

出國報告（出國類別：進修）

2024 美國癌症研究協會年會暨德州大學/匹茲堡大學研習研究心得報告

服務機關：臺中榮民總醫院

姓名職稱：蕭自宏研究員

派赴國家/地區：美國/聖地牙哥、聖安東尼奧與匹茲堡

出國期間：2024 年 4 月 5 日至 2024 年 4 月 21 日

報告日期：2024 年 5 月 13 日

2024 美國癌症研究協會年會暨德州大學/匹茲堡大學研習研究心得報告心得報告

目次

摘要	6
目的	7
過程	8
心得	13
建議	14

摘要

2024 年美國癌症研究協會年會聚焦了癌症研究的最新進展，主題為 "Inspiring Science • Fueling Progress • Revolutionizing Care"。會議由史丹佛大學和波士頓麻省總醫院的專家擔任主席，探討了人工智慧、計算腫瘤學、空間生物學和多模態數據科學等領域。在開幕的 Plenary Talk 中，研究者們展示了最新的 3D 組織成像技術，並強調了計算生物學在癌症研究中的重要性。會議還討論了大型語言模型在提取組織病理報告結構化資訊中的應用，以及人工智慧在藥物發現、個性化治療和臨床試驗設計中的潛力。這些技術的應用可以幫助改善癌症診斷、治療和預防。

我們在會議上貼出了一張海報，展示了關於運用單細胞定序探討食道癌腫瘤異質性和生物標記的研究成果。通過與其他研究者的交流，我們深入了解了腫瘤微環境的影響，並探討了如何使用單細胞定序技術來分析腫瘤內部結構。這次年會為我們提供了許多寶貴的知識和見解，並為未來的研究方向帶來了啟發。

至德州大學以及匹茲堡大學的研習涵蓋了空間生物學和深度學習於生物資訊領域的研習，包含了深度學習在 HLA 配對抗原預測中的應用；空間基因體分析與單細胞定序實驗；深度學習 embedding 的應用與研究探討等。這次研習提供了關於相關研究細節的深入了解。本次研習也體會深度學習在生物資訊分析領域的迅速發展。計劃將此次研習中的學習和經驗應用到本院的研究中，並尋求進一步的國際合作。

關鍵字：深度學習、空間生物學、多模態數據

目的

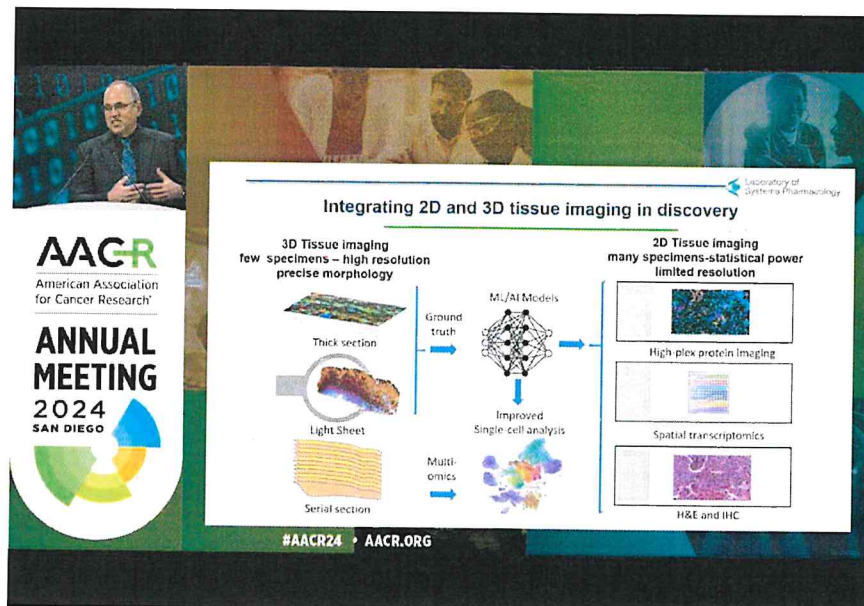
本次參訪旨在技術的精進與拓展與國際研究界的合作機會，特別是在空間生物學以及癌症研究領域。藉由參訪提升本院空間基因體研究以及精準癌症研究的能量。此外，參與頂尖研究機構的研究對本院醫學與基因體學能量有所提升。

2024AACR 美國癌症研究協會年會

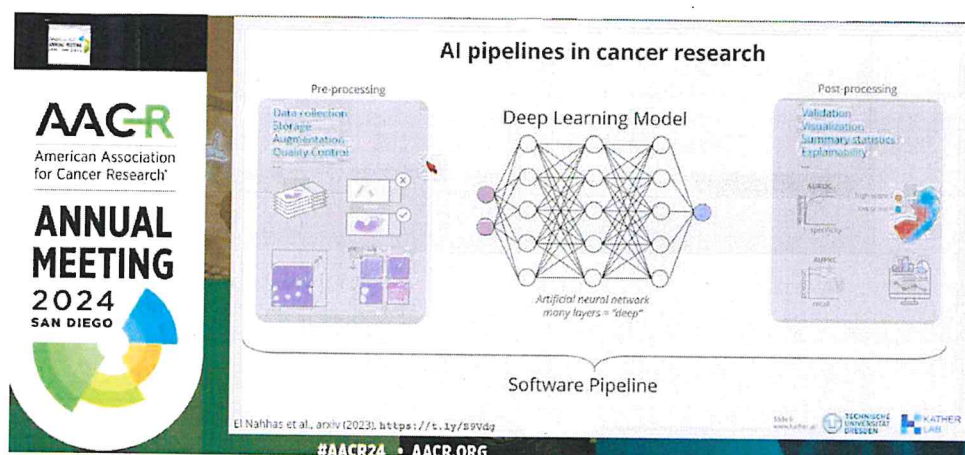
在今年的美國癌症研究學會（AACR）2024 年會上，主題為 "Inspiring Science • Fueling Progress • Revolutionizing Care"，以傳遞科學創新和推動進步的精神為核心。本次參加了這個會議，體驗到了許多先進的前沿研究和技術進展。本次會議由史丹佛大學醫學、遺傳學和生物醫學數據科學教授 Christina Curtis 以及波士頓麻省總醫院癌症中心的 Keith T. Flaherty 共同擔任主席。Christina Curtis 教授在史丹佛大學負責人工智慧、癌症基因體學以及乳癌轉譯研究。Keith T. Flaherty 專長為新藥開發、設計。今年的會議特別關注科技研究進展”領域，其中包括機器學習、計算腫瘤學與空間生物學。本次開幕前兩天的 Planty talk 都圍繞三個重點:spatial biology 空間生物學、AI 人工智慧、和 multimodality 多模態數據科學等領域。前兩天的 plenary talk 上，多個世界頂尖機構包含哈佛大學、歐洲 EBI 等教授都展示了建構在結合數位病理、基因體學、跟多重抗體螢光染色再加上醫療資訊結合而成的空間生物學和多模式數據科學癌症研究。過往生物資訊以及資訊分析在癌症研究都是扮演協助分生實驗的角色，但 2024 年整個 plenary talk 幾乎每位講者的研究都有大量計算生物學的研究成果，本次 AACR 大會揭露了在癌症研究，資料科學分析佔的角色越來越重要。

其中哈佛大學系統生物學的 Peter Sorger 教授展示了最新的 3D 空間基因體學的方法，他共解釋了四種不同的 3D 組織成像技術及其應用，包括光學切片成像、連續薄切片重建、共聚焦光學切片成像和循環光片螢光顯微成像等。每種技術都有不同的解析度、成像深度和適用樣本厚度範圍。圖片中顯示了這些技術的示例成像結果，清晰可見腫瘤細胞、血管等組織結構和生物標記物的分布情況。他並在演講中展示他的研究成果。顯示現在的空間生物學將從二維推展到 3D，進而完整了解腫瘤在組織層級的運作。當然需要結合多種不同型態的資料，也是這次大會非常關注的議題 multimodality。如何使用 AI 去解析這些高度複雜的資料也是本次大會的重點之一。

最新的 3D 組織成像技術在癌症研究中具有多方面的應用價值：利用高解析度的 3D 成像可以清晰展現腫瘤內部的異質性結構，精細觀察腫瘤細胞的形態、分佈、基因表達和信號傳導途徑等分子特徵，幫助深入理解腫瘤發生發展機制。3D 成像可以詳細展現腫瘤與其周圍的血管、免疫細胞、基質細胞等微環境的空間分布和相互作用，揭示微環境對腫瘤生長、轉移等行為的影響。

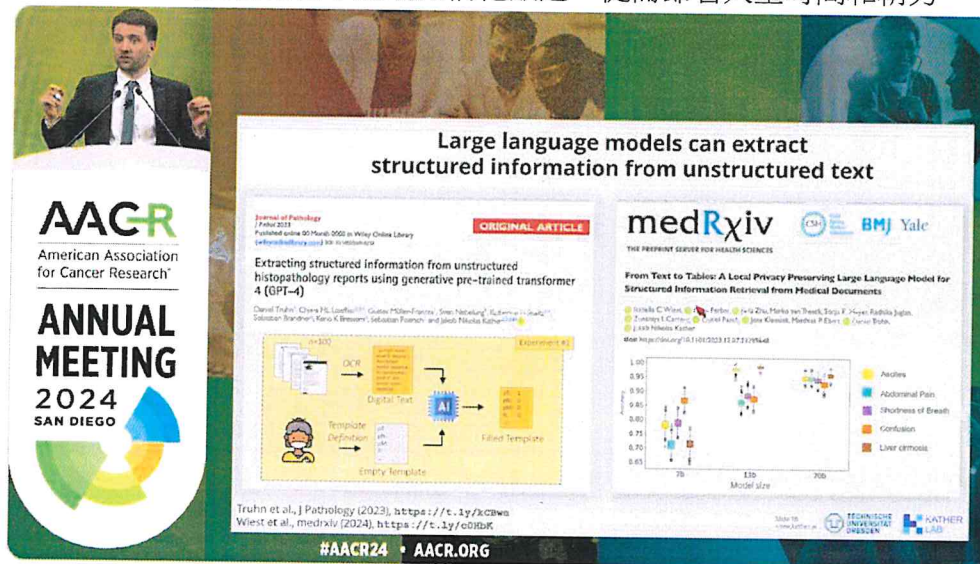


另外與會的專家也闡釋了人工智慧在癌症研究中的具體應用，例如：人工智慧演算法可以分析大量基因體和分子數據，以識別潛在的新療法標靶。可以用於設計和優化新的癌症療法，如藥物和免疫療法、開發早期癌症檢測、個性化治療計劃和監測治療反應的工具。人工智慧驅動的癌症可能的研究工具包含基於人工智慧的圖像分析工具：這些工具可以分析醫學圖像，如 X 光和 CT 掃描，以識別癌細胞和其他異常情況。人工智慧驅動的藥物發現平台：這些平台可以加速藥物發現的過程，通過識別潛在的藥物候選物並預測其效用和安全性。人工智慧驅動的臨床試驗設計工具：這些工具可以幫助研究人員設計更有效和更高效的臨床試驗。演講中並闡釋了人工智慧在癌症研究中的未來展望，強調其有潛力在癌症預防、診斷和治療中帶來革命性變革。

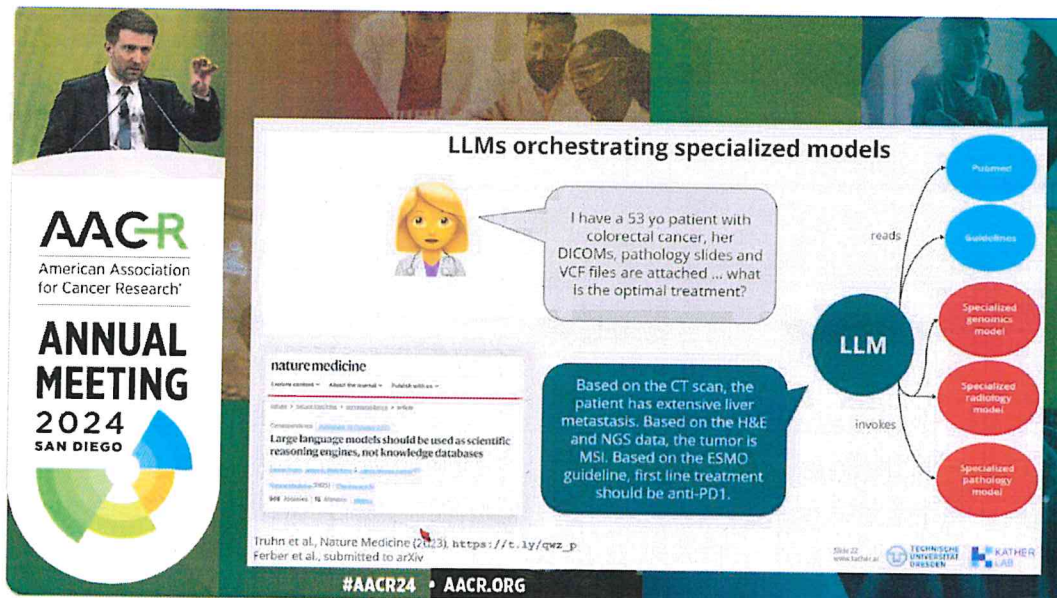


本次會議的專家也揭示了如何使用大型語言模型（GPT-4）從非結構化組織病理報告中擷取結構化資訊。組織病理報告是重要的臨床資料，但由於其非結構化特性，一般分析經常面臨困難。使用大型語言模型來從非結構化組織病理報告中提取結構化資料進而進行資訊分析。大型語言模型的普及將克服許多傳統方法的局限性。提供準確又高效的資料擷取。將可能徹底改變組織病理報告的分析方式。

自動化從組織病理報告中提取結構化茲運，從而節省大量時間和精力。



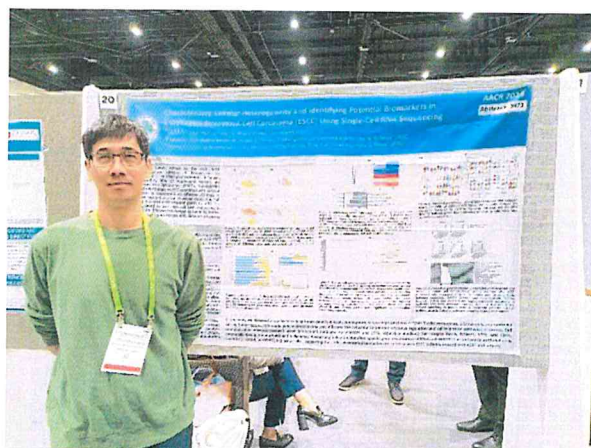
更進一步地，專家闡釋了「大型語言模型 (LLM) 實現精準腫瘤學的假設的情景，例如一名 53 歲的結腸直腸癌患者。患者的 DICOM 文件、病理切片和 VCF 文件被輸入到大型語言模型中，該模型協調了專業的基因體學、放射學和病理學模型，最終生成個性化的治療建議。演講強調了 LLM 可以作為整合和協調精準腫瘤學中專業模型的中心樞紐的潛力。以及未來 LLM 如何利用多種數據源提供全面且量身定制化的治療建議。本次演講傳達了一個對精準腫瘤學未來的引人注目的願景，其中 LLM 在整合和協調專業模型以提供個性化、數據驅動的治療建議中扮演關鍵角色。



本次在 2024 年美國癌症研究協會年會上，團隊張貼了一張海報展示了最新的研究成果，海報題目為” Characterizing Cellular Heterogeneity and Identifying Potential Biomarkers in Esophageal Squamous Cell Carcinoma (ESCC) Using Single-Cell RNA

Sequencing”，本次海報展示我們研究團隊運用單細胞定序去探討食道癌的腫瘤細胞，並探討對於腫瘤各類細胞的基因表達。在海報的張貼過程中與很多與會的學者討論，也得到很多反饋意見，例如我們找到一個腫瘤高表達基因 SPP1，在本次會議的演講中，其他團隊的空間基因體資料顯示其為癌症相關的 Cancer-Associated Fibroblasts 所釋放，顯示腫瘤微環境對於癌症的發展有重大的影響，相關心得之後將用於本研究之精進。

與會期間，我收取癌症醫學最新的發展。尤其是對於 AI 跟 multimodality 的注重。也觀察到其他研究單位如何使用新穎的基因體技術例如單細胞定序、空間生物學學等等探索癌症醫學更嶄新的領域。這些領域的最新進展不僅擴展了我的學術視野，同時為未來的研究方向提供了寶貴的參考。



德州大學與匹茲堡大學研究

本次兩所大學的深入研究包括以下主要方面：

深度學習演算法於人類 HLA 配對抗原之預測

本次研習針對使用深度神經網絡來預測主要組織相容性複合體 (MHC) 的抗原呈現，並利用遷移學習方法來預測新抗原的免疫原性。這類的生物資訊演算法可以幫助找尋能引發適應性免疫反應的抗原，未來的應用找尋腫瘤新生抗原 (neoantigens)，此為癌症個人化疫苗開發的重要關鍵，也可以用於了解免疫疾病可能引發的抗原。

本次深入研究了 BigMHC 這個深度神經網路方法，此方法先在質譜檢測的胜肽 MHC 配體數據上進行訓練，並通過遷移學習抗原特異性免疫反應的實驗數據。BigMHC 在訓練過程中使用多種深度神經網絡架構，包括雙向 LSTM、錨點塊等，來準確學習抗原呈現和免疫原性。結果呈現高準確度，具備於臨床研究中應用的潛力。本次研究理解此新抗原預測方法之原理，以及如何運用於免疫疾病可能的抗原預測以及癌症免疫療法引發所需的腫瘤新生抗原(neoantigens)預測。

空間基因體分析實作：

本院預計於五月開始進行空間基因體實驗，實驗下機之後要開始進行資料分析，本次研習使用德州大學產出之空間基因體資料，運用德州大學建立的分析流程進行學習，學習的過程包含資料前處理、細胞分型、UMAP 空間位點分型、差異表達基因鑑別等步驟。

參訪定序實驗室以及了解所用的單細胞實驗系統

本次參訪了德州的定序實驗室並了解其如何進行單細胞定序實驗以及空間基因體實驗，在參訪過程中得知該實驗室最近測試一家新公司提供單細胞定序耗材，其價格為目前我院在使用 10x 公司試劑的一半不到，回國之後我也聯繫相關的代理商，並詢問德州大學的測試結果，如果後續品質穩定，本院將更換耗材提供廠商以節省成本。

深度學習 embedding 的應用

Embedding 是將來源數據透過函數映射到另一個空間的向量表示方式，大型語言模型先將語言文本轉換成多維 embedding 向量之後再進行 transformer 的深度學習模型處理。目前最常用的檢索增強生成 (Retrieval-Augmented Generation, RAG) 技術背後的原理就是使用 embedding。目前常用的是使用 Word2vec 來進行轉換，此演算法運用 skip-gram 與 Continuous Bag of Word (CBOW) 的方法來實作，核心是一個極為淺層的類神經網路。透過使每個字詞與前後字詞的向量相近，來訓練出含有每個字詞語義的字詞向量。目前很多資料科學研究是基於 embedding 向量

來進行數值運算，在匹茲堡癌症中心的實驗室，他們將基因功能的描述的文字轉換成 embedding 之後在數值空間進行分群，進而想運用描述來找到相近功能的基因。本次研習參與了它們的研究討論，並一同探討在 embedding 空間可能的研究發展方向。

參與茲卡病毒于小鼠腦部感染之空間基因體研究

本次研習參與了一個多團隊的討論會議。進行的相關研究的緣由是因為茲卡病毒非洲株跟亞洲株的感染率差異很大。所以病毒實驗室運用小鼠感染兩株不同的茲卡病毒，再將不同時間點的小鼠大腦進行空間基因體學的實驗，然後將數據交由生物資訊實驗室進行分析，本次討論會就是研究感染的生物實驗室與生物資訊實驗室的 join meeting。兩邊就數據分析與生物機轉進行相互的討論，進而規畫後續的研究方向，空間基因體學是非常新穎的基因體技術，大家都還在摸索如何使用這樣新穎的技術，兩邊在分析以及如何運用層面進行討論，看如何闡釋產出的 data，推動生物醫學的進展。

參訪空間蛋白質體實驗室

我拜訪了 Dr. Wayne Stallaert，他專注於空間蛋白質體研究，他實驗室有一台萊卡的 Cell Dive 高階儀器。此機器為病理切片多重抗體螢光染色系統。可高通量檢測上百個抗體於一片病理切片上。他目前最高一片病理切片可同時使用 150 個抗體去偵測 150 蛋白質的表達量。我跟他請教了這種多重抗體染色實驗需要注意的地方。他也提及後續產出的資料量非常的龐大，一百五十個抗體就是一百五十張高解析度的螢光染色圖片，後續影像處理跟資料分析所要耗費的心力比前端實驗要多上很多倍。在此次的討論當中，他教導了很多後續的分析細節。

參與深度學習 journal club

我也參與了每兩周一次的 journal club of deep learning, 本次討論的論文題目是”CancerGPT for few shot drug pair synergy prediction using large pretrained language models”。因應深度學習的快速進步，匹茲堡的團隊有固定的 journal club，研習深度學習於生物資訊以及基因體領域的最新進展。

研習多重抗體螢光染色分析流程

本次我也跟匹茲堡的生物資訊團隊請教病理多重螢光染色的分析。以下是它們使用各種分析工具的流程：

圖像導入：使用 QuPath 或其他圖像處理工具將多重螢光染色的病理圖像導入到分析環境中。確保圖像解析度和色彩深度足夠，以進行後續的分析。

細胞分割：使用 CellPose 或 QuPath 中的分割功能，對病理圖像進行細胞分割。這步驟的目的是識別和標記圖像中的各個細胞。調整分割參數以確保準確的細胞分割，避免過度或不足分割。

細胞提取：使用 Cell Cutter 從已分割的圖像中提取單個細胞。這樣可以更深入地

分析細胞的特徵和染色模式。根據需要，對提取的細胞進行篩選，排除可能的雜質或非目標細胞。

特徵提取：使用 QuPath 或其他工具，提取每個細胞的特徵，如形狀、大小、螢光強度等。根據不同的螢光染色，分析不同的生物標記物在細胞中的分布和表達。

giga SOM 分析：使用 giga SOM 對提取的細胞進行聚類分析，以識別細胞群體中的模式和關聯。設定聚類參數，確保聚類結果的穩定性和準確性。

多重螢光染色分析：使用 CyLinter 檢查多重螢光染色的圖像，以確保染色的品質和一致性。分析不同染色之間的相關性，並識別可能的異常或錯誤。這個流程可以作為我們以後發展的基礎框架，再根據我們的需求進行調整。

本次參訪收穫分為兩個面向，其中一個是空間基因體技術以及多重抗體螢光染色細節的深入了解，跟著國外的團隊以它們現在正在進行的研究進行探討，了解對方進行資料分析之軟體以及分析流程，後續將以此次參訪的分析流程為基礎，作為本院發展的技術模板。此外，了解兩個研究機構對於深度學習的高度關注，了解他們如何將深度學習的技術導入生物資訊分析領域。此次的開會與也感受到國外進步的速度飛快，每一年都有不同的進展，回來將加強本院的發展，也持續與國外機構保持國際合作，加速本院進展。

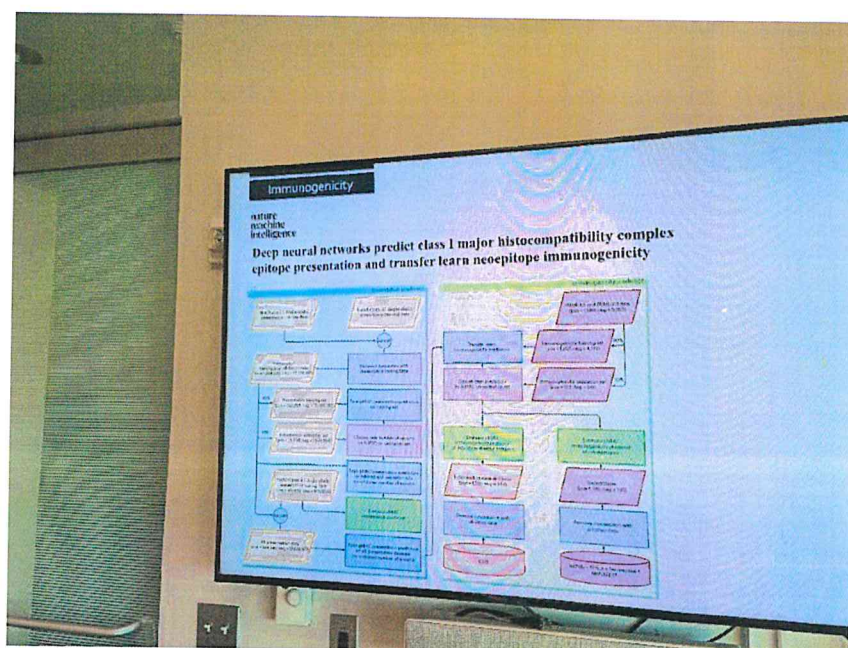
心得

出國會議與研習最大的心得是感受到深度學習對於研究領域的影響，美國癌症年會充滿了 AI 的演講。兩間大學的生物資訊人員都在探討如何使用深度學習來強化本身自我的研究。除了單純的應用之外，國外的研究都深入了解深度學習的理論基礎，進而將此技術量身打造去推動自我專注的研究領域。此外本次也感受到國外進步的速度，例如空間生物學已經從單純的基因或者蛋白質資料，轉化成更複雜的 3D 空間生物學，或者結合其他種類資料進行多模態(multimodal)分析。

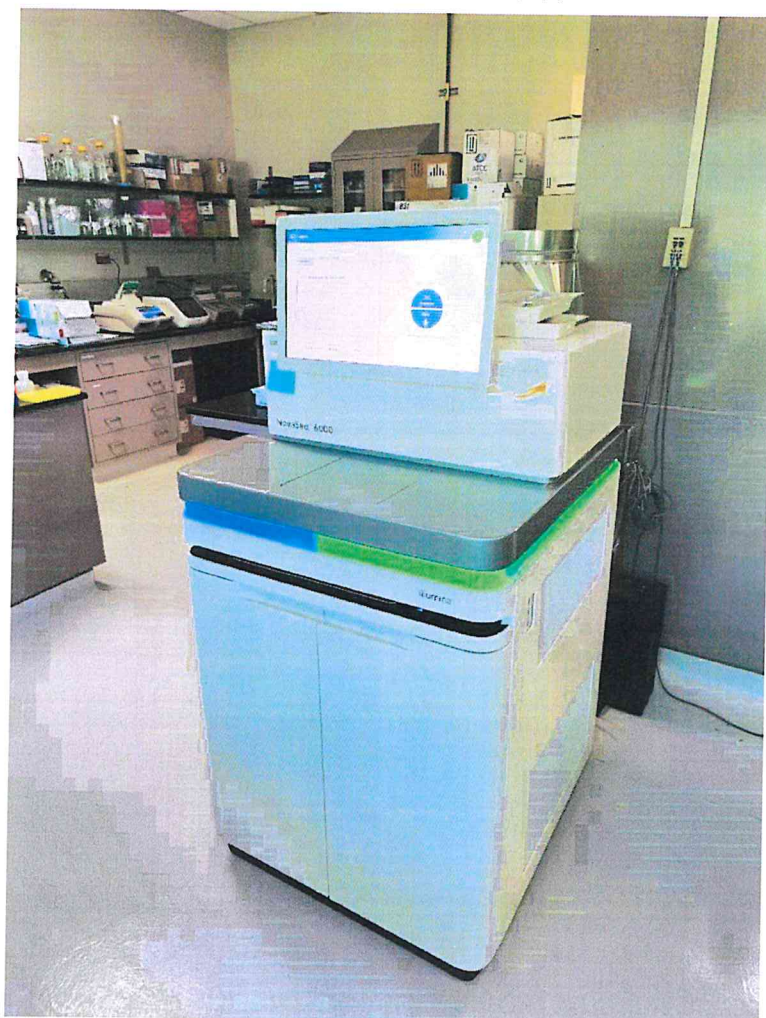
此次研習也理解了 embedding 對於文本資料研究的潛力，將在 TPMI 的資料建構 embedding 的分析方式，進而強化研究能量。

建議

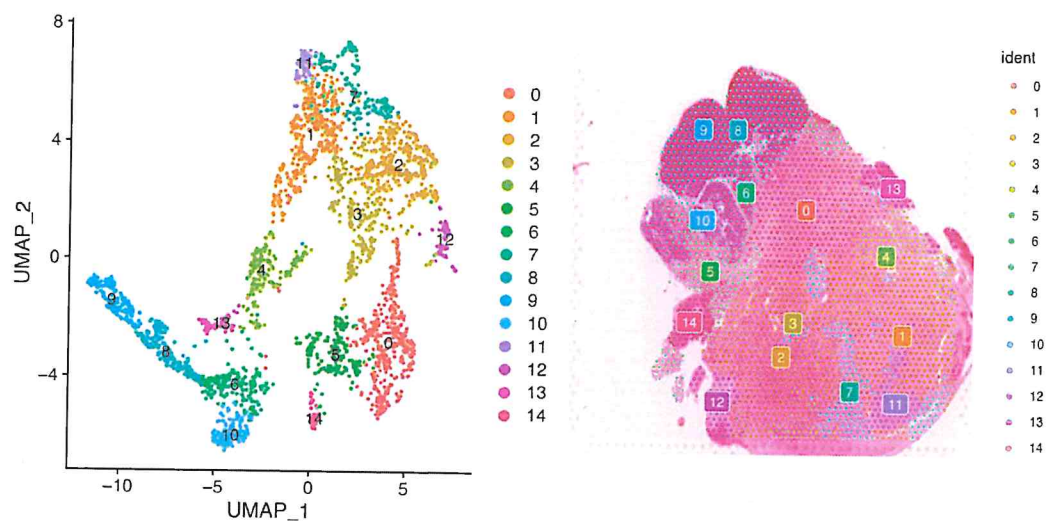
1. 建構本院核心算力：深度學習對於資料科學與醫學研究扮演越來越重要的角色，算力是深度學習的基礎需求，建議本院建立基礎算力以供未來發展使用。
2. 強化深度學習在基因體學的研發能量：建議建構深度學習相關專業研究能量，深入理解數學模型，進而驅動如 transformer 或者 embedding 等深度學習的前沿發展與本院的基因資料做結合，並延伸與醫療資訊整合，將可強化本院之研發能量。
3. 關注基因體學多模態 multi-modal 資料分析的發展：本次 AACR 已有多個頂尖團隊發表了空間生物學的 multi-modal analysis 的成果。除空間生物學之外，LLM 等領域也在討論多模態的資料處理。本院還在發展基礎的空間基因體學，多重螢光染色技術也還在建立當中。須持續專注多模態在空間生物學未來更進一步的發展
4. 關注 embedding 與檢索增強生成 (Retrieval-Augmented Generation, RAG) 技術於精準醫學的相關發展：本院 TPMI 資料庫須結合很多文本資料，例如 SOAP 或者出院摘要，可以使用 RAG 進行相關應用，或者使用 embedding 進行分析研究，將針對此技術找尋研究方向。



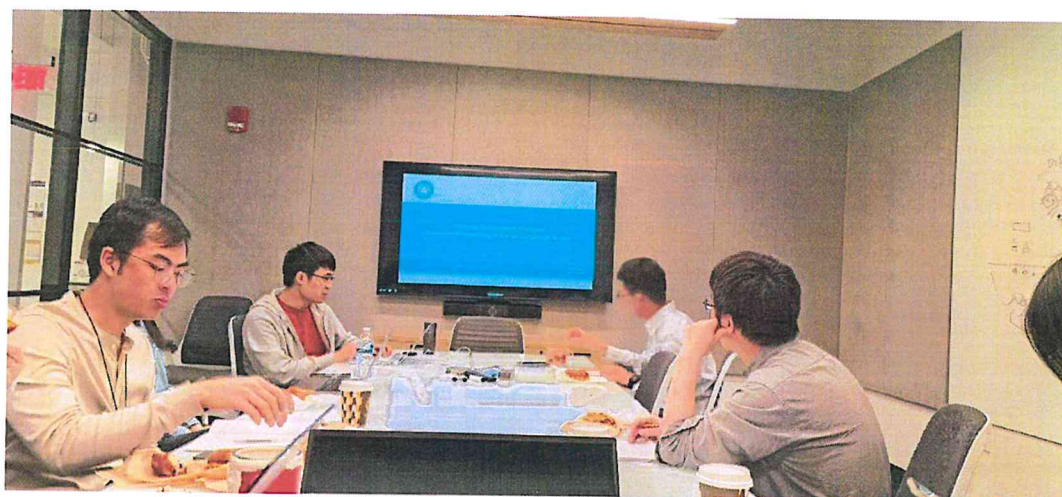
圖一：深度學習於人類 MHC 抗原討論會議



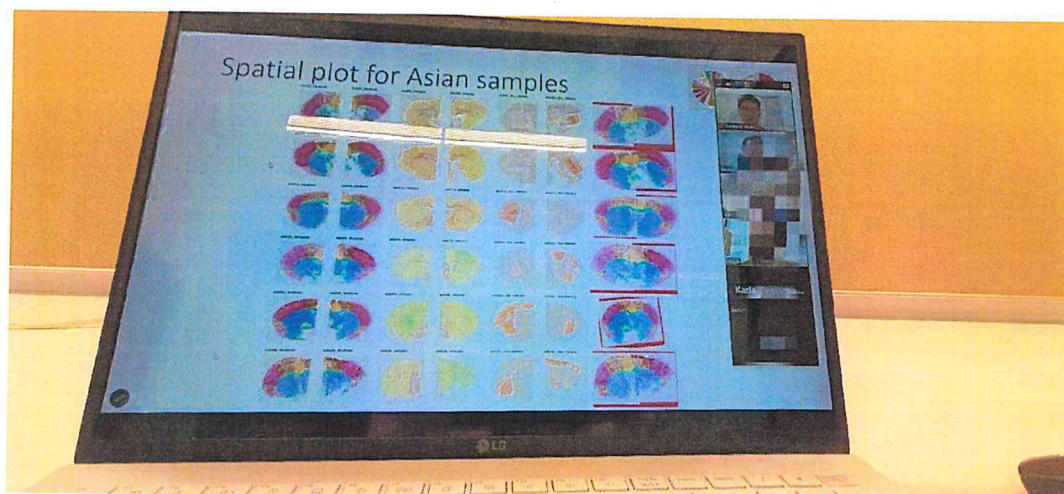
圖二:參觀德州大學基因體實驗室



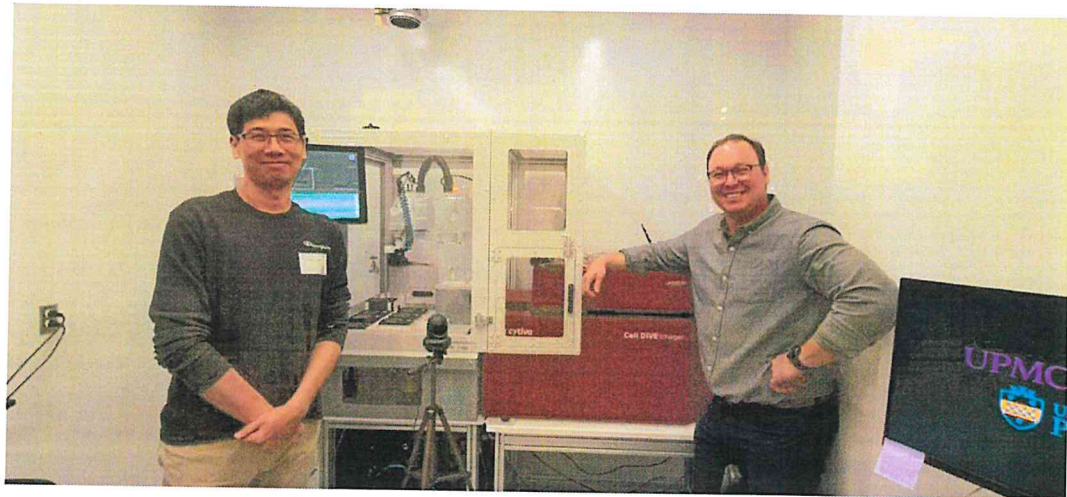
圖三:空間基因體資料實作分析



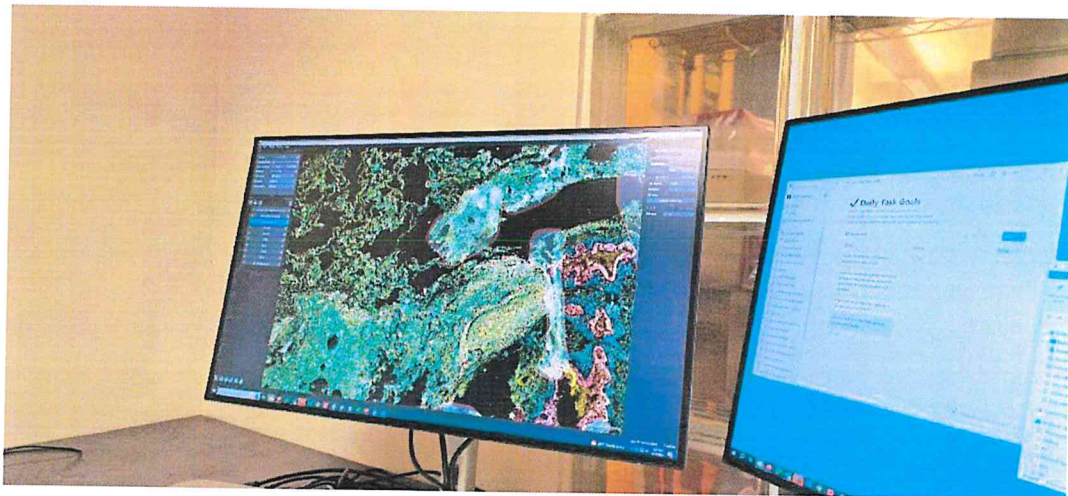
圖四:深度學習 embedding 應用討論會議



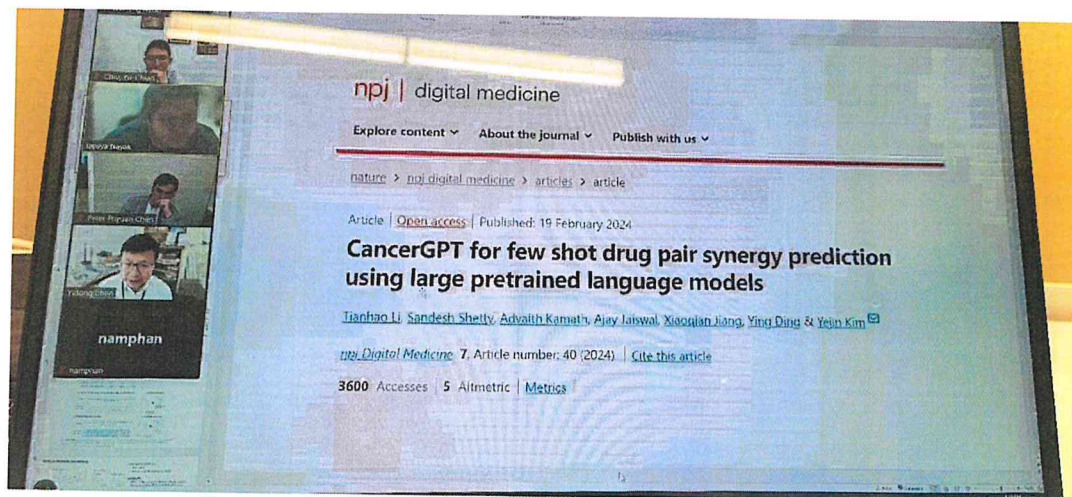
圖五: 茲卡病毒空間基因體研究討論會議



圖五: 參訪多重抗體螢光染色實驗室



圖六: 見習實作分析多重抗體螢光染色分析



圖七: 深度學習 journal club