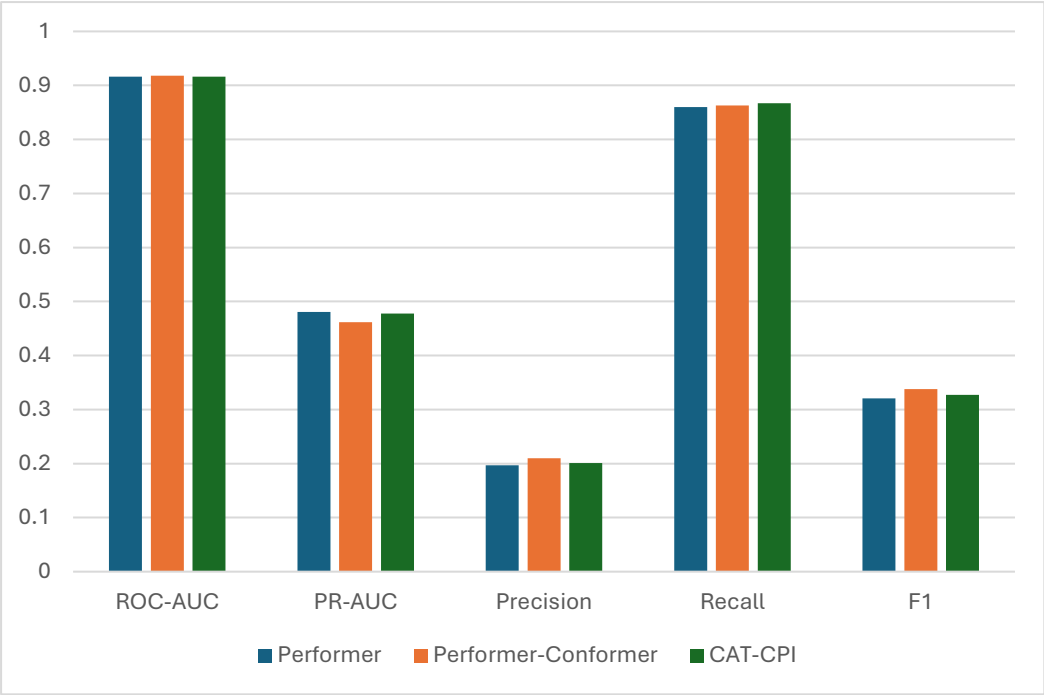


圖 21 以 Davis 資料集進行三種不同架構的實驗結果

Celegans 中的效能略低於 Performer，但差距極小；而在樣本不平衡且資料複雜度高的 Davis 資料集上，Performer-Conformer 的 F1 分數（0.338）高於 Performer（0.321）與 CAT-CPI（0.327），且 Recall 值達 0.863，顯示其在回收正樣本的能力上表現突出。這樣的結果反映出融合 Conformer 的架構能更有效捕捉蛋白質與化合物之間的局部結構關係，有助於提升模型在艱困預測場景下的表現。

至於 CAT-CPI 作為基準模型，雖然在三個資料集上皆展現穩定的預測能力，但缺乏分子圖像與多模態結構輔助，其效能普遍略低於改良架構。特別是在 Celegans 與 Davis 中，與其他模型的 F1 差距明顯，說明傳統單一模態的輸入在複雜交互作用建模中可能面臨性能瓶頸。

若以效能為首要考量，Performer 是本研究中最具預測能力的模型架構；而若需兼顧資料泛化與回收率，Performer-Conformer 更具彈性與穩定性。此分析結果提供了模型架構選擇的重要依據，也驗證了多模態輸入與融合式架構在 CPI 預測任務中的實用性與潛力。

第五章 結論與展望

本研究針對化合物－蛋白質交互作用（CPI）預測任務中常見的幾項挑戰——包括模態資訊利用不足、模型效能與效率難以兼顧，以及樣本不平衡導致的分類困難等問題——進行系統性分析與改進，並提出一套具彈性擴展性、可支援多模態輸入與多種模型架構的 CPI 預測框架。研究中以 CAT-CPI 為基礎，分別在輸入特徵設計、Transformer 架構優化與損失函數選擇三個面向進行實驗與比較。

在特徵層面，我們將原本僅使用分子圖像與蛋白質序列的 CAT-CPI 架構擴充為三模態輸入，加入由 SMILES 轉換而成的分子圖（透過 GCN 萃取原子與鍵結資訊），並以圖像形式重新映射後與其他模態共同輸入模型。實驗結果顯示，在 Human 與 Celegans 等資料集中，此類多模態融合可穩定提升 F1-score、Precision 與 AUC 指標，證明原子結構訊息能補充影像模態之不足。然而在如 Davis 這類樣本分布極度不平衡且結構較雜的資料中，多模態效果則較不穩定，顯示未來應進一步優化模態選擇與權重分配機制。

在模型架構方面，我們比較了原始 Transformer、Performer、Conformer 以及自設計的 Performer-Conformer 混合架構。實驗結果顯示，Performer 架構在多數資料集上展現優異的預測效能，能兼顧準確率與運算效率；Conformer 在結構表示上具潛力，尤其有助於捕捉蛋白質序列的局部特徵；而 Performer-Conformer 混合架構則整合兩者優勢，在樣本分布極端或特徵複雜的情境下表現穩定，具備良好的泛化能力。綜合分析指出，以 Performer 為基礎的架構能提供強勁的效能，而融合 Conformer 所建構的混合模型則更具適應性與彈性，可作為未來 CPI 預測模型優化的重要方向。

在損失函數實驗中，我們發現 Cross Entropy 在資料分布均衡的情況下仍為穩定選擇，而在 Davis 資料中，Hybrid Loss（CE＋Focal）在 AUC、Precision 與 F1 上表現最佳，顯示混合型損失有助於強化模型對少數類的辨識能力，適合用於真實世界中常見的極度不平衡場景。

整體而言，本研究所提出之多模態高效 CPI 預測框架，具備以下主要貢獻：一、融合分子圖、圖像與蛋白質資訊，有效提升特徵表達能力；二、導入可替換注意力模組與擴展架構，實現效能與資源消耗的折衷選擇；三、採用多種損失設計，提升模型在各類資料分佈下的泛化性與穩定性。

未來研究可延伸以下方向：一、引入 AlphaFold2/3 等結構預測工具所提供之三維結構或交互特徵作為額外輸入；二、結合注意力可視化與模態注意力機制，以理解不同輸入特徵對預測結果的貢獻程度；三、進一步擴充資料來源至非標準配體、RNA 或蛋白質複合體，推進 CPI 模型之應用範疇與生物解釋性。

參考文獻

[1] Meng, X. Y., Zhang, H. X., Mezei, M., & Cui, M. (2011). Molecular docking: A powerful approach for structure-based drug discovery. Current Computer-Aided Drug Design, 7(2), 146–157. https://doi.org/10.2174/157340911795677602

[2] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1609.02907

[3] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306. https://doi.org/10.1016/j.physd.2019.132306

[4] Koch, G. R. (2015). Siamese neural networks for one-shot image recognition.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. arXiv. https://doi.org/10.48550/arXiv.1706.03762

[6] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., ... & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. In Interspeech 2020. https://doi.org/10.48550/arXiv.2005.08100

[7] Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., ... & Weller, A. (2020). Rethinking attention with performers. arXiv. https://doi.org/10.48550/arXiv.2009.14794

[8] Wang, S., Li, B. Z., Khabsa, M., Fang, H., & Ma, H. (2020). Linformer: Self-attention with linear complexity. arXiv. https://doi.org/10.48550/arXiv.2006.04768

[9] Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv. https://doi.org/10.48550/arXiv.2312.00752

[10] de Boer, P.-T., Kroese, D. P., Mannor, S., & Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. Annals of Operations Research, 134(1), 19–67. https://doi.org/10.1007/s10479-005-5724-z

[11] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. arXiv. https://doi.org/10.48550/arXiv.1708.02002

[12] Ben-Baruch, E., Ridnik, T., Zamir, N., Noy, A., Friedman, I., Protter, M., & Zelnik-Manor, L. (2020). Asymmetric loss for multi-label classification. arXiv. https://doi.org/10.48550/arXiv.2009.14119

[13] Lim, S., Lu, Y., Cho, C. Y., Sung, I., Kim, J., Kim, Y., Park, S., & Kim, S. (2021). A review on compound-protein interaction prediction methods: Data, format, representation and model. Computational and Structural Biotechnology Journal, 19, 1541–1556. https://doi.org/10.1016/j.csbj.2021.03.004

[14] Goldberg, Y., & Levy, O. (2014). Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv. https://doi.org/10.48550/arXiv.1402.3722

[15] Chen, L., Tan, X., Wang, D., Zhong, F., Liu, X., Yang, T., Luo, X., Chen, K., Jiang, H., & Zheng, M. (2020). TransformerCPI: Improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics, 36(16), 4406–4414. https://doi.org/10.1093/bioinformatics/btaa524

[16] Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., ... & Kavukcuoglu, K. (2020). Improved protein structure prediction using potentials from deep learning. Nature, 577, 706–710. https://doi.org/10.1038/s41586-019-1923-7

[17] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold.

Nature, 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2

[18] Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Jumper, J., ... & Hassabis, D. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 630, 493–500. https://doi.org/10.1038/s41586-024-07487-w

[19] Cai, T., Lim, H., Abbu, K. A., Qiu, Y., Nussinov, R., & Xie, L. (2021). MSA-regularized protein sequence transformer toward predicting genome-wide chemical-protein interactions: Application to GPCRome deorphanization. Journal of Chemical Information and Modeling, 61(4), 1570–1582. https://doi.org/10.1021/acs.jcim.0c01285

[20] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for self-supervised learning of language representations. In International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1909.11942

[21] Ying, Q., Wu, J., & Zhang, Q. (2022). CAT-CPI: Combining CNN and transformer to learn compound image features for predicting compound-protein interactions. Frontiers in Molecular Biosciences, 9, 963912. https://doi.org/10.3389/fmolb.2022.963912

[22] Qian, Y., Li, X., Wu, J., Zhang, Q., & Ying, Q. (2023). MCL-DTI: Using drug multimodal information and bi-directional cross-attention learning method for predicting drug–target interaction. BMC Bioinformatics, 24, 323. https://doi.org/10.1186/s12859-023-05447-1

[23] Wei, L., Long, W., & Wei, L. (2022). MDL-CPI: Multi-view deep learning model for compound-protein interaction prediction. Methods, 204, 91–98. https://doi.org/10.1016/j.ymeth.2022.01.008

[24] Nguyen, N.-Q., Jang, G., Kim, H., & Kang, J. (2023). Perceiver CPI: A nested cross-attention network for compound–protein interaction prediction.

Bioinformatics, 39(1), btac731. https://doi.org/10.1093/bioinformatics/btac731

[25] Jaegle, A., Gimeno, F., Brock, A., Zisserman, A., Vinyals, O., & Carreira, J. (2021). Perceiver IO: A general architecture for structured inputs & outputs. In International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2107.14795

[26] Zeng, X., Chen, W., & Lei, B. (2024). CAT-DTI: Cross-attention and transformer network with domain adaptation for drug-target interaction prediction. BMC Bioinformatics, 25, 141. https://doi.org/10.1186/s12859-024-05753-2

[27] Liu, S., Liu, Y., Xu, H., Xia, J., & Li, S. Z. (2025). SP-DTI: Subpocket-informed transformer for drug–target interaction prediction. Bioinformatics, 41(3), btaf011. https://doi.org/10.1093/bioinformatics/btaf011

[28] Liu, H., Sun, J., Guan, J., Zheng, J., & Zhou, S. (2015). Improving compound–protein interaction prediction by building up highly credible negative samples. Bioinformatics, 31(Supplement_1), i221–i229. https://doi.org/10.1093/bioinformatics/btv256

[29] Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., ... & Zarrinkar, P. P. (2011). Comprehensive analysis of kinase inhibitor selectivity. Nature Biotechnology, 29(11), 1046–1051. https://doi.org/10.1038/nbt.1990

[30] van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(11), 2579–2605. http://www.jmlr.org/papers/v9/vandermaaten08a.html