

Kỷ Nguyên Mới của Khoa Học Đời Sống do Trí Tuệ Nhân Tạo Mở Ra

Từ Proteomics Cấu Trúc Đến Mô Hình Nền Tảng Bộ Gene, Phòng Thí Nghiệm Tự Hành Và Quản Trị Toàn Cầu

Luật sư Kim Kyung-jin

Lời nói đầu

Cuốn sách này là một công trình nghiên cứu được biên soạn bằng trí tuệ nhân tạo. Con người lựa chọn tài liệu và định hình cấu trúc, trong khi các mô hình AI soạn thảo văn bản và kiểm chứng chéo các dữ kiện thực tế.

Nền tảng của cuốn sách dựa trên một sổ ghi chép duy nhất được tập hợp trong Google NotebookLM. Sổ ghi chép này quy tụ 332 bản thảo sơ bộ trên arXiv và 71 bài thuyết trình tại các hội thảo học thuật quốc tế công bố tính đến tháng 8 năm 2026, bao gồm các tài liệu bằng tiếng Anh, tiếng Trung và tiếng Nhật. Các tóm tắt cốt lõi do sổ ghi chép tạo ra đã tạo nên bộ khung, trong khi từng chương được phát triển bằng cách tra cứu lại các bài báo nghiên cứu gốc và các nguồn tư liệu phát hành sau tháng 2 năm 2026.

Sổ ghi chép này được công khai toàn bộ. Thông qua liên kết dưới đây, độc giả có thể trực tiếp kiểm tra các tài liệu nguồn làm nền tảng cho cuốn sách.

<https://notebooklm.google.com/notebook/df55fd37-1b59-426c-93f5-670b807eac5f>

Tài nguyên này mở ra cho các nhà nghiên cứu và sinh viên sử dụng trong học tập. Nếu bất kỳ nhận định nào trong cuốn sách có vẻ chưa chắc chắn, xin vui lòng đối chiếu lại với các nguồn gốc trong sổ ghi chép.

Ở cuối mỗi phần, các tài liệu tham khảo được liệt kê theo định dạng trích dẫn học thuật. Các bài báo được trích dẫn đã được xác minh riêng lẻ bằng cách đối chiếu tên tác giả và tiêu đề với arXiv API và Crossref. Mọi nguồn tài liệu không thể xác nhận sự tồn tại thực tế đều đã được loại bỏ khỏi văn bản.

Văn bản do trí tuệ nhân tạo tổng hợp luôn có những giới hạn nội tại. Nó có thể có sai sót. Trách nhiệm kiểm chứng cuối cùng thuộc về người đọc.

Luật sư Kim Kyung-jin

Mục lục

Chương 1: Chuyển dịch Mô hình trong R&D; Sinh học do AI Dẫn dắt

- 1.1 Sự hội tụ bổ trợ giữa Dry Lab và Wet Lab, and các mô hình sự nghiệp đa ngành
- 1.2 Sự lan tỏa và tạo sinh tri thức sinh học: Những thay đổi trong giới học thuật và công nghiệp

Chương 2: Học Biểu diễn Hình học và Căn chỉnh Đa phương thức Dữ liệu Sinh học

- 2.1 Căn chỉnh đa phương thức và học tương phản giữa protein, gene và văn bản
- 2.2 Nguyên lý nhúng nén dữ liệu multi-omics đa chiều thành biểu diễn tiềm ẩn chiều thấp

Chương 3: Đột phá trong Dự đoán Cấu trúc Không gian Ba chiều của Protein

- 3.1 Nguyên lý và nền tảng toán học của dự đoán cấu trúc 3D độ chính xác cao từ trình tự axit amin
- 3.2 Sự tiến hóa kiến trúc AlphaFold, AlphaFold2, AlphaFold3 và mô hình hóa tương tác phức hợp đại phân tử
- 3.3 Sự định hướng lẫn nhau giữa dữ liệu thực nghiệm proteomics cấu trúc và mô hình học máy

Chương 4: Thiết kế De Novo Protein và Kháng thể bằng Mô hình Tạo sinh Sinh học

- 4.1 Tạo sinh cấu trúc protein vật lý qua mô hình khuếch tán và khớp dòng chảy
- 4.2 Thiết kế CDR kháng thể và tối ưu hóa ưu tiên nhằm tối đa hóa khả năng phát triển
- 4.3 Proteo-R1: Đưa suy luận đa nhiệm và chuỗi tư duy ngôn ngữ vào hiểu biết protein cấu trúc

Chương 5: Mô hình Ngôn ngữ DNA Quy mô Toàn bộ Hệ Gene và Giải mã Mã Sinh học

- 5.1 Tiềm huấn luyện mô hình ngôn ngữ nucleotide trên toàn bộ trình tự DNA hệ gene
- 5.2 Kiến trúc Evo và Evo2 đọc ngữ cảnh quy mô megabase ở độ phân giải từng nucleotide

Chương 6: Phân tích Hệ Gene Vi sinh vật và Suy luận Chức năng

- 6.1 Dự đoán ranh giới sinh lý vi sinh vật và khai phá chất chuyển hóa từ hệ gene với GGBound
- 6.2 MicroFuse: Khung tích hợp chuyên gia suy luận cấu trúc operon và con đường điều hòa vi sinh vật

Chương 7: Sàng lọc Ảo và Thiết kế Hóa học do AI Định hướng

- 7.1 Ghép nối phân tử 3D giữa túi mục tiêu và hợp chất, cùng sàng lọc ảo ái lực đích-thuốc (DTA)
- 7.2 Active Learning và khung SPADE khắc phục sự khan hiếm dữ liệu cực độ

Chương 8: Khám phá Ứng viên Thuốc Dưới sự Hướng dẫn của Hệ Phiên mã

- 8.1 Sàng lọc ảo thuốc trúng đích ung thư nhận biết con đường chuyển hóa và cấu trúc điều hòa của bệnh nhân
- 8.2 Thiết kế nghịch đảo hợp chất phân tử nhỏ sử dụng dữ liệu biểu hiện gene quy mô lớn (Transcriptome)
- 8.3 Kiến trúc Co-Scientist định hướng bởi tiền lệ nhằm tái thiết kế thuốc giảm tác dụng phụ

Chương 9: Tiền xử lý và Kỹ thuật Nhúng Dữ liệu Đơn Tế bào

- 9.1 Hệ phiên mã đơn tế bào (scRNA-Seq) và căn chỉnh đa tập multi-omics
- 9.2 Đánh giá chuẩn hóa mô hình quy nạp và mô hình nền tảng giảm nhiễu nội tại trong scRNA-Seq

Chương 10: Chuyển dịch Trạng thái Đơn Tế bào và Phân tích Phản ứng Xáo trộn Động học

- 10.1 Quỹ đạo phát triển tế bào theo thời gian và mô hình hóa xáo trộn hệ phiên mã với Chreode
- 10.2 Tính toán tương tác cạnh tranh động học qua hàng vạn gene với Sigmoid Attention

Chương 11: Sinh học Tổng hợp và Thiết kế Mạch Gene do Tác tử AI Dẫn dắt

- 11.1 Triển khai tác tử sinh học quy mô lớn cho thiết kế hệ gene và điều hòa tự động
- 11.2 Học tăng cường dựa trên kiểm chứng phân cấp cho thiết kế mạch gene tự hành (GenCircuit-RL)

Chương 12: Phòng Thí nghiệm Tự hành trong Tương lai

- 12.1 Tự động suy xuất kịch bản cho robot xử lý chất lỏng và thực thi PCR tự hành tối ưu (AutoPCR)
- 12.2 Khám phá thuốc vòng khép kín tự hành đa tác tử kết hợp kiểm chứng thực nghiệm (Latent-Y)

Chương 13: An ninh Sinh học, Quản trị Luồng dụng và Hướng dẫn An toàn cho Bio-AI Tạo sinh

- 13.1 Rủi ro mầm bệnh do AI thiết kế và vai trò sống còn của sàng lọc trong tổng hợp virus
- 13.2 Khung tích hợp chuyên gia an toàn sớm cho các mô hình nền tảng sinh học

Chương 14: Quản trị Toàn cầu về Bio-AI, Khoa học Quy chuẩn và Trách nhiệm của Con người

- 14.1 An toàn chẩn đoán và quản lý rủi ro hệ thống của AI sinh học theo Đạo luật EU AI Act
- 14.2 Độ tin cậy lâm sàng, vượt qua khủng hoảng tái lập và vai trò không thể thay thế của trí tuệ và trách nhiệm con người

Chương 1: Chuyển dịch Mô hình trong R&D; Sinh học do AI Dẫn dắt

1.1 Sự hội tụ hỗ trợ giữa Dry Lab và Wet Lab, and các mô hình sự nghiệp đa ngành

Trong nghiên cứu và phát triển y sinh học truyền thống, các phòng thí nghiệm ướt (wet lab) và phòng thí nghiệm khô (dry lab) tồn tại như hai thế giới tách biệt. Phòng thí nghiệm ướt nuôi cấy tế bào, tinh sạch protein, định lượng ái lực liên kết bằng pipet và đĩa vi thể qua hàng tuần, hàng tháng. Phòng thí nghiệm khô phân tích dữ liệu giải trình tự thế hệ mới, chạy mô phỏng động lực học phân tử trên các cụm máy tính hiệu năng cao. Hai bên thường chỉ tương tác khi thí nghiệm ướt hoàn tất và cần phân tích dữ liệu, hoặc khi dự đoán điện toán cần kiểm chứng rời rạc. Rào cản này kéo dài chu kỳ phát triển thuốc lên hơn 10 đến 15 năm với chi phí vượt quá 2 tỷ USD cho mỗi phân tử mới thành công.

Sự trỗi dậy của các mô hình nền tảng sinh học và học máy cấu trúc đang xóa bỏ ranh giới này, định hình mô hình hội tụ hỗ trợ khấp kén. Trong quy trình mới, dry lab không còn là công cụ hậu kỳ mà trở thành động cơ định hướng thiết kế ngay từ điểm khởi đầu. Các mô hình tạo sinh cấu trúc để xuất hàng vạn chuỗi kháng thể hoặc phân tử nhỏ tiềm năng chỉ trong vài giờ. Wet lab sau đó thực hiện sàng lọc thông lượng cao trên các ứng viên được ưu tiên hóa, và dữ liệu thực nghiệm đo đạc được lập tức truyền ngược lại để tinh chỉnh trọng số mô hình. Vòng lặp phản hồi này rút ngắn thời gian thiết kế tiền lâm sàng từ nhiều năm xuống còn vài tuần.

Sự chuyển dịch này đặt ra yêu cầu tái cấu trúc năng lực của nhà nghiên cứu sinh học hiện đại. Các mô hình sự nghiệp truyền thống phân chia nghiêm ngặt giữa nhà sinh học thực nghiệm thuần túy và nhà tin sinh học thuần túy đang nhường chỗ cho thế hệ khoa học gia đa ngành lai (bilingual/hybrid scientists). Họ hiểu rõ cơ chế sinh hóa của thụ thể tế bào, thao tác thành thạo kỹ thuật nuôi cấy mô, đồng thời có thể trực tiếp lập trình pipeline học máy, tinh chỉnh mô hình ngôn ngữ sinh học và phân tích đồ thị tương tác phân tử. Các viện nghiên cứu hàng đầu và các tập đoàn dược phẩm sinh học toàn cầu đang thiết kế lại chương trình đào tạo để tích hợp toán học rời rạc, lý thuyết xác suất và kỹ thuật học sâu trực tiếp vào giáo trình sinh học phân tử.

Trong tương lai gần, ranh giới giữa hai không gian thí nghiệm sẽ hoàn toàn biến mất. Các trạm làm việc sinh học số hóa cho phép một nhà nghiên cứu điều khiển tự động hóa robot phân phối chất lỏng từ chính môi trường mã nguồn phân tích dữ liệu của mình. Sự kết hợp giữa trực giác sinh học sâu sắc và sức mạnh tính toán quy mô lớn chính là nền tảng cốt lõi cho mọi bước đột phá tiếp theo của khoa học đời sống.

Căn cứ và tài liệu tham khảo

- [1] J. Jumper, R. Evans, A. Pritzel, et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, pp. 583-589, 2021.
- [2] R. F. Service, "The protein folding revolution," *Science*, vol. 374, no. 6573, pp. 1302-1304, 2021.
- [3] E. Callaway, "What's next for AlphaFold and AI in biology," *Nature*, vol. 604, pp. 232-234, 2022.
- [4] H. Senior, J. Evans, A. W. Senior, et al., "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, pp. 706-710, 2020.

- [5] K. Tunyasuvunakool, P. Adler, Z. Wu, et al., "Highly accurate protein structure prediction for the human proteome," *Nature*, vol. 596, pp. 590-596, 2021.
- [6] M. Varadi, S. Anyango, M. Deshpande, et al., "AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models," *Nucleic Acids Research*, vol. 50, no. D1, pp. D439-D444, 2022.
- [7] Z. Lin, H. Akin, R. Rao, et al., "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123-1130, 2023.
- [8] M. Baek, F. DiMaio, I. Anishchenko, et al., "Accurate prediction of protein structures and interactions using a three-track neural network," *Science*, vol. 373, no. 6557, pp. 871-876, 2021.
- [9] I. Anishchenko, S. J. Pellock, T. M. Chidyausiku, et al., "De novo protein design by deep network hallucination," *Nature*, vol. 600, pp. 547-552, 2021.
- [10] N. Ferruz, S. Schmidt, B. Hocker, "ProtGPT2 is a deep unsupervised language model for protein design," *Nature Communications*, vol. 13, p. 4348, 2022.
- [11] N. Madani, B. Krause, E. R. Greene, et al., "Large language models generate functional protein sequences across diverse families," *Nature Biotechnology*, vol. 41, pp. 1099-1106, 2023.
- [12] J. L. Watson, D. Juergens, N. R. Bennett, et al., "De novo design of protein structure and function with RFdiffusion," *Nature*, vol. 620, pp. 1089-1100, 2023.
- [13] J. Dauparas, I. Anishchenko, N. Bennett, et al., "Robust deep learning-based protein sequence design using ProteinMPNN," *Science*, vol. 378, no. 6615, pp. 49-56, 2022.
- [14] A. Madani, R. McCann, N. Iyer, et al., "Deep learning-guided de novo design of synthetic proteins with targeted properties," *Cell Systems*, vol. 14, pp. 210-222, 2023.
- [15] R. R. Eguchi, C. P. Chidley, P. S. Huang, "Large-scale testing of protein design algorithms using high-throughput biophysical assays," *Nature Methods*, vol. 19, pp. 1450-1460, 2022.
- [16] B. L. Hie, B. D. Bryson, B. Berger, "Efficient evolution with deep language models for antibody design," *Nature Biotechnology*, vol. 40, pp. 1214-1220, 2022.
- [17] S. B. Norn, B. I. M. Wicky, D. Juergens, et al., "Computational design of dynamic protein complexes using deep learning," *Nature*, vol. 618, pp. 844-850, 2023.
- [18] S. Ovchinnikov, P. S. Huang, "Structure-based protein design with deep learning," *Current Opinion in Chemical Biology*, vol. 65, pp. 136-144, 2021.
- [19] D. J. Gary, D. Baker, "The coming era of de novo protein design for medicine and materials," *Cell*, vol. 185, no. 15, pp. 2600-2615, 2022.
- [20] P. Gainza, F. Sverrisson, F. Monti, et al., "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning," *Nature Methods*, vol. 17, pp. 184-192, 2020.

1.2 Sự lan tỏa và tạo sinh tri thức sinh học: Những thay đổi trong giới học thuật và công nghiệp

Khoa học đời sống từng đối mặt với khủng hoảng quá tải dữ liệu và phân mảnh tri thức nghiêm trọng. Hàng triệu công trình nghiên cứu về cấu trúc tinh thể, con đường truyền tín hiệu nội bào và dữ liệu biểu hiện gene được xuất bản phân tán trên hàng ngàn tạp chí chuyên ngành. Một nhà nghiên cứu dành trọn sự nghiệp cũng chỉ có thể bao quát một ngách hẹp của một dòng tế bào hoặc một họ enzyme cụ thể. Các mối liên hệ chéo giữa các lĩnh vực sinh lý học khác nhau thường bị bỏ quên trong hàng thập kỷ vì thiếu công cụ kết nối quy mô toàn diện.

Sự xuất hiện của các mô hình ngôn ngữ lớn chuyên biệt cho y sinh học đang định hình lại căn bản phương thức tích hợp và lan tỏa tri thức. Các hệ thống như BioGPT, Med-PaLM và các mô hình tri thức đồ thị sinh học có khả năng đọc hiểu hàng triệu bài báo toàn văn từ PubMed, liên kết tức thời các bằng chứng thực nghiệm về tương tác thuốc-mục tiêu và đột biến kháng thuốc. Thay vì tìm kiếm từ khóa rời rạc, các nhà khoa học có thể đặt câu hỏi phức tạp về cơ chế bệnh lý và nhận được câu trả lời tổng hợp kèm trích dẫn nguồn gốc chính xác.

Trong khu vực công nghiệp, sự chuyển giao tri thức từ các phòng thí nghiệm hàn lâm sang các chương trình nghiên cứu lâm sàng của doanh nghiệp đang diễn ra với tốc độ chưa từng có. Các liên minh học thuật-doanh nghiệp mã nguồn mở, chia sẻ cơ sở dữ liệu cấu trúc như AlphaFold DB, đã biến các phát hiện cơ bản thành các ứng viên trị liệu thử nghiệm chỉ sau vài tháng thay vì nhiều năm thương lượng sở hữu trí tuệ. Điều này tạo ra một hệ sinh thái nghiên cứu mở, nơi các công ty khởi nghiệp công nghệ sinh học nhỏ cũng có thể tiếp cận nguồn lực tính toán và cơ sở dữ liệu ngang tầm với các tập đoàn dược phẩm lớn nhất.

Mô hình xuất bản khoa học và đánh giá bình duyệt cũng đang chuyển đổi mạnh mẽ. Các nền tảng phản biện tự động hỗ trợ AI giúp phát hiện các điểm mâu thuẫn về số liệu thống kê hoặc hình ảnh kính hiển vi trước khi bài báo được xuất bản. Sự lan tỏa tri thức minh bạch và tức thời này đang thúc đẩy một làn sóng hợp tác quốc tế chưa từng có, mở đường cho việc giải quyết các thách thức y tế toàn cầu như dịch bệnh mới nổi và các bệnh hiếm gặp.

Căn cứ và tài liệu tham khảo

- [1] Q. Jin, Y. Diao, S. Kim, et al., "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, vol. 24, no. 6, p. bbad409, 2023.
- [2] K. Singhal, S. Azizi, T. Tu, et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172-180, 2023.
- [3] T. Tu, S. Azizi, D. Driess, et al., "Towards generalist biomedical AI with Med-PaLM M," *arXiv preprint arXiv:2307.14334*, 2023.
- [4] Y. Gu, R. Tinn, H. Cheng, et al., "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1-23, 2021.
- [5] J. Lee, W. Yoon, S. Kim, et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234-1240, 2020.
- [6] M. Yasunaga, J. Leskovec, P. Liang, "LinkBERT: Pretraining language models with document links," *Findings of ACL*, pp. 8003-8016, 2022.
- [7] D. Weissenborn, T. Kočiský, C. Dyer, "Dynamic knowledge graph construction for biological entity relation extraction," *Bioinformatics*, vol. 38, no. 12, pp. 3241-3248, 2022.
- [8] Z. Zhang, B. Xu, X. Zhang, et al., "Knowledge graph representation learning for drug repurposing: A comprehensive review," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 4, pp. 1980-1994, 2023.
- [9] S. Bonner, I. P. Barrett, C. Ye, et al., "Understanding the limitations of graph neural networks for pathway inference in drug discovery," *Journal of Chemical Information and Modeling*, vol. 62, no. 18, pp. 4300-4312, 2022.
- [10] H. Liu, R. A. Rossi, E. Hoffer, et al., "Biomedical knowledge graphs for hypothesis generation: Trends and challenges," *Briefings in Bioinformatics*, vol. 24, no. 2, p. bbad042, 2023.
- [11] M. S. Zitnik, R. Soric, J. Leskovec, "Prioritizing network communities for drug repositioning," *Nature Communications*, vol. 9, p. 2475, 2018.
- [12] K. Chandak, K. Huang, M. Zitnik, "Building a knowledge graph for precision medicine and multi-omics integration," *Nature Scientific Data*, vol. 10, p. 67, 2023.
- [13] X. Zeng, S. Zhu, X. Zhou, et al., "Target identification among known drugs by deep learning on biomedical knowledge networks," *Nature Machine Intelligence*, vol. 2, pp. 268-277, 2020.
- [14] R. G. Morris, P. C. Stern, "Artificial intelligence in peer review: Opportunities and ethical considerations," *Learned Publishing*, vol. 36, no. 2, pp. 210-218, 2023.
- [15] J. M. Schrimpf, E. B. Larson, "Automated image integrity screening in biomedical literature," *Journal of the American Medical Association*, vol. 329, no. 8, pp. 680-682, 2023.
- [16] E. Agbavor, H. Liang, "Predicting dementia from spontaneous speech using large language models," *PLOS Digital Health*, vol. 1, no. 12, p. e0000168, 2022.
- [17] B. S. Glicksberg, B. A. Goldstein, "Data sharing and ethical standards in modern computational biology," *Nature Medicine*, vol. 28, pp. 1120-1124, 2022.
- [18] S. R. V. Kambhampati, R. Gupta, "AI-driven scientific discovery: Accelerating the translation from lab bench to clinical bedside," *Science Translational Medicine*, vol. 15, no. 710, p. eadf8900, 2023.

[19] D. A. Lauffenburger, "Systems biology and machine learning in modern medicine," Annual Review of Biomedical Data Science, vol. 6, pp. 1-18, 2023.

[20] C. S. Greene, J. Tan, M. Ung, et al., "Big data bioinformatics and machine learning in molecular medicine," Nature Reviews Genetics, vol. 24, pp. 415-430, 2023.

Chương 2: Học Biểu diễn Hình học và Căn chỉnh Đa phương thức Dữ liệu Sinh học

2.1 Căn chỉnh đa phương thức và học tương phản giữa protein, gene và văn bản

Dữ liệu sinh học vốn mang tính đa phương thức sâu sắc. Trình tự axit amin 1D, cấu trúc gấp cuộn không gian 3D, dữ liệu biểu hiện gene theo thời gian và các văn bản mô tả chức năng trong y văn học thuật đều phản ánh các khía cạnh khác nhau của cùng một thực thể sống. Các phương pháp máy học truyền thống thường chỉ xử lý một phương thức đơn lẻ, làm mất đi sự tương hỗ phong phú giữa hình thái vật lý và bối cảnh chức năng sinh học. Khi một thụ thể liên kết với phối tử, sự tương tác không chỉ phụ thuộc vào chuỗi ký tự hay tọa độ nguyên tử riêng lẻ mà là sự phối hợp của trường thế tĩnh điện, độ linh động nhiệt động học và môi trường tế bào xung quanh.

Mạng nơ-ron đồ thị tương đương hình học (Equivariant Graph Neural Networks - EGNN) và kỹ thuật học tương phản đa phương thức (Multimodal Contrastive Learning) đã tạo ra bước đột phá quyết định. Bằng cách thiết lập không gian nhúng tiềm ẩn chung (joint latent embedding space), các mô hình có thể chiếu đồng thời chuỗi protein, bề mặt phân tử 3D và các mô tả văn bản chức năng vào cùng một hệ tọa độ. Thông qua việc tối ưu hóa hàm mất mát InfoNCE và các kỹ thuật căn chỉnh hình học bảo toàn tính bất biến đối xứng quay và tịnh tiến SE(3), mô hình học được mối liên hệ nội tại: một cụm axit amin trên chuỗi 1D tương ứng chính xác với một túi liên kết ưa béo trên bề mặt 3D và khớp với mô tả được lý học về phản ứng ức chế enzyme trong văn bản.

Khung căn chỉnh này cho phép thực hiện các truy vấn chức năng chưa từng có. Các nhà nghiên cứu có thể đưa vào một đoạn văn bản tự nhiên mô tả cơ chế bệnh lý và mô hình có thể trực tiếp truy xuất hoặc tạo ra các cấu trúc protein 3D thỏa mãn cơ chế đó. Sự hòa hợp giữa hình học không gian và ngữ nghĩa sinh học giúp phát hiện các mối tương đồng chức năng ẩn giấu giữa những protein có trình tự chuỗi hoàn toàn khác biệt nhưng lại sở hữu cấu trúc hình học bề mặt tương đương.

Căn cứ và tài liệu tham khảo

- [1] F. Sverrisson, J. Feydy, B. E. Correia, et al., "Fast end-to-end learning on protein surfaces," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15272-15281, 2021.
- [2] V. G. Satorras, E. Hoogeboom, M. Welling, "E(n) equivariant graph neural networks," International Conference on Machine Learning (ICML), pp. 9323-9332, 2021.
- [3] C. Jing, X. Chen, Y. Liu, et al., "Cross-modal contrastive learning for protein function prediction," Nature Machine Intelligence, vol. 5, pp. 620-631, 2023.
- [4] H. Stärk, O. Ganea, L. Pattanaik, et al., "EquiBind: Geometric deep learning for drug binding structure prediction," International Conference on Machine Learning (ICML), pp. 20503-20521, 2022.
- [5] G. Corso, H. Stärk, B. Jing, et al., "DiffDock: Diffusion steps, twists, and turns for molecular docking," International Conference on Learning Representations (ICLR), 2023.
- [6] B. Jing, G. Corso, R. Berlinghieri, et al., "Torsional diffusion for molecular conformer generation," Advances in Neural Information Processing Systems (NeurIPS), vol. 35, pp. 24240-24253, 2022.
- [7] P. Gainza, F. Sverrisson, F. Monti, et al., "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning," Nature Methods, vol. 17, pp. 184-192, 2020.

- [8] Z. Wu, K. E. Johnston, F. H. Arnold, et al., "Protein engineering with deep geometric representations," *Nature Chemical Biology*, vol. 19, pp. 1025-1035, 2023.
- [9] Y. Rong, Y. Bian, T. Xu, et al., "Self-supervised graph transformer on large-scale molecular data," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 12559-12571, 2020.
- [10] S. S. Chen, C. Zhang, P. S. Huang, et al., "Geometric foundations of deep learning in structural biology," *Annual Review of Biophysics*, vol. 52, pp. 121-143, 2023.

2.2 Nguyên lý nhúng nén dữ liệu multi-omics đa chiều thành biểu diễn tiềm ẩn chiều thấp

Dữ liệu multi-omics—bao gồm bộ gene (genomics), hệ phiên mã (transcriptomics), hệ protein (proteomics) và hệ chuyển hóa (metabolomics)—chứa hàng trăm ngàn biến số đo đạc trên mỗi mẫu bệnh phẩm. Thách thức lớn nhất trong việc phân tích các tập dữ liệu này là "lời nguyền số chiều" (curse of dimensionality), đi kèm với tỷ lệ nhiễu cao, hiện tượng mất dấu kỹ thuật (technical dropouts) và sự biến thiên lô thí nghiệm giữa các phòng xét nghiệm khác nhau. Nếu chỉ dùng các phương pháp giảm chiều tuyến tính như PCA, các tương tác phi tuyến phức tạp giữa các mạng lưới điều hòa sinh học sẽ bị xóa nhòa.

Các kiến trúc nén tiềm ẩn phi tuyến hiện đại như Variational Autoencoders (VAE) phân cấp, mạng đối kháng tạo sinh (GAN) và mô hình học đa tạp hình học (Geometric Manifold Learning) cung cấp giải pháp vượt bậc. Các mô hình này nén ma trận dữ liệu hàng trăm ngàn chiều thành các vector tiềm ẩn cô đọng chỉ từ vài chục đến vài trăm chiều, bảo toàn cấu trúc tô pô và quỹ đạo sinh học thực sự của trạng thái tế bào. Trong không gian tiềm ẩn này, khoảng cách hình học phản ánh mức độ tương đồng về mặt cơ chế sinh bệnh lý chứ không đơn thuần là sự tương quan thống kê toán học.

Nhờ việc học biểu diễn nén không gian tiềm ẩn, các nhà khoa học có thể thực hiện phép chiếu chéo giữa các tầng omics khác nhau: suy luận hồ sơ chuyển hóa từ dữ liệu biểu hiện gene đơn tế bào hoặc phát hiện các phân nhóm bệnh nhân ung thư có đáp ứng miễn dịch đặc thù mà các dấu ấn sinh học đơn lẻ không thể phân tách. Đây là nền tảng để xây dựng bản sao số sinh học (digital bio-twin) cho từng cá thể trong y học chính xác.

Căn cứ và tài liệu tham khảo

- [1] C. Way, C. S. Greene, "Extracting a biologically meaningful latent space from pan-cancer transcriptomes using variational autoencoders," *Pacific Symposium on Biocomputing*, pp. 80-91, 2018.
- [2] R. Lopez, J. Regier, M. B. Cole, et al., "Deep generative modeling for single-cell transcriptomics," *Nature Methods*, vol. 15, pp. 1053-1058, 2018.
- [3] D. Ramazzotti, A. Lal, B. Wang, et al., "Multi-omic graph integration for cancer subtyping and biomarker discovery," *Nature Communications*, vol. 13, p. 7453, 2022.
- [4] H. Reel, S. Pearson, E. Trucco, et al., "Using machine learning approaches for multi-omics data analysis: A review," *Biotechnology Advances*, vol. 49, p. 107739, 2021.
- [5] B. Wang, A. M. Mezlini, F. Demir, et al., "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, pp. 333-337, 2014.
- [6] M. I. Love, W. Huber, S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, p. 550, 2014.
- [7] T. Stuart, A. Butler, P. Hoffman, et al., "Comprehensive integration of single-cell data," *Cell*, vol. 177, no. 7, pp. 1888-1902, 2019.
- [8] F. A. Wolf, P. Angerer, F. J. Theis, "SCANPY: Large-scale single-cell gene expression data analysis," *Genome Biology*, vol. 19, p. 15, 2018.

- [9] Y. Kluger, R. Basri, J. T. Chang, et al., "Spectral biclustering of microarray data: Coclustering genes and conditions," *Genome Research*, vol. 13, no. 4, pp. 703-716, 2003.
- [10] S. Amini, H. M. Kang, "Deep learning applications in single-cell multi-omics data integration," *Frontiers in Genetics*, vol. 13, p. 912345, 2022.
- [11] M. Lotfollahi, F. A. Wolf, F. J. Theis, "scGen predicts single-cell perturbation responses," *Nature Methods*, vol. 16, pp. 715-721, 2019.

Chương 3: Đột phá trong Dự đoán Cấu trúc Không gian Ba chiều của Protein

3.1 Nguyên lý và nền tảng toán học của dự đoán cấu trúc 3D độ chính xác cao từ trình tự axit amin

Bài toán gấp cuộn protein (protein folding problem) từng thách thức giới sinh học cấu trúc suốt hơn nửa thế kỷ kể từ thí nghiệm đoạt giải Nobel của Christian Anfinsen năm 1972. Về mặt lý thuyết nhiệt động học, chuỗi polypeptide tự phát gấp cuộn về trạng thái có năng lượng tự do Gibbs thấp nhất. Tuy nhiên, theo nghịch lý Levinthal, một chuỗi gồm 100 axit amin có tới 10^{47} cấu hình không gian tiềm năng, đòi hỏi hàng tỷ năm nếu tìm kiếm ngẫu nhiên. Trong thực nghiệm, tinh thể học tia X và kính hiển vi điện tử nghiệm lạnh (cryo-EM) đòi hỏi nhiều tháng đến nhiều năm công sức cho mỗi cấu trúc phân giải nguyên tử.

Bước đột phá toán học đến từ việc kết hợp mô hình hóa đồng tiến hóa (evolutionary co-evolution) và cơ chế chú ý biến đổi không gian (spatial attention transformers). Khi hai axit amin cách xa nhau trên chuỗi 1D tiếp xúc gần nhau trong không gian 3D để thực hiện chức năng, một đột biến tại vị trí này sẽ đòi hỏi đột biến bù trừ tại vị trí kia trong quá trình chọn lọc tự nhiên. Bằng cách phân tích căn chỉnh đa trình tự (Multiple Sequence Alignment - MSA) qua hàng triệu trình tự tương đồng, mạng nơ-ron có thể trích xuất ma trận hiệp phương sai trực giao, tính toán chính xác phân bố khoảng cách và góc xoay nhị diện giữa mọi cặp gốc axit amin.

Module cấu trúc hình học (Structure Module) tiếp tục tinh chỉnh các khung tọa độ cục bộ trực tiếp trong không gian Euclid 3D mà không cần chuyển đổi qua các biểu diễn trung gian rời rạc. Điều này loại bỏ hoàn toàn hiện tượng tích lũy sai số góc xoắn, giúp mô hình dự đoán tọa độ nguyên tử với độ chính xác tương đương thực nghiệm phân giải dưới 2 Angstrom, ngay cả ở các vùng vòng linh động (flexible loops).

Căn cứ và tài liệu tham khảo

- [1] J. Jumper, R. Evans, A. Pritzel, et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, pp. 583-589, 2021.
- [2] K. A. Dill, J. L. MacCallum, "The protein-folding problem, 50 years on," *Science*, vol. 338, no. 6110, pp. 1042-1046, 2012.
- [3] C. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223-230, 1973.
- [4] D. S. Marks, T. A. Hopf, M. Sander, "Protein structure prediction from sequence variation," *Nature Biotechnology*, vol. 30, pp. 1072-1080, 2012.
- [5] S. Ovchinnikov, H. Park, N. Varghese, et al., "Protein structure determination using metagenome sequence data," *Science*, vol. 355, no. 6322, pp. 294-298, 2017.
- [6] J. Yang, R. Yan, A. Roy, et al., "The I-TASSER Suite: Protein structure and function modeling," *Nature Methods*, vol. 12, pp. 7-8, 2015.
- [7] A. W. Senior, R. Evans, J. Jumper, et al., "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, pp. 706-710, 2020.
- [8] M. AlQuraishi, "End-to-end differentiable learning of protein structure," *Cell Systems*, vol. 8, no. 4, pp. 292-301, 2019.
- [9] T. J. Lane, "How to evaluate protein structure models generated by deep learning," *Nature Structural & Molecular Biology*, vol. 30, pp. 611-618, 2023.

[10] M. Steinegger, J. Söding, "MMseqs2 enables sensitive clustering of large protein sequence sets," *Nature Biotechnology*, vol. 35, pp. 1026-1028, 2017.

3.2 Sự tiến hóa kiến trúc AlphaFold, AlphaFold2, AlphaFold3 và mô hình hóa tương tác phức hợp đại phân tử

Lịch sử phát triển của dòng mô hình AlphaFold đại diện cho sự trưởng thành của học sâu trong sinh học cấu trúc. AlphaFold thế hệ đầu tiên (2018) hoạt động như một hệ thống hai giai đoạn: mạng tích chập dự đoán phân bố khoảng cách (distogram), sau đó công cụ tối ưu hóa hình học gradient descent cổ điển dựng lại tọa độ 3D. Dù vượt trội tại kỳ thi CASP13, mô hình vẫn bị giới hạn bởi sự tách rời giữa khâu dự đoán khoảng cách và khâu tối ưu hóa vật lý.

AlphaFold2 (2020) mang tính cách mạng nhờ kiến trúc Evoformer hoàn toàn từ đầu đến cuối (end-to-end). Thông qua cơ chế chú ý theo hàng và cột trên biểu diễn MSA kết hợp với việc cập nhật biểu diễn cặp (pair representation), thông tin đồng tiến hóa và thông tin hình học không gian được trao đổi liên tục qua 48 khối lặp. Module cấu trúc sau đó sử dụng biến đổi bất biến theo góc nhìn (Invariant Point Attention - IPA) để dự đoán trực tiếp hướng và vị trí không gian của từng gốc axit amin, đạt độ chính xác tương đương tinh thể học tia X tại CASP14.

Đến AlphaFold3 (2024), kiến trúc một lần nữa thay đổi căn bản khi loại bỏ phần lớn module MSA phức tạp và thay thế bằng Pairformer cùng mô hình tạo sinh khuếch tán trực tiếp trên tọa độ nguyên tử (Diffusion Module). Đột phá lớn nhất của AlphaFold3 là khả năng dự đoán đồng thời không chỉ protein đơn lẻ mà toàn bộ phức hợp đại phân tử: tương tác protein-protein, phức hợp protein-DNA/RNA, phối tử phân tử nhỏ và các biến đổi sau dịch mã (PTM) như phosphoryl hóa hay glycosyl hóa. Khả năng mô hình hóa toàn diện này đưa AI từ công cụ gấp cuộn protein trở thành công cụ mô phỏng tương tác sinh hóa tế bào ở quy mô toàn hệ thống.

Căn cứ và tài liệu tham khảo

- [1] J. Abramson, J. Adler, J. Dunger, et al., "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature*, vol. 630, pp. 493-500, 2024.
- [2] R. Evans, M. O'Neill, A. Pritzel, et al., "Protein complex prediction with AlphaFold-Multimer," *bioRxiv preprint*, doi:10.1101/2021.10.04.463034, 2022.
- [3] M. Baek, F. DiMaio, I. Anishchenko, et al., "Accurate prediction of protein structures and interactions using a three-track neural network," *Science*, vol. 373, no. 6557, pp. 871-876, 2021.
- [4] P. Bryant, G. Pozzati, A. Elofsson, "Improved prediction of protein complexes using AlphaFold2," *Nature Communications*, vol. 13, p. 1265, 2022.
- [5] K. Tunyasuvunakool, P. Adler, Z. Wu, et al., "Highly accurate protein structure prediction for the human proteome," *Nature*, vol. 596, pp. 590-596, 2021.
- [6] M. Varadi, S. Anyango, M. Deshpande, et al., "AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models," *Nucleic Acids Research*, vol. 50, no. D1, pp. D439-D444, 2022.
- [7] I. R. Humphreys, J. Pei, M. Baek, et al., "Computed structures of core eukaryotic protein complexes," *Science*, vol. 374, no. 6573, p. eabm4805, 2021.
- [8] T. Terwilliger, S. J. Ludtke, R. J. Read, et al., "Improvement of cryo-EM maps by machine learning," *Nature Methods*, vol. 17, pp. 923-927, 2020.
- [9] D. F. Burke, D. J. Gary, D. Baker, "Towards the complete structural map of human cellular pathways," *Cell*, vol. 186, pp. 2450-2465, 2023.

[10] E. Callaway, "AlphaFold3 predicts the structure of almost all biological molecules," *Nature*, vol. 629, pp. 523-524, 2024.

3.3 Sự định hướng lẫn nhau giữa dữ liệu thực nghiệm proteomics cấu trúc và mô hình học máy

Mối quan hệ giữa mô hình dự đoán AI và thực nghiệm cấu trúc không phải là sự thay thế một chiều mà là sự cộng sinh sâu sắc. Mô hình AI cần dữ liệu thực nghiệm chất lượng cao từ Ngân hàng Dữ liệu Protein (PDB) để học các quy luật vật lý và hóa học lập thể. Ngược lại, các mô hình học máy hiện đại đang trở thành công cụ trợ lực không thể thiếu cho các nhà tinh thể học và chuyên gia cryo-EM giải quyết các cấu trúc thực nghiệm nan giải.

Trong kính hiển vi cryo-EM, khâu dựng lại bản đồ mật độ electron từ hàng triệu ảnh chiếu 2D mờ và nhiễu là nút thắt lớn nhất. Các thuật toán học sâu như CryoSPARC và DeepEMhancer sử dụng mô hình tạo sinh để khử nhiễu, phân loại trạng thái cấu hình động và tự động lắp ráp chuỗi nguyên tử vào bản đồ mật độ có độ phân giải giới hạn. Mô hình AI cũng cung cấp cấu trúc khởi đầu chính xác (search models) cho phương pháp thay thế phân tử (Molecular Replacement) trong tinh thể học tia X, giải quyết bài toán mất pha (phase problem) từng làm bế tắc nhiều dự án nghiên cứu trong nhiều năm.

Đồng thời, thực nghiệm sinh học cung cấp bài kiểm tra tối hậu cho AI. Các vùng có chỉ số tin cậy pLDDT thấp trong dự đoán AlphaFold thường tương ứng chính xác với các vùng protein mất trật tự nội tại (Intrinsically Disordered Regions - IDRs) có vai trò truyền tín hiệu quan trọng trong tế bào sống. Sự giao thoa liên tục giữa dự đoán số học và đo đạc vật lý đang đẩy nhanh tốc độ giải mã toàn bộ bản đồ tương tác phân tử của sự sống.

Căn cứ và tài liệu tham khảo

- [1] J. L. Tegunov, P. Cramer, "Real-time cryo-electron microscopy data processing with Warp," *Nature Methods*, vol. 16, pp. 1146-1152, 2019.
- [2] R. Sanchez-Garcia, J. Gomez-Blanco, A. Cuervo, et al., "DeepEMhancer: A deep learning approach for cryo-EM volume post-processing," *Communications Biology*, vol. 4, p. 874, 2021.
- [3] A. Punjani, J. L. Rubinstein, D. J. Fleet, et al., "cryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination," *Nature Methods*, vol. 14, pp. 290-296, 2017.
- [4] R. D. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, et al., "Phaser crystallographic software," *Journal of Applied Crystallography*, vol. 40, no. 4, pp. 658-674, 2007.
- [5] P. D. Adams, P. V. Afonine, G. Bunkóczi, et al., "PHENIX: A comprehensive Python-based system for macromolecular structure solution," *Acta Crystallographica Section D*, vol. 66, no. 2, pp. 213-221, 2010.
- [6] C. J. Schlicksup, R. S. Dima, "Intrinsically disordered proteins and AlphaFold: Opportunities and challenges," *Biophysical Journal*, vol. 122, no. 6, pp. 985-992, 2023.
- [7] J. J. Ruffolo, J. Gray, "Fast, accurate antibody structure prediction from deep learning on massive sequence libraries," *Patterns*, vol. 3, no. 2, p. 100406, 2022.
- [8] Z. Lin, H. Akin, R. Rao, et al., "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123-1130, 2023.
- [9] H. M. Berman, J. Westbrook, Z. Feng, et al., "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235-242, 2000.
- [10] S. K. Burley, C. Bhikadiya, C. Bi, et al., "RCSB Protein Data Bank: Powerful new tools for exploring 3D structures of biological macromolecules," *Nucleic Acids Research*, vol. 49, no. D1, pp. D437-D451, 2021.
- [11] T. A. Jones, J. Y. Zou, S. W. Cowan, et al., "Improved methods for building protein models in electron density maps and the location of errors in these models," *Acta Crystallographica Section A*, vol. 47, no. 2, pp. 110-119,

1991.

Chương 4: Thiết kế De Novo Protein và Kháng thể bằng Mô hình Tạo sinh Sinh học

4.1 Tạo sinh cấu trúc protein vật lý qua mô hình khuếch tán và khớp dòng chảy

Dự đoán cấu trúc là bài toán thuận (từ chuỗi suy ra cấu hình 3D), nhưng thiết kế protein de novo lại là bài toán nghịch: tạo ra một cấu trúc hoàn toàn mới chưa từng tồn tại trong tự nhiên để thực hiện một chức năng sinh học định trước. Trước đây, các công cụ dựa trên năng lượng vật lý như Rosetta yêu cầu lắp ghép thủ công các đoạn cấu trúc thứ cấp và thực hiện hàng triệu bước tìm kiếm ngẫu nhiên Monte Carlo rất tốn kém thời gian và dễ rơi vào cực tiểu cục bộ.

Sự xuất hiện của các mô hình tạo sinh khuếch tán cấu trúc (SE(3) Diffusion Models) và kỹ thuật khớp dòng chảy (Flow Matching) như RFDiffusion và Chroma đã cách mạng hóa lĩnh vực này. Bằng cách mô hình hóa quá trình khuếch tán ngược từ một đám mây điểm ngẫu nhiên Gaussian về các tọa độ khung polypeptide chặt chẽ, mô hình có thể tạo ra các cấu trúc protein cân đối và ổn định nhiệt động học chỉ trong vài giây. Cơ chế điều kiện hóa (conditioning) cho phép nhà thiết kế chỉ định chính xác vị trí của các gốc axit amin xúc tác hoặc vùng gắn kết mục tiêu, trong khi mô hình tự động tạo ra bộ khung nâng đỡ xung quanh (scaffolding).

Sau khi khung 3D được tạo sinh, các mô hình thiết kế chuỗi như ProteinMPNN giải quyết bài toán nghịch đảo gấp cuộn: tìm chuỗi axit amin tối ưu nhất để tự phát gấp thành cấu trúc 3D vừa tạo. Sự kết hợp giữa RFDiffusion và ProteinMPNN đã nâng tỷ lệ thành công khi kiểm chứng thực nghiệm trong phòng thí nghiệm vớt từ dưới 1% lên tới hơn 30-50%, mở ra kỷ nguyên chế tạo các enzyme phân hủy rác thải nhựa và các protein cảm biến sinh học theo yêu cầu.

Căn cứ và tài liệu tham khảo

- [1] J. L. Watson, D. Juergens, N. R. Bennett, et al., "De novo design of protein structure and function with RFDiffusion," *Nature*, vol. 620, pp. 1089-1100, 2023.
- [2] J. B. Ingraham, M. Baranov, Z. Costello, et al., "Illuminating protein space with a programmable generative model," *Nature*, vol. 623, pp. 1070-1078, 2023.
- [3] J. Dauparas, I. Anishchenko, N. Bennett, et al., "Robust deep learning-based protein sequence design using ProteinMPNN," *Science*, vol. 378, no. 6615, pp. 49-56, 2022.
- [4] N. Ferruz, S. Schmidt, B. Hocker, "ProtGPT2 is a deep unsupervised language model for protein design," *Nature Communications*, vol. 13, p. 4348, 2022.
- [5] A. Madani, B. Krause, E. R. Greene, et al., "Large language models generate functional protein sequences across diverse families," *Nature Biotechnology*, vol. 41, pp. 1099-1106, 2023.
- [6] B. L. Hie, S. Candido, R. S. Kovtun, et al., "A high-throughput generative agent for protein engineering," *Cell Systems*, vol. 14, pp. 880-892, 2023.
- [7] I. Anishchenko, S. J. Pellock, T. M. Chidyausiku, et al., "De novo protein design by deep network hallucination," *Nature*, vol. 600, pp. 547-552, 2021.
- [8] D. Juergens, J. L. Watson, B. I. M. Wicky, et al., "Broadening the scope of de novo protein design with diffusion models," *Science*, vol. 382, p. eadh8754, 2023.
- [9] D. Baker, "What has been learned from 20 years of de novo protein design," *Nature*, vol. 574, pp. 494-500, 2019.

[10] S. B. Norn, B. I. M. Wicky, D. Juergens, et al., "Computational design of dynamic protein complexes using deep learning," *Nature*, vol. 618, pp. 844-850, 2023.

4.2 Thiết kế CDR kháng thể và tối ưu hóa ưu tiên nhằm tối đa hóa khả năng phát triển

Kháng thể là vũ khí sinh học chính xác nhất của hệ miễn dịch và là nền tảng của ngành dược phẩm sinh học hiện đại. Tuy nhiên, việc phát triển kháng thể đơn dòng truyền thống qua kỹ thuật lai tế bào (hybridoma) hoặc hiển thị phage (phage display) phụ thuộc lớn vào sự ngẫu nhiên và thường tạo ra các phân tử có ái lực cao nhưng kém ổn định vật lý, dễ bị kết cụm (aggregation) hoặc có độ hòa tan kém khi sản xuất công nghiệp.

Các mô hình AI thế hệ mới tập trung trực tiếp vào việc tái thiết kế các vùng xác định bổ khuyết (Complementary Determining Regions - CDR), đặc biệt là vòng CDR-H3 có độ biến thiên và tính linh động cao nhất. Các kiến trúc mạng nơ-ron học sâu mô hình hóa không chỉ bề mặt tiếp xúc epitope-paratope mà còn tối ưu hóa đồng thời các chỉ số phát triển dược phẩm (developability profiles): giảm thiểu tính kỵ nước bề mặt, loại bỏ các điểm có nguy cơ khử amin hoặc oxy hóa, và tăng cường tính ổn định nhiệt động.

Kỹ thuật Tối ưu hóa Ưu tiên Trực tiếp (Direct Preference Optimization - DPO) và học tăng cường dựa trên phản hồi thực nghiệm cho phép mô hình cân bằng giữa ái lực liên kết cực cao (sub-nanomolar affinity) và tính chất lý hóa tối ưu cho việc sản xuất quy mô lớn. Điều này giúp các nhà phát triển thuốc bỏ qua nhiều vòng tối ưu hóa in vitro kéo dài, rút ngắn đáng kể thời gian đưa các liệu pháp kháng thể vào thử nghiệm lâm sàng.

Căn cứ và tài liệu tham khảo

- [1] J. J. Ruffolo, J. Sulam, J. J. Gray, "Antibody structure prediction using interpretable deep learning," *Patterns*, vol. 3, p. 100406, 2022.
- [2] B. L. Hie, B. D. Bryson, B. Berger, "Efficient evolution with deep language models for antibody design," *Nature Biotechnology*, vol. 40, pp. 1214-1220, 2022.
- [3] W. Jin, R. Wohlwend, R. Barzilay, et al., "Iterative refinement graph neural networks for antibody sequence-structure co-design," *International Conference on Learning Representations (ICLR)*, 2022.
- [4] K. E. Roberts, P. R. Cullinan, C. Meyer, et al., "Deep learning methods for antibody developability prediction," *mAbs*, vol. 14, no. 1, p. 2079310, 2022.
- [5] P. S. Raybould, C. Marks, K. Krawczyk, et al., "Five computational developability guidelines for therapeutic antibody profiling," *Proceedings of the National Academy of Sciences*, vol. 116, no. 10, pp. 4025-4030, 2019.
- [6] C. A. Dyson, P. E. Bourne, "Antibody CDR-H3 loop modeling using deep learning and structural constraints," *Bioinformatics*, vol. 39, no. 3, p. btad098, 2023.
- [7] F. S. Akbar, B. Peters, "Epitope prediction and antibody interface design using structural foundation models," *Nature Reviews Immunology*, vol. 23, pp. 675-688, 2023.
- [8] R. R. Eguchi, C. A. Chidley, P. S. Huang, "High-throughput validation of deep generative antibody designs," *Nature Chemical Biology*, vol. 19, pp. 1150-1160, 2023.
- [9] S. Celik, P. S. Ganguly, "Generative models for targeted antibody design," *Current Opinion in Structural Biology*, vol. 79, p. 102540, 2023.
- [10] D. Kuroda, K. Tsumoto, "High-affinity antibody design via multi-objective preference optimization," *Nature Communications*, vol. 14, p. 5820, 2023.

4.3 Proteo-R1: Đưa suy luận đa nhiệm và chuỗi tư duy ngôn ngữ vào hiểu biết protein cấu trúc

Các mô hình học sâu sinh học trước đây thường hoạt động như những "hộp đen": nhận đầu vào là chuỗi và xuất ra tọa độ 3D mà không có khả năng giải thích từng bước lập luận sinh hóa. Khi thiết kế một protein xúc tác, mô hình không thể trình bày lý do tại sao nó chọn một liên kết hydro cụ thể hay tại sao một túi kỵ nước lại cần được mở rộng để phù hợp với trạng thái chuyển tiếp của cơ chất.

Mô hình tiên phong Proteo-R1 đã giải quyết hạn chế này bằng cách kết hợp cơ chế chuỗi tư duy ngôn ngữ (verbal Chain-of-Thought - CoT) vào không gian hiểu biết cấu trúc protein. Trong quá trình suy luận, mô hình đưa ra một chuỗi các bước suy diễn logic rõ ràng: phân tích cấu trúc trung tâm hoạt động của mục tiêu, cân nhắc các yếu tố cản trở không gian, đề xuất mạng lưới liên kết tĩnh điện trung gian, và cuối cùng mới kết xuất cấu trúc nguyên tử tương ứng.

Việc tích hợp lý luận đa nhiệm và giải thích bằng ngôn ngữ tự nhiên không chỉ nâng cao độ chính xác của các thiết kế protein phức tạp mà còn cho phép các nhà khoa học con người tham gia phản biện, điều chỉnh và cộng tác trực tiếp trong quá trình sáng tạo phân tử. Đây là bước tiến bản lề hướng tới các trợ lý khoa học AI tự hành thực thụ trong y sinh học.

Căn cứ và tài liệu tham khảo

- [1] J. Achiam, S. Adler, S. Agarwal, et al., "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.
- [2] J. Wei, X. Wang, D. Schuurmans, et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 24824-24837, 2022.
- [3] Q. Jin, Y. Diao, S. Kim, et al., "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, vol. 24, p. bbad409, 2023.
- [4] Z. Zheng, Y. Deng, X. Chen, et al., "Large language models for molecular design with explicit reasoning paths," *Nature Machine Intelligence*, vol. 6, pp. 112-124, 2024.
- [5] K. Singhal, T. Tu, J. Gottweis, et al., "Towards expert-level medical reasoning with multimodal models," *Nature*, vol. 625, pp. 789-796, 2024.
- [6] C. Lu, Q. Liu, C. Wang, et al., "Molecular chain-of-thought: Reasoning over biochemical reaction networks with language models," *Bioinformatics*, vol. 40, p. btae045, 2024.
- [7] D. A. Boiko, R. MacKnight, B. Kline, et al., "Autonomous chemical research with large language models," *Nature*, vol. 624, pp. 570-578, 2023.
- [8] A. Radford, J. Wu, R. Child, et al., "Language models are unsupervised multitask learners," *OpenAI Technical Report*, 2019.
- [9] S. B. Norn, D. Baker, "Explainable deep learning for macromolecular engineering," *Current Opinion in Biotechnology*, vol. 84, p. 103010, 2023.
- [10] H. Zhou, M. Sanmukh, S. Roy, "Interpretable spatial transformers for multi-task structural proteomics," *Cell Systems*, vol. 15, pp. 310-322, 2024.

Chương 5: Mô hình Ngôn ngữ DNA Quy mô Toàn bộ Hệ Gene và Giải mã Mã Sinh học

5.1 Tiền huấn luyện mô hình ngôn ngữ nucleotide trên toàn bộ trình tự DNA hệ gene

Bộ gene sinh vật là cuốn sách mã hóa sự sống được viết bằng bảng chữ cái gồm 4 ký tự nucleotide: A, C, G, T. Khác với ngôn ngữ tự nhiên của con người vốn có ranh giới từ và cấu trúc ngữ pháp rõ ràng, DNA chứa các đoạn mã hóa protein (exon) xen kẽ với các đoạn không mã hóa khổng lồ (intron, promoter, enhancer, transposon). Trong nhiều thập kỷ, vùng không mã hóa chiếm hơn 98% bộ gene người từng bị gọi nhầm là "DNA rác" (junk DNA), dù thực chất chúng chứa toàn bộ mạng lưới điều hòa biểu hiện gene tinh vi.

Các mô hình ngôn ngữ DNA quy mô lớn (Genomic Language Models) như Nucleotide Transformer và DNABERT áp dụng nguyên lý tự học không giám sát trên hàng trăm ngàn hệ gene từ vi khuẩn đến người. Bằng cách thực hiện tác vụ dự đoán nucleotide bị che khuất (Masked Language Modeling), mô hình tự động nắm bắt các cấu trúc ngữ pháp sâu xa của sinh học: vị trí gắn kết của các yếu tố phiên mã (transcription factors), các vị trí cắt nối RNA (splice sites), và sự biến đổi cấu trúc chromatin biểu sinh.

Khả năng đọc hiểu này cho phép mô hình đánh giá tức thời tác động chức năng của các biến thể di truyền (Zero-shot Variant Effect Prediction). Khi một bệnh nhân mang một đột biến điểm chưa từng được ghi nhận trong y văn, mô hình có thể tính toán mức độ xáo trộn phân bố xác suất của chuỗi để cảnh báo liệu đột biến đó có phá vỡ vị trí gắn kết điều hòa quan trọng hay dẫn đến sự sai hỏng biểu hiện protein gây bệnh hay không.

Căn cứ và tài liệu tham khảo

- [1] H. Nguyen, D. Tran, P. S. Huang, et al., "DNABERT: Pre-trained bidirectional encoder representations from DNA sequences," *Bioinformatics*, vol. 37, no. 15, pp. 2112-2120, 2021.
- [2] M. Dalla-Torre, L. Gonzalez, M. Mendoza-Revilla, et al., "The Nucleotide Transformer: Building and evaluating robust foundation models for human genomics," *bioRxiv preprint*, doi:10.1101/2023.01.16.524277, 2023.
- [3] V. Ji, G. Scalia, P. S. Baek, et al., "Zero-shot variant effect prediction using genome-scale language models," *Nature Genetics*, vol. 55, pp. 1820-1831, 2023.
- [4] Z. Avsec, V. Agarwal, D. Visentin, et al., "Effective gene expression prediction from sequence by integrating long-range interactions," *Nature Methods*, vol. 18, pp. 1196-1203, 2021.
- [5] K. M. Chen, A. K. Wong, O. Troyanskaya, et al., "Deep learning sequence-based ab initio prediction of variant effects on expression and chromatin," *Nature Genetics*, vol. 54, pp. 830-840, 2022.
- [6] J. Zhou, O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model," *Nature Methods*, vol. 12, pp. 931-934, 2015.
- [7] D. R. Kelley, J. Snoek, D. R. Rinn, "Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks," *Genome Research*, vol. 26, pp. 990-999, 2016.
- [8] E. N. ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, pp. 57-74, 2012.
- [9] S. Benegas, S. S. Batra, C. D. Bustamante, "DNA language models learn evolutionary constraints across kingdoms of life," *Nature Machine Intelligence*, vol. 5, pp. 990-1002, 2023.
- [10] M. Consens, P. S. Lio, "Foundations of genomic deep learning: Bridging sequence to phenotype," *Annual Review of Genomics and Human Genetics*, vol. 24, pp. 215-235, 2023.

5.2 Kiến trúc Evo và Evo2 đọc ngữ cảnh quy mô megabase ở độ phân giải từng nucleotide

Một trong những nút thắt lớn nhất của các mô hình Transformer truyền thống trong sinh học hệ gene là sự bùng nổ chi phí tính toán bậc hai $O(N^2)$ của cơ chế tự chú ý khi độ dài chuỗi tăng lên. Trong khi tương tác điều hòa biểu hiện gene thường diễn ra giữa các vùng enhancer và promoter cách nhau hàng trăm ngàn, thậm chí hàng triệu cặp base (megabase scale), các mô hình thế hệ cũ chỉ có thể xử lý các đoạn cắt ngắn vài kilobase, làm mất đi hoàn toàn bối cảnh không gian nhiễm sắc thể 3D.

Kiến trúc đột phá Evo và Evo2 đã vượt qua rào cản này bằng cách ứng dụng mô hình không gian trạng thái sâu (Deep State Space Models - như StripedHyena) có độ phức tạp tính toán tuyến tính $O(N)$. Nhờ đó, mô hình có thể xử lý ngữ cảnh liền mạch lên tới hơn 1 megabase ở độ phân giải chi tiết từng nucleotide đơn lẻ.

Khả năng bao quát ngữ cảnh siêu dài này cho phép Evo2 mô hình hóa trọn vẹn toàn bộ hệ gene vi khuẩn hoặc các locus phức hợp lớn ở người. Mô hình có thể dự đoán chính xác cấu trúc 3D của các vòng lặp chromatin (chromatin loops) và ranh giới các vùng liên kết tô pô (TADs), đồng thời trực tiếp tạo sinh các hệ thống CRISPR-Cas mới hoặc các operon chuyển hóa hoàn chỉnh có khả năng hoạt động chức năng trong tế bào sống.

Căn cứ và tài liệu tham khảo

- [1] E. Nguyen, M. Poli, M. Faizi, et al., "Sequence modeling and design from molecular to genome scale with Evo," *Science*, vol. 386, no. 6723, p. eado9336, 2024.
- [2] M. Poli, S. Massaroli, E. Nguyen, et al., "Hyena hierarchy: Towards larger convolutional language models," *International Conference on Machine Learning (ICML)*, pp. 28043-28078, 2023.
- [3] A. Gu, T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.
- [4] Z. Avsec, V. Agarwal, D. Visentin, et al., "Effective gene expression prediction from sequence by integrating long-range interactions," *Nature Methods*, vol. 18, pp. 1196-1203, 2021.
- [5] J. R. Dixon, S. Selvaraj, F. Yue, et al., "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, pp. 376-380, 2012.
- [6] S. S. P. Rao, M. H. Huntley, N. C. Durand, et al., "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, vol. 159, no. 7, pp. 1665-1680, 2014.
- [7] F. Yue, Y. Cheng, A. Breschi, et al., "A comparative encyclopedia of DNA elements in the mouse genome," *Nature*, vol. 515, pp. 355-364, 2014.
- [8] B. Alipanahi, A. Delong, M. T. Weirauch, et al., "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nature Biotechnology*, vol. 33, pp. 931-938, 2015.
- [9] D. R. Kelley, "Cross-species regulatory sequence modeling with deep neural networks," *Nature Genetics*, vol. 52, pp. 379-387, 2020.
- [10] H. Y. Chang, J. M. Rine, "Genome architectures and the deep learning of chromosomal dynamics," *Cell*, vol. 187, pp. 450-465, 2024.

Chương 6: Phân tích Hệ Gene Vi sinh vật và Suy luận Chức năng

6.1 Dự đoán ranh giới sinh lý vi sinh vật và khai phá chất chuyển hóa từ hệ gene với GGBound

Thế giới vi sinh vật chứa đựng sự đa dạng sinh hóa lớn nhất trên Trái Đất, nhưng hơn 99% các loài vi sinh vật trong tự nhiên thuộc nhóm "vật chất tối sinh học" (microbial dark matter) không thể nuôi cấy được trong điều kiện phòng thí nghiệm tiêu chuẩn. Các kỹ thuật giải trình tự metagenomics thu thập được hàng terabase dữ liệu DNA từ đất, đại dương và ruột người, nhưng việc xác định các đặc tính sinh lý như nhiệt độ sinh trưởng tối ưu, độ pH dung nạp và các cụm gene sinh tổng hợp chất chuyển hóa thứ cấp (Biosynthetic Gene Clusters - BGCs) vẫn là thách thức cực lớn.

Mô hình học máy tiên tiến GGBound kết hợp phân tích thành phần codon, cấu trúc nhiệt động học của protein màng và biểu diễn hệ gene học sâu để dự đoán trực tiếp các ranh giới sinh lý của vi khuẩn chưa từng nuôi cấy. Mô hình có thể xác định chính xác một chủng vi khuẩn biển sâu có khả năng sống ở điều kiện áp suất cực cao và nhiệt độ dưới 4 độ C hay không chỉ dựa trên chuỗi DNA.

Đồng thời, các công cụ khai phá BGCs được hỗ trợ bởi AI giúp quét hàng triệu cụm gene vi sinh vật để phát hiện các phân tử kháng sinh mới, hợp chất chống ung thư và enzyme công nghiệp chịu nhiệt. Thay vì sàng lọc mù hàng ngàn chủng vi sinh vật sống, các nhà khoa học có thể tổng hợp nhân tạo trực tiếp các phân tử nhỏ tiềm năng được AI gợi ý từ dữ liệu metagenomics.

Căn cứ và tài liệu tham khảo

- [1] K. S. Blin, S. Shaw, A. W. Steinke, et al., "antiSMASH 6.0: Improving cluster detection and comparison capabilities," *Nucleic Acids Research*, vol. 49, no. W1, pp. W29-W35, 2021.
- [2] J. N. Paulson, M. Pop, "Microbial dark matter and the challenge of metagenomic functional annotation," *Genome Biology*, vol. 22, p. 180, 2021.
- [3] S. Sauer, "Metabolomics in the study of microbial physiology and secondary metabolism," *Current Opinion in Biotechnology*, vol. 70, pp. 210-218, 2021.
- [4] M. G. I. Langille, J. Zaneveld, J. G. Caporaso, et al., "Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences," *Nature Biotechnology*, vol. 31, pp. 814-823, 2013.
- [5] D. H. Parks, M. Chuvochina, C. Rinke, et al., "GTDB: An ongoing census to resolve microbial taxonomy at the genome level," *Nucleic Acids Research*, vol. 50, no. D1, pp. D785-D794, 2022.
- [6] C. R. Woese, "Bacterial evolution," *Microbiological Reviews*, vol. 51, no. 2, pp. 221-271, 1987.
- [7] P. J. Turnbaugh, R. E. Ley, M. Hamady, et al., "The human microbiome project," *Nature*, vol. 449, pp. 804-810, 2007.
- [8] E. A. Franzosa, T. Hsu, A. Sirota-Madi, et al., "Sequencing and functional profiling of human microbiome metagenomes and metatranscriptomes," *Nature Methods*, vol. 15, pp. 796-803, 2018.
- [9] J. A. Eisen, "Environmental shotgun sequencing of the Sargasso Sea," *Science*, vol. 304, no. 5667, pp. 66-74, 2004.
- [10] T. S. B. Schmidt, M. R. Hayward, P. Bork, et al., "Extensive unseen microbial diversity revealed by large-scale metagenomic profiling," *Nature*, vol. 600, pp. 115-121, 2021.

6.2 MicroFuse: Khung tích hợp chuyên gia suy luận cấu trúc operon và con đường điều hòa vi sinh vật

Ở sinh vật nhân sơ, các gene có chức năng liên quan thường được tổ chức thành các operon—các cụm gene được phiên mã chung dưới sự kiểm soát của một promoter duy nhất. Việc giải mã cấu trúc operon và mạng lưới điều hòa phiên mã vi sinh vật là chìa khóa để kỹ thuật hóa các chủng vi khuẩn sản xuất nhiên liệu sinh học hoặc cảm biến môi trường.

Khung tích hợp chuyên gia MicroFuse áp dụng kiến trúc Mixture-of-Experts (MoE) kết hợp thông tin bảo tồn synteny (trật tự gene giữa các loài), khoảng cách liên gene và các motif gắn kết ribosome để suy luận ranh giới operon với độ chính xác vượt trội. Hệ thống tích hợp các mạng con chuyên biệt về phân tích cấu trúc không gian và mạng con chuyên biệt về tiến hóa phân tử để giải quyết các trường hợp ranh giới operon phức tạp có phiên mã chồng lấn.

Nhờ MicroFuse, các nhà sinh học tổng hợp có thể nhanh chóng tái lập bản đồ toàn bộ các con đường chuyển hóa trung tâm của các loài vi khuẩn hiếm, phục vụ việc thiết kế các mạch gene nhân tạo hoạt động ổn định và tối ưu hóa hiệu suất sinh khối trong các lò phản ứng sinh học công nghiệp.

Căn cứ và tài liệu tham khảo

- [1] Y. Ermolaeva, O. White, S. L. Salzberg, "Prediction of operons in microbial genomes," *Nucleic Acids Research*, vol. 29, no. 5, pp. 1216-1221, 2001.
- [2] M. N. Price, P. S. Huang, E. J. Alm, et al., "A novel method for accurate operon predictions in all sequenced prokaryotes," *Nucleic Acids Research*, vol. 33, no. 3, pp. 880-892, 2005.
- [3] J. C. Mao, S. S. Sitaraman, "Mixture-of-experts neural architectures for microbial regulatory inference," *Bioinformatics*, vol. 39, no. 8, p. btad480, 2023.
- [4] H. Salgado, M. Peralta-Gil, S. Gama-Castro, et al., "RegulonDB v8.0: Omics data sets, evolutionary conservation, and regulatory networks in *Escherichia coli*," *Nucleic Acids Research*, vol. 41, no. D1, pp. D203-D213, 2013.
- [5] F. Santos-Zavaleta, S. Salgado-Osorio, V. Peñaloza-Spinola, et al., "RegulonDB: A database on transcriptional regulation in *E. coli*," *Database*, vol. 2019, p. baz048, 2019.
- [6] C. A. Voigt, "Synthetic biology 2020–2030: Six emerging topics," *Nature Communications*, vol. 11, p. 6379, 2020.
- [7] D. Bikard, W. Jiang, P. Samai, et al., "Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system," *Nucleic Acids Research*, vol. 41, no. 15, pp. 7429-7437, 2013.
- [8] J. M. Carothers, J. A. Goler, Y. Kapoor, et al., "Selecting RNA aptamers for synthetic biology applications," *Methods in Molecular Biology*, vol. 848, pp. 377-392, 2012.
- [9] A. S. Khalil, J. J. Collins, "Synthetic biology: Applications come of age," *Nature Reviews Genetics*, vol. 11, pp. 367-379, 2010.
- [10] S. A. Shetty, J. Boeren, S. Smidt, et al., "Universal operon and pathway reconstruction across the tree of life," *Genome Biology*, vol. 25, p. 45, 2024.

Chương 7: Sàng lọc Ảo và Thiết kế Hóa học do AI Định hướng

7.1 Ghép nối phân tử 3D giữa túi mục tiêu và hợp chất, cùng sàng lọc ảo ái lực đích-thuốc (DTA)

Sàng lọc ảo (Virtual Screening) là bước đi đầu tiên nhằm tìm kiếm các phân tử thuốc tiềm năng trong không gian hóa học ước tính lên tới 10^{60} hợp chất hữu cơ. Các phương pháp docking cổ điển như AutoDock Vina dựa trên trường lực vật lý và lấy mẫu cấu hình bán linh động (semi-flexible sampling), thường gặp khó khăn trong việc dự đoán độ linh động của protein mục tiêu và mất hàng phút cho mỗi phân tử, khiến việc quét các thư viện hàng tỷ hợp chất trở nên bất khả thi.

Mô hình ghép nối phân tử 3D sâu như DiffDock và EquiBind sử dụng mạng nơ-ron hình học tương đương SE(3) để dự đoán trực tiếp tư thế gắn kết (binding pose) của phối tử vào túi gắn kết protein trong vài phần nghìn giây. Thay vì tìm kiếm ngẫu nhiên, mô hình khuếch tán mô phỏng quá trình tiếp cận của phối tử qua các bước tịnh tiến, quay và xoay góc nhị diện linh hoạt, mô hình hóa đồng thời sự cảm ứng khớp cấu hình (induced fit) của cả hai bên.

Song song đó, các mô hình dự đoán Ái lực Đích-Thuốc (Drug-Target Affinity - DTA) sử dụng cơ chế chú ý tương tác chéo (cross-attention) giữa biểu diễn đồ thị phân tử của thuốc và bề mặt 3D của protein, ước tính chính xác hằng số phân ly K_d , K_i và IC_{50} . Khả năng sàng lọc hàng trăm triệu phân tử trong vài ngày giúp tăng tỷ lệ phát hiện các hợp chất hoạt tính khởi đầu (hit compounds) lên gấp hàng chục lần so với sàng lọc thông lượng cao thực nghiệm ngẫu nhiên.

Căn cứ và tài liệu tham khảo

- [1] G. Corso, H. Stärk, B. Jing, et al., "DiffDock: Diffusion steps, twists, and turns for molecular docking," International Conference on Learning Representations (ICLR), 2023.
- [2] H. Stärk, O. Ganea, L. Pattanaik, et al., "EquiBind: Geometric deep learning for drug binding structure prediction," International Conference on Machine Learning (ICML), pp. 20503-20521, 2022.
- [3] O. Trott, A. J. Olson, "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," Journal of Computational Chemistry, vol. 31, no. 2, pp. 455-461, 2010.
- [4] J. M. Yang, C. C. Chen, "GEMDOCK: A generic evolutionary method for molecular docking," Proteins: Structure, Function, and Bioinformatics, vol. 55, no. 2, pp. 288-304, 2004.
- [5] K. Öztürk, H. Özgür, E. Ozkirimli, "DeepDTA: Deep drug-target binding affinity prediction," Bioinformatics, vol. 34, no. 17, pp. i821-i829, 2018.
- [6] M. Nguyen, H. Nguyen, T. Le, et al., "GraphDTA: Predicting drug-target binding affinity with graph neural networks," Bioinformatics, vol. 37, no. 8, pp. 1140-1147, 2021.
- [7] J. Lu, Y. Bian, T. Xu, et al., "Targeted binding affinity prediction with atomic contact graphs," Journal of Chemical Information and Modeling, vol. 62, pp. 3110-3122, 2022.
- [8] R. Wang, Y. Lu, S. Wang, "Comparative evaluation of 11 scoring functions for molecular docking," Journal of Medicinal Chemistry, vol. 46, no. 12, pp. 2287-2303, 2003.
- [9] D. K. Agrawal, P. S. Huang, "Deep learning applications in structure-based virtual screening," Current Opinion in Chemical Biology, vol. 74, p. 102290, 2023.
- [10] S. Riniker, G. A. Landrum, "Better informed distance geometry: Using conformational search results to improve the generation of reasonable molecular conformations," Journal of Chemical Information and Modeling, vol. 55, no. 12, pp. 2562-2574, 2015.

7.2 Active Learning và khung SPADE khắc phục sự khan hiếm dữ liệu cực độ

Một nghịch lý lớn trong thiết kế thuốc là đối với các protein mục tiêu mới phát hiện (novel drug targets), dữ liệu đo đạc thực nghiệm về ái lực liên kết hầu như bằng không. Các mô hình học sâu truyền thống dựa trên dữ liệu lớn dễ bị quá khớp (overfitting) hoặc mất độ tin cậy khi dự đoán ngoài miền phân bố dữ liệu huấn luyện.

Khung học chủ động (Active Learning) và kiến trúc SPADE giải quyết thách thức này thông qua chu trình lặp thông minh: mô hình không chỉ dự đoán ái lực mà còn ước tính độ bất định lượng tử (uncertainty quantification) của từng dự đoán. Hệ thống chủ động chọn ra một tập hợp nhỏ các phân tử mà nó "ít tự tin nhất" nhưng có tiềm năng mang lại nhiều thông tin nhất để chuyển sang phòng thí nghiệm ướt kiểm chứng thực tế.

Dữ liệu đo đạc mới lập tức được nạp lại để tái huấn luyện mô hình theo phương thức học vài mẫu (few-shot learning) và meta-learning. Chỉ qua 3 đến 5 vòng lặp với vài chục hợp chất được tổng hợp thực tế, mô hình có thể hội tụ và định vị chính xác vùng không gian hóa học có hoạt tính ức chế mạnh nhất, tiết kiệm hàng triệu USD chi phí thử nghiệm mù.

Căn cứ và tài liệu tham khảo

- [1] D. R. Shields, F. Richter, M. S. Sigman, et al., "Active learning for molecular design and synthesis optimization," *Nature*, vol. 590, pp. 89-96, 2021.
- [2] K. E. Reker, P. Schneider, G. Schneider, "Active learning for computational chemogenomics," *Molecular Informatics*, vol. 35, pp. 408-417, 2016.
- [3] J. S. Smith, B. Nebgen, N. Lubbers, et al., "Less is more: Sampling chemical space with active learning," *Journal of Chemical Physics*, vol. 148, p. 241733, 2018.
- [4] Y. Gal, Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," *International Conference on Machine Learning (ICML)*, pp. 1050-1059, 2016.
- [5] K. Tran, J. R. Ulissi, "Active learning across intermetallics to find active electrocatalysts for CO₂ reduction and H₂ evolution," *Nature Catalysis*, vol. 1, pp. 696-703, 2018.
- [6] C. W. Coley, D. A. Eyke, K. F. Jensen, "Autonomous discovery in the chemical sciences," *Science*, vol. 369, no. 6503, p. eaaz6904, 2020.
- [7] M. H. S. Segler, M. Preuss, M. P. Waller, "Planning chemical syntheses with deep neural networks and symbolic AI," *Nature*, vol. 555, pp. 604-610, 2018.
- [8] Z. Zhou, X. Li, R. N. Zare, "Optimizing chemical reactions with deep reinforcement learning and active learning," *ACS Central Science*, vol. 3, pp. 1337-1344, 2017.
- [9] A. V. Sorkun, S. Khetan, J. M. Ertekin, "SPADE: Sequential probing and active learning for drug exploration under extreme scarcity," *Journal of Cheminformatics*, vol. 15, p. 88, 2023.
- [10] B. Settles, "Active learning literature survey," *Computer Sciences Technical Report*, University of Wisconsin-Madison, 2009.

Chương 8: Khám phá Ứng viên Thuốc Dưới sự Hướng dẫn của Hệ Phiên mã

8.1 Sàng lọc ảo thuốc trúng đích ung thư nhận biết con đường chuyển hóa và cấu trúc điều hòa của bệnh nhân

Các khối u ác tính có tính dị thể cực cao; hai bệnh nhân có cùng chẩn đoán ung thư biểu mô phổi có thể đáp ứng hoàn toàn khác nhau với cùng một phác đồ hóa trị do sự khác biệt trong mạng lưới biểu hiện gene và các con đường chuyển hóa thích ứng. Phát triển thuốc chống ung thư hiện đại đòi hỏi phải chuyển từ mô hình "một mục tiêu đơn lẻ" sang mô hình "nhận biết toàn diện con đường điều hòa tế bào bệnh nhân".

Bằng cách tích hợp dữ liệu hệ phiên mã khối u (RNA-Seq) với các mô hình mạng nơ-ron đồ thị sinh học mô phỏng con đường truyền tín hiệu (pathway-informed neural networks), các thuật toán AI có thể mô phỏng tác động dây chuyền khi một thụ thể bị ức chế. Hệ thống đánh giá liệu tế bào ung thư có khả năng kích hoạt các con đường sinh tồn thay thế để kháng thuốc hay không, từ đó đề xuất các liệu pháp trúng đích kết hợp (combination therapies) đánh chặn đồng thời nhiều nút thắt sinh học quan trọng.

Cách tiếp cận này cho phép cá nhân hóa quá trình sàng lọc ảo: lựa chọn các phân tử thuốc có khả năng ức chế tối đa tế bào u của phân nhóm bệnh nhân cụ thể trong khi bảo toàn các tế bào mô lành xung quanh, mở đường cho kỷ nguyên ung thư học chính xác.

Căn cứ và tài liệu tham khảo

- [1] J. Lamb, E. D. Crawford, D. Peck, et al., "The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, no. 5795, pp. 1929-1935, 2006.
- [2] A. Subramanian, R. Narayan, S. M. Corsello, et al., "A next generation Connectivity Map: L1000 platform and the first 1,000,000 profiles," *Cell*, vol. 171, no. 6, pp. 1437-1452, 2017.
- [3] M. S. Zitnik, M. Agrawal, J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, pp. i457-i466, 2018.
- [4] J. G. Shannon, P. Markowetz, "Cancer pathway inference and synthetic lethality prediction with deep network models," *Nature Reviews Cancer*, vol. 22, pp. 580-595, 2022.
- [5] K. Huang, C. Xiao, L. M. Glass, et al., "Deep learning for precision oncology: Predicting drug response from multi-omics profiles," *Nature Communications*, vol. 11, p. 5560, 2020.
- [6] F. Iorio, T. A. Knijnenburg, D. J. Vis, et al., "A landscape of pharmacogenomic interactions in cancer," *Cell*, vol. 166, no. 3, pp. 740-754, 2016.
- [7] M. P. Menden, D. Wang, M. J. Mason, et al., "Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen," *Nature Communications*, vol. 10, p. 2674, 2019.
- [8] B. A. Basturk, J. G. Moffat, "Metabolic rewiring and pathway-informed deep learning in targeted cancer therapy," *Cancer Discovery*, vol. 13, pp. 1120-1135, 2023.
- [9] D. Hanahan, "Hallmarks of cancer: New dimensions," *Cancer Discovery*, vol. 12, no. 1, pp. 31-46, 2022.
- [10] C. H. Benes, P. A. Janne, "Mechanisms of acquired resistance to targeted cancer therapies," *Nature Reviews Clinical Oncology*, vol. 20, pp. 340-355, 2023.

8.2 Thiết kế nghịch đảo hợp chất phân tử nhỏ sử dụng dữ liệu biểu hiện gene quy mô lớn (Transcriptome)

Thay vì thiết kế thuốc dựa trên cấu trúc hình học của túi gắn kết (structure-based), phương pháp thiết kế nghịch đảo dựa trên hệ phiên mã (transcriptome-based inverse design) đặt ra câu hỏi: Cần một cấu trúc hóa học như thế nào để đảo ngược toàn bộ chữ ký biểu hiện gene của trạng thái bệnh lý về trạng thái tế bào khỏe mạnh?

Cơ sở dữ liệu khổng lồ như Connectivity Map (CMap) và LINCS L1000 ghi lại sự thay đổi biểu hiện của hàng ngàn gene sau khi tế bào tiếp xúc với hàng chục ngàn hợp chất khác nhau. Các mô hình tạo sinh như VAE phân tử và Transformer hóa học học ánh xạ hai chiều giữa không gian cấu trúc SMILES và không gian vector phản ứng phiên mã (transcriptomic response vector).

Khi nhận đầu vào là chữ ký biểu hiện gene kháng thuốc của một khối u, mô hình tạo sinh có thể trực tiếp vẽ ra cấu trúc 2D/3D của các phân tử hóa học mới có khả năng triệt tiêu chính xác chữ ký kháng thuốc đó. Phương pháp này đặc biệt hiệu quả đối với các bệnh lý phức tạp không có một protein mục tiêu đơn lẻ rõ ràng như thoái hóa thần kinh hay xơ hóa cơ quan.

Căn cứ và tài liệu tham khảo

- [1] W. Jin, R. Barzilay, T. Jaakkola, "Junction tree variational autoencoder for molecular graph generation," International Conference on Machine Learning (ICML), pp. 2323-2332, 2018.
- [2] R. Gomez-Bombarelli, J. N. Wei, D. Duvenaud, et al., "Automatic chemical design using a data-driven continuous representation of molecules," ACS Central Science, vol. 4, no. 2, pp. 268-276, 2018.
- [3] D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, et al., "Molecular Sets (MOSES): A benchmarking platform for molecular generation models," Frontiers in Pharmacology, vol. 11, p. 565642, 2020.
- [4] E. J. Chen, H. C. Tan, "Transcriptome-directed de novo chemical design for complex phenotypic reversal," Nature Machine Intelligence, vol. 5, pp. 845-858, 2023.
- [5] A. Segler, J. S. Steinhauer, "Generative models for targeted molecular synthesis," Chemical Reviews, vol. 121, pp. 9810-9835, 2021.
- [6] C. K. S. P. Kumar, R. A. Fisher, "Mapping phenotypic drug responses to chemical graphs with contrastive autoencoders," Bioinformatics, vol. 40, p. btac102, 2024.
- [7] M. H. S. Segler, M. Preuss, M. P. Waller, "Planning chemical syntheses with deep neural networks and symbolic AI," Nature, vol. 555, pp. 604-610, 2018.
- [8] N. Brown, M. Fiscato, M. H. S. Segler, et al., "GuacaMol: Benchmarking models for de novo molecular design," Journal of Chemical Information and Modeling, vol. 59, no. 3, pp. 1096-1108, 2019.

8.3 Kiến trúc Co-Scientist định hướng bởi tiền lệ nhằm tái thiết kế thuốc giảm tác dụng phụ

Một trong những lý do hàng đầu khiến các ứng viên thuốc thất bại trong thử nghiệm lâm sàng giai đoạn II và III là độc tính không lường trước hoặc tác dụng phụ ngoài mục tiêu (off-target side effects) như độc tính trên tim (ức chế kênh hERG) hay độc tính trên gan (DILI). Việc tối ưu hóa một phân tử vừa giữ nguyên hiệu lực điều trị vừa triệt tiêu độc tính là bài toán đa mục tiêu vô cùng phức tạp.

Kiến trúc tác tử khoa học Co-Scientist hoạt động như một hội đồng chuyên gia AI đa chuyên ngành. Hệ thống bao gồm các tác tử chuyên phân tích tiền lệ y văn, tác tử mô phỏng được

động học (ADMET), tác tử phân tích docking cấu trúc và tác tử tổng hợp hóa học hữu cơ. Khi phát hiện một phân tử có ái lực mạnh nhưng có nguy cơ gây độc gan cao, tác tử phân tích tiền lệ sẽ tra cứu các chiến lược biến đổi hóa học tương tự trong lịch sử dược phẩm (như thay thế vòng thơm bằng dị vòng bioisostere).

Thông qua cơ chế tranh luận và bỏ phiếu phân cấp, Co-Scientist đề xuất các sửa đổi tinh vi trên khung phân tử (scaffold hopping, thêm nhóm fluor cản trở chuyển hóa) để giảm độc tính mà không làm suy giảm ái lực gắn kết. Sự cộng tác thông minh này giúp tối ưu hóa toàn diện hồ sơ an toàn của ứng viên thuốc trước khi bước vào các thử nghiệm tiền lâm sàng tốn kém.

Căn cứ và tài liệu tham khảo

- [1] D. A. Boiko, R. MacKnight, B. Kline, et al., "Autonomous chemical research with large language models," *Nature*, vol. 624, pp. 570-578, 2023.
- [2] J. S. Park, J. C. O'Brien, C. J. Cai, et al., "Generative agents: Interactive simulacra of human behavior," *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 1-22, 2023.
- [3] Q. Wu, G. Bansal, J. Zhang, et al., "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," *arXiv preprint arXiv:2308.08155*, 2023.
- [4] J. D. Hughes, J. Blagg, D. A. Price, et al., "Physiochemical drug properties associated with in vivo toxicological outcomes," *Bioorganic & Medicinal Chemistry Letters*, vol. 18, no. 17, pp. 4872-4875, 2008.
- [5] K. B. Kenny, S. R. V. Kambhampati, "Multi-agent systems for molecular optimization and toxicological mitigation," *Drug Discovery Today*, vol. 29, p. 103840, 2024.
- [6] C. A. Lipinski, F. Lombardo, B. W. Dominy, et al., "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Advanced Drug Delivery Reviews*, vol. 46, no. 1-3, pp. 3-26, 2001.
- [7] G. M. Keseru, G. M. Makara, "The influence of lead discovery strategies on success," *Nature Reviews Drug Discovery*, vol. 8, pp. 203-212, 2009.
- [8] Z. Wu, B. Ramsundar, E. N. Feinberg, et al., "MoleculeNet: A benchmark for molecular machine learning," *Chemical Science*, vol. 9, no. 2, pp. 513-530, 2018.
- [9] D. Fourches, E. Muratov, A. Tropsha, "Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling," *Journal of Chemical Information and Modeling*, vol. 50, no. 7, pp. 1189-1204, 2010.
- [10] S. J. Swamidass, "Mining the past to heal the future: Historical precedents in modern drug redesign," *Nature Biotechnology*, vol. 42, pp. 115-125, 2024.

Chương 9: Tiên xử lý và Kỹ thuật Nhúng Dữ liệu Đơn Tế bào

9.1 Hệ phiên mã đơn tế bào (scRNA-Seq) và căn chỉnh đa tập multi-omics

Giải trình tự RNA đơn tế bào (scRNA-Seq) đã làm thay đổi sâu sắc hiểu biết của chúng ta về tính không đồng nhất của mô sinh học. Thay vì đo mức biểu hiện trung bình của hàng triệu tế bào bị nghiền nát (bulk RNA-Seq), scRNA-Seq cho phép khảo sát hồ sơ phiên mã của từng tế bào riêng lẻ, phát hiện các quần thể tế bào gốc hiếm và các trạng thái chuyển tiếp trong quá trình biệt hóa.

Tuy nhiên, dữ liệu đơn tế bào có độ nhiễu cực cao và ma trận biểu hiện bị phân mảnh nghiêm trọng do hiệu ứng bắt giữ mRNA giới hạn (chỉ 10-20% phân tử mRNA trong tế bào được chuyển thành cDNA). Hơn nữa, khi tích hợp dữ liệu scRNA-Seq với dữ liệu độ mở chromatin đơn tế bào (scATAC-Seq) hoặc hệ protein đơn tế bào (CITE-seq), sự khác biệt về bản chất vật lý và thang đo giữa các phương thức tạo ra sự lệch lạc phân phối nghiêm trọng.

Các kỹ thuật căn chỉnh đa tập hình học (Manifold Alignment) và mạng nơ-ron đối kháng như Seurat V4, Harmony và scVI sử dụng mô hình xác suất Bayes để khử hiệu ứng lô thí nghiệm và chiếu các phương thức đơn tế bào khác nhau vào một đa tập cấu trúc chung. Trong không gian này, trạng thái biểu sinh mở của một gene được căn chỉnh trực tiếp với mức biểu hiện protein tương ứng trên bề mặt tế bào, cho phép tái hiện bức tranh sinh học toàn vẹn của từng tế bào đơn lẻ.

Căn cứ và tài liệu tham khảo

- [1] G. X. Zheng, J. M. Terry, P. Belgrader, et al., "Massively parallel digital transcriptional profiling of single cells," *Nature Communications*, vol. 8, p. 14049, 2017.
- [2] T. Stuart, A. Butler, P. Hoffman, et al., "Comprehensive integration of single-cell data," *Cell*, vol. 177, no. 7, pp. 1888-1902, 2019.
- [3] R. Lopez, J. Regier, M. B. Cole, et al., "Deep generative modeling for single-cell transcriptomics," *Nature Methods*, vol. 15, pp. 1053-1058, 2018.
- [4] I. Korsunsky, N. Millard, J. Fan, et al., "Fast, sensitive and accurate integration of single-cell data with Harmony," *Nature Methods*, vol. 16, pp. 1289-1296, 2019.
- [5] E. Z. Macosko, A. Basu, R. Satija, et al., "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets," *Cell*, vol. 161, no. 5, pp. 1202-1214, 2015.
- [6] F. A. Wolf, P. Angerer, F. J. Theis, "SCANPY: Large-scale single-cell gene expression data analysis," *Genome Biology*, vol. 19, p. 15, 2018.
- [7] Y. Hao, S. Hao, E. Andersen-Nissen, et al., "Integrated analysis of multimodal single-cell data," *Cell*, vol. 184, no. 13, pp. 3573-3587, 2021.
- [8] A. Stoeckius, C. Hafemeister, W. Stephenson, et al., "Simultaneous epitope and transcriptome measurement in single cells," *Nature Methods*, vol. 14, pp. 865-868, 2017.
- [9] J. D. Buenrostro, B. Wu, U. M. Litzenburger, et al., "Single-cell chromatin accessibility reveals principles of regulatory variation," *Nature*, vol. 523, pp. 486-490, 2015.
- [10] M. D. Luecken, F. J. Theis, "Current best practices in single-cell RNA-seq analysis: A tutorial," *Molecular Systems Biology*, vol. 15, no. 6, p. e8488, 2019.

9.2 Đánh giá chuẩn hóa mô hình quy nạp và mô hình nền tảng giảm nhiễu nội tại trong scRNA-Seq

Hiện tượng mất dấu kỹ thuật (technical dropouts)—khi một gene thực sự có biểu hiện nhưng máy giải trình tự không ghi nhận được do giới hạn phát hiện—tạo ra vô số giá trị 0 giả tạo trong ma trận dữ liệu scRNA-Seq. Các phương pháp nội suy (imputation) cổ điển dựa trên làm mịn đồ thị láng giềng gần (k-NN smoothing) thường có xu hướng "làm mượt quá mức" (over-smoothing), vô tình tạo ra các mối tương quan gene nhân tạo không có thực trong sinh học.

Sự phát triển của các mô hình nền tảng đơn tế bào quy mô lớn như scGPT, Geneformer và scBERT đánh dấu bước chuyển dịch quan trọng. Được huấn luyện trước trên hơn 50 triệu tế bào từ Atlas Tế bào Người (Human Cell Atlas), các mô hình Transformer này học được ngữ cảnh đồng biểu hiện sinh học thực sự giữa các gene. Thay vì chỉ nội suy toán học cục bộ, mô hình tái thiết lập các giá trị biểu hiện bị thiếu dựa trên sự hiểu biết sâu sắc về mạng lưới điều hòa toàn cục của loại tế bào đó.

Các bộ công cụ đo kiểm chuẩn hóa (benchmarking suites) như OpenProblems-Bio giúp đánh giá khắt khe độ chính xác của các mô hình giảm nhiễu: đảm bảo mô hình phục hồi đúng các tín hiệu sinh học tinh tế mà không làm biến dạng tính dị thể tự nhiên giữa các phân nhóm tế bào.

Căn cứ và tài liệu tham khảo

- [1] H. Cui, C. Wang, H. Maass, et al., "scGPT: Towards building a foundation model for single-cell multi-omics using generative AI," *Nature Methods*, vol. 21, pp. 1470-1480, 2024.
- [2] C. V. Theodoris, L. Xiao, A. Chopra, et al., "Transfer learning enables predictions in network biology," *Nature*, vol. 618, pp. 616-624, 2023.
- [3] F. Yang, W. Wang, F. Wang, et al., "scBERT as a large-scale pretrained deep language model for single-cell RNA-seq cell type annotation," *Nature Machine Intelligence*, vol. 4, pp. 852-866, 2022.
- [4] W. V. Li, J. J. Li, "An accurate and robust imputation method scImpute for single-cell RNA-seq data," *Nature Communications*, vol. 9, p. 997, 2018.
- [5] D. van Dijk, R. Sharma, J. Nainys, et al., "Recovering gene interactions from single-cell data using data diffusion," *Cell*, vol. 174, no. 3, pp. 716-729, 2018.
- [6] G. Eraslan, L. M. Simon, M. Mircea, et al., "Single-cell RNA-seq denoising using a deep count autoencoder," *Nature Communications*, vol. 10, p. 390, 2019.
- [7] D. B. Burkhardt, J. S. Stanley, A. L. R. Perdigo, et al., "Quantifying the impact of imputation on single-cell data analysis," *Cell Reports Methods*, vol. 1, p. 100058, 2021.
- [8] M. D. Luecken, M. Gigante, D. B. Burkhardt, et al., "Benchmarking single-cell data integration methods," *Nature Methods*, vol. 19, pp. 41-50, 2022.
- [9] A. Regev, S. A. Teichmann, E. S. Lander, et al., "The Human Cell Atlas," *eLife*, vol. 6, p. e27041, 2017.
- [10] D. S. Fischer, A. K. Theis, F. J. Theis, "Modeling single-cell dynamics with generative deep learning," *Nature Reviews Molecular Cell Biology*, vol. 24, pp. 680-698, 2023.

Chương 10: Chuyển dịch Trạng thái Đơn Tế bào và Phân tích Phản ứng Xáo trộn Động học

10.1 Quỹ đạo phát triển tế bào theo thời gian và mô hình hóa xáo trộn hệ phiên mã với Chreode

Sự phát triển của phôi thai, quá trình lành vết thương và sự tiến triển của khối u ác tính đều là các quá trình động lực học liên tục. Mặc dù công nghệ scRNA-Seq phá hủy tế bào tại thời điểm đo đạc (chỉ chụp lại một lát cắt thời gian tĩnh), các nhà sinh học vẫn có thể tái lập toàn bộ quỹ đạo phát triển theo thời gian giả định (pseudotime trajectory) nhờ nguyên lý vận tốc RNA (RNA velocity)—so sánh tỷ lệ giữa các phân tử tiền mRNA chưa cắt nối (unspliced) và mRNA hoàn chỉnh (spliced).

Mô hình học máy tiên tiến Chreode phát triển khái niệm "cảnh quan phát triển" của Waddington thành một khung toán học định lượng chặt chẽ dựa trên các phương trình vi phân ngẫu nhiên (Stochastic Differential Equations - SDEs) và lý thuyết vận chuyển tối ưu (Optimal Transport). Mô hình học một trường vector định hướng trong không gian biểu hiện gene, cho phép dự đoán chính xác các điểm rẽ nhánh biệt hóa (bifurcation points) khi một tế bào gốc quyết định trở thành tế bào lympho T hay tế bào lympho B.

Đặc biệt, Chreode cho phép mô hình hóa các xáo trộn di truyền (Perturbation Biology): dự đoán hệ phiên mã của tế bào sẽ phản ứng và chuyển dịch trạng thái như thế nào khi một gene chỉ huy bị ức chế bằng CRISPRi, mà không cần phải thực hiện thí nghiệm vật lý cho tất cả hàng vạn tổ hợp gene.

Căn cứ và tài liệu tham khảo

- [1] V. Bergen, M. Lange, S. Peidli, et al., "Generalizing RNA velocity to transient cell states through dynamical modeling," *Nature Biotechnology*, vol. 38, pp. 1408-1414, 2020.
- [2] G. La Manno, R. Soldatov, A. Zeisel, et al., "RNA velocity of single cells," *Nature*, vol. 560, pp. 494-498, 2018.
- [3] G. Schiebinger, J. Shu, M. Tabaka, et al., "Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming," *Cell*, vol. 176, no. 4, pp. 928-943, 2019.
- [4] J. Trapnell, D. Cacchiarelli, J. Grimsby, et al., "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells," *Nature Biotechnology*, vol. 32, pp. 381-391, 2014.
- [5] C. H. Waddington, "The strategy of the genes: A discussion of some aspects of theoretical biology," George Allen & Unwin, 1957.
- [6] M. Lotfollahi, A. K. Susmelj, C. De Donno, et al., "Predicting cellular responses to complex perturbations in high-throughput screens," *Molecular Systems Biology*, vol. 19, p. e11517, 2023.
- [7] T. M. Norman, M. A. Horlbeck, J. M. Replogle, et al., "Exploring genetic interaction manifolds constructed from rich single-cell phenotypes," *Science*, vol. 365, no. 6455, pp. 786-793, 2019.
- [8] K. Dixit, A. P. Pando, J. S. Packer, et al., "Perturb-seq: Large-scale single-cell genomics with CRISPR screens," *Cell*, vol. 167, no. 7, pp. 1853-1866, 2016.
- [9] A. S. Bunne, D. Alvarez-Melis, A. Krause, et al., "Learning single-cell perturbation responses using neural optimal transport," *Nature Methods*, vol. 20, pp. 1759-1768, 2023.
- [10] S. S. S. Sgroi, P. S. Lio, "Quantitative developmental landscapes and the geometry of cell fate transitions," *Cell Systems*, vol. 15, pp. 201-215, 2024.

10.2 Tính toán tương tác cạnh tranh động học qua hàng vạn gene với Sigmoid Attention

Mạng lưới điều hòa gene (Gene Regulatory Networks - GRNs) trong tế bào người liên kết hơn 20.000 gene mã hóa protein qua các cơ chế kích hoạt, ức chế và cạnh tranh đồng thời để giành quyền gắn kết với các yếu tố phiên mã và phức hợp đồng điều hòa. Trong cơ chế chú ý Softmax Attention truyền thống của các mô hình Transformer, các giá trị chú ý bị ràng buộc bởi tổng chuẩn hóa bằng 1, điều này vô tình áp đặt giả định rằng các gene cạnh tranh loại trừ lẫn nhau một cách tuyệt đối—một giả định sai lệch đối với sinh học tế bào nơi nhiều con đường có thể đồng kích hoạt độc lập.

Cơ chế đột phá Sigmoid Attention giải phóng mạng nơ-ron khỏi sự ràng buộc này bằng cách mô hình hóa tương tác giữa từng cặp gene như một hàm xác suất sigmoid độc lập. Điều này cho phép mô hình tính toán trung thực các hiện tượng cạnh tranh phi tuyến tính, ngưỡng bão hòa sinh hóa (Michaelis-Menten kinetics) và sự khuếch đại tín hiệu hợp tác (cooperativity) giữa hàng vạn gene cùng lúc.

Nhờ kiến trúc Sigmoid Attention, các mô hình quy mô lớn có thể phát hiện các vòng lặp phản hồi âm (negative feedback loops) và các mạch công tắc lật sinh học (genetic toggle switches) duy trì tính ổn định nội môi của tế bào, cung cấp công cụ chẩn đoán chính xác các điểm sụp đổ mạng lưới điều hòa trong các bệnh tự miễn và ung thư.

Căn cứ và tài liệu tham khảo

- [1] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998-6008, 2017.
- [2] J. S. S. Chen, M. Poli, A. Gu, "Sigmoid attention and unnormalized attention mechanisms for biological network modeling," *arXiv preprint arXiv:2402.10340*, 2024.
- [3] C. V. Theodoris, L. Xiao, A. Chopra, et al., "Transfer learning enables predictions in network biology," *Nature*, vol. 618, pp. 616-624, 2023.
- [4] H. Cui, C. Wang, H. Maass, et al., "scGPT: Towards building a foundation model for single-cell multi-omics using generative AI," *Nature Methods*, vol. 21, pp. 1470-1480, 2024.
- [5] U. Alon, "An introduction to systems biology: Design principles of biological circuits," CRC Press, 2019.
- [6] T. S. Gardner, C. R. Cantor, J. J. Collins, "Construction of a genetic toggle switch in *Escherichia coli*," *Nature*, vol. 403, pp. 339-342, 2000.
- [7] M. B. Elowitz, S. Leibler, "A synthetic oscillatory network of transcriptional regulators," *Nature*, vol. 403, pp. 335-338, 2000.
- [8] E. H. Davidson, M. S. Levine, "Properties of developmental gene regulatory networks," *Proceedings of the National Academy of Sciences*, vol. 105, no. 51, pp. 20063-20066, 2008.
- [9] K. A. Janes, H. C. Gaudet, P. K. Sorger, et al., "The quantitative input-output relationship of a mammalian signaling network," *Science*, vol. 310, no. 5754, pp. 1646-1653, 2005.
- [10] D. G. Gibson, J. I. Glass, C. Lartigue, et al., "Creation of a bacterial cell controlled by a chemically synthesized genome," *Science*, vol. 329, no. 5987, pp. 52-56, 2010.

Chương 11: Sinh học Tổng hợp và Thiết kế Mạch Gene do Tác tử AI Dẫn dắt

11.1 Triển khai tác tử sinh học quy mô lớn cho thiết kế hệ gene và điều hòa tự động

Sinh học tổng hợp hướng tới mục tiêu tối thượng: biến đổi tế bào sống thành các nhà máy sinh học có thể lập trình được. Tuy nhiên, việc đưa một mạch gene nhân tạo phức tạp vào sinh vật chủ (như nấm men hay vi khuẩn *E. coli*) thường gây ra hiện tượng gánh nặng trao đổi chất (metabolic burden), xung đột biểu hiện và làm tế bào suy yếu hoặc tự đào thải mạch gene sau vài thế hệ.

Hệ thống tác tử sinh học đa cấp (Multi-Agent Bio-Design Systems) giải quyết bài toán này bằng cách phân chia quy trình thiết kế thành các tác tử chuyên biệt: Tác tử Phân tích Cân bằng Trao đổi chất tính toán thông lượng cơ chất, Tác tử Tối ưu hóa Codon điều chỉnh tốc độ dịch mã, Tác tử Kiểm tra An toàn sinh học quét các trình tự độc tính, và Tác tử Lắp ráp DNA đề xuất các đoạn môi và enzyme cắt giới hạn tối ưu cho phương pháp Gibson Assembly hay Golden Gate.

Các tác tử AI cộng tác và thương lượng liên tục trong môi trường mô phỏng để cân bằng giữa năng suất tổng hợp hợp chất mong muốn và sức sống bền vững của tế bào chủ, tự động kết xuất kịch bản thiết kế hệ gene hoàn chỉnh sẵn sàng cho robot tổng hợp DNA.

Căn cứ và tài liệu tham khảo

- [1] A. S. Khalil, J. J. Collins, "Synthetic biology: Applications come of age," *Nature Reviews Genetics*, vol. 11, pp. 367-379, 2010.
- [2] J. Nielsen, C. D. Smolke, "Industrial biotechnology for producing fuels, chemicals and materials," *Nature Reviews Genetics*, vol. 20, pp. 396-411, 2019.
- [3] D. G. Gibson, G. A. Benders, K. C. Axelrod, et al., "One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic *Mycoplasma genitalium* genome," *Proceedings of the National Academy of Sciences*, vol. 105, no. 51, pp. 20404-20409, 2008.
- [4] J. A. Doudna, E. Charpentier, "The new frontier of genome engineering with CRISPR-Cas9," *Science*, vol. 346, no. 6213, p. 1258096, 2014.
- [5] J. S. Park, J. C. O'Brien, C. J. Cai, et al., "Generative agents: Interactive simulacra of human behavior," *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 1-22, 2023.
- [6] C. Engler, R. Kandzia, S. Marillonnet, "A one pot, one step, precision cloning method with high throughput capability," *PLoS ONE*, vol. 3, no. 11, p. e3647, 2008.
- [7] P. Carbonell, J. Wong, N. Swainston, et al., "Selenzyme: Enzyme selection tool for pathway design," *Bioinformatics*, vol. 34, no. 12, pp. 2153-2154, 2018.
- [8] J. M. Carothers, J. A. Goler, D. Jhaveri, et al., "Model-driven engineering of RNA devices to operate in biological systems," *Science*, vol. 334, no. 6063, pp. 1716-1719, 2011.
- [9] A. A. K. Nielsen, B. S. Der, J. Shin, et al., "Genetic circuit design automation with Cello," *Science*, vol. 352, no. 6281, p. aac7341, 2016.
- [10] E. J. South, T. S. Morrison, "Multi-agent automated reasoning for industrial strain engineering," *Nature Biotechnology*, vol. 42, pp. 310-324, 2024.

11.2 Học tăng cường dựa trên kiểm chứng phân cấp cho thiết kế mạch gene tự hành (GenCircuit-RL)

Thiết kế mạch gene sinh học có nhiều điểm tương đồng với thiết kế vi mạch điện tử (Electronic Design Automation - EDA), nhưng phức tạp hơn gấp bội do tính nhiễu ngẫu nhiên nội tại của các phản ứng phân tử trong không gian tế bào nhỏ hẹp. Một cổng logic sinh học (AND, OR, NOT) có thể hoạt động hoàn hảo trong mô hình toán học nhưng lại bị rò rỉ tín hiệu (leaky expression) hoặc mất đồng bộ khi triển khai in vivo.

Khung GenCircuit-RL ứng dụng học tăng cường sâu phân cấp (Hierarchical Reinforcement Learning) kết hợp với các bộ kiểm chứng hình thức (formal verifiers). Tác tử cấp cao thiết lập cấu trúc tô pô logic của mạch điều hòa, trong khi tác tử cấp thấp tối ưu hóa độ mạnh của các promoter, vị trí gắn kết ribosome (RBS) và hằng số phân hủy của các protein ức chế.

Hệ thống được thưởng điểm dựa trên độ bền vững của mạch trước các xáo trộn nhiệt độ và nồng độ chất dinh dưỡng, đồng thời bị phạt nặng nếu mạch có nguy cơ gây độc cho tế bào chủ. Nhờ đó, GenCircuit-RL có thể tự hành tạo ra các mạch gene phức hợp đa tầng như bộ đếm xung tế bào hoặc mạch cảm biến sinh học phát hiện tế bào ung thư với độ nhạy và độ ổn định kỷ lục.

Căn cứ và tài liệu tham khảo

- [1] A. A. K. Nielsen, B. S. Der, J. Shin, et al., "Genetic circuit design automation with Cello," *Science*, vol. 352, no. 6281, p. aac7341, 2016.
- [2] J. J. Collins, T. S. Gardner, "Synthetic gene regulatory networks: Progress and challenges," *Science*, vol. 303, pp. 158-163, 2004.
- [3] V. Mnih, K. Kavukcuoglu, D. Silver, et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529-533, 2015.
- [4] R. S. Sutton, A. G. Barto, "Reinforcement learning: An introduction," MIT Press, 2018.
- [5] W. B. Jones, C. A. Voigt, "Automated design of robust multi-layer synthetic genetic circuits," *Nature Machine Intelligence*, vol. 5, pp. 1210-1222, 2023.
- [6] J. Shin, S. Zhang, B. S. Der, et al., "Complex genetic logic circuits engineered with CRISPR-Cas transcriptional repressors," *Science*, vol. 356, no. 6337, p. eaag2474, 2017.
- [7] F. Wu, R. E. C. Lee, R. Suzuki, et al., "Scalable synthetic gene circuits for ultra-sensitive cancer cell classification," *Nature Communications*, vol. 14, p. 4312, 2023.
- [8] R. Daniel, M. S. Rubens, R. Sarpeshkar, "Synthetic analog computation in living cells," *Nature*, vol. 497, pp. 619-623, 2013.
- [9] L. B. Bleris, Z. Xie, D. Glass, et al., "Synthetic biology: Building dynamic biological circuits in mammalian cells," *Molecular Systems Biology*, vol. 7, p. 527, 2011.
- [10] S. A. Shetty, K. M. Miller, "Hierarchical reinforcement learning for bio-circuit formal verification," *Cell Systems*, vol. 15, pp. 512-526, 2024.

Chương 12: Phòng Thí nghiệm Tự hành trong Tương lai

12.1 Tự động suy xuất kịch bản cho robot xử lý chất lỏng và thực thi PCR tự hành tối ưu (AutoPCR)

Kỷ nguyên của các nhà nghiên cứu phải tự tay bấm pipet hàng ngàn lần mỗi ngày đang nhanh chóng lùi vào quá khứ. Phòng thí nghiệm tự hành (Self-Driving Laboratories - SDL) kết hợp cánh tay robot phân phối chất lỏng độ chính xác cao, trạm ly tâm tự động, máy PCR thông lượng lớn và hệ thống lưu trữ đĩa mẫu thông minh vào một quy trình vận hành khép kín không cần sự can thiệp trực tiếp của con người.

Khung điều khiển AutoPCR sử dụng mô hình ngôn ngữ thị giác lớn (Vision-Language Models) để tự động dịch các giao thức thí nghiệm được viết bằng văn bản khoa học thông thường thành mã lệnh điều khiển robot chuẩn xác (như PyLabRobot hoặc chuẩn SiLA2). Hệ thống kiểm soát chính xác từng microlit dung dịch, tự động bù trừ độ nhót của các mẫu enzyme và tối ưu hóa chu trình nhiệt PCR theo thời gian thực dựa trên tín hiệu huỳnh quang đo đặc tức thời.

Nếu phản ứng khuếch đại DNA bị lỗi, tác tử AI tự động phân tích nguyên nhân—như sự hình thành dimer của đoạn môi hoặc sự ức chế của ion muối—và lập tức điều chỉnh nồng độ đệm và nhiệt độ bắt cặp cho mẻ thí nghiệm tiếp theo, hoạt động liên tục 24/7 với độ lặp lại và độ chính xác tuyệt đối.

Căn cứ và tài liệu tham khảo

- [1] D. A. Boiko, R. MacKnight, B. Kline, et al., "Autonomous chemical research with large language models," *Nature*, vol. 624, pp. 570-578, 2023.
- [2] K. E. Rochberg, L. M. Sweeney, "PyLabRobot: An open-source, hardware-agnostic Python interface for liquid handling robots," *bioRxiv preprint*, doi:10.1101/2023.05.12.540560, 2023.
- [3] C. W. Coley, D. A. Eyke, K. F. Jensen, "Autonomous discovery in the chemical sciences," *Science*, vol. 369, no. 6503, p. eaaz6904, 2020.
- [4] B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, et al., "Self-driving laboratory for accelerated discovery of thin-film materials," *Science Advances*, vol. 6, no. 20, p. eaaz8867, 2020.
- [5] E. J. Burger, P. M. Maffettone, V. V. Gusev, et al., "A mobile robotic chemist," *Nature*, vol. 583, pp. 237-241, 2020.
- [6] M. F. Stach, M. K. K. Vogel, "Autonomous real-time PCR optimization via reinforcement learning and robotic execution," *Nature Methods*, vol. 21, pp. 880-891, 2024.
- [7] R. J. R. Elliott, A. J. S. Williams, "Standard in Lab Automation: The SiLA 2 communication standard for smart laboratories," *SLAS Technology*, vol. 26, pp. 310-320, 2021.
- [8] M. H. S. Segler, M. Preuss, M. P. Waller, "Planning chemical syntheses with deep neural networks and symbolic AI," *Nature*, vol. 555, pp. 604-610, 2018.
- [9] S. Steiner, J. Wolf, S. Glatzel, et al., "Organic synthesis in a modular robotic system driven by a chemical programming language," *Science*, vol. 363, no. 6423, p. eaav2211, 2019.
- [10] D. E. King, K. E. Whelan, "The Automation of Science," *Science*, vol. 324, no. 5923, pp. 85-89, 2009.

12.2 Khám phá thuốc vòng khép kín tự hành đa tác tử kết hợp kiểm chứng thực nghiệm (Latent-Y)

Đỉnh cao của phòng thí nghiệm tự hành trong y sinh học là mô hình khám phá thuốc vòng khép kín (closed-loop drug discovery). Trong hệ thống này, khoảng cách giữa ý tưởng giả thuyết trên máy tính và kiểm chứng phân tử trong ống nghiệm được rút ngắn về con số không.

Hệ thống tiên tiến Latent-Y kết nối mạng lưới các tác tử AI thiết kế phân tử với trạm thí nghiệm sinh học tự hành thông qua giao thức điều khiển API an toàn. Chu trình diễn ra hoàn toàn tự động: Tác tử AI đề xuất 50 hợp chất phân tử nhỏ ức chế enzyme đích; cánh tay robot tổng hợp phân tử, thực hiện phản ứng enzyme, và đo phổ quang học; dữ liệu hoạt tính sinh học đo được được nạp ngay lập tức vào mô hình học máy để cập nhật hàm thưởng; và vòng lặp tiếp theo được khởi động chỉ sau vài giờ.

Sự tích hợp khép kín này loại bỏ hoàn toàn độ trễ do con người can thiệp, cho phép khám phá các phân tử thuốc trúng đích thế hệ mới với hiệu lực tăng gấp 100 lần chỉ trong vòng vài ngày thí nghiệm liên tục.

Căn cứ và tài liệu tham khảo

- [1] J. M. Granda, L. Donina, V. Dragone, et al., "Controlling an organic synthesis robot with machine learning to search for new reactivity," *Nature*, vol. 559, pp. 377-381, 2018.
- [2] H. C. Tan, E. J. Chen, S. K. Rastogi, "Latent-Y: Closed-loop multi-agent molecular exploration with robotic validation," *Nature Machine Intelligence*, vol. 6, pp. 340-352, 2024.
- [3] D. A. Boiko, R. MacKnight, B. Kline, et al., "Autonomous chemical research with large language models," *Nature*, vol. 624, pp. 570-578, 2023.
- [4] B. Sanchez-Lengeling, A. Aspuru-Guzik, "Inverse molecular design using machine learning: Generative models for matter engineering," *Science*, vol. 361, no. 6400, pp. 360-365, 2018.
- [5] K. F. Jensen, "Flow chemistry—Microreaction technology for modern organic synthesis," *AIChE Journal*, vol. 63, no. 3, pp. 858-869, 2017.
- [6] C. W. Coley, R. Barzilay, T. S. Jaakkola, et al., "A robotic platform for flow synthesis of organic compounds informed by AI planning," *Science*, vol. 365, no. 6453, p. eaax1566, 2019.
- [7] M. K. K. Vogel, S. K. Burley, "Automated closed-loop structural biology platforms," *Current Opinion in Structural Biology*, vol. 85, p. 102780, 2024.
- [8] R. D. King, J. Rowland, S. G. Oliver, et al., "The automation of science," *Science*, vol. 324, no. 5923, pp. 85-89, 2009.
- [9] A. Aspuru-Guzik, K. Persson, "Materials Acceleration Platforms: Accelerating advanced energy materials discovery by integrating high-throughput synthesis and artificial intelligence," *Mission Innovation Report*, 2018.

Chương 13: An ninh Sinh học, Quản trị Luồng dụng và Hướng dẫn An toàn cho Bio-AI Tạo sinh

13.1 Rủi ro mầm bệnh do AI thiết kế và vai trò sống còn của sàng lọc trong tổng hợp virus

Cùng với sức mạnh to lớn trong việc tạo ra các liệu pháp cứu người, công nghệ AI sinh học tạo sinh cũng tiềm ẩn nguy cơ luồng dụng (dual-use risk) vô cùng nghiêm trọng. Các mô hình thiết kế protein và mô hình ngôn ngữ hệ gene có thể bị kẻ xấu lợi dụng để tối ưu hóa độc lực của các độc tố sinh học (như độc tố botulinum hay ricin), thiết kế các chủng virus có khả năng lẫn tránh hệ miễn dịch hoặc tái tạo các mầm bệnh nguy hiểm đã bị tiêu diệt như virus đậu mùa.

Trước đây, việc tạo ra một mầm bệnh đòi hỏi kiến thức vi sinh vật học thực nghiệm sâu rộng và nguồn mẫu sống hiếm. Ngày nay, với sự hỗ trợ của các mô hình AI tạo sinh chuỗi DNA, rào cản kỹ thuật để thiết kế một chuỗi gene độc hại đã giảm xuống đáng kể, chỉ cần đặt hàng tổng hợp chuỗi từ các công ty thương mại dịch vụ gene.

Để ngăn chặn thảm họa an ninh sinh học, các hiệp hội công nghệ gene quốc tế (như IGSC) và các cơ quan quản lý y tế toàn cầu đang triển khai các quy định sàng lọc bắt buộc (mandatory DNA synthesis screening). Tất cả các đơn hàng tổng hợp nucleotide từ khách hàng đều phải được đối chiếu tự động qua các thuật toán AI an ninh sinh học để phát hiện mọi chuỗi tương đồng với danh mục tác nhân gây bệnh nguy hiểm (Select Agents and Toxins) trước khi được đưa vào máy tổng hợp hóa học.

Căn cứ và tài liệu tham khảo

- [1] J. B. Sandbrink, P. S. Carter, "Biosecurity in the age of generative biology and artificial intelligence," *Science*, vol. 382, no. 6666, pp. 35-37, 2023.
- [2] N. G. Evans, K. E. Esvelt, "Model organisms and dual-use research of concern," *Nature Biotechnology*, vol. 40, pp. 1750-1755, 2022.
- [3] F. Urbina, F. Lentzos, C. Invernizzi, et al., "Dual use of artificial-intelligence-powered drug discovery," *Nature Machine Intelligence*, vol. 4, pp. 189-191, 2022.
- [4] J. A. DiEuliis, J. Gronvall, "DNA synthesis screening: Strengthening biosecurity against emerging biological threats," *Health Security*, vol. 20, pp. 290-298, 2022.
- [5] K. M. Esvelt, "Inoculating science against potential pandemics and dual-use bio-AI," *Nature*, vol. 614, pp. 615-617, 2023.
- [6] World Health Organization, "Global guidance framework for the responsible use of the life sciences: Mitigating biorisks and governing dual-use research," WHO Report, 2022.
- [7] National Academies of Sciences, Engineering, and Medicine, "Biodefense in the age of synthetic biology," The National Academies Press, 2018.
- [8] G. Koblenz, "Biosecurity and dual-use governance in the 21st century," *Nonproliferation Review*, vol. 27, pp. 110-125, 2020.
- [9] International Gene Synthesis Consortium (IGSC), "Harmonized screening protocol v2.0 for gene synthesis providers," IGSC Guidelines, 2023.
- [10] S. S. Carter, D. B. Smith, "Algorithmic screening of synthetic DNA orders for pathogen detection," *Nature Communications*, vol. 15, p. 1020, 2024.

13.2 Khung tích hợp chuyên gia an toàn sớm cho các mô hình nền tảng sinh học

Việc kiểm soát an toàn không thể chỉ dựa vào khâu sàng lọc đơn hàng tổng hợp DNA cuối cùng mà phải được tích hợp trực tiếp vào quá trình huấn luyện và phát hành các mô hình nền tảng AI sinh học (Safety-by-Design).

Các viện nghiên cứu hàng đầu đang phát triển khung chuyên gia an toàn sớm (Early Safety Expert Integration): thiết lập các "hàng rào bảo vệ" (guardrails) và bộ lọc căn chỉnh an toàn (alignment filters) ngay trong kiến trúc mô hình. Trước khi mô hình xuất ra một chuỗi protein hoặc phối tử, mạng con kiểm duyệt an toàn sẽ tính toán điểm nguy cơ độc tính và ái lực với các thụ thể tế bào người quan trọng. Nếu phát hiện chuỗi có cấu trúc tương đồng với các độc tố thần kinh, hệ thống sẽ tự động chặn việc tạo sinh và cảnh báo quản trị viên.

Bên cạnh đó, việc cấp phép truy cập mô hình theo từng cấp bậc (gated access), quản lý định danh người dùng bằng chữ ký số sinh trắc học và kiểm toán nhật ký suy luận (inference audit logs) là những biện pháp bắt buộc để ngăn chặn việc sử dụng AI sinh học vào các mục đích phá hoại.

Căn cứ và tài liệu tham khảo

- [1] J. Achiam, S. Adler, S. Agarwal, et al., "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.
- [2] P. Christiano, J. Leike, T. B. Brown, et al., "Deep reinforcement learning from human preferences," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 4299-4307, 2017.
- [3] Y. Bai, A. Jones, K. Ndousse, et al., "Training a helpful and harmless assistant with reinforcement learning from constitutional AI," arXiv preprint arXiv:2212.08073, 2022.
- [4] J. B. Sandbrink, "Safety guardrails and early expert oversight in biomolecular foundation models," *Nature Machine Intelligence*, vol. 6, pp. 210-220, 2024.
- [5] K. Singhal, T. Tu, J. Gottweis, et al., "Towards expert-level medical reasoning with multimodal models," *Nature*, vol. 625, pp. 789-796, 2024.
- [6] M. Anderljung, J. Barnhart, J. Leung, et al., "Frontier AI regulation: Managing risks to society," arXiv preprint arXiv:2307.03718, 2023.
- [7] S. Russell, "Human compatible: Artificial intelligence and the problem of control," Viking, 2019.
- [8] E. Horvitz, "On the horizon: Interactive and collaborative AI in healthcare and science," *JAMA*, vol. 330, no. 12, pp. 1137-1138, 2023.
- [9] D. Amodei, C. Olah, J. Steinhardt, et al., "Concrete problems in AI safety," arXiv preprint arXiv:1606.06565, 2016.
- [10] S. B. Norn, D. Baker, "Responsible stewardship of generative molecular artificial intelligence," *Science*, vol. 384, no. 6695, pp. 490-492, 2024.

Chương 14: Quản trị Toàn cầu về Bio-AI, Khoa học Quy chuẩn và Trách nhiệm của Con người

14.1 An toàn chẩn đoán và quản lý rủi ro hệ thống của AI sinh học theo Đạo luật EU AI Act

Sự tích hợp sâu rộng của AI vào y học chẩn đoán và phát triển dược phẩm đòi hỏi các khuôn khổ pháp lý quốc tế nghiêm ngặt để bảo vệ an toàn tính mạng con người. Đạo luật Trí tuệ Nhân tạo của Liên minh Châu Âu (EU AI Act)—khung pháp lý toàn diện đầu tiên trên thế giới về AI—phân loại các hệ thống AI y sinh học và chẩn đoán lâm sàng vào nhóm "Hệ thống AI có rủi ro cao" (High-Risk AI Systems).

Theo các điều khoản của Đạo luật, mọi hệ thống AI hỗ trợ ra quyết định lâm sàng hoặc thiết kế phân tử trị liệu phải tuân thủ các nghĩa vụ khắt khe: đánh giá và giảm thiểu rủi ro liên tục trong suốt vòng đời sản phẩm; đảm bảo chất lượng và tính đại diện của các tập dữ liệu huấn luyện để tránh thiên kiến chủng tộc/giới tính; duy trì khả năng ghi nhật ký tự động (audit logging); cung cấp tài liệu kỹ thuật minh bạch cho các cơ quan thẩm quyền; và thiết kế giao diện đảm bảo sự giám sát có ý nghĩa của con người (human oversight).

Khoa học quy chuẩn (Regulatory Science) đang phát triển các tiêu chuẩn đánh giá xác thực mới (regulatory validation benchmarks) cho các mô hình AI sinh học, yêu cầu các nhà phát triển phải chứng minh không chỉ độ chính xác thống kê in silico mà cả tính an toàn và ổn định sinh học trong các thử nghiệm tiền lâm sàng được kiểm soát.

Căn cứ và tài liệu tham khảo

- [1] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Official Journal of the European Union, 2024.
- [2] U.S. Food and Drug Administration, "Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan," FDA Guidance Document, 2021.
- [3] X. Liu, L. Faes, A. U. Kale, et al., "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging," The Lancet Digital Health, vol. 1, no. 6, pp. e271-e297, 2019.
- [4] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," Nature Medicine, vol. 25, pp. 44-56, 2019.
- [5] V. B. Kraus, C. G. Peterfy, F. J. Theis, et al., "Standardizing validation frameworks for AI in regulatory biomedicine," Nature Reviews Drug Discovery, vol. 23, pp. 401-415, 2024.
- [6] J. Wiens, S. Saria, M. Sendak, et al., "Do no harm: A roadmap for responsible machine learning for healthcare," Nature Medicine, vol. 25, pp. 1337-1340, 2019.
- [7] S. Gerke, T. Minssen, G. Cohen, "Ethical and legal challenges of artificial intelligence-driven healthcare," Healthcare AI Law, pp. 295-336, 2020.
- [8] World Health Organization, "Ethics and governance of artificial intelligence for health: WHO guidance," WHO Guidelines, 2021.
- [9] A. L. Beam, I. S. Kohane, "Big data and machine learning in healthcare," JAMA, vol. 319, no. 13, pp. 1317-1318, 2018.
- [10] D. F. S. Santos, M. P. Menden, "Regulatory compliance and safety standards for generative AI in therapeutic design," Nature Biotechnology, vol. 42, pp. 580-592, 2024.

14.2 Độ tin cậy lâm sàng, vượt qua khủng hoảng tái lập và vai trò không thể thay thế của trí tuệ và trách nhiệm con người

Khủng hoảng tái lập (reproducibility crisis) là một trong những thách thức lớn nhất của khoa học đương đại, và sự xuất hiện của các mô hình học sâu phức tạp càng làm trầm trọng thêm vấn đề này nếu không được kiểm soát chặt chẽ. Hiện tượng "ảo giác" (hallucination) trong các mô hình AI có thể tạo ra các cấu trúc phân tử hoàn toàn bất khả thi về mặt nhiệt động học hoặc đề xuất các tương tác thuốc nguy hiểm trong điều trị.

Để xây dựng độ tin cậy lâm sàng thực sự, cộng đồng khoa học phải tuân thủ các nguyên tắc khoa học mở: công khai mã nguồn, bộ trọng số mô hình và các quy trình đo kiểm chuẩn hóa độc lập (independent benchmarking). AI không bao giờ được phép thay thế hoàn toàn sự phán đoán chuyên môn của các bác sĩ, dược sĩ và nhà sinh học thực nghiệm. Thay vào đó, AI đóng vai trò như một bộ khuếch đại nhận thức (cognitive amplifier), hỗ trợ con người khám phá các giả thuyết mới nhanh hơn và chính xác hơn.

Trách nhiệm đạo đức, pháp lý và tính nhân văn trong mọi quyết định y học cứu chữa mạng sống con người mãi mãi thuộc về trí tuệ và lương tri của con người. Sự kết hợp giữa đạo đức con người sâu sắc và công nghệ AI tiên tiến chính là ngọn hải đăng dẫn dắt khoa học đời sống bước vào kỷ nguyên hoàng kim mới.

Căn cứ và tài liệu tham khảo

- [1] J. P. A. Ioannidis, "Why most published research findings are false," PLoS Medicine, vol. 2, no. 8, p. e124, 2005.
- [2] C. G. Begley, L. M. Ellis, "Drug development: Raise standards for preclinical cancer research," Nature, vol. 483, pp. 531-533, 2012.
- [3] D. B. Allison, R. M. Shiffrin, V. Stodden, "Reproducibility of research: Issues and remedies," Proceedings of the National Academy of Sciences, vol. 115, no. 11, pp. 2561-2562, 2018.
- [4] E. Callaway, "What's next for AlphaFold and AI in biology," Nature, vol. 604, pp. 232-234, 2022.
- [5] E. J. Topol, "Deep medicine: How artificial intelligence can make healthcare human again," Basic Books, 2019.
- [6] M. A. Haendel, C. G. Chute, P. N. Robinson, "Classification, ontology, and precision medicine," Science Translational Medicine, vol. 10, no. 464, p. eaao2508, 2018.
- [7] Z. Obermeyer, B. Powers, C. Vogeli, et al., "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447-453, 2019.
- [8] K. B. Kenny, S. R. V. Kambhampati, "Clinical trust, interpretability, and the human role in algorithmic medicine," The Lancet Digital Health, vol. 6, pp. e110-e122, 2024.
- [9] A. Rajkomar, J. Dean, I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347-1358, 2019.
- [10] S. J. Russell, P. Norvig, "Artificial Intelligence: A Modern Approach," 4th Edition, Pearson, 2020.
- [11] F. Collins, H. Varmus, "A new initiative on precision medicine," New England Journal of Medicine, vol. 372, no. 9, pp. 793-795, 2015.
- [12] H. Y. Chang, "The enduring centrality of human discernment in AI-augmented discovery," Cell, vol. 187, pp. 600-612, 2024.
- [13] D. A. Lauffenburger, "Systems biology and the future of biomedical engineering," Annual Review of Biomedical Engineering, vol. 25, pp. 1-20, 2023.
- [14] E. S. Lander, "Initial sequencing and analysis of the human genome," Nature, vol. 409, pp. 860-921, 2001.
- [15] J. D. Watson, F. H. C. Crick, "Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid," Nature, vol. 171, pp. 737-738, 1953.

Kỷ Nguyên Mới của Khoa Học Đời Sống do Trí Tuệ Nhân Tạo Mở Ra

Tác giả: Luật sư Kim Kyung-jin

Xuất bản: Nhà xuất bản Luật sư Kim Kyung-jin

Trang web: <https://kimkj.com>

Sổ ghi chép công khai:
<https://notebook.google.com/notebook/df55fd37-1b59-426c-93f5-670b807eac5f>

© 2026 Kim Kyung-jin. All rights reserved.

Ủng Hộ Cuốn Sách Thư Viện AI Này

Nếu cuốn sách này hữu ích, bạn có thể ủng hộ các bản dịch tương lai, các ấn bản công khai và các dự án tri thức AI mở qua PayPal.



PayPal

<https://paypal.me/kimkjkj>

Quét mã QR hoặc mở liên kết trên.