Improving Protein-Protein Interaction Extraction via Intermediate-Task Transfer Learning with Pretrained Language Models
運用中介任務遷移學習於預訓練語言模型以改善蛋白質交互作用擷取

尹懌碩 Yi-Shuo Yin[1]
亞洲大學生物資訊與醫學工程學系
yin940222@gmail.com

林浚瑋 Chun-Wei Lin[1]
亞洲大學生物資訊與醫學工程學系
linjunwei76@gmail.com

陳鯨太 Ching-Tai Chen[2]
臺北市立大學資訊科學系
ctchen@utaipei.edu.tw

摘要

蛋白質交互作用辨識為生物醫學文本探勘中的關鍵任務，對藥物開發、癌症免疫及精準醫療皆具重要意義。本文實作遷移學習架構，並結合 LoRA (Low-Rank Adaptation) 微調技術與 Optuna 超參數自動化搜索，於 BioInfer、HPRD50、LLL 三個廣泛使用的 PPI 語料庫上進行實驗，結果顯示，BioBERT 學習效能全面優於 BERT；而多階層遷移學習架構整體表現優於單階層基準，三個資料集之最佳 F1-score 分別提升 3.8%、1.2%、6.8%，MCC 也分別提升 5.4%、0.8%及 14%。分析指出，中介任務資料集之特性與句法複雜度影響遷移成效，小型語料在多層遷移也展現較大型語料更高的準確力提升。錯誤分析歸納出三類主要錯誤型態，分別為否定與對比失敗、共存/pathway 誤判、以及調控性語言誤判，本研究除了驗證多層遷移的可行性，也為後續 PPI 辨識模型之改進提供參考方向。

Abstract

Protein-protein interaction (PPI) extraction is a critical task in biomedical text mining, with significant implications for drug development, cancer immunology, and precision medicine. This paper implements a transfer learning framework combined with LoRA (Low-Rank Adaptation) fine-tuning and automated hyperparameter search via Optuna, evaluated on three widely used PPI corpora: BioInfer, HPRD50, and LLL. Experimental results show that BioBERT consistently outperforms BERT across all settings, and that the multi-stage transfer learning architecture achieves better overall performance than single-stage baselines, yielding F1-score improvements of 3.8%, 1.2%, and 6.8% on the three datasets respectively, along with corresponding MCC improvements of 5.4%, 0.8%, and 14%. Further analysis reveals that the characteristics and syntactic complexity of the intermediate-task datasets influence transfer effectiveness, and that smaller corpora exhibit greater accuracy gains from multi-stage transfer than larger ones. Error analysis identifies three major categories of errors: failures in negation and contrast handling, misclassification of co-occurrence versus pathway relations, and misinterpretation of regulatory language. In addition to validating the feasibility of multi-stage transfer learning, this study provides directions for future improvements in PPI extraction models.

關鍵字：生醫文獻探勘、蛋白質交互作用、遷移學習、BioBERT、LoRA、中介任務

Keywords: biomedical text mining, protein-protein interaction, transfer learning, BioBERT, LoRA, intermediate task

1 Introduction

近年來，生物醫學研究領域中的書目資料數量持續大幅成長，這些文獻蘊含大量重要研究成果，對生醫領域具有重要參考價值。然而，面對浩如江海的資料量，研究人員往往需要投入大量人力與時間，才能從中擷取關鍵資訊。為提升生物醫學文獻探勘的效率，近年來自然語言處理 （Natural Language Processing,NLP）技術已廣泛應用於相關研究，並對自動化擷取

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 343–352
November 6-7, 2026 ©2026 Association for Computational Linguistics

生物實體及其相互作用資訊具有重要貢獻（Warikoo et al., 2018）。其中，蛋白質交互作用 (Protein–Protein Interaction, PPI) 的辨識，對於生醫研究、新藥開發、癌症免疫及精準治療等領域皆提供相當重要的資訊。

預訓練語言模型 (Pretrained Language Models, PLM) 在多種自然語言處理任務中展現出強大性能，並逐漸被應用於生物醫學文本探勘相關研究，顯示其在知識擷取與語意理解方面具有良好潛力 (Minaee et al., 2025)。在 PPI 辨識相關工作中，除了傳統以規則或統計方法為基礎的系統外，也已有研究提出以多粒度語意融合的方法來提升蛋白質交互作用關係抽取效能（Li et al., 2021）；此外，亦有研究探索結合語言學樣式與依存樹結構之方法，用以提升生物醫學關係抽取及 PPI 辨識效能（Warikoo et al., 2018）。在較早期的系統中，亦已有如 PIE (Protein Interaction information Extraction system) 之類的 PPI 自動預測工具，作為以傳統方法進行蛋白質交互作用辨識的代表性研究（Kim et al., 2008）。然而，這些方法在面對龐大且持續成長的生醫文獻時，往往仍需要相當成本的特徵設計與系統調整，顯示在有限資源條件下仍存在進一步優化的空間。

為因應上述挑戰，遷移學習 (Transfer Learning) 提供了一種有效的解決途徑。其核心概念在於將既有模型或資料集所學得的知識，轉移並應用於相關任務或目標資料上，以提升模型的學習效率與整體表現。相較於直接從頭訓練，遷移學習使模型能先在大規模語料中學得較具通用性的語言表示，再進一步適應特定任務的需求，因此，在標註資料有限的情況下，通常能透過重用預訓練模型所學得的通用表示，降低訓練成本並提升模型效能 (Phang et al., 2019)。然而，當目標任務的訓練資料規模較小時，若直接進行微調，模型仍可能因資料不足而出現訓練不穩定、過擬合，甚至災難性遺忘等問題。

先前研究表示在生物醫學關係抽取任務中，若結合句法資訊並採用適當的遷移學習策略，模型可在資料稀少的情境下獲得更佳表現，顯示來源任務與句法結構的選擇會影響知識轉移成效 (Legrand et al., 2021)。這種中介任務式遷移學習（intermediate-task transfer learning, ITTL）在多數情況下能有效提升目標任務表現，特別是在資料規模有限或任務之間具有一定相似性的情境下更為明顯（Ekakristi et al., 2025）。然而，中介任務的選擇並非必然帶來正向效果，不同任務間的複雜度、詞彙重疊與任務類型差異，皆可能影響知識轉移成效，因此仍需審慎設計與評估（Pruksachatkun et al., 2020），同時中介任務的效果往往取決於其與目標任務之間的語意層次與可轉移知識的相似程度 (Nkhata and Gauch, 2025)。

基於上述背景，本研究聚焦於將遷移學習應用於蛋白質交互作用辨識任務，並系統性探討不同遷移策略對模型效能之影響。透過比較多種訓練設定與模型適應方式，本文分析遷移學習對 PPI 辨識準確率之提升效果，進一步驗證其於生物醫學文本探勘中的應用潛力，並為後續相關研究提供可參考之方法與實驗基礎。

2 Related work

2.1 預訓練語言模型(Pretrained Language Models, PLM)

語言模型的發展始於 Transformer 架構的提出，不同模型依據其架構設計與訓練目標，發展出各具特色的預訓練典範。GPT 採用 Transformer 解碼器架構，以自迴歸方式進行單向語言模型預訓練，開啟了「生成式預訓練＋微調」的範式，後續 GPT-2、GPT-3 延續此架構並大幅擴增模型與資料規模，展現出強大的生成能力。BERT (Devlin et al., 2019) 則基於 Transformer 雙向編碼器架構，透過 Masked Language Model 與 Next Sentence Prediction 兩項任務進行預訓練，使模型能同時利用左右上下文資訊，因而在多項自然語言處理任務上取得突破性表現。T5 進一步將所有 NLP 任務統一轉換為「文字到文字 (text-to-text)」的形式，以單一 Encoder-Decoder 架構處理分類、翻譯、摘要等多種任務，展現出高度的任務彈性與統一性。然而，由於通用語料庫與生物醫學語料庫之詞彙分佈差異甚大，上述模型在處理專業醫學文本時性能受限，因此 BioBERT (Lee et al., 2020) 應運而生。BioBERT 延續 BERT 之雙向編碼器架構，並進一步以 PubMed 摘要與 PMC 全文文章進行持續預訓練，使其在生物醫學領域上顯著提升性能表現。

2.2 蛋白質交互作用擷取 (PPI Extraction)

PPI 擷取方法涵蓋統計與規則、機器學習、深度學習、大型語言模型等技術路徑，各方法

並存發展而非單純汰換。統計規則方法如 Kim et al.(2008) 提出的 PIE 系統整合共現統計與句型規則進行線上預測，Fundel et al.(2007) 之 Relation Extraction 則以依存句法樹之最短路徑為判斷依據，但均仰賴人工特徵設計、泛化能力有限。機器學習方法以核方法緩解此限制：Airola et al.(2008) 利用完整依存圖取代單一最短路徑，結合多種句法剖析器與核函數提升跨語料泛化；Warikoo et al.(2018) 之 LPTK (Linguistic Pattern-aware dependency Tree Kernel) 則結合語言學樣式與依存句法樹 Kernel，強化句法結構對關係抽取之貢獻。深度學習方法改以神經網路自動學習語意表徵：Peng and Lu.(2017) 之 McDepCNN (Multichannel Dependency-based Convolutional Neural Network) 結合詞向量與依存路徑資訊，於跨語料評估中優於核方法；Li et al.(2021) 則以 multigranularity semantic fusion 融合多層次語意特徵。近年研究則聚焦如何以提示工程 (prompt engineering) 引導通用 PLM 執行 PPI 擷取：Rehana et al.(2023) 顯示 GPT-4 未經生醫領域微調即可取得與頂尖 BERT 模型相當之效能；Chang et al.(2025) 則系統性設計多層次提示情境，驗證提示設計品質對 PLM 於 PPI 任務之關鍵影響。上述發展中，統計規則方法奠定基本框架，機器學習與深度學習逐步以自動化特徵學習取代人工設計，PLM 時代則開展免微調應用之可能性，但現有研究多聚焦單一階段模型比較，較少探討跨技術路徑之遷移學習策略，此亦為本研究採用 BERT／BioBERT 並導入中介任務遷移學習架構作為主要方向。

2.3 中介任務遷移學習（Intermediate-Task Transfer Learning,ITTL）

標註資料有限時，直接微調預訓練模型易導致訓練不穩定或過擬合。Pruksachatkun et al.(2020) 提出的 STILTs (Supplementary Training on Intermediate Labeled-data Tasks)架構，即在通用預訓練與目標任務之間加入資料量較大的中介任務進行訓練，為 ITTL 研究奠定基礎；Pruksachatkun et al.(2020) 進一步指出，中介任務與目標任務間的複雜度與詞彙重疊程度會影響轉移成效，選擇不當甚至造成負遷移，Legrand et al.(2021) 則證實結合句法資訊的遷移策略能提升生醫關係抽取在低資源情境下的表現。近年 Ekakristi et al.(2025) 與 Nkhata and Gauch(2025) 亦分別於印尼語任務與立場偵測任務中驗證，ITTL 成效高度依賴中介任務與目標任務間的相似程度，此亦為本研究設計多階層遷移路徑的理論依據。

3 Method

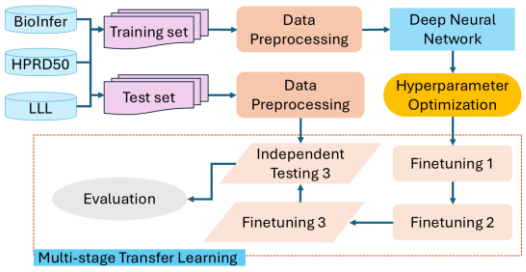

圖 1. 實驗流程圖

本研究採用單階層遷移學習 (Single-Stage Transfer Learning, SSTL) 與多階層遷移學習 (Multi-Stage Transfer Learning, MSTL) 兩種架構針對蛋白質交互作用 (PPI) 進行比較。圖 1 為整體實驗流程圖，包含數據預處理、超參數最佳化、模型訓練與評估四個主要階段。在預處理階段對五個 PPI Dataset 類別標籤進行轉換，統一將原始標籤 Ture 與 False 分別轉變為 1 和 0，並將每個句子中的任兩個蛋白質或基因名稱分別替換為 PROTEIN1 和 PROTEIN2 標籤，以利模型學習，接著將實驗分成 SSTL 與 MSTL 架構，SSTL 採用 BERT (Devlin et al., 2019) 與 BioBERT (Lee et al., 2020) 模型對目標任務進行 Optuna (Akiba et al., 2019) 超參數自動化搜索並利用找到最佳參數組合進行 Fine-Tune。而在 MSTL 架構，PLM 模型先使用 Optuna 進行經歷超參數自動化搜索，再經歷 1~2 階層的中間任務訓練並遷移到目標任務上訓練並測試，最後透過多階層架構與模型之間的比較驗證方法有效性。

3.1 Dataset

本研究採用三個在 PPI 研究領域中廣泛使用的語料庫:BioInfer (Pyysalo et al., 2007)、HPRD50 (Fundel et al., 2007)、LLL (Nédellec, 2005)。BioInfer 從 Medline 摘要選出 1,100 個句子，其標註範圍不侷限於人類，而包含生物基因、蛋白質與 RNA。該語料庫由實體、關係與依存句法三種層次組成複雜的語義標註系統。此外，BioInfer 具有複雜的嵌套關係結構，除

了描述實體間關係外，也能表示表示不同層次關係之間的連結，例如:「蛋白質 A 抑制蛋白質 B 與蛋白質 C 的結合」，其中蛋白質 B 與 C 結合屬於第一層關係，然而蛋白質 A 抑制 B 與 C 則建立於第一層關係之上，形成第二層關係，且 BioInfer 單一句子可能包含大量生醫實體，使其詳細程度達到了領域最豐富、複雜度最高。HPRD50 語料庫基於 Human Protein Reference Database，研究者從中隨機抽取 50 篇結合 ProMiner 系統自動化識別蛋白質實體且由人類專家標註。LLL 語料庫主要來自 Medline 數據庫，該語料庫專注於枯草桿菌基因與蛋白質，其標註僅納入包含明確動作 (如啟動或抑制)、結合作用以及調節子成員關係之描述，由 77 句訓練集和 87 句測試集組成。由於測試集包含較多僅提及單一蛋白質或無明確相互作用之語句，因此最終測試集僅提取出 61 筆實體對，而訓練集則提取出 269 筆實體對。完整資料集分布情況詳見表 1。

Dataset	Training Set		Test Set	
	Pos	Neg	Pos	Neg
BioInfer	2028	5339	409	1132
HPRD50	137	226	26	44
LLL	134	135	30	31

表 1. 資料集大小

3.2 Single-Stage Transfer Learning

本研究使用 google-bert/bert-base-uncased (BERT) 預訓練模型、dmis-lab/biobert-v1.1 (BioBERT-V1.1) 預訓練模型與 michiyasunaga/BioLinkBERT-base (BioLinkBERT) 進行 Fine-tuning 分析評估。

數據預處理: 為確保資料格式與深度學習框架之相容性，並符合計算損失函數 (Loss Function) 與模型輸入之張量要求，本研究將三項 PPI Data 的標籤值轉換為 0/1。訓練集切分為訓練集 (80%) 驗證集 (20%)，並確保訓練集與驗證集正負比例一致，分別透過 BERT 和 BioBERT 各自 Tokenizer 轉換為模型所需的輸入格式。

我們將每個句子中的任兩個蛋白質和基因名稱分別替換為”PROTEIN1”和”PROTEIN2”。然而有些句子含超過兩個蛋白質與基因，對於其他蛋白質基因名稱我們採用”PROTEIN”進行替換。以保持句子語意結構，減少非目標實體對 BERT、BioBERT 與 BioLinkBERT 的影響，進而使模型注意力機制更聚焦於描述蛋白質互動之上下文上。

本研究採用 LORA 微調技術，在有限的運算資源下高效的對 BERT、BioBERT 與 BioLinkBERT 進行 SSTL (Hu et al., 2021)。LORA 原理就是在預訓練模型權重旁掛上兩個 low-rank 可訓練矩陣。在訓練過程中，凍結原始模型的權重、僅調整旁邊兩個小矩陣參數。此方法可以大幅減少訓練參數量，降低記憶體消耗與訓練時間，也能降低在小資料集上 overfitting。

我們採用 Optuna (Akiba et al., 2019) 自動化搜索最佳超參數，搭配預設 TPE (Tree-structured Parzen Estimator) 採樣器 (Bergstra et al., 2011)，並以驗證集 validation loss 作為最佳化的目標，確保模型在各資料集皆能達到最佳效能。此方法較傳統的 Grid Search 更有效率找到最佳參數組合。

Target	BERT			BioBERT			BioLinkBERT		
	BioInfer	HPRD50	LLL	BioInfer	HPRD50	LLL	BioInfer	HPRD50	LLL
P	0.602	0.600	0.615	**0.646**	**0.690**	**0.743**	0.642	0.617	0.710
R	0.681	0.692	0.800	0.648	**0.744**	**0.844**	**0.694**	0.679	0.833
ACC	0.794	0.694	0.650	0.811	**0.776**	**0.776**	**0.814**	0.724	0.743
F1	0.638	0.637	0.692	0.645	**0.707**	**0.786**	**0.664**	0.646	0.761
MCC	0.498	0.401	0.324	0.518	**0.538**	**0.566**	**0.539**	0.422	0.506

表 2. BERT、BioBERT、BioLinkBERT 模型在 SSTL 架構下五項指標 Precision (P)、Recall (R)、Accuracy (ACC)、F1-Score 及 MCC 對比。BERT、BioBERT 與 BioLinkBERT 分數相比較高者以粗體標示。

本研究針對單階層遷移學習 Optuna 設置以下超參數進行搜索，在 batch_size 設置 8, 16, 32 數字範圍；learning rate 於 1e-6 到 1e-3 區間內搜索。針對 LoRA 之參數，LoRA Rank 範圍設置 8 到 24 並以步長為 4 區間搜索；LoRA Alpha 從 4, 8, 16, 32, 64 中篩選；LoRA Dropout 則於 0.0 到 0.5 之間，以步長為 0.1 進行搜索，每個資料集都獨立進行 20 次搜尋。

訓練過程使用 Hugging Face 的 Trainer 類別對目標數據集進行 100 epoch 訓練。為了防止 overfitting，我們設置 Early Stopping 機制，當 validation loss 連續 15 epoch 沒有進步時，訓練會自動停止並載入表現最佳的模型權重，並利用測試集評估模型在鑑別蛋白質交互作用的效能。

3.3 Multi-Stage Transfer Learning

本研究參考文獻將傳統 SSTL 流程擴展為包含中間任務訓練遷移到目標任務上 Fine-tuning 的 MSTL 架構 (Ekakristi et al., 2025)。為了試驗在同領域不同資料集上交互訓練是否能夠提升模型效能。本研究將多階層遷移學習架構分為二階與三階，除目標資料集外，其餘中間任務根據訓練資料集規模採取由大到小遷移策略。此策略歸於 Phang et al. (2019) 證實利用資料豐富的中間監督任務來緩解小資料集在微調時的脆弱性，進而提升最終目標任務模型的穩健性與有效性。因此，本研究以此方法引導模型進行知識的遞進式累積，以最大化遷移學習效益。

針對多階層遷移學習，本研究使用 Optuna 對超參數進行搜索。其中 Batch Size、Learning Rate 與 LoRA 搜索範圍均與 SSTL 相同。兩者差異在於多階層中不同階層之訓練任務採取各別獨立學習率，例如，在二階層遷移學習中，分別定義第一階層學習率 lrA 與第二階層的學習率 lrB。

訓練流程為確保模型能在不同階段任務達到最佳性能，針對每一次中間任務微調皆經過獨立完整 20 次的 Optuna 超參數自動化搜索，並搭配 LoRA 微調機制。而訓練配置如 Trainer 設置、訓練週期與 Early Stopping 機制皆沿用 SSTL 設定。相異之處在於 MSTL 架構訓練時若前一階層 validation loss 連續 15 epoch 沒有進步時，訓練會自動停止並載入表現最佳的模型權重，進而執行下一階層訓練任務，以此類推，直到目標階層訓練停止並載入最佳模型權重，

Target	Transfer Learning Type	P	R	ACC	F1	MCC
BioInfer	SSTL	0.646	0.648	0.811	0.645	0.518
	MSTL: H-B	0.668	0.641	0.820	0.654	0.532
	MSTL: L-B	0.657	0.618	0.812	0.636	0.510
	MSTL: H-L-B	**0.695**	**0.675**	**0.834**	**0.683**	**0.572**
	GK	0.567	0.672	-	0.613	-
HPRD50	SSTL	0.690	0.744	0.776	0.707	0.538
	MSTL: B-H	0.635	0.731	0.743	0.680	0.481
	MSTL: L-H	0.542	**0.833**	0.676	0.660	0.413
	MSTL: B-L-H	0.691	0.756	**0.781**	**0.719**	**0.546**
	GK	0.643	0.658	-	0.634	-
	LPTK	**0.727**	0.622	-	0.671	-
LLL	SSTL	0.743	0.844	0.776	0.786	0.566
	MSTL: B-L	0.805	**0.911**	0.847	**0.860**	0.677
	MSTL: H-L	0.766	0.822	0.765	0.793	0.564
	MSTL: B-H-L	**0.833**	0.878	**0.852**	0.854	**0.706**
	GK	0.725	0.872	-	0.768	-
	LPTK	0.789	0.721	-	0.753	-

表 3. 二、三階層遷移學習與 SSTL 五項指標 Precision (P)、Recall (R)、Accuracy (ACC)、F1-Score 及 MCC 對比。其中 Transfer Learning Type 代表各階層遷移學習之訓練路徑，針對二階層以上皆採用字首字母作為各資料集代表（B 代表 BioInfer、H 代表 HPRD50、L 表示 LLL）。各數據集之最高分以粗體表示。

再透過測試集評估模型效能。為了降低單次實驗分數之偶然與數據的可信性，本研究實驗均採用三次實驗之平均值作為最終評估分數。最後，為了驗證多階層遷移學習的有效性，MSTL 結果將與 3.2 節 SSTL 架構中的 BioBERT 基準模型進行比較。

3.4 Evaluation

我們使用多項評估指標，包含 Accuracy (1)、Precision (2)、Recall (3)、F1-score (4) 及 MCC (5) 進行評估，公式如下：

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (4)$$

$$MCC = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

其中 TP 為 True Positive，TN 為 True Negative、FP 為 False Positive，FN 為 False Negative。

4 Result and Discussion

4.1 SSTL 實驗結果分析

表 2 呈現 BERT、BioBERT 與 BioLinkBERT 模型在 SSTL 方法下三個 PPI Data 的 F1-score 比較。實驗結果發現 BioBERT 效能整體而言優於 BERT 與 BioLinkBERT，主因在於 BioBERT 模型基於 BERT 強大的語意理解架構，並進一步於大規模生物醫學文獻上進行預訓練，使其在生醫領域具有專業術語與生醫文本之上下文理解能力。綜合上述結果，BioBERT 表現出更佳的泛化能力，因此本研究後續所有 MSTL 實驗皆以 BioBERT 模型進行。

4.2 MSTL 實驗結果分析

表 3 呈現 SSTL 與 MSTL 在各項評估指標之實驗結果，並且與外部的方法進行比較，GK 為 Graph Kernel(Airola et al., 2008)根據句子的語法結構並建立及分析形式化的圖譜識別 PPI；LPTK (Warikoo et al., 2018)為利用語義分析樹特徵提取 PPI。綜觀整體實驗結果，可發現各資料在不同階層的 MSTL 架構下分數大多優於 SSTL，並且優於其他方法。為進一步了解 MSTL 架構對各目標資料集之影響，本研究以目標資料集規模劃分為大型資料集 (BioInfer) 與小型資料集 (HPRD50、LLL) 兩類進行最佳遷移路徑分析，並探討效能提升與下降之原因。

4.2.1 大型目標資料集遷移效能分析

BioInfer 實驗結果：以 BioInfer 結果來說，隨著遷移階層的增加模型效能也呈現逐步成長之趨勢。在 MSTL 二階中則呈現兩種截然不同的遷移效果。若中介任務使用 HPRD50 (MSTL: H-B)，其 F1 分數從 0.645 提升至 0.654，進步了 0.9%。本研究推測因 HPRD50 與 BioInfer 標註內容皆聚焦於人類蛋白質與基因上，且 HPRD50 包含清晰高質量的實體關係上下文特徵，使模型面對 BioInfer 標註內容廣域且複雜的嵌套句法結構時，能夠有效識別實體間的關係進而提升模型效能。反之，若中介任務使用 LLL 時(MSTL: L-B)，整體 F1 (0.636)與 MCC (0.510)僅微幅下滑 0.9%與 0.8%，此現象可能因 LLL 數據規模極小且專於枯草桿菌領域之同位語及三類核心動作 (明確動作、結合作用、調節子家族) 特定句法模板，在缺乏廣域的知識引導下，難以有效幫助 BioInfer，進而造成模型效能小幅度下滑。當延伸至 MSTL: H-L-B 架構時，模型效能具有顯著得提升，且各項指標更達到該資料集之最佳效能。其表現可能源自於模型吸取 HPRD50 多樣化的語意知識與

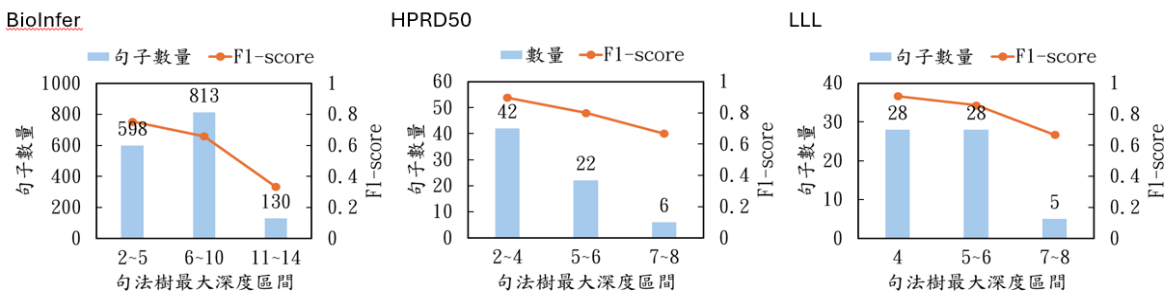

圖 2. 各資料集最大句法樹深度圖

LLL 明確的句法模板，接續遷移學習後，面對標註類型較廣泛且關係結構更複雜 BioInfer 具備較佳的泛化效果。因此，在最終測試中 MSTL: H-L-B 相較於 SSTL，ACC、F1、MCC 分別進步 2.3%、3.8%與 5.4%。

4.2.2 小型目標資料集遷移效能分析

HPRD50 實驗結果：整體實驗結果顯示二階遷移 MSTL: B-H 與 MSTL:L-H 相比於 SSTL 均呈現負遷移趨勢。此現象歸於中介任務與目標資料集在句法結構、句子平均長度及正負樣本比例差異。在句法結構上，BioInfer 多為複雜嵌套結構，LLL 則多為明確動作之語句，兩者皆與 HPRD50 以最短依存路徑為核心語句結構有所差異；此外，依據本研究統計，BioInfer 與 LLL 訓練集句子平均長度介於 34 至 35 個單詞之間，然而 HPRD50 平均僅 27 個單詞，甚至測試集平均僅 22 個單詞；且 HPRD50 正負比為 1:1.63，BioInfer 為 1:2.57 而 LLL 則趨近於 1:1。此多重數據落差致使中介任務學習之特徵分布無法有效對齊目標資料集，進而造成負遷移。反之在 MSTL: B-L-H 中，模型 F1、MCC 達到 0.719 與 0.546，取得該資料集最佳效能。此現象源於前置任務 BioInfer 提供了豐富的語意知識，且訓練具有足夠的路徑收斂模型，進而校正 MSTL 二階帶來的負遷移影響。最終與 SSTL 架構相比整體模型的 F1、MCC 分別小幅度提升 1.2%及 0.8%，展現出更佳的泛化能力。

LLL 實驗結果：以 LLL 作為目標資料集之實驗結果顯示，由於 LLL 本身數據規模為三者最小，使 SSTL 效能最為受限，而導入 MSTL 架構後，各項指標均獲得不同程度的提升。MSTL: B-L 結果提升顯著優於 MSTL: H-L。主要得益於 BioInfer 與 LLL 資料集均為 PPI 領域之複雜語句結構，且兩者句子長度分布相近。在此前提下，在 BioInfer 所提供的廣義與多物種的生物醫學知識鋪墊下，且未受到中介任務干擾，使模型 F1 與 MCC 分別達到 0.86 與 0.677 (分別提升 7.4%與 11.1%)。然而整體而言，MSTL: B-H-L 則為實驗之最佳效能，其 Precision (0.833)、Accuracy (0.852)及 MCC (0.706)均達到最佳表現，F1 也僅以 0.6%微幅落後 MSTL: B-L。相比 SSTL，三階層遷移學習架構下 F1 提升 6.8%，MCC 提升 14%，此結果證明多階層的架構可以幫助模型提升穩定性與泛化能力。

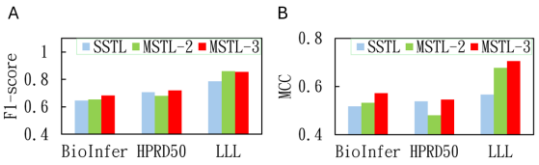

圖 3. 不同資料集在各階層最佳 F1 與 MCC 統計圖。(A)各資料集之最佳 F1-score；(B)各資料集之最佳 MCC。其中 SSTL、MSTL-2、MSTL-3 分別代表單階層、二階層及三階層遷移學習路徑。

4.3 Dependency Parse Tree 分析

我們進一步分析各資料集之句子複雜度，使用基於 Spacy 框架且為生醫優化處理工具之 Scispacy (v0.6.2) (Neumann et al., 2019)，並搭配 en_core_sci_md 模型 (v0.5.4) 對生醫文本進行處理；在句法分析上，透過 Spacy 的依存句法分析器對文本生成依存句法樹 (Dependency Parse Tree) 並結合 token.ancestors 追溯各節點通往根節點 (ROOT) 之依存路徑，最後計算最大長度作為最大句法樹深度。根據圖 2 趨勢可觀察到，當句子最大句法樹深度越長時，F1-score 隨之下降，其中 BioInfer 最大句法樹深度明顯大於其餘兩者，它在深度 11-14 之間的句法有 130 句，佔整體資料集 8.4%，而另兩個資料集最大深度皆為 8。這樣的現象也反映在其 SSTL 之 F1-score, MCC 為三個資料集最低。

綜合而言，本研究提出之 MSTL 架構表現大多優於 SSTL。其中在 BioInfer、HPRD50、LLL 三個資料集之最佳模型相比 SSTL，F1-score 分別顯著提升 3.8%、1.2%、6.8%，MCC 也分別提升 5.4%、0.8%及 14%，而資料集最小的 LLL 分數提升最多。此外，圖 3 顯示三個資料集在各階層之 F1 與 MCC 最佳效能。觀察可知，目標資料集 BioInfer、LLL 均在 MSTL-3 有顯著的提升，其中 BioInfer 歸功於 HPRD50 提供多樣化的語意知識與 LLL 明確的句法模板，能夠有效增強模型泛化能力；而 LLL 則歸功於先前任務提供豐富的上下文語義引導與更為嚴苛的句法篩選，有效彌補 LLL 在 SSTL 下數據量不足的限制。然而 HPRD50 雖在 MSTL-3 達到最佳效能，但其 F1 與 MCC 仍與 SSTL 極為接近。此現象可能源於在第二階段專注於特定領域結構後，遷移至 HPRD50 校準先前任務領域不相同帶來的負遷移，模型效能

回歸至基準表現，進而呈現與 SSTL 相似的結果。

5 Error analysis

我們將模型在 BioInfer、HPRD50、LLL 三個資料集上的錯誤預測 (FN/FP) 歸納為三大類。否定與對比失敗共 40 句，其中 BioInfer:37 句 (FN 20、FP 17)、HPRD50:3 句 (FP 3) ；共存/pathway 誤判為互作共 53 句，全數集中在 BioInfer (FN 35、FP 18)；調控/濃度變化誤判為互作共 43 句，BioInfer:42 句 (FN 20、FP 22)、LLL:1 句 (FP 1)。

5.1 Negation and Contrast Failure

此錯誤為句子中出現否定詞或對比結構（如 not、unable to），但模型未能正確判斷否定詞的作用對象，導致誤判互作關係是否存在，例如：*Expressions of PROTEIN A, and B1 in Hodgkin and Reed-Sternberg cells: not suppressed by PROTEIN1 PROTEIN2 expression.* 此句為 FN，雖出現「not suppressed by」，但否定的對象是「抑制效果」而非「兩蛋白之間的關聯」，「沒有抑制作用」本身仍是一種功能性關係描述。這類錯誤反映模型缺乏判斷否定詞實際作用對象的能力，同樣的 not/no 可能否定交互作用本身、蛋白身分的相同性、或某個效應的存在，三者字面相似但表達的意思完全不同。

5.2 Co-occurrence and Shared Pathway Confusion

此錯誤為當句子描述兩個蛋白質共同出現在同一生物路徑或功能過程中，但並未直接說明兩者之間存在物理結合，模型因缺乏明確的互作動詞而誤判為無關，例如：*Here we show that the PROTEIN1 dependent proteolysis system is involved in the regulation of PROTEIN2 turnover.* 此句為 FN，句中並未出現 bind、interact 等直接互作動詞，模型可能僅憑此判斷兩蛋白無關。此類錯誤中兩蛋白確實同屬一個功能路徑，問題不在「關係是否存在」，而在模型無法分辨「功能性參與」與「直接物理結合」這兩種強度不同的關係描述，尤其 BioInfer 語料中 involved in、participate in、pathway 等功能性語言本就比 binding 動詞更常見，容易造成系統性漏判。

5.3 Regulatory/Secretory/Concentration-Change Confusion

此錯誤為句子描述某種處理條件下生理指標的變化（如施用劑量、酵素活性升降、蛋白濃度增減），而非兩蛋白質之間的直接物理互作，模型卻因調控性詞彙的表面訊號而產生誤判，例如：*Addition of Acanthamoeba PROTEIN1 in molar ratios to PROTEIN2 of about 1.1:1 and 2.3:1 retarded the initial kinetics of assembly (induced by addition of 2mM Mg+2) and reduced the fraction of PROTEIN incorporated into filaments.* 此句為 FP，PROTEIN1、PROTEIN2 在此僅是實驗設計中各自獨立的劑量與量測變因，並非彼此結合的對象，但 addition、retarded、reduced 等調控性詞彙的表面訊號可能被模型誤認為互動線索。與 5.2 不同的是，此類錯誤中兩蛋白往往連「同屬一個路徑」都談不上，關係基礎更薄弱，也使模型在這一類上出現雙向誤判 (劑量描述誤判為 positive，調控語言與 binding 線索混雜時又可能反過來漏判 negative)。

5.4 遷移學習帶來的影響

我們比較遷移學習前後模型在三個資料集上的預測結果。在 SSTL 預測錯誤但在 MSTL 預測正確的數目共計有 135 筆 (在 3 組資料分別為 BioInfer 123 筆,HPRD50 3 筆, 與 LLL 9 筆)。在 135 筆改善案例中，70%屬於 FP 修正為 TN，顯示遷移學習的主要貢獻在於抑制正例過度預測。改善集中在三類語言特徵：(1)含強交互動詞 (bind、interact、complex) 的句子佔改善案例 35%，遷移後模型學會這類動詞不一定表示當前 PROTEIN1–PROTEIN2 這一對有直接關係； (2)含否定詞 (not、unable to、independent of) 的句子佔 23%，FP 誤判明顯減少； (3)共存／pathway 描述 (involved in、signaling pathway) 佔 18%，亦以 FP 修正為主。遷移學習帶來 86 筆退步案例中，56%為 TP 變為 FN，44%為 TN 變為 FP，可以大概分成兩類語言特徵：(1)佔退步案例 52%的複合修飾與嵌套子句 (含 which、that、whereas)，當蛋白質間的作用被包覆於多重從屬子句時，遷移後容易忽略深層的語意脈絡而未能成功擷取真實關係；(2)佔退步案例 28% 的結構與異構物描述(含 domain、isoform、subunit)，模型容易將分子結構或片段的共存關係誤判為直接相互

作用，這顯示模型在面對深層文脈與複雜語意干擾時，判斷穩定度仍存在局限。

6 Conclusion

本研究將遷移學習應用於 PPI 辨識任務，比較單階層與多階層遷移學習架構，於 BioInfer、HPRD50、LLL 三個語料庫進行實驗。結果顯示，BioBERT 在單階層架構下全面優於 BERT；多階層遷移學習則整體優於單階層基準，F1-score 分別提升 3.8%、1.2%、6.8%，MCC 也分別提升 5.4%、0.8%及 14%。驗證了中介任務式遷移學習的有效性。分析發現，遷移成效高度取決於中介任務與目標任務間的領域相似性、句法複雜度與資料規模。三階層路徑在大型語料庫上展現逐步限縮語意判斷範圍的優勢，穩定性與泛化能力 (MCC) 皆優於單階層與二階層架構。

錯誤分析進一步顯示，遷移學習主要透過抑制模型對強交互動詞、否定詞及共存描述等詞彙表面線索的過度泛化，有效減少 FP 誤判；然而遷移同時帶來一定程度的保守化傾向，使部分真實正例轉為漏判，顯示模型在 Precision 與 Recall 間仍存在權衡空間，對複雜語意結構的判斷穩定度亦有待提升。未來研究可納入更多樣化的中介任務組合，並延伸至其他生醫關係抽取任務，驗證架構之泛化性。

References

Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. 2008. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics*, 9(Suppl 11):S2. https://doi.org/10.1186/1471-2105-9-S11-S2

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. arXiv:1907.10902 [cs]. https://doi.org/10.48550/arXiv.1907.10902

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for Hyper-Parameter Optimization. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Yung-Chun Chang, Ming-Siang Huang, Yi-Hsuan Huang, and Yi-Hsuan Lin. 2025. The influence of prompt engineering on large language models for protein–protein interaction identification in biomedical literature. *Scientific Reports*, 15:15493. https://doi.org/10.1038/s41598-025-99290-4

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Adrianus Saga Ekakristi, Alfan Farizki Wicaksono, and Rahmad Mahendra. 2025. Intermediate-task transfer learning for Indonesian NLP tasks. *Natural Language Processing Journal*, 12:100161. https://doi.org/10.1016/j.nlp.2025.100161

Katrin Fundel, Robert Küffner, and Ralf Zimmer. 2007. RelEx—Relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371. https://doi.org/10.1093/bioinformatics/btl616

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL].https://doi.org/10.48550/arXiv.2106.09685

S. Kim, S.-Y. Shin, I.-H. Lee, S.-J. Kim, R. Sriram, and B.-T. Zhang. 2008. PIE: an online prediction system for protein-protein interactions from text. *Nucleic Acids Research*, 36(Web Server):W411–W415. https://doi.org/10.1093/nar/gkn281

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Joël Legrand, Yannick Toussaint, Chedy Raïssi, and Adrien Coulet. 2021. Syntax-based transfer learning for the task of biomedical relation extraction. *Journal of Biomedical Semantics*, 12:16. https://doi.org/10.1186/s13326-021-00248-y

Yijing Li, Yanping Chen, Yongbin Qin, Ying Hu, Ruizhang Huang, and Qinghua Zheng. 2021. Protein-protein interaction relation extraction based on multigranularity semantic fusion. *Journal of Biomedical Informatics*, 123:103931. https://doi.org/10.1016/j.jbi.2021.103931

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. Large Language Models: A Survey. arXiv:2402.06196 [cs.CL]. https://doi.org/10.48550/arXiv.2402.06196

Claire Nédellec. 2005. Learning Language in Logic - Genic Interaction Extraction Challenge. *Proceedings of the Learning Language in Logic 2005 Workshop at the International Conference on Machine Learning*.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *Proceedings of the 18th BioNLP Workshop and Shared Task*:319–327. https://doi.org/10.18653/v1/W19-5034

Gibson Nkhata and Susan Gauch. 2025. Intermediate-Task Transfer Learning: Leveraging Sarcasm Detection for Stance Detection. arXiv:2503.03172 [cs]. https://doi.org/10.48550/arXiv.2503.03172

Yifan Peng and Zhiyong Lu. 2017. Deep learning for extracting protein-protein interactions from biomedical literature. In Kevin Bretonnel Cohen, Dina Demner-Fushman, Sophia Ananiadou, and Junichi Tsujii, editors, *Proceedings of the 16th BioNLP Workshop*, pages 29–38, Vancouver, Canada,. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-2304

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2019. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. arXiv:1811.01088 [cs]. https://doi.org/10.48550/arXiv.1811.01088

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020a. Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.467

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020b. Intermediate-Task Transfer Learning with Pretrained Models for Natural Language Understanding: When and Why Does It Work? arXiv:2005.00628 [cs]. https://doi.org/10.18653/v1/2020.acl-main.467

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8:50. https://doi.org/10.1186/1471-2105-8-50

Hasin Rehana, Nur Bengisu Çam, Mert Basmaci, Jie Zheng, Christianah Jemiyo, Yongqun He, Arzucan Özgür, and Junguk Hur. 2023. Evaluation of GPT and BERT-based models on identifying proteinprotein interactions in biomedical text. *ArXiv*:arXiv:2303.17728v2. https://doi.org/10.48550/arXiv.2303.17728

Neha Warikoo, Yung-Chun Chang, and Wen-Lian Hsu. 2018. LPTK: a linguistic pattern-aware dependency tree kernel approach for the BioCreative VI CHEMPROT task. *Database*, 2018. https://doi.org/10.1093/database/bay108