出國報告（出國類別：進修）

學習全基因體定序分析與臨床應用

服務機關：國立成功大學醫學院附設醫院
姓名職稱：林伯昱 主治醫師
派赴國家：美國
出國期間：113 年 8 月 14 日至 114 年 8 月 13 日
報告日期：114 年 9 月 15 日

摘要

本次出國進修於美國紐約大學醫學資訊碩士學程，主要學習全基因體定序分析流程及其臨床應用，並結合人工智慧探討基因變異致病性判讀。課程涵蓋程式設計、生物資訊學與機器學習。研究部分於 Brandes 教授實驗室，專注於基因語言模型的應用，包括：比較不同模型在致病性預測的效能、分析蛋白質語言模型 ESM1b 在不同基因與功能區域的表現，以及探討如何依循臨床指引有效整合多模型使用。其中，我提出的共同校準轉換方法 P-KNN，能優於單一模型提供更強的判讀效力，並獲美國人類遺傳學會接受為海報發表。研究成果亦獲紐約大學頒發傑出研究獎，並已投稿論文。同時，本次進修觀摩國際機構在算力管理、資料規範及跨領域教育的經驗，對我國醫療資訊發展具重要借鏡。

關鍵字：基因體定序、基因語言模型、校準轉換

目次

目的

1. 學習全基因體定序從原始資料到基因變異的分析流程。
2. 學習使用最新機器學習技術判讀基因變異的致病性。
3. 學習把各項基因變異判讀的證據合理使用在臨床上。

過程

過去這一年，我透過就讀紐約大學的醫學資訊碩士學位，學習如何把全基因體定序資料應用在臨床照護上，過程可以分成課程與研究兩方面：

在課程方面，我修習了「python」與「R」的程式語言，建立資料分析的基礎能力；接著，我也修習了「基因體生物資訊學」 (bioinformatics) 以了解生物資訊的運算理論、「基因定序資訊分析」 (applied sequencing informatics) 以了解基因體資料的分析步驟以及現有工具的使用方法以及分析流程 (目的 1) 、「生物資訊資料庫管理」以學習關聯式資料庫的建置與維持；此外，我還修習了「機器學習」與「深度學習」兩門課，以了解最新的人工智慧技術以及他們在生物醫學方面的應用，。

在研究方面，我加入了 Nadav Brandes 教授的實驗室，Nadav Brandes 教授的專長是基因語言模型，也就是把 DNA 的核甘酸序列或蛋白質的氨基酸序列，使用跟語言人工智慧類似的方式來處理分析。在過去這一年，我主要參與探究了三個主題，第一個主題，是比較各個不同的基因語言模型預測基因變異臨床致病性的能力[1] (目的 2) ，此研究已經投稿正在同儕評審中，第二個主題，是評估 ESM1b[2,3] 蛋白質語言模型在不同功能的基因、蛋白質功能區域預測基因變異臨床致病性的能力的變化，第三個主題，則是用符合臨床治療指引的方式 (目的 3) ，如何有效使用多個致病性預測模型。上述研究的主題三最後成為我的碩士論文，這篇論文獲得本年度紐約大學醫學資訊碩士學程頒發的傑出研究獎 (全年度一名) ，也已經投稿正在同儕評審中。

圖一：畢業時獲得傑出研究獎

圖二：與指導老師 Nadav Brandes 合照

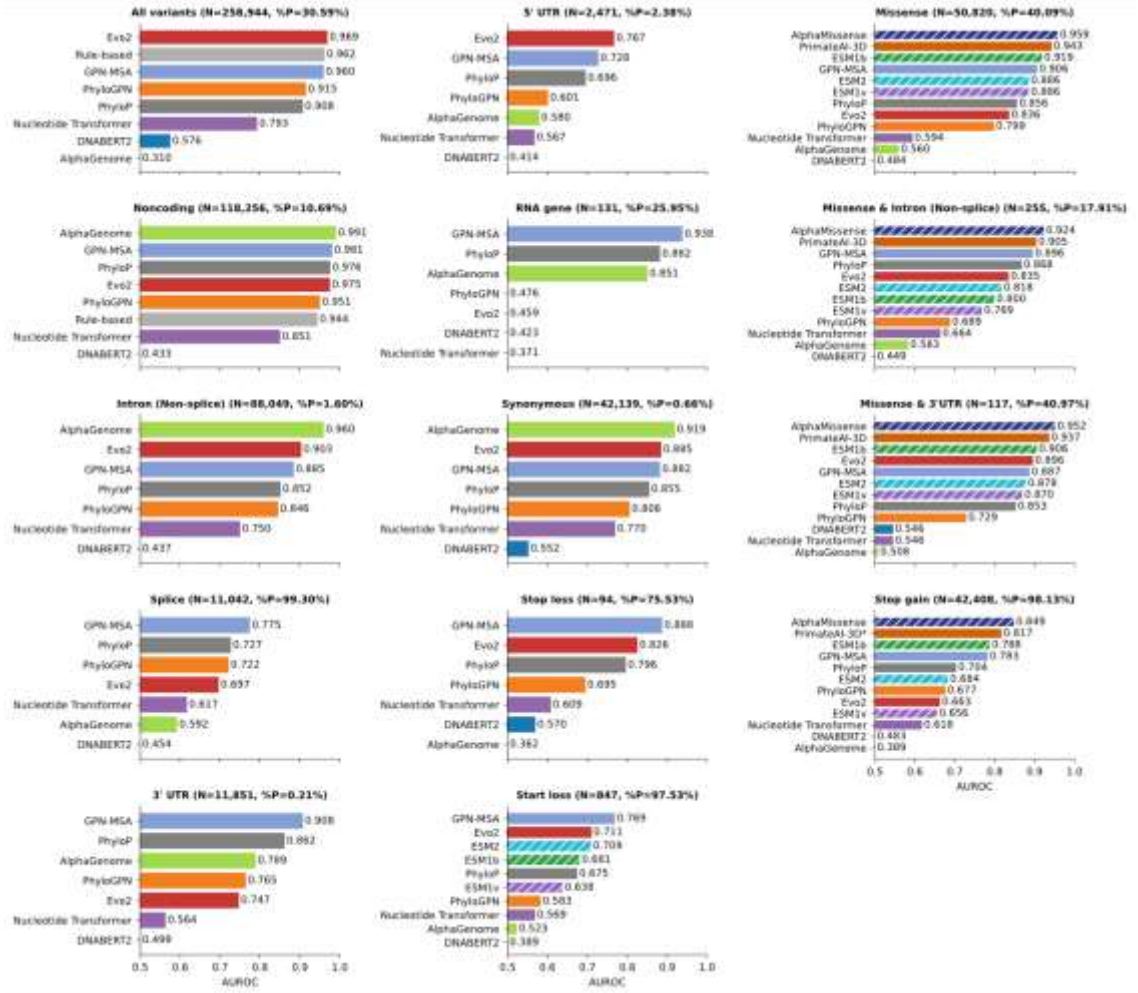
心得

臨床上，當我們用全基因定序找出一位病患的大量基因變異，我們需要區分哪一個變異和他的臨床疾病相關。而電腦預測模型，可以用來預測基因變異是否會造成人類疾病。得益於人工智慧近幾年的突破性發展，有許多基因語言模型被發表，這些模型把 DNA 的核甘酸序列和蛋白質的氨基酸序列，使用語言模型的訓練方式，以不同物種的序列以及大量的細胞、動物實驗室數據來進行訓練，進而能預測基因序列中每個位置，是不同核甘酸或胺基酸的機率。從演化上來看，越低機率的核甘酸代表越不容易在演化中被保留，也就是可能造成演化劣勢，甚至是臨床疾病，因此，我們可以使用基因語言模型，來預測基因變異的致病性[4]。

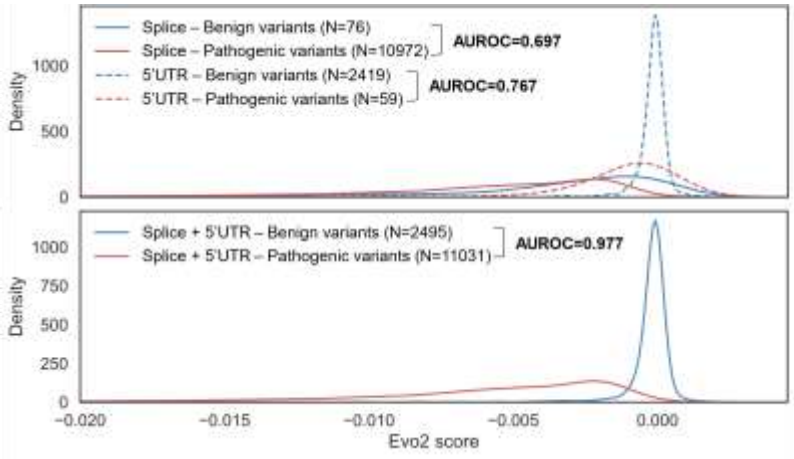
目前，有越來越多的基因語言模型被發表，且他們大都宣稱可以良好的預測基因、蛋白質序列中的各項機率，然而，每篇研究使用的測試基因變異資料集差異很大，到底哪一個模型真正最強呢？基於這個問題，我的第一個研究主題，是想要公平的比較各個不同的基因語言模型預測變異臨床致病性的能力，

我們發現，如果把各基因變異依照他所在的功能區域分類成子族群，如 5’不轉譯區域 (untranslated region)、外顯子 (exon) 、內含子 (intron) 等，各個子族群中致病力判斷最強的模型各自不同，因此，根據我們面對的基因變異子族群，最合適的作法是客製化選擇使用該子族群中表現最好的基因語言模型 (圖三) 。

更重要的是，基因語言模型雖然在原始的發表文章中都宣稱預測能力良好[2,3,5-13]，但在各個子族群中，模型在這些子分類中的預測能力時常大幅下降 (圖三) 。我們發現這個現象背後的原因，是統計學上的辛普森悖論 (Simpson paradox) ，因為基因語言模型學到的知識，偏向使用子族群分類，以及子族群中致病基因變異的比例來給出機率估計，舉例來說 (圖四) ，剪接基因變異 (splice variant) 大多都會造成疾病，5’不轉譯區域的變異大多不會造成疾病，而 Evo2[8] 語言模型會給所有剪接基因變異比較低的機率，也就是預測他們可能致病，給所有 5’不轉譯區域的變異比較高的機率，也就是比較不會致病，因此，儘管我們在 (圖四上半) 可以看到基因語言模型在剪接區域以及 5’不轉譯區域子族群內，致病基因變異和非致病基因的分數互相重疊很大，無法良好區分，致病性預測能力不佳，但當我們把剪接區域以及 5’不轉譯區域兩組合併時 (圖四下半) ，預測能力就會看起來有顯著提升。但這樣的高預測機率，根據使用者的需求，有時候是誤導性的，如果我們想判斷的是已經知道子族群的基因變異，建議使用該子族群中表現最好的人工智慧模型，而非整體表現最好的模型。



圖三：各基因語言模型在不同功能區域的變異致病性預測能力（AUROC: ROC 曲線下面積)



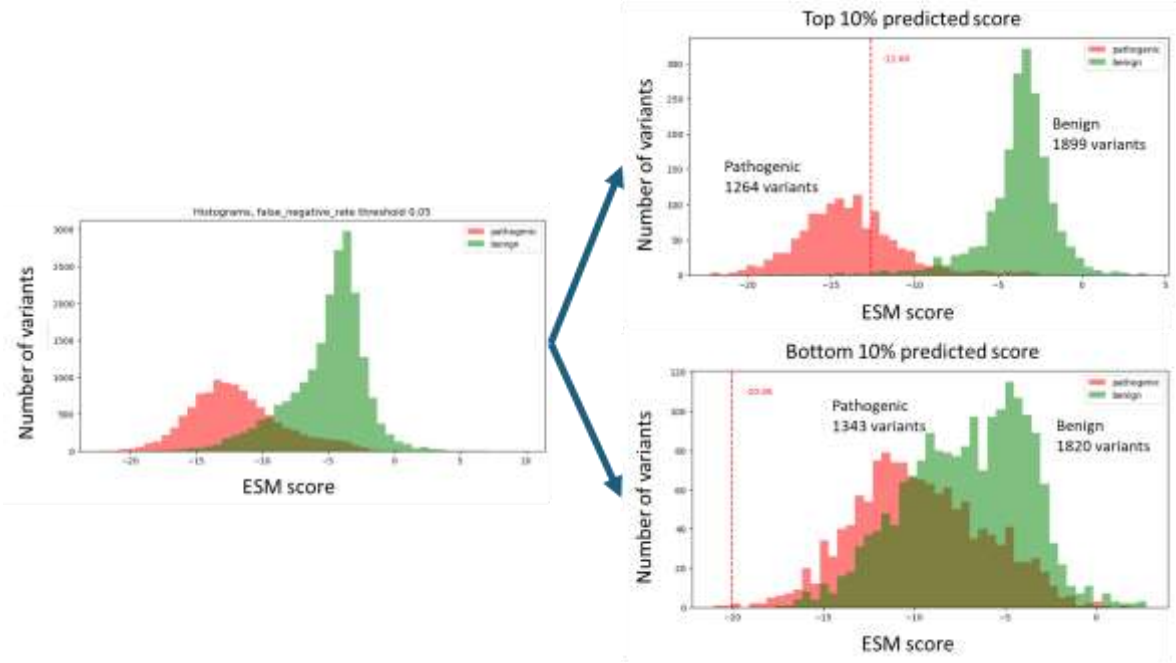
圖四：基因語言模型在剪接區域 (Splice)、5’不轉譯區域 (5’UTR)，以及兩組合併的基因變異致病性預測能力，我們可以看到，對於剪接區域的基因變異，絕大多數是致病性的，而不管變異是否致病，Evo2 都給了較低的分數，相較之下 5’不轉譯區域絕大多數是不致病的基因變異，而不論這類基因變異是否致病，Evo2 都給了較高的分數，因此當我們把剪接區域和 5’不轉譯區域合在一起時，致病和非致病的變異就被分開了。(AUROC: ROC 曲線下面積)

基因變異臨床致病性判斷的指引 (American College of Medical Genetics and Genomics / Association for Molecular Pathology guideline) 正在經歷數據化的重構[14]，從本來直覺化的致病性支持證據組合，變成基於貝氏定理的數學架構 (Bayesian framework)[15,16] 。電腦預測模型，包括基因語言模型，可以用來預測基因變異是否會造成人類疾病，提供判斷的支持證據。然而，在新的貝氏定理的數學架構下，所有的支持證據必須額外經過一個步驟，被校準轉換 (calibrate) 成對數似然比 (log-likelihood ratio) 的形式，才能使用。

目前診斷指引撰寫團隊發展出來的校準轉換的方法，只能一次校準一個預測工具[17]。但是，每個預測模型擅長的範圍不同，如前面 (研究一) 展現了不同的基因功能區域的表現不同，也有其他研究團隊發現比如不同的蛋白質結構區域[18,19]、甚至是不同的臨床疾病[20]，模型的表現也會不同，如何針對病患的情況，客製化的選擇要用哪個電腦預測模型，就成為留給臨床醫師面對的大問題，更糟的是，這個選擇必須在看到各個模型的預測結果之前就先進行，否則如果我們看了預測結果，然後選支持證據最強的，會因為先射箭才畫靶，破壞校準轉換的準確性。因此，我的後續兩個研究，主要都專注在解決這個臨床困境。

我想到的第一個解決方法，是找出各個模型不擅長預測的基因變異子族群，並且避免在這些子族群內避免使用表現不好的模型。更詳細的說，我先嘗試分析 ESM1b[2,3] 蛋白質語言模型在不同功能的基因、蛋白質功能區域預測基因變異臨床致病性的能力的變化，我們發現在不同基因、蛋白質序列中的不同位置、不同的胺基酸變化等變因，ESM1b 的預測能力都會不同，且這個能力起伏的模式相當複雜，難以歸納出簡單的原則。因此，我們建立了一個機器學習模型，用來預測 ESM1b 能準確分類的基因變異和不能準確分類致病性的基因變異。依照這個模型，我們就可以找出 ESM1b 能最準確區分致病性的基因變異，和最無法準確區分的基因變異 (圖五) ，使用這個機器學習的模型，我們就能在面對需要判讀的基因變異前 (當然也更是看到基因變異的模型預測結果前)，事先知道哪些基因變異的子族群適合或不適合使用 ESM1b 進行預測，如果我們對於各個預測模型的強項和弱點都有所了解，就能知道在臨床遇到一個病人時，依照他的情況要如何選擇合適的基因變異致病性預測模型。

本研究尚未發表，我計畫在回國後繼續分析其他預測模型的弱點子族群，然後建立一個教我們如何選擇預測模型的決策方法，來提供臨床使用的參考建議。

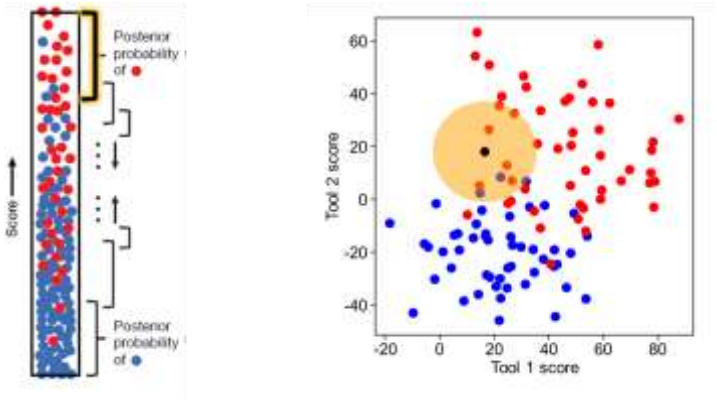


圖五：使用機器學習模型，區分 ESM1b 最能準確分類致病性的基因變異 (右上) ，和 ESM1b 最不能準確分類的基因變異 (右下)

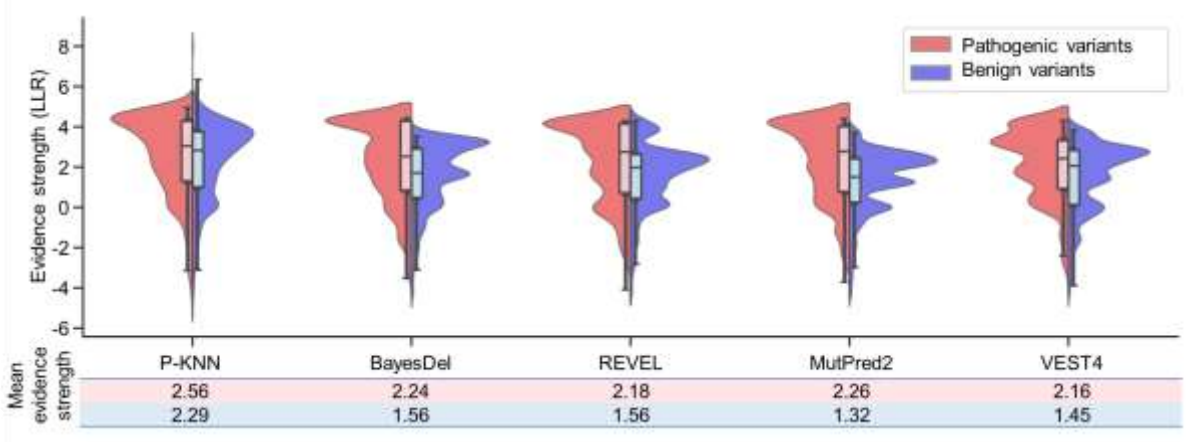
我想到的第二個解決辦法，是改變現行的校準轉換方式，既然一次只能校準轉換一個預測模型的限制另臨床苦惱，我們何不能一次共同校準 (jointly calibrate) 所有的預測模型呢？因此，我發展了一個共同校準多個預測工具的方法。

現行的校準方法，是使用一組已知會致病或不會致病的基因變異，把他們依照預測模型的分數排列 (圖六) ，接下來，對於預測模型的某個分數，我們可以藉由尋找這個分數附近的一群基因變異，並以其中會致病的變異的比例，作為這個分數轉換成的基因變異致病機率；相較之下，我發展的共同校準方法 (P-KNN, Pathogenicity-K Nearest Neighbors)，是把原本只用一個預測模型分數形成的一維空間，擴展成每個模型的分數定義一個維度的多維空間 (圖六右) ，偕著，與現行校準方法同樣的，我們使用一組已知致病性的基因變異散佈在這個多維空間中，然後當我們想知道一個新的基因變異是否致病時，就把她依照各個預測模型的分數放進這個空間中，找尋距離他最近的一群基因變異，然後以其中會致病的變異的比例，作為基因變異致病機率。

當我們嘗試使用 P-KNN 來共同校準之前診斷指引撰寫團隊校準過的 13 個預測模型，發現我們得到的證據強度，會比 13 個中任何一個預測模型都要更好 (圖七) ，顯示共同校準不只免去了臨床醫師需要選擇預測模型的困擾，同時也有效的整合了各個預測模型的強項。這項內容已經作為海報被美國人類遺傳學會的年會接受，正式論文也已經投稿，在同儕評審中。



圖六：現行校準轉換方法 (左) 與我發展的共同校準方法 (右) (紅點：已知致病的基因變異，藍點，已知不會致病的基因變異)



圖七：共同校準方法(P-KNN)提供的證據強度優於四個最好的預測模型 (LLR: 對數似然比)

建議事項

1. 建立友善使用者的算力：紐約大學的醫學資訊研究十分盛行，很大程度需要歸功於友善使用者的算力，友善展現在公平、免費、簡便三個方向 (見 2-4) ，我覺得都很值得學習。
2. 公平的算力：所有院內的研究者，只要申請帳號，都可以獲得算力的使用權，所有寫好的程式碼統一送出計算預約，並由排程系統依據之前的用量以及當下的需求決定運行的優先順序。
3. 免費的算力：在基本的運算資源使用下，使用者不需要額外付費。這裡的基本，其實一點都不差，我可以同時使用 8 張 NVIDIA A100 80GB 顯示卡連續數十個小時，都不需要額外付費，對於大型研究的測試非常有幫助。
4. 簡便的算力：除了命令列 (command line) 以外，也建立了簡單的使用者介面，讓寫程式基礎較弱的研究者也能快速上手設計自己的資料分析。
5. 減少取得資料的障礙：這個問題可以分使用方法、申請手續與費用兩方面來討論 (見

6-7) 。

6. 資料使用方法：若想要堅持資料不落地，就必須像英國生物資料庫 (UK biobank)一樣，在使用資料的平台上提供足夠的算力讓研究者免費使用或以合理價格租用。否則，除非允許研究者將申請的資料攜出，不然研究將無法進行。
7. 申請手續與費用：管理資料的各單位應該分工明確，比如醫療資料的隱私問題交給倫理審查委員會進行，資料交付及訂價則由資訊單位進行，兩者應盡量避免干涉對方的決策權，以避免雙重審查標準，令研究者無所適從。
8. 建立醫學院資訊相關的訓練課程：跨領域人才對於科技的應用至關重要，為了從醫學系跨進資訊的領域，建議在醫學系甚至是醫學院其他系建立相關的選修學程，或是輔修，來訓練出更能適應新世代的醫療人員。
9. 不要讓自然語言模型進入現實世界實作：**這是最重要的一點**，自然語言模型在預訓練時，常使用網路上抓取的資料，其中充滿仇恨的言論就構成模型的潛在思想，我建議所有語言模型的操作，都必須經過人類的審核。 (接 10)
10. 有許多研究 [21,22]，都顯示靠後續微調無法完全改進自然語言模型預訓練留下的問題，甚至模型還會反抗微調的效果，稍微再次調整，就重新展現出育訓練的仇恨思想(如果想要看實際例子，請看台大李宏毅老師授課錄影 https://www.youtube.com/watch?v=QLiKmca4kzI&t=4704s)。

附錄-參考資料

1 Lu, B., Liu, X., Lin, P.-Y. & Brandes, N. Genomic heterogeneity inflates the performance of variant pathogenicity predictions. *bioRxiv*, 2025.2009.2005.674459 (2025). https://doi.org:10.1101/2025.09.05.674459

2 Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat Genet* **55**, 1512-1522 (2023). https://doi.org:10.1038/s41588-023-01465-0

3 Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* **118** (2021). https://doi.org:10.1073/pnas.2016239118

4 Frazer, J. *et al.* Publisher Correction: Disease variant prediction with deep generative models of evolutionary data. *Nature* **601**, E7 (2022). https://doi.org:10.1038/s41586-021-04207-6

5 Gao, H. *et al.* The landscape of tolerated genetic variation in humans and primates. *Science* **380** (2023). https://doi.org:10.1126/science.abn8197

6 Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023). https://doi.org:10.1126/science.adg7492

7 Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021.2007.2009.450648 (2021). https://doi.org:10.1101/2021.07.09.450648

8 Brixi, G. *et al.* Genome modeling and design across all domains of life with Evo 2. (2025). https://doi.org:10.1101/2025.02.18.638918

9 Benegas, G., Albors, C., Aw, A. J., Ye, C. & Song, Y. S. A DNA language model based on multispecies alignment predicts the effects of genome-wide variants. *Nat Biotechnol* (2025). https://doi.org:10.1038/s41587-024-02511-w

10 Avsec, Ž. *et al.* AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model. *bioRxiv*, 2025.2006.2025.661532 (2025). https://doi.org:10.1101/2025.06.25.661532

11 Albors, C., Li, J. C., Benegas, G., Ye, C. & Song, Y. S. in *International Conference on Research in Computational Molecular Biology.* 99-117 (Springer).

12 Dalla-Torre, H. *et al.* Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat Methods* (2024). https://doi.org:10.1038/s41592-024-02523-z

13 Zhou, Z. *et al.* Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006* (2023).

14 Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405-424 (2015). https://doi.org:10.1038/gim.2015.30

15 Tavtigian, S. V., Harrison, S. M., Boucher, K. M. & Biesecker, L. G. Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines. *Hum Mutat* **41**, 1734-1737 (2020). https://doi.org:10.1002/humu.24088

16 Tavtigian, S. V. *et al.* Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet Med* **20**, 1054-1060 (2018). https://doi.org:10.1038/gim.2017.210

17 Pejaver, V. *et al.* Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am J Hum Genet* **109**, 2163-2177 (2022). https://doi.org:10.1016/j.ajhg.2022.10.013

18 Ramadane-Morchadi, L. *et al.* ACMG/AMP interpretation of BRCA1 missense variants: Structure-informed scores add evidence strength granularity to the PP3/BP4 computational evidence. *Am J Hum Genet* (2025). https://doi.org:10.1016/j.ajhg.2024.12.011

19 Luppino, F., Lenz, S., Chow, C. F. W. & Toth-Petroczy, A. Deep learning tools predict variants in disordered regions with lower sensitivity. *BMC Genomics* **26**, 367 (2025). https://doi.org:10.1186/s12864-025-11534-9

20 Anderson, D. & Lassmann, T. An expanded phenotype centric benchmark of variant prioritisation tools. *Hum Mutat* **43**, 539-546 (2022). https://doi.org:10.1002/humu.24362

21 Ji, J. *et al.* Language models resist alignment: Evidence from data compression. *arXiv preprint arXiv:2406.06144* (2024).

22 *Toward understanding and preventing misalignment generalization*, <https://openai.com/index/emergent-misalignment/?fbclid=IwY2xjawMygSVleHRuA2FlbQIxMQABHn1bkZ1yPh2GJ_oAL5hAdmEIVGlY1seMvP8eBc4ATlNT92tBSQrfo2man6uU_aem_9P9zH3M8KRZUISa3x83G3w> (