# 인공지능이 여는
# 생명과학의 새 시대

김경진 변호사

# 서문

이 책은 인공지능으로 종합 정리한 연구서입니다. 자료를 고르고 구조를 정한 것은 사람이고, 문장을 쓰고 사실을 대조하는 작업은 인공지능 모델이 맡았습니다.

바탕이 된 것은 구글 노트북엘엠(NotebookLM)에 만들어 둔 노트 하나입니다. 2026년 8월까지 공개된 arXiv 논문 332편과 국제 학술 세미나 영상 71건을 그 노트에 모았습니다. 영어, 중국어, 일본어 자료가 함께 들어 있습니다. 노트가 정리해 낸 장별 요지를 뼈대로 삼았고, 각 장의 본문은 원문 논문과 2026년 2월 이후의 자료를 다시 검색해 채웠습니다.

노트는 공개해 두었습니다. 아래 주소로 들어가면 이 책이 어떤 자료 위에 서 있는지 직접 확인할 수 있습니다.

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

연구자와 학생이 학습에 쓰라고 열어 둔 것입니다. 책의 서술이 미덥지 않은 대목이 있으면 노트의 원문으로 돌아가 확인하시기 바랍니다.

각 절 끝에는 그 절이 근거로 삼은 자료 목록을 논문 표기 방식으로 붙였습니다. 인용한 논문은 arXiv 응용 프로그래밍 인터페이스와 크로스레프로 저자와 제목을 한 건씩 대조했습니다. 실재를 확인하지 못한 자료는 본문에서 뺐습니다.

인공지능이 정리한 글에는 한계가 있습니다. 틀릴 수 있습니다. 확인은 읽는 사람의 몫입니다.

김경진

# 목차

# 제1장: 인공지능 기반 바이오 R&D의 패러다임 전환

## 1.1 드라이랩(Dry Lab)과 웨트랩(Wet Lab)의 상호 보완적 융합 및 다학제적 커리어 모델

새벽 두 시의 실험실은 조용합니다. 형광등은 절반만 켜져 있고 원심분리기가 낮게 웅웅거립니다. 대학원생 한 명이 96개 구멍이 뚫린 플레이트에 시약을 한 칸씩 옮겨 담습니다. 손목이 저려 올 때쯤 그는 몸을 돌려 옆자리 노트북을 봅니다. 화면에서는 학습 곡선이 한 칸씩 내려가고 있습니다. 피펫과 키보드를 번갈아 잡는 이 장면이 지금 생명과학 연구실의 평범한 밤입니다.

생물학은 오랫동안 손으로 만지는 학문이었습니다. 그레고어 멘델은 완두콩을 하나하나 교배해 유전의 규칙을 찾아냈습니다. 1953년 로절린드 프랭클린이 찍은 X선 회절 사진 한 장이 DNA 이중나선의 모양을 드러냈습니다. 2003년 인간 게놈 프로젝트는 30억 쌍의 염기 서열을 끝까지 읽어 냈습니다. 이 성과들의 무대는 언제나 시험관과 배양기가 놓인 물리적 실험실, 웨트랩(Wet Lab)이었습니다. 웨트랩은 살아 있는 세포가 실제로 어떻게 반응하는지 눈으로 확인하는 자리입니다. 그 자리를 대신할 곳은 아직 없습니다.

문제는 값입니다. 신약 하나가 실험대에서 약국 선반까지 가는 데 평균 10년에서 15년이 걸립니다. 들어가는 연구개발비는 2조 원을 넘습니다. 임상 1상에 들어간 후보 물질 열 개 중 아홉 개는 허가를 받지 못하고 사라집니다. 수만 개의 화합물을 하나씩 합성해 세포에 넣어 보는 방식은 정직합니다. 그리고 느립니다. 실패는 대부분 실험대 위에서 확인되는데, 그때는 이미 돈과 시간이 다 들어간 뒤입니다.

드라이랩(Dry Lab)은 시약 대신 데이터를 다루는 실험실입니다. 냉장고에 든 것은 시료가 아니라 서버실의 열기이고, 재료는 페타바이트 단위의 오믹스 측정값과 수억 건의 단백질 원자 좌표입니다. 화학적으로 만들 수 있는 분자의 수는 10의 60제곱 정도로 추정합니다. 지구상의 모래알을 다 세어도 이 숫자 근처에 가지 못합니다. 드라이랩은 이 거대한 창고를 물리 법칙과 기하학적 딥러닝(Geometric Deep Learning)으로 훑습니다. 분자를 문장이 아니라 3차원 도형으로 다루는 학습 방식입니다. 예전에는 몇 달이 걸리던 후보 선별이 몇 시간으로 줄었습니다.

드라이랩이 연구실의 변두리에서 한복판으로 옮겨 온 순간을 하나만 고르라면 2021년입니다. 그해 딥마인드는 아미노산 서열만 넣으면 단백질이 접히는 3차원 모양을

예측하는 알파폴드(AlphaFold)를 공개했습니다[7]. 그전까지 단백질 구조 하나를 밝히려면 결정을 키우고 X선을 쏘아 몇 년을 매달려야 했습니다. 알파폴드는 그 일을 몇 분으로 줄였고, 공개된 예측 구조는 2억 개를 넘었습니다. 전 세계 연구자 수백만 명이 이 데이터베이스를 씁니다. 구조를 아는 것과 기능을 아는 것은 다르다는 반론은 지금도 유효합니다. 그래도 출발선의 위치는 완전히 옮겨졌습니다. 예전에는 구조를 얻는 것이 논문 한 편이었고, 지금은 구조가 실험의 준비물입니다.

진짜 변화는 둘 중 하나가 이기는 데 있지 않습니다. 두 실험실이 서로에게 답장을 보내기 시작한 데 있습니다. 드라이랩이 수억 개의 가상 분자에서 유망한 후보 열댓 개를 골라 보내면, 웨트랩은 그것을 실제로 합성해 결합 상수와 세포 독성을 잽니다. 그 측정값은 다시 숫자가 되어 모델로 돌아갑니다. 모델은 자기가 틀렸던 지점을 확인하고 가중치를 고칩니다. 이 왕복을 닫힌 루프(Closed Loop)라고 부릅니다. 편지가 오가는 주기가 짧을수록 연구는 빨라집니다.

루프 한 바퀴가 실제로 어떻게 도는지 보면 감이 옵니다. 월요일 아침, 모델이 표적 단백질의 결합 자리에 들어갈 분자 열두 개를 추립니다. 화요일에 합성이 걸리고, 목요일에 결합 세기를 재는 실험이 끝납니다. 열두 개 중 아홉 개는 예측보다 훨씬 약하게 붙었습니다. 예전 같으면 여기서 낙담하고 끝냈을 결과입니다. 지금은 이 아홉 번의 실패가 값비싼 학습 자료가 됩니다. 모델은 자기가 어떤 원자 배치에서 과신했는지 확인하고 다음 주 후보 목록을 다시 짭니다. 실패가 데이터가 되는 순간, 실험실의 시간은 직선이 아니라 나선이 됩니다.

2026년의 연구실은 이 편지를 사람이 아니라 기계가 부치는 단계에 들어섰습니다. 프로토파일럿(ProtoPilot)은 실험 의도를 받아 프로토콜 문서를 쓰고, 그것을 장비가 읽는 코드로 옮긴 뒤, 실행 결과를 보고 프로토콜 자체를 고쳐 쓰는 다중 에이전트 체계입니다[1]. 라텐트-와이(Latent-Y)는 문장 하나로 시작해 문헌 조사, 표적 분석, 항체 후보 설계, 실험 검증까지 이어지는 항체 개발 과정을 스스로 굴렸습니다[2]. 로봇 팔이 피펫을 잡는 웨트랩 조작을 미리 시뮬레이션으로 학습시키는 연구도 나왔고[3], 여러 제조사의 장비를 하나의 운영체제로 묶는 자율 실험실 소프트웨어도 공개되었습니다[4]. 진코 바이오웍스는 2026년 말까지 모든 연구개발을 자율 실험 설비 위로 옮기겠다고 밝혔습니다[10]. 버클리의 무기재료 합성 실험실은 사람의 개입 없이 17일을 연속으로 돌았습니다.

여기서 오해가 하나 생깁니다. 로봇이 다 하니 사람은 빠진다는 오해입니다. 실제로는 반대입니다. 자율 실험실이 늘어날수록 무엇을 물어볼지 정하는 사람의 몫이 커집니다. 기계는 실험을 빨리 반복할 뿐, 물어볼 가치가 있는 질문을 스스로 만들지 못합니다. 세포 반응을 컴퓨터 안에서 미리 재현하는 가상 세포 모델의 성능을 재는 벤치마크가 나왔지만, 실제 실험값과 어긋나는 구간이 여전히 넓습니다[6]. 부서를 나누지 않는 인공지능 중심 바이오텍이 가능한지를 따진 연구도 조직도를 그대로 에이전트로 옮기는 방식으로는 안 된다고 지적합니다[5].

루프가 끊기는 자리도 분명합니다. 컴퓨터 안에서 잘 붙던 분자가 실제 세포막을 통과하지 못하는 일은 흔합니다. 쥐에게는 안전했는데 사람 간세포에서 독성이 나오는 일도 그렇습니다. 예측 모델이 학습한 데이터는 이미 성공한 실험 쪽으로 기울어 있습니다. 실패한 실험은 논문이 되지 않아 기록에 남지 않기 때문입니다. 모델은 세상에 없는 데이터를 배울 수 없습니다. 자율 실험실이 값진 이유가 여기에도 있습니다. 사람이 부끄러워 버렸을 실패 기록을 로봇은 빠짐없이 남깁니다.

그래서 연구실의 사람 구성이 바뀌었습니다. 예전 제약사는 분자생물학자, 유기화학자, 생물정보학자가 각자의 층에서 순서대로 일하는 컨베이어 벨트였습니다. 지금은 두 언어를 함께 쓰는 사람이 프로젝트의 앞에 섭니다. 실험을 아는 사람은 모델이 내놓은 예측값을 그대로 믿지 않고, 그 숫자가 생물학적으로 말이 되는지 되묻습니다. 계산을 아는 사람은 배양 접시마다 달라지는 배치 효과와 시약 순도 편차를 모델의 불확실성 항으로 옮겨 적습니다. 이 두 사람이 한 사람이면 회의 시간이 사라집니다.

노동 시장의 숫자가 이 변화를 그대로 보여 줍니다. 2026년 2분기 미국 생물정보학 채용 공고는 631건이었고, 공개된 연봉은 14만 8천 달러에서 21만 5천 달러 사이였습니다[8]. 단백질 모델링과 실서비스 수준의 기계학습을 함께 다루는 선임 연구자는 기본급만 24만 5천 달러에서 32만 5천 달러를 받습니다[9]. 같은 해 미국 바이오제약 업계에서 감원된 인원은 약 4만 2700명으로, 2024년의 2만 9천 명보다 늘었습니다. 그런데 조사에 응한 바이오제약 기업의 64%는 지금도 사람을 뽑고 있고, 41%는 2026년에 채용 자리가 더 늘어날 것으로 봅니다[9]. 한쪽에서는 사람이 나가고 다른 쪽에서는 자리를 못 채웁니다. 제약사의 43%가 필요한 디지털 역량을 갖춘 지원자를 찾지 못한다고 답했습니다.

대학의 교육 과정도 뒤늦게 따라오고 있습니다. 생명과학과 학생이 미분방정식과 파이썬을 함께 듣고, 컴퓨터공학과 학생이 세포생물학 실습에 들어가는 복수 전공 과정이 늘었습니다. 문제는 학점표가 아니라 손입니다. 코드는 온라인 강의로 배울 수 있지만, 배지에 곰팡이가 피는 냄새와 원심분리기 균형을 못 맞췄을 때 나는 소리는 그 자리에 서 봐야 압니다. 반대로 실험만 십 년 한 사람은 자기 데이터가 왜 모델 학습에 쓸 수 없는 형식인지 모릅니다. 두 감각을 한 사람 안에 넣는 데는 시간이 걸립니다. 지금 채용난의 정체가 이것입니다.

저는 이 격차가 당분간 좁혀지지 않는 쪽에 걸겠습니다. 세포 배양을 십 년 한 사람이 파이썬을 배우는 데는 몇 달이면 됩니다. 그러나 모델이 왜 그 분자를 골랐는지 의심하고, 그 의심을 다음 주 실험 설계로 바꾸는 감각은 논문 몇 편으로 생기지 않습니다. 새벽 두 시에 피펫과 키보드를 번갈아 잡던 그 대학원생이 앞으로 십 년을 끌고 갈 사람입니다.

**참고자료**


[1] Jiang, Y., et al. (2026). A Self-Evolving Agentic System for Automated Generation and Execution of Biological Protocols. arXiv preprint arXiv:2606.31763. https://arxiv.org/abs/2606.31763

[2] Latent Labs Team, et al. (2026). Latent-Y: A Lab-Validated Autonomous Agent for De Novo Drug Design. arXiv preprint arXiv:2603.29727. https://arxiv.org/abs/2603.29727

[3] Liu, Z., et al. (2026). An Embodied Simulation Platform, Benchmark, and Data-Efficient Augmentation Framework for Wet-Lab Robotics. arXiv preprint arXiv:2606.12936. https://arxiv.org/abs/2606.12936

[4] Gao, J., et al. (2025). UniLabOS: An AI-Native Operating System for Autonomous Laboratories. arXiv preprint arXiv:2512.21766. https://arxiv.org/abs/2512.21766

[5] Wang, Y. (2026). Do AI-Native Biotechs Need Departments? Benchmarking Company World Models for AI-Driven Drug Development. arXiv preprint arXiv:2607.18696. https://arxiv.org/abs/2607.18696

[6] Brouwer, E. D., et al. (2026). AssayBench: An Assay-Level Virtual Cell Benchmark for LLMs and Agents. arXiv preprint arXiv:2605.10876. https://arxiv.org/abs/2605.10876

[7] Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589. https://doi.org/10.1038/s41586-021-03819-2

[8] CompBioJobs. (2026년 7월). Bioinformatics Job Market Q2 2026: 631 Jobs, $148K-$215K Avg, Salary Transparency Surges. https://www.compbiojobs.com/blog/bioinformatics-job-market-q2-2026

[9] KORE1. (2026). Biotech AI Hiring 2026: Where Bio Meets ML. https://www.kore1.com/biotech-ai-hiring-2026/

[10] Ginkgo Bioworks. (2026). Autonomous Lab. https://www.ginkgo.bio/autonomous-lab

## 1.2 생물학적 지식의 전파와 창출: 학계와 산업계의 변화

아침 여덟 시, 연구자는 커피를 앞에 두고 문헌 검색창을 엽니다. 어제 하루 사이에 올라온 관련 논문이 마흔 편입니다. 초록만 읽어도 두 시간이 걸립니다. 그중 서른 편은 읽지 않아도 됐을 논문이고, 읽지 않은 열 편 중 하나에 그가 6개월 동안 헤맨 문제의 답이 들어 있습니다. 어느 쪽이 어느 쪽인지는 읽어 봐야 압니다.

지식의 양은 이미 사람의 읽기 속도를 넘어섰습니다. 2026년 전 세계 학술 논문 발표량은 600만 편을 넘길 것으로 봅니다. 2025년의 550만 편에서 한 해 만에 50만 편이 늘어난 셈입니다[10]. 심사를 거치지 않고 먼저 공개하는 사전 공개 서버도 같은 속도로 부풀었습니다. 바이오아카이브(bioRxiv)와 메드아카이브(medRxiv)에는 2025년 한 해 동안 6만 4229편의 새 원고가 올라왔습니다. 2024년의 5만 6천여 편보다 많습니다. 이 중 약 80%는 나중에 학술지에 실리는데, 게재까지 평균 250일이 걸립니다. 논문이 되기 전에 이미 반년 넘게 세상에 나와 있는 것입니다.

읽는 속도가 아니라 쓰는 속도가 먼저 바뀌었습니다. 미국암연구학회에 2025년 제출된 원고 7177편 가운데 36%에서 인공지능이 쓴 문장이 검출되었습니다. 바이오아카이브에 논문을 올리는 생명과학 연구자의 생산성은 52.9% 늘었습니다[10]. 심사자를 구하지 못해 편집자들이 애를 먹는 상황도 같이 왔습니다. 제출은 늘었는데 읽어 줄 사람은 그대로입니다. 심사를 인공지능에 맡겨도 되는지 따진 연구는 모델이 논문의 결함을 잡아내는 능력과 잡아냈다고 주장하는 능력이 서로 다르다고 경고합니다[6].

옛 방식의 지식 습득은 손으로 쌓는 일이었습니다. 표적 단백질 하나를 파악하려면 선행 논문 수백 편을 직접 읽고 정리해야 했습니다. 그 과정에서 다른 팀이 이미 겪은 독성 실패 보고를 놓치고 같은 벽에 다시 부딪히는 일이 흔했습니다. 물리학이나 재료공학 쪽에서 나온 측정 기술은 학과 건물 하나만 건너면 되는 거리에 있는데도 몇 년씩 전해지지 않았습니다. 지식이 칸막이 안에 갇히는 이 현상을 지식 사일로(Knowledge Silo)라고 부릅니다. 같은 층에서 일하면서 서로 다른 층에 사는 상태입니다.

칸막이를 무너뜨린 것은 대규모 언어 모델(Large Language Model)과 생의학 지식 그래프입니다. 지식 그래프는 유전자, 단백질, 약물, 질병을 점으로 놓고 관계를 선으로 이어 붙인 지도입니다. 지하철 노선도를 떠올리면 쉽습니다. 역이 유전자이고 노선이

상호작용입니다. 어떤 약이 어느 역에서 출발해 몇 정거장 뒤의 질병에 닿는지 눈으로 따라갈 수 있습니다. 사람이 논문을 읽어 이 지도를 그리려면 평생이 걸리지만, 모델은 수백만 편의 논문 본문과 특허와 임상시험 보고서를 한꺼번에 읽어 선을 긋습니다. 문제는 이 지도가 매달 새로 그려진다는 점입니다. 생의학 지식 그래프가 커지고 형태가 바뀔 때 모델이 앞서 배운 것을 잊지 않도록 이어 학습시키는 방법이 2026년에 제시되었습니다[3]. 언어 모델이 생의학 전용 도구를 직접 호출하도록 훈련시킨 데이터셋도 공개되었습니다[1]. 대사체 분야에서는 흩어진 자료를 이어 학습한 전용 모델이 대사물질 사이의 관계 그래프를 예측해 내놓기 시작했습니다[2]. 미생물이 어떤 화합물을 만드는지 논문에서 뽑아내 구조화하는 문헌 채굴 체계도 나왔습니다[4].

이제 질문의 형태가 바뀝니다. 연구자는 "KRAS G12D 변이 췌장암에서 면역관문억제제가 듣지 않게 만드는 대사 우회 경로가 무엇이고, 그 경로를 막을 합성 치사 표적을 찾아 달라"고 씁니다. 모델은 효소 반응 네트워크, 동물 모델의 전사체 데이터, 실패한 임상 기록을 함께 뒤져 몇 초 만에 후보 목록과 실험 설계 초안을 내놓습니다. 몇 년 전이라면 박사후연구원 한 명이 두 달을 매달릴 일이었습니다.

여기서 한 걸음 더 나간 것이 인공지능 연구자 에이전트입니다. 퓨처하우스가 만든 로빈(Robin)은 가설을 세우고 실험을 설계하고 결과를 해석하는 과정을 스스로 돌려, 녹내장 약으로 쓰이던 리파수딜을 건성 황반변성 치료 후보로 지목했습니다. 착상에서 논문까지 걸린 시간은 두 달 반이었습니다[8]. 같은 회사의 코스모스(Kosmos)는 연구자 한 명이 6개월에 걸쳐 할 분석을 하루에 마친다고 밝혔는데, 결과의 정확도는 약 80%로 추정합니다. 나머지 20%가 문제입니다. 대장암 취약점을 찾는 다중 규모 에이전트 군집 연구는 언어 모델의 그럴듯한 추론과 포유류 세포의 실제 물리가 어긋나는 지점을 어떻게 메울지가 병목이라고 짚습니다[7]. 에이전트 시대의 과학 발견을 다시 정의하자는 제안도 같은 해에 나왔습니다[5].

지식이 퍼지는 통로도 바뀌었습니다. 예전에는 학회장 복도가 정보의 중심이었습니다. 발표 뒤 커피를 들고 서서 나누는 십 분짜리 대화에서 다음 연구의 방향이 정해졌습니다. 지금은 사전 공개 서버에 원고가 올라가는 순간 세계 어디의 연구실에서든 같은 날 그 방법을 따라 해 봅니다. 코드와 학습된 모델 가중치가 논문과 함께 공개되는 것이 관례가 되면서, 재현에 걸리는 시간이 몇 달에서 하루로 줄었습니다. 대신 검증되지 않은 주장이 그만큼 빨리 퍼집니다. 심사를 통과하지 않은 원고가 여섯 달 넘게 세상을 돌아다니는 동안, 그 결과를

인용한 후속 연구가 이미 여러 편 쌓입니다. 속도를 얻고 안전장치를 잃은 거래입니다.

산업의 돈은 이 흐름을 먼저 읽었습니다. 알파벳 산하 아이소모픽 랩스는 21억 달러 규모의 시리즈 B 투자를 받았습니다. 2026년 상반기 이 분야 벤처 투자액의 약 87%가 이 한 건이었습니다. 같은 회사는 일라이 릴리, 노바티스와 손잡았습니다. 릴리는 선급금 4500만 달러에 성과 연동 마일스톤 최대 17억 달러, 노바티스는 선급금 3750만 달러에 마일스톤 12억 달러 조건입니다. 두 계약을 합치면 30억 달러에 가깝습니다[9]. 2026년 1월에는 일라이 릴리와 엔비디아가 5년간 10억 달러를 함께 넣어 신약 발굴 전용 인공지능 연구소를 세우기로 했습니다. 사노피는 자가면역 분야 계약 하나에 선급금 1억 6천만 달러와 마일스톤 25억 6천만 달러를 걸었습니다. 리커전은 경쟁사 엑센시아를 6억 8800만 달러에 흡수했습니다.

숫자만 보면 축포 같지만, 임상 성적표는 아직 냉정합니다. 인공지능이 발굴한 프로그램은 173건이 임상 단계에 올라와 있고, 최종 허가를 받은 약은 아직 하나도 없습니다. 임상 1상 통과율은 80%를 넘지만 2상에서 40% 근처로 떨어집니다. 그래도 저는 이 변화가 거품이 아니라고 봅니다. 인실리코 메디슨은 표적 발굴부터 임상 2상까지 30개월 안에 도달했습니다. 같은 구간을 사람의 손으로 걸으면 6년에서 8년이 걸립니다. 약이 듣는지는 아직 모릅니다. 그러나 시계는 이미 다르게 갑니다.

학계와 산업계의 관계도 뒤집혔습니다. 예전에는 대학이 기초 원리를 밝혀 논문을 내면, 몇 년 뒤 제약사가 그것을 주워 개발에 들어갔습니다. 지금은 계산 자원을 쥔 기업이 기초 생물학 모델을 먼저 만들고, 대학이 그 모델을 빌려 씁니다. 스크리닝 장비 하나 없는 소규모 연구실도 클라우드에서 수억 개의 가상 분자를 걸러 볼 수 있습니다. 반대로 대형 제약사는 대학 병원이 모은 환자 검체 데이터를 학습 재료로 받아 갑니다. 서로가 서로의 부품이 되었습니다.

이 뒤바뀜에는 그늘도 있습니다. 기초 생물학 모델을 훈련시키는 데 필요한 계산 자원은 웬만한 대학의 한 해 연구비를 넘어섭니다. 모델을 만드는 쪽과 빌려 쓰는 쪽이 갈리면, 무엇을 연구할지 정하는 힘도 만드는 쪽으로 기웁니다. 기업이 공개하지 않기로 한 학습 데이터는 논문에 실리지 않고, 실리지 않은 것은 검증되지 않습니다. 그래서 공개 데이터셋과 공개 벤치마크를 유지하는 일이 예전보다 중요해졌습니다. 누구나 같은 시험지로 채점할 수 있어야 주장과 광고를 구별할 수 있습니다.

연구자의 하루도 모양이 달라졌습니다. 아침에 문헌을 읽던 두 시간이 삼십 분으로 줄고, 대신 모델이 정리해 준 요약이 맞는지 원문을 되짚는 시간이 생겼습니다. 요약이 틀리는 지점은 늘 정해져 있습니다. 실험 조건의 예외, 통계 검정의 전제, 저자가 조심스럽게 붙여 둔 단서 문장입니다. 이 세 곳을 확인하는 습관이 있느냐 없느냐가 연구자의 실력을 가릅니다. 도구가 빨라질수록 의심하는 훈련이 값이 나갑니다.

그래서 연구자에게 남는 질문은 하나로 좁혀집니다. 무엇을 외우고 있느냐가 아니라, 무엇을 물어볼 것이냐입니다. 답을 찾는 일은 기계가 빨라졌습니다. 아직 아무도 묻지 않은 질문을 알아보는 눈은 여전히 사람 쪽에 있습니다.

**참고자료**


[1] Gao, X., et al. (2026). BioTool: A Comprehensive Tool-Calling Dataset for Enhancing Biomedical Capabilities of Large Language Models. arXiv preprint arXiv:2605.05758. https://arxiv.org/abs/2605.05758

[2] Ku, D., et al. (2026). MetaboLLM: a metabolomics-specialized large language model for biochemical knowledge integration and predictive metabolite graph construction. arXiv preprint arXiv:2608.06253. https://arxiv.org/abs/2608.06253

[3] Radwan, Y. A., et al. (2026). CMKL: Modality-Aware Continual Learning for Evolving Biomedical Knowledge Graphs. arXiv preprint arXiv:2605.10510. https://arxiv.org/abs/2605.10510

[4] Sun, X., et al. (2025). Literature Mining System for Nutraceutical Biosynthesis: From AI Framework to Biological Insight. arXiv preprint arXiv:2512.22225. https://arxiv.org/abs/2512.22225

[5] Zheng, Y., et al. (2026). Rethinking Scientific Discovery in the Agentic Era. arXiv preprint arXiv:2607.03863. https://arxiv.org/abs/2607.03863

[6] Wang, J., et al. (2026). When AI reviews science: Can we trust the referee?. arXiv preprint arXiv:2604.23593. https://arxiv.org/abs/2604.23593

[7] Baker, C., et al. (2026). Autonomous mechanistic discovery of colorectal cancer vulnerabilities via multi-scale AI swarms. arXiv preprint arXiv:2607.16262. https://arxiv.org/abs/2607.16262

[8] Chemical & Engineering News. (2026년 5월). AI companies introduce new agent-based tools for scientific discovery. https://cen.acs.org/pharmaceuticals/drug-discovery/ai-companies-introduce-agent-based-research-tools/104/web/2026/05

[9] BioSpace. (2026). Lilly, Novartis Sign AI Partnership with Alphabet's Isomorphic. https://www.biospace.com/lilly-novartis-sign-ai-partnership-with-alphabet-s-isomorphic

[10] Futurity Publishing. (2026). AI Is Reshaping Scientific Publishing in 2026: Biology and Medicine Are Growing Fastest. https://futurity-publishing.com/ai-scientific-publishing-2026/

# 제2장: 생물학 데이터의 기하학적 표현 학습과 다중 모달 정렬

## 2.1 단백질, 유전자, 텍스트 언어 간의 모달리티 정렬(Multimodal Alignment)과 대조 학습

실험실 화이트보드 앞에 세 사람이 서 있습니다. 같은 단백질 하나를 두고 각자 다른 노트를 폅니다. 서열을 다루는 사람은 스무 종 아미노산의 약자를 줄줄이 적습니다. 구조를 다루는 사람은 원자마다 x, y, z 좌표 세 숫자를 적습니다. 논문을 읽는 사람은 이렇게 씁니다. "이 효소는 세포 주기 진입 단계에서 표적 단백질의 세린 잔기에 인산기를 붙인다." 세 노트는 글자 하나 겹치지 않습니다. 그런데 셋은 같은 분자를 가리킵니다.

생명은 정보를 여러 번 갈아입힙니다. 유전 정보는 네 글자로 된 1차원 염기 서열에 보관됩니다. 전사와 번역을 거치면 아미노산 사슬로 바뀝니다. 사슬은 원자 사이의 끌어당김과 밀어냄에 따라 3차원 공간에서 접힙니다. 접힌 단백질은 다른 분자와 붙어 신호를 주고받습니다. 그 결과를 인간은 논문의 문장으로 적어 둡니다. 형태는 네 번 바뀌었지만 실체는 하나입니다.

지난 십 년의 모델은 이 중 하나만 붙들었습니다. 서열만 읽는 언어 모델이 있었고, 좌표만 읽는 구조 모델이 있었습니다. 문헌만 읽는 텍스트 모델은 따로 놀았습니다. 그래서 서열은 알지만 구조가 없는 단백질 앞에서 멈췄습니다. 구조는 풀렸지만 기능 설명이 없는 신규 표적 앞에서도 멈췄습니다. 세 노트를 나란히 놓고 읽을 사람이 없었던 셈입니다.

모달리티 정렬(Multimodal Alignment)은 형태가 다른 데이터를 하나의 잠재 공간(Latent Space)에 함께 꽂아 넣는 방법입니다. 도서관 사서가 책을 주제별로 붙여 꽂는 일과 닮았습니다. 표지 색이 달라도 같은 주제면 옆자리에 둡니다. 이 배치를 학습으로 만드는 알고리즘이 대조 학습(Contrastive Learning)입니다. 같은 단백질에서 나온 서열 벡터, 구조 벡터, 설명문 벡터를 긍정 쌍(Positive Pair)으로 묶습니다. 관계없는 단백질끼리는 부정 쌍(Negative Pair)이 됩니다. 학습은 InfoNCE 손실 함수로 진행합니다 [1]. 긍정 쌍은 서로 당기고, 부정 쌍은 서로 밀어냅니다. 이미지와 문장을 이 방식으로 붙인 CLIP이 2021년에 길을 열었고 [2], 생물학은 그 문법을 그대로 가져왔습니다 [3].

정렬이라는 말이 추상적으로 들릴 수 있으니 숫자로 옮겨 보겠습니다. 서열 인코더는 아미노산 사슬을 받아 1,024개 숫자로 된 벡터 하나를 뱉습니다. 구조 인코더는 원자 좌표를 받아 같은 길이의 벡터를 뱉습니다. 텍스트 인코더도 문장을 받아 같은 길이의

벡터를 뱉습니다. 세 벡터는 처음에는 아무 관계가 없습니다. 학습이 진행되면 같은 단백질에서 나온 세 벡터의 코사인 유사도가 0.1에서 0.9 근처까지 올라갑니다. 손실 함수 안에는 온도라 불리는 작은 숫자가 하나 들어 있습니다. 이 값을 낮추면 모델은 헷갈리는 부정 쌍에 더 매섭게 반응합니다. 값을 높이면 느슨하게 봅니다. 논문 한 편의 성패가 이 숫자 하나에 갈리는 일이 실제로 벌어집니다.

성과는 검색에서 먼저 나왔습니다. 서열과 구조와 기능 텍스트를 한꺼번에 대조 학습한 ProTrek은 Swiss-Prot에서 고른 정밀 쌍 1,400만 건과 UniRef50에서 거른 잡음 쌍 2,500만 건, 합쳐 4,000만 건으로 학습했습니다 [4]. 서열에서 기능을 찾는 검색은 기존 대비 30배, 기능 설명에서 서열을 찾는 검색은 60배 좋아졌습니다. 단백질끼리 비교하는 속도는 Foldseek과 MMseqs2보다 100배 빨랐습니다. 열한 개 하위 과제 중 아홉 개에서 ESM-2를 앞섰습니다. 이 모델의 공개 서버는 50억 개 단백질의 임베딩을 미리 계산해 두고 있습니다.

학습 데이터의 질이 결과를 가릅니다. Swiss-Prot의 설명문은 사람이 손으로 검토한 문장입니다. UniRef50에서 긁어 온 2,500만 건은 그렇지 않습니다. 잡음 쌍을 넣으면 정밀도가 떨어질 것 같지만, 실제로는 반대였습니다. 넓고 거친 데이터가 좁고 깨끗한 데이터의 빈틈을 메웁니다. 사전을 만들 때 표준어만 모으면 현장에서 쓸 수 없는 것과 같습니다. 이 모델의 공개 서버에는 NCBI의 단백질 서열 7억 건이 나중에 더 붙었습니다.

2026년 6월에는 세 모달리티를 아예 한 몸으로 만든 모델이 나왔습니다. BioMatrix는 분자 서열과 분자 구조, 단백질 서열과 단백질 구조, 그리고 자연어를 모두 같은 이산 토큰 공간으로 옮깁니다 [5]. 외부 인코더도, 모달리티별 출력 헤드도 없습니다. Qwen3 언어 모델의 17억과 40억 파라미터 판을 바탕으로 3,044억 토큰을 더 학습했습니다. 여섯 범주 80개 과제 중 77개에서 최고 성능이나 그에 준하는 성능을 냈습니다. 읽을 수 있는 모달리티를 그대로 생성까지 할 수 있다는 점이 이전 모델과 갈리는 지점입니다.

구조를 토큰으로 바꾸는 기술도 가벼워졌습니다. 2026년 5월에 공개된 Yeti는 조회 없는 양자화(Look-up Free Quantization)로 원자 좌표를 이산 토큰으로 바꿉니다 [6]. 파라미터가 ESM3의 10분의 1인데 코드북 사용률과 토큰 다양성이 앞섰습니다. 사전 학습 없이 처음부터 서열과 구조를 함께 생성하도록 훈련했는데, 10배 큰 모델과 비슷한 결과를 냈습니다. 무거운 좌표를 가벼운 글자로 바꾸는 일이 이제 연구실 한 대의 장비로

가능해졌습니다.

이 정렬이 완성되면 검색창이 바뀝니다. 연구자가 "인산화효소의 활성 포켓에 붙으면서 수소 결합을 셋 이상 만드는 도메인"이라고 문장으로 칩니다. 모델은 그 문장의 벡터와 가까운 자리에 놓인 구조와 서열을 꺼내 옵니다. 실험을 한 번도 하지 않은 단백질도 후보에 오릅니다. 단백질 쌍이 어떻게 맞물리는지를 자유 문장으로 설명해 내는 모델도 2026년 5월에 나왔고 [7], RNA 설계에서도 같은 정렬 방식이 제로샷 조건부 생성으로 이어졌습니다 [8].

그런데 정렬이 곧 합체는 아닙니다. 대조 학습으로 붙여 놓은 두 모달리티는 실제로 같은 자리에 겹치지 않습니다. 이미지 벡터와 텍스트 벡터는 각자 좁은 원뿔 모양 영역에 몰려 있고, 그 사이에는 빈 띠가 남습니다. 이것을 모달리티 간극(Modality Gap)이라 부릅니다 [9]. 신경망은 층을 지날수록 벡터 사이 각도를 좁히는 성질이 있어서, 초기화 단계에서 이미 두 덩어리가 갈라진 채 굳습니다. 코사인 유사도로 순위를 매기는 데는 문제가 없습니다. 두 공간이 정말 하나가 되었다고 말하는 순간 틀립니다.

왜 이 간극이 남는지에 대한 설명은 아직 하나로 모이지 않았습니다. 초기화 단계의 기하 때문이라는 설명이 있고, 대조 손실 자체에 내재한 성질이라는 설명이 있습니다. 학습이 덜 되어서라는 설명도, 데이터에 실린 편향 때문이라는 설명도 있습니다. 온도 값, 벡터 길이, 두 모달리티의 데이터 양 차이가 모두 간극의 폭을 움직입니다. 실무에서 이것이 문제가 되는 지점은 검색이 아니라 생성입니다. 텍스트 벡터가 놓인 자리에서 곧장 구조를 만들어 내려 하면, 그 자리는 구조 벡터들이 사는 동네가 아닙니다. 빈 띠 위에 집을 짓는 셈입니다.

이산 토큰으로 옮기는 과정에서 치르는 대가도 측정되었습니다. 2026년 4월 논문은 이것을 기하 정렬 세금(Geometric Alignment Tax)이라 이름 붙였습니다 [10]. 연속된 곡면을 억지로 이산 칸에 밀어 넣으면 원래의 거리 관계가 망가집니다. 같은 인코더에서 교차 엔트로피 대신 연속 출력 헤드를 쓰자 기하 왜곡이 8.5배까지 줄었습니다. 더 뼈아픈 숫자는 따로 있습니다. 연속 목적함수에서 세 구조의 차이는 1.3배였는데, 이산 토큰화에서는 3,000배로 벌어졌습니다. 토큰화는 공짜가 아닙니다.

정렬된 공간은 실험 순서도 바꿉니다. 예전에는 표적 단백질을 정한 뒤 결합 후보를 수십만 개 합성해 걸러 냈습니다. 지금은 원하는 성질을 문장으로 적고, 그 문장 근처에 놓인 후보 수백 개만 골라 실험합니다. 2026년 4월에 나온 연구는 촉매 잔기를 미리 지정하지 않고도

새 효소를 설계하는 데까지 나아갔습니다 [11]. 서열과 구조와 기능 설명을 함께 다루는 모델이 있었기에 가능한 일입니다. 화학이 탐색해 온 반응은 진화가 만들어 낸 것의 일부였고, DNA가 적어 낼 수 있는 화학은 그보다 훨씬 넓습니다.

저는 이 기술을 번역으로 부르는 편이 정직하다고 봅니다. 세 언어를 나란히 놓고 서로 옮겨 적는 사전이 생겼습니다. 사전이 있으니 검색은 빨라졌고, 실험실이 감당할 후보 목록은 짧아졌습니다. 하지만 좋은 사전이 있다고 두 언어가 한 언어가 되지는 않습니다. 다음 세대의 과제는 모델을 키우는 쪽이 아니라, 세금을 덜 내는 표현 방식을 찾는 쪽에 있습니다.

**참고자료**


[1] van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation Learning with Contrastive Predictive Coding. arXiv preprint arXiv:1807.03748. https://arxiv.org/abs/1807.03748

[2] Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020. https://arxiv.org/abs/2103.00020

[3] Xu, M., Yuan, X., Miret, S., & Tang, J. (2023). ProtST: Multi-Modality Learning of Protein Sequences and Biomedical Texts. arXiv preprint arXiv:2301.12040. https://arxiv.org/abs/2301.12040

[4] Su, J., He, Y., You, S., et al. (2025). A trimodal protein language model enables advanced protein searches. Nature Biotechnology. https://doi.org/10.1038/s41587-025-02836-0

[5] Pei, Q., et al. (2026). BioMatrix: Towards a Comprehensive Biological Foundation Model Spanning the Modality Matrix of Sequences, Structures, and Language. arXiv preprint arXiv:2606.22138. https://arxiv.org/abs/2606.22138

[6] Giri, N., Farrell, S., & Bouchard, K. E. (2026). Yeti: A compact protein structure tokenizer for reconstruction and multi-modal generation. arXiv preprint arXiv:2605.09981. https://arxiv.org/abs/2605.09981

[7] Fei, X., et al. (2026). PPI2Text: Captioning Protein-Protein Interactions with Coordinate-Aligned Pair-Map Decoding. arXiv preprint arXiv:2605.08924. https://arxiv.org/abs/2605.08924

[8] Klypa, R., et al. (2026). Multimodal Alignment and Preference Optimization for Zero-Shot Conditional RNA Generation. arXiv preprint arXiv:2605.23961. https://arxiv.org/abs/2605.23961

[9] Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., & Zou, J. (2022). Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. arXiv preprint arXiv:2203.02053. https://arxiv.org/abs/2203.02053

[10] Raju, P. C. (2026). The Geometric Alignment Tax: Tokenization vs. Continuous Geometry in Scientific Foundation Models. arXiv preprint arXiv:2604.04155. https://arxiv.org/abs/2604.04155

[11] Rector-Brooks, J., et al. (2026). General Multimodal Protein Design Enables DNA-Encoding of Chemistry. arXiv preprint arXiv:2604.05181. https://arxiv.org/abs/2604.05181

## 2.2 고차원 오믹스 데이터를 저차원 잠재 표현(Latent Representation) 공간으로 압축하는 임베딩 원리

세포 하나를 시퀀싱 장비에 넣으면 숫자 2만 개가 나옵니다. 유전자 하나마다 몇 번 읽혔는지를 센 값입니다. 화면에 뜬 표는 가로로 끝없이 이어집니다. 그런데 그 2만 칸 중 대부분은 0입니다. 유전자가 정말 꺼져 있어서 0인 칸과, 켜져 있었는데 장비가 놓쳐서 0인 칸이 섞여 있습니다. 뒤쪽을 드롭아웃(Dropout)이라고 부릅니다. 표를 처음 열어 본 대학원생은 이 두 종류의 0을 구별할 방법이 없다는 사실에 먼저 좌절합니다.

숫자가 많다고 정보가 많지는 않습니다. 2만 차원 공간에서는 모든 점이 서로 비슷하게 멀어집니다. 거리를 재도 순위가 잘 갈리지 않습니다. 이것을 차원의 저주(Curse of Dimensionality)라고 부릅니다. 잡음까지 얹히면 두 세포가 같은 종류인지 다른 종류인지 판정하는 일조차 흔들립니다.

다행히 유전자들은 제각기 놀지 않습니다. 세포 주기에 묶인 유전자 무리는 함께 켜집니다. 에너지 대사에 묶인 무리도, 면역 반응에 묶인 무리도 그렇습니다. 사무실 조명 2만 개가 있지만 스위치는 수십 개인 상황과 비슷합니다. 표면의 차원은 2만이어도, 실제로 움직이는 손잡이는 훨씬 적습니다. 이 얇은 곡면을 저차원 매니폴드(Low-Dimensional Manifold)라고 부릅니다.

변분 오토인코더(VAE, Variational Autoencoder)는 그 손잡이를 찾아내는 장치입니다. 인코더 신경망이 2만 차원 벡터를 받아 32차원이나 64차원의 확률 분포로 줄입니다. 디코더 신경망이 그 짧은 벡터만 보고 원래의 2만 개 숫자를 되살립니다. 학습은 두 가지를 함께 줄입니다. 복원이 얼마나 틀렸는지를 재는 재구성 오차, 그리고 잠재 분포가 표준 분포에서 얼마나 벗어났는지를 재는 쿨백-라이블러 발산입니다. 단일 세포 데이터에 이 구조를 처음 제대로 얹은 모델이 2018년의 scVI입니다 [1]. 음이항 분포로 읽힘 수를 모형화해 드롭아웃과 실제 0을 갈라 놓았습니다.

디코더가 하는 일을 조금 더 들여다보면 이 구조가 왜 잡음에 강한지 보입니다. 디코더는 32개 숫자만 들고 2만 개 숫자를 만들어 내야 합니다. 정보가 턱없이 부족합니다. 그래서 모델은 유전자들이 함께 움직이는 규칙을 외울 수밖에 없습니다. 세포 주기 유전자 하나가 켜져 있으면 나머지도 켜져 있을 확률이 높다는 규칙 같은 것입니다. 이 규칙이 몸에 배면,

장비가 놓쳐 0으로 찍힌 칸도 주변 유전자를 보고 메울 수 있습니다. 압축은 결과적으로 결측값 보정 장치가 됩니다.

압축은 잡음을 버리고 뜻을 남깁니다. 장비 배치가 달라 생긴 차이는 복원에 별 도움이 안 되므로 잠재 벡터에 자리를 잡지 못합니다. 반대로 세포의 분화 단계, 조직 정체성, 종양의 진행 정도는 복원에 필요하므로 살아남습니다. 짧아진 벡터 위에서는 거리를 재는 일이 다시 뜻을 가집니다. 어떤 희귀암 환자의 전사체 벡터가 면역항암제 반응군 무리 가까이에 놓인다면, 같은 약을 시도해 볼 근거가 됩니다. 정상 세포에서 종양으로 가는 궤적을 거꾸로 따라가면 갈림길에 선 전사 인자를 지목할 수도 있습니다.

잠재 공간은 그림으로 보는 순간 사람을 홀립니다. 2만 차원을 두 개의 축으로 눌러 그린 UMAP 지도는 색색의 섬처럼 보입니다 [10]. 섬 사이 거리가 세포 사이 관계처럼 읽힙니다. 그런데 이 지도는 이웃 관계만 지키고 전체 거리는 지키지 않습니다. 두 섬이 멀리 떨어져 그려졌다고 두 세포 무리가 실제로 먼 것은 아닙니다. 발표 자료에 쓰기 좋은 그림일수록 조심해서 읽어야 합니다. 판정은 눌러 그린 2차원이 아니라 32차원 벡터 위에서 해야 합니다.

여기까지가 교과서의 이야기입니다. 2023년 이후 이 자리에는 훨씬 큰 모델들이 들어왔습니다. 3,300만 개 세포로 학습한 scGPT 같은 단일 세포 파운데이션 모델이 대표입니다 [2]. 유전자를 토큰으로 보고 트랜스포머에 넣는 방식입니다. 그런데 2026년의 검증 결과는 이 흐름에 제동을 걸었습니다.

2026년 2월 18일 공개된 논문 한 편이 뼈아픕니다 [3]. 보스턴대의 두 연구자는 정규화를 꼼꼼히 하고 선형 방법만 쓴 파이프라인으로 여러 벤치마크에서 파운데이션 모델과 같거나 나은 점수를 냈습니다. 학습에 없던 세포 유형과 없던 종을 다루는 분포 밖 과제에서는 큰 모델을 앞섰습니다. 세포 정체성의 생물학이 선형 표현으로 상당 부분 잡힌다는 뜻입니다. 같은 해 다른 벤치마크들도 비슷한 말을 했습니다. 제로샷 임베딩이 고변이 유전자만 골라 쓴 기준선보다 못한 경우가 여러 생물계에서 나왔습니다.

왜 이런 결과가 나왔을까요. 자연어와 달리 유전자 발현 표에는 순서가 없습니다. 문장에서 단어의 자리는 뜻을 바꾸지만, 표에서 유전자 열의 자리는 아무 뜻도 없습니다. 트랜스포머가 잘하는 일은 순서와 문맥을 읽는 일인데, 그 재능이 쓰일 자리가 애초에 좁았던 셈입니다. 게다가 벤치마크 자체가 헐거웠습니다. 세포 유형이 뚜렷하게 갈리는

데이터에서는 주성분 분석만으로도 점수가 잘 나옵니다. 어려운 것은 세포 무리가 서로 겹치는 조직이고, 그런 데이터에서는 모든 방법이 함께 흔들립니다.

큰 모델 안에 뭐가 들어 있는지도 파헤쳐졌습니다. 2026년 2월 25일 논문은 141개의 기하·위상 가설을 52차례에 걸쳐 자동으로 세우고 검증했습니다 [4]. 27건은 긍정, 21건은 애매, 63건은 확실한 부정이었습니다. 제일 엄한 기준을 적용하면 15건도 살아남지 못했고, 견고한 양성 비율은 10퍼센트 언저리로 떨어졌습니다. 그래도 남은 것은 있습니다. 유전자 임베딩의 이웃 구조에는 진짜 위상이 있었고, 곡면을 따라 잰 거리가 유클리드 거리보다 조절 관계를 잘 찾아냈습니다. scGPT와 Geneformer를 정준상관분석으로 맞춰 보면 상관이 0.80까지 나왔습니다. 두 모델은 유전자 공간의 모양에는 동의했습니다. 어느 유전자가 어디에 놓이는지에는 동의하지 않았습니다.

그러니 이 모델들을 버릴 일도 아닙니다. 2026년 3월에는 scGPT 내부에서 조혈 계통의 분화를 설명하는 곡면을 찾아내고, 그 곡면에서 실제로 쓸 만한 알고리즘을 뽑아낸 연구가 나왔습니다 [5]. 구조를 조금 바꾸는 것만으로 표현이 좋아지기도 합니다. 소프트맥스 어텐션을 시그모이드 어텐션으로 바꿨더니 여섯 개 데이터셋에서 세포 유형 분리도가 25퍼센트 올라갔습니다 [6]. 희소 오토인코더로 내부를 뜯어보면 생물학 지식은 정연하게 정리되어 있고, 조절 논리는 거의 없다는 진단도 나왔습니다 [7].

이 대목에서 저는 한쪽 편을 들겠습니다. 지난 삼 년 동안 이 분야는 파라미터 개수를 성적표로 삼았습니다. 그 습관이 논문을 늘렸지만 임상 현장에는 별로 닿지 않았습니다. 병원에서 필요한 것은 처음 보는 환자의 조직에서도 흔들리지 않는 표현입니다. 학습 데이터에 없던 세포 유형 앞에서 선형 방법이 이겼다는 결과는, 큰 모델이 데이터의 결을 외웠을 뿐 생물학의 규칙을 배우지는 못했다는 신호로 읽힙니다.

돈과 관심은 모델 아래층으로 내려가는 중입니다. 1억 개 세포와 6만 건 약물 처리 실험을 담은 Tahoe-100M이 공개되었고, 암 모델 50종에 약물 1,100종 이상을 걸어 본 기록이 들어 있습니다 [8]. Arc Institute의 가상 세포 아틀라스는 6억 개가 넘는 세포를 모았습니다. 이미지 쪽에서는 JUMP Cell Painting 한 세트가 115테라바이트에 이릅니다 [9]. 2026년 가상 세포 챌린지에는 114개국에서 5,000명 넘게 등록했는데, 상위권을 차지한 접근은 거대 모델이 아니라 의사벌크 요약 통계와 순위 검정이었습니다.

측정 방식도 함께 늘고 있습니다. 유전자 발현만이 아니라 염색질이 얼마나 열려 있는지, 단백질에 인산기가 붙었는지를 같은 세포에서 동시에 읽습니다. 조직 절편 위 좌표까지 붙인 공간 전사체도 흔해졌습니다. 축이 늘어나면 압축은 더 어려워집니다. 서로 다른 측정치를 하나의 잠재 벡터에 넣으려면, 어느 축을 얼마나 믿을지부터 정해야 합니다. 앞 절에서 본 모달리티 간극이 여기서도 그대로 나타납니다. 세포 하나를 두고 만든 두 종류의 벡터가 같은 동네에 살지 않습니다.

압축이라는 방향은 옳았습니다. 2만 개 숫자를 수십 개로 줄이는 일은 여전히 이 분야의 중심 작업입니다. 다만 압축기를 크게 만드는 일과 잘 압축하는 일은 별개였습니다. 2026년의 성적표는 그 둘을 갈라 놓았습니다. 좋은 정규화, 정직한 기준선, 분포 밖 검증. 이 세 가지를 건너뛴 모델은 벤치마크에서 이겨도 새 조직 앞에서 무너집니다.

**참고자료**


[1] Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nature Methods, 15(12), 1053-1058. https://doi.org/10.1038/s41592-018-0229-2

[2] Cui, H., Wang, C., Maan, H., et al. (2024). scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nature Methods, 21(8), 1470-1480. https://doi.org/10.1038/s41592-024-02201-0

[3] Souza, H., & Mehta, P. (2026). Parameter-free representations outperform single-cell foundation models on downstream benchmarks. arXiv preprint arXiv:2602.16696. https://arxiv.org/abs/2602.16696

[4] Kendiukhov, I. (2026). What Topological and Geometric Structure Do Biological Foundation Models Learn? Evidence from 141 Hypotheses. arXiv preprint arXiv:2602.22289. https://arxiv.org/abs/2602.22289

[5] Kendiukhov, I. (2026). Discovery of a Hematopoietic Manifold in scGPT Yields a Method for Extracting Performant Algorithms from Biological Foundation Model Internals. arXiv preprint arXiv:2603.10261. https://arxiv.org/abs/2603.10261

[6] Sadashivaiah, V., et al. (2026). Better Models, Faster Training: Sigmoid Attention for single-cell Foundation Models. arXiv preprint arXiv:2604.27124. https://arxiv.org/abs/2604.27124

[7] Kendiukhov, I. (2026). Sparse autoencoders reveal organized biological knowledge but minimal regulatory logic in single-cell foundation models: a comparative atlas of Geneformer and scGPT. arXiv preprint arXiv:2603.02952. https://arxiv.org/abs/2603.02952

[8] Zhang, J., et al. (2025). Tahoe-100M: A Giga-Scale Single-Cell Perturbation Atlas for Context-Dependent Gene Function and Cellular Modeling. bioRxiv. https://doi.org/10.1101/2025.02.20.639398

[9] Muñoz, A. F., et al. (2026). JUMP-lite: Compact, reproducible benchmarking of cell representations. arXiv preprint arXiv:2608.07632. https://arxiv.org/abs/2608.07632

[10] Becht, E., McInnes, L., Healy, J., et al. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnology, 37(1), 38-44. https://doi.org/10.1038/nbt.4314

# 제3장: 단백질 3차원 구조 예측 혁신

## 3.1 아미노산 서열 기반의 고정밀 입체 구조 예측 원리와 수학적 기초

이야기는 시험관 하나에서 시작합니다. 1950년대 후반, 미국 국립보건원의 크리스천 안핀센(Christian Anfinsen)은 리보뉴클레이스(Ribonuclease)라는 효소를 요소 용액에 담갔습니다. 리보뉴클레이스는 RNA를 잘라 내는 가위 같은 단백질입니다. 요소는 접혀 있던 사슬을 풀어 헤칩니다. 실험대 위 유리관 안에서 효소는 엉킨 실타래가 되었습니다. RNA를 자르는 능력도 함께 사라졌습니다.

안핀센은 여기서 멈추지 않았습니다. 그는 요소를 천천히 씻어 냈습니다. 풀렸던 사슬이 스스로 다시 접혔습니다. 가위는 다시 RNA를 잘랐습니다. 아무도 손을 대지 않았는데 사슬이 제 형태를 찾아간 것입니다. 이 실험에서 나온 결론이 안핀센의 도그마(Anfinsen's Dogma)입니다. 단백질의 3차원 형태는 아미노산 서열 안에 이미 적혀 있습니다. 자연 상태의 구조는 자유 에너지가 제일 낮은 골짜기입니다 [1]. 안핀센은 1972년에 노벨 화학상을 받았습니다.

단백질에서 형태는 곧 기능입니다. 효소는 기질이 딱 들어맞는 홈을 가지고 있어야 반응을 촉매합니다. 항체는 병원체 표면과 정확히 포개지는 팔을 가져야 병원체를 붙잡습니다. 콜라겐은 세 가닥이 밧줄처럼 꼬여야 피부와 힘줄을 버팁니다. 서열이 같아도 형태가 무너지면 그 단백질은 아무 일도 하지 못합니다. 그래서 서열에서 형태를 읽어 내는 일은 생명과학의 오래된 숙제였습니다.

문제는 계산이었습니다. 1969년 사이러스 레빈탈(Cyrus Levinthal)은 종이 위에서 경우의 수를 세어 보았습니다 [2]. 아미노산 100개짜리 작은 단백질을 생각합니다. 주쇄의 파이 각과 프사이 각, 곁사슬의 카이 각이 저마다 몇 가지 값을 가질 수 있습니다. 각도를 아주 거칠게 잡아도 가능한 형태의 수가 10의 47제곱을 넘어섭니다. 한 형태를 확인하는 데 1조 분의 1초가 걸린다고 해도, 전부 훑으려면 우주의 나이보다 긴 시간이 필요합니다. 그런데 실제 세포 안에서 단백질은 1초도 안 되어 접힙니다. 이것이 레빈탈의 역설입니다.

역설의 해답은 지형에 있었습니다. 단백질이 모든 형태를 하나씩 확인하는 것이 아닙니다. 자유 에너지 지형은 평평한 벌판이 아니라 깔때기 모양입니다. 사슬은 깔때기 벽을 따라 미끄러집니다. 어느 방향으로 굴러도 아래로 내려가고, 아래로 내려갈수록 갈 수 있는 길이 좁아집니다. 계단을 하나씩 세어 가며 내려가는 것이 아니라, 미끄럼틀을 타고 바닥에 닿는

셈입니다. 그러니 서열 안에 형태가 적혀 있다는 말은 이렇게 바꿔 읽을 수 있습니다. 서열이 깔때기의 모양을 정합니다.

돌파구는 물리 계산이 아니라 역사에서 나왔습니다. 지금 우리 손에 있는 단백질 서열은 수억 년 진화의 기록입니다. 사람의 헤모글로빈과 물고기의 헤모글로빈은 서열이 절반쯤 다릅니다. 그런데 접힌 모양은 거의 같습니다. 형태가 살아남는 동안 글자만 바뀐 것입니다. 여기서 단서가 나옵니다. 3차원 공간에서 서로 맞닿아 있는 두 자리는 함께 변합니다.

지하철 좌석을 떠올리면 쉽습니다. 나란히 앉은 두 사람 중 한 사람이 커다란 배낭을 무릎에 올리면, 옆 사람은 어깨를 움츠려 자리를 내줍니다. 두 사람이 붙어 앉아 있기 때문에 생기는 반응입니다. 단백질도 같습니다. 15번 자리의 아미노산이 부피가 큰 것으로 바뀌면, 3차원에서 그 옆에 놓인 80번 자리는 작은 아미노산으로 바뀝니다. 반대로 두 자리가 멀리 떨어져 있으면 서로 눈치를 볼 이유가 없습니다. 이 짝지어진 변화를 공진화(Co-evolution)라고 부릅니다.

딥러닝 모델은 이 신호를 대량으로 읽습니다. UniRef나 BFD 같은 거대한 서열 창고에서 목표 단백질과 닮은 서열 수만 개를 끌어옵니다. 그것들을 위아래로 줄 세운 표가 다중 서열 정렬(MSA, Multiple Sequence Alignment)입니다. 도서관 열람실 책상에 같은 문장의 판본 수만 장을 나란히 펼쳐 놓은 모습입니다. 어떤 칸은 어느 판본에서나 똑같습니다. 어떤 두 칸은 늘 짝을 이루어 함께 바뀝니다.

문제는 신호가 번져 나간다는 점입니다. A와 B가 붙어 있고 B와 C가 붙어 있으면, A와 C도 붙어 있는 것처럼 보입니다. 2011년에 마르코스 연구진과 마크스 연구진은 이 간접 신호를 걷어 내는 직접 결합 분석(DCA, Direct Coupling Analysis)을 제안했습니다 [3][4]. 통계물리학의 최대 엔트로피 모형을 빌려 와서, 두 자리 사이의 직접 결합만 남기고 나머지를 정리하는 방법입니다. 마크스 연구진은 실험 구조를 전혀 보지 않고 서열만으로 15종 단백질의 접힘을 맞혔습니다. 오차는 대체로 3에서 5 앙스트롬 수준이었습니다. 원자 지름이 대략 1에서 2 앙스트롬이니, 형태의 큰 줄기를 잡아낸 셈입니다.

신경망은 이 결과를 확률 지도로 바꿉니다. 두 잔기 사이의 거리를 하나의 숫자로 찍지 않고, 2 앙스트롬에서 22 앙스트롬까지의 구간을 64칸으로 쪼갠 뒤 각 칸에 확률을 매깁니다. 이것이 거리 분포도(Distogram)입니다. 모델은 "8 앙스트롬"이라고 단정하는 대신 "7에서 9 앙스트롬일 확률이 높다"고 말합니다. 딥마인드가 2020년에 발표한 1세대

알파폴드는 이 확률 지도를 물리 에너지 함수에 넣고 경사하강법으로 좌표를 복원했습니다 [5].

이 방식에는 대가가 따릅니다. 공진화 신호를 읽으려면 친척 서열이 많아야 합니다. 열람실 책상에 판본이 세 장뿐이면 어느 칸이 짝을 이루는지 알아낼 방법이 없습니다. 실제로 정렬에 모인 서열이 수백 개 아래로 떨어지면 예측 정확도가 눈에 띄게 떨어집니다. 바이러스의 신생 유전자나 항체의 초가변 영역처럼 진화의 기록이 짧은 서열이 여기 해당합니다. 사람이 설계한 인공 단백질도 마찬가지입니다. 세상에 친척이 없기 때문입니다.

그래서 다른 길도 열렸습니다. 단백질 언어 모델은 정렬 표를 만들지 않습니다. 수억 개 서열을 문장처럼 읽어 아미노산 배열의 규칙을 통째로 외웁니다. 사람이 문법책 없이 책을 많이 읽어 문장 감각을 얻는 것과 비슷합니다. 정렬을 찾는 검색 단계가 사라지니 속도가 몇십 배 빨라집니다. 정확도는 정렬을 쓰는 모델보다 조금 낮지만, 친척이 없는 서열에서는 오히려 유리합니다 [8].

예측이 틀릴 수 있으니 모델은 자기 확신도 함께 내놓습니다. pLDDT는 잔기마다 0에서 100까지 매겨지는 점수입니다. 90을 넘으면 곁사슬 방향까지 믿어도 좋고, 70에서 90 사이면 골격은 맞다고 봅니다. 50 아래는 형태가 정해지지 않은 구역일 가능성이 큽니다. 이 점수가 붙어 있어서 연구자는 구조 그림을 색으로 읽습니다. 파란 부분은 믿고, 주황 부분은 실험으로 확인합니다.

여기서 수학이 등장합니다. 단백질을 통째로 90도 돌려도 그것은 같은 단백질입니다. 책상 위 머그컵을 돌려놓아도 머그컵인 것과 같습니다. 3차원 회전과 평행 이동을 모아 놓은 집합을 특수 유클리드 군(Special Euclidean Group, SE(3))이라고 부릅니다. 좋은 모델이라면 이 변환에 흔들리지 않아야 합니다. 원자 사이의 거리나 결합 에너지 같은 값은 아무리 돌려도 그대로여야 합니다. 이것을 불변성(Invariance)이라고 합니다. 반대로 좌표를 출력할 때는 입력을 돌린 만큼 출력도 함께 돌아가야 합니다. 이것을 등변성(Equivariance)이라고 합니다.

이 규칙을 데이터로 가르칠 수도 있습니다. 같은 단백질을 수천 번 돌려 가며 학습시키면 됩니다. 그러나 그것은 낭비입니다. 대신 규칙을 신경망 설계 안에 못으로 박아 넣는 길이 있습니다. 불변 포인트 어텐션(IPA, Invariant Point Attention)은 각 잔기에 좌표계를

하나씩 붙이고, 그 좌표계 안에서만 거리를 재도록 만듭니다. 등변 그래프 신경망(EGNN)은 좌표 갱신을 두 원자를 잇는 방향 벡터로만 표현합니다 [6]. 회전을 학습으로 배우지 않으니 학습 데이터도 덜 듭니다. 예측한 구조가 물리적으로 뒤틀리는 일도 줄어듭니다.

여기까지 읽으면 구조 표현이 언제나 서열 표현을 이길 것 같습니다. 2026년 6월에 나온 한 분석은 그렇지 않다고 말합니다 [7]. 이 연구는 알파폴드2의 내부 표현과 서열만 보는 단백질 언어 모델 ESM-2의 표현을 세 가지 과제에서 맞붙였습니다. 결합 친화도 예측에서는 ESM-2가 이겼습니다. 상관계수가 0.449 대 0.307이었습니다. 형태 유연성 예측에서도 ESM-2가 앞섰습니다. AUROC가 0.824 대 0.764였습니다. 알파폴드2가 이긴 과제는 알로스테릭 부위 찾기 하나였습니다. AUROC 0.548 대 0.485로 근소한 차이였습니다. 이 연구는 데이터를 잔기 단위로 무심하게 나누면 성능이 27에서 39퍼센트까지 부풀려진다는 사실도 함께 지적했습니다.

안핀센이 옳았습니다. 서열은 구조를 결정합니다. 그러나 구조를 계산해 냈다고 해서 그 구조가 모든 질문에 답해 주지는 않습니다. 단백질이 얼마나 세게 붙는지, 얼마나 잘 흔들리는지는 서열 자체에 더 진하게 남아 있기도 합니다. 구조 예측은 목적지가 아니라 갈림길입니다.

**참고자료**


[1] Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. Science, 181(4096), 223-230. https://doi.org/10.1126/science.181.4096.223

[2] Levinthal, C. (1969). How to fold graciously. Mossbauer Spectroscopy in Biological Systems: Proceedings of a Meeting held at Allerton House, 22-24.

[3] Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., & Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS ONE, 6(12), e28766. https://doi.org/10.1371/journal.pone.0028766

[4] Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., & Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. PNAS, 108(49), E1293-E1301. https://doi.org/10.1073/pnas.1111471108

[5] Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. Nature, 577(7792), 706-710. https://doi.org/10.1038/s41586-019-1923-7

[6] Satorras, V. G., Hoogeboom, E., & Welling, M. (2021). E(n) Equivariant Graph Neural Networks. arXiv preprint arXiv:2102.09844. https://arxiv.org/abs/2102.09844

[7] Chauhan, K. (2026). When Does Structure Help? The Information Bonus of AlphaFold2 Representations over Protein Language Models. arXiv preprint arXiv:2606.04228. https://arxiv.org/abs/2606.04228

[8] Zhang, Y., Deng, N., Song, X., Bi, Z., Wang, T., Yao, Z., et al. (2025). Advanced Deep Learning Methods for Protein Structure Prediction and Design. arXiv preprint arXiv:2503.13522. https://arxiv.org/abs/2503.13522

## 3.2 AlphaFold, AlphaFold2, AlphaFold3 아키텍처 및 거대 단백질 복합체 상호작용 예측의 진화

2020년 11월 30일, 화상 회의 화면 앞에 구조생물학자 수백 명이 앉아 있었습니다. 격년으로 열리는 단백질 구조 예측 대회 CASP14의 결과 발표 자리였습니다. 이 대회는 아직 공개되지 않은 실험 구조를 문제로 내고, 참가 팀이 서열만 보고 답을 제출하게 합니다. 채점표에 알파폴드2의 점수가 떴을 때 회의실은 조용해졌습니다. 대회를 30년 가까이 이끌어 온 존 몰트(John Moult)는 이 문제가 어떤 의미에서는 풀렸다고 말했습니다.

숫자가 그 말을 뒷받침했습니다. 알파폴드2의 전체 표적 중앙값 GDT_TS 점수는 92.4점이었습니다. 100점이 만점이고, 90점을 넘으면 실험으로 푼 구조와 구별하기 어렵다고 봅니다. 백본 원자의 평균 오차는 0.96 앙스트롬이었습니다 [1]. 탄소 원자 하나의 지름보다 작은 오차입니다. 2위 팀의 중앙값은 70점대 초반이었습니다.

2년 전에는 사정이 달랐습니다. 2018년 CASP13에 처음 나온 1세대 알파폴드는 두 단계로 나뉜 기계였습니다. 앞단에서 합성곱 신경망이 잔기 사이의 거리 분포도를 예측하고, 뒷단에서 로제타 계열의 물리 에너지 최소화가 그 지도를 보며 좌표를 빚었습니다 [2]. 성적은 좋았지만 두 단계 사이에 벽이 있었습니다. 뒷단의 실수가 앞단으로 되돌아가지 못했습니다.

알파폴드2는 그 벽을 허물었습니다. 서열이 들어가면 좌표가 나오는 하나의 신경망으로 다시 지었습니다. 심장은 이보포머(Evoformer)라는 블록입니다. 이보포머는 장부 두 권을 동시에 씁니다. 한 권은 다중 서열 정렬 표입니다. 진화가 남긴 판본들의 기록입니다. 다른 한 권은 잔기 쌍 표입니다. i번과 j번 잔기의 관계를 담은 격자입니다. 48개 블록을 지나는 동안 두 장부는 서로를 계속 고쳐 씁니다. 서열 쪽에서 새로 읽은 신호가 쌍 장부로 넘어가고, 쌍 장부에서 정리된 기하 정보가 다시 서열 쪽 해석을 바로잡습니다.

쌍 장부를 다듬는 규칙 중에 삼각형 부등식이 있습니다. 지도 위 세 도시를 생각하면 됩니다. 서울에서 대전이 140킬로미터이고 대전에서 대구가 150킬로미터라면, 서울에서 대구가 500킬로미터일 수는 없습니다. 잔기 A와 B가 가깝고 B와 C가 가깝다면, A와 C도 일정 범위 안에 들어와야 합니다. 이보포머는 모든 잔기 삼각형에 이 제약을 걸어 모순된 거리 조합을 걸러 냅니다. 물리 법칙을 계산하지 않고도 물리적으로 말이 되는 지도를 얻는 방법입니다.

정리된 지도는 구조 모듈로 넘어갑니다. 구조 모듈은 각 잔기를 3차원 공간에 떠 있는 작은 삼각판으로 봅니다. 판마다 위치와 방향이 있습니다. 불변 포인트 어텐션이 판들을 서로 당기고 밀어 배치를 잡아 갑니다. 곁사슬의 회전각까지 한 번에 나옵니다. 그리고 완성된 구조를 다시 입력으로 되먹여 세 번 더 돌립니다. 초고를 쓰고, 읽고, 고쳐 쓰는 과정과 닮았습니다.

이 모델이 세상을 바꾼 방식은 논문보다 데이터베이스 쪽이 컸습니다. 딥마인드와 유럽생물정보학연구소는 알파폴드 단백질 구조 데이터베이스를 열고 예측 구조를 무료로 풀었습니다. 2024년 기준 등록된 구조는 2억 1400만 개가 넘습니다. 2021년 첫 공개 때보다 500배 늘어난 규모입니다. 이용자는 유엔 회원국 전역에서 240만 명을 넘었습니다[3]. 실험실 한 곳이 구조 하나를 푸는 데 몇 년이 걸리던 시절이 있었습니다. 이제 대학원생이 점심시간에 검색창에 이름을 넣고 구조를 받습니다.

그사이에 중간 단계가 하나 있었습니다. 알파폴드2는 사슬 하나를 잘 접었지만 사슬 둘이 서로 붙는 문제에는 약했습니다. 연구자들은 두 사슬 사이에 가짜 링커를 끼워 하나처럼 속이는 편법을 썼습니다. 2022년에 나온 알파폴드 멀티머는 이 문제를 정면으로 다시 학습했습니다[12]. 사슬 경계를 모델에 알려 주고, 여러 종의 서열 정렬을 종별로 짝지어 넣는 방식입니다. 두 단백질이 같은 생물에서 함께 진화했다면 결합면의 잔기들도 함께 변했을 것이라는 생각입니다. 이때부터 예측 대상이 분자 하나에서 분자들의 관계로 옮겨 갔습니다.

2024년의 알파폴드3는 대상을 통째로 넓혔습니다. 단백질 하나가 아니라 생체 분자 복합체 전체를 원자 수준에서 다룹니다[4]. 구조도 바뀌었습니다. 무거운 이보포머는 페어포머(Pairformer)로 가벼워졌습니다. 다중 서열 정렬을 다루는 블록은 4개로 줄었습니다. 대신 뒷단에 확산 모듈(Diffusion Module)이 들어왔습니다. 확산 모델은 잡음으로 뒤덮인 점구름에서 잡음을 조금씩 걷어 내며 형체를 드러내는 방식입니다. 흐린 새벽 안개가 걷히면서 건물 윤곽이 나타나는 장면과 같습니다.

덕분에 알파폴드3는 단백질끼리의 결합뿐 아니라 DNA 이중나선, RNA 가닥, 인산화 같은 번역 후 변형, 신약 후보가 되는 저분자 리간드, 아연과 철 같은 금속 이온까지 한 장면 안에 놓고 계산합니다. 단백질과 리간드가 어떻게 맞물리는지를 겨루는 포즈버스터스(PoseBusters) 시험에서 알파폴드3는 76.4퍼센트의 성공률을

기록했습니다. 전통적인 도킹 프로그램들의 성적은 50퍼센트대였습니다 [4].

그렇다고 모든 항목이 좋아지지는 않았습니다. 2025년에 나온 독립 벤치마크는 냉정한 그림을 보여 줍니다 [5]. 단량체 단백질에서 알파폴드3의 TM-score는 0.817, 알파폴드2는 0.812였습니다. 차이가 통계적으로 뚜렷하지 않았습니다. LDDT는 0.792 대 0.787로 알파폴드3가 조금 앞섰습니다. 복합체에서는 DockQ가 0.589 대 0.568이었고, 접촉면을 제대로 맞힌 비율은 78.2퍼센트였습니다. 핵산 복합체와 항원 항체 복합체에서는 이전 모델보다 뚜렷하게 나았습니다. 확장이 곧 정확도는 아니었지만, 다룰 수 있는 분자의 종류가 늘어난 것 자체가 성과였습니다.

대회 성적표도 같은 이야기를 합니다. CASP16에서는 세 단계에 걸쳐 85개 표적이 나왔고, 사람 전문가 팀 125개와 서버 36개가 답을 냈습니다 [6]. 평가진의 결론은 명확했습니다. 단일 도메인 단백질의 접힘을 틀린 경우는 하나도 없었습니다. 상위권은 대부분 알파폴드2와 알파폴드3를 쓰되 다중 서열 정렬 입력을 손질한 팀들이었습니다. 반면 복합체는 표적의 30퍼센트 넘게 어려운 문제로 남았습니다. 항체와 항원이 붙는 자리를 맞히는 일이 제일 힘들었습니다.

2026년의 흐름은 개방과 속도입니다. 알파폴드3는 상업적 이용에 제약이 있습니다. 그 빈자리를 오픈소스 모델들이 메웠습니다. MIT 출신 연구진이 만든 볼츠(Boltz) 계열은 MIT 라이선스로 모델과 가중치, 학습 코드를 모두 공개했습니다. 볼츠2는 구조와 함께 결합 친화도를 예측하는데, 자유 에너지 섭동 계산에 가까운 정확도를 1000배 넘게 빠른 속도로 냅니다. 오픈폴드 컨소시엄은 2026년 3월 13일에 오픈폴드3의 대규모 갱신과 함께 학습 데이터 전체를 공개했습니다. 대부분의 항목에서 알파폴드3와 겨룰 만한 성적을 냈고, 2026년의 우선 과제로 항체와 항원 예측 개선을 내걸었습니다 [7].

속도와 규모를 다루는 연구도 쏟아졌습니다. 2026년 5월에 나온 DCFold는 확산 과정을 반복하지 않고 한 번의 순방향 계산으로 구조를 만들어 추론 속도를 15배 끌어올렸습니다 [8]. 정확도는 알파폴드3와 견줄 만한 수준을 지켰습니다. 같은 해 3월의 Fold-CP는 문맥 병렬화로 메모리 벽을 넘었습니다. NVIDIA B300 GPU 64장을 묶어 3만 잔기가 넘는 거대 조립체를 계산했고, 단백질 복합체 목록인 CORUM 데이터베이스의 90퍼센트 이상을 채점했습니다 [9]. 몇천 잔기에서 멈추던 천장이 사라진 것입니다.

설계 쪽 성과가 예측의 수준을 증언합니다. 차이 디스커버리(Chai Discovery)는 항체 결합 사례가 하나도 없는 표적 52개를 골라, 표적마다 후보를 20개 이하만 만들어 실험했습니다. 16퍼센트가 실제로 붙었습니다. 미니 단백질에서는 표적 5개에서 68퍼센트의 성공률이 나왔습니다 [10]. 수만 개를 걸러 하나를 찾던 방식과 비교하면 접시 한 판으로 끝나는 규모입니다.

다음 시험은 이미 예고되어 있습니다. CASP17은 2026년 대회를 준비하며 9월 1일까지 실험 좌표를 받고 있습니다. 새로 힘을 주는 분야는 형태 앙상블입니다. 하나의 정답 구조가 아니라, 한 단백질이 오갈 수 있는 여러 형태를 함께 내놓으라는 요구입니다 [11]. 정지 사진의 시대가 끝나가고 있습니다.

**참고자료**


[1] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589. https://doi.org/10.1038/s41586-021-03819-2

[2] Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. Nature, 577(7792), 706-710. https://doi.org/10.1038/s41586-019-1923-7

[3] Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., et al. (2024). AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Research, 52(D1), D368-D375. https://doi.org/10.1093/nar/gkad1011

[4] Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 630(8016), 493-500. https://doi.org/10.1038/s41586-024-07487-w

[5] Peng, C., Ni, W., Liu, Q., Hu, G., & Zheng, W. (2025). A comprehensive benchmarking of AlphaFold3 for predicting biomacromolecules and their interactions. Briefings in Bioinformatics, 26(6), bbaf616. https://doi.org/10.1093/bib/bbaf616

[6] Yuan, R., Zhang, J., Kryshtafovych, A., Schaeffer, R. D., Zhou, J., Cong, Q., & Grishin, N. V. (2025). CASP16 protein monomer structure prediction assessment. bioRxiv. https://doi.org/10.1101/2025.05.29.656942

[7] OpenFold Consortium. (2026년 3월 13일). OpenFold Consortium Announces Major OpenFold3 Update and Public Release of Training Data for Reproducible Biomolecular AI. https://www.businesswire.com/news/home/20260313170622/en/

[8] Zhang, Z., Feng, Y., Song, Y., Qiu, K., Zhou, H., & Ma, W.-Y. (2026). DCFold: Efficient Protein Structure Generation with Single Forward Pass. arXiv preprint arXiv:2605.17899. https://arxiv.org/abs/2605.17899

[9] Lin, D., Chu, S., Iyer, V., Lee, Y., St John, J., Boyd, K., et al. (2026). Fold-CP: A Context Parallelism Framework for Biomolecular Modeling. arXiv preprint arXiv:2603.14806. https://arxiv.org/abs/2603.14806

[10] Chai Discovery Team. (2025). Zero-shot antibody design in a 24-well plate. bioRxiv. https://doi.org/10.1101/2025.07.05.663018

[11] Protein Structure Prediction Center. (2026). Call for Targets: CASP17. https://predictioncenter.org/casp17/callForTargets.cgi

[12] Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., et al. (2022). Protein complex prediction with AlphaFold-Multimer. bioRxiv. https://doi.org/10.1101/2021.10.04.463034

## 3.3 실험적 프로테오믹스(Structural Proteomics) 데이터와 기계학습 모델의 구조적 상호 유도

단백질 데이터 은행(PDB)에 등록된 구조 파일을 열면 원자 좌표가 줄줄이 나옵니다. 모든 좌표가 한 순간에 멈춰 있습니다. 결정 안에서 얼려 붙잡은 자세이거나, 전자현미경 아래에서 급속 냉동된 자세입니다. 앨범에 붙은 증명사진과 같습니다. 그런데 세포 안의 단백질은 증명사진처럼 살지 않습니다. 헥소키네이스는 포도당이 다가오면 두 개의 엽이 조개껍데기처럼 닫힙니다. 수송 단백질은 막 안쪽과 바깥쪽을 향해 번갈아 문을 엽니다. 신호 전달 단백질은 상대가 붙는 순간 전체 골격을 비틉니다.

인공지능 모델은 이 앨범을 보고 배웠습니다. 그래서 앨범이 가진 편향을 그대로 물려받습니다. 알파세뉴클레인, 타우, p53의 전사활성 영역 같은 단백질은 세포 안에서 형태가 계속 변합니다. 알파폴드에 넣으면 조각난 저신뢰도 구조가 나옵니다. 모델이 틀린 것이 아닙니다. 정답이 하나가 아닌 문제에 하나의 답을 요구한 것이 문제였습니다.

2026년 네이처 메소드에 실린 한 논평은 이 지점을 앞으로의 중심 과제로 짚었습니다 [1]. 저자들은 구조 하나를 잘 맞히는 시대에서, 형태들이 각각 얼마의 비율로 존재하는지까지 맞히는 시대로 넘어가야 한다고 주장합니다. 그러려면 정답 데이터가 필요합니다. 앙상블을 채점할 기준과 도구가 아직 없다는 것이 이들의 지적입니다.

모델 안에서 답을 찾으려는 시도가 먼저 있었습니다. 알파폴드2에 넣는 서열 정렬을 통째로 쓰지 않고 일부만 무작위로 뽑아 넣으면, 같은 단백질에서 서로 다른 형태가 나옵니다. 정렬을 진화 계통별로 묶어 넣으면 그 묶음이 지지하는 형태가 따로 나옵니다. 2024년 네이처에 실린 연구는 이 방법으로 카이네이스와 수송 단백질의 열린 형태와 닫힌 형태를 모두 뽑아냈습니다 [11]. 사진 한 장을 뽑는 대신 앨범을 만든 것입니다. 그러나 어느 형태가 세포 안에서 몇 퍼센트를 차지하는지는 이 방법으로 알 수 없습니다. 비율을 정하려면 실험이 필요합니다.

실험이 그 빈자리를 메웁니다. 질량 분석 기반 화학적 교차 결합(XL-MS, Cross-linking Mass Spectrometry)이 대표적입니다. 단백질 용액에 짧은 화학 팔을 가진 시약을 넣습니다. 자주 쓰는 DSS 시약의 팔 길이는 11.4 앙스트롬입니다. 이 팔은 두 라이신(Lysine) 잔기의 곁사슬 끝을 붙잡아 묶습니다. 라이신 곁사슬 자체가 길기 때문에,

묶인 두 잔기의 중심 탄소 사이 거리는 대략 30 앙스트롬 이내입니다. 그다음 단백질을 잘게 잘라 질량 분석기에 넣습니다. 붙어 나온 두 조각의 무게를 읽으면 어느 자리와 어느 자리가 가까웠는지 알 수 있습니다 [2][3]. 세포를 통째로 실로 묶어 놓고 매듭만 찾아 거리를 재는 방식입니다.

수소 중수소 교환 질량 분석(HDX-MS)은 다른 정보를 줍니다. 단백질 골격의 아마이드 수소는 물에 노출되면 중수소로 바뀝니다. 표면에서 헐렁하게 흔들리는 부위는 몇 초 만에 바뀝니다. 안쪽에 단단히 묻힌 부위는 몇 시간이 지나도 그대로입니다. 무거워진 정도를 시간별로 재면 어느 구역이 열려 있고 어느 구역이 닫혀 있는지가 드러납니다 [4]. 좌표는 알려 주지 않지만 움직임은 알려 줍니다.

두 실험은 각자 불완전합니다. XL-MS는 거리 몇십 개를 줄 뿐 나머지 원자는 비워 둡니다. HDX-MS는 구역 단위로만 말합니다. 딥러닝 모델은 반대로 모든 원자의 좌표를 채우지만 어느 형태가 맞는지 고르지 못합니다. 그래서 둘을 붙이는 연구가 나왔습니다.

2026년 5월에 공개된 AIMS-Fold가 그 예입니다 [5]. 이 연구는 XL-MS의 거리 제약과 HDX-MS의 용매 노출 정보를 확산 기반 공동 접힘 모델 안으로 밀어 넣습니다. 실험값을 사후 검증에 쓰는 것이 아니라 생성 과정 자체를 그쪽으로 끕니다. 연구진은 두 실험 정보가 각각 정확도를 올리고, 함께 넣으면 상승 효과가 난다고 보고했습니다. 결과가 두드러진 곳은 유도 근접(induced proximity) 표적이었습니다. 프로탁(PROTAC)이나 분자 접착제처럼 원래 만나지 않던 두 단백질을 억지로 붙이는 약물이 여기 속합니다. 학습 데이터에 비슷한 사례가 거의 없는 문제입니다. AIMS-Fold는 이 표적들에서 실험 정보를 쓰지 않는 볼츠2보다 높은 정확도를 냈습니다.

저온전자현미경(cryo-EM)에서도 같은 일이 벌어집니다. 유럽 생물정보학연구소가 운영하는 전자현미경 데이터 은행에는 2026년 8월 12일 기준 60,895건의 밀도 지도가 쌓여 있습니다 [6]. 밀도 지도는 안개 낀 사진 같습니다. 분자가 어디에 있는지는 보이지만 원자가 어디에 있는지는 흐릿합니다. 연구자는 그 안개 속에 모델을 밀어 넣고 맞춰 갑니다. 이 정련 작업은 오래 걸립니다. 피닉스나 로제타 같은 전통 도구는 계산이 무겁고 사람 손을 많이 탑니다.

2026년 2월에 나온 CryoNet.Refine는 이 과정을 한 걸음 확산 모델로 바꿨습니다 [7]. 밀도 지도를 손실 함수에 넣고 화학 결합의 기하 제약을 함께 걸어, 모델과 지도의 상관도와

기하 품질을 동시에 올렸습니다. 이 연구는 ICLR 2026에 채택되었고 웹 서버로도 열려 있습니다. 같은 연구실에서 나온 CryoLVM은 밀도 지도 자체를 자기지도 학습으로 미리 배웁니다 [8]. 지도를 선명하게 만들고, 해상도를 올리고, 촬영 한계로 빠진 부분을 메웁니다. 형태 이질성을 정면으로 다루는 CryoDyna 같은 연구도 나왔습니다 [9]. 하나의 실험 데이터에서 여러 형태를 갈라내는 시도입니다.

현미경은 이제 시험관 밖으로 나갔습니다. 저온전자단층촬영(cryo-ET)은 세포를 정제하지 않고 얼린 뒤, 얇게 깎아 여러 각도에서 찍습니다. 그렇게 얻은 3차원 영상 안에는 리보솜과 단백질 분해 복합체가 실제로 놓여 있던 자리 그대로 들어 있습니다. 문제는 영상이 아주 어둡고 지저분하다는 것입니다. 잡음 속에서 분자 하나를 골라내는 입자 선별 단계에 딥러닝이 들어갑니다. 예측 구조를 본보기 삼아 영상 안을 훑고, 닮은 밀도를 찾아 표시합니다. 세포 사진 한 장 안에서 단백질의 주소록을 뽑아내는 셈입니다.

화살은 반대 방향으로도 날아갑니다. X선 결정학에는 위상 문제라는 오래된 난관이 있습니다. 회절 무늬의 세기는 측정되지만 위상은 측정되지 않습니다. 위상을 얻으려면 비슷한 구조를 미리 하나 넣어 맞춰 보는 분자 치환이 필요했고, 그 비슷한 구조가 없으면 실험이 몇 년씩 멈췄습니다. 지금은 예측 모델을 넣습니다. 오차 1 앙스트롬 안쪽의 초안이 있으면 위상이 풀립니다. 저온전자현미경 쪽도 같습니다. 흐릿한 밀도 지도에 예측 모델을 얹으면 사람이 손으로 사슬을 따라가던 며칠이 몇 시간으로 줄어듭니다 [10].

여러 실험을 한 판에 올리는 통합 모델링도 자리를 잡았습니다. 소각 X선 산란은 분자 전체의 부피와 모양을 알려 줍니다. 핵자기공명은 원자 수준의 거리와 각도를 알려 주되 큰 분자에는 잘 듣지 않습니다. 교차 결합은 접촉점을, 저온전자현미경은 전체 윤곽을 알려 줍니다. 어느 하나도 혼자서는 구조를 완성하지 못합니다. 신경망은 이 서로 다른 측정값들을 같은 좌표계 위에 놓고 저울질하는 자리에서 힘을 냅니다. 각각의 실험이 남긴 제약을 모두 만족시키는 형태들의 집합을 찾는 문제로 바꾸는 것입니다.

이 관계를 한쪽이 다른 쪽을 대체하는 구도로 읽으면 곤란합니다. 실험은 모델에게 닻을 내려 줍니다. 모델은 실험이 비워 둔 좌표를 채웁니다. 사람은 그 사이에서 판단합니다. 지금 나온 결과들이 보여 주는 그림은 이렇습니다. 실험 정보를 넣은 모델이 넣지 않은 모델을 이깁니다. 그리고 모델이 만든 초안 위에서 실험 정련이 훨씬 빨라집니다.

한 가지는 분명히 해 두려 합니다. 예측 구조는 아직 실험 구조를 대신하지 못합니다. 모델은 자기가 틀린 지점을 스스로 알지 못합니다. 신뢰도 점수는 높은데 실제로는 어긋난 사례가 계속 보고됩니다. 유도 근접 표적처럼 학습 데이터가 얇은 영역에서 이 위험은 커집니다. 교차 결합 실험에서 나온 거리 제약 서른 줄이 그 위험을 잘라 냅니다. 모델이 그려 놓은 그럴듯한 배치가 실제 측정값과 어긋나는 순간, 연구자는 다시 계산할 이유를 손에 쥡니다. 시험관과 신경망은 경쟁자가 아니라 같은 문제를 앞뒤에서 미는 두 사람입니다. 앞선 사람은 어디로 갈지 알고, 뒤선 사람은 얼마나 밀렸는지 압니다.

**참고자료**


[1] Wankowicz, S. A., & Bonomi, M. (2026). From possibility to precision in macromolecular ensemble prediction. Nature Methods, 23, 1100-1108. https://doi.org/10.1038/s41592-026-03084-z

[2] Leitner, A., Faini, M., Stengel, F., & Aebersold, R. (2016). Crosslinking and Mass Spectrometry: An Integrated Approach to Understand the Structure and Function of Molecular Machines. Trends in Biochemical Sciences, 41(1), 20-32. https://doi.org/10.1016/j.tibs.2015.10.008

[3] Rappsilber, J. (2011). The beginning of a beautiful friendship: Cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. Journal of Structural Biology, 173(3), 530-540. https://doi.org/10.1016/j.jsb.2010.10.014

[4] Engen, J. R. (2009). Analysis of protein conformation and dynamics by hydrogen/deuterium exchange MS. Analytical Chemistry, 81(19), 7870-7875. https://doi.org/10.1021/ac901154s

[5] Shtrikman, A., Simchi, N., Ran Shchory, M., Brodsky, S., Seger, E., & Pevzner, K. (2026). Co-folding model guided by structural proteomics. arXiv preprint arXiv:2605.26192. https://arxiv.org/abs/2605.26192

[6] EMBL-EBI. (2026년 8월 12일). Electron Microscopy Data Bank (EMDB) 통계. https://www.ebi.ac.uk/emdb/

[7] Huang, F., Yu, X., Xu, K., & Zhang, Q. C. (2026). CryoNet.Refine: A One-step Diffusion Model for Rapid Refinement of Structural Models with Cryo-EM Density Map Restraints. arXiv preprint arXiv:2602.22263. https://arxiv.org/abs/2602.22263

[8] Fu, W., Shu, K., Xu, K., & Zhang, Q. C. (2026). CryoLVM: Self-supervised Learning from Cryo-EM Density Maps with Large Vision Models. arXiv preprint arXiv:2602.02620. https://arxiv.org/abs/2602.02620

[9] Zhang, C., et al. (2025). CryoDyna: Multiscale end-to-end modeling of cryo-EM macromolecule dynamics with physics-aware neural network. arXiv preprint arXiv:2510.16510. https://arxiv.org/abs/2510.16510

[10] Terashi, G., Wang, X., Prasad, P. R., & Kihara, D. (2023). Deep learning methods for protein structure modeling in cryo-EM density maps. Current Opinion in Structural Biology, 79, 102534. https://doi.org/10.1016/j.sbi.2023.102534

[11] Wayment-Steele, H. K., Ojoawo, A., Otten, R., Apitz, J. M., Pitsawong, W., Hömberger, M., Ovchinnikov, S., Colwell, L., & Kern, D. (2024). Predicting multiple conformations via sequence clustering and AlphaFold2. Nature, 625(7996), 832-839. https://doi.org/10.1038/s41586-023-06832-9

# 제4장: 생성형 바이오 모델을 활용한 단백질 및 항체 신규 설계 (De Novo Design)

## 4.1 디퓨전(Diffusion) 모델 및 플로우 매칭(Flow Matching) 기반의 물리적 단백질 구조 생성

화면에는 점들이 흩어져 있습니다. 좌표계 안에 아무 규칙 없이 뿌려진 수백 개의 점입니다. 연구자가 실행 버튼을 누릅니다. 점들이 조금씩 자리를 옮깁니다. 계산이 쉰 번쯤 돌아가면 점들은 나선을 그리고, 그 옆에 병풍 같은 판이 붙고, 마지막에는 집게처럼 벌어진 홈 하나가 남습니다. 지구에 한 번도 존재한 적 없던 단백질이 그렇게 만들어집니다. 걸린 시간은 몇 분입니다.

자연에 있는 단백질은 수십억 년짜리 시행착오의 결과물입니다. 진화는 생존과 번식이라는 두 가지 시험만 통과하면 되었습니다. 암세포만 골라 붙잡는 분자를 만들라거나, 페트병을 분해하는 효소를 만들라는 주문은 진화의 시험 범위 밖입니다. 그래서 20세기의 단백질 공학은 자연산 단백질을 조금씩 고쳐 쓰는 데 머물렀습니다. 지향적 진화(Directed Evolution)라고 부릅니다. 아미노산을 무작위로 바꾼 변이체를 수십만 개 만들어 놓고 그중 쓸 만한 것을 골라내는 방식입니다. 도서관 한 층 전체를 훑어야 하는데 첫 번째 서가만 더듬은 셈입니다. 아미노산 100개짜리 단백질 하나가 가질 수 있는 서열의 가짓수는 20의 100제곱입니다. 우주의 원자 수보다 많습니다.

생성형 인공지능은 이 서가 문제를 다르게 풀었습니다. 확산 모델(Diffusion Model)이 그 출발점입니다. 흐릿한 안개에서 형체를 조금씩 걷어내는 방식입니다. 학습할 때는 실제 단백질 구조에 잡음을 조금씩 뿌려 완전한 무질서로 만듭니다. 쓸 때는 그 과정을 거꾸로 돌립니다. 무질서한 원자 구름에서 시작해 한 단계씩 잡음을 덜어내면 구조가 드러납니다. 이때 표적 단백질의 원자 좌표를 조건(Conditioning)으로 넣어 줍니다. 조건이란 그림을 그리기 전에 미리 정해 놓은 밑그림입니다. 표적의 결합 홈 모양을 밑그림으로 주면, 모델은 그 홈에 맞춰 들어갈 결합체(Binder)의 백본(Backbone) 골격을 만들어 냅니다. 백본은 아미노산의 뼈대에 해당하는 N, Cα, C, O 네 원자의 사슬입니다.

이 방식이 세상에 알려진 것은 2023년입니다. 워싱턴대 데이비드 베이커 연구실이 RFdiffusion을 네이처에 발표했습니다. 구조 예측 모델인 RoseTTAFold의 속을 뜯어 확산 모델로 다시 학습시킨 것입니다. 이미 접힌 단백질을 알아보도록 배운 신경망은 접힘의 규칙을 몸에 익히고 있었습니다. 그 규칙을 거꾸로 돌려 구조를 만들게 했더니, 실험실에서 실제로 접히는 단백질이 나왔습니다[1]. 발표 이후 3년 사이에 파생 모델이

쏟아졌습니다.

이 계산에는 지켜야 할 규칙이 있습니다. 단백질을 통째로 돌리거나 옮겨도 그 단백질은 같은 단백질입니다. 모델이 이 사실을 모르면 같은 분자를 서로 다른 둘로 착각하고 헛수고를 합니다. 그래서 구조 생성 모델에는 등변성(Equivariance)이라는 성질을 넣습니다. 입력을 돌리면 출력도 같은 각도로 따라 도는 성질입니다. 지하철 노선도를 뒤집어 들어도 역과 역의 순서는 그대로인 것과 같습니다. 여기에 물리 규칙이 더 얹힙니다. 아미노산을 잇는 펩타이드 결합의 길이는 1.33옹스트롬 근처에서 거의 움직이지 않습니다. 결합각도 정해진 범위 밖으로 벗어나지 못합니다. 이 제약을 어긴 구조는 화면에서는 그럴듯해 보여도 시험관에서는 접히지 않습니다. 알파 나선과 베타 병풍이라는 두 가지 기본 모양이 저절로 나타나는 것도 이 규칙을 모델이 지키기 때문입니다.

설계 과제 가운데 모티프 스캐폴딩(Motif Scaffolding)이라는 것이 있습니다. 기능을 맡은 짧은 조각 하나를 붙박아 놓고, 그 조각이 정확한 모양을 유지하도록 나머지 몸통을 새로 지어 주는 일입니다. 바이러스 표면에서 항체가 달라붙는 몇 잔기만 남기고 나머지를 새로 지어 백신 항원을 만드는 작업이 여기에 해당합니다. 고정된 조각이 못이라면, 새로 만드는 몸통은 그 못을 제자리에 붙들어 두는 액자입니다.

플로우 매칭(Flow Matching)은 그다음 세대의 방법입니다. 확산 모델이 지그재그로 흔들리며 목적지를 찾아간다면, 플로우 매칭은 출발점에서 도착점까지 곧은 길을 미리 배웁니다. 확률 미분 방정식 대신 상미분 방정식 궤적을 학습합니다. 길이 곧으니 걸음 수가 줄고, 걸음 수가 줄면 계산 시간이 줄어듭니다. 2026년 3월에 나온 PI-Mamba는 이 길을 상태 공간 모델과 묶어 잔기 2,000개가 넘는 큰 단백질을 24기가바이트짜리 그래픽카드 한 장으로 생성했습니다. 국소 결합 기하 위반율은 0퍼센트, 설계 가능성 지표인 scTM 점수는 0.91이었습니다[3]. 같은 해 6월의 ProHiFlo는 굵은 골격을 먼저 잡고 원자를 나중에 채우는 계단식 방법으로 효소 활성 부위 뼈대 설계 성공률을 58.9퍼센트까지 올렸습니다. 같은 과제에서 RFdiffusion은 41.2퍼센트였습니다[4].

길이를 미리 정해 줘야 한다는 제약도 풀렸습니다. 2026년 7월의 GPFlow는 단백질 길이 자체를 생성 과정에서 결정하도록 만들었습니다. 모티프 스캐폴딩 과제 16개 가운데 10개에서 1위를 차지했습니다[5]. 설계자가 "이 정도 길이면 되겠지"라고 미리 찍어야 했던 부담이 사라진 것입니다.

3차원 골격이 나왔다고 끝은 아닙니다. 그 골격을 실제로 유지해 줄 아미노산 서열을 알아내야 합니다. 이 작업을 역접힘(Inverse Folding)이라고 합니다. 접힌 모양을 먼저 보고 원래 글자를 되짚는 일입니다. 이 자리를 오래 지킨 것은 2022년의 ProteinMPNN입니다. 주어진 골격에 맞는 서열을 예측했을 때 자연이 실제로 쓰던 아미노산과 52.4퍼센트가 일치했습니다. 물리 계산에 기대던 로제타는 같은 시험에서 32.9퍼센트였습니다[2]. 계산 시간도 초 단위로 줄었습니다. 2026년 3월의 Refold는 여기서 한 걸음 더 갔습니다. 데이터베이스에서 비슷한 구조 조각을 찾아오는 방식과 신경망이 처음부터 예측하는 방식을 붙여, 국소 정확도와 전체 일관성을 함께 잡았습니다[6].

현장의 판도를 바꾼 사건은 2025년 12월 3일에 일어났습니다. 데이비드 베이커 연구실이 RFdiffusion3를 공개했습니다. 이전 판인 RFdiffusion2가 잔기와 원자를 섞어 다루었다면, 3판은 원자 하나를 설계의 기본 단위로 삼습니다. 어느 원자와 어느 원자 사이에 수소 결합을 놓을지까지 지정할 수 있습니다. 전원자 계산이라 무거워질 법한데도 추론 속도는 2판보다 열 배 빨랐습니다. 금속가수분해효소를 설계해 96개를 실제로 합성했더니 5개가 제 기능을 했고, 그중 ZETA_1이라는 설계품은 기존 설계 효소보다 촉매 효율이 몇 자릿수 높았습니다. 시스테인 가수분해효소에서는 190개를 시험해 35개가 활성을 보였습니다[7].

여기서 숫자를 정직하게 읽어야 합니다. 96개 중 5개는 5퍼센트입니다. 190개 중 35개는 18퍼센트입니다. 2025년 8월에 나온 메타 분석은 표적 15종에 대해 실험으로 검증된 결합체 3,766개를 모아 성공률을 계산했습니다. 11.6퍼센트였습니다[8]. 컴퓨터 안에서는 그럴듯한 구조가 쏟아지지만, 시험관에서 실제로 붙는 것은 열에 하나입니다. 연구비를 신청하는 자리에서는 "성공률 80퍼센트"라는 말이 자주 들립니다. 그 숫자는 대개 컴퓨터가 컴퓨터를 채점한 결과입니다. 웨트랩의 채점표는 다릅니다.

그래서 최근 연구의 무게중심은 "더 잘 만들기"에서 "덜 버리기"로 옮겨 가고 있습니다. 2026년 3월의 Proteina-Complexa는 생성 모델과 구조 예측기를 이용한 서열 최적화를 하나로 합쳤습니다. 컴퓨터로 예측한 단일체 구조에서 도메인 사이 접촉면을 긁어모아 Teddymer라는 대규모 합성 학습 데이터를 만들고, 여기에 실험으로 확인된 복합체를 더해 기반 모델을 학습시킵니다. 그다음 생성 시점에 계산 자원을 더 부어 후보를 다듬습니다[9]. 같은 해 7월의 Chamaileon은 표적 하나, 상태 하나라는 오래된 전제를

깼습니다. 하나의 서열이 여러 표적과 여러 형태에 두루 붙도록 설계하는 문제로 다시 정의하고, CROSS라는 새 평가 기준을 함께 내놓았습니다[10]. 표적이 모양을 바꾸며 도망 다니는 바이러스 단백질을 상대하려면 이런 접근이 필요합니다.

평가 방식도 함께 바뀌고 있습니다. 예전에는 설계한 서열을 구조 예측기에 다시 넣어 원래 의도한 모양이 나오는지 보는 것으로 채점을 마쳤습니다. 문제는 설계 모델과 채점 모델이 같은 데이터로 배웠다는 데 있습니다. 같은 선생님에게 배운 학생이 같은 선생님에게 시험을 봅니다. 그래서 2026년의 연구들은 AlphaFold3나 Boltz-1처럼 계열이 다른 예측기를 여러 개 세워 놓고 교차 채점을 합니다. 3,766개 결합체를 다시 채점한 앞의 메타 분석도 이 방식을 썼고, 접촉면에 초점을 맞춘 지표가 기존의 전체 신뢰도 점수보다 실험 성공을 잘 맞혔다고 보고했습니다[8].

안개에서 형체를 꺼내는 기술은 이미 손에 들어왔습니다. 남은 일은 그 형체가 유리병 속에서도 살아남는지 확인하는 것입니다. 그리고 그 확인은 여전히 사람이 피펫을 들어야 끝납니다.

**참고자료**


[1] Watson, J. L., Juergens, D., Bennett, N. R., et al. (2023). De novo design of protein structure and function with RFdiffusion. Nature, 620(7976), 1089-1100. https://doi.org/10.1038/s41586-023-06415-8

[2] Dauparas, J., Anishchenko, I., Bennett, N., et al. (2022). Robust deep learning-based protein sequence design using ProteinMPNN. Science, 378(6615), 49-56. https://doi.org/10.1126/science.add2187

[3] Wu, T., & Zhu, L. (2026). PI-Mamba: Linear-Time Protein Backbone Generation via Spectrally Initialized Flow Matching. arXiv preprint arXiv:2603.26705. https://arxiv.org/abs/2603.26705

[4] Wang, C., Cleti, M., & Jano, P. (2026). ProHiFlo: Hierarchical Flow Matching with Functional Guidance for De Novo Protein Generation. arXiv preprint arXiv:2606.11243. https://arxiv.org/abs/2606.11243

[5] Cheng, C., Ni, Z., Qu, Y., et al. (2026). Variable-Length Generative Protein Design via Generalized Poisson Flow. arXiv preprint arXiv:2607.09039. https://arxiv.org/abs/2607.09039

[6] Zhu, Y., et al. (2026). Refold: Refining Protein Inverse Folding with Efficient Structural Matching and Fusion. arXiv preprint arXiv:2603.14350. https://arxiv.org/abs/2603.14350

[7] Genetic Engineering & Biotechnology News. (2025년 12월 3일). RFdiffusion3 Now Open Source, Designs DNA Binders and Advanced Enzymes. https://www.genengnews.com/topics/artificial-intelligence/rfdiffusion3-now-open-source-designs-dna-binders-and-advanced-enzymes/

[8] Overath, M. D., Rygaard, A., Jacobsen, C. P., et al. (2025). Predicting Experimental Success in De Novo Binder Design: A Meta-Analysis of 3,766 Experimentally Characterised Binders. bioRxiv. https://doi.org/10.1101/2025.08.14.670059

[9] Didi, K., Zhang, Z., Zhou, G., et al. (2026). Scaling Atomistic Protein Binder Design with Generative Pretraining and Test-Time Compute. arXiv preprint arXiv:2603.27950. https://arxiv.org/abs/2603.27950

[10] Cao, H., Cheng, S., Liu, M., et al. (2026). Chamaileon: Cross-Context Binder Design with Contextualized Modeling and Mixed Sampling. arXiv preprint arXiv:2607.23518. https://arxiv.org/abs/2607.23518

## 4.2 항체 CDR 설계와 생체 적합성 극대화를 위한 선호도 정렬(Preference Optimization) 기술

시험관 하나가 실험대 위에 서 있습니다. 어제까지 맑았던 액체가 오늘 아침에는 뿌옇습니다. 바닥에 흰 앙금이 가라앉아 있습니다. 항체가 서로 엉겨 붙은 것입니다. 결합력 측정에서 나노몰 단위를 기록해 연구실이 들썩였던 후보였습니다. 그 후보는 그날로 폐기됩니다.

항체(Antibody)는 Y자 모양의 단백질입니다. 두 팔 끝에 손가락처럼 튀어나온 고리가 여섯 개 있는데, 이것이 상보성 결정 부위(CDR, Complementarity-Determining Region)입니다. 항원의 표면을 더듬어 붙잡는 부분입니다. 여섯 고리 중 중쇄 세 번째 고리인 CDR-H3가 문제아입니다. 길이가 제각각이고 모양도 자유롭게 흔들립니다. 항원을 알아보는 힘의 상당 부분이 이 고리에서 나오는데, 바로 그 자유로움 때문에 예측이 어렵습니다.

숫자로 보면 이 고리의 무게가 드러납니다. 세계 의약품 매출 상위 열 개 품목 가운데 절반 이상이 항체 계열입니다. 하나의 항체 신약을 만드는 데 십 년 넘는 시간과 수천억 원이 들어갑니다. 그 비용의 상당 부분은 후보를 만드는 데가 아니라 만들어 놓은 후보를 버리는 데서 나옵니다. 실험실에서 수만 개를 걸러 열 개를 남기고, 그 열 개 중 아홉 개가 동물시험이나 초기 임상에서 탈락합니다. 설계 단계에서 탈락 사유를 미리 알아낼 수 있다면 아낄 수 있는 것은 돈만이 아닙니다.

여기에 인공지능을 붙이는 방법은 여러 갈래로 갈렸습니다. 2026년 2월의 AbFlow는 플로우 매칭을 항체 전체 원자 구조 설계에 가져왔습니다. 항원 표면의 정보를 읽는 등변 인코더를 속도장 신경망에 붙여 CDR-H3 부위를 다듬습니다[1]. 같은 해 7월의 ABOPD는 학습 방식 자체를 손봤습니다. 기존 확산 모델은 실제 구조에 잡음을 뿌려 만든 상태만 보고 배웁니다. 그런데 실제로 생성할 때 모델이 지나가는 중간 상태는 자기가 만든 상태입니다. 이 어긋남 때문에 고리 모양이 궤적을 따라 조금씩 밀립니다. ABOPD는 학습 중에 모델이 스스로 지나간 상태를 가져다가 실제 구조를 정답지로 놓고 지도합니다. 정책 증류(On-Policy Distillation)라고 부릅니다. 그 결과 RAbD 평가에서 CDR-H3 구조 오차가 2.37옹스트롬에서 1.95옹스트롬으로 0.42옹스트롬 줄었습니다[2]. 옹스트롬은 1억분의 1센티미터입니다. 원자 하나 크기의 절반쯤 되는 차이가 붙느냐 마느냐를

가릅니다.

그런데 강하게 붙는 항체를 만드는 일은 절반입니다. 나머지 절반은 약이 될 수 있느냐입니다. 이것을 약물성(Developability)이라고 부릅니다. 첫째로 엉기지 않아야 합니다. 둘째로 고농도에서 끈적이지 않아야 합니다. 피하주사 한 방에 150밀리그램을 넣으려면 액체가 물처럼 흘러야 합니다. 셋째로 CHO 세포 배양기에서 넉넉하게 만들어져야 합니다. 넷째로 사람 면역계가 그 항체를 침입자로 오해하지 않아야 합니다. 이 네 가지는 서로 밀고 당깁니다. 소수성 잔기를 늘리면 결합력은 올라가지만 응집 위험도 같이 올라갑니다. 도시락 반찬 칸을 늘리면 밥 칸이 줄어드는 것과 같습니다.

모델이 빠지는 함정도 드러났습니다. 2026년 5월의 EvoStruct는 등변 그래프 신경망 계열의 항체 설계 모델들이 어휘 붕괴를 겪는다고 지적했습니다. 서열 회수율 점수는 높은데, 실제로 뱉는 아미노산은 티로신과 글리신 몇 종류에 쏠려 있었습니다. 항체 고리에 이 둘이 흔하다 보니, 아무 데나 그 둘을 적어 넣는 것만으로도 점수가 올라간 것입니다. 시험 범위를 외운 학생이 답안지에 같은 답만 반복해 적는 것과 같습니다. EvoStruct는 진화 정보를 담은 단백질 언어 모델을 구조 정보와 함께 붙여 이 쏠림을 풀었습니다[9]. 항원 정보를 제대로 읽게 만드는 다중 모달 기반 모델도 같은 해 7월에 나왔습니다[10].

선호도 정렬(Preference Optimization)이 여기서 등장합니다. 원리는 사람에게 언어 모델을 맞출 때 쓰던 것과 같습니다. 정답을 하나 정해 주는 대신, 두 후보를 나란히 놓고 어느 쪽이 나은지만 알려 줍니다. 직접 선호도 최적화(DPO, Direct Preference Optimization)는 이 비교 쌍만으로 모델을 밀고 당깁니다. 좋은 쪽이 나올 확률은 올리고 나쁜 쪽이 나올 확률은 내립니다. 별도의 보상 모델을 따로 학습시키지 않아도 됩니다. 이 방식이 항체 설계와 잘 맞는 이유가 있습니다. 응집이나 점도 같은 성질은 절대적인 정답값을 매기기가 어렵습니다. 측정 장비와 완충액 조성에 따라 숫자가 흔들립니다. 그러나 같은 조건에서 두 후보를 나란히 놓고 어느 쪽이 덜 엉겼는지 고르는 일은 실험자가 어렵지 않게 합니다. 사람이 답하기 쉬운 질문으로 문제를 바꿔 놓은 것입니다.

항체 실험 데이터는 이 방식과 잘 맞습니다. 정확한 발현량 수치를 가진 서열은 늘 부족하지만, "이건 잘 나왔고 저건 안 나왔다"는 정도의 기록은 많기 때문입니다. 2026년 6월에 나온 연구는 이 점을 파고들었습니다. 정량 데이터가 붙은 서열은 1,254개뿐이었지만, 낙타류 면역 실험에서 나온 표지 없는 서열 400만 개를 약한 감독

신호로 끌어왔습니다. 길이가 제각각인 항체 서열을 다루려고 IMGT 번호 체계로 자리를 맞추고, 가려진 위치의 로그 우도를 합쳐 쓰는 근사법을 만들었습니다. 발현량 순위 예측에서 기존 방법을 대부분 항목에서 앞섰습니다[3].

부작용도 있습니다. 선호도 정렬을 세게 걸면 모델이 원래 잘하던 일을 잊습니다. 결합력만 쫓다가 접히지도 않는 서열을 뱉는 식입니다. 2026년 5월의 ProteinOPD는 여러 목표를 각각 맡은 교사 모델들의 지식을 학생 모델 하나에 토큰 단위로 옮겨 붙이는 방법으로 이 망각을 막았습니다. 보상 신호로 시행착오를 반복하는 기존 정렬 방식보다 학습 속도가 8배 빨랐습니다[4].

평가 기준을 세우려는 움직임도 있습니다. 항체 설계 논문은 지난 3년 사이 수십 편이 쏟아졌는데, 저마다 다른 시점의 SAbDab 데이터로 다른 지표를 재고 있었습니다. 2026년 3월의 CHIMERA-Bench는 에피토프별로 정리한 공통 시험지를 내놓았습니다[5]. 다목적 최적화 쪽에서는 BOAT가 여러 예측기를 베이지안 최적화로 묶었고[6], CrossAbSense는 성질마다 전용 예측기를 두어 값비싼 생물리 실험 횟수를 줄였습니다[7]. 그럼에도 약물성 예측은 결합력 예측만큼 자리를 잡지 못했습니다. 같은 항체를 두고 예측기마다 다른 답을 내는 일이 흔합니다[6]. 결합력을 잘 맞히니 약물성도 잘 맞히겠지 하고 넘어가면 시험관은 다시 뿌예집니다.

면역원성 문제는 성격이 다릅니다. 나머지 세 가지가 물리와 화학의 문제라면, 이것은 사람마다 다른 면역계의 문제입니다. 항체를 사람 몸에 넣었는데 몸이 그 항체를 침입자로 보면 항약물항체(ADA)가 만들어집니다. 그러면 약효가 사라지고, 심하면 알레르기 반응이 옵니다. 그래서 설계 단계에서 사람 항체가 흔히 쓰는 골격에 최대한 가깝게 맞춥니다. 이것을 인간화(Humanization)라고 합니다. 그런데 사람 골격에 가깝게 맞출수록 설계의 자유도가 줄고, 자유도가 줄면 결합력이 떨어집니다. 밀고 당기기가 여기서도 반복됩니다.

그래도 사람 몸에 도달한 사례가 나왔습니다. 앱사이(Absci)의 ABS-201은 프로락틴 수용체를 표적으로 하는 항체입니다. 회사의 생성형 인공지능 설계 결과물이 사람 임상에 들어간 첫 사례로, 2025년 12월에 첫 투여가 시작되었습니다. 2026년 6월 24일에 나온 1상 중간 결과는 건강한 성인 32명을 150, 450, 900, 1800밀리그램 네 개 용량 집단으로 나눠 정맥 투여한 것으로, 중대한 이상반응은 없었습니다. 약동학 자료는 6개월에 몇 차례 주사하는 간격도 가능하다고 시사했습니다[8]. 남성형 탈모와 자궁내막증처럼 주사제가

없던 질환이 목표입니다.

이 사례를 읽을 때 조심할 것이 있습니다. 1상은 안전성을 보는 시험입니다. 약이 듣는지는 아직 모릅니다. 그리고 인공지능이 만든 항체라고 소개되는 후보 중에는 처음부터 새로 만든 것이 아니라 기존 항체를 다듬은 것이 섞여 있습니다. 결합력을 조금 올리는 친화도 성숙 작업은 지금 기술이 제일 믿을 만하게 해내는 일입니다. 백지에서 시작해 사람에게 들어간 항체는 아직 손에 꼽습니다. 그 사실을 흐리지 않고 세는 것이 이 분야를 정직하게 읽는 방법입니다.

뿌옇게 흐려진 시험관은 지금도 매주 나옵니다. 달라진 것은 그 시험관이 나오기 전에 모델이 먼저 손을 든다는 점입니다. 손을 드는 정확도는 아직 사람의 눈만 못합니다. 그래도 백 개를 다 만들어 보고 아흔아홉 개를 버리던 시절과는 다릅니다.

**참고자료**


[1] Wang, W., Zhang, Y., Wei, Z., & Huang, W. (2026). AbFlow: End-to-end Paratope-Centric Antibody Design by Interaction Enhanced Flow Matching. arXiv preprint arXiv:2602.07084. https://arxiv.org/abs/2602.07084

[2] Yang, Z., He, J., Xie, J., et al. (2026). ABOPD: Antibody CDR Design via On-Policy Distillation. arXiv preprint arXiv:2607.18835. https://arxiv.org/abs/2607.18835

[3] Sun, J. Q., Babaie, M., Hou, W., Crowley, M., & Young, D. (2026). Preference-based Antibody Expression Ranking: Scaling with Large-scale Weak Supervision. arXiv preprint arXiv:2607.16263. https://arxiv.org/abs/2607.16263

[4] Zhang, Y., Cao, H., Jiang, Z., et al. (2026). ProteinOPD: Towards Effective and Efficient Preference Alignment for Protein Design. arXiv preprint arXiv:2605.10189. https://arxiv.org/abs/2605.10189

[5] Ahmed, M., et al. (2026). CHIMERA-Bench: A Benchmark Dataset for Epitope-Specific Antibody Design. arXiv preprint arXiv:2603.13431. https://arxiv.org/abs/2603.13431

[6] Rao, J., et al. (2026). BOAT: Navigating the Sea of In Silico Predictors for Antibody Design via Multi-Objective Bayesian Optimization. arXiv preprint arXiv:2604.13980. https://arxiv.org/abs/2604.13980

[7] Crouzet, S. J. (2026). Biologically-Grounded Multi-Encoder Architectures as Developability Oracles for Antibody Design. arXiv preprint arXiv:2604.09369. https://arxiv.org/abs/2604.09369

[8] Absci Corp. (2026년 6월 24일). Absci Announces Positive Interim Phase 1 Data from the HEADLINE Trial of ABS-201, a Novel Antibody Targeting the Prolactin Receptor (PRLR). https://investors.absci.com/news-releases/news-release-details/absci-announces-positive-interim-phase-1-data-headlinetm-trial/

[9] Ahmed, M., et al. (2026). EvoStruct: Bridging Evolutionary and Structural Priors for Antibody CDR Design via Protein Language Model Adaptation. arXiv preprint arXiv:2605.21485. https://arxiv.org/abs/2605.21485

[10] Shi, X., et al. (2026). Antigen-specific Antibody Multi-modal Foundation Model for Functional Antibody Design. arXiv preprint arXiv:2607.20057. https://arxiv.org/abs/2607.20057

## 4.3 멀티태스크 추론과 언어적 CoT 기술을 단백질 구조 이해에 주입한 Proteo-R1 모델

박사과정 학생이 결과 파일을 엽니다. 숫자 하나가 적혀 있습니다. 마이너스 1.8. 이 돌연변이가 결합을 그만큼 약하게 만든다는 뜻입니다. 학생은 지도교수에게 이유를 묻습니다. 교수도 모릅니다. 모델이 알려 준 것은 숫자뿐이기 때문입니다. 왜 210번 잔기가 중요한지, 어떤 상호작용이 끊어졌는지는 신경망 가중치 안에 잠겨 있습니다. 학생은 그 숫자를 믿을지 말지 스스로 정해야 합니다.

지난 몇 년의 생성 모델은 성능이 좋아질수록 이 답답함이 커졌습니다. 원자 수준의 정밀도로 구조를 만들어 내면서도, 어느 잔기가 기능에 결정적인지를 밖으로 내놓지 않았습니다. 설계 결정이 연속적인 잡음 제거 과정 안에 녹아 있었기 때문입니다. 사람이 개입할 손잡이가 없었습니다.

이것이 왜 문제가 되는지는 실패했을 때 드러납니다. 설계한 결합체 백 개를 합성했는데 아흔 개가 붙지 않았다고 합시다. 연구자는 다음 실험을 계획해야 합니다. 그런데 왜 실패했는지 모르면 고칠 곳을 정할 수 없습니다. 남은 방법은 조건을 조금 바꿔 다시 백 개를 만드는 것뿐입니다. 실험 예산은 유한한데 탐색은 눈을 가린 채 이어집니다. 계산이 빨라진 만큼 실험실의 병목이 도드라졌습니다.

2026년 5월 1일 공개된 Proteo-R1은 이 지점을 정면으로 겨냥합니다. 스탠퍼드대와 여러 기관의 연구자 스물다섯 명이 참여했고, 같은 해 ICML에서 발표되었습니다. 이름 뒤의 R1은 추론(Reasoning)을 뜻합니다. 핵심 발상은 이해와 생성을 떼어 놓는 것입니다.

기존 방식은 이해와 생성을 한 덩어리로 묶었습니다. 하나의 신경망이 표적을 보고 곧바로 좌표를 뱉습니다. 빠르지만 중간이 없습니다. 요리로 치면 재료를 넣고 뚜껑을 덮었다가 완성된 음식을 꺼내는 것과 같습니다. 간을 보는 순간이 없으니 짠지 싱거운지도 다 끓고 나서야 압니다.

Proteo-R1은 두 전문가로 이루어집니다. 앞쪽은 이해 전문가입니다. 멀티모달 거대 언어 모델(MLLM)이 단백질 서열과 구조와 문헌 설명을 함께 읽습니다. 그리고 결합과 특이성을 좌우하는 핵심 잔기를 지목합니다. 뒤쪽은 생성 전문가입니다. AlphaFold3 계열의 확산 모델이 앞에서 지목한 잔기를 고정된 제약으로 받아 구조와 서열을 함께 만듭니다. 사람

전문가가 일하는 순서와 같습니다. 어떤 접촉이 중요한지 먼저 따지고, 그다음에 기하 구조를 다듬습니다.

여기서 생각 사슬(CoT, Chain-of-Thought)이 쓰이는 방식이 남다릅니다. 언어 모델의 생각 사슬은 답을 내기 전에 중간 과정을 글로 적어 보는 기법입니다[2]. 사람이 암산 대신 종이에 풀어 쓰는 것과 같습니다. 그런데 생성 모델의 잡음 제거 궤적 안에 이 텍스트를 그대로 밀어 넣으면 불안정해집니다. 말은 그럴듯한데 구조는 흔들립니다. Proteo-R1은 추론의 결과를 문장이 아니라 잔기 번호 목록으로 굳혀서 넘깁니다. 추론이 제안이 아니라 약속이 되는 것입니다[1].

숫자로 보면 이렇습니다. SAbDab 항체 복합체로 학습한 뒤 여러 CDR을 동시에 다시 설계하는 시험에서, CDR-H1 구조 오차는 1.33옹스트롬으로 직전 최고 모델인 MFDesign의 1.61옹스트롬보다 낮았습니다. CDR-H2는 1.13옹스트롬 대 1.44옹스트롬이었습니다. RAbD 자료로 CDR-H3만 다시 설계했을 때는 복합체 품질 지표인 DockQ가 0.801을 기록했습니다. 같은 시험에서 dyMEAN은 0.407이었습니다. 국소 구조 일치도인 lDDT는 0.9693으로 비교 대상 중 제일 높았습니다[1].

같은 표에 낮은 숫자도 있습니다. 아미노산 회수율은 10.75퍼센트였습니다. 원래 항체가 쓰던 아미노산과 열에 하나만 겹쳤다는 뜻입니다. 이 숫자를 실패로 읽을지 성공으로 읽을지는 갈립니다. 자연 항체를 그대로 베끼는 것이 목표라면 낮은 값입니다. 같은 자리에서 같은 일을 하는 다른 답을 찾는 것이 목표라면 이야기가 달라집니다. 구조가 맞고 결합면이 맞는데 글자가 다르다면, 그것은 특허 회피와 면역원성 회피에서 오히려 쓸모가 있습니다. 저는 후자로 읽습니다. 서열 회수율을 성적표의 첫 줄에 두는 관행은 평가 기준이 생성 모델을 아직 못 따라간 흔적입니다.

이런 모델을 가르치려면 재료가 필요합니다. 단백질 구조 데이터는 좌표의 나열이고, 논문은 사람의 문장입니다. 둘을 이어 붙인 자료는 오랫동안 없었습니다. 2026년 2월에 공개된 AFD-Instruction이 그 빈자리를 메웠습니다. 항체의 기능 설명을 자연어 지시문 형태로 정리한 대규모 자료로, 언어 모델이 "이 항체는 무엇을 하는가"라는 질문에 답할 수 있게 만드는 교재입니다[8]. 좌표만 보던 모델에게 말을 가르친 셈입니다.

추론을 붙이려는 시도는 Proteo-R1 하나가 아닙니다. 2026년 7월의 S1-Omni는 과학 영역 전반에서 이해와 예측과 생성을 하나의 멀티모달 추론 모델로 묶으려 했습니다[3].

3월의 Agent Rosetta는 거대 언어 모델에게 로제타라는 오래된 단백질 설계 도구를 손에 쥐여 주고 과제를 스스로 풀게 했습니다[4]. 1월의 AutoBinder Agent는 흩어져 있던 설계 도구들을 MCP라는 공통 규약으로 연결해 결합체 설계를 처음부터 끝까지 자동으로 돌립니다[5]. Latent-Y는 여기서 한 발 더 나가 실험실 검증까지 포함한 자율 설계 순환을 보고했습니다[6]. 기능 명세에서 서열까지를 한 번에 잇는 공동 생성 방식도 나왔습니다[7]. 산업 현장에서는 설계와 선별을 한 줄로 이어 붙인 대량 처리 플랫폼이 자리를 잡았습니다[10]. 연구실의 시제품과 공장의 생산 라인 사이 거리가 좁아지고 있습니다.

다만 이 분야의 문헌은 아직 흩어져 있습니다. 어떤 논문은 서열 회수율로, 어떤 논문은 구조 오차로, 또 다른 논문은 결합 자유 에너지로 자기 모델을 자랑합니다. 표현 방식도 제각각입니다. 잔기 단위로 다루는 모델과 원자 단위로 다루는 모델이 같은 표에 놓입니다. 2026년 3월에 나온 정리 논문은 이 난립을 지적하면서, 표현 방식과 조건 부여 방법과 평가 기준을 하나의 틀로 묶어 보려 했습니다[9]. 기술이 앞서가고 자로 재는 법이 뒤따라오는 중입니다.

이 흐름에는 이름이 하나 더 붙습니다. 도구를 쓰는 인공지능입니다. 예전의 모델은 자기 안에서 답을 끝냈습니다. 지금의 모델은 밖에 있는 계산기를 부릅니다. 구조 예측기를 돌리고, 결합 에너지를 계산하고, 결과가 나쁘면 설계를 고쳐 다시 돌립니다. 사람이 하던 반복 작업을 그대로 옮겨 놓은 것입니다. 밤새 실험 계획을 다시 짜던 일이 새벽 두 시의 자동 실행 기록으로 바뀝니다.

그러나 조심할 지점이 있습니다. 추론 텍스트는 읽기 좋습니다. 읽기 좋은 글은 믿음을 부릅니다. 언어 모델이 "이 잔기는 2.8옹스트롬 거리에서 수소 결합을 형성합니다"라고 적어 놓으면 사람은 그 문장을 검증 없이 넘기기 쉽습니다. 그 거리는 계산된 값일 수도 있고 학습 데이터에서 흔히 보던 문장의 재생일 수도 있습니다. Proteo-R1의 설계가 영리한 지점이 여기입니다. 추론의 출력을 잔기 번호로 못 박아 두면, 그 지목이 맞았는지 틀렸는지를 뒤쪽 생성 결과로 검증할 수 있습니다. 말잔치가 아니라 확인 가능한 주장이 됩니다.

남은 숙제도 분명합니다. 이해 전문가가 지목한 잔기가 틀렸을 때, 생성 전문가는 그 틀린 지시를 그대로 따릅니다. 하드 제약이란 곧 되돌릴 수 없는 명령이기 때문입니다. 논문에 실린 절제 실험이 이 점을 보여 줍니다. 정답 접촉 잔기를 넣어 주면 구조 오차가 더 내려가고

원자끼리 부딪히는 일도 줄어듭니다[1]. 지금의 성능은 이해 전문가의 눈만큼입니다. 앞단이 좋아지면 뒷단이 따라 좋아지고, 앞단이 헛짚으면 뒷단은 정성껏 헛것을 만듭니다.

앞의 박사과정 학생 이야기로 돌아갑니다. 이제 결과 파일에는 숫자 하나 대신 잔기 목록이 함께 옵니다. 학생은 그 목록을 보고 "여기는 동의하지만 저기는 아니다"라고 말할 수 있습니다. 그 반대 의견을 제약으로 다시 넣고 모델을 돌리면 다른 설계가 나옵니다. 실험대 위에서 판정이 나기까지 걸리는 시간은 여전히 몇 주입니다. 달라진 것은 그 몇 주 동안 무엇을 시험하고 있는지 사람이 안다는 점입니다.

**참고자료**


[1] Wu, F., Xuan, W., Qi, H., et al. (2026). Proteo-R1: Reasoning Foundation Models for De Novo Protein Design. arXiv preprint arXiv:2605.02937. https://arxiv.org/abs/2605.02937

[2] Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv preprint arXiv:2201.11903. https://arxiv.org/abs/2201.11903

[3] Zhao, J., et al. (2026). S1-Omni: A Unified Multimodal Reasoning Model for Scientific Understanding, Prediction, and Generation. arXiv preprint arXiv:2607.15686. https://arxiv.org/abs/2607.15686

[4] Teneggi, J., et al. (2026). Protein Design with Agent Rosetta: A Case Study for Specialized Scientific Agents. arXiv preprint arXiv:2603.15952. https://arxiv.org/abs/2603.15952

[5] Ge, F., et al. (2026). AutoBinder Agent: An MCP-Based Agent for End-to-End Protein Binder Design. arXiv preprint arXiv:2602.00019. https://arxiv.org/abs/2602.00019

[6] Latent Labs Team, et al. (2026). Latent-Y: A Lab-Validated Autonomous Agent for De Novo Drug Design. arXiv preprint arXiv:2603.29727. https://arxiv.org/abs/2603.29727

[7] Chen, X., et al. (2026). Co-Generative De Novo Functional Protein Design. arXiv preprint arXiv:2605.00948. https://arxiv.org/abs/2605.00948

[8] Luo, L., et al. (2026). AFD-INSTRUCTION: A Comprehensive Antibody Instruction Dataset with Functional Annotations for LLM-Based Understanding and Design. arXiv preprint arXiv:2602.04916. https://arxiv.org/abs/2602.04916

[9] Wanasekara, S. H., et al. (2026). Generative Modeling in Protein Design: Neural Representations, Conditional Generation, and Evaluation Standards. arXiv preprint arXiv:2603.26378. https://arxiv.org/abs/2603.26378

[10] Gao, J., et al. (2025). HelixDesign-Binder: A Scalable Production-Grade Platform for Binder Design Built on HelixFold3. arXiv preprint arXiv:2505.21873. https://arxiv.org/abs/2505.21873

# 제5장: 게놈 수준의 DNA 언어 모델과 생물학적 코드 해석

## 5.1 DNA 유전체 전체 서열을 학습하는 뉴클레오타이드 기반 언어 모델의 사전 학습

새벽 세 시, 연구실 모니터에는 글자 네 개만 흐릅니다. A, C, G, T입니다. 스크롤을 아무리 내려도 다섯 번째 글자는 나오지 않습니다. 아데닌, 시토신, 구아닌, 티민. 이 네 글자가 사람 한 명을 짓는 설명서의 전부입니다. 분량은 30억 쌍입니다. 200쪽짜리 책으로 찍어 내면 5천 권이 넘습니다.

그 5천 권 가운데 단백질을 만들라고 지시하는 문장은 1.5%가 되지 않습니다. 나머지 98.5%는 오랫동안 쓰레기 DNA라고 불렸습니다. 지금 그 이름을 쓰는 연구자는 없습니다. 그 98.5%가 어떤 유전자를 언제 켜고, 어느 조직에서 켜고, 얼마나 세게 켤지 정하는 운영체제이기 때문입니다. 비암호화 유전체(Non-coding Genome)입니다. 요리법이 아니라 요리사의 근무표에 가깝습니다. 근무표가 어긋나면 요리법이 아무리 정확해도 주방은 멈춥니다.

게놈 언어 모델(Genomic Foundation Model)은 이 근무표를 읽으려고 만든 기계입니다. 수천 종의 박테리아, 식물, 포유류, 사람의 전장 유전체를 신문 기사 뭉치처럼 쌓아 놓고 통째로 읽힙니다. 읽히기 전에 서열을 잘라야 합니다. 이 자르기를 토큰화(Tokenization)라고 부릅니다. 6글자씩 묶는 k-mer 방식, 자주 나오는 조각을 사전으로 만드는 바이트 방식, 그리고 한 글자를 그대로 한 토큰으로 두는 단일 염기 방식이 있습니다. 셋 사이에는 분명한 맞바꿈이 있습니다. 크게 자르면 빠르고, 잘게 자르면 정확합니다. 염기 하나가 바뀌어 병이 생기는 문제를 다루려면 잘게 잘라야 합니다.

학습 방식은 빈칸 채우기 시험지와 같습니다. 마스크 언어 모델링(Masked Language Modeling)이라고 부릅니다. 서열의 15%를 무작위로 가리고, 가린 자리에 무엇이 들어갈지 맞히게 합니다. 답을 알려 주는 사람은 없습니다. 원래 서열이 곧 정답지입니다. 사람 손이 붙인 정답표가 필요 없어서 이 방식을 자기 지도 학습(Self-Supervised Learning)이라고 합니다. 모델은 가린 자리 하나를 맞히려고 앞뒤로 수만에서 수십만 글자를 훑습니다. 그 훈련을 수조 번 반복합니다.

이 지루한 반복 끝에 기계는 아무도 가르쳐 주지 않은 문법을 스스로 익힙니다. 유전자의 시동 스위치인 프로모터(Promoter), 멀리서 소리를 키우는 인핸서(Enhancer), 신호가

옆방으로 새지 않게 막는 인슐레이터(Insulator), 문장을 자르고 붙이는 자리인 스플라이스 접합 부위(Splice Site)를 구분해 냅니다. 생화학 실험으로 수십 년에 걸쳐 밝혀낸 규칙들입니다. 2021년의 DNABERT가 이 길을 열었고[1], 2024년 말 네이처 메소드에 실린 뉴클레오타이드 트랜스포머는 850종의 유전체를 읽혀 사람 유전체 과제 열여덟 가지에서 기준선을 갈아 치웠습니다[2].

모델이 여러 종의 유전체를 함께 읽는 데에는 이유가 있습니다. 사람과 생쥐가 갈라진 지 8천만 년이 넘었는데도 어떤 구간의 글자가 거의 그대로라면, 그 구간은 바뀔 때마다 개체가 죽었다는 뜻입니다. 진화가 40억 년 동안 돌린 실험의 결과가 서열 안에 눌려 있습니다. 사람 유전체 하나만 읽은 모델은 이 실험 기록을 보지 못합니다. 850종을 읽은 모델은 어느 자리가 함부로 건드리면 안 되는 자리인지 감을 잡습니다. 종간 비교라는 오래된 도구를 기계가 자기 방식으로 다시 발명한 셈입니다.

그 뒤로 판이 다시 바뀌었습니다. 2025년 12월 공개된 뉴클레오타이드 트랜스포머 3세대(NTv3)는 파라미터가 800만에서 6억 5천만 개입니다[3]. 크기만 보면 요즘 기준으로 작습니다. 그런데 한 번에 보는 창은 100만 염기입니다. 그것도 한 글자 해상도를 잃지 않은 채로 봅니다. 사전 학습에 쓴 염기는 8조 개가 넘습니다. 여기에 동식물 24종에서 모은 기능 트랙과 주석 라벨 약 1만 6천 개를 얹어 다시 학습시켰습니다. 결과는 7개 종 106개 과제로 짜인 자체 벤치마크에서 나왔습니다. 유전자 구조를 찾는 일에서 오랜 표준이던 AUGUSTUS와 SpliceAI, SegmentNT를 앞섰고, 파라미터가 10억 개인 농업 특화 모델 AgroNT보다도 유전자 발현과 단백질 양을 잘 맞혔습니다. 100만 염기 창을 처리하는 속도는 하이에나DNA나 카두세우스보다 열 배 가까이 빨랐습니다[4].

같은 시기 구글 딥마인드는 다른 길로 갔습니다. 2026년 1월 28일 네이처 649권에 실린 알파게놈은 DNA 100만 염기를 입력받아 유전자 발현, 전사 개시, 염색질 접근성, 히스톤 변형, 전사 인자 결합, 염색질 접촉 지도, 스플라이스 부위 사용까지 수천 개의 기능 트랙을 한 번에 예측합니다[5]. 변이 효과 예측 평가 26개 가운데 25개에서 기존 최고 모델과 같거나 그보다 나은 성적을 냈습니다. 모든 항목을 하나의 모델로 동시에 예측한 것은 알파게놈뿐이었습니다.

이 능력이 병원에서 어떤 장면으로 바뀌는지 보겠습니다. 원인을 모르는 발달 지연 아이의 유전체를 읽었습니다. 단백질을 만드는 부위에는 아무 문제가 없습니다. 그런데 어떤

유전자에서 8만 염기 떨어진 자리에 C가 T로 바뀐 곳이 하나 있습니다. 단일 염기 다형성(SNP)입니다. 예전에는 이 한 글자를 두고 몇 달간 세포 실험을 했습니다. 게놈 언어 모델은 바뀌기 전 서열과 바뀐 뒤 서열을 각각 집어넣고, 모델 내부의 숫자 묶음이 얼마나 흔들리는지 잽니다. 흔들림이 크면 전사 인자가 앉을 자리가 무너졌다는 뜻입니다. 판정에 걸리는 시간은 1초입니다.

이 1초가 임상유전학의 오래된 병목을 건드립니다. 병원에서 유전자 검사를 받으면 결과지에 세 가지 판정이 찍힙니다. 병을 일으킨다, 무해하다, 그리고 모르겠다입니다. 세 번째 칸의 이름은 의미 불명 변이(Variant of Uncertain Significance)입니다. 이 칸에 걸린 환자와 가족은 몇 년을 기다립니다. 수술을 할지 말지, 형제를 검사할지 말지 정하지 못한 채 기다립니다. 비암호화 영역의 변이는 이 칸에 걸리는 비율이 훨씬 높습니다. 실험으로 확인할 방법이 마땅치 않기 때문입니다. 게놈 언어 모델의 진짜 값어치는 여기 있습니다. 새로운 사실을 발견하는 쪽이 아니라, 판정을 미뤄 둔 서랍을 여는 쪽입니다.

성적표만 보면 문제가 끝난 것 같습니다. 그렇지 않습니다. 2025년 말 공개된 NABench는 고속 실험 162건에서 260만 개의 변이 서열을 모아 표준 분할과 메타데이터를 붙였습니다. 그 위에서 대표 모델 29개를 제로샷, 퓨샷, 전이 학습, 지도 학습 네 가지 방식으로 돌렸습니다[6]. 결과는 고르지 않았습니다. 어떤 모델은 DNA 과제에서 잘하고 RNA 과제에서 무너졌습니다. 2026년 5월의 ViroBench는 더 아픈 곳을 짚었습니다. 바이러스 유전체 과제 18가지에서 뉴클레오타이드 모델 66개를 돌렸더니, 계통이 멀어지거나 채집 시점이 달라지는 순간 성능이 주저앉았습니다. 생성 과제에서는 모델이 높은 확률을 매긴 서열이 실제로는 기능하지 않는 일이 잦았습니다. 통계적 그럴듯함과 생물학적 작동은 다른 것이었습니다. 사전 학습 데이터의 분류학적 다양성을 넓힌 가벼운 모델이 원래 모델보다 67.5% 나은 성적을 냈습니다. 몸집보다 식단이 중요했습니다[7].

2026년 4월에는 더 근본적인 반론이 나왔습니다. 기계론적 불변성 검사(Mechanistic Invariance Test)라는 이름의 650개 서열 벤치마크입니다[8]. 연구진은 자기회귀형, 마스크형, 양방향 상태 공간형까지 다섯 개 모델을 세워 놓고 조절 요소의 위치를 옮기거나 간격을 벌리거나 가닥 방향을 뒤집었습니다. 모델들은 위치를 알아보지 못했습니다. 겉으로 보이던 민감도는 전부 서열의 A와 T 비율에 딸려 온 것이었고, 상관계수가 0.78에서 0.96까지 나왔습니다. Evo2-1B와 카두세우스는 조절 요소가 엉뚱한 자리에 놓인 서열에 더 높은 점수를 주었습니다. 생물학의 사실을 거꾸로 뒤집은 셈입니다. 구성 성분이

위치보다 46배 크게 작용했습니다. 그리고 파라미터 100개짜리 위치 인식 행렬이 이 시험을 거의 만점으로 통과했습니다. 모델이 커질수록 편향은 옅어지지 않고 짙어졌습니다.

저는 이 결과를 실패담으로 읽지 않습니다. 문법책이 절반쯤 쓰였다는 증거로 읽습니다. 기계는 어휘를 외웠고 어순은 아직 모릅니다. 그래서 지금 필요한 일은 파라미터를 열 배 늘리는 쪽이 아니라 학습 목표와 구조를 다시 짜는 쪽입니다. 실용 연구는 이미 그쪽으로 움직입니다. 2026년 3월의 한 연구는 거대 게놈 모델의 mRNA 표현을 임베딩 수준에서 옮겨 담아 모델을 200분의 1로 줄이고도 같은 급에서 최고 성적을 냈습니다[9]. 8월에는 게놈 딥러닝의 불확실성 추정이 얼마나 믿을 만한지 따진 실증 연구가 나왔습니다[10]. 판정을 내리는 일보다 판정을 얼마나 믿을지 말하는 일이 앞선다는 뜻입니다. 환자 앞에 앉은 의사에게는 0.97이라는 점수보다, 이 점수를 믿어도 되는지가 먼저입니다.

**참고자료**


[1] Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics, 37(15), 2112-2120. https://doi.org/10.1093/bioinformatics/btab083

[2] Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., et al. (2025). Nucleotide Transformer: building and evaluating robust foundation models for human genomics. Nature Methods, 22(2), 287-297. https://doi.org/10.1038/s41592-024-02523-z

[3] Boshar, S., Evans, B., Tang, Z., et al. (2025). A foundational model for joint sequence-function multi-species modeling at scale for long-range genomic prediction. bioRxiv. https://doi.org/10.64898/2025.12.22.695963

[4] InstaDeep. (2026년 2월 18일). Modelling the Genome with NTv3. https://instadeep.com/2026/02/modelling-the-genome-with-ntv3/

[5] Avsec, ⊠., Latysheva, N., Cheng, J., Novati, G., Taylor, K. R., et al. (2026). Advancing regulatory variant effect prediction with AlphaGenome. Nature, 649(8099), 1206-1218. https://doi.org/10.1038/s41586-025-10014-0

[6] Li, Z., Ma, R., Tan, J., Tan, C., & Zheng, S. (2025). NABench: Large-Scale Benchmarks of Nucleotide Foundation Models for Fitness Prediction. arXiv preprint arXiv:2511.02888. https://arxiv.org/abs/2511.02888

[7] Ye, D., Hu, F., Hu, H., Hu, S., Tan, Y., Ouyang, W., Li, S. Z., Cui, J., & Dong, N. (2026). ViroBench: Benchmarking Nucleotide Foundation Models on Viral Genomics Tasks. arXiv preprint arXiv:2605.25388. https://arxiv.org/abs/2605.25388

[8] Cheng, B., & Zhang, J. (2026). The Mechanistic Invariance Test: Genomic Language Models Fail to Learn Positional Regulatory Logic. arXiv preprint arXiv:2604.06549. https://arxiv.org/abs/2604.06549

[9] Haidari, R., Martin, S., & Allard, M. (2026). Distilling Genomic Models for Efficient mRNA Representation Learning via Embedding Matching. arXiv preprint arXiv:2604.08574. https://arxiv.org/abs/2604.08574

[10] Saran, S., Ghanbari, M., & Ohler, U. (2026). Uncertainty-Aware Deep Learning for Genomics Applications: Insights from an Empirical Study. arXiv preprint arXiv:2608.11054. https://arxiv.org/abs/2608.11054

## 5.2 100만 염기쌍 스케일의 초장기 맥락을 단일 염기 해상도로 읽는 Evo 및 Evo2 아키텍처

도서관에서 장편소설의 줄거리를 맞히는 시험을 본다고 해 봅시다. 조건이 하나 붙습니다. 아무 데나 펼친 두 쪽만 읽을 수 있습니다. 인물 이름은 알아낼 수 있습니다. 문장이 아름다운지도 알 수 있습니다. 그러나 3부에서 왜 그 사람이 죽는지는 끝내 모릅니다. 2023년까지의 유전체 딥러닝 모델이 이 시험을 치르고 있었습니다.

원인은 트랜스포머의 어텐션에 있습니다. 어텐션은 입력의 모든 글자쌍을 서로 비교합니다. 글자가 2배로 늘면 계산량은 4배가 됩니다. 서열 길이의 제곱으로 비용이 붑니다. 그래서 실전 모델의 창은 수천 염기쌍에 머물렀습니다. 문제는 진핵생물의 유전체가 그렇게 짧은 호흡으로 움직이지 않는다는 데 있습니다. 인핸서 하나가 수십만에서 수백만 염기쌍 떨어진 프로모터를 켭니다. 그 먼 거리는 실이 꼬이듯 3차원으로 접혀 좁혀집니다. 염색질 루핑(Chromatin Looping)입니다. 지하철 노선도 위에서는 멀어 보여도 실제로는 환승 한 번이면 닿는 두 역과 같습니다. 창이 좁은 모델은 그 환승 통로를 보지 못합니다.

돌파구는 어텐션을 버리는 데서 나왔습니다. 2023년의 하이에나DNA는 어텐션 대신 긴 합성곱(Long Convolution)을 썼습니다[1]. 같은 해 등장한 맘바(Mamba)는 상태 공간 모델(State Space Model)이라는 오래된 제어공학 도구를 딥러닝에 되살렸습니다[2]. 상태 공간 모델은 지나온 서열을 하나의 요약 상태에 눌러 담고 앞으로 밀고 나갑니다. 계산량이 길이에 정비례해서 늘어납니다. 제곱이 아니라 정비례입니다. 창을 100배 늘려도 비용은 100배만 늡니다.

이 아이디어를 유전체에 부은 결과가 Evo입니다. 2024년 11월 사이언스에 실린 첫 Evo는 원핵생물 유전체 2조 7천억 염기를 읽고 분자 수준과 게놈 수준을 한 모델에서 다뤘습니다[3]. 그다음이 Evo 2입니다. 2025년 2월 프리프린트로 나와 2026년 3월 4일 네이처 652권에 정식으로 실렸습니다[4]. 숫자를 그대로 옮기겠습니다. 학습에 쓴 염기는 9조 3천억 개, 종은 12만 8천 개가 넘습니다. 세균과 고균에 머물던 첫 모델과 달리 사람, 식물, 곰팡이까지 진핵생물이 들어왔습니다. 크기는 두 종류입니다. 7B 모델은 2조 4천억 토큰을, 40B 모델은 9조 3천억 토큰을 읽었습니다. 한 번에 보는 창은 100만 토큰입니다. 그 창 안에서 해상도는 염기 하나입니다. 구조는 스트립드 하이에나 2(StripedHyena 2)이고, 최적화된 트랜스포머보다 학습이 3배 가까이 빨랐습니다.

스트립드 하이에나라는 이름은 구조를 그대로 설명합니다. 얼룩말의 줄무늬처럼 서로 다른 층을 번갈아 쌓았다는 뜻입니다. 대부분의 층은 어텐션 없이 합성곱과 상태 공간 연산으로 채우고, 사이사이에 어텐션 층을 몇 개만 끼웁니다. 값싼 층이 긴 거리를 실어 나르고, 비싼 층이 결정적인 자리에서 정밀하게 짚습니다. 화물열차 사이에 특급 객차를 몇 량 붙인 편성과 같습니다. 전부 특급으로 만들면 요금을 감당하지 못하고, 전부 화물로 만들면 정확도가 떨어집니다. 이 배합을 찾아낸 것이 Evo 2 학습 속도의 비결입니다.

100만 염기 창이 왜 중요한지는 CRISPR 치료제 설계실에서 드러납니다. 유전자 하나를 고치려고 자를 자리를 정합니다. 자르는 순간 그 자리만 바뀌는 것이 아닙니다. 30만 염기 떨어진 인핸서가 방향을 잃을 수 있고, 옆 유전자의 발현이 두 배로 뛸 수 있습니다. 창이 좁은 모델에게는 이런 질문 자체를 던질 수 없습니다. Evo 2는 편집 전후의 100만 염기를 통째로 받아 파장을 미리 계산합니다.

이 모델이 실제로 무엇을 해냈는지는 발표 1주년에 정리되었습니다[5]. 박테리오파지 설계에서는 만들어 본 285개 중 16개가 실제로 증식했고, 노린 숙주만 감염했으며 다른 균주는 건드리지 않았습니다. 후성유전체 설계에서는 예측한 염색질 접근성과 실험으로 잰 값의 AUROC가 0.92에서 0.95 사이로 맞았고, 세포 유형 특이 설계 24개 중 4개가 목표 세포에서 두 배 넘는 접근성 차이를 만들었습니다. 가축 유전체의 변이 판정에서는 변이 종류를 가린 조건에서 AUROC 0.921을 기록했습니다. 사람 쪽에서는 BRCA1 유전자의 포화 돌연변이 자료가 시험대였습니다. 유방암과 난소암 위험을 가르는 유전자입니다. Evo 2는 학습한 적 없는 이 과제에서 기능 상실 변이와 무해한 변이를 우도만으로 갈라냈고, 내부 표현 위에 얹은 분류기는 AUROC 0.95, AUPRC 0.88을 냈습니다.

읽는 일과 쓰는 일은 다릅니다. Evo 2는 두 가지를 다 합니다. 모델에게 앞부분만 주고 계속 이어 쓰라고 하면 염색체 규모의 서열을 뽑아냅니다. 여기에 추론 시점 탐색(Inference-time Search)을 붙이면 결과가 달라집니다. 모델이 후보 서열을 여러 개 만들고, 그중 원하는 성질을 가진 것만 남기며 앞으로 나아가는 방식입니다. 연구진은 이 방법으로 특정 세포에서만 염색질이 열리도록 설계한 서열을 만들었습니다. 유전체를 읽는 기계에서 유전체를 쓰는 기계로 넘어간 지점입니다.

여기서 무거운 질문이 따라옵니다. 서열을 쓸 줄 아는 기계는 병원체도 쓸 수 있습니다. Evo 2 연구진은 학습 데이터를 만들 때 진핵생물을 감염시키는 바이러스 서열을 미리 걷어

냈습니다[4]. 사람과 동물의 병원체를 모델이 배우지 못하게 막은 조치입니다. 앞 절에서 본 ViroBench의 지적이 여기에 겹칩니다. 우도가 높다고 기능하는 것은 아니지만, 그 격차가 언제까지 안전판 노릇을 할지는 아무도 모릅니다. 학습 데이터를 걸러 내는 일은 모델을 만드는 사람의 선택입니다. 선택은 다음 사람이 뒤집을 수 있습니다.

쓰는 사람들의 숫자도 남았습니다. 깃허브 다운로드 8만 8천 회, 포크 380개, 허깅페이스 모델 다운로드 10만 회 이상, 7B 모델 API 요청 200만 건과 40B 모델 600만 건, 학습 데이터 OpenGenome2 내려받기 4만 8천 회, 프리프린트 인용 200회 이상입니다. 논문 한 편이 아니라 인프라가 되었다는 뜻입니다.

그리고 이 인프라의 바닥에 금이 하나 있습니다. 2026년 4월, 한 연구자가 Evo2-7B를 세워 놓고 3차원 염색질을 아는지 물었습니다[6]. 시험 대상은 TAD 경계와 서로 마주 보는 CTCF 루프였습니다. 창은 1메가베이스였으니 TAD 하나가 통째로 들어갑니다. 방법은 두 가지였습니다. 기능적으로 의미 있는 자리를 흔들었을 때 모델의 우도가 떨어지는지 보았고, 모델에게 직접 서열을 만들게 해 마주 보는 CTCF 루프가 나오는지 보았습니다. 두 시험 모두 통과하지 못했습니다. Evo 2는 기능적 교란과 무작위 대조군을 구별하지 못했고, TAD 경계는 일부만 되찾았습니다. CTCF의 국소 문법은 배웠는데 그 위에 얹힌 3차원 살림살이는 배우지 못한 것입니다. 학습 데이터에 3차원 구조가 없는 원핵생물이 많이 섞였다는 사정도 있습니다. 연구자의 결론은 분명했습니다. 창을 더 늘리는 길이 아니라, 세포 유형과 3차원 접촉을 함께 넣는 양방향 구조가 답이라는 것입니다.

같은 모델에서 뜻밖의 능력도 나왔습니다. 2025년 11월의 연구는 Evo 2에 예시 몇 개를 문맥으로 넣어 주면 규칙을 스스로 잡아낸다는 사실을 보였습니다[7]. 예시 수를 늘릴수록 성적이 로그 선형으로 올랐습니다. 사람 언어로 학습한 모델에서만 나온다고 여겨지던 문맥 내 학습이 A와 T와 C와 G만 먹은 모델에서도 자랐습니다. 문맥 내 학습이 언어의 특성이 아니라 대규모 예측 학습의 부산물이라는 이야기가 됩니다.

응용 쪽은 이미 발이 빠릅니다. CRISPR-Cas12의 가이드 RNA 효율을 예측하는 연구는 게놈 파운데이션 모델의 표현에 염색질 접근성 데이터를 붙여, 학습 데이터가 적은 상황에서도 성능을 지켜 냈습니다[8]. 세포 유형별로 켜지는 조절 DNA를 설계하는 Ctrl-DNA는 언어 모델의 생성 능력에 제약 조건을 건 강화 학습(Reinforcement Learning)을 얹어, 원하는 세포에서만 켜지는 서열을 뽑아냈습니다[9]. 저는 이 흐름에

한쪽 편을 듭니다. 서열만 더 먹이는 확장은 이제 수확이 줄어듭니다. 실험 측정값을 학습 목표 안으로 끌어들이는 반기계론적 설계가 다음 자리를 가져갑니다[10]. 창을 100만 염기로 넓힌 일은 대단한 성취입니다. 그러나 넓은 창으로 내다본다고 바깥의 문법을 아는 것은 아닙니다. 창은 열렸고, 읽는 법은 이제 배우는 중입니다.

### 참고자료


[1] Nguyen, E., Poli, M., Faizi, M., Thomas, A., Birch-Sykes, C., Wornow, M., et al. (2023). HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. arXiv preprint arXiv:2306.15794. https://arxiv.org/abs/2306.15794

[2] Gu, A., & Dao, T. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv preprint arXiv:2312.00752. https://arxiv.org/abs/2312.00752

[3] Nguyen, E., Poli, M., Durrant, M. G., Kang, B., Katrekar, D., et al. (2024). Sequence modeling and design from molecular to genome scale with Evo. Science, 386(6723), eado9336. https://doi.org/10.1126/science.ado9336

[4] Brixi, G., Durrant, M. G., Ku, J., Naghipourfar, M., Poli, M., et al. (2026). Genome modelling and design across all domains of life with Evo 2. Nature, 652, 1349-1361. https://doi.org/10.1038/s41586-026-10176-5

[5] Arc Institute. (2026년 3월 4일). Evo 2: One Year Later. https://arcinstitute.org/news/evo-2-one-year-later

[6] Lee, U. (2026). Probing 3D Chromatin Structure Awareness in Evo2 DNA Language Model. arXiv preprint arXiv:2604.07196. https://arxiv.org/abs/2604.07196

[7] Breslow, N., Mishra, A., Revsine, M., Schatz, M. C., Liu, A., & Khashabi, D. (2025). Genomic Next-Token Predictors are In-Context Learners. arXiv preprint arXiv:2511.12797. https://arxiv.org/abs/2511.12797

[8] Dehghani Amirabad, A., Zhang, Y., Moskalev, A., Rajesh, S., Mansi, T., Li, S., Prakash, M., & Liao, R. (2025). Multimodal Modeling of CRISPR-Cas12 Activity Using Foundation Models and Chromatin Accessibility Data. arXiv preprint arXiv:2506.11182. https://arxiv.org/abs/2506.11182

[9] Chen, X., Ma, S., Lin, R., Lin, J., & Wang, B. (2025). Ctrl-DNA: Controllable Cell-Type-Specific Regulatory DNA Design via Constrained RL. arXiv preprint arXiv:2505.20578. https://arxiv.org/abs/2505.20578

[10] Kova□evi□, L., Gaudelet, T., Opzoomer, J., Triendl, H., Whittaker, J., Uhler, C., Edwards, L., & Taylor-King, J. P. (2025). No Foundations without Foundations: Why semi-mechanistic models are essential for regulatory biology. arXiv preprint arXiv:2501.19178. https://arxiv.org/abs/2501.19178

# 제6장: 미생물의 유전체 분석 및 기능적 추론

## 6.1 유전체 서열 기반 미생물 생리적 지표 예측 및 대사 물질 탐색을 위한 GGBound 에이전트

숲 바닥에서 흙 한 숟갈을 퍼 옵니다. 그 안에는 수십억 마리의 세균이 들어 있습니다. 실험실로 가져와 배지 접시에 얇게 펴 바르고 사흘을 기다립니다. 접시 위에 자라난 콜로니는 몇 개일까요. 운이 좋아야 백 개 남짓입니다. 나머지 수십억은 접시 위에서 아무 일도 하지 않습니다. 죽어서가 아닙니다. 우리가 차려 준 밥상이 그들 입맛에 맞지 않았을 뿐입니다.

배지에서 자라지 않는 이런 미생물을 난배양성 미생물(Unculturable Microorganisms)이라고 부릅니다. 배양이 어려운 미생물이라는 뜻입니다. 어떤 균은 산소가 조금만 닿아도 죽습니다. 어떤 균은 옆에 사는 다른 균이 만들어 주는 아미노산 하나가 없으면 자라지 못합니다. 미생물학자들은 오랫동안 이 사실을 알면서도 손을 쓰지 못했습니다. 자라지 않으면 관찰할 수 없고, 관찰할 수 없으면 연구할 수 없었기 때문입니다.

메타게놈 시퀀싱(Metagenomics)이 그 벽을 옆으로 돌아갔습니다. 균을 키우는 대신 흙에 든 DNA를 통째로 뽑아 읽고, 컴퓨터로 조각을 이어 붙여 개별 종의 유전체를 되살리는 방법입니다. 이렇게 복원한 유전체를 메타게놈 조립 유전체(MAG)라고 합니다. 2026년 1월 학술지에 실린 gcMeta 데이터베이스에는 시료 104,266개에서 나온 MAG 270만 개가 들어 있습니다. 이들을 종 수준으로 묶으면 109,586개 무리가 되고, 그중 69,248개가 이름조차 붙은 적 없는 새 분류군입니다. 전체의 63퍼센트입니다 [4]. 록펠러 대학 연구진은 숲 토양 시료 하나에서 2.5테라베이스쌍을 읽어 완전한 세균 유전체 수백 개를 얻었는데, 99퍼센트 넘게 처음 보는 종이었습니다 [5][10].

문제는 그다음입니다. 서열을 손에 쥐어도 그 균이 어떤 세상에서 사는지는 알 수 없습니다. 섭씨 4도의 심해에서 사는지, 80도 열수구에서 사는지. pH 3의 산성 온천을 좋아하는지, 바닷물 염도를 견디는지. 무엇을 먹고 사는지. 이 값들을 알아내려면 결국 균을 키워 봐야 했습니다. 온도를 바꿔 가며, 소금 농도를 바꿔 가며, 탄소원을 바꿔 가며 수백 번을 반복해야 표 한 줄이 채워집니다.

2026년 5월 14일 공개된 GGBound가 이 순서를 뒤집자고 제안합니다 [1]. 상하이교통대학 등의 연구진이 만든 이 에이전트는 유전체 서열을 입력으로 받아 균주의

생존 경계선을 곧바로 답합니다. 살 수 있는 온도 구간, 최적 pH, 견딜 수 있는 염도, 먹을 수 있는 기질, 세포 모양까지 하나의 문제로 묶었습니다. 연구진은 국제 미생물 분류 학술지(IJSEM) 논문과 NCBI, BacDive 데이터베이스에서 균주 1,525종에 대한 문항 6,448개를 뽑아 시험지를 만들었습니다.

GGBound의 속을 열어 보면 세 덩어리가 보입니다. 유전체 언어 모델 LucaOne이 서열을 숫자 벡터로 바꿉니다. 이 벡터를 얼려 둔 채 Qwen 언어 모델의 입력 토큰 사이에 끼워 넣습니다. 언어 모델은 비슷한 균주를 찾아오는 검색 모듈과, 유전체 규모 대사 모델(GEM)을 직접 돌려 보는 도구를 함께 씁니다. 대사 모델은 균의 대사 반응을 방정식으로 적어 둔 지도입니다. 에이전트는 여기서 특정 영양분을 빼 보고 균이 자라는지 계산합니다. 사람이 배지를 바꿔 가며 하던 실험을 컴퓨터 안에서 먼저 해 보는 셈입니다.

학습 방식에 재미있는 장치가 하나 있습니다. 연구진은 강화학습 단계에서 모델에게 가짜 유전체 벡터를 몰래 넣어 봅니다. 진짜 유전체를 넣었을 때와 답이 똑같으면, 그 답은 유전체를 읽고 낸 답이 아닙니다. 언어 모델이 원래 갖고 있던 상식으로 찍은 답입니다. GGBound는 이런 답에 상을 주지 않습니다. 진짜 유전체가 정답을 만드는 데 실제로 보탬이 됐을 때만 보상을 줍니다. 시험지에 서열을 붙여 놓고 안 보고 푸는 학생을 걸러내는 장치입니다.

결과는 40억 파라미터짜리 모델이 훨씬 큰 범용 모델들과 맞서는 그림입니다. DeepSeek-V3.2, DeepSeek-R1, GLM-4.7, Kimi-K2를 상대로 생존 경계 예측에서 대등하거나 앞섰습니다 [1]. 파라미터를 백 배 늘리는 대신 유전체를 제대로 읽게 만든 쪽이 이겼습니다.

답하는 항목 가운데 기질 이용성이 실무에서 눈에 띕니다. 어떤 당을 먹는지, 어떤 아미노산을 스스로 못 만드는지 알면 배지에 무엇을 넣어야 할지 좁혀집니다. 배양 방법을 찾아내는 연구 분야를 컬처로믹스(Culturomics)라고 부릅니다. 균 하나를 키우려고 배지 조성 수백 가지를 만들어 놓고 기다리는 일입니다. 사람이 감으로 짜던 조합을 대사 모델이 먼저 걸러 주면, 접시 개수가 수백에서 수십으로 줄어듭니다.

유전체에서 캐낼 것이 생리 지표만은 아닙니다. 세균은 항생물질, 색소, 독소 같은 2차 대사산물을 만드는 유전자들을 게놈 한자리에 나란히 모아 둡니다. 이 묶음을 생합성 유전자 클러스터(BGC, Biosynthetic Gene Clusters)라고 합니다. 공장 한 동에 컨베이어

벨트를 순서대로 늘어놓은 모습과 닮았습니다. 폴리케타이드 합성효소(PKS)와 비리보솜 펩타이드 합성효소(NRPS)는 벨트 위에서 재료를 하나씩 이어 붙이는 기계입니다. 어떤 기계가 어떤 순서로 놓였는지 읽으면 나올 물질의 뼈대를 종이 위에 그릴 수 있습니다.

antiSMASH는 이 작업을 자동으로 하는 도구입니다. 2025년 나온 8.0 버전은 찾아낼 수 있는 클러스터 유형을 81개에서 101개로 늘렸습니다 [2]. 같은 팀이 관리하는 antiSMASH 데이터베이스 5판에는 세균 54,800종, 고균 833종, 진균 421종의 유전체가 담겼고, 여기서 뽑아낸 BGC 영역이 497,429개입니다. 직전 판의 231,534개에서 두 배 넘게 늘었습니다 [3]. 이 가운데 대부분은 어떤 물질을 만드는지 아무도 모릅니다. 지도는 있는데 지명이 비어 있는 상태입니다.

빈칸을 채우는 방법도 나왔습니다. 록펠러 대학 연구진은 배양하지 못한 토양 세균의 클러스터 서열을 읽고, 그 서열이 지시하는 펩타이드를 화학적으로 직접 합성했습니다. 균을 키우지 않고 균의 설계도만 빌린 것입니다. 이렇게 만든 물질 가운데 에루타시딘(erutacidin)은 세균 막에 있는 카디올리핀이라는 지질에 달라붙어 막을 무너뜨립니다. 기존 항생제가 듣지 않는 내성균에도 효과를 보였습니다 [5].

분자를 아예 새로 그리는 쪽도 빨라졌습니다. 맥마스터 대학과 스탠퍼드 연구진이 2026년 발표한 SyntheMol-RL은 강화학습으로 항생제 후보를 설계합니다 [6]. 시판되는 빌딩블록 약 15만 개와 화학 반응 50가지를 조합하면 460억 개의 분자를 만들 수 있는데, 모델은 그 공간을 뒤져 실제로 합성 가능한 후보만 골라냅니다. 연구진은 58개를 합성해 시험했고, 그중 6개가 아시네토박터 바우마니를 비롯한 ESKAPE 병원균에 강한 활성을 보였습니다.

유전체 바깥의 지식을 끌어오는 작업도 함께 갑니다. 2025년 12월 공개된 한 연구는 언어 모델로 논문에서 미생물 균주와 화합물의 짝을 뽑아내는 시스템을 만들었습니다 [7]. 코리네박테리움 글루타미쿰과 대장균이 기능성 물질 생산에서 되풀이해 등장하는 주역이라는 결과를 확인했습니다. 실험 데이터와 논문 문장을 같은 표에 올려놓는 일입니다. 발굴한 균주를 공장으로 옮기려면 용어부터 맞아야 합니다. 2026년 2월 18일 공개된 PREFER는 정밀 발효 연구 공동체를 위한 온톨로지입니다 [8]. 바이오파운드리의 고속 발효 장비가 쏟아 내는 데이터를 서로 다른 실험실이 같은 이름으로 부르게 만드는 사전입니다. 이름이 어긋나면 데이터를 합칠 수 없고, 합칠 수 없으면 모델을 학습시킬 수 없습니다.

서열이 곧 생리는 아닙니다. GGBound의 답도 결국 예측입니다. 유전체 딥러닝 모델의 불확실성을 따져 본 2026년 8월 연구는 모델이 내놓는 확신의 정도가 실제 정확도와 잘 맞지 않는 경우를 여럿 보고했습니다 [9]. 그래도 저는 이 방향에 표를 던집니다. 배지 조성을 바꿔 가며 3년을 헤매던 자리에, 먼저 시도해 볼 후보 열 가지를 순서대로 적어 준 목록이 놓입니다. 실험을 없애는 기술이 아니라 실험의 순서를 정해 주는 기술입니다. 그 차이가 3년과 3개월을 가릅니다.

**참고자료**


[1] Huang, H., Gong, X., Wang, J., Bai, L., Xiao, X., Zhao, W., & Liang, S. (2026). GGBound: A Genome-Grounded Agent for Microbial Life-Boundary Prediction. arXiv preprint arXiv:2605.14442. https://arxiv.org/abs/2605.14442

[2] Blin, K., Shaw, S., Vader, L., et al. (2025). antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. Nucleic Acids Research, 53(W1), W32-W38. https://doi.org/10.1093/nar/gkaf334

[3] The antiSMASH database version 5. (2026). Nucleic Acids Research, 54(D1), D522-D526. https://doi.org/10.1093/nar/gkaf1210

[4] Sun, Y., et al. (2025). gcMeta 2025: a global repository of metagenome-assembled genomes enabling cross-ecosystem microbial discovery and function research. Nucleic Acids Research, 54(D1), D724. https://academic.oup.com/nar/article/54/D1/D724/8307362

[5] Burian, J., et al. (2025). Bioactive molecules unearthed by terabase-scale long-read sequencing of a soil metagenome. Nature Biotechnology. https://doi.org/10.1038/s41587-025-02810-w

[6] Swanson, K., et al. (2026). SyntheMol-RL: a flexible reinforcement learning framework for designing easily synthesizable antibiotics. Molecular Systems Biology. https://doi.org/10.1038/s44320-026-00206-9

[7] Sun, X., Sarmah, N., & Guo, M. (2025). Literature Mining System for Nutraceutical Biosynthesis: From AI Framework to Biological Insight. arXiv preprint arXiv:2512.22225. https://arxiv.org/abs/2512.22225

[8] Amigó, T., Tan, S. Z. K., Luu Phanthanourak, A., et al. (2026). PREFER: An Ontology for the PREcision FERmentation Community. arXiv preprint arXiv:2602.16755. https://arxiv.org/abs/2602.16755

[9] Saran, S., Ghanbari, M., & Ohler, U. (2026). Uncertainty-Aware Deep Learning for Genomics Applications: Insights from an Empirical Study. arXiv preprint arXiv:2608.11054. https://arxiv.org/abs/2608.11054

[10] Rockefeller University. (2025년 9월). Hundreds of new bacteria, and two potential antibiotics, found in soil. https://www.rockefeller.edu/news/38239-hundreds-of-new-bacteria-and-two-potential-antibiotics-found-in-soil/

## 6.2 오페론(Operon) 구조 및 미생물 조절 경로 해석을 위한 전문가 융합 프레임워크(MicroFuse)

지하철 객차 여러 량이 한 줄로 붙어 기관사 한 사람의 손에 움직입니다. 세균의 유전자도 그렇게 붙어 있습니다. 젖당을 분해하는 데 필요한 효소 세 개의 유전자가 게놈 위에 나란히 놓이고, 앞쪽 스위치 하나가 켜지면 셋이 함께 읽힙니다. 이 한 묶음을 오페론(Operon)이라고 부릅니다. 프로모터 하나 아래 여러 유전자가 하나의 mRNA로 전사되는 단위입니다.

1961년 프랑수아 자코브(François Jacob)와 자크 모노(Jacques Monod)가 대장균의 젖당 오페론(Lac Operon)을 밝혀냈습니다 [1]. 배지에 젖당을 넣으면 스위치가 열리고, 포도당이 함께 있으면 닫힙니다. 세균은 필요할 때만 효소를 만들어 재료와 힘을 아낍니다. 이 발견 이후 오페론은 원핵생물이 환경 변화에 맞춰 대사를 켜고 끄는 기본 단위로 자리 잡았습니다.

모노가 남긴 성장 곡선 그림에는 계단이 두 개 있습니다. 포도당과 젖당을 함께 넣은 배지에서 대장균은 먼저 포도당만 먹고 한 번 자랍니다. 포도당이 떨어지면 잠시 멈춥니다. 그 정지 구간 동안 세균은 젖당 오페론을 켭니다. 그러고 나서 두 번째 계단을 오릅니다. 멈춰 선 그 몇 분이 유전자 스위치가 돌아가는 시간이었습니다. 눈에 보이지 않는 조절 회로를 곡선 하나로 그려 낸 실험입니다.

서열만 보고 오페론의 경계를 찾는 일은 생각보다 고약합니다. 유전자와 유전자 사이 빈 구간의 길이가 일정하지 않습니다. 짧으면 같은 묶음일 가능성이 높지만 예외가 흔합니다. 묶음 안쪽에 보조 프로모터가 숨어 있어 상황에 따라 뒷부분만 따로 읽히기도 합니다. 전사를 끝내는 머리핀 구조가 애매하게 생긴 경우도 많습니다. 규칙 몇 개를 적어 두고 기계적으로 자르면 반드시 틀립니다.

2001년부터 2009년 사이에 나온 방법들은 유전자 간 거리, 기능 주석의 일치, 발현 상관관계를 확률 모형으로 묶었습니다 [2][3][4]. 대장균과 고초균처럼 실험 자료가 두껍게 쌓인 종에서는 잘 맞았습니다. 걸림돌은 새로 복원한 메타게놈 유전체였습니다. 이름도 없는 문(Phylum)에 속한 균의 유전자에는 참고할 주석도, 발현 데이터도 없습니다.

그래서 유전체 자체를 언어처럼 배우는 모델이 나왔습니다. 2025년 7월 공개된 Bacformer는 세균 게놈 하나를 단백질의 나열로 봅니다 [5]. 문장이 단어의 나열이듯 게놈은 단백질의 나열이라는 관점입니다. 게놈 130만 개, 단백질 서열 30억 개로 학습했습니다. 각 단백질은 ESM-2로 먼저 벡터가 되고, 그 벡터들이 순서 정보를 달고 트랜스포머에 들어갑니다. 연구진은 이 모델이 예측한 오페론 구조를 롱리드 RNA 시퀀싱으로 확인했고, 단백질 상호작용 예측은 AlphaFold3로 대조했습니다.

2026년 5월 9일 공개된 MicroFuse는 여기서 한 걸음 더 나갑니다 [6]. 두 종류의 증거가 있습니다. 하나는 단백질 자체가 무엇인가에 관한 증거입니다. 구조까지 배운 단백질 언어 모델 ProstT5가 이 쪽을 맡습니다. 다른 하나는 그 유전자가 게놈 어디에 어떻게 놓였는가에 관한 증거입니다. Bacformer가 이 쪽을 맡습니다. 두 벡터를 그냥 이어 붙이면 될 것 같지만, 그러면 둘이 서로 다른 말을 할 때 무슨 일이 벌어지는지 모델이 배우지 못합니다.

MicroFuse는 전문가 넷을 둡니다. 단백질 전문가, 게놈 맥락 전문가, 둘이 같은 말을 할 때를 맡는 합의 전문가, 둘이 어긋날 때를 맡는 충돌 전문가입니다. 라우터가 두 벡터를 함께 보고 넷에게 발언권을 나눠 줍니다. 회의실에 네 사람을 앉혀 놓고, 의견이 갈릴 때 갈등을 다루는 사람에게 마이크를 더 주는 구조입니다. 학습에는 두 모달리티를 같은 공간으로 끌어당기는 InfoNCE 정렬과, 의견이 갈리는 사례에 가중치를 더 얹는 대조 학습을 함께 씁니다.

두 증거가 어긋나는 일은 드물지 않습니다. 세균은 다른 종에게서 유전자 덩어리를 통째로 받아 옵니다. 수평 유전자 이동입니다. 항생제 내성 유전자가 이렇게 옮겨 다닙니다. 파지가 게놈 한가운데에 자기 DNA를 밀어 넣고 나가기도 합니다. 이런 자리에서는 옆에 나란히 놓인 두 유전자가 하는 일이 서로 무관합니다. 반대로 기능이 같은 두 유전자가 게놈 양쪽 끝에 멀리 떨어져 있기도 합니다. 배치를 믿으면 틀리고 기능을 믿어도 틀리는 자리, 충돌 전문가가 맡는 곳이 바로 거기입니다.

평가용 데이터도 새로 만들었습니다. OG-Operon100K는 OMG 메타게놈 코퍼스에서 뽑은 유전자 쌍 10만 개입니다. 같은 가닥에 놓이고 사이 간격이 짧은 쌍 5만 개를 양성으로, 간격이 크거나 배치가 어긋난 쌍 5만 개를 음성으로 삼았습니다. 학습 68,692쌍, 검증 14,169쌍, 시험 17,139쌍으로 나누되 스캐폴드 단위로 갈라 같은 조각이 양쪽에 걸치지

않게 했습니다.

스캐폴드 단위로 나눈 대목을 그냥 지나치면 안 됩니다. 메타게놈 조각 하나에서 나온 유전자 쌍이 학습과 시험에 함께 들어가면, 모델은 문제를 푸는 대신 답을 기억합니다. 시험 점수는 올라가고 실력은 그대로입니다. 같은 문제집을 풀고 같은 문제집으로 시험을 보는 셈입니다. 이 논문이 조각 단위로 칼을 댄 것은 점수를 낮추더라도 정직한 숫자를 얻겠다는 선택입니다.

숫자를 그대로 옮기겠습니다. MicroFuse의 AUROC는 0.6616입니다. 두 벡터를 그냥 이어 붙인 기준선이 0.6587, ProstT5만 쓴 모델이 0.5884, Bacformer만 쓴 모델이 0.5841이었습니다 [6]. 융합이 이겼지만 기준선과의 차이는 0.003입니다. 한쪽만 쓴 모델과 비교하면 0.07 넘게 벌어집니다. 두 시선을 합치는 일에는 분명한 값이 있고, 합치는 방식을 정교하게 다듬는 데서 나오는 이득은 아직 작습니다. 논문은 서열 유사도가 오히려 사람을 헷갈리게 만드는 어려운 사례에서 이득이 컸다고 적었습니다.

0.66이라는 숫자가 낮아 보인다면 맞습니다. 같은 해 5월 21일 발표된 트랜스포머 기반 오페론 예측 모델은 대장균에서 정확도 87.5퍼센트, 혼합 종 조건에서 F1 84.96퍼센트를 기록했습니다 [7]. 다만 두 연구가 푸는 문제가 다릅니다. 뒤쪽은 실험 주석이 촘촘한 모델 생물의 게놈을 다뤘고, MicroFuse는 이름 없는 메타게놈 조각을 다뤘습니다. 교과서를 펴 놓고 푸는 시험과 백지 위에서 푸는 시험을 같은 점수표로 볼 수는 없습니다.

유전체 언어 모델을 향한 경고도 나왔습니다. 2026년 4월 8일 공개된 한 연구는 조절 요소의 위치를 일부러 어긋나게 넣는 시험을 만들어 유전체 언어 모델 다섯 개를 검사했습니다 [8]. 모델들은 틀린 자리에 놓인 조절 요소에 더 높은 점수를 줬습니다. 파라미터 100개짜리 작은 모델이 수십억 파라미터 모델을 앞질렀습니다. 크기의 문제가 아니라 구조의 문제라는 뜻입니다. 게놈에서 위치는 배경이 아니라 내용입니다. 오페론 예측이 어려운 까닭도 여기에 있습니다.

그럼에도 이 지도를 기다리는 사람들이 있습니다. 합성생물학 연구자가 균주에 인공 대사 경로를 심을 때, 삽입 지점이 기존 오페론 한가운데면 원래 회로가 끊깁니다. 유전자 회로 설계를 강화학습으로 자동화하려는 시도가 2026년 5월에도 나왔는데 [9], 이런 도구가 내놓는 설계안이 실제 균주 안에서 돌아가려면 내인성 회로의 지도가 먼저 있어야 합니다. 상수도 생물여과기의 성능을 물리 법칙과 함께 예측하는 모델도 미생물 군집의 기능 구성을

입력으로 씁니다 [10]. 미생물 군집을 다루는 공학은 결국 누가 누구와 한 묶음으로 켜지는지를 아는 데서 출발합니다.

앞 절에서 다룬 생합성 유전자 클러스터도 대부분 오페론입니다. 항생물질 하나를 만들려면 효소 열 개가 순서대로 일해야 하고, 그 효소들의 유전자는 한 스위치 아래 묶여 있습니다. 클러스터의 시작과 끝을 잘못 자르면 조립 라인의 앞부분만 들고 오는 꼴이 됩니다. 다른 미생물에 옮겨 심어도 물질이 나오지 않습니다. 배양되지 않는 균의 유전체에서 새 항생제를 캐내려는 시도가 오페론 경계 문제와 한 줄로 이어지는 이유입니다.

오페론 예측은 화려한 분야가 아닙니다. 논문 한 편이 올린 AUROC가 0.003이면 기사가 되지 않습니다. 그런데 저는 그 0.003이 정직해 보입니다. 두 종류의 증거가 어긋나는 순간에 따로 이름을 붙이고, 그 순간을 맡을 전문가를 따로 둔 것. 그것이 이 논문이 남긴 자리입니다. 접시 위에서 자라지 않는 균의 유전자 배치를 읽어 내는 일은 이제 막 시작됐습니다.

**참고자료**


[1] Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. Journal of Molecular Biology, 3(3), 318-356. https://doi.org/10.1016/S0022-2836(61)80072-7

[2] Ermolaeva, M. D., White, O., & Salzberg, S. L. (2001). Prediction of operons in microbial genomes. Nucleic Acids Research, 29(5), 1216-1221. https://doi.org/10.1093/nar/29.5.1216

[3] Price, M. N., Huang, K. H., Alm, E. J., & Arkin, A. P. (2005). A novel method for predicting operons in genome sequences. Nucleic Acids Research, 33(3), 880-892. https://doi.org/10.1093/nar/gki232

[4] Mao, F., Dam, P., Chou, J., Olman, V., & Xu, Y. (2009). DOOR: a database for prokaryotic operons. Nucleic Acids Research, 37(Database issue), D459-D463. https://doi.org/10.1093/nar/gkn757

[5] Wiatrak, M., et al. (2025). A contextualised protein language model reveals the functional syntax of bacterial evolution. bioRxiv. https://doi.org/10.1101/2025.07.20.665723

[6] Cho, S. (2026). MicroFuse: Protein-to-Genome Expert Fusion for Microbial Operon Reasoning. arXiv preprint arXiv:2605.08815. https://arxiv.org/abs/2605.08815

[7] Assaf, R., & Fakhri, B. (2026). Transformer-based operon prediction using textual representations of gene pairs. Bioinformatics Advances, 6(1), vbag140. https://doi.org/10.1093/bioadv/vbag140

[8] Cheng, B., & Zhang, J. (2026). The Mechanistic Invariance Test: Genomic Language Models Fail to Learn Positional Regulatory Logic. arXiv preprint arXiv:2604.06549. https://arxiv.org/abs/2604.06549

[9] Flynn, N. (2026). GenCircuit-RL: Reinforcement Learning from Hierarchical Verification for Genetic Circuit Design. arXiv preprint arXiv:2605.14215. https://arxiv.org/abs/2605.14215

[10] Uzma, Cholet, F., Quinn, D., Smith, C., You, S., & Sloan, W. (2025). EnviroPiNet: A Physics-Guided AI Model for Predicting Biofilter Performance. arXiv preprint arXiv:2504.18595. https://arxiv.org/abs/2504.18595

# 제7장: AI 기반 컴퓨터 가상 스크리닝 및 화학물질 설계

## 7.1 표적 결합 포켓과 화합물 간 3차원 분자 도킹(Docking) 및 약물-표적 친화도(DTA) 가상 스크리닝

밤 열한 시의 실험실입니다. 로봇 팔이 구멍 384개가 뚫린 투명한 플레이트 위를 움직입니다. 구멍마다 서로 다른 화합물이 한 방울씩 떨어집니다. 다음 날 아침, 형광 판독기가 어느 구멍의 빛이 꺼졌는지 알려 줍니다. 빛이 꺼진 자리가 후보 물질입니다. 이것이 고속 대량 스크리닝(HTS, High-Throughput Screening)입니다. 실제 화합물을 실제 시험관에 넣고 실제로 반응을 재는 방법입니다.

신약 개발의 첫 단추는 표적 단백질의 활성 부위(Binding Pocket)에 열쇠처럼 들어맞는 저분자 유기 화합물(Small Molecule)을 찾는 일입니다. 자물쇠 구멍은 울퉁불퉁합니다. 열쇠는 딱 하나만 맞습니다. HTS는 열쇠 수십만 개를 한 개씩 꽂아 보는 방식입니다. 시약값으로 수억 원이 나가고, 판독까지 몇 달이 걸립니다. 그렇게 뒤져도 회사가 보유한 라이브러리 수백만 개를 벗어나지 못합니다.

문제는 화학 공간의 크기입니다. 약처럼 생긴 분자의 개수는 10의 60제곱 개로 추정합니다. 수백만은 그 앞에서 물 한 방울입니다. 지난 몇 해 사이 이 격차를 메운 것이 주문 제작형 가상 라이브러리입니다. Enamine REAL Space는 2026년 8월 기준으로 945억 개의 분자를 검색하고 주문할 수 있게 열어 두었습니다 [6]. 2025년 9월에는 830억 개였습니다. 1년이 안 되어 100억 개 넘게 불어난 셈입니다. 이 분자들은 검증된 병렬 합성 규칙 169가지를 재고 시약 20만여 종에 적용해 만듭니다. 주문하면 3주에서 4주 안에 도착하고, 합성 성공률은 80퍼센트를 넘습니다. 종이 위의 분자가 아니라 택배로 오는 분자입니다. 무료로 열린 ZINC-22에도 주문 가능한 분자가 370억 개 넘게 들어 있고, 도킹에 바로 쓸 수 있는 3차원 형태 45억 개가 미리 계산되어 있습니다 [5].

945억 개를 로봇 팔로 시험할 방법은 없습니다. 컴퓨터 안에서 열쇠를 꽂아 보는 수밖에 없습니다. 전통적인 분자 도킹(Molecular Docking) 프로그램인 AutoDock Vina는 물리 역장을 놓고 몬테카를로 탐색으로 리간드의 자세를 무작위로 흔들어 봅니다 [4]. 좋은 자세가 나오면 남기고 나쁜 자세는 버립니다. 정확하지만 느립니다. 리간드 하나에 수십 초가 걸립니다. 10억 개를 다 처리하려면 컴퓨터 수천 대가 몇 주를 돌아야 합니다.

도킹에는 두 가지 어려움이 겹쳐 있습니다. 하나는 자세를 맞히는 일입니다. 리간드가 포켓 안에서 어떤 각도로 앉는지 알아내야 합니다. 다른 하나는 점수를 매기는 일입니다. 그 자세가 얼마나 단단히 붙는지를 숫자로 바꿔야 합니다. 자세는 그럴듯한데 점수가 엉망인 경우가 흔합니다. 원자끼리 밀어내는 반데르발스 반발력, 수소 결합, 물을 밀어내며 붙는 소수성 접촉, 방향족 고리가 겹쳐 쌓이는 파이-파이 적층까지 다 계산에 넣어야 합니다. 이 힘들의 합을 몇 개의 항으로 요약하는 순간 오차가 따라붙습니다.

DiffDock은 이 계산을 다른 문제로 바꿔 놓았습니다 [1]. 도킹을 점수 계산이 아니라 생성 문제로 봅니다. 확산 모델(Diffusion Model)입니다. 흐릿한 안개에서 형체를 조금씩 걷어내는 방식입니다. DiffDock은 리간드가 움직일 수 있는 자유도를 셋으로 나눕니다. 앞뒤로 옮기는 병진, 축을 도는 회전, 결합을 비트는 비틀림 각도입니다. 이 셋이 이루는 곡면 위에서 확산 과정을 거꾸로 돌려 결합 자세를 직접 그려 냅니다. 성적은 분명했습니다. PDBBind 시험에서 1순위 예측이 정답과 2옹스트롬 안으로 맞은 비율이 38퍼센트였습니다. 전통 도킹은 23퍼센트, 기존 딥러닝은 20퍼센트였습니다. 실험으로 푼 구조 대신 컴퓨터가 접은 구조를 넣었을 때 차이는 더 벌어집니다. 기존 방법이 10.4퍼센트에서 멈출 때 DiffDock은 21.7퍼센트를 지켰습니다.

이런 모델의 뼈대에는 등변 그래프 신경망(Equivariant Graph Neural Network)이 들어갑니다. 등변이란 입력을 돌리면 출력도 같이 돌아간다는 뜻입니다. 책상 위의 분자 모형을 손으로 90도 돌려도 원자 사이의 거리는 그대로입니다. 사람에게는 당연한 이 사실을 신경망에 미리 심어 주면, 모델은 회전 각도마다 처음부터 다시 배우느라 데이터를 낭비하지 않습니다. DiffDock은 같은 리간드를 여러 번 흔들어 후보 자세를 수십 개 만든 뒤, 별도의 신뢰도 모델로 순위를 매깁니다. 자기가 자신 없는 예측에 스스로 낮은 점수를 매기는 구조입니다.

여기서 박수만 치면 안 됩니다. PoseBusters 연구진은 딥러닝 도킹 모델이 내놓은 자세를 물리와 화학의 잣대로 다시 재 보았습니다 [7]. 정답과의 거리는 가까운데 원자끼리 겹치거나 결합각이 말이 안 되는 자세가 수두룩했습니다. 훈련에서 본 적 없는 서열의 단백질 앞에서는 성적이 무너졌습니다. 결론은 냉정합니다. 딥러닝 도킹은 아직 고전 도구를 모든 면에서 이기지 못합니다. 그래서 현장에서는 딥러닝이 그린 자세를 물리 엔진으로 한 번 다듬어 내보냅니다. 저는 이 절충이 부끄러운 타협이 아니라 옳은 설계라고 봅니다. 속도는 신경망에서 얻고, 물리 법칙은 물리에게 맡기면 됩니다.

자세를 찾았으면 세기를 재야 합니다. 약물-표적 친화도(DTA, Drug-Target Affinity) 예측이 그 일을 합니다 [2]. 결합 해리 상수, 저해 상수, 반수 최대 억제 농도 같은 숫자로 나옵니다. 숫자가 작을수록 잘 붙습니다. 마이크로몰은 약하고, 나노몰이면 쓸 만하고, 피코몰이면 아주 강합니다. 초기 모델은 단백질 서열 문자열과 화합물 문자열만 보고 이 값을 맞혔습니다. 지금은 3차원 좌표를 그대로 읽습니다. TerraBind는 구조를 거칠게 요약하는 방식으로 기존 최고 모델보다 추론이 26배 빠르면서 친화도 정확도를 20퍼센트가량 올렸습니다 [8]. 빠르다는 말은 곧 많이 볼 수 있다는 말입니다.

속도의 의미는 Boltz-2에서 또렷해집니다 [3]. 결합 자유 에너지를 물리로 계산하는 자유 에너지 섭동(FEP)은 정확도의 기준이지만 화합물 하나에 몇 시간이 듭니다. Boltz-2는 FEP+ 벤치마크에서 피어슨 상관 0.62를 냈습니다. 널리 쓰이는 오픈소스 FEP 파이프라인과 어깨를 나란히 하면서 1000배 넘게 빨랐습니다. CASP16 친화도 과제에서는 복합체 140개를 놓고 제출된 모든 방법을 앞섰습니다. 몇 시간이 몇 초로 줄면 하던 일의 종류가 바뀝니다. 후보 열 개를 검산하던 도구가, 후보 백만 개를 훑는 도구가 됩니다.

숫자가 실험대 위에서 증명된 사례도 나왔습니다. 2026년 8월 5일, Deep Origin은 자체 도킹·채점 구조로 가상 화합물 800억 개를 훑은 결과를 공개했습니다 [9]. 암세포가 면역 공격을 피하는 데 관여하는 CD73을 표적으로 골랐습니다. 예측 상위 183개를 실제로 사서 시험했더니 56개가 활성을 보였습니다. 적중률 30.6퍼센트입니다. 같은 표적에 앞서 수행된 기계학습 스크린은 335개를 시험해 약한 저해제 하나를 건졌습니다. 적중률 0.3퍼센트였습니다. 100배 차이입니다. 대표 후보 하나는 세포 수준에서 570나노몰의 효능을 냈고, 기존 약물처럼 뉴클레오타이드를 흉내 낸 구조가 아니었습니다. 화학적으로 새로운 출발점이 생겼다는 뜻입니다. 시약 값으로 따지면 화합물 183개는 실험실 한 곳이 한 달에 감당할 만한 규모입니다. 예전 같으면 같은 수확을 얻으려 수만 개를 사야 했습니다.

모든 표적이 이렇게 풀리지는 않습니다. KRAS와 MYC처럼 포켓이 얕고 밋밋한 단백질 앞에서는 생성 모델이 여전히 헤맵니다. 깊은 주머니에는 열쇠를 밀어 넣을 수 있지만, 접시처럼 평평한 표면에는 붙잡을 데가 없습니다. ShallowBench는 이 취약점만 골라 재는 벤치마크를 만들어, 얕은 포켓에서 모델 성능이 얼마나 떨어지는지를 눈에 보이게 했습니다 [10]. 어려운 문제를 어렵다고 적어 두는 일도 진보입니다.

그림을 그리면 이렇습니다. 아침에 표적 구조를 넣습니다. 낮 동안 컴퓨터가 수십억 개를 훑습니다. 저녁에 이름 50개에서 100개가 적힌 목록이 나옵니다. 로봇 팔은 그 목록만 집습니다. 384개 구멍을 밤새 채우던 그 팔이, 이제는 컴퓨터가 골라 준 자리에만 시약을 떨어뜨립니다.

### 참고자료


[1] Corso, G., Stärk, H., Jing, B., Barzilay, R., & Jaakkola, T. (2022). DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. arXiv preprint arXiv:2210.01776. https://arxiv.org/abs/2210.01776

[2] Khan, J., & Shuvo, M. H. (2026). Recent Advances in Deep Learning-Based Drug-Target Binding Affinity Prediction. arXiv preprint arXiv:2608.13797. https://arxiv.org/abs/2608.13797

[3] Passaro, S., Corso, G., Wohlwend, J., et al. (2025). Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. bioRxiv. https://doi.org/10.1101/2025.06.14.659707

[4] Trott, O., & Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry, 31(2), 455-461. https://doi.org/10.1002/jcc.21334

[5] Tingle, B. I., Tang, K. G., Castanon, M., et al. (2023). ZINC-22: A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery. Journal of Chemical Information and Modeling, 63(4), 1166-1176. https://doi.org/10.1021/acs.jcim.2c01253

[6] Enamine. (2026년 8월 18일). REAL Space. https://enamine.net/compound-collections/real-compounds/real-space-navigator

[7] Buttenschoen, M., Morris, G. M., & Deane, C. M. (2023). PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. arXiv preprint arXiv:2308.05777. https://arxiv.org/abs/2308.05777

[8] Rossi, M., Pederson, R., Wang-Henderson, M., et al. (2026). TerraBind: Fast and Accurate Binding Affinity Prediction through Coarse Structural Representations. arXiv preprint arXiv:2602.07735. https://arxiv.org/abs/2602.07735

[9] Deep Origin. (2026년 8월 5일). Deep Origin Introduces Novel Virtual Screening Architecture; Achieves 100-fold Hit Rate Jump on Challenging Target. https://deeporigin.com/news/virtual-screening-architecture-100-fold-hit-rate/

[10] Reddy, S., & Liu, S. (2026). ShallowBench: Benchmarking Generative Drug Design Models on Shallow-Pocket Targets. arXiv preprint arXiv:2606.06717. https://arxiv.org/abs/2606.06717

## 7.2 데이터 희소성(Data Scarcity) 극대화를 극복하는 능동 학습(Active Learning)과 SPADE 프레임워크

화이트보드에 표적 단백질 이름이 적혀 있습니다. 그 아래 숫자는 12입니다. 지금까지 세상에 보고된, 이 단백질에 붙는다고 확인된 화합물의 개수입니다. 열두 개. 손가락으로 셀 수 있는 숫자입니다. 딥러닝 모델 하나를 제대로 훈련하려면 보통 수만 건이 듭니다. 희귀 유전 질환, 새로 나타난 감염병, 이제 막 구조가 풀린 미개척 표적(Undrugged Target) 앞에서 연구실은 늘 이 화이트보드를 마주합니다.

데이터가 열두 개일 때 모델을 훈련하면 무슨 일이 생길까요. 모델은 그 열두 개에 공통으로 붙어 있는 작용기 하나를 외웁니다. 그리고 그 작용기가 보이면 무조건 높은 점수를 줍니다. 과적합(Overfitting)입니다. 시험 범위만 외운 학생이 응용 문제 앞에서 무너지는 것과 같습니다. 처음 보는 화학 골격(Scaffold)이 들어오면 예측값은 근거 없는 숫자가 됩니다. 그런데 신약 개발에서 값진 후보는 대개 처음 보는 골격에서 나옵니다. 이미 알려진 골격은 다른 회사가 특허로 덮어 두었기 때문입니다.

능동 학습(Active Learning)은 이 막다른 길을 다르게 빠져나갑니다. 모델에게 답을 다 주지 않습니다. 대신 모델이 스스로 묻게 합니다. 모델은 후보 화합물마다 예측값과 함께 자신이 얼마나 헷갈리는지를 같이 내놓습니다. 그리고 제일 헷갈리는 지점, 실험 한 번으로 배울 것이 많은 지점을 골라 사람에게 넘깁니다. 웨트랩이 그 몇 개만 실제로 만들어 재고, 결과를 다시 모델에 먹입니다. 이 고리를 몇 바퀴 돌립니다. 도서관에서 아무 책이나 뽑는 대신, 자기가 모르는 서가부터 찾아가는 사람과 같습니다.

고르는 방식에도 성격이 있습니다. 점수가 높게 나온 것만 골라 실험하면 당장은 성과가 납니다. 대신 모델은 이미 아는 동네만 맴돕니다. 헷갈리는 것만 골라 실험하면 지도는 넓어지지만 당장 쓸 물질이 안 나옵니다. 실전에서는 두 전략을 섞습니다. 한 회차에 확실한 후보 절반, 낯선 후보 절반을 함께 넣는 식입니다. 탐색과 수확 사이의 비율을 정하는 일이 연구자의 판단이 개입하는 자리입니다.

효과는 숫자로 나옵니다. HASTEN 방식은 Enamine REAL의 리드 유사 화합물 15억 6천만 개를 대상으로 삼았습니다 [1]. 전체를 다 도킹하지 않고 1퍼센트만 도킹했습니다. 그러고도 실제 상위 1000개 후보 가운데 90퍼센트를 되찾았습니다. 도킹 계산의

99퍼센트를 지운 셈입니다. 2025년 3월 Nature Computational Science에 실린 연구는 이 방식을 수십억 규모의 공간에서 반복해, 실험으로 확인된 새 화합물까지 손에 넣었습니다 [2].

여기에도 그늘이 있습니다. 능동 학습에서 예측을 맡는 대리 모델은 대개 2차원 구조만 봅니다. 수용체가 어떻게 생겼는지, 리간드가 어떤 자세로 앉는지는 모릅니다. 한 연구진이 표적 여섯 개를 놓고 이 대리 모델들이 실제로 무엇을 배우는지 뜯어보았습니다 [3]. 답은 씁쓸했습니다. 모델은 점수가 높게 나온 화합물들에 흔한 구조 패턴을 외우고 있었습니다. 그래도 실전에서는 쓸모가 있었습니다. DUD-E 자료에서 진짜 활성 화합물을 찾아냈고, 처음 훈련에 쓴 집합보다 훨씬 큰 Enamine REAL에서 점수 높은 화합물을 건져 올렸습니다. 원리를 알고 쓴다는 조건이 붙습니다.

2026년 5월에 나온 SPADE는 이 고리를 표본 효율의 문제로 정면에서 다룹니다 [4]. 논문이 던지는 첫 문장이 아픕니다. 후보 리간드 가운데 신약 개발의 초기 단계라도 통과하는 비율이 5퍼센트가 안 됩니다. SPADE는 사전 데이터가 하나도 없는 새 단백질에서 출발하도록 설계되었습니다. 성적은 이렇습니다. 품질 좋은 리간드 10개를 확보하는 데 평균 40번의 시험이 들었습니다. 딥러닝 기법과 베이지안 최적화 기법을 하나씩 맞붙였을 때 표본 효율이 중앙값 기준으로 7퍼센트에서 32퍼센트 앞섰습니다. 후보를 채점하는 속도는 바로 뒤를 쫓는 경쟁 기법보다 10배 빨랐습니다. 연구진은 자료와 코드를 공개했습니다. 속도가 왜 중요한지는 고리를 한 바퀴 돌려 보면 압니다. 실험 결과가 들어올 때마다 모델을 다시 학습시키고, 후보 수억 개를 처음부터 다시 채점해야 합니다. 이 재채점이 하루씩 걸리면 웨트랩은 그동안 손을 놓습니다. 계산이 몇 시간으로 줄면 실험과 계산이 같은 주 안에서 번갈아 돌아갑니다.

40번이라는 숫자를 실험대의 언어로 옮겨 보겠습니다. 화합물 하나를 합성하고 결합 친화도를 재는 데 며칠과 수십만 원이 듭니다. 40번이면 한 사람이 두 달 안에 끝냅니다. 무작위로 뒤졌다면 같은 결과를 얻는 데 수천 번이 필요했을 일입니다. 연구비 신청서에 적히는 숫자가 바뀌고, 박사 과정 학생의 3년이 바뀝니다. 실패한 실험 수천 건을 줄인다는 말은 버려지는 시약과 폐액도 그만큼 줄어든다는 뜻입니다.

능동 학습의 심장은 불확실성을 정직하게 재는 능력입니다. 모델이 모르면서 안다고 말하면 고리 전체가 엉뚱한 곳으로 굴러갑니다. 그래서 신경망 여러 개를 함께 돌려 예측이

흩어지는 정도를 재거나, 학습 때 쓰던 드롭아웃을 예측할 때도 켜 두고 여러 번 물어보는 방법을 씁니다. 분자 확산 모델이 만들어 낸 구조마다 신뢰도를 붙이는 사후 추정 기법도 2026년 6월에 나왔습니다 [5]. 결합력만 보는 것도 위험합니다. 약은 잘 붙기만 하면 되는 물건이 아닙니다. 물에 녹아야 하고, 간에서 견뎌야 하고, 독성이 없어야 합니다. BOAT는 여러 예측기를 한꺼번에 놓고 다목적 베이지안 최적화로 이 줄다리기를 조율합니다 [6].

고리의 반대편에는 분자를 새로 그리는 모델이 서 있습니다. 요즘은 거대 언어 모델이 분자식을 문장처럼 써 내려갑니다. 문제는 그렇게 나온 문자열이 화학적으로 말이 되지 않는 경우가 많다는 점입니다. 원자가가 맞지 않거나 고리가 닫히지 않습니다. 그래서 기계가 자동으로 검증할 수 있는 규칙을 보상으로 걸고, 규칙을 지킨 출력에 점수를 주며 모델을 다시 훈련합니다 [8]. 답을 사람이 일일이 채점하지 않아도 되니 훈련을 오래 돌릴 수 있습니다. 생성기가 후보를 만들고, 능동 학습이 그중 무엇을 실험할지 고르고, 실험 결과가 다시 생성기를 고칩니다. 세 바퀴가 맞물려 돌아갑니다.

이 고리에는 관리해야 할 위험도 있습니다. 모델이 추천하는 화합물은 회차가 지날수록 서로 닮아 갑니다. 첫 회차에 우연히 잘 나온 골격 하나가 다음 회차의 훈련 데이터를 채우고, 그 데이터가 다시 같은 골격을 추천하게 만듭니다. 눈덩이가 굴러가듯 편향이 커집니다. 막는 방법은 촌스럽지만 확실합니다. 매 회차마다 추천 목록의 일부를 골격이 서로 다른 화합물로 강제 배정하는 것입니다. 실패할 것이 뻔한 실험을 일부러 섞어 넣는 셈입니다. 그 손해가 지도의 정확도를 지킵니다.

마지막 관문은 늘 시험관입니다. 생성 모델이 그려 낸 분자가 진짜 세포에서도 듣는지를 따로 검증한 연구가 있습니다 [7]. 컴퓨터 안의 점수와 시험관의 결과는 자주 어긋납니다. 그 간격을 좁힌 산업계 사례가 2026년에 나왔습니다. Insilico Medicine은 TNIK을 표적으로 가상 저해제 7만 8천 개를 만들고, 여러 목표를 함께 걸어 걸러 낸 뒤 상위 60개만 합성했습니다. 적중률이 16.7퍼센트였습니다. 그렇게 나온 렌토서팁은 2026년 7월 특발성 폐섬유증을 대상으로 임상 3상에 들어갔습니다 [9]. AI가 설계한 분자가 사람에게 대규모로 투여되는 시험대에 처음 오른 것입니다.

같은 고리는 화학 밖에서도 돕니다. 유전자 두 개, 세 개를 짝지어 세포에 넣는 조합 CRISPR 실험은 경우의 수가 폭발합니다. 연구진은 능동 학습으로 다음에 시험할 조합을 골라 실험 횟수를 크게 줄였습니다 [10]. 화합물이든 유전자든 원리는 하나입니다. 다 해 보지

않습니다. 무엇을 모르는지 먼저 계산하고, 거기부터 실험합니다.

화이트보드로 돌아가 봅니다. 숫자는 여전히 12에서 시작합니다. 달라진 것은 다음 실험을 고르는 손입니다. 예전에는 연구자의 직감이 골랐습니다. 지금은 모델이 자기가 모르는 지점을 짚어 주고, 연구자가 그 목록을 보며 판단합니다. 그렇게 백 번이 안 되는 실험으로 나노몰 수준의 리드 화합물(Lead Compound)이 손에 들어옵니다.

**참고자료**


[1] Sivula, T., Yetukuri, L., Kalliokoski, T., Käsnänen, H., Poso, A., & Pöhner, I. (2023). Machine Learning-Boosted Docking Enables the Efficient Structure-Based Virtual Screening of Giga-Scale Enumerated Chemical Libraries. Journal of Chemical Information and Modeling, 63(18), 5773-5783. https://doi.org/10.1021/acs.jcim.3c01239

[2] Luttens, A., Cabeza de Vaca, I., Sparring, L., et al. (2025). Rapid traversal of vast chemical space using machine learning-guided docking screens. Nature Computational Science, 5(4), 301-312. https://doi.org/10.1038/s43588-025-00777-x

[3] Kim, J., Nam, J., & Ryu, S. (2024). Understanding active learning of molecular docking and its applications. arXiv preprint arXiv:2406.12919. https://arxiv.org/abs/2406.12919

[4] Nandakumar, R., Fauber, B., & Chakrabarti, D. (2026). SPADE: Faster Drug Discovery by Learning from Sparse Data. arXiv preprint arXiv:2605.05370. https://arxiv.org/abs/2605.05370

[5] Seij, P., Naesseth, C. A., Mandt, S., et al. (2026). Uncertainty Estimation for Molecular Diffusion Models. arXiv preprint arXiv:2606.13451. https://arxiv.org/abs/2606.13451

[6] Rao, J., Gonzalez Hernandez, F., Gerard, L., et al. (2026). BOAT: Navigating the Sea of In Silico Predictors for Antibody Design via Multi-Objective Bayesian Optimization. arXiv preprint arXiv:2604.13980. https://arxiv.org/abs/2604.13980

[7] Osman, N., Lembo, V., Bottegoni, G., et al. (2025). From In Silico to In Vitro: Evaluating Molecule Generative Models for Hit Generation. arXiv preprint arXiv:2512.22031. https://arxiv.org/abs/2512.22031

[8] Ouyang, M., Lan, H., & Lin, W. (2026). Adopting Reinforcement Learning with Verifiable Rewards for Molecular Generation. arXiv preprint arXiv:2607.19044. https://arxiv.org/abs/2607.19044

[9] Insilico Medicine. (2026년 7월). Insilico Initiates Phase III Clinical Trial for Rentosertib, Its AI-Empowered TNIK Inhibitor for Idiopathic Pulmonary Fibrosis. https://insilico.com/news/xmjsn4l091-insilico-initiates-phase-iii-clinical-tr

[10] Qin, J., Wessels, H.-H., Fernandez-Granda, C., et al. (2024). Active learning for efficient discovery of optimal gene combinations in the combinatorial perturbation space. arXiv preprint arXiv:2411.12010. https://arxiv.org/abs/2411.12010

# 제8장: 전사체 매핑 가이드 신약 후보물질 발굴

## 8.1 환자의 대사 경로와 조절 구조를 인지하는 표적 항암제 가상 탐색 기법

배양 접시 위의 배지는 원래 맑은 붉은색입니다. 암세포를 넣고 하루가 지나면 색이 노랗게 바뀝니다. 젖산이 쌓였다는 신호입니다. 1920년대 독일의 오토 바르부르크는 이 색깔 변화를 보고 고개를 갸웃했습니다. 배양기 안에는 산소가 충분했습니다. 산소가 있으면 세포는 포도당 한 분자에서 서른 개가 넘는 에너지 화폐를 뽑아낼 수 있습니다. 그런데 암세포는 그 길을 버리고 두 개만 건지는 발효 쪽으로 달려갔습니다. 손해를 보면서 속도를 산 것입니다. 이 현상을 와버그 효과(Warburg Effect)라고 부릅니다. 암세포는 효율을 포기하고 분열 속도를 삽니다.

암세포가 바꾸는 것은 포도당 경로 하나가 아닙니다. 아미노산을 끌어오는 통로, 지방산을 만드는 조립 라인, 세포막 재료를 뽑아내는 곁가지까지 통째로 다시 짭니다. 이것을 대사 재프로그래밍이라고 부릅니다. 분열하는 세포에게 필요한 것은 에너지만이 아닙니다. 새 세포 한 개를 만들려면 DNA 재료와 지질 막과 단백질이 그만큼 더 있어야 합니다. 발효 경로는 에너지 화폐를 적게 내주는 대신 그 중간 산물들을 잔뜩 흘려 줍니다. 암세포는 발전소를 줄이고 자재 창고를 늘린 셈입니다. 많은 종양이 글루타민에 목을 매는 이유도 여기 있습니다. 질소와 탄소를 한꺼번에 대는 재료이기 때문입니다.

다시 짠 회로에는 반드시 약한 고리가 생깁니다. 정상 세포는 쓰지 않는데 암세포만 매달리는 효소가 있습니다. 그 효소를 막으면 암세포만 굶어 죽습니다. 이론은 아름답습니다.

문제는 성적표입니다. 미토콘드리아 복합체 1을 막는 IACS-010759, 대사 회로를 겨냥한 데비미스타트가 2026년까지 임상에서 잇따라 주저앉았습니다[1]. 실패의 이유는 하나로 모입니다. 같은 대장암 환자 열 명을 놓아도 열 명의 대사 회로가 제각각입니다. KRAS 변이가 있는 환자와 없는 환자가 다르고, 종양 덩어리 안쪽의 산소 없는 자리와 바깥쪽의 혈관 근처가 다릅니다. 종양 미세환경(TME, Tumor Microenvironment)은 암세포를 둘러싼 면역세포, 혈관, 섬유아세포까지 포함한 동네를 뜻합니다. 이 동네가 대사를 함께 굴립니다. 억제제 하나가 모든 환자에게 통할 리 없습니다.

여기서 게놈 규모 대사 모델(GEM, Genome-scale Metabolic Model)이 등장합니다. 지하철 노선도를 떠올리면 쉽습니다. 역은 대사물이고 선로는 효소 반응입니다. 사람

세포의 노선도에는 선로가 만 개를 넘습니다. 여기에 플럭스 밸런스 분석(FBA, Flux Balance Analysis)을 붙입니다. 플럭스는 각 선로에 몇 대의 열차가 지나가는지를 뜻합니다. FBA는 들어온 열차 수와 나간 열차 수가 역마다 맞아떨어져야 한다는 조건만 걸고, 세포가 증식을 최대로 하려면 열차를 어떻게 배차할지 계산합니다[2]. 실험을 하지 않고 방정식만 풀어 세포의 살림살이를 추정하는 방식입니다.

노선도는 사람마다 같지만 운행표는 사람마다 다릅니다. 환자 종양 조직의 전사체(Transcriptome)를 읽으면 어떤 효소 유전자가 얼마나 켜져 있는지 알 수 있습니다. 발현량이 바닥인 효소의 선로는 닫고, 치솟은 효소의 선로는 넓힙니다. 사람 손으로는 못 할 일입니다. 만 개가 넘는 선로마다 문을 열지 닫을지 정해야 하고, 문 하나를 잘못 닫으면 노선도 전체가 멈춰 버립니다. 그래서 발현량을 반응 제약 조건으로 바꾸는 알고리즘이 따로 발달했습니다. 조직 사진 한 장을 받아 그 조직만의 살아 있는 노선도를 되살리는 작업입니다.

이렇게 만든 환자 맞춤 노선도로 유전자를 하나씩 지워 보면 어느 효소가 그 환자의 암세포에게 급소인지 나옵니다. 2014년 연구는 이 방식으로 암 세포주별 대사 약점을 찾아냈고, 예측한 표적 중 상당수가 실험에서도 세포 증식을 막았습니다[3]. 종양 미세환경도 이 노선도 위에서 다뤄집니다. 종양 덩어리 한가운데는 혈관에서 멀어 산소가 희박합니다. 산소 유입량을 낮춰 놓고 계산을 다시 돌리면 그 자리의 세포가 어떤 선로로 갈아타는지 화면에 뜹니다. 같은 종양 안에서도 바깥과 안쪽이 다른 약에 죽는다는 뜻입니다.

계산만으로는 부족한 대목도 드러났습니다. FBA는 세포 안의 물질 수지가 늘 균형을 이룬다고 가정합니다. 실제 세포는 그렇게 얌전하지 않습니다. 효소는 조절 신호를 받아 켜지고 꺼지며, 같은 유전자가 있어도 단백질까지 만들어지지 않는 경우가 흔합니다. 방정식이 놓치는 이 여백을 데이터로 메우려는 시도가 이어졌습니다. 2026년 삼중음성 유방암 연구진은 대사 조정 최소화(MOMA) 계산이 뱉어낸 플럭스 값을 숫자 특징으로 바꾸어 랜덤 포레스트에 먹였습니다. 계산만 썼을 때 필수 유전자를 찾아내는 민감도는 0.37이었습니다. 기계학습을 얹자 0.55로 올라갔습니다[4]. 물리 법칙이 밑그림을 그리고 학습 모델이 빈칸을 메운 셈입니다.

암세포가 우회로를 갖고 있다는 점이 이 분야의 진짜 벽입니다. 경로 A를 막으면 경로 B를 켭니다. 그래서 두 곳을 동시에 막아야 합니다. 두 유전자를 각각 지웠을 때는 멀쩡한데 둘을 함께 지우면 죽는 조합을 합성 치사(Synthetic Lethality)라고 부릅니다. 다리가 두 개인 강을 떠올리면 됩니다. 하나만 끊으면 우회하지만 둘을 끊으면 건널 수 없습니다. BRCA 유전자가 망가진 유방암에 PARP 억제제가 듣는 이유가 그것입니다. 2026년에는 대사 노선도와 유전자 조절 회로를 함께 놓고 이런 치명적 조합을 계산하는 방법이 나왔습니다. 유전자를 끄는 조합뿐 아니라 켜는 조작까지 섞어 따집니다[5].

2026년 6월, 벨파스트 퀸스대 연구진은 옥토퍼스(Octopus)라는 이름의 다중 규모 자율 발견 엔진을 내놓았습니다[6]. 언어모델 여러 대가 무리를 이루어 가설을 만들고, 그 가설을 물리 기반 대사 계산과 CRISPR 의존성 데이터가 곧바로 검증합니다. 말만 그럴듯한 가설은 계산 단계에서 걸러집니다. 이 시스템이 대장암 전사체를 아무 정답 없이 훑은 뒤 지목한 표적은 인슐린 유사 성장인자 2(IGF2)였습니다. 항암제 5-플루오로우라실에 암세포가 버티게 만드는 고리라는 것입니다.

숫자를 봅니다. 거짓 발견율은 0.0292였습니다. 이 유전자의 발현이 높은 환자와 낮은 환자의 생존 곡선을 가른 로그랭크 검정값은 0.0007이었습니다. 연구진은 여기서 멈추지 않고 별도의 쥐 집단에서 종양 크기를 쟀습니다. 유의미한 축소가 나왔고 검정값은 0.0373이었습니다[6]. 컴퓨터가 혼자 세운 가설이 살아 있는 동물에서 확인된 것입니다.

같은 연구실은 석 달 앞서 다른 각도로 대장암을 건드렸습니다. 신경-기호 결합 방식의 세계 모델을 만들어 유전자 변이를 컴퓨터 안에서 고쳐 보는 실험입니다[7]. 표본은 83건뿐이었고 전사체 예측의 상관계수는 0.268에 그쳤습니다. 그런데 나온 이야기가 반전입니다. 변이 KRAS는 5-플루오로우라실 저항성을 오히려 눌러 주는 방패였습니다. 반대로 PIK3CA 변이를 정상으로 되돌리자 저항성이 올라갔습니다. 고쳤는데 나빠진 것입니다. 보상 회로가 생존 신호를 되살렸기 때문입니다.

저는 이 대목에서 한쪽 편을 듭니다. 상관계수 0.268은 예측 성능으로 내세울 숫자가 아닙니다. 표본 83건도 적습니다. 그런데도 이 연구가 값진 이유는 틀릴 수 있는 기전을 말로 내놓았기 때문입니다. 정확도만 높고 이유를 말하지 못하는 모델은 종양내과 의사의 처방을 바꾸지 못합니다. 왜 그런지를 말하는 모델은 다음 실험을 설계하게 만듭니다. 대사 지식을 언어모델에 통째로 먹여 대사물 관계망을 짓는 시도도 2026년 8월에

나왔습니다[8]. 계산과 언어가 같은 책상에 앉는 중입니다.

정상 세포가 반드시 써야 하는 선로는 건드리지 않고 암세포만 매달리는 선로를 골라 끊는 일. 대사 네트워크를 아는 인공지능이 하려는 일은 이 한 문장으로 줄어듭니다. 전신 독성이 줄어드는 이유도 여기 있습니다. 표적이 정확하면 화합물을 적게 써도 되고, 적게 쓰면 간과 심장이 견딜 여유가 생깁니다.

물론 화면 속 노선도가 환자의 몸은 아닙니다. 계산이 지목한 급소가 실제 조직에서도 급소인지는 세포 실험과 동물 실험이 답합니다. 옥토퍼스 연구진이 쥐를 데려온 이유가 그것입니다. 계산에서 실험으로 넘어가는 이 다리를 짧게 만드는 일이 앞으로 몇 년의 숙제입니다. 다음 절에서 다룰 방식은 아예 다른 문을 두드립니다. 대사 노선도를 그리는 대신, 세포가 뱉어내는 유전자 발현 자체를 약이 되돌려 놓게 만드는 길입니다.

**참고자료**


[1] Frontiers in Immunology. (2026). Cancer metabolism: from the Warburg effect to precision therapy. https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2026.1793553/full

[2] Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? Nature Biotechnology, 28(3), 245-248. https://doi.org/10.1038/nbt.1614

[3] Yizhak, K., Gaude, E., Le Dévédec, S., et al. (2014). Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer. Nature Communications, 5, 4349. https://doi.org/10.1038/ncomms5349

[4] Kim, B. K., et al. (2026). Integrating Genome-Scale Metabolic Modeling with Machine Learning Improves Gene Essentiality Prediction in Triple-Negative Breast Cancer. International Journal of Molecular Sciences. https://doi.org/10.3390/ijms27115059

[5] Barrena, N., et al. (2025). Beyond synthetic lethality in large-scale metabolic and regulatory network models via genetic minimal intervention set. Bioinformatics Advances, 6(1), vbaf319. https://academic.oup.com/bioinformaticsadvances/article/6/1/vbaf319/8384237

[6] Baker, C., Ren, T., Rafferty, K., Wang, H., & McDade, S. (2026). Autonomous mechanistic discovery of colorectal cancer vulnerabilities via multi-scale AI swarms. arXiv preprint arXiv:2607.16262. https://arxiv.org/abs/2607.16262

[7] Baker, C., Ren, T., Rafferty, K., & Wang, H. (2026). Contextual Invertible World Models: A Neuro-Symbolic Agentic Framework for Colorectal Cancer Drug Response. arXiv preprint arXiv:2603.02274. https://arxiv.org/abs/2603.02274

[8] Ku, D., Kwak, M. G., Pasquel, F. J., & Li, J. (2026). MetaboLLM: a metabolomics-specialized large language model for biochemical knowledge integration and predictive metabolite graph construction. arXiv preprint arXiv:2608.06253. https://arxiv.org/abs/2608.06253

[9] Hanahan, D. (2022). Hallmarks of cancer: new dimensions. Cancer Discovery, 12(1), 31-46. https://doi.org/10.1158/2159-8290.CD-21-1059

## 8.2 대규모 유전자 발현(Transcriptome) 데이터를 활용한 소분자 화합물 역설계

도서관에서 책 한 권을 잘못 꽂으면 그 책만 사라집니다. 서가 전체가 무너지지는 않습니다. 신약 개발은 오랫동안 그 한 권을 찾는 일이었습니다. 병을 일으키는 단백질 하나를 고르고, 그 단백질의 주머니에 딱 맞는 분자를 깎았습니다. 잘 통한 사례도 많습니다. 그런데 알츠하이머병, 류머티즘 관절염, 섬유증 앞에서는 이 방식이 번번이 미끄러졌습니다. 이 병들은 책 한 권이 잘못 꽂힌 상태가 아닙니다. 서가 스무 개의 순서가 통째로 흐트러진 상태입니다.

그래서 질문을 바꾼 사람들이 있습니다. 어떤 단백질을 막을까가 아니라, 세포 전체가 어떤 상태로 돌아가야 하는가를 먼저 물었습니다. 이것을 표현형 기반 신약 개발(Phenotypic Drug Discovery)이라고 부릅니다. 세포의 상태를 읽는 자는 전사체입니다. 세포 안에서 어떤 유전자가 얼마나 켜져 있는지를 숫자 벡터로 적은 것입니다. 병든 세포의 벡터와 건강한 세포의 벡터를 나란히 놓으면 차이가 보입니다. 그 차이를 정확히 뒤집는 약을 찾으면 됩니다.

2006년 브로드 연구소의 연결성 지도(Connectivity Map, CMap)가 이 발상을 처음 실물로 만들었습니다. 화합물 164종을 세포에 넣고 유전자 발현이 어떻게 흔들리는지 기록한 작은 목록이었습니다[1]. 규모는 작았지만 논리는 그대로였습니다. 병의 시그니처와 정반대 시그니처를 만드는 약을 고른다. 2017년 L1000 플랫폼이 이 목록을 백만 건 넘는 프로파일로 키웠습니다[2]. 만 개가 넘는 유전자를 전부 재는 대신 서로 정보를 나눠 갖는 랜드마크 유전자 978개만 측정하고 나머지는 계산으로 복원하는 방식입니다. 값이 백분의 일로 떨어졌습니다. 2025년 공개된 검색 엔진 L2S2는 화합물과 CRISPR 유전자 제거 실험을 합쳐 167만 8천 건의 시그니처를 한 번에 뒤집니다[3].

978개만 재고 나머지 만 개를 계산으로 채운다는 말이 미덥지 않게 들릴 수 있습니다. 근거는 있습니다. 유전자들은 서로 손을 잡고 움직입니다. 한 무리가 함께 켜지고 함께 꺼집니다. 도서관에서 어느 서가 앞에 사람이 몰렸는지만 봐도 그날 시험 과목을 짐작하는 것과 비슷합니다. 대표 몇 권의 대출 기록으로 서가 전체의 움직임을 복원하는 셈입니다. 물론 복원값은 복원값입니다. 그래도 백만 건을 실제로 다 재는 일과 비교하면 이 타협은 남는 장사였습니다.

이 창고가 실제로 약을 찾아 준 사례도 쌓였습니다. 브로드 연구소 연구진은 항암제로 개발된 적이 없는 약 4,518종을 암세포주에 붓고 어느 것이 세포를 죽이는지 훑었습니다. 당뇨약, 항생제, 소염제 사이에서 암세포를 죽이는 물질이 여럿 나왔습니다[10]. 이미 사람에게 써 본 약이라 안전성 자료가 두툼하다는 점이 큽니다. 새로 만드는 대신 이미 있는 것의 쓸모를 바꾸는 길입니다.

여기까지는 창고에서 물건을 고르는 일입니다. 창고에 없는 물건이 필요하면 어떻게 할까요. 만들면 됩니다. 전사체 역설계는 원하는 발현 변화 벡터를 조건으로 넣고 화학 구조를 거꾸로 조립하는 방식입니다. 정답 분자를 보여 주고 흉내 내게 하는 것이 아니라, 결과만 정해 놓고 원인을 만들어 내라고 시키는 셈입니다.

조건부 확산 모델(Conditional Diffusion Model)이 이 일을 맡습니다. 확산 모델은 흐릿한 안개에서 형체를 조금씩 걷어내는 방식으로 그림을 그립니다. 분자에서는 안개가 원자 좌표와 결합 정보입니다. 여기에 전사체 벡터를 조건으로 매달면, 안개를 걷어내는 방향 자체가 원하는 세포 상태 쪽으로 기웁니다. 2026년 5월 발표된 CURE는 처치 전과 처치 후의 전사체 상태에서 목적 지향 섭동 임베딩을 뽑아내고, 그것을 화학 표현과 나란히 맞춘 뒤 확산 과정을 이끕니다[4]. 연구진은 이 문제를 생성형 역문제라고 불렀습니다. 표적 구조를 모르는 상황에서도 세포가 어떻게 변해야 하는지만 알면 분자를 설계한다는 뜻입니다.

두 달 뒤인 7월에는 조건을 여러 개 동시에 거는 JoPMol이 나왔습니다[5]. 생물학적 조건과 화학적 조건을 함께 잡습니다. 세포 상태만 맞추면 합성이 불가능한 괴물 분자가 나오고, 화학 규칙만 맞추면 약효가 없는 평범한 분자가 나옵니다. 두 손잡이를 동시에 돌리는 일이 이 분야의 실무입니다.

생성 모델이 만든 구조를 실험실에서 실제로 만들 수 있느냐는 늘 따라붙는 질문입니다. 화면 위에서는 어떤 원자든 붙일 수 있습니다. 플라스크 안에서는 그렇지 않습니다. 합성 경로가 스무 단계를 넘거나 중간체가 불안정하면 그 분자는 종이 위에서 끝납니다. 그래서 요즘 모델들은 합성 접근성 점수와 약물 유사성 지표를 조건에 함께 걸어 둡니다. 상상력에 재갈을 물리는 장치입니다.

역설계가 제대로 서려면 예측이 정확해야 합니다. 내가 만든 분자를 세포에 넣으면 유전자 발현이 어떻게 변할지를 미리 알아야 합니다. 2026년 5월 공개된 셀 월드 모델

크레오데(Chreode)는 생쥐 배아 데이터 240만 세포로 학습한 뒤, 노먼 퍼터브시퀀스 데이터에서 예측 오차를 0.2121에서 0.1858로 낮췄습니다. 12.4퍼센트 개선입니다[6]. 같은 해 3월에는 검색을 붙인 방식이 나왔습니다. 예측하려는 세포 유형과 비슷한 과거 실험을 먼저 찾아와 참고한 뒤 답을 내는 구조입니다. 세포 유형을 무시하고 아무거나 끌어오는 방식은 처참하게 실패했습니다[7]. 어떤 이웃을 데려오느냐가 답을 갈랐습니다.

한 알의 약이 단백질 하나만 건드린다는 가정도 흔들리고 있습니다. 2026년 7월 나온 종설은 다중 표적 설계를 정면으로 다룹니다. 암과 퇴행성 뇌 질환에서는 보상 신호 경로와 내성 때문에 표적 하나를 잡는 방식이 밀린다는 것입니다[8]. 전사체를 조건으로 쓰는 설계는 여기에 잘 맞습니다. 몇 개의 단백질을 건드리든 세포의 최종 상태만 원하는 자리에 놓이면 되기 때문입니다.

빈틈도 정직하게 적겠습니다. 시그니처를 뒤집는다는 발상 자체에 함정이 있습니다. 병든 세포에서 올라간 유전자를 전부 내리고 내려간 유전자를 전부 올리면 병이 낫는다는 보장이 없습니다. 그 변화 중 일부는 병의 원인이지만 일부는 몸이 병과 싸우며 낸 반응이기 때문입니다. 싸우는 쪽까지 함께 눌러 버리면 상태가 나빠집니다. 열이 나는 환자에게 해열제만 주는 일과 비슷합니다. 어느 유전자가 원인이고 어느 유전자가 반응인지 가르는 작업이 계속 필요합니다.

데이터 자체의 한계도 있습니다. L1000 데이터는 배양 접시의 세포주에서 나왔습니다. 사람의 간 조직이나 뇌 조직과 같지 않습니다. 측정하는 유전자는 978개뿐이고 나머지는 추정치입니다. 세포 하나하나를 보는 대신 수만 개를 갈아서 평균을 냅니다. 그래서 유전자 발현 말고 다른 창을 함께 여는 흐름이 생겼습니다. 세포를 여섯 가지 염색으로 찍어 모양 변화를 읽는 셀 페인팅 데이터가 그것입니다. 공개된 JUMP 데이터만 115테라바이트에 이릅니다[9]. 숫자 벡터와 사진이 같은 세포를 다른 각도에서 증언합니다.

수만 개 세포를 갈아서 평균을 낸다는 대목은 다시 짚고 넘어가겠습니다. 종양 조직 한 조각에는 암세포도 있고 면역세포도 있고 혈관 세포도 있습니다. 이들을 함께 갈면 서로 다른 목소리가 하나로 뭉개집니다. 암세포에서 뚝 떨어진 유전자가 면역세포에서 치솟으면 평균은 아무 일 없었다고 말합니다. 단일 세포 단위로 읽는 기술이 이 문제를 풀고 있습니다. 세포 하나하나의 발현을 따로 재고, 그 세포가 어느 종류인지 표시한 뒤 약을 넣습니다. 어떤 세포 집단이 약에 반응했는지가 그제야 보입니다.

전사체 역설계의 매력은 겸손함에 있습니다. 표적 단백질의 3차원 구조를 모른다고 손을 놓지 않습니다. 세포가 어떤 상태에서 어떤 상태로 옮겨 가야 하는지만 알면 됩니다. 병의 원인을 완전히 이해하지 못한 채로도 약을 설계하기 시작하는 것입니다. 그 점이 불편하다는 사람도 있습니다. 저는 이쪽에 한 표를 던집니다. 알츠하이머병의 원인을 백 퍼센트 알아낼 때까지 환자가 기다려 주지 않기 때문입니다. 아스피린도 그랬습니다. 버드나무 껍질이 왜 통증을 줄이는지 아무도 모르던 시절에 사람들은 그것을 달여 마셨습니다. 작동 기전이 밝혀진 것은 그로부터 백 년 뒤였습니다. 세포의 상태를 읽고 그 상태를 되돌릴 분자를 그리는 일은, 껍질을 달이던 손이 이제 계산기를 든 모습에 가깝습니다.

**참고자료**


[1] Lamb, J., Crawford, E. D., Peck, D., et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science, 313(5795), 1929-1935. https://doi.org/10.1126/science.1132939

[2] Subramanian, A., Narayan, R., Corsello, S. M., et al. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell, 171(6), 1437-1452. https://doi.org/10.1016/j.cell.2017.10.049

[3] Marino, G. B., et al. (2025). L2S2: chemical perturbation and CRISPR KO LINCS L1000 signature search engine. Nucleic Acids Research, 53(W1), W338. https://academic.oup.com/nar/article/53/W1/W338/8123452

[4] Xu, Z., Zhang, Z., Wang, L., Liu, Z., Liu, Q., Wu, S., & Wang, L. (2026). Reading the Cell, Designing the Cure: Perturbation-Conditioned Molecular Diffusion for Function-Oriented Drug Design. arXiv preprint arXiv:2605.15243. https://arxiv.org/abs/2605.15243

[5] Yuan, H., Li, C., Ma, W., Murata, T., & Jiang, Y. (2026). Gene Expression-Informed Jointly Controlled Generative Modeling for Precision Molecular Design. arXiv preprint arXiv:2607.11978. https://arxiv.org/abs/2607.11978

[6] Qiu, M., Zheng, G., Xu, Y., Zhang, R., Ding, Y., Long, Q., & Chen, T. (2026). Chreode: A Cell World Model for One-Step Temporal Dynamics and Perturbation Prediction. arXiv preprint arXiv:2605.28111. https://arxiv.org/abs/2605.28111

[7] Di Francesco, A. G., Rubbi, A., & Liò, P. (2026). Retrieval-Augmented Generation for Predicting Cellular Responses to Gene Perturbation. arXiv preprint arXiv:2603.07233. https://arxiv.org/abs/2603.07233

[8] Han, T., Pan, Z., Ge, W., & Zhao, Q. (2026). Beyond SBDD: Geometric Deep Learning in Polypharmacology and Multi-target Drug Design. arXiv preprint arXiv:2607.20550. https://arxiv.org/abs/2607.20550

[9] Muñoz, A. F., et al. (2026). JUMP-lite: Compact, reproducible benchmarking of cell representations. arXiv preprint arXiv:2608.07632. https://arxiv.org/abs/2608.07632

[10] Corsello, S. M., Nagari, R. T., Spangler, R. D., et al. (2020). Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. Nature Cancer, 1(2), 235-248. https://doi.org/10.1038/s43018-019-0018-6

## 8.3 부작용 방지 재설계를 가이드하는 Precedent-Guided Co-Scientist 아키텍처

임상 1상 병동의 심전도 화면에는 파형이 규칙적으로 흐릅니다. 어느 날 그 파형의 한 구간이 조금 길어집니다. QT 간격입니다. 담당 의사의 손이 멈춥니다. 이 신호를 놓치면 며칠 뒤 환자의 심장이 제멋대로 떨리는 부정맥으로 이어질 수 있습니다. 신약 개발에서 제일 잔인한 순간은 약이 듣지 않을 때가 아닙니다. 약이 잘 듣는데 다른 데를 부술 때입니다.

임상에 들어간 후보 물질 열 개 중 아홉 개가 끝을 보지 못합니다. 그중 상당수를 독성이 잡아먹습니다. 화합물이 심장 근육 세포의 hERG 칼륨 통로를 막으면 QT 간격이 늘어나고 토르사드 드 푸앵트라는 치명적 부정맥이 옵니다. hERG 통로가 유독 자주 사고를 내는 데는 이유가 있습니다. 통로 안쪽 주머니가 넓고 물을 싫어하는 성질이라, 기름기 있는 고리 구조에 질소가 붙은 흔한 약물 골격이면 대충 걸려듭니다. 약을 만드는 사람들이 즐겨 쓰는 모양이 하필 이 통로가 좋아하는 모양입니다.

간에서는 다른 사고가 납니다. 간세포의 시토크롬 P450(CYP) 효소는 몸에 들어온 낯선 분자를 잘게 부수는 해체 반장입니다. 어떤 약은 이 반장을 붙잡아 일을 못 하게 만듭니다. 그러면 함께 먹은 다른 약이 분해되지 못하고 혈중에 쌓입니다. 반대 방향의 사고도 있습니다. 해체 과정에서 원래 분자보다 훨씬 사나운 조각이 튀어나와 간세포 단백질에 들러붙습니다. 약물 유도 간 손상(DILI)은 한 제약사가 해마다 3억 5천만 달러를 태우는 항목입니다. 사람 만 명 중 한 명에게만 나타나기도 해서 임상에서 걸러내기가 어렵고, 동물 실험은 사람에게 간 손상을 일으키는 약의 절반가량을 놓칩니다[1].

인공지능은 이 두 가지 사고를 예측하는 일에 오래 매달렸습니다. hERG 차단 예측은 그중 성숙한 축에 듭니다. 분자 지문을 쓴 앙상블 모델이 교차 검증에서 정확도 84.9퍼센트, 곡선하면적 0.887을 기록했습니다. 그런데 처음 보는 화합물 집합에서는 곡선하면적이 0.786으로 내려앉았습니다[1]. 간 손상 쪽에서는 2025년에 약물 300종의 사람 세포 반응을 모은 독성유전체 자원이 공개되었고, 특이도를 100퍼센트로 묶었을 때 민감도 88퍼센트로 간 손상을 짚어냈습니다[2]. 숫자는 좋습니다. 그런데 이 모델들은 "이 분자는 위험하다"까지만 말합니다. 그다음에 뭘 해야 하는지는 말해 주지 않습니다.

2026년 7월 3일 아카이브에 올라온 PRECEDE는 그 빈칸을 겨냥합니다[3]. 이름 그대로 선례를 앞세운 공동 연구자라는 뜻입니다. 김유진과 홍참길이 만들었고 국제기계학습학회 2026년 과학 인공지능 워크숍에 채택되었습니다. 하는 일은 분명합니다. 이미 있는 모약물을 받아, 지정된 부작용 하나를 줄이면서 약효는 그대로 두는 방향으로 고쳐 씁니다. 새 분자를 무에서 뽑아내는 일이 아닙니다. 고치는 일입니다.

핵심은 근거의 출처입니다. PRECEDE는 약물과 부작용을 잇는 연관 데이터, 생의학 지식 그래프, 그리고 과거에 안전성 때문에 구조를 손본 선례들을 함께 놓고 추론합니다. 언어모델 조율자가 이 재료들을 엮되, 명시적인 규칙을 지키고 중간중간 사람이 확인하는 지점을 통과해야 합니다. 연구진이 논문에서 내세운 말이 오래 남습니다. 가설은 감사 가능하고, 반증 가능하며, 기존 약리학의 테두리 안에 머물러야 한다는 것입니다[3]. 생성 모델에게 자유를 주지 않겠다는 선언입니다.

지식 그래프라는 말이 낯설 수 있습니다. 약과 단백질과 부작용과 유전자를 점으로 찍고, 둘 사이에 관계가 있으면 선을 긋습니다. 어떤 약이 어떤 효소를 막는다는 선, 그 효소가 어떤 경로에 속한다는 선, 그 경로가 망가지면 어떤 증상이 나온다는 선이 이어집니다. 화합물 하나를 던져 넣으면 이 그물망 위에서 몇 다리 건너 어떤 부작용에 닿는지를 따라갈 수 있습니다. 왜 위험한지를 문장으로 설명할 수 있다는 뜻입니다. 점수 하나만 뱉는 예측 모델과 갈리는 지점이 여기입니다.

고쳐 쓰는 과정은 정비소 작업대와 닮았습니다. 설계 담당이 구조를 하나 내놓습니다. 조율자가 그 구조를 그물망 위에 올려놓고 위험한 조각이 있는지 훑습니다. 걸리는 조각이 나오면 과거에 같은 조각으로 무너진 약이 있었는지, 그때 화학자들이 어떻게 손봤는지를 끌어옵니다. 대체안을 붙여 다시 점수를 냅니다. 약효가 떨어졌으면 되돌리고, 유지되었으면 사람에게 넘깁니다. 한 바퀴가 끝날 때마다 어떤 근거로 무엇을 바꿨는지가 기록에 남습니다.

왜 선례가 그렇게 중요할까요. 의약화학자들은 수십 년 동안 같은 수법을 반복해 왔기 때문입니다. 대사 과정에서 반응성 퀴논 이민으로 바뀌어 간세포 단백질을 붙잡는 작용기가 있으면, 그 자리에 불소를 붙여 대사를 막습니다. 산화되기 쉬운 고리는 옥세탄 고리로 갈아 끼웁니다. 크기와 전자 성질은 비슷하되 몸속에서 다르게 행동하는 대체 작용기를 바이오아이소스테어(Bioisostere, 생물학적 동배체)라고 부릅니다[4]. 약효를 내는 핵심

부위는 손대지 않고 사고를 내는 부위만 바꿔 끼우는 것입니다. 낡은 정비소 같은 일이지만 실제로 이렇게 살아난 약이 여럿입니다.

같은 시기 다른 진영도 움직였습니다. 구글의 코사이언티스트(Co-Scientist)는 저자 51명이 붙은 논문으로 2026년 네이처에 실렸습니다[5]. 여러 전문 에이전트가 가설을 만들고, 비평 담당 에이전트가 동료 심사자처럼 깎아내리고, 감독 에이전트가 이를 조율합니다. 약물 재창출과 새 표적 발굴, 항생제 내성 기전에서 검증했고 급성 골수성 백혈병 후보를 시험관 실험으로 확인했습니다. 라텐트 랩스의 라텐트-Y는 항체 설계 캠페인 전체를 스스로 돌려, 표적 아홉 개 중 여섯 개에서 실험으로 확인된 결합체를 얻었습니다. 성공률 67퍼센트이고 전문가 예상 소요 시간보다 56배 빨랐으며 결합력은 한 자릿수 나노몰이었습니다[6]. 만드는 속도는 이미 사람을 앞질렀습니다.

속도가 붙을수록 안전 장치가 급해집니다. 2026년 7월 나온 벤치마크 MolSafeEval은 생성 모델이 뱉어낸 분자가 독성이나 반응성 측면에서 위험한지를 네 가지 과제 유형에 걸쳐 따집니다[7]. 지금까지의 벤치마크가 새로움과 성능만 재 왔다는 지적입니다. 만드는 능력과 안 만들 줄 아는 능력은 다른 문제입니다.

성적표를 봅니다. 2026년 미국임상종양학회에서 발표된 분석은 63개 회사의 인공지능 기반 후보 117건이 사람 대상 임상에 들어갔다고 셌습니다. 이 중 1상을 마친 것이 60건으로 51.3퍼센트, 2상을 마친 것은 8건으로 6.8퍼센트였습니다[8]. 1상 성공률은 80에서 90퍼센트로 역사적 평균 50퍼센트를 크게 웃돕니다. 그런데 2상 성공률은 40퍼센트 언저리로 예전과 다르지 않습니다[8]. 허가를 받은 약은 아직 없고, 렌토서팁 한 건이 3상에 올라가 있습니다.

이 숫자들이 한 방향을 가리킵니다. 인공지능은 분자를 안전하고 잘 빠지게 다듬는 일을 확실히 잘합니다. 1상은 그 능력을 재는 시험입니다. 2상은 다릅니다. 그 표적을 건드리면 정말 병이 낫는지를 묻습니다. 여기서 인공지능은 아직 힘을 못 씁니다. 표적 선택이 틀렸으면 분자를 아무리 곱게 깎아도 소용이 없기 때문입니다.

임상 2상의 벽은 화학의 문제가 아니라 생물학의 문제입니다. 이 단백질을 막으면 이 병이 낫는다는 가설이 틀렸다면, 그 가설 위에 세운 분자는 아무리 곱게 다듬어도 무너집니다. 8.1에서 본 대사 노선도와 8.2에서 본 전사체 시그니처가 값진 이유가 여기 있습니다. 둘 다 표적을 고르는 단계에 개입하는 도구입니다. 화합물을 다듬는 도구와 표적을 고르는 도구가

한 줄로 이어질 때 2상 성적이 움직일 것입니다.

그래서 PRECEDE가 사람의 검토 지점을 구조 안에 못 박아 둔 선택이 옳다고 봅니다. 인상적인 자율성보다 되짚어 볼 수 있는 기록이 필요한 단계입니다. 어떤 선례를 근거로 이 작용기를 바꾸었는지, 그 근거가 어느 논문 어느 표에서 나왔는지 사람이 따라갈 수 있어야 합니다. 2상 문턱에서 무너지는 후보를 줄이는 길은 화려한 생성 모델이 아닙니다. 왜 그렇게 고쳤는지 설명할 줄 아는 시스템입니다.

임상 1상 병동의 심전도 화면으로 돌아가 봅니다. 파형이 늘어지는 그 순간에 필요한 것은 확률값 하나가 아닙니다. 이 분자의 어느 부위가 통로에 걸렸는지, 과거 어떤 약이 같은 자리에서 무너졌는지, 그때 무엇을 바꿔 살아났는지를 말해 주는 기록입니다. 인공지능이 신약 개발에 남길 진짜 자산은 빠른 설계가 아니라 그 기록일지 모릅니다.

**참고자료**


[1] Anand Phani Kumar, K. (2025). Recent advances in AI-based toxicity prediction for drug discovery. Frontiers in Chemistry. https://doi.org/10.3389/fchem.2025.1632046

[2] Bergen, V., et al. (2025). A large-scale human toxicogenomics resource for drug-induced liver injury prediction. Nature Communications. https://www.nature.com/articles/s41467-025-65690-3

[3] Kim, Y., & Hong, C. (2026). A Precedent-Guided Co-Scientist for Side-Effect-Aware Drug Redesign. arXiv preprint arXiv:2607.02944. https://arxiv.org/abs/2607.02944

[4] Meanwell, N. A. (2011). Synopsis of some recent tactical application of bioisosteres in drug design. Journal of Medicinal Chemistry, 54(8), 2529-2591. https://doi.org/10.1021/jm1013693

[5] Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Sirkovic, P., et al. (2026). Accelerating scientific discovery with Co-Scientist. Nature. https://www.nature.com/articles/s41586-026-10644-y

[6] Latent Labs Team (Schmon, S. M., Pretorius, D., Mathis, S., et al.). (2026). Latent-Y: A Lab-Validated Autonomous Agent for De Novo Drug Design. arXiv preprint arXiv:2603.29727. https://arxiv.org/abs/2603.29727

[7] Xu, T., Cao, X., Zhu, Z., Ding, K., & Chen, H. (2026). MolSafeEval: A Benchmark for Uncovering Safety Risks in AI-Generated Molecules. arXiv preprint arXiv:2607.00464. https://arxiv.org/abs/2607.00464

[8] Khairil, L., et al. (2026). AI in Drug Discovery: Clinical Failures, Regulatory Reality, and the Validation Crisis Behind the Hype. Pharmaceuticals, 19(6), 916. https://www.mdpi.com/1424-8247/19/6/916

[9] Redfern, W. S., Carlsson, L., Davis, A. S., et al. (2003). Relationships between preclinical cardiac electrophysiology, clinical QT interval prolongation and torsade de pointes for a broad range of drugs. Cardiovascular Research, 58(1), 32-45. https://doi.org/10.1016/S0008-6363(02)00846-5

# 제9장: 단일 세포 데이터 전처리와 임베딩 기술

## 9.1 단일 세포 전사체 분석(scRNA-Seq) 기술 및 이종 대형 오믹스 데이터의 인접성 정렬

과일 가게 앞에서 믹서가 돌아갑니다. 딸기와 바나나와 시금치가 한꺼번에 갈려 초록빛 주스가 됩니다. 이 주스를 아무리 정밀하게 분석해도 딸기 한 알의 맛은 되찾을 수 없습니다. 2010년대 이전의 전사체 분석이 꼭 이랬습니다. 벌크 RNA 시퀀싱(Bulk RNA-Seq)이라는 방식입니다. 수백만 개의 세포가 섞인 조직을 통째로 갈아서 mRNA를 뽑고, 조직 전체의 평균 성분표를 읽습니다. 문제는 평균이 소수를 지운다는 데 있습니다. 종양 조직 안에 1퍼센트도 안 되는 비율로 숨어 있는 암 줄기세포, 항암제를 견뎌내는 내성 세포, 면역 반응의 방향을 정하는 희귀 조절 T세포의 신호는 나머지 99퍼센트의 아우성에 파묻혔습니다. 환자를 다시 아프게 만드는 세포는 바로 그 1퍼센트인데도 그랬습니다. 평균은 점잖은 거짓말을 합니다. 병의 씨앗은 늘 평균 바깥에 있었습니다.

단일 세포 RNA 시퀀싱(scRNA-Seq)은 믹서를 치우고 세포를 한 알씩 다루는 기술입니다. 지하철 개찰구가 승객을 한 명씩 통과시키듯, 미세유체(Microfluidics) 칩이 세포를 한 개씩 통과시킵니다. 머리카락보다 가는 관 속에서 세포 하나가 기름에 둘러싸인 작은 물방울, 곧 액적(Droplet) 안에 갇힙니다. 물방울 하나가 세포 하나만을 위한 실험실이 되는 셈입니다. 액적 안에는 세포만 들어가지 않습니다. 표면에 짧은 DNA 가닥 수백만 개가 붙은 작은 구슬이 함께 들어갑니다. 이 DNA 가닥이 세포의 mRNA를 낚아채면서 두 가지 꼬리표를 붙입니다. 하나는 세포 바코드입니다. 한 구슬 위의 가닥들은 모두 같은 세포 바코드를 가지고 있어서, 나중에 수백만 개의 서열이 뒤섞여도 어느 세포에서 나온 분자인지 되짚을 수 있습니다. 다른 하나는 분자 고유 식별자(UMI, Unique Molecular Identifier)입니다. 도서관이 같은 책 열 권에 서로 다른 등록번호를 붙이는 것과 같은 원리입니다. 시퀀싱 전에 mRNA를 수만 배로 복사하는데, UMI가 없으면 복사본을 원본으로 잘못 세게 됩니다. UMI 덕분에 복사본이 아무리 많아도 원본 분자 수를 정확히 셀 수 있습니다. 물론 이 공정도 완벽하지는 않습니다. 세포 두 개가 한 방울에 같이 들어가면 세상에 없는 잡종 세포가 데이터에 생겨납니다. 터진 세포에서 흘러나온 mRNA가 남의 방울에 스며들기도 합니다. 죽어가는 세포는 미토콘드리아 유전자 비율이 치솟아 정체를 드러냅니다. 그래서 분석의 첫 단계는 언제나 이런 불량 액적과 병든 세포를 골라내는 품질 검사입니다. 실험 장면은 뜻밖에 조용합니다. 손톱만 한 칩 위에서 물방울이 초당 수천 개씩 만들어지는 동안, 연구자는 튜브 몇 개를 옮겨 꽂을 뿐입니다. 시끄러운 일은 전부 눈에

보이지 않는 관 속에서 벌어집니다.

속도는 무서울 만큼 빨라졌습니다. 2015년 연구진은 드롭시크(Drop-seq)로 생쥐 망막 세포 44,808개를 한 번에 읽어 39종의 세포 유형을 갈라냈습니다[1]. 같은 호 같은 학술지에 실린 인드롭(inDrop)은 배아줄기세포를 같은 원리로 분석했습니다[2]. 그로부터 9년 뒤인 2024년 10월, 10x 지노믹스는 GEM-X 아키텍처로 한 번에 최대 256만 개 세포를 처리하고 세포당 비용을 1센트 아래로 내렸다고 발표했습니다[3]. 세포 하나의 유전자 2만 개를 읽는 값이 껌 한 통 값보다 싸진 것입니다. 10년 사이에 실험 한 번의 규모가 수만 개에서 수백만 개로 커졌습니다. 인간 유전체 초안 하나에 13년과 수조 원이 들던 시대에서 보면 아득한 속도입니다.

읽어낸 데이터는 세포 하나마다 2만 개의 숫자가 붙은 표입니다. 사람의 눈은 2만 차원을 볼 수 없습니다. 그래서 분석의 중심에 임베딩(Embedding)이 놓입니다. 2만 개의 숫자를 그 세포의 성격이 담긴 수십 개의 좌표로 눌러 담는 기술입니다. 지구본을 평평한 세계지도로 펼치는 일과 비슷합니다. 왜곡을 완전히 피할 수는 없지만, 가까운 도시는 지도에서도 가깝게 놓아야 합니다. 잘 만든 임베딩 공간에서는 비슷한 세포끼리 모여 섬을 이루고, 줄기세포에서 성숙한 세포로 이어지는 분화의 길이 해안선처럼 드러납니다. 연구자의 모니터에 뜨는 점구름 그림에서 점 하나가 세포 하나입니다. 이 장에서 다루는 전처리 기술은 결국 이 지도를 믿을 수 있게 만드는 기술입니다.

그런데 규모가 커지자 새로운 골칫거리가 나타났습니다. 배치 효과(Batch Effect)입니다. 실험 묶음 사이에 끼어드는 체계적인 기술 오차를 가리키는 말입니다. 같은 사람의 혈액이라도 시약 묶음이 다르거나, 시퀀싱 장비가 다르거나, 검체를 받은 병원과 날짜가 다르면 데이터가 다르게 나옵니다. 같은 노래를 서로 다른 스피커로 틀면 음색이 달라지는 것과 비슷합니다. 노래는 그대로인데 소리가 다릅니다. 이 편차를 그대로 두면 컴퓨터는 서울 병원의 T세포와 부산 병원의 T세포를 서로 다른 세포 유형으로 착각합니다. 임베딩 지도 위에서는 병원별로 섬이 갈라져, 생물학이 아니라 행정구역이 지도를 그리게 됩니다. 여러 병원의 데이터를 모을수록 지도가 선명해지는 것이 아니라 병원 수만큼 조각나는 역설이 벌어집니다. 상호 최근접 이웃(MNN, Mutual Nearest Neighbors) 방법은 두 데이터 묶음에서 서로를 이웃으로 지목하는 세포 쌍을 찾아 그 차이만큼 좌표를 이동시킵니다[4]. 하모니(Harmony)는 세포를 저차원 공간에 심은 뒤 군집화와 보정을 번갈아 반복하는데, 노트북 한 대로도 수십만 세포를 몇 분 만에 정렬할 만큼 가볍습니다[5].

scVI는 아예 확률 모델을 세웁니다. 배치 정보를 조건으로 넣은 변분 오토인코더가 기술 편차와 생물학 신호를 처음부터 분리해서 배웁니다[6]. 다만 보정에도 절제가 필요합니다. 스피커의 음색 차이를 지우려다 노래의 가사까지 지우면, 곧 환자와 건강한 사람의 진짜 차이까지 지우면 분석은 실패합니다. 보정은 삭제가 아니라 번역이어야 합니다.

측정 종류가 늘면서 정렬 문제는 한 겹 더 깊어졌습니다. 유전자 발현을 읽는 scRNA-Seq, 염색질이 얼마나 열려 있는지를 읽는 scATAC-Seq, 세포 표면 단백질을 항체로 세는 CITE-Seq은 저마다 단위도 다르고 0의 분포도 다릅니다[7]. 서로 다른 언어로 쓰인 세 권의 지도책을 한 장의 지도로 포개는 일과 같습니다. LIGER 같은 행렬 분해 기법은 데이터 종류마다 공유 성분과 고유 성분을 나누어, 모든 측정이 함께 쓰는 공통 좌표를 찾아냅니다[8]. 요즘은 최적 수송(Optimal Transport) 이론도 널리 쓰입니다. 이삿짐 트럭들의 총 이동 거리가 제일 짧아지는 배차 계획을 짜듯, 한 데이터의 세포 분포를 다른 데이터의 분포로 옮기는 비용이 제일 작아지는 짝을 계산하는 방법입니다. 시간 순서로 찍힌 정지 사진들 사이를 최적 수송으로 이어 세포의 분화 궤적을 복원하는 연구가 2025년에도 이어졌습니다[9].

이 모든 전처리 기술이 모여 만든 결과물이 휴먼 셀 아틀라스(Human Cell Atlas)입니다. 사람 몸의 모든 세포 유형을 지도로 만들겠다는 국제 협력입니다. 2016년에 시작해 2026년 현재 100여 개 나라에서 3,600명이 넘는 연구자가 참여하고 있습니다. 약 1만 명의 기증자에게서 얻은 1억 개가 넘는 세포를 분석했고, 관련 논문이 450편을 넘었습니다. 폐, 신경계, 심장, 면역계를 포함한 18개 생물학 네트워크로 나뉜 첫 완성 초안이 2026년 공개를 목표로 조립되고 있습니다[10]. 초안 다음의 목표는 더 큽니다. 지구촌 인구를 고루 대표하는 기증자들에게서 수십억 개 세포를 모아 모든 장기와 조직을 아우르는 종합 아틀라스를 만드는 것입니다[10]. 수백 개 연구실이 서로 다른 장비와 시약으로 십 년 가까이 쌓은 데이터를 한 장의 세포 지도로 합치는 일입니다. 배치 효과를 지우고 이종 측정을 한 좌표계에 정렬하는 기술이 없었다면, 이 지도는 시작조차 못 했을 것입니다. 해부학 교과서가 장기의 생김새를 그렸다면, 이 아틀라스는 장기를 이루는 세포 하나하나의 이름과 상태를 적습니다. 믹서 앞에서 사라지던 딸기 한 알이, 이제는 지도 위의 좌표 하나로 살아남았습니다.

**참고자료**

[1] Macosko, E. Z., Basu, A., Satija, R., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell, 161(5), 1202-1214. https://doi.org/10.1016/j.cell.2015.05.002

[2] Klein, A. M., Mazutis, L., Akartuna, I., et al. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell, 161(5), 1187-1201. https://doi.org/10.1016/j.cell.2015.04.044

[3] 10x Genomics. (2024년 10월 15일). 10x Genomics Delivers 'Single Cell for a Single Cent' with New Chromium Launches. https://www.prnewswire.com/news-releases/10x-genomics-delivers-single-cell-for-a-single-cent-with-new-chromium-launches-302276878.html

[4] Haghverdi, L., Lun, A. T. L., Morgan, M. D., & Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbours. Nature Biotechnology, 36(5), 421-427. https://doi.org/10.1038/nbt.4091

[5] Korsunsky, I., Millard, N., Fan, J., et al. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nature Methods, 16(12), 1289-1296. https://doi.org/10.1038/s41592-019-0619-0

[6] Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nature Methods, 15(12), 1053-1058. https://doi.org/10.1038/s41592-018-0229-2

[7] Stoeckius, M., Hafemeister, C., Stephenson, W., et al. (2017). Simultaneous epitope and transcriptome measurement in single cells. Nature Methods, 14(9), 865-868. https://doi.org/10.1038/nmeth.4380

[8] Welch, J. D., Kozareva, V., Ferreira, A., et al. (2019). Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell, 177(7), 1873-1887. https://doi.org/10.1016/j.cell.2019.05.006

[9] Ling, Y., et al. (2025). CellStream: Dynamical Optimal Transport Informed Embeddings for Reconstructing Cellular Trajectories from Snapshots Data. arXiv preprint arXiv:2511.13786. https://arxiv.org/abs/2511.13786

[10] Human Cell Atlas. (2026년 8월 18일 확인). About the Human Cell Atlas. https://www.humancellatlas.org/learn-more/about-the-human-cell-atlas/

## 9.2 전사체 데이터 고유의 노이즈 극복을 위한 결측값 대치(Imputation) 벤치마킹 분석

출석부를 상상해 보겠습니다. 학생은 분명히 교실에 앉아 있었는데 출석부에는 결석으로 적혀 있습니다. 선생님이 이름을 부르는 순간에 하필 고개를 숙이고 있었기 때문입니다. 학생 탓이 아닙니다. 부르는 방식의 한계입니다. 단일 세포 데이터에서 이런 일이 매일 벌어집니다. 드롭아웃(Dropout)이라고 부르는 현상입니다. 세포 안에서 활발히 일하던 유전자가 측정 과정에서 검출되지 않아 발현량 0으로 기록되는 결측입니다. 왜 이런 일이 생길까요. 세포 하나에 든 mRNA는 다 합쳐도 1피코그램, 곧 1조분의 1그램 수준에 지나지 않습니다. 이 미미한 양을 역전사 효소가 DNA로 바꿔 붙잡는 과정에서 상당수 분자가 그냥 사라집니다. 액적 방식의 분자 포획 효율은 열에 한둘 수준에 머뭅니다. 시퀀싱 예산도 세포 하나하나에 얇게 나뉘어, 세포당 읽기 횟수가 벌크 방식과 견줄 수 없이 적습니다. 조금만 발현되는 유전자는 복권 추첨에서 번번이 빠지는 번호처럼 기록에서 누락됩니다. 그 결과 세포 수만 개에 유전자 2만 개를 곱한 거대한 발현 행렬의 80퍼센트에서 95퍼센트가 0으로 채워집니다. 종이에 비유하면 백지에 가까운 표입니다. 피해는 실제적입니다. 도움 T세포의 이름표 격인 CD4 유전자조차 실제 도움 T세포의 상당수에서 0으로 찍힙니다. 이름표가 지워진 세포는 분류에서 엉뚱한 자리로 밀려나고, 희귀 세포 집단은 통째로 흩어져 보이지 않게 됩니다.

여기서 까다로운 질문이 나옵니다. 그 0은 진짜일까요. 어떤 0은 유전자가 실제로 꺼져 있어서 생긴 생물학적 침묵입니다. 간세포에서 신경세포 유전자가 0인 것은 당연하고, 그 0이 곧 세포의 정체성입니다. 어떤 0은 유전자가 켜져 있었는데 측정이 놓친 기술적 결측입니다. 겉보기에는 둘 다 똑같은 0이라서, 눈으로는 구별할 방법이 없습니다. 그런데 이 구별이 분석의 성패를 가릅니다. 기술적 결측을 진짜 침묵으로 믿으면 희귀 세포의 핵심 유전자를 놓치고, 진짜 침묵을 결측으로 오해해 값을 채우면 세상에 없는 세포가 만들어집니다. 결측값 대치(Imputation)는 이 둘을 통계적으로 갈라내고, 놓친 값을 추정해 채워 넣는 계산 기술입니다.

채우는 방법은 세 갈래로 발전했습니다. 첫 갈래는 이웃에게 물어보는 방식입니다. MAGIC은 발현 패턴이 비슷한 세포끼리 그래프로 잇고, 그 위에서 잉크가 물에 번지듯 발현값을 확산시켜 빈칸을 이웃의 값으로 부드럽게 메웁니다[1]. 결석으로 적힌 학생의 성적을 짝꿍들의 성적으로 짐작하는 셈입니다. 확산을 몇 걸음 진행할지 정하는 값 하나가

결과를 크게 바꿉니다. 짧게 걸으면 빈칸이 남고, 오래 걸으면 반 전체가 한 사람처럼 변합니다. 둘째 갈래는 확률 분포를 세우는 방식입니다. 유전자 발현량 같은 횟수 데이터는 음이항(Negative Binomial) 분포라는 통계 모델을 잘 따릅니다. 평균과 퍼짐 정도를 따로 조절할 수 있는 횟수 분포입니다. 딥 카운트 오토인코더 DCA와 scVI는 이 분포에 0이 유난히 많이 나오는 사정까지 얹은 모델을 신경망으로 학습해서, 유전자마다 평균 발현량과 함께 이 0이 결측일 확률을 계산합니다[2]. 이웃의 값을 그대로 베끼는 대신, 데이터 전체가 따르는 법칙을 먼저 배우고 그 법칙으로 빈칸을 읽는 방식입니다. 셋째 갈래가 요즘 화제의 주인공인 단일 세포 파운데이션 모델입니다. scGPT는 3,300만 개 세포를, Geneformer는 3,000만 개 세포를 미리 학습한 트랜스포머입니다[3][4]. 두 모델은 세포를 문장처럼 읽습니다. 유전자 하나가 단어 하나입니다. scGPT는 유전자 이름과 발현량 구간을 토큰으로 바꿔 넣고, Geneformer는 발현량이 높은 유전자부터 낮은 유전자까지 순위를 매겨 줄을 세웁니다. 언어 모델이 문장의 가려진 단어를 앞뒤 문맥으로 맞히듯, 이 모델들은 한 세포에서 가려진 유전자의 발현량을 나머지 유전자들의 문맥으로 맞히도록 훈련받았습니다. 훈련의 부산물로 세포 하나를 수백 차원의 좌표로 요약하는 임베딩 능력이 생기고, 그 좌표가 결측값 추정에도 쓰입니다.

그런데 채우는 일에는 값비싼 부작용이 따라옵니다. 과대평활화(Over-smoothing)입니다. 이웃의 값으로 빈칸을 메우다 보면 세포들 사이의 진짜 차이까지 뭉개진다는 뜻입니다. 반 평균으로 빈 성적을 채우면 모두의 성적표가 서로 닮아가는 것과 같습니다. 2018년의 한 연구는 대치가 원본에 없던 가짜 상관관계와 거짓 마커 유전자를 만들어낼 수 있음을 실험으로 보였습니다[5]. 데이터를 좋게 만들려던 손질이 없던 신호를 지어낸 것입니다. 2020년의 체계적 벤치마킹은 대치 방법 18종을 세포 군집화, 분화 궤적 분석, 차등 발현 분석이라는 실제 과제에 걸쳐 비교했습니다. 결과는 차가웠습니다. 상당수 방법이 아무것도 채우지 않은 원본보다 나쁜 결과를 내는 경우가 확인되었고, 모든 과제에서 이기는 방법은 없었습니다[6]. 논문 그림 속 순위표는 과제가 바뀔 때마다 뒤집혔습니다. 군집화에서 앞서던 방법이 차등 발현 분석에서는 바닥으로 떨어졌습니다. 2026년 3월에 나온 대규모 비교 분석도 다른 데이터로 같은 그림을 다시 그렸습니다. 어떤 방법이 이기느냐는 데이터와 과제에 따라 갈렸습니다[7]. 그래서 요즘 분석 관행은 조심스러워졌습니다. 채워 넣은 값을 최종 결론에 직접 쓰기보다, scVI 같은 확률 모델의 잠재 공간에서 군집화와 시각화를 하고, 유전자 발현의 유무를 따지는 결정적 판단은 원시 데이터로 되돌아가 확인합니다. 저는 이 방향이 옳다고 봅니다. 0을 지우는

일은 지우개가 아니라 수술칼로 해야 하는 일입니다. 채우기 전과 후를 견주는 대조 실험 없이 대치부터 돌리는 분석은, 출석부를 예쁘게 만들려고 결석 기록을 몽땅 출석으로 고치는 일과 다르지 않습니다.

파운데이션 모델을 둘러싼 논쟁은 한층 뜨겁습니다. 수천만 세포를 학습했으니 어떤 과제든 잘하리라는 기대가 앞서 나갔고, 검증이 그 뒤를 쫓았습니다. 2025년 네이처 메서즈에 실린 검증 연구는 유전자 교란 반응 예측에서 딥러닝 모델들이 선형 회귀 같은 간단한 기준선을 아직 이기지 못한다고 보고했습니다[8]. 2026년 2월에는 학습 파라미터가 하나도 없는 표현 방식이 여러 하위 과제에서 파운데이션 모델을 앞선다는 분석이 나왔습니다[9]. 2026년 6월 공개된 VCBench는 Geneformer, scGPT, UCE, TranscriptFormer, Arc State 다섯 모델을 미리 등록해 둔 기준선과 정면으로 맞붙였습니다. 평가 축은 교란 반응 예측, 종을 건너뛰는 지식 이전, 유전자 조절 네트워크 추론, 발생 시간 순서 맞히기처럼 이 분야가 아끼는 과제들이었습니다. 결과는 다섯 개 평가 축 가운데 네 축에서 선형 모델과 최근접 이웃 기준선이 파운데이션 모델과 대등하거나 더 나았습니다. 한 교란 예측 과제에서는 덧셈 기준선이 0.890점을 받는 동안 Geneformer는 0.627점, scGPT는 0.545점에 그쳤습니다[10]. 수천만 세포를 삼킨 거대 모델이 덧셈 하나에 진 것입니다. 교란을 가하지 않은 대조군 발현에 교란별 평균 변화량을 더하기만 한 값이, 수억 개의 파라미터를 가진 모델보다 정답에 가까웠습니다.

그래서 파운데이션 모델이 쓸모없다는 뜻일까요. 저는 그렇게 보지 않습니다. 모델 내부를 뜯어보면 세포 유형과 유전자 프로그램에 관한 지식이 정돈된 형태로 들어 있다는 해석 연구가 쌓이고 있고, 주석이 없는 새 데이터에 세포 이름표를 붙이는 일처럼 문맥 이해가 필요한 과제에서는 이 모델들이 제 몫을 합니다. 언어 모델도 처음에는 맞춤법 검사기보다 못하다는 평가를 받던 시절이 있었습니다. 데이터와 학습 방법이 무르익으면 판이 바뀔 수 있습니다. 문제는 모델이 아니라 검증의 순서였습니다. 비싼 모델을 쓰겠다면 값싼 기준선을 먼저 이겨야 합니다. 초라해 보이는 덧셈 기준선을 벤치마크에 반드시 넣는 습관, 그것이 이 분야가 2026년에 얻은 제일 값진 교훈입니다. 출석부의 결석 표시를 고치는 손은 화려할수록 좋은 것이 아니라 정확할수록 좋습니다. 백지에 가까운 그 표 안에, 아직 이름을 얻지 못한 세포들의 이야기가 남아 있기 때문입니다.

**참고자료**

[1] van Dijk, D., Sharma, R., Nainys, J., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. Cell, 174(3), 716-729. https://doi.org/10.1016/j.cell.2018.05.061

[2] Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., & Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nature Communications, 10, 390. https://doi.org/10.1038/s41467-018-07931-2

[3] Cui, H., Wang, C., Maan, H., et al. (2024). scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nature Methods, 21(8), 1470-1480. https://doi.org/10.1038/s41592-024-02201-0

[4] Theodoris, C. V., Xiao, L., Chopra, A., et al. (2023). Transfer learning enables predictions in network biology. Nature, 618(7965), 616-624. https://doi.org/10.1038/s41586-023-06139-1

[5] Andrews, T. S., & Hemberg, M. (2018). False signals induced by single-cell imputation. F1000Research, 7, 1740. https://doi.org/10.12688/f1000research.16613.2

[6] Hou, W., Ji, Z., Ji, H., & Hicks, S. C. (2020). A systematic evaluation of single-cell RNA-sequencing imputation methods. Genome Biology, 21, 218. https://doi.org/10.1186/s13059-020-02132-x

[7] Iwashita, Y., et al. (2026). A Large-Scale Comparative Analysis of Imputation Methods for Single-Cell RNA Sequencing Data. arXiv preprint arXiv:2603.24626. https://arxiv.org/abs/2603.24626

[8] Ahlmann-Eltze, C., Huber, W., & Anders, S. (2025). Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. Nature Methods, 22. https://www.nature.com/articles/s41592-025-02772-6

[9] Souza, H., & Mehta, P. (2026). Parameter-free representations outperform single-cell foundation models on downstream benchmarks. arXiv preprint arXiv:2602.16696. https://arxiv.org/abs/2602.16696

[10] Weidener, L., et al. (2026). VCBench: A Multi-Dimensional Benchmark for Single-Cell Foundation Models. bioRxiv. https://www.biorxiv.org/content/10.64898/2026.06.18.733146v1

# 제10장: 단일 세포 상태 전이와 동적 섭동 반응 분석

## 10.1 시간 흐름에 따른 세포 발달 궤적 및 외부 섭동 자극 시 전사체 변화 모델링 (Chreode)

실험대 위에 플레이트 하나가 놓여 있습니다. 웰마다 같은 날 같은 배지에서 자란 세포가 들어 있습니다. 연구자는 6시간째에 첫 줄을 꺼냅니다. 세포막을 터뜨리고 안에 있던 mRNA를 뽑아냅니다. 그 세포는 그 순간 죽습니다. 12시간째에 두 번째 줄을 꺼냅니다. 그 세포도 죽습니다. 24시간, 48시간, 72시간. 줄이 하나씩 비어 갑니다. 실험이 끝나면 데이터 파일 다섯 개가 남고, 살아서 끝까지 간 세포는 하나도 없습니다.

단일 세포 시퀀싱은 이렇게 작동합니다. 세포를 부수어야 그 안의 전사체를 읽을 수 있습니다. 그래서 우리가 손에 쥔 것은 한 세포가 자라는 동영상이 아닙니다. 서로 다른 세포 수만 개를 한 순간에 찍은 단체 사진 다섯 장입니다. 사진 속 인물은 매번 다른 사람입니다. 그런데 우리가 알고 싶은 것은 사진이 아니라 시간입니다. 줄기세포가 심근세포로 갈지 신경세포로 갈지, 정상 간세포가 염증 자극을 받고 며칠 뒤 어디에 가 있을지는 연속된 시간 위에서 벌어지는 일입니다.

발생생물학자 콘래드 워딩턴은 1957년에 이 문제를 그림 한 장으로 정리했습니다. 언덕 꼭대기에서 공을 굴립니다. 언덕에는 골짜기가 여러 갈래로 패여 있습니다. 공은 아무 데로나 굴러가지 않고 골짜기를 따라 내려갑니다. 그는 이 지형을 후성유전적 지형(Epigenetic Landscape)이라 불렀고, 골짜기 하나하나를 크레오드(Chreode)라 이름 붙였습니다. 크레오드는 그리스어로 '정해진 길'이라는 뜻입니다. 세포는 자유롭게 아무 상태나 되는 것이 아니라, 이미 파여 있는 골짜기를 따라 흘러갑니다 [1]. 70년 전 그림이지만 지금도 이 분야의 모든 모델이 이 그림 위에서 이야기를 시작합니다.

지형을 데이터에서 되찾으려는 첫 시도는 RNA 벨로시티(RNA Velocity)였습니다. 갓 만들어진 mRNA에는 인트론이 남아 있고, 성숙한 mRNA에는 인트론이 잘려 나가 있습니다. 두 종류의 비율을 재면 그 유전자의 발현이 지금 오르는 중인지 내리는 중인지 알 수 있습니다. 사진 한 장에서 화살표를 뽑아내는 기술입니다 [2]. 2020년의 scVelo는 정상 상태 가정을 걷어내고 동역학 모형을 붙여 이 방법을 넓혔습니다 [3].

문제는 화살표가 늘 같은 곳을 가리키지 않는다는 데 있습니다. 2026년 6월 PLOS Computational Biology에 실린 비교 연구는 다섯 가지 벨로시티 방법을 세 개

데이터셋에 돌려 보고, 국소 일관성과 방법 간 일치도, 시퀀싱 깊이에 대한 안정성을 따졌습니다. 같은 데이터에 다른 알고리즘을 넣으면 다른 답이 나왔습니다 [4]. 2026년 3월 Cell Reports Methods에 실린 더 큰 벤치마크는 열다섯 가지 방법을 열일곱 개 독립 연구 데이터에 적용했습니다. 정확도와 안정성, 사용성 세 축을 모두 앞선 방법은 하나도 없었습니다 [5]. 사람 태아 전뇌 데이터에서는 두 방법이 같은 세포 집단에 정반대 방향의 흐름을 그리기도 했습니다. 화살표는 인트론 비율이라는 얇은 근거 하나에 기대고 있고, 그 근거는 유전자마다 다르게 무너집니다.

그래서 최근의 흐름은 화살표 대신 수송으로 옮겨 갔습니다. 6시간째 사진의 세포 무리를 24시간째 사진의 세포 무리로 옮기려면 어떻게 옮기는 것이 비용이 적을까. 이것이 최적 수송(Optimal Transport)입니다. 이삿짐을 옮길 때 제일 가까운 창고부터 채우는 계산과 같습니다. 개별 세포를 일대일로 추적하지 않고 무리 전체를 무리 전체로 옮깁니다. CellStream은 동적 최적 수송을 임베딩 학습에 넣어, 띄엄띄엄한 스냅샷에서 세포 궤적을 복원했습니다 [6].

벡터장이라는 말이 여기서 나옵니다. 고차원 전사체 공간의 점 하나가 세포 하나입니다. 그 점마다 화살표를 하나씩 붙여 놓은 것이 벡터장입니다. 어느 점에 놓이든 그 자리의 화살표를 따라가면 다음 상태가 나옵니다. 신경망으로 이 화살표 함수를 배우고 미분방정식 풀이기에 넘기는 방식을 뉴럴 ODE(Neural Ordinary Differential Equations)라 부릅니다. 층을 하나씩 쌓는 대신 시간을 연속으로 다루는 신경망입니다. 세포 궤적은 이 틀에 잘 맞습니다. 세포는 층을 건너뛰며 변하지 않고, 시간을 따라 조금씩 변하기 때문입니다.

여기까지 오면 계산량이라는 벽이 나옵니다. 궤적을 미분방정식으로 풀면 시작점에서 도착점까지 시간을 잘게 쪼개 여러 번 적분해야 합니다. 데이터셋이 바뀌면 처음부터 다시 풉니다. 밤에 잡을 걸어 두고 아침에 출근해 로그를 확인하는 생활이 여기서 나옵니다. 2026년 5월 27일 공개된 Chreode는 이 순서를 뒤집었습니다. 이름 그대로 워딩턴의 골짜기를 모델 구조로 옮긴 세포 세계 모델(Cell World Model)입니다. 무거운 계산을 학습 시간으로 밀어 넣고, 추론할 때는 한 번의 전이 연산으로 끝냅니다. 전이 연산자는 세 갈래로 쪼개져 있습니다. 골짜기를 따라 내려가는 흐름, 접평면 위에서 도는 회전 운동, 그리고 퍼져 나가는 확률적 흩어짐입니다 [7]. 워딩턴의 그림에 있던 경사와 흔들림이 그대로 항으로 남아 있습니다.

숫자를 보겠습니다. Chreode는 생쥐 배아 아틀라스 240만 세포, 데이터셋 일곱 개로 사전 학습했습니다. Weinreb 조혈 데이터와 Veres 췌도 분화 데이터에서 처음부터 학습한 모델과 PI-SDE, PRESCIENT보다 나은 싱크혼 거리를 얻었습니다. 눈여겨볼 대목은 옮겨 심기입니다. 발달 궤적으로 배운 표현을 섭동 예측 모델 GEARS에 유전자 상태 임베딩으로 넣자, Norman Perturb-seq 데이터의 상위 20개 차등 발현 유전자 오차가 0.2121에서 0.1858로 내려갔습니다. 12.4퍼센트 개선입니다 [7]. 분화를 배운 모델이 CRISPR 유전자 넉아웃 반응도 조금 더 잘 맞혔다는 뜻입니다. 발달과 섭동은 결국 같은 지형 위의 이동이라는 가설에 힘이 실립니다. 사전 학습한 뼈대는 Weinreb 데이터의 클론 운명 점수를 한 번도 배우지 않고 맞히는 시험에서도 동적 최적 수송 기반 방법들과 나란히 섰습니다 [7].

데이터 쪽도 규모가 달라졌습니다. Tahoe-100M은 암세포주 50종에 약물 1,100여 종을 걸어 6만 건 가까운 섭동 실험을 돌리고 1억 개 세포를 읽었습니다 [8]. 2026년 1월 Tahoe와 Arc Institute, Biohub는 여기에 네 배 넘게 섭동을 늘려 세포 1억 2천만 개, 섭동 22만 5천 건 규모의 데이터를 함께 만들겠다고 발표했습니다 [8]. 이 데이터를 감당하려는 모델도 나왔습니다. SCALE은 섭동 예측을 조건부 수송 문제로 다시 쓰고 사전 학습을 12.51배 빠르게 돌렸으며, Tahoe-100M 평가에서 이전 최고 성능보다 PDCorr을 12.02퍼센트, 차등 발현 유전자 중첩을 10.66퍼센트 끌어올렸습니다 [9].

세계 모델(World Model)이라는 이름은 로봇과 게임 인공지능에서 건너왔습니다. 로봇이 팔을 어느 각도로 움직이면 컵이 어디로 쓰러질지를 머릿속에서 먼저 굴려 보는 모형이 세계 모델입니다. 행동을 넣으면 다음 상태가 나옵니다. 세포에 이 틀을 옮기면 행동 자리에 약물과 유전자 넉아웃이 들어가고, 상태 자리에 전사체가 들어갑니다. 실험 대신 계산으로 결과를 먼저 보는 것입니다.

이런 모델이 실제로 무엇을 해 주는지는 약을 두 개 섞는 자리에서 드러납니다. 암 줄기세포에 후보 물질을 걸었을 때 세포가 사멸 쪽 골짜기로 굴러가는지, 아니면 옆 골짜기로 새어 나가 내성 상태에 자리를 잡는지를 컴퓨터 안에서 먼저 굴려 봅니다. 실험실에서 조합을 다 시험하려면 후보가 100개일 때 짝만 4,950가지입니다. 예측이 절반만 맞아도 실험 순서를 바꿀 수 있습니다.

그래도 저는 여기서 축배를 들지 않으려 합니다. Arc Institute가 연 첫 가상 세포 챌린지의 결과가 그 이유입니다. 114개국에서 5천 명이 등록하고 1,200개 팀이 제출했지만, 최종 순위표는 예측 모델이 소박한 기준선을 모든 지표에서 앞서지는 못한다는 사실을 그대로 보여 주었습니다. 우승한 팀들은 딥러닝에 고전 통계 특징을 섞어 썼습니다 [10]. 순수한 종단 학습만으로 이 문제가 풀리지 않았다는 뜻입니다. 두 번째 대회는 2026년 8월 20일 시작하며, 상금은 10만 달러이고 평가 범위를 여러 유전자를 동시에 끄는 조합 섭동과 학습에 없던 세포 유형으로 넓혔습니다 [10]. 어려운 쪽으로 시험 문제를 옮긴 것입니다.

이 장면이 저는 마음에 듭니다. 죽은 세포의 사진 다섯 장에서 시간을 되찾아 오는 일은 원래 무리한 요구입니다. 그런데 사람들은 그 무리한 요구를 워딩턴의 언덕이라는 오래된 그림 위에 얹어 놓고, 240만 개 세포로 골짜기의 모양을 배웠습니다. 지도는 아직 거칠고 화살표는 자주 흔들립니다. 그래도 데이터에 시간 축을 돌려놓은 쪽이 옳습니다. 정지 화면끼리 짝지어 놓고 그것을 생명이라 부르는 것보다는 낫습니다.

**참고자료**


[1] Waddington, C. H. (1957). The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology. George Allen & Unwin.

[2] La Manno, G., Soldatov, R., Zeisel, A., et al. (2018). RNA velocity of single cells. Nature, 560(7719), 494-498. https://doi.org/10.1038/s41586-018-0414-6

[3] Bergen, V., Lange, M., Peidli, S., Wolf, F. A., & Theis, F. J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. Nature Biotechnology, 38(12), 1408-1414. https://doi.org/10.1038/s41587-020-0591-3

[4] Ancheta, S., et al. (2026). Challenges and progress in RNA velocity: Comparative analysis across multiple biological contexts. PLOS Computational Biology, 22(6), e1014303. https://doi.org/10.1371/journal.pcbi.1014303

[5] Luo, Y., et al. (2025). Benchmarking RNA velocity methods across 17 independent studies. Cell Reports Methods. https://www.cell.com/cell-reports-methods/fulltext/S2667-2375(26)00067-6

[6] Ling, Y., Zhang, P., Zhang, Z., & Zhou, P. (2025). CellStream: Dynamical Optimal Transport Informed Embeddings for Reconstructing Cellular Trajectories from Snapshots Data. arXiv preprint arXiv:2511.13786. https://arxiv.org/abs/2511.13786

[7] Qiu, M., Zheng, G., Xu, Y., Zhang, R., Ding, Y., Long, Q., & Chen, T. (2026). Chreode: A Cell World Model for One-Step Temporal Dynamics and Perturbation Prediction. arXiv preprint arXiv:2605.28111. https://arxiv.org/abs/2605.28111

[8] Arc Institute. (2026년 1월 20일). Tahoe Therapeutics, Arc Institute, and Biohub Partner to Generate the Largest Perturbation Dataset for Virtual Cell Models. https://arcinstitute.org/news/tahoe-arc-biohub

[9] Chen, S., Yu, L., Jin, K., et al. (2026). SCALE: Scalable Conditional Atlas-Level Endpoint transport for virtual cell perturbation prediction. arXiv preprint arXiv:2603.17380. https://arxiv.org/abs/2603.17380

[10] Arc Institute. (2025년 12월 9일). Virtual Cell Challenge 2025 Wrap-Up: Winners and Reflections. https://arcinstitute.org/news/virtual-cell-challenge-2025-wrap-up

## 10.2 수만 개의 유전자 간 동적 경쟁 상호작용을 계산하는 시그모이드 어텐션(Sigmoid Attention)

학부 3학년 수업에서 어텐션 히트맵을 처음 띄웠던 날을 기억합니다. 화면에는 문장 하나가 가로로도 세로로도 놓여 있었고, 칸마다 밝기가 달랐습니다. 'it'이라는 단어의 행을 짚자 밝은 칸은 딱 하나였습니다. 앞에 나온 명사 하나만 환하게 빛나고 나머지는 전부 어두웠습니다. 조교가 말했습니다. 이 행의 숫자를 다 더하면 1이 됩니다. 저는 그 문장이 예쁘다고 생각했습니다. 그때는 그것이 나중에 문제가 될 줄 몰랐습니다.

행의 합이 1이 되도록 만드는 함수가 소프트맥스(Softmax)입니다. 각 값에 지수를 취한 뒤 전체 합으로 나누어 확률처럼 만드는 계산입니다. 2017년 트랜스포머 논문이 어텐션에 이 함수를 넣은 이유는 분명합니다 [1]. 언어에서 대명사는 하나만 가리켜야 합니다. 'it'이 고양이를 가리키면 다른 명사는 가리키지 않습니다. 한 표를 한 후보에게만 주는 반장 선거와 같습니다. 어느 한쪽이 표를 더 가져가면 다른 쪽은 반드시 잃습니다. 제로섬입니다.

세포 안에서는 이 규칙이 통하지 않습니다. 전사 인자 p53 하나가 켜지면 하위 표적 유전자 수백 개가 함께 켜집니다. MYC도 그렇습니다. 유전자 A가 켜졌다고 해서 유전자 B가 꺼져야 할 이유가 없습니다. 세포는 표를 나눠 갖는 선거장이 아니라, 스위치 여러 개가 동시에 올라가는 배전반에 가깝습니다. 그런데 어텐션에 소프트맥스를 쓰면 모델은 배전반을 선거장으로 착각합니다. 한 유전자에 높은 가중치를 주려면 다른 유전자의 몫을 깎아야 합니다. 실제로는 나란히 켜지는 조절 관계가 모델 안에서 서로를 밀어냅니다.

숫자로 보면 압박이 뚜렷합니다. 사람의 단백질 코딩 유전자는 약 2만 개입니다. 한 행의 합이 1이라면 칸 하나에 돌아가는 평균 몫은 2만 분의 1, 곧 0.00005입니다. p53처럼 표적이 300개인 조절자가 그 관계를 모두 살리려면 한 칸당 0.003쯤을 나눠 가져야 하고, 그러면 정작 중요한 몇 개도 함께 흐려집니다. 반대로 몇 개만 밝히면 나머지 관계는 0에 가깝게 눌립니다. 실제로 존재하는 조절 관계가 계산 규칙 때문에 사라지는 일이 벌어집니다.

시그모이드 어텐션은 이 나눗셈을 없앱니다. 질의 벡터와 키 벡터를 곱해 차원의 제곱근으로 나눈 값에, 유전자 쌍마다 따로 시그모이드 함수를 씌웁니다. 각 칸은 0과 1 사이에 놓이지만 행의 합은 1일 필요가 없습니다. 어떤 유전자는 300개 표적과 동시에

강하게 이어지고, 어떤 유전자는 아무와도 이어지지 않습니다. 옆 칸의 사정이 내 칸의 값을 바꾸지 않습니다. 결합의 세기가 그대로 숫자에 남습니다.

애플 연구진은 2024년에 이 교체가 언어 모델에서도 성립한다는 것을 보였습니다. 시그모이드 어텐션을 쓴 트랜스포머 역시 보편 함수 근사기이며, 학습 초반의 큰 어텐션 노름만 잡아 주면 소프트맥스와 성능이 같았습니다. 이들이 만든 FlashSigmoid 커널은 H100 GPU에서 FlashAttention-2보다 추론이 17퍼센트 빨랐습니다 [2]. 이 결과가 나오기 전까지 시그모이드 어텐션은 몇 번이나 시도되었다가 학습이 무너져 접힌 방법이었습니다.

생물학 쪽 결론은 2026년 4월 29일에 나왔습니다. 단일 세포 파운데이션 모델의 소프트맥스를 시그모이드로 그대로 갈아 끼운 연구입니다. 서로 다른 여섯 개 단일 세포 데이터셋에서 세포 유형 분리도가 25퍼센트 올라갔고, 학습 속도는 소프트맥스 대비 최대 10퍼센트 빨라졌습니다 [3]. 이유는 미분에 있습니다. 시그모이드의 도함수는 어디서나 0.25를 넘지 않습니다. 야코비 행렬도 대각 구조를 지킵니다. 소프트맥스는 한 칸을 건드리면 같은 행의 모든 칸이 함께 흔들리는 조밀한 결합 구조를 가집니다. 1억 6천만 매개변수 모델로 스트레스 시험을 돌렸을 때 소프트맥스 쪽은 기울기가 1만 배 폭발하며 학습이 무너졌고, 시그모이드 쪽은 끝까지 버텼습니다. 같은 연구진이 만든 TritonSigmoid 커널은 H100에서 515 TFLOPS를 냈습니다 [3].

규모를 가늠해 보겠습니다. 유전자 하나를 토큰 하나로 놓으면 어텐션 행렬은 2만 곱하기 2만, 곧 4억 칸이 됩니다. 이 행렬을 통째로 메모리에 올리면 GPU가 버티지 못합니다. FlashAttention은 행렬을 타일로 쪼개 고속 메모리 안에서만 계산하고 중간 결과를 저장하지 않는 방식으로 이 벽을 넘었습니다 [4]. 시그모이드 어텐션이 실전에 들어온 것은 이런 커널 기술이 먼저 자리를 잡았기 때문입니다. 수식을 바꾸는 일과 그 수식을 하드웨어에서 굴리는 일은 따로 놀지 않습니다.

이 차이가 실제 결과물에서 어떻게 보이는지 이야기해 보겠습니다. 유전자 조절 네트워크(Gene Regulatory Network)를 추론하면 화면에 표 하나가 뜹니다. 왼쪽에 조절자, 오른쪽에 표적, 그 옆에 점수가 붙은 목록입니다. 소프트맥스로 학습한 모델에서 이 목록을 뽑으면 상위 몇 줄은 그럴듯한데 그 아래로 점수가 절벽처럼 떨어집니다. 실험으로 확인된 관계가 200번째 줄에 가 있는 일이 생깁니다. 검증된 관계를 놓치는 오류를 허위

음성(False Negative)이라 부릅니다. 시그모이드로 바꾸면 목록의 점수 분포가 완만해집니다. 강한 관계가 여러 개 있으면 여러 개가 함께 높은 점수를 받습니다. 목록을 눈으로 훑는 사람에게는 이 차이가 상당히 큽니다.

여기서 한 가지를 분명히 하고 싶습니다. 어텐션 가중치가 곧 유전자 조절 관계는 아닙니다. 2026년 3월에 나온 회로 추적 연구는 Geneformer의 5번째 층에서 살아 있는 희소 오토인코더 특징 4,065개를 하나도 빠뜨리지 않고 추적해, 유의한 하류 연결 약 140만 개를 지도로 그렸습니다. 연결은 고르게 퍼져 있지 않았습니다. 전체 특징의 1.8퍼센트가 허브 노릇을 했고, 상위 허브 20개 가운데 8개는 어떤 생물학적 의미도 붙일 수 없었습니다. 특징 세 개를 동시에 지우는 실험에서 중복 비율은 0.59, 두 개일 때는 0.74였고 시너지는 0이었습니다 [5]. 같은 연구자가 Geneformer와 scGPT를 나란히 놓고 본 다른 연구에서는, 희소 오토인코더가 정돈된 생물학 지식은 꺼내 주지만 인과적 조절 논리는 거의 담고 있지 않았습니다 [6]. 어텐션이 밝힌 칸은 함께 발현된다는 통계이지, A가 B를 켠다는 인과가 아닙니다.

경고는 더 있습니다. 2026년 2월 18일 공개된 연구는 매개변수를 하나도 학습하지 않은 표현이 여러 하류 벤치마크에서 단일 세포 파운데이션 모델을 앞섰다고 보고했습니다 [7]. 2025년 Nature Methods에 실린 벤치마크는 GEARS와 scGPT, scFoundation을 얄팍한 선형 모델과 나란히 세웠는데, 단일 유전자 섭동에서도 이중 섭동에서도 딥러닝 모델은 선형 기준선을 넘지 못했습니다 [8]. 왜 그런지에 대한 답 하나는 2026년 5월 ICML에 실린 연구가 내놓았습니다. 유전자 발현 데이터의 대부분은 섭동과 무관한 신호이고, 섭동이 남기는 흔적은 아주 희박합니다. 큰 신호에 파묻힌 작은 신호를 따로 떼어내지 못하면 모델 크기를 키워도 소용이 없습니다 [9].

수식을 갈아 끼우는 시도는 시그모이드 하나로 끝나지 않았습니다. 2025년 12월에 나온 SpikGPT는 스파이킹 신경망의 발화 방식을 어텐션에 붙여 세포 유형 주석 문제를 풀었습니다 [10]. 어떤 연구는 어텐션을 아예 걷어내고 그래프 위의 시간 변화를 직접 모델링합니다. 공통점이 하나 있습니다. 자연어의 문법을 그대로 세포에 씌우지 않겠다는 태도입니다.

제가 이 대목을 좋아하는 까닭은 따로 있습니다. 소프트맥스를 시그모이드로 바꾸는 일은 코드 몇 줄입니다. 함수 이름 하나를 고치면 끝납니다. 그런데 그 몇 줄을 고치기 위해

누군가는 유전자 조절의 물리를 다시 들여다봐야 했고, 누군가는 H100 커널을 새로 짜야 했으며, 누군가는 1억 6천만 매개변수 모델을 일부러 망가뜨려 가며 기울기를 관찰해야 했습니다. 짧은 수정 뒤에 긴 이해가 있습니다.

그럼에도 저는 시그모이드 쪽 손을 들어 줍니다. 이유는 성능표가 아닙니다. 소프트맥스는 언어를 위해 만들어진 수식이고, 언어에는 그 수식이 맞았습니다. 세포에는 맞지 않습니다. 지금까지 우리는 자연어 처리에서 잘 돌아간 부품을 떼어다 유전자에 그대로 붙여 왔습니다. 25퍼센트라는 숫자보다 중요한 것은, 이제 사람들이 생물학의 사실에 맞춰 수식을 뜯어고치기 시작했다는 점입니다. 배전반을 배전반으로 다루는 일. 그게 시작입니다.

**참고자료**


[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. NeurIPS 2017. arXiv preprint arXiv:1706.03762. https://arxiv.org/abs/1706.03762

[2] Ramapuram, J., Danieli, F., Dhekane, E., et al. (2024). Theory, Analysis, and Best Practices for Sigmoid Self-Attention. arXiv preprint arXiv:2409.04431. https://arxiv.org/abs/2409.04431

[3] Sadashivaiah, V., Dasoulas, G., Mueller, J., & Ghosh, S. (2026). Better Models, Faster Training: Sigmoid Attention for single-cell Foundation Models. arXiv preprint arXiv:2604.27124. https://arxiv.org/abs/2604.27124

[4] Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022). FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. arXiv preprint arXiv:2205.14135. https://arxiv.org/abs/2205.14135

[5] Kendiukhov, I. (2026). Exhaustive Circuit Mapping of a Single-Cell Foundation Model Reveals Massive Redundancy, Heavy-Tailed Hub Architecture, and Layer-Dependent Differentiation Control. arXiv preprint arXiv:2603.11940. https://arxiv.org/abs/2603.11940

[6] Kendiukhov, I. (2026). Sparse autoencoders reveal organized biological knowledge but minimal regulatory logic in single-cell foundation models: a comparative atlas of Geneformer and scGPT. arXiv preprint arXiv:2603.02952. https://arxiv.org/abs/2603.02952

[7] Souza, H., & Mehta, P. (2026). Parameter-free representations outperform single-cell foundation models on downstream benchmarks. arXiv preprint arXiv:2602.16696. https://arxiv.org/abs/2602.16696

[8] Ahlmann-Eltze, C., Huber, W., & Anders, S. (2025). Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. Nature Methods. https://doi.org/10.1038/s41592-025-02772-6

[9] Jiang, W., Liu, Y., Cai, Y., Gao, E., Dong, J., Abbasnejad, E., Yao, L., & Shi, J. Q. (2026). What Makes a Representation Good for Single-Cell Perturbation Prediction? ICML 2026. arXiv preprint arXiv:2605.19343. https://arxiv.org/abs/2605.19343

[10] Huang, M., & Kamaleswaran, R. (2025). SpikGPT: A High-Accuracy and Interpretable Spiking Attention Framework for Single-Cell Annotation. arXiv preprint arXiv:2512.03286. https://arxiv.org/abs/2512.03286

# 제11장: AI 에이전트 구동형 합성생물학 및 유전자 회로 설계

## 11.1 유전체 및 조절 설계 자동화를 위한 대규모 생물학 에이전트 구동

새벽 두 시의 실험실에는 형광등 소리만 남습니다. 책상 위에 노트 한 권이 펼쳐져 있고, 대학원생은 프라이머 길이를 세고 있습니다. 어닐링 온도를 손으로 계산하고, 깁슨 어셈블리 반응액의 비율을 다시 적고, 지웠다가 또 적습니다. 다음 날 아침 실험이 실패하면 원인은 대개 이 노트 어딘가에 숨어 있습니다. 숫자 하나가 2도 틀렸을 뿐인데 사흘이 사라집니다.

합성생물학(Synthetic Biology)은 생명체를 부품으로 나누어 다시 조립하는 공학입니다. 대장균이나 효모의 유전자를 규격화된 조각으로 쪼개고, 그 조각을 이어 붙여 사람이 원하는 기능을 넣습니다. 석유에서 뽑던 플라스틱 원료를 미생물이 만들게 하는 일, 암세포의 표지 단백질을 만났을 때만 면역 물질을 뿜는 세포를 만드는 일이 여기에 속합니다. 조각의 규격을 처음 문서로 정한 것이 바이오브릭(BioBrick) 표준입니다 [1]. 모든 부품의 앞뒤에 똑같은 절단 자리를 붙여 두면, 어느 조각을 어느 조각에 이어도 같은 효소로 자르고 붙일 수 있습니다. 콘센트 모양을 나라마다 통일해 두면 어느 가전이든 꽂히는 것과 같은 발상입니다.

이 규격을 학생들이 실제로 쓰게 만든 무대가 iGEM 대회입니다. 2026년 대회에는 전 세계에서 약 300개 팀이 등록했고, 참가 마을 가운데 하나가 소프트웨어와 인공지능입니다 [2]. 팀들은 새로 만든 부품을 표준 생물학 부품 등록소(Registry of Standard Biological Parts)에 제출합니다. 11월 13일부터 16일까지 파리에서 열리는 그랜드 잼버리가 그 결과를 모으는 자리입니다. 등록소는 도서관과 닮았습니다. 누군가 반납한 책이 다음 사람의 출발점이 됩니다. 부품이 쌓일수록 새로 들어온 학부생이 처음부터 만들어야 할 것이 줄어듭니다.

생성 인공지능이 이 도서관에 새 서가를 붙였습니다. 2026년 초 셀 시스템스에 실린 검토 논문은 지금을 변곡점으로 부릅니다 [3]. 부품과 회로와 유전체를 데이터로부터 새로 만들어 내는 일이 실험실 밖에서 먼저 끝난다는 뜻입니다. 서열 하나를 설계할 때 고려할 것은 많습니다. 숙주 세포가 어떤 코돈을 잘 읽는지에 맞추어 서열을 다시 쓰는 코돈 최적화(Codon Optimization)가 있습니다. 같은 아미노산을 만드는 세 글자 조합이 여러 개인데, 세포마다 잘 읽는 조합이 다르기 때문입니다. mRNA가 스스로 접혀서 리보솜이 올라타지 못하게 막는 자리를 피해야 하고, 번역 시작 신호와 종결 신호, 폴리아데닐화

신호까지 제자리에 놓아야 합니다.

연구자가 에이전트에게 던지는 말은 이제 이런 모양입니다. 햄스터 난소 세포에서 치료용 항체의 무거운 사슬과 가벼운 사슬이 1대 2 비율로 만들어지도록 프로모터 회로를 짜고, 검증 절차서까지 붙여 달라는 요청입니다. 두 사슬의 비율이 어긋나면 항체가 제대로 접히지 않고 세포 안에 찌꺼기가 쌓입니다. 예전에는 이 비율을 맞추려고 프로모터 세기를 바꿔 가며 몇 달을 돌렸습니다. 에이전트는 서열 최적화와 부품 배치와 실행 코드를 한 묶음으로 내놓습니다. 사람이 할 일은 그 묶음이 틀렸는지 따지는 쪽으로 옮겨 갑니다.

문제는 설계도가 아니라 손입니다. 회로를 그리는 데는 하루가 걸리지만, 그것을 세포 안에 넣는 절차서에는 수백 개의 숫자가 들어갑니다. 피펫으로 몇 마이크로리터를 옮길지, 써모사이클러를 몇 도에서 몇 초 유지할지, 원심분리기를 몇 회전으로 돌릴지, 항생제를 어느 농도로 깔지가 전부 적혀 있어야 합니다. 예전에는 이 숫자를 논문 부록과 시약 회사 매뉴얼에서 하나씩 긁어모았습니다. 대규모 언어 모델 에이전트가 이 자리를 파고들었습니다. 사람이 문장으로 목표를 말하면, 에이전트가 절차서를 쓰고 그것을 로봇이 읽는 코드로 바꿉니다.

2026년 6월에 공개된 프로토파일럿(ProtoPilot)은 이 흐름을 숫자로 보여 줍니다 [4]. 연구진은 검증된 표준 프로토콜 98건에서 합성생물학과 분자생물학 과제 294개를 뽑아 시험대를 만들었습니다. 에이전트 하나가 다 하는 구조가 아닙니다. 실험 의도를 읽는 쪽, 수치와 장비 제약을 맞추는 쪽, 코드를 쓰는 쪽이 나뉘어 있고, 성공한 절차는 기술 서고에 쌓여 다음 과제에 다시 불려 나옵니다. 결과를 보면 에이전트가 만든 절차서가 전문가 평가에서 상위 세 개 안에 들어간 비율이 90.2퍼센트였습니다. 절차서를 실행 코드로 옮긴 뒤 검증을 통과한 비율은 89.5퍼센트, 오픈트론스 장비에서 그대로 돌아간 비율은 88.24퍼센트였습니다. 같은 조건에서 기존 오픈트론스 인공지능 도구는 32.35퍼센트에 그쳤습니다. 격차를 만든 것은 문장력이 아니라 장비 단위의 검증 관문이었습니다. 연구진은 여기서 멈추지 않고 실제 실험대에서 돌렸습니다. 에이전트가 짠 조립 반응의 산물을 생어 시퀀싱으로 읽어 서열이 맞는지 확인했습니다.

로봇 팔 쪽도 따라오고 있습니다. 2026년 6월 발표된 파이펫(Pipette) 시험대는 시료 다루기, 배양 용기 옮기기, 장비 조작, 정밀 배치까지 12개 과제로 이루어져 있습니다 [5]. 과제마다 사람의 시범을 30번만 보여 준 조건에서 한 모델은 평균 성공률 60.3퍼센트를

기록했고, 시뮬레이션으로 데이터를 늘리자 다른 모델은 40.4퍼센트에서 71.8퍼센트로 올라갔습니다. 사람이 손목으로 익히는 감각을 기계에 옮기는 데 시범 30번이면 충분하지 않습니다. 부족한 부분을 가상 실험실이 메웁니다. 장비 업계도 움직였습니다. 오픈트론스는 2026년 2월 보스턴에서 열린 학회에서 자연어 지시로 qPCR 작업 흐름을 짜는 시연을 했고, 4월에 배포한 앱 9.0에 인공지능이 만든 프로토콜을 미리 돌려 보는 기능을 넣었습니다 [6]. 코드가 로봇을 움직이기 전에 화면에서 먼저 움직여 봅니다.

절차서를 기계가 읽으려면 형식부터 정리해야 합니다. 논문에 적힌 실험 방법은 본문과 표와 그림과 부록에 흩어져 있고, 사람은 그것을 눈으로 이어 붙여 읽습니다. 기계는 그러지 못합니다. 그래서 실험 절차를 정해진 항목으로 나누어 적는 구조화 규격을 만드는 작업이 함께 굴러갑니다. 시약 이름, 농도, 시간, 온도, 장비 모델이 각자 제자리를 갖는 순간 절차서는 문장이 아니라 데이터가 됩니다.

여기서 배울 점은 프롬프트가 전부가 아니라는 사실입니다. 로제타 에이전트(Agent Rosetta) 연구진은 단백질 설계 소프트웨어에 언어 모델을 붙이면서, 문장만 잘 써서는 올바른 명령을 만들지 못한다는 것을 확인했습니다 [7]. 도구와 모델 사이에 제대로 된 환경을 짜 넣어야 전문가 수준에 닿았습니다. 명령어 하나를 잘못 쓰면 프로그램이 조용히 엉뚱한 결과를 뱉기 때문입니다. 항체 설계 에이전트 레이턴트-Y는 표적 아홉 개 가운데 67퍼센트에서 성공했고, 실험실에서 확인한 결합력은 한 자릿수 나노몰까지 내려갔습니다. 전문가가 이 에이전트를 쓰면 설계 캠페인이 56배 빨라졌습니다 [8]. 몇 주가 몇 시간이 됩니다.

저는 이 대목에서 편을 듭니다. 에이전트의 값어치는 글 솜씨가 아니라 자기 결과를 스스로 반려하는 관문에 있습니다. 그럴듯한 절차서는 아무 모델이나 씁니다. 그 절차서를 장비 규격에 넣어 보고, 안 되면 고치고, 실험 결과를 받아 다시 고치는 고리가 있어야 새벽 두 시의 노트가 사라집니다. 설계와 제작과 시험과 학습이 한 바퀴로 이어지는 순간, 실험실은 도구 창고가 아니라 스스로 굴러가는 기계가 됩니다.

경계도 같이 커졌습니다. 2026년 6월 공개된 ABC-Bench는 액체 취급 로봇 코딩, DNA 조립 설계, DNA 합성 심사 우회라는 세 과제에서 시험한 모든 언어 모델 에이전트가 사람 전문가의 중앙값을 넘었다고 보고했습니다 [9]. 세 번째 과제가 마음에 걸립니다. 실력이 올라간 만큼 위험한 서열을 걸러 내는 문턱도 낮아졌다는 뜻이기 때문입니다. 시장은 이미

그 속도에 반응하고 있습니다. 생물 파운드리 서비스 시장은 2025년 16억 5천만 달러에서 2026년 20억 1천만 달러로 늘었고, 연 21.3퍼센트로 자라고 있습니다 [10]. 이제 실험 기록은 지우개 자국이 아니라 서버의 로그 파일에 남습니다. 어느 쪽이 더 정직한지는, 실패한 실험을 되짚어 본 사람이면 압니다.

### 참고자료


[1] Shetty, R. P., Endy, D., & Knight, T. F. (2008). Engineering BioBrick vectors from BioBrick parts. Journal of Biological Engineering, 2, 5. https://doi.org/10.1186/1754-1611-2-5

[2] iGEM Foundation. (2026). About the iGEM Competition. https://competition.igem.org/about

[3] Kim, N., Carluccio, G. D., Zhang, K., & Collins, J. J. (2026). Generative AI for synthetic biology: Designing biological parts, circuits, and genomes. Cell Systems, 17, 101533. https://doi.org/10.1016/j.cels.2026.101533

[4] Jiang, Y., et al. (2026). A Self-Evolving Agentic System for Automated Generation and Execution of Biological Protocols. arXiv preprint arXiv:2606.31763. https://arxiv.org/abs/2606.31763

[5] Liu, Z., Jin, H., Du, Z., et al. (2026). Pipette: An Embodied Simulation Platform, Benchmark, and Data-Efficient Augmentation Framework for Wet-Lab Robotics. arXiv preprint arXiv:2606.12936. https://arxiv.org/abs/2606.12936

[6] The Robot Report. (2026년 4월). Opentrons introduces dynamic simulation, visualization for AI-generated lab workflows. https://www.therobotreport.com/opentrons-introduces-dynamic-simulation-visualization-ai-generated-lab-workflows/

[7] Teneggi, J., Turzo, S. M. B. A., Marwah, T., Bietti, A., Renfrew, P. D., Mulligan, V. K., & Golkar, S. (2026). Protein Design with Agent Rosetta: A Case Study for Specialized Scientific Agents. arXiv preprint arXiv:2603.15952. https://arxiv.org/abs/2603.15952

[8] Schmon, S. M., et al. (2026). Latent-Y: A Lab-Validated Autonomous Agent for De Novo Drug Design. arXiv preprint arXiv:2603.29727. https://arxiv.org/abs/2603.29727

[9] Liu, A. B., Nedungadi, S., Cai, B., Kleinman, A., Bhasin, H., & Donoughe, S. (2026). ABC-Bench: An Agentic Bio-Capabilities Benchmark for Biosecurity. arXiv preprint arXiv:2606.11150. https://arxiv.org/abs/2606.11150

[10] The Business Research Company. (2026). Biofoundry-as-a-Service Global Market Report 2026. https://www.researchandmarkets.com/reports/6241453/biofoundry-as-a-service-global-market-report

## 11.2 자율적 유전자 회로 설계를 위한 계층적 검증 기반 강화 학습(GenCircuit-RL)

컴퓨터 화면에서는 신호가 깔끔한 사각파로 뜁니다. 켜짐과 꺼짐 사이에 망설임이 없습니다. 같은 회로를 대장균에 넣고 현미경 아래 형광을 재면 곡선은 뭉개져 있습니다. 어떤 세포는 켜지고 옆 세포는 꺼져 있습니다. 한 배양접시 안에서 같은 설계도를 받은 세포들이 서로 다른 답을 냅니다. 합성생물학을 하는 사람이라면 이 장면 앞에서 한 번쯤 커피를 식혔습니다.

유전자 회로는 전자 회로의 문법을 세포 안으로 옮긴 것입니다. 2000년에 두 편의 논문이 같은 호에 실리며 이 분야가 열렸습니다. 하나는 두 개의 억제자가 서로를 누르며 켜짐과 꺼짐 두 상태를 기억하는 토글 스위치(Toggle Switch)였고 [1], 다른 하나는 세 개의 억제자가 돌아가며 서로를 눌러 주기적으로 깜빡이는 리프레실레이터(Repressilator)였습니다 [2]. 첫 번째는 기억 소자이고 두 번째는 시계입니다. 세포가 스스로 상태를 붙잡고, 스스로 박자를 셉니다. 2016년에는 첼로(Cello)가 나왔습니다. 회로 도면을 하드웨어 기술 언어로 적으면 소프트웨어가 DNA 서열을 뽑아 주는 도구입니다 [3]. 전사 억제자를 NOT 게이트와 NOR 게이트로 쓰면 어떤 논리식이든 조립됩니다. 반도체 설계자가 쓰던 방식이 그대로 세포로 넘어왔습니다.

그런데 세포는 회로 기판이 아닙니다. 분자들은 열 때문에 쉬지 않고 흔들리고, 유전자가 켜지는 순간도 확률로 정해집니다. 억제자 단백질이 몇 개밖에 없는 상황에서는 한 개가 붙느냐 떨어지느냐로 출력이 뒤집힙니다. 이 요동을 정확히 다루려면 화학 마스터 방정식과 길레스피(Gillespie) 확률 시뮬레이션이 필요합니다 [4]. 분자 하나하나가 언제 반응할지를 주사위를 굴려 따라가는 계산법입니다. 자원 다툼도 있습니다. 리보솜과 RNA 중합효소는 정해진 양뿐인데, 인공 회로가 이것을 많이 가져가면 세포는 자라기를 늦춥니다. 앞단 모듈에 뒷단 모듈을 붙이는 순간 앞단의 동작이 바뀌는 되먹임(Retroactivity)도 생깁니다 [5]. 전자 회로에서는 부품을 이어 붙여도 각자 하던 일을 하지만, 세포 안에서는 이웃이 생기면 성격이 변합니다. 지하철 환승 통로에 사람이 몰리면 반대편 승강장의 흐름까지 느려지는 것과 비슷합니다.

2026년 5월 14일 공개된 GenCircuit-RL은 이 문제를 코드 생성 문제로 다시 정의했습니다 [6]. 모델은 DNA 서열을 직접 뱉지 않습니다. 파이썬 라이브러리 pysbol3를 부르는 코드를

쓰고, 그 코드가 실행되면서 합성생물학 개방 언어(SBOL) 형식의 회로 설계 파일을 만듭니다. 부품과 부품의 연결이 기계가 읽을 수 있는 형식으로 남는다는 뜻입니다. 이 선택이 왜 중요한지는 검토 단계에서 드러납니다. 서열 뭉치를 받으면 사람은 눈으로 읽지 못하지만, 설계 파일은 어느 프로모터가 어느 유전자를 켜는지 그대로 보여 줍니다. 그림 대신 조립 설명서를 쓰게 시킨 셈입니다.

핵심은 보상을 주는 방식에 있습니다. 이 프레임워크는 옳고 그름을 다섯 계층으로 쪼갭니다. 코드가 오류 없이 실행되는지부터 시작해서, 결과 파일이 규격을 지켰는지, 부품이 제자리에 놓였는지, 그리고 회로의 연결 구조가 과제가 요구한 위상과 맞는지까지 층층이 따져 점수를 나눠 줍니다. 학습은 네 단계 커리큘럼으로 진행합니다. 쉬운 과제부터 시작해 어려운 설계로 올라갑니다. 연구진은 시험대로 SynBio-Reason을 함께 내놓았습니다. 여섯 종류의 표준 회로에서 뽑은 회로 4,753개와, 남이 짜다 만 코드를 고치는 일부터 백지에서 설계하는 일까지 아홉 개 과제로 이루어져 있습니다. 계층 보상을 쓴 모델은 기능 추론 과제에서 맞았는지 틀렸는지만 알려 주는 이진 보상보다 14에서 16퍼센트포인트 앞섰습니다. 커리큘럼을 빼면 성능이 무너졌습니다.

대사 부담은 눈으로도 보입니다. 인공 회로를 잔뜩 넣은 대장균은 배양액에서 흐리게 자랍니다. 회로가 잘 돌아갈수록 세포는 힘들어집니다. 사흘이면 끝날 배양이 닷새로 늘고, 어떤 세포주는 회로를 스스로 잘라 버리고 편해집니다. 그래서 설계 점수에는 논리가 맞는지만이 아니라 세포가 견딜 수 있는지도 들어가야 합니다. 신호 대 잡음비도 같은 자리에 놓입니다. 켜짐 상태의 형광이 꺼짐 상태의 형광보다 몇 배나 높은지, 그 차이가 세포마다 얼마나 벌어지는지를 재야 회로가 쓸 만한지 압니다. 화면에서 논리표를 통과했다는 사실은 출발점이지 도착점이 아닙니다.

숫자보다 이 결과가 말하는 태도가 중요합니다. 채점표가 0점 아니면 100점이면 학생은 어디서 틀렸는지 모릅니다. 부분 점수가 있는 시험지를 받아야 다음번에 고칠 자리를 찾습니다. 강화학습 에이전트도 똑같습니다. 회로가 작동하지 않았다는 한 줄보다, 코드는 돌았고 부품 배치도 맞았는데 연결 위상이 틀렸다는 정보가 학습을 끌고 갑니다. 이 모델은 학습에서 본 적 없는 부품으로 바꿔 시험했을 때도 구조가 맞는 회로를 만들었고, 교과서에 실린 표준 설계를 아무도 알려 주지 않은 상태에서 스스로 다시 찾아냈습니다. 사람이 20년에 걸쳐 쌓은 답을 기계가 되짚어 온 셈입니다.

다른 방향에서도 같은 답이 나옵니다. 2026년 1월에 나온 GenAI-Net은 반응을 제안하는 에이전트와 시뮬레이션 평가를 짝지어 생체분자 네트워크를 만듭니다 [7]. 사람이 원하는 동작을 말로 적으면, 에이전트가 반응식을 하나씩 얹어 보고 시뮬레이터가 점수를 매깁니다. 용량 반응 곡선, 논리 게이트, 진동자, 외부 교란에도 출력이 제자리로 돌아오는 완전 적응 회로까지 결정론과 확률론 양쪽 환경에서 설계했습니다. 같은 기능을 내는 서로 다른 구조를 여러 개 뽑아 준다는 점이 실험자에게 값집니다. 하나가 실험대에서 망가져도 다음 후보가 남기 때문입니다.

조절 DNA 쪽에서는 Ctrl-DNA가 제약 조건을 건 강화학습으로 목표 세포에서만 켜지는 프로모터와 인핸서를 만들었습니다 [8]. 원하는 세포에서 활성을 올리면서 다른 세포로 새어 나가는 발현은 누르는, 두 목표를 동시에 잡는 방식입니다. 만들어진 서열 안에는 그 세포에서만 일하는 전사 인자가 붙는 자리가 실제로 들어 있었습니다. 회로를 밖에서 조종하는 연구도 나란히 갑니다. 빛으로 유전자 발현을 켜고 끄는 광유전학(Optogenetics)은 반응이 너무 가팔라서 중간 세기를 맞추기 어려웠습니다. 2026년 국제 제어 학회에 나온 연구는 빛을 완전히 켜고 끄기를 짧게 반복하는 펄스 폭 변조로 이 문제를 풀었습니다 [9]. 형광등을 흐리게 만드는 대신, 아주 빠르게 깜빡여 평균 밝기를 맞추는 방식입니다.

그래도 시뮬레이터는 세포가 아닙니다. 보상 함수가 아무리 정교해도 그 안의 상수는 사람이 넣은 값입니다. 이 바닥을 실측으로 갈아 끼운 연구가 2026년 1월 14일 네이처에 실렸습니다 [10]. 라이스대학교 연구진은 CLASSIC이라는 방법으로 수십만 개가 넘는 회로 설계를 한꺼번에 사람 세포에 집어넣었습니다. 긴 읽기 서열 분석으로 어느 세포가 어떤 회로를 받았는지 알아내고, 짧은 읽기 서열 분석으로 그 회로가 얼마나 세게 작동했는지 재서, 설계와 성능을 잇는 한 장의 지도를 만들었습니다. 그 지도로 학습한 모델이 시험해 본 적 없는 회로의 성능을 예측했고, 연구진이 40개를 골라 직접 만들어 확인하자 40개 모두 예측과 맞았습니다.

이 지도가 생기면서 설계자의 하루가 바뀝니다. 예전에는 후보를 서른 개쯤 고르고, 클로닝에 두 주를 쓰고, 결과를 보고 다시 서른 개를 골랐습니다. 이제는 모델에게 후보 수만 개의 성능을 먼저 물어보고, 그중에서 실험대에 올릴 몇 개를 고릅니다. 실험이 사라지지는 않습니다. 실험이 놓이는 자리가 앞에서 뒤로 옮겨 갔을 뿐입니다. 예측을 확인하는 실험은 예측을 만드는 실험보다 훨씬 적게 듭니다.

편을 들자면 이쪽입니다. 유전자 회로 설계에서 인공지능이 가져온 진짜 변화는 더 똑똑한 생성 모델이 아니라, 틀린 곳을 층층이 짚어 주는 채점표와 그 채점표를 실측 데이터로 다시 쓰는 습관입니다. 화면 속 사각파와 배양접시 속 뭉개진 곡선 사이의 거리는 아직 남아 있습니다. 다만 그 거리를 재는 자가 생겼습니다. 자가 있으면 줄일 수 있습니다.

### 참고자료


[1] Gardner, T. S., Cantor, C. R., & Collins, J. J. (2000). Construction of a genetic toggle switch in Escherichia coli. Nature, 403(6767), 339-342. https://doi.org/10.1038/35002131

[2] Elowitz, M. B., & Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. Nature, 403(6767), 335-338. https://doi.org/10.1038/35002125

[3] Nielsen, A. A. K., Der, B. S., Shin, J., et al. (2016). Genetic circuit design automation. Science, 352(6281), aac7341. https://doi.org/10.1126/science.aac7341

[4] Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry, 81(25), 2340-2361. https://doi.org/10.1021/j100540a008

[5] Del Vecchio, D., Ninfa, A. J., & Sontag, E. D. (2008). Modular cell biology: retroactivity and insulation. Molecular Systems Biology, 4, 161. https://doi.org/10.1038/msb4100204

[6] Flynn, N. (2026). GenCircuit-RL: Reinforcement Learning from Hierarchical Verification for Genetic Circuit Design. arXiv preprint arXiv:2605.14215. https://arxiv.org/abs/2605.14215

[7] Filo, M., Rossi, N., Fang, Z., & Khammash, M. (2026). GenAI-Net: A Generative AI Framework for Automated Biomolecular Network Design. arXiv preprint arXiv:2601.17582. https://arxiv.org/abs/2601.17582

[8] Chen, X., Ma, S., Lin, R., Lin, J., & Wang, B. (2025). Ctrl-DNA: Controllable Cell-Type-Specific Regulatory DNA Design via Constrained RL. arXiv preprint arXiv:2505.20578. https://arxiv.org/abs/2505.20578

[9] Espinel-Ríos, S. (2026). Switching-time bioprocess control with pulse-width-modulated optogenetics. arXiv preprint arXiv:2511.22893. https://arxiv.org/abs/2511.22893

[10] Rai, K., O'Connell, R., et al. (2026). Ultra-high-throughput mapping of genetic design space. Nature, 650, 1035-1044. https://doi.org/10.1038/s41586-025-09933-9

# 제12장: 미래형 자율 로보틱 실험실 (Self-Driving Lab)

## 12.1 액체 핸들링 로봇 자동화 스크립트 도출 및 PCR/실험 실행 자동화 최적화 (AutoPCR)

새벽 두 시의 실험실을 떠올려 봅니다. 대학원생 한 명이 384웰 플레이트 앞에 앉아 있습니다. 오른손 엄지는 이미 감각이 무뎠습니다. 피펫 버튼을 누르고, 떼고, 다시 누릅니다. 이 동작을 오늘만 4천 번 넘게 했습니다. 웰 하나에 들어가는 시약은 5마이크로리터입니다. 눈물 한 방울의 십분의 일쯤 되는 양입니다.

문제는 손목이 아니라 숫자에 있었습니다. 아침에 뜬 사람과 밤에 지친 사람이 만든 플레이트는 결과가 달랐습니다. 피펫을 기울인 각도, 팁을 시약에 담근 깊이, 흡입 버튼을 떼는 속도가 조금씩 달랐기 때문입니다. 같은 실험을 같은 사람이 화요일과 금요일에 해도 값이 어긋났습니다. 생명과학의 재현성 위기는 거창한 이론에서 시작하지 않았습니다. 손끝의 떨림에서 시작했습니다.

이 위기의 규모는 2016년에 숫자로 드러났습니다. 네이처가 연구자 1,576명에게 물었습니다. 남이 발표한 실험을 따라 해 보고 실패한 적이 있느냐. 70퍼센트가 넘는 사람이 그렇다고 답했습니다. 자기가 예전에 했던 실험조차 다시 재현하지 못한 적이 있다고 답한 사람이 절반을 넘었습니다 [6]. 논문에 적힌 문장은 대개 이렇습니다. "시약을 잘 섞은 뒤 실온에서 30분 배양했다." 잘 섞는다는 게 몇 번 위아래로 흔드는 것인지, 실온이 22도인지 26도인지는 적혀 있지 않습니다. 그 빈칸을 각 실험실의 손버릇이 채웠습니다.

액체마다 성격도 다릅니다. 물은 순순히 빨려 들어옵니다. 글리세롤은 끈적해서 천천히 올라오고 팁 벽에 달라붙습니다. 계면활성제가 든 용액은 거품을 물고 올라옵니다. 사람은 이 차이를 손으로 배웁니다. 로봇은 배우지 못합니다. 그래서 액체 클래스(Liquid Class)라는 설정을 만듭니다. 시약 종류마다 흡입 속도, 대기 시간, 토출 압력, 팁 끝에 남은 방울을 떨어내는 동작을 미리 정해 두는 표입니다. 이 표가 잘 짜여야 로봇의 정확도가 사람을 앞섭니다. 표가 엉성하면 로봇은 같은 실수를 만 번 반복합니다.

로봇은 지루함을 느끼지 않습니다. 이것이 액체 핸들링 로봇의 첫 번째 미덕입니다. 이 생각 자체는 오래됐습니다. 2009년 영국 애버리스트위스 대학교의 로스 킹 연구팀은 애덤(Adam)이라는 로봇에게 효모 유전자의 기능에 관한 가설을 스스로 세우고, 실험을 설계하고, 결과로 가설을 고르게 했습니다. 사람이 손대지 않은 상태에서 기계가 새로운

과학 지식을 만든 첫 사례로 기록됐습니다 [7]. 그로부터 17년이 지났습니다. 문제는 로봇마다 말이 달랐다는 데 있었습니다. 해밀턴 로봇에 쓴 코드는 테칸 로봇에서 돌지 않았습니다. 실험실을 옮기면 프로토콜을 처음부터 다시 짰습니다. 파이랩로봇(PyLabRobot)이 이 벽을 허물었습니다. 하드웨어에 매이지 않는 파이썬 인터페이스입니다. 해밀턴 STAR, 테칸 EVO, 오픈트론스 OT-2를 같은 명령어 한 벌로 움직입니다 [1]. 데크 배치와 랩웨어를 공통 언어로 적어 두니, 한 실험실에서 검증한 코드를 다른 실험실이 그대로 내려받아 돌릴 수 있게 됐습니다. 프로토콜이 논문 속 문장에서 실행 가능한 파일로 바뀐 순간입니다.

여기서 한 걸음 더 나간 실험이 2025년 말에 나왔습니다. 연구진은 유전자 증폭(PCR)을 온도를 바꾸는 장치의 일이 아니라, 움직임을 제어하는 공정으로 다시 정의했습니다 [2]. 열순환기 안에서 블록 온도를 올렸다 내리는 대신, 온도가 일정하게 유지되는 기름 욕조에 밀봉된 피펫 팁을 담갔다 뺐다 반복합니다. 팁이 깊이 들어가면 변성 온도, 얕게 들어가면 결합 온도입니다. 로봇 팔의 상하 운동이 곧 온도 사이클이 됩니다. 결과는 고성능 상용 열순환기와 견줄 만한 증폭 효율과 시퀀싱 정확도였습니다. 팁을 밀봉했으니 교차 오염도 줄었습니다. 처리량을 늘리려면 열순환기를 더 사지 않고, 로봇 스케줄만 촘촘히 짜면 됩니다.

그렇다면 사람이 쓴 프로토콜을 누가 코드로 옮길까요. 2026년 6월에 공개된 프로토파일럿(ProtoPilot)은 그 일을 여러 인공지능 에이전트에게 나눠 맡겼습니다 [3]. 논문에 적힌 절차를 읽고, 장비가 실제로 할 수 있는 동작인지 따지고, 실행 코드를 쓰고, 실패하면 스스로 고쳐 씁니다. 검증은 말잔치로 끝나지 않았습니다. 표준 프로토콜 98건에서 뽑은 294개 과제로 시험했습니다. 전문가들은 이 시스템이 내놓은 상위 세 개 후보 가운데 90.2퍼센트를 쓸 만하다고 평가했습니다. 코드 검증 관문 통과율은 89.5퍼센트였습니다. 오픈트론스 장비에서 실제로 돌려 본 성공률은 88.24퍼센트로, 기준 모델의 32.35퍼센트를 크게 앞섰습니다. DNA 조립 결과는 시퀀싱으로 확인했습니다.

말과 기계 사이에는 아직 통역이 필요합니다. 2026년 6월 제안된 실험실 에이전트 프로토콜(LAP)은 그 통역 규칙을 표준으로 만들자고 제안합니다 [4]. 모델 컨텍스트 프로토콜(MCP)은 에이전트와 소프트웨어 도구를 잇고, A2A는 에이전트끼리를 잇습니다. 정작 에이전트와 물리 장비를 잇는 규격이 비어 있었습니다. 장비 조작은 되돌릴 수 없습니다. 시약을 쏟으면 끝입니다. LAP은 장비마다 성능과 물리적 한계를 적은 서명된

명세서를 두고, 한 번에 한 작업만 장비를 점유하도록 예약을 걸고, 위험한 조작 앞에서는 사람의 확인 토큰을 요구합니다. 측정값에는 단위와 보정 정보와 불확실성을 함께 붙입니다. 숫자 하나가 그냥 숫자로 떠돌지 않게 하려는 장치입니다.

감시하는 눈도 붙었습니다. 2026년 상용 액체 핸들러에는 카메라가 기본으로 들어갑니다. 웰 안의 액면 높이를 실시간으로 재고, 기포가 생기거나 팁이 막히면 동작을 멈춥니다. 데크에 올린 플레이트가 몇 밀리미터 틀어져 있어도 영상으로 잡아냅니다. 예전에는 이런 어긋남을 다음 날 아침 결과지를 보고서야 알았습니다. 밤새 돌린 플레이트 스무 장이 통째로 쓰레기가 되는 일이 흔했습니다.

로봇이 무결점이라는 말은 아닙니다. 팁이 제대로 안 꽂히고, 그리퍼가 플레이트를 놓치고, 시약통이 예상보다 빨리 비는 일이 실제로 벌어집니다. 자동화의 값어치는 실수를 없애는 데 있지 않습니다. 실수가 언제 어디서 일어났는지 기록이 남는 데 있습니다. 사람은 자기 실수를 대개 기억하지 못합니다. 로봇은 로그를 남깁니다.

그 기록을 서로 읽을 수 있게 만드는 작업도 진행 중입니다. 2026년 7월 공개된 사이스키마(SciSchema.org)는 실험 절차를 구조화된 형식으로 적는 표준 묶음을 내놓았습니다 [8]. 실험 조건은 논문 본문, 표, 그림, 보충자료에 흩어져 있습니다. 사람이 읽어도 헷갈리는데 기계는 더 헤맵니다. 절차를 정해진 칸에 나눠 담으면, 다른 실험실의 로봇이 그 칸을 그대로 읽어 실행할 수 있습니다. 논문이 읽는 물건에서 돌리는 물건으로 바뀝니다.

규모가 커지면 풍경이 달라집니다. 긴코 바이오웍스는 보스턴에 로봇 랙 105대를 연결한 자율 실험실을 세우고 2026년 3월 클라우드 랩 서비스를 열었습니다 [5]. 한 시연에서 이 시스템은 반응 3만 6천 건을 돌려 데이터 15만 점을 만들었습니다. 사람 손으로는 엄두를 낼 수 없는 양입니다. 회사가 내세운 숫자 가운데 눈에 띄는 것은 가동 시간입니다. 사람이 일하는 실험실은 주당 40시간 돌아갑니다. 로봇 실험실은 168시간, 곧 일주일 내내 돌아갑니다. 같은 작업을 하는 데 필요한 공간은 3분의 1로 줄었습니다.

문턱도 낮아졌습니다. 예전에는 로봇 한 대에 수억 원이 들었고, 전담 엔지니어가 붙어야 돌아갔습니다. 2026년 시장에서는 해밀턴 STAR V가 고속 산업용 자리를 지키는 한편, 오픈트론스 플렉스 같은 장비가 작은 실험실의 진입 문턱을 끌어내렸습니다. 대학원생이 파이썬 스무 줄을 써서 자기 프로토콜을 돌립니다. 자동화가 대기업의 특권이던 시절이

끝나가고 있습니다.

새벽 두 시의 대학원생 이야기로 돌아갑니다. 그 학생이 로봇 덕에 잠을 자게 됐다는 결말은 절반만 맞습니다. 로봇을 돌리려면 액체 클래스를 잡고, 데크를 배치하고, 코드를 짜고, 실패 로그를 읽어야 합니다. 일이 사라진 게 아니라 자리를 옮겼습니다. 엄지로 하던 일을 이제 머리로 합니다.

로봇이 사람보다 똑똑해서가 아닙니다. 로봇은 같은 동작을 열 번째나 만 번째나 똑같이 합니다. 실험 노트에 적히지 않던 손버릇이 코드로 적히면, 그 실험은 지구 반대편에서도 다시 돌아갑니다.

**참고자료**


[1] Wierenga, R. P., Golas, S. M., Ho, W., Coley, C. W., & Esvelt, K. M. (2023). PyLabRobot: An open-source, hardware-agnostic interface for liquid-handling robots and accessories. Device, 1(4), 100111. https://doi.org/10.1016/j.device.2023.100111

[2] Beguin, V., Grétillat, J., Kaminskait□, K., Juzenas, S., Burgi, P.-Y., & Charmet, J., et al. (2025). High-fidelity robotic PCR amplification. arXiv preprint arXiv:2512.23877. https://arxiv.org/abs/2512.23877

[3] Jiang, Y., et al. (2026). A Self-Evolving Agentic System for Automated Generation and Execution of Biological Protocols (ProtoPilot). arXiv preprint arXiv:2606.31763. https://arxiv.org/abs/2606.31763

[4] Zhu, L., et al. (2026). LAP: An Agent-to-Instrument Protocol for Autonomous Science. arXiv preprint arXiv:2606.03755. https://arxiv.org/abs/2606.03755

[5] Ginkgo Bioworks. (2026년 3월 2일). Ginkgo Bioworks launches Ginkgo Cloud Lab, powered by autonomous lab infrastructure. https://www.prnewswire.com/news-releases/ginkgo-bioworks-launches-ginkgo-cloud-lab-powered-by-autonomous-lab-infrastructure-302700458.html

[6] Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature, 533(7604), 452-454. https://doi.org/10.1038/533452a

[7] King, R. D., Rowland, J., Oliver, S. G., et al. (2009). The automation of science. Science, 324(5923), 85-89. https://doi.org/10.1126/science.1165620

[8] D'Souza, J., et al. (2026). SciSchema.org: A Multidisciplinary Collection of Schemas for Structured Scientific Process Descriptions. arXiv preprint arXiv:2607.27955. https://arxiv.org/abs/2607.27955

[9] Holland, I., & Davies, J. A. (2020). Automation in the life science research laboratory. Frontiers in Bioengineering and Biotechnology, 8, 571777. https://doi.org/10.3389/fbioe.2020.571777

## 12.2 다중 에이전트 자율 루프 기반 신약 물질 발굴 및 실험 Validation 결합 루프 (Latent-Y)

2026년 3월 말, 런던의 한 연구팀이 인공지능에게 문장 하나를 건넸습니다. 특정 표적 단백질에 붙는 항체를 설계하라는 지시였습니다. 사람은 그 뒤로 손을 대지 않았습니다. 시스템은 관련 논문을 읽고, 표적 구조를 뜯어보고, 어느 부위에 붙을지 정하고, 후보 서열을 만들고, 계산으로 걸러내고, 실험실에 보낼 목록을 스스로 골랐습니다.

이 시스템의 이름이 레이턴트와이(Latent-Y)입니다 [1]. 결과는 실험대 위에서 판가름 났습니다. 표적 아홉 개 가운데 여섯 개에서 실제로 붙는 나노바디를 얻었습니다. 표적 기준 성공률 67퍼센트입니다. 결합 세기는 한 자릿수 나노몰 수준까지 내려갔습니다. 사람이 후보를 골라 주거나 중간에 방향을 틀어 주지 않았습니다. 전문가들이 같은 설계 과제를 맡았을 때 걸릴 것으로 예상한 시간과 견주면, 레이턴트와이와 함께 일한 팀은 56배 빨리 끝냈습니다 [2]. 몇 주가 몇 시간으로 접혔습니다.

여기서 쓰인 설계 엔진은 레이턴트엑스투(Latent-X2)입니다. 원자 수준에서 생물학적 분자를 만들어 내는 생성 모델입니다. 같은 구조를 조금 바꾸면 고리형 펩타이드나 소형 결합체 설계에도 그대로 쓰입니다. 하나의 에이전트가 여러 치료 양식을 넘나듭니다.

67퍼센트라는 숫자를 어떻게 읽어야 할까요. 아홉 번 시도해서 세 번은 빈손으로 끝났다는 뜻이기도 합니다. 사람 연구팀도 늘 성공하지는 않습니다. 다만 사람은 실패한 표적에 몇 달을 쓰고 나서야 포기합니다. 자율 루프는 며칠 만에 포기하고 다음 표적으로 넘어갑니다. 속도의 값어치는 성공을 빨리 얻는 데만 있지 않습니다. 실패를 빨리 확인하는 데 있습니다.

자율 실험실의 심장은 하나의 원이었습니다. 설계하고, 만들고, 시험하고, 배우고, 다시 설계합니다. 이 원을 사람 없이 돌린 첫 사례들은 화학에서 나왔습니다. 2020년 리버풀 대학교의 이동형 로봇 화학자는 키 1.75미터의 팔로 실험실을 돌아다니며 8일 동안 광촉매 실험 688건을 수행했습니다. 그리고 처음 물질보다 6배 활성이 좋은 조합을 찾아냈습니다 [3]. 2023년 로런스 버클리 국립연구소의 에이랩(A-Lab)은 17일을 쉬지 않고 돌면서 목표 물질 58개 가운데 41개를 실제로 합성했습니다 [4]. 같은 해 카네기멜런 대학교의 코사이언티스트(Coscientist)는 언어 모델이 문헌을 읽고 실험 계획을 세운 뒤 로봇에게 명령을 내리는 구조를 선보였습니다 [5].

이 원은 네 칸으로 나뉩니다. 설계(Design), 제작(Build), 시험(Test), 학습(Learn)입니다. 앞 글자를 따 DBTL 루프라고 부릅니다. 도서관 계단을 오르내리는 장면과 닮았습니다. 한 층 올라가 책을 뽑고, 읽고, 아니다 싶으면 내려와 다른 층으로 갑니다. 사람이 하면 층마다 며칠이 걸립니다. 로봇이 하면 몇 시간입니다.

시험 칸에서 벌어지는 일이 이 루프의 무게중심입니다. 설계한 항체 서열을 DNA로 합성하고, 세포에 넣어 단백질로 만들고, 정제한 뒤 표적과 붙여 봅니다. 얼마나 세게 붙었는지는 표면 플라스몬 공명(SPR)으로 잽니다. 금속 막 위에 표적을 깔고 항체를 흘려보내면서 반사되는 빛의 각도 변화를 읽는 방법입니다. 붙는 속도와 떨어지는 속도가 실시간 곡선으로 나옵니다. 이 곡선이 다음 설계의 재료가 됩니다. 예측값과 실측값의 차이가 클수록 모델은 크게 배웁니다.

붙지 않은 후보들도 버려지지 않습니다. 실패한 설계는 어디를 건드리면 안 되는지 알려 줍니다. 사람이 하던 시절에는 이런 음성 결과가 논문에 실리지 않았습니다. 실험 노트 구석에 적혔다가 졸업과 함께 사라졌습니다. 자율 루프는 성공과 실패를 같은 형식으로 저장합니다. 실패의 값어치를 처음으로 회수하기 시작한 셈입니다.

원을 돌리는 머리 역할은 베이지안 최적화(Bayesian Optimization)가 맡습니다. 다음에 무엇을 시험할지 정하는 방법입니다. 안개 낀 산에서 정상을 찾는다고 생각하면 됩니다. 아무 데나 파 보는 대신, 지금까지 밟아 본 지점들의 높이로 지형을 짐작하고 다음 발을 놓습니다. 2026년 디지털 디스커버리에 실린 자율 실험실 벤치마킹 연구는 이 방법이 얼마나 절약해 주는지 실측했습니다. 한 사례에서는 같은 성능의 구조를 찾는 데 필요한 실험 횟수가 18분의 1로 줄었습니다 [6]. 항체처럼 따져야 할 조건이 여럿일 때는 다목적 베이지안 최적화가 쓰입니다. 결합력만 높이면 되는 게 아니라 용해도, 안정성, 면역원성을 함께 맞춰야 하기 때문입니다 [7].

루프가 발견까지 밀고 간 사례도 나왔습니다. 2026년 6월 공개된 다중 규모 인공지능 군집 연구는 대장암 전사체를 감독 없이 훑어, 5-플루오로우라실 내성과 얽힌 취약점으로 인슐린 유사 성장인자2(IGF2)를 지목했습니다 [8]. 통계 보정을 거친 뒤에도 유의했고, 독립적으로 진행된 생쥐 실험의 종양 축소 결과와 맞아떨어졌습니다. 이 연구팀이 붙든 문제는 정확도가 아니라 이유였습니다. 언어 모델은 그럴듯한 문장을 잘 만듭니다. 세포는 문장을 읽지 않습니다. 물리 엔진과 설명 기법을 붙여 예측의 근거를 추적한 이유가 여기에

있습니다.

들뜬 소식만 있는 것은 아닙니다. 자율성에도 등급이 있습니다. 0단계부터 5단계까지 나누는 척도가 쓰이는데, 자율 실험실을 표방하는 시스템 대다수는 2단계에 머뭅니다. 프로토콜이 디지털로 적히고 데이터가 기계가 읽을 수 있는 형태로 나오는 수준입니다. 폐루프를 돌리며 이상을 스스로 잡아내는 3단계는 소수입니다. 4단계는 잘 정의된 좁은 문제에서만 증명됐습니다. 5단계는 아직 없습니다 [6]. 베이지안 최적화는 조건 몇 개를 훑는 문제에서 힘을 냅니다. 단계가 길고 오차가 쌓이는 프로토콜에서는 무너집니다. 리버풀의 앤드루 쿠퍼는 진정한 의미의 완전 자율 실험실은 아직 존재하지 않는다고 말합니다.

위험도 같이 커집니다. 24시간 돌아가는 설계 엔진은 약을 만드는 데도, 독을 만드는 데도 같은 속도를 냅니다. 2026년 7월 공개된 몰세이프이밸(MolSafeEval)은 이 빈틈을 겨눈 평가 기준입니다 [9]. 기존 분자 생성 벤치마크는 새로움과 물성만 봤습니다. 만들어진 분자가 독성을 띠는지, 반응성이 위험한지는 채점표에 없었습니다. 로봇이 설계와 합성을 이어 붙이면, 화면 속 구조식은 그날 밤 실제 물질이 됩니다. 안전 장치를 설계 단계에 심어야 하는 이유가 여기 있습니다. 실행 단계에서 막으면 늦습니다.

규제도 따라 움직였습니다. 미국 식품의약국과 유럽 의약품청은 2026년 1월 인공지능이 만든 데이터를 규제 심사에 쓰는 원칙을 함께 내놓았습니다. 로봇이 만든 데이터가 허가 서류에 실리려면, 어느 장비가 언제 어떤 조건으로 측정했는지가 추적돼야 합니다. 12.1에서 본 표준화 작업이 학문적 취향의 문제가 아닌 이유입니다.

자율 루프가 잘 도는 분야와 그렇지 않은 분야도 갈립니다. 무기물 합성처럼 재료를 섞고 굽는 공정은 로봇이 다루기 쉽습니다. 온도와 시간과 비율이 전부이기 때문입니다. 포유류 세포를 다루는 실험은 다릅니다. 세포는 같은 조건에서도 컨디션에 따라 다르게 자랍니다. 배양 접시를 옮기는 손길 하나에 반응합니다. 항체 발굴이 화학보다 늦게 자동화된 이유가 여기에 있습니다. 레이턴트와이가 의미를 갖는 지점도 같습니다. 다루기 까다로운 생물 쪽에서 실험 검증까지 밀고 나갔기 때문입니다.

돈과 시간의 셈법은 그럼에도 바뀌고 있습니다. 캐나다 액셀러레이션 컨소시엄은 정부 지원금 2억 캐나다달러를 받아, 신물질 하나를 찾는 데 드는 1천만 달러와 10년을 100만 달러와 1년으로 줄이겠다는 목표를 내걸었습니다. 로봇 랙 한 대를 붙이는 데 드는 비용은

장비당 1만에서 5만 달러 수준이고, 연결하는 소프트웨어 작업에 10만에서 50만 달러가 더 듭니다. 클라우드 랩을 표방했던 스트라티오스가 방향을 틀었다는 사실도 함께 봐야 합니다. 기술이 되는 것과 사업이 되는 것은 다른 문제였습니다.

사람이 밀려나는 그림이 아닙니다. 레이턴트와이에게 건넨 것은 결국 문장 하나였습니다. 어떤 표적을 왜 공략할 것인가. 그 질문을 고르는 일은 여전히 사람 몫입니다. 로봇은 답을 찾는 속도를 바꿨습니다. 질문의 값어치는 바꾸지 못했습니다.

**참고자료**


[1] Schmon, S. M., Pretorius, D., Mathis, S., et al. (2026). Latent-Y: A lab-validated autonomous agent for de novo drug design. arXiv preprint arXiv:2603.29727. https://arxiv.org/abs/2603.29727

[2] Latent Labs. (2026년 3월 23일). Latent-Y: The autonomous AI agent for drug design at scale. https://www.latentlabs.com/press-release/latent-y/

[3] Burger, B., Maffettone, P. M., Gusev, V. V., et al. (2020). A mobile robotic chemist. Nature, 583(7815), 237-241. https://doi.org/10.1038/s41586-020-2442-2

[4] Szymanski, N. J., Rendy, B., Fei, Y., et al. (2023). An autonomous laboratory for the accelerated synthesis of novel materials. Nature, 624(7990), 86-91. https://doi.org/10.1038/s41586-023-06734-w

[5] Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. Nature, 624(7992), 570-578. https://doi.org/10.1038/s41586-023-06792-0

[6] Benchmarking self-driving labs. (2026). Digital Discovery, 5(1), 14. https://doi.org/10.1039/D5DD00337G

[7] Rao, J., et al. (2026). BOAT: Navigating the Sea of In Silico Predictors for Antibody Design via Multi-Objective Bayesian Optimization. arXiv preprint arXiv:2604.13980. https://arxiv.org/abs/2604.13980

[8] Baker, C., Ren, T., Rafferty, K., Wang, H., & McDade, S. (2026). Autonomous mechanistic discovery of colorectal cancer vulnerabilities via multi-scale AI swarms. arXiv preprint arXiv:2607.16262. https://arxiv.org/abs/2607.16262

[9] Xu, T., et al. (2026). MolSafeEval: A Benchmark for Uncovering Safety Risks in AI-Generated Molecules. arXiv preprint arXiv:2607.00464. https://arxiv.org/abs/2607.00464

[10] Abolhasani, M., & Kumacheva, E. (2023). The rise of self-driving labs in chemical and materials sciences. Nature Synthesis, 2(6), 483-492. https://doi.org/10.1038/s44160-022-00231-0

# 제13장: 생성형 바이오 AI의 이중적 사용 (Dual-Use) 방지 및 안전 가이드라인

## 13.1 AI 설계 병원체 생성 위험 및 바이러스 합성 시나리오에 대응하는 필터링의 중요성

2021년 가을, 미국 노스캐롤라이나의 작은 신약 회사 연구진이 국제 회의 발표 자료를 준비하고 있었습니다. 주제는 인공지능 신약 설계의 오용 가능성이었습니다. 발표에 쓸 사례가 마땅치 않았습니다. 그래서 평소 쓰던 분자 생성 모델의 설정을 하나만 뒤집었습니다. 독성 점수가 낮은 분자를 골라내던 규칙을, 독성 점수가 높은 분자를 골라내는 규칙으로 바꾼 것입니다. 코드 한 줄 수준의 변경이었습니다. 그날 저녁 계산을 걸어 두고 퇴근했습니다.

여섯 시간이 지나기 전에 모델은 4만 개의 후보 분자를 뱉어 놓았습니다. 그 목록에는 이미 알려진 신경작용제가 들어 있었고, 공개된 화학무기보다 독성 예측값이 더 높은 새 분자들도 섞여 있었습니다. 연구진은 결과 파일을 외부와 공유하지 않고 봉인했습니다. 그리고 만드는 방법이 아니라 자신들이 겪은 일만 짧은 논문으로 냈습니다 [1]. 그 논문은 두 쪽이 조금 넘습니다. 분량은 짧았지만 파장은 길었습니다.

이 사건이 보여 준 것은 도구의 성격입니다. 병을 고치는 분자를 찾는 계산과 사람을 해치는 분자를 찾는 계산은 같은 계산입니다. 방향만 반대입니다. 부엌칼로 요리도 하고 사람도 다치게 할 수 있다는 오래된 비유가 여기서는 정확하게 들어맞습니다. 과학계는 이것을 이중적 사용의 딜레마(Dual-Use Dilemma)라고 부릅니다. 하나의 기술이 이로움과 해로움을 동시에 품고 있어서, 해로움만 떼어내 버릴 수 없는 상태를 말합니다.

다행히 설계와 실물 사이에는 넓은 강이 하나 있습니다. 화면에 뜬 서열은 아직 아무것도 아닙니다. 그것이 실제 위협이 되려면 누군가 DNA를 합성하고, 세포에 넣고, 배양하고, 확산시키는 여러 단계를 통과해야 합니다. 그 강을 건너는 다리 가운데 제일 좁고 눈에 잘 띄는 지점이 상업용 DNA 합성 주문입니다. 방어선을 세운다면 그 다리 위가 유리합니다. 그래서 각국 정부와 산업계는 모델의 입출력 단계와 합성 주문 단계, 두 곳에 검문소를 세우는 쪽으로 움직였습니다.

합성 비용이 내려간 것도 이 지형을 바꿨습니다. 예전에는 유전자 한 조각을 주문하는 값이 실험실 예산에서 눈에 띄었습니다. 지금은 소모품 값에 가깝습니다. 값이 내려가면 주문 건수가 늘고, 건수가 늘면 살펴야 할 서류도 함께 늘어납니다. 방어선을 사람 손만으로 지킬

수 없게 된 이유가 여기 있습니다.

다중 방어선(Multi-layered Biosecurity Screening)이라는 말은 이 검문소를 여러 겹 세운다는 뜻입니다. 학습 데이터를 고르는 자리에서 한 겹, 모델 내부에서 생성 흐름을 눌러 주는 자리에서 또 한 겹, 들어오는 요청과 나가는 서열을 살피는 자리에서 다시 한 겹입니다. 그 뒤에 합성 주문 심사가 있고, 요즘은 인터넷으로 실험을 맡기는 원격 실험실 단계까지 붙습니다. 겹마다 놓치는 것이 다릅니다. 데이터 정제는 배포 뒤에 새로 붙는 미세조정을 막지 못합니다. 출력 심사는 여러 번 나눠 들어오는 주문 앞에서 무뎌집니다. 어느 한 겹을 완벽하게 만드는 일보다 겹의 수를 지키는 일이 중요한 이유입니다.

그런데 2025년 10월 2일, 그 검문소가 생각보다 헐겁다는 보고가 학술지 사이언스에 실렸습니다. 마이크로소프트 연구진과 국제 유전자 합성 컨소시엄 소속 전문가들이 함께 쓴 논문입니다 [2]. 연구진은 공개된 단백질 설계 도구로 우려 단백질의 변형본을 만들어 보았습니다. 서열은 달라졌지만 기능은 유지되도록 다시 그린 것입니다. 합성 기업들이 쓰던 심사 소프트웨어는 자연에서 유래한 원본은 잘 잡아냈습니다. 그러나 인공지능이 다시 그린 변형본 앞에서는 자주 눈을 감았습니다.

연구진의 대응 방식이 이 사건의 진짜 교훈입니다. 그들은 2023년 말부터 약 2년 동안 이 작업을 비공개로 진행했습니다. 취약점을 확인한 뒤에는 열 달에 걸쳐 합성 기업, 생물보안 단체, 정책 담당자에게 조용히 알리고 심사 소프트웨어의 보완판을 배포했습니다. 보완판이 현장에 충분히 깔린 다음에야 논문을 공개했습니다. 소프트웨어 보안에서 쓰는 제로데이 취약점 처리 절차를 생물 심사 영역에 그대로 옮겨 온 것입니다. 위험을 발견한 사람이 그 위험을 먼저 떠들지 않고 먼저 막았습니다.

필터가 왜 여러 겹이어야 하는지는 심사 방식의 역사에서도 드러납니다. 초기 심사는 주문받은 서열을 이미 알려진 위험 병원체 서열과 얼마나 닮았는지 비교하는 방식이었습니다. 닮은 정도로 판정하는 방식은 글자를 조금만 바꿔도 흔들립니다. 그래서 심사는 서열의 겉모습이 아니라 그 서열이 만들어 낼 단백질의 구조와 기능을 추정하는 쪽으로 옮겨 가고 있습니다. 모델 안쪽에서도 방어가 시도됩니다. 2026년 3월에 나온 연구는 단백질 언어모델을 특정 분류군에 맞춰 미세조정하기만 해도 독성 단백질이 튀어나올 수 있음을 보이고, 추론 시점에 위험한 출력을 눌러 주는 기법을 제시했습니다 [3]. 2026년 7월에는 생성된 분자의 안전 위험을 재는 전용 평가 기준이 공개되었습니다 [4].

모델이 실제로 어디까지 도와주는지 재 보는 일도 필터의 일부입니다. 2025년 5월 아프리카 연구진은 자사 생물 에이전트에서 안전 장치를 끈 상태로 시험한 결과를 공개했습니다. 장치를 끄자 독성 화합물 설계 요청이 그대로 통과했습니다 [7]. 이런 시험은 공격을 익히려고 하는 것이 아닙니다. 방어선의 두께를 자로 재려고 합니다. 소방 훈련에서 일부러 연기를 피우는 것과 같습니다. 연구진은 통과 여부만 숫자로 보고하고, 통과한 요청의 내용은 공개하지 않았습니다. 이 절제가 없으면 안전 연구가 설명서로 뒤집힙니다.

추론 시점 필터만 믿는 방식에는 한계가 뚜렷합니다. 2025년 8월 연구는 기존 생물안전 대책이 대부분 생성 순간에 걸리는 필터에 기대고 있다고 짚었습니다 [8]. 필터는 모델 바깥에 붙은 부품입니다. 부품은 떼어 낼 수 있습니다. 가중치를 손에 넣은 쪽이 필터를 벗기는 데 걸리는 시간은 길지 않습니다. 그래서 방어는 모델 안쪽과 물리적 공급망 양쪽으로 갈라져야 합니다. 2025년 10월에는 여러 기관 연구자 열아홉 명이 생성형 인공지능이 생물학의 문턱을 낮추면서 새 취약점을 만든다고 정리하고, 평가와 접근 통제와 국제 공조를 한 묶음으로 다루자는 로드맵을 냈습니다 [9].

제도 쪽 숫자도 함께 봐야 합니다. 미국 백악관 과학기술정책실이 2024년 9월에 낸 합성 핵산 심사 프레임워크는 200뉴클레오타이드 이상 주문을 심사 대상으로 잡았습니다 [5]. 이 기준은 2025년 10월 13일부터 50뉴클레오타이드로 좁혀졌습니다. 짧은 조각을 여러 번 나눠 주문해 이어 붙이는 우회로를 막기 위해서입니다. 국제 유전자 합성 컨소시엄은 회원사가 늦어도 2026년 10월 24일까지 이 짧은 기준에 맞추도록 협약을 고쳤습니다 [6]. 2025년 5월 5일 행정명령은 2024년 프레임워크를 개정하거나 대체하라고 지시했는데, 2026년 8월 현재 후속 문서는 아직 나오지 않았습니다. 규칙이 비어 있는 구간에서는 업계 자율 협약이 실질적인 방어선 노릇을 합니다.

심사에 걸린 뒤에 벌어지는 일도 알아 둘 만합니다. 소프트웨어가 붉은 표시를 띄운다고 주문이 곧바로 반려되지는 않습니다. 담당자가 주문서를 열어 봅니다. 이 서열이 왜 필요한지, 주문한 사람이 어느 기관에 속해 있는지, 배송지가 연구 시설이 맞는지 확인합니다. 대학 실험실이 백신 연구용으로 시킨 조각과 주소가 불분명한 개인이 시킨 같은 조각은 다르게 다뤄집니다. 기계는 닮은 정도를 재고, 사람은 맥락을 읽습니다. 둘이 함께 있어야 심사가 굴러갑니다. 그래서 기준을 좁히는 규칙 개정은 담당자 훈련과 인력 확보를 같이 요구합니다. 규칙만 고치고 사람을 늘리지 않으면 붉은 표시가 쌓이다가 무시됩니다.

국경도 빈틈입니다. 심사를 의무로 정한 나라와 권고로 남겨 둔 나라가 섞여 있습니다. 주문서 한 장은 국경을 쉽게 넘습니다. 책상 위에 놓고 쓰는 소형 합성기가 퍼지면서 구멍은 더 커졌습니다. 미국 프레임워크는 제조사가 이런 장비에 심사 기능을 넣도록 권고합니다 [5]. 권고는 의무가 아닙니다. 2025년 11월 군축 전문가들은 소형 합성기 규제 공백을 공개적으로 지적했습니다 [10]. 규칙이 닿지 않는 장비가 늘어날수록 심사의 무게중심은 주문 접수 창구에서 장비 판매 단계로 옮겨 갑니다.

필터를 두고 연구 속도를 깎아 먹는 세금이라고 부르는 목소리가 있습니다. 저는 반대편에 섭니다. 화학 분야에서 4만 개 목록이 여섯 시간 만에 나온다는 사실이 알려진 뒤에도 신약 설계 연구는 멈추지 않았습니다. 연구진이 결과를 봉인하고 절차를 만들었기 때문입니다. 사이언스 논문의 저자들이 취약점을 먼저 알리고 먼저 고쳤기 때문에, 단백질 설계 도구는 지금도 실험실에서 돌아갑니다. 필터는 문을 잠그는 장치가 아니라 문을 계속 열어 두기 위한 조건입니다. 이 조건을 스스로 지키지 못하면 규칙은 바깥에서, 훨씬 거칠게 들어옵니다.

**참고자료**


[1] Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. Nature Machine Intelligence, 4(3), 189-191. https://doi.org/10.1038/s42256-022-00465-9

[2] Wittmann, B. J., Alexanian, T., Bartling, C., Beal, J., Clore, A., Diggans, J., Flyangolts, K., Gemler, B. T., Mitchell, T., Murphy, S. T., Wheeler, N. E., & Horvitz, E. (2025). Strengthening nucleic acid biosecurity screening against generative protein design tools. Science, 390(6768), 82-87. https://doi.org/10.1126/science.adu8578

[3] Fernández Burda, M., Aranguri, S., Arcuschin Moreno, I., & Ferrante, E. (2026). Inference-Time Toxicity Mitigation in Protein Language Models. arXiv preprint arXiv:2603.04045. https://arxiv.org/abs/2603.04045

[4] Xu, T., Cao, X., Zhu, Z., Ding, K., & Chen, H. (2026). MolSafeEval: A Benchmark for Uncovering Safety Risks in AI-Generated Molecules. arXiv preprint arXiv:2607.00464. https://arxiv.org/abs/2607.00464

[5] Office of Science and Technology Policy. (2024년 9월). Framework for Nucleic Acid Synthesis Screening. https://aspr.hhs.gov/S3/Documents/OSTP-Nucleic-Acid-Synthesis-Screening-Framework-Sep2024.pdf

[6] International Gene Synthesis Consortium. Harmonized Screening Protocol v3.0. https://genesynthesisconsortium.org/wp-content/uploads/IGSC-Harmonized-Screening-Protocol-v3.0-1.pdf

[7] Hattoh, G., Ayensu, J., Ofori, N. P., Eshun, S., & Akogo, D. (2025). Can Large Language Models Design Biological Weapons? Evaluating Moremi Bio. arXiv preprint arXiv:2505.17154. https://arxiv.org/abs/2505.17154

[8] Feldman, J., & Feldman, T. (2025). Resilient Biosecurity in the Era of AI-Enabled Bioweapons. arXiv preprint arXiv:2509.02610. https://arxiv.org/abs/2509.02610

[9] Zhang, Z., Chakraborty, S., Bedi, A. S., Mathew, E., Saravanan, V., Cong, L., Velasquez, A., Lin-Gibson, S., Blewett, M., Hendrycks, D., London, A. J., Zhong, E., Raphael, B., Dieng, A. B., Ma, J., Xing, E., Altman, R., Church, G., & Wang, M. (2025). Generative AI for Biosciences: Emerging Threats and Roadmap to Biosecurity. arXiv preprint arXiv:2510.15975. https://arxiv.org/abs/2510.15975

[10] Arms Control Association. (2025년 11월 24일). Regulatory Gaps in Benchtop Nucleic Acid Synthesis Create Biosecurity Vulnerabilities. https://www.armscontrol.org/blog/2025-11-24/regulatory-gaps-benchtop-nucleic-acid-synthesis-create-biosecurity-vulnerabilities

## 13.2 바이오 위협 사전 통제를 위한 AI 기초 모델 구축 초기 안전 전문가 참여 프레임워크

건물을 다 짓고 나서 소방 점검관을 부르면 어떤 일이 벌어질까요. 점검관은 비상구가 하나 더 필요하다고 말합니다. 그런데 그 자리에는 이미 배관이 지나갑니다. 벽을 뜯자니 공사비가 새로 짓는 값에 가깝습니다. 결국 소화기를 몇 대 더 놓고, 안내판을 크게 붙이고, 서류에는 점검 완료라고 적습니다. 인공지능 안전 관리가 오랫동안 이런 모습이었습니다. 학습이 끝난 모델을 배포 직전에 시험하는 사후 레드팀(Post-hoc Red Teaming) 방식입니다. 다 만든 다음에 위험한 질문을 던져 보고 답이 새면 그 구멍만 막는 것입니다.

이 방식이 생물 영역에서 얼마나 버티는지 재 본 연구가 2026년 6월에 나왔습니다. 연구진은 인공지능 에이전트에게 실험 자동화 로봇용 코드 작성, DNA 조각 설계, 심사 시스템의 허점 판별이라는 세 가지 과제를 시켰습니다. 세 과제 모두에서 시험한 모든 에이전트가 사람 전문가 집단의 중간 성적을 넘겼습니다. 실험실 검증에서는 한 모델이 짜 준 스크립트가 실제 액체 취급 로봇 위에서 그대로 돌아갔습니다 [1]. 방어자가 읽어야 할 대목은 성적표가 아니라 방향입니다. 위험은 지식을 알려 주는 데서 멈추지 않고, 손을 대신 움직여 주는 쪽으로 넘어왔습니다.

수백억 개의 값으로 이루어진 거대 생물 기초 모델(Biological Foundation Models)에서 사후 점검은 더 흔들립니다. 모델은 서열을 통째로 외우지 않습니다. 아미노산이 이어지는 규칙, 구조가 접히는 규칙을 잠재 공간(Latent Space)이라는 내부 좌표계에 압축해 둡니다. 도서관에 비유하면 책을 꽂아 둔 상태가 아니라 책을 쓰는 문법을 익힌 상태입니다. 그러니 위험한 책 제목을 목록에서 지운다고 그 문법이 사라지지 않습니다. 앞 절에서 본 2026년 3월 연구가 그 증거입니다. 독성을 학습 목표로 삼지 않았는데도, 특정 생물 분류군에 맞춰 미세조정하자 독성 단백질이 흘러나왔습니다.

그래서 방어선을 앞으로 당기는 설계 기반 보안(Security-by-Design)이 자리를 잡고 있습니다. 안전 담당자를 배포 회의가 아니라 데이터 수집 기획 회의에 앉히는 것입니다. 어떤 서열 데이터베이스를 학습에 넣고 뺄지, 우려 병원체 관련 자료를 미리 걸러 낼지, 걸러 낸 뒤에도 남는 능력을 무엇으로 잴지를 처음부터 정합니다. 학습이 끝난 모델에서 특정 지식만 지우는 기계 망각(Machine Unlearning) 기법도 연구되지만, 지운 뒤에 남는 흔적을 전부 확인하기는 어렵습니다. 재료를 넣기 전에 고르는 쪽이 요리를 다 한 뒤에

재료를 골라내는 쪽보다 쉽습니다.

데이터 정제는 말처럼 깔끔하지 않습니다. 우려 병원체 서열만 골라 빼면 될 것 같지만, 그와 비슷한 조각은 무해한 미생물 유전체에도 널려 있습니다. 넓게 빼면 모델이 백신 연구에 쓸 지식까지 잃습니다. 좁게 빼면 남은 조각들이 다시 이어 붙습니다. 이 눈금을 맞추려면 정제 기준 회의에 미생물학자가 앉아야 합니다. 어떤 서열을 남길지 판단하려면 그 서열이 자연에서 무슨 일을 하는지 알아야 하기 때문입니다. 정제 기준을 정한 회의록을 남기는 일도 중요합니다. 나중에 문제가 생겼을 때 어떤 판단이 어디서 어긋났는지 되짚을 수 있어야 합니다.

위험의 종류를 구분하는 일도 설계 단계의 몫입니다. 2023년 6월에 나온 분석은 대화형 언어모델과 생물 설계 도구의 위험이 서로 다르다고 정리했습니다 [6]. 언어모델은 이미 공개된 지식을 모아 정리해 주면서 진입 장벽을 낮춥니다. 생물 설계 도구는 세상에 없던 분자를 만들어 냅니다. 앞의 것은 아는 사람의 수를 늘리고, 뒤의 것은 없던 것의 수를 늘립니다. 막는 방법도 달라야 합니다. 대화 내용을 살피는 장치는 언어모델에 듣지만, 아미노산 문자열만 주고받는 설계 도구에는 잘 듣지 않습니다.

평가를 설계하는 일 자체가 어렵습니다. 무엇을 재야 위험을 잰 것이 되는지 분명하지 않기 때문입니다. 시험 문제를 잘 푸는 모델이 실제 실험을 성공시키는 모델과 같은지 아무도 확신하지 못합니다. 2026년 4월에 나온 평가 연구는 네 개의 최신 모델에 73개의 초심자 시늉 문항을 던졌습니다 [7]. 어떤 모델은 정당한 학습용 질문까지 거절했고, 어떤 모델은 맥락을 놓친 답을 내놓았습니다. 조이면 학생의 숙제가 막히고, 풀면 방어선이 사라집니다. 이 눈금을 맞추는 자리에 생물학자와 안전 담당자가 함께 붙어야 합니다.

가중치 공개 문제에서 순서는 되돌릴 수 없는 성질을 갖습니다. 서비스형 모델은 문제가 생기면 접속을 끊으면 됩니다. 가중치를 인터넷에 올린 모델에는 회수 버튼이 없습니다. 안전 장치를 얹어 두어도 내려받은 쪽에서 미세조정 몇 시간이면 벗겨 냅니다. 연구자 100명이 넘는 인원이 공개 전 생물보안 시험을 지지하는 성명에 이름을 올렸지만, 실제로 그런 시험을 거쳐 공개된 모델은 전체 공개 건수의 2.5%에 못 미칩니다. 2025년 12월에 공개된 단백질 설계 도구의 새 판본은 이전 판본보다 열 배 빠릅니다. 속도는 열 배로 오르는데 검증 관행은 제자리입니다.

모델 안쪽만 조이는 방식에도 구조적인 한계가 있습니다. 2026년 2월 5일 공개된 연구는 키워드 차단, 출력 심사, 내용 기반 거부 같은 모델 수준 통제가 생물 설계 도구에는 잘 맞지 않는다고 지적합니다 [2]. 같은 서열이 백신 연구에서는 정당하고 다른 맥락에서는 위험합니다. 요청 문장만 보고는 판정할 수 없습니다. 그래서 이 연구는 무엇을 요청했는지가 아니라 누가 요청했는지를 확인하는 고객 확인(Know Your Customer) 절차를 생물보안 기반 시설로 삼자고 제안합니다. 소속 기관, 연구 윤리 심의 통과 여부, 배송 주소를 확인하는 일은 은행이 이미 수십 년째 해 온 일입니다.

이 생각은 관리형 접근(Managed Access)이라는 이름으로 산업 현장에 들어왔습니다. 가중치는 공개하지 않고, 웹 화면과 API로 능력만 내어 주는 방식입니다. 핵위협방지구상이 낸 관리형 접근 프레임워크는 도구의 위험도를 나누고, 등급마다 공개 범위와 이용자 확인 절차를 다르게 정하도록 제안합니다 [9]. 구조 예측처럼 널리 쓰이는 기능은 문을 열어 둡니다. 없던 기능을 처음부터 만들어 내는 도구는 소속과 목적을 확인한 뒤에 내어 줍니다. 기록도 남깁니다. 사고가 났을 때 어디서 무엇이 나갔는지 되짚어야 다음 방어선을 고칠 수 있습니다. 같은 기관이 2026년 봄에 낸 전망 보고서는 여기에 시한을 붙였습니다. 앞으로 18개월이 기업의 자율 안전 관행과 국제 조율이 기술 속도를 따라잡는지 판가름하는 구간이라는 것입니다 [3].

제도는 조금씩 자리를 잡고 있습니다. 2026년 5월 5일 미국 국립표준기술연구소 산하 인공지능 표준혁신센터는 구글 딥마인드, 마이크로소프트, xAI와 사전 배포 시험 협약을 맺었습니다. 기존의 오픈AI, 앤트로픽을 더해 다섯 개 기업이 대상이 되었고, 미공개 모델을 포함해 완료된 평가가 40건을 넘었습니다. 생물보안과 화학무기 위험 평가 가운데 일부는 기밀 환경에서 부처 합동 조직이 수행합니다 [8]. 기업 쪽에서는 앤트로픽이 2025년 5월 화학·생물 위험에 대응하는 안전 등급을 발동하면서 가중치 탈취 방지와 배포 제한을 함께 걸었습니다 [10]. 유럽에서는 생물 기초 모델이 유럽연합 인공지능법의 체계적 위험 범용 모델 의무에 걸리는지를 두고 2026년 5월 분석이 나왔습니다 [4]. 미국 상원에는 2026년 2월 4일 생물보안 현대화 법안이 발의되었습니다. 2026년 2월 24일 공개된 국제 인공지능 안전 보고서는 29개국이 참여한 전문가 패널의 검토를 담았습니다 [5].

저는 이 흐름에서 한쪽 편을 듭니다. 안전 전문가를 마지막에 부르는 관행을 버려야 합니다. 데이터 목록을 짜는 첫 회의에 생물안전 담당자가 앉아 있으면, 나중에 벽을 뜯을 일이 줄어듭니다. 비용도 그쪽이 쌉니다. 사이언스 논문의 사례에서 취약점을 고치는 데 열 달이

걸렸습니다. 그 열 달은 이미 세상에 나간 도구를 뒤쫓느라 쓴 시간입니다. 같은 사람들이 처음부터 한 자리에 앉았다면 몇 주로 끝났을 일입니다.

### 참고자료


[1] Liu, A. B., Nedungadi, S., Cai, B., Kleinman, A., Bhasin, H., & Donoughe, S. (2026). ABC-Bench: An Agentic Bio-Capabilities Benchmark for Biosecurity. arXiv preprint arXiv:2606.11150. https://arxiv.org/abs/2606.11150

[2] Feldman, J., Feldman, T., & Anton, A. I. (2026). Know Your Scientist: KYC as Biosecurity Infrastructure. arXiv preprint arXiv:2602.06172. https://arxiv.org/abs/2602.06172

[3] Nuclear Threat Initiative. (2026년 봄). AIxBio Horizon Scan: Spring 2026. https://www.nti.org/analysis/articles/aixbio-horizon-scan-spring-2026/

[4] Qureshi, T., Griffith, J., Holtman, K., Mir Teijeiro, M., Chin, Z. S., & Gipi⌧kis, R. (2026). The Case for ESM3 as a General-Purpose AI Model with Systemic Risk Under the EU AI Act. arXiv preprint arXiv:2605.01611. https://arxiv.org/abs/2605.01611

[5] Bengio, Y. et al. (2026). International AI Safety Report 2026. arXiv preprint arXiv:2602.21012. https://arxiv.org/abs/2602.21012

[6] Sandbrink, J. B. (2023). Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. arXiv preprint arXiv:2306.13952. https://arxiv.org/abs/2306.13952

[7] Richter, M. (2026). Model Capability Assessment and Safeguards for Biological Weaponization. arXiv preprint arXiv:2604.19811. https://arxiv.org/abs/2604.19811

[8] OpenAI. (2026). Working with US CAISI and UK AISI to build more secure AI systems. https://openai.com/index/us-caisi-uk-aisi-ai-update/

[9] Nuclear Threat Initiative. A Framework for Managed Access to Biological AI Tools. https://www.nti.org/analysis/articles/a-framework-for-managed-access-to-biological-ai-tools/

[10] Anthropic. (2025년 5월 22일). Activating ASL-3 protections. https://www.anthropic.com/news/activating-asl3-protections

# 제14장: 바이오 AI 글로벌 거버넌스 및 인간의 책임

## 14.1 유럽연합(EU) AI Act 규제 하의 생물학적 인공지능 안전성 진단 및 시스템 위험 관리

2026년 7월 24일, 유럽연합 관보에 규정 하나가 실렸습니다. 사흘 뒤인 7월 27일에 발효했습니다. 디지털 옴니버스(Digital Omnibus)라는 이름이 붙은 이 규정은 인공지능법의 고위험 조항 적용일을 뒤로 미뤘습니다 [1]. 그 시점은 원래 마감일인 8월 2일을 엿새 앞둔 때였습니다. 유럽 곳곳의 바이오테크 규제 담당자들은 여름휴가를 반납하고 서류를 맞추던 중이었습니다. 책상 위에 쌓아 둔 기술 문서 더미가 갈 곳을 잃었습니다.

미뤄졌다는 말은 없어졌다는 뜻이 아닙니다. 부속서 III에 열거된 독립형 고위험 AI 시스템의 의무는 2027년 12월 2일로, 이미 유럽 제품안전법의 규율을 받는 제품에 내장된 AI는 2028년 8월 2일로 넘어갔습니다 [1][2]. 의료기기 소프트웨어가 바로 뒤쪽 경로에 속합니다. 유럽 의료기기규정(MDR)에 따라 IIa 등급 이상으로 분류된 진단 보조 AI는 인공지능법 제6조 제1항을 통해 고위험으로 편입되며, 시한은 2028년 8월 2일입니다 [3]. 인증 절차가 두 개로 늘어나지는 않습니다. 인공지능법의 요구사항이 기존 인증기관(Notified Body) 심사 안으로 들어갑니다. 심사표의 항목이 늘어날 뿐입니다.

유예되지 않은 것들도 있습니다. 생성형 AI가 만든 이미지와 음성, 영상에 표시를 붙이라는 제50조의 투명성 의무는 예정대로 2026년 8월 2일에 적용됐습니다 [1]. 범용 인공지능 모델(GPAI)에 관한 의무는 그보다 1년 앞선 2025년 8월 2일에 이미 발효했고, 집행위원회가 실제로 과징금을 물릴 수 있게 된 날짜가 2026년 8월 2일입니다 [4]. 유럽연합 집행위원회는 2025년 7월 10일 범용 AI 행동강령(Code of Practice)을 공개했습니다. 투명성, 저작권, 안전·보안 세 장으로 이루어진 자율 규범입니다. 서명은 강제가 아니지만, 서명한 기업은 법상 의무를 이행한 것으로 추정받습니다 [4].

생명과학 모델은 어느 칸에 들어갈까요. 단백질 언어모델을 두고 이 질문을 정면으로 다룬 논문이 2026년 5월에 나왔습니다. 연구진은 ESM3 같은 최전선 생물학 기반모델을 생물학적 위해 사슬(biorisk chain)의 각 단계에 대응시켜 본 뒤, 이 모델들이 시스템적 위험을 지닌 범용 AI에 해당한다고 결론지었습니다 [5]. 법이 정한 추정 기준은 학습에 쓴 연산량 10의 25제곱 부동소수점 연산입니다. 숫자 하나로 그은 선이라 논쟁이 많습니다. 연산량이 작아도 위험한 모델이 있고, 연산량이 커도 무해한 모델이 있습니다. 그래도 선을

긋지 않으면 아무것도 시작되지 않습니다.

고위험으로 분류된 시스템에는 기록 의무가 따라붙습니다. 제12조는 시스템이 작동하는 동안 로그를 자동으로 남기라고 요구합니다. 제19조는 그 로그를 최소 여섯 달 보관하라고 정합니다. 무엇을 넣어야 하는지도 구체적입니다. 사용 기간, 조회한 참조 데이터베이스, 일치 판정을 부른 입력 데이터, 그리고 결과를 검증한 사람의 신원입니다 [6]. 부속서 IV의 기술 문서는 10년입니다. 유전체 변이 판독 소프트웨어를 떠올려 보십시오. 어느 환자의 어느 변이를 어떤 데이터베이스와 대조해 병원성으로 판정했는지, 그 판정을 누가 확인했는지가 6개월 뒤에도 남아 있어야 합니다.

초안 단계에서 흔히 잘못 알려진 대목이 과징금입니다. 세계 매출의 7퍼센트는 제5조가 금지한 행위를 했을 때의 상한입니다. 3,500만 유로와 세계 연매출 7퍼센트 중 큰 금액이 적용됩니다. 고위험 의무를 어기거나 문서를 갖추지 못한 경우의 상한은 1,500만 유로 또는 세계 연매출 3퍼센트입니다 [6]. 감사 로그를 남기지 않았다고 7퍼센트를 물지는 않습니다. 액수를 정확히 아는 일이 규제 대응의 첫 단추입니다. 겁을 먹는 것과 준비를 하는 것은 다른 일입니다.

기록만 남기면 되는 것도 아닙니다. 제10조는 학습과 검증, 시험에 쓴 데이터의 품질을 요구합니다. 데이터가 어디서 왔는지, 어떤 집단이 얼마나 들어 있는지, 빠진 집단은 없는지를 문서로 보여야 합니다 [2]. 유전체 데이터에서 이 조항은 아프게 다가옵니다. 공개된 대규모 유전체 데이터베이스의 참가자 다수는 유럽계 조상을 가졌습니다. 그런 자료로 학습한 변이 병원성 예측기는 다른 조상 집단의 환자에게서 성능이 떨어집니다. 제9조는 시스템의 전 생애에 걸친 위험 관리 절차를 요구하고, 제15조는 정확도와 사이버 보안 수준을 요구합니다. 조항 번호를 외울 일은 아닙니다. 실험실에서 할 일은 하나입니다. 데이터의 출처와 한계를 스스로 적어 두는 습관을 들이는 것입니다.

이 문서 작업을 사람 손으로만 감당하기는 어렵습니다. 모델 하나가 여섯 달 사이에 세 번 갱신되는 분야입니다. 워드 파일로 관리하는 기술 문서는 저장하는 순간부터 낡습니다. 그래서 문서 자체를 기계가 읽을 수 있는 형식으로 만들자는 제안이 나왔습니다. 2025년에 공개된 AI 모델 여권(AI Model Passport)은 보건·의료 AI의 데이터 출처, 학습 조건, 평가 결과, 책임자를 정해진 항목으로 묶어 모델과 함께 옮겨 다니게 하는 틀입니다 [7]. 식품 포장지 뒷면의 성분표와 비슷합니다. 무엇이 들었는지, 언제 만들었는지, 누가

만들었는지가 한 면에 적혀 있습니다.

규제 대응은 공짜가 아닙니다. 대형 제약사에는 규제 담당 부서가 있습니다. 변호사와 품질 관리자가 상주하고, 인증기관과 통화할 사람이 정해져 있습니다. 대학 실험실에서 막 창업한 다섯 명짜리 회사에는 그런 부서가 없습니다. 논문을 쓰던 사람이 기술 문서를 쓰고, 코드를 짜던 사람이 로그 보관 정책을 만듭니다. 적용일이 2027년 12월과 2028년 8월로 밀린 배경에는 이런 사정이 있었습니다. 조화 표준이 아직 다 나오지 않았다는 지적도 컸습니다. 지킬 기준이 확정되지 않았는데 지키라고 요구하기는 어렵습니다.

한국은 유럽보다 먼저 움직였습니다. 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법이 2026년 1월 22일 시행됐고, 대통령령 제36053호가 같은 날 함께 시행됐습니다 [8]. 의료, 금융, 생체 인식처럼 국민의 생명과 권익에 직결되는 영역은 고영향 인공지능으로 묶입니다. 유럽이 고위험 조항을 2027년 12월로 미룬 상황을 놓고 보면, 규범을 실제로 전면 적용한 첫 나라는 한국이 됐습니다. 다만 정부는 최소 1년의 계도기간을 두었습니다. 과태료 부과가 2027년 1월까지 유예됩니다. 의무가 유예된 것이 아니라 제재가 유예된 것입니다. 과태료 상한은 3,000만 원으로, 유럽의 액수와는 자릿수가 다릅니다.

신약 쪽은 규제의 말투가 조금 다릅니다. 미국 식품의약국(FDA)은 2025년 1월, 의약품의 안전성과 유효성, 품질에 관한 규제 판단을 뒷받침하는 데 AI를 쓸 때 그 모델의 신뢰도를 어떻게 세우고 보일지에 관한 지침 초안을 냈습니다. 2016년 이후 AI 요소가 들어간 허가 제출이 500건을 넘긴 뒤였습니다. 2026년 1월 14일에는 FDA와 유럽 의약품청(EMA)이 의약품 개발에서의 좋은 AI 실무 원칙 10가지를 함께 발표했습니다 [9]. 사람 중심 설계, 위험 기반 접근, 사용 맥락의 명확화, 데이터 거버넌스와 문서화, 생애주기 관리가 그 안에 들어 있습니다. 모델을 금지하는 규범이 아닙니다. 어디까지 믿을 수 있는지를 스스로 증명해 보이라는 요구입니다.

제일 곤란한 자리는 생물보안입니다. 2025년 10월 2일 사이언스에 실린 연구가 그 틈을 드러냈습니다. 연구진은 공개된 단백질 설계 모델 세 종으로 우려 대상 단백질 72종을 흉내 낸 유전자 서열 76,080개를 만들었습니다. 합성 업체들이 쓰는 서열 대조 검색 도구는 이 서열들을 상당수 걸러내지 못했습니다. 서열은 달라도 접히는 모양과 기능은 비슷하기 때문입니다. 연구진은 결과를 공개하기 전에 보완 패치를 만들어 배포했습니다 [10]. 모델 출력만 막아서는 소용이 없다는 진단에서, 합성 주문자의 신원 확인을 생물보안의 기반

시설로 삼자는 제안도 나왔습니다 [11]. 생성된 분자의 독성과 반응성을 미리 점검하는 안전성 평가 기준도 만들어지고 있습니다 [12].

이 문제가 서류 규제와 다른 점은 되돌릴 수 없다는 데 있습니다. 잘못 붙인 라벨은 떼고 다시 붙이면 됩니다. 합성돼 배송된 DNA는 회수할 방법이 없습니다. 그래서 방어를 한 겹으로 짜지 않습니다. 모델 쪽에서는 유해한 서열을 만들지 않도록 학습 단계에서 선호를 조정합니다. 합성 업체 쪽에서는 주문 서열을 검색합니다. 그 앞단에서는 주문한 사람이 누구인지 확인합니다. 어느 한 겹도 완전하지 않습니다. 겹쳐 놓으면 뚫기가 어려워집니다. 문 하나에 자물쇠를 세 개 다는 방식입니다.

법은 늦게 오고, 자주 미뤄지고, 나라마다 다르게 옵니다. 그래도 방향은 한곳을 봅니다. 모델의 성능 지표 옆에 기록과 책임의 칸이 생겼습니다. 이제 실험실 서버 랙 옆에는 로그 저장소가 함께 놓입니다. 예전에는 모델을 배포하면 일이 끝났습니다. 지금은 배포한 날부터 기록이 쌓이기 시작합니다. 어느 날 감사관이 찾아와 여섯 달 전 화요일 오후에 그 시스템이 무엇을 보고 무엇이라고 답했는지 물으면, 답이 나와야 합니다. 그 답을 준비하는 사람은 법무팀이 아니라 시스템을 만든 사람입니다.

**참고자료**


[1] Gibson Dunn. (2026년 7월). EU AI Act Omnibus Agreement: Postponed High-Risk Deadlines and Other Key Changes. https://www.gibsondunn.com/eu-ai-act-omnibus-agreement-postponed-high-risk-deadlines-and-other-key-changes/

[2] European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[3] European Commission. Article 6: Classification Rules for High-Risk AI Systems. EU Artificial Intelligence Act. https://artificialintelligenceact.eu/article/6/

[4] European Commission. (2025년 7월 10일). The General-Purpose AI Code of Practice. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

[5] Qureshi, T., et al. (2026). The Case for ESM3 as a General-Purpose AI Model with Systemic Risk Under the EU AI Act. arXiv preprint arXiv:2605.01611. https://arxiv.org/abs/2605.01611

[6] European Commission. Article 19: Automatically Generated Logs. EU Artificial Intelligence Act. https://artificialintelligenceact.eu/article/19/

[7] Kalokyri, V., et al. (2025). AI Model Passport: Data and System Traceability Framework for Transparent AI in Health. arXiv preprint arXiv:2506.22358. https://arxiv.org/abs/2506.22358

[8] 한국지능정보사회진흥원(NIA). (2026년 1월 22일). AI기본법 시행 및 하위법령 공개. https://nia.or.kr/site/nia_kor/ex/bbs/View.do?cbIdx=99835&bcIdx=28987&parentSeq=28987

[9] U.S. Food and Drug Administration. Artificial Intelligence for Drug Development. Center for Drug Evaluation and Research. https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/artificial-intelligence-drug-development

[10] Wittmann, B. J., Alexanian, T., et al. (2025). Strengthening nucleic acid biosecurity screening against generative protein design tools. Science. https://doi.org/10.1126/science.adu8578

[11] Feldman, J., et al. (2026). Know Your Scientist: KYC as Biosecurity Infrastructure. arXiv preprint arXiv:2602.06172. https://arxiv.org/abs/2602.06172

[12] Xu, T., et al. (2026). MolSafeEval: A Benchmark for Uncovering Safety Risks in AI-Generated Molecules. arXiv preprint arXiv:2607.00464. https://arxiv.org/abs/2607.00464

## 14.2 AI가 제시하는 임상적 발견의 신뢰성, 재현성 위기 극복 및 인간의 최종 책임과 지혜의 몫

96웰 플레이트를 배양기에 넣고 문을 닫습니다. 화면에는 결합 확률 99퍼센트라는 숫자가 떠 있습니다. 사흘 뒤 플레이트를 꺼내 판독기에 올립니다. 아무 일도 일어나지 않았습니다. 대조군과 구분되지 않는 수치가 96개 칸에 고르게 깔려 있습니다. 실험을 다시 짭니다. 시약을 바꾸고, 세포주를 바꾸고, 농도를 바꿉니다. 결과는 같습니다.

이런 밤은 드물지 않습니다. 2016년 네이처가 연구자 1,576명에게 물었습니다. 70퍼센트가 다른 사람의 실험을 재현하는 데 실패한 적이 있다고 답했습니다. 절반이 넘는 사람은 자기 실험을 자기가 재현하지 못한 경험을 털어놨습니다 [1]. 그보다 11년 전, 존 이오아니디스는 출판된 연구 결과의 상당수가 거짓일 수밖에 없는 이유를 통계로 보였습니다 [2]. 재현성 위기라는 말은 인공지능이 오기 전부터 있었습니다. 인공지능은 그 위기를 만들지 않았습니다. 다만 속도를 붙였습니다.

기계학습이 이 위기에 얹는 문제는 성격이 다릅니다. 사람의 실수가 아니라 평가 설계의 허점에서 옵니다. RNA 이차구조 예측을 예로 들겠습니다. 2026년 3월에 나온 연구는 학습 데이터와 시험 데이터를 서열 유사도 기준으로 제대로 갈라 놓았을 때 순위표가 뒤집힌다는 사실을 보였습니다. 기존 벤치마크에서 앞서던 딥러닝 모델과 RNA 기반모델의 일반화 능력이 과대평가돼 있었습니다 [3]. 단일세포 분야에서도 비슷한 일이 벌어졌습니다. 2026년 2월 연구는 학습 파라미터가 하나도 없는 표현 방식이 대규모 기반모델을 여러 하위 과제에서 앞선다고 보고했습니다 [4]. 유전체 언어모델이 유전자 조절의 위치 논리를 정말 배웠는지 검사한 연구도 있습니다. 조절 요소의 위치를 바꾸면 예측도 따라 바뀌어야 하는데, 모델들은 따라 바뀌지 않았습니다 [5]. 상관관계를 외웠을 뿐 작동 원리를 배우지 못한 것입니다.

더 아픈 사례는 이미지 쪽에 있습니다. 극저온 전자현미경(cryo-EM)으로 단백질 구조를 볼 때, 첫 단계는 사진에서 입자 후보를 골라내는 일입니다. 이 작업은 주형 대조나 딥러닝으로 자동화돼 있습니다. 2025년 연구는 순수한 잡음 이미지에 주형을 대고 입자를 고르면 잡음에서 그럴듯한 구조가 만들어져 나온다는 것을 보였습니다 [6]. 없는 것을 본 셈입니다. 찾으려는 모양을 알려 주면 알고리즘은 그 모양을 찾아 옵니다. 확증 편향이 사람의 눈에서 코드로 옮겨 갔습니다.

전처리 단계도 조용한 함정입니다. 대사체 분석에서는 시료를 기계에 넣고 나서 결과표를 만들기까지 정해야 할 항목이 수십 개입니다. 잡음 제거 방식, 정규화 방법, 결측치 처리 기준이 각각 여러 갈래로 갈립니다. 어느 쪽을 골라도 논문에 쓸 수 있는 근거가 있습니다. 연구자는 보통 한 갈래만 고르고 결과를 냅니다. 2026년 7월에 나온 연구는 여러 갈래를 모두 돌려 본 뒤, 어떤 경로를 거쳐도 살아남는 특징만 최종 목록에 올리는 방식을 내놓았습니다 [7]. 답을 하나 내는 대신, 답이 얼마나 흔들리는지를 먼저 재는 방식입니다.

평가 기반 시설 자체를 다시 짓는 움직임도 있습니다. 세포 이미지 프로파일링 분야의 공개 자료 JUMP는 용량이 115테라바이트에 이릅니다. 이 정도 규모면 연구실마다 필요한 부분만 잘라 쓰게 되고, 그러면 서로의 결과를 나란히 놓고 비교하기 어려워집니다. 2026년 8월에 나온 JUMP-lite는 이 자료를 작고 표준화된 평가 묶음으로 줄여, 누구나 같은 조건에서 세포 표현형 모델을 견주도록 만든 시도입니다 [8]. 벤치마크를 작게 만드는 일은 논문 제목으로 화려하지 않습니다. 그래도 이런 작업이 없으면 순위표의 1위는 아무 뜻이 없습니다.

그래서 요즘 좋은 모델은 답과 함께 자신 없음의 크기를 내놓습니다. 유전체 응용에서 여러 불확실성 추정 기법을 나란히 놓고 비교한 2026년 8월 연구는, 기법마다 신뢰도가 크게 다르며 어떤 것은 실제 오차와 잘 맞지 않는다고 보고했습니다 [9]. 독성 예측에서도 과거 성적을 같은 조건으로 다시 돌려 정리한 재현 가능한 순위표가 나왔습니다 [10]. 점수를 올리는 경쟁에서, 그 점수가 무엇을 뜻하는지 따지는 작업으로 무게가 옮겨 가는 중입니다. 확률 0.99와 확률 0.51을 같은 얼굴로 내미는 모델은 실험대 앞에서 쓸모가 적습니다.

실험실을 벗어나는 순간 조건이 달라진다는 문제도 남습니다. 학습에 쓰는 라벨이 정확한 정답이 아니라 대리 지표인 경우가 흔합니다. 세포 생존율로 약효를 대신 재고, 결합 세기로 치료 효과를 대신 잽니다. 그런데 병원이 바뀌고 장비가 바뀌면 그 대리 지표와 실제 결과 사이의 관계도 함께 바뀝니다. 2026년 4월 연구는 이 현상에 감독 표류(supervision drift)라는 이름을 붙였습니다 [11]. 정답의 정의 자체가 환경마다 움직인다는 뜻입니다. 그런 세상에서도 버티는 예측기를 어떻게 학습시킬지가 지금의 숙제입니다.

임상 성적표는 이 이야기를 숫자로 확인해 줍니다. AI로 발굴한 분자들의 임상 1상 성공률은 80에서 90퍼센트로 업계 평균을 크게 웃돕니다. 그런데 2상 성공률은 40퍼센트 안팎으로, 기존 신약과 거의 같습니다 [12]. 1상은 안전성을 봅니다. 분자의 물리화학적 성질을 잘

설계하면 넘을 수 있습니다. 2상은 효능을 봅니다. 이 병의 원인이 정말 이 표적인지를 사람의 몸이 판정합니다. 인공지능은 분자를 빨리 만들었습니다. 표적이 옳은지는 대신 답해 주지 못했습니다. 2026년 현재 임상에 들어간 AI 발굴 프로그램은 173건을 넘었지만, 아직 승인까지 간 약은 없습니다. 발굴 기간은 몇 년씩 줄었습니다. 임상 2상과 3상의 시계는 그대로 흘러갑니다. 사람에게 약을 먹이고 결과를 기다리는 시간은 알고리즘으로 줄일 수 있는 종류의 시간이 아닙니다.

여기서 갈리는 것이 상관관계와 인과관계입니다. 인공지능은 데이터 안의 규칙을 찾아내는 계산기입니다. 데이터에 담기지 않은 개입의 결과를 스스로 알지는 못합니다. 유전자 조절 생물학에서 제대로 된 기반모델을 만들려면 반쯤 기계론적인 모형, 그러니까 알려진 생물학적 작동 방식을 구조 안에 심어 넣은 모형이 있어야 한다는 주장이 이미 나와 있습니다 [13]. 데이터를 아무리 쌓아도 인과의 사다리를 저절로 올라가지는 못한다는 뜻입니다. 그 사다리를 올리는 일은 실험 설계자의 몫입니다. 무엇을 고정하고 무엇을 바꿀지 정하는 사람, 대조군을 어디에 둘지 정하는 사람이 인과를 만듭니다.

실험 설계 회의의 풍경을 떠올려 보겠습니다. 화이트보드에 후보 열 개가 적혀 있습니다. 누군가 손을 들고 묻습니다. 이 중 셋은 같은 회사 화합물 라이브러리에서 왔는데, 그 라이브러리의 용매가 나머지와 다르지 않으냐고요. 그 한마디로 대조군이 하나 늘어납니다. 용매만 넣은 웰이 플레이트 구석에 자리를 잡습니다. 사흘 뒤 그 웰에서 신호가 나옵니다. 후보 셋의 효과는 화합물이 아니라 용매의 것이었습니다. 모델은 이 질문을 하지 않습니다. 라이브러리의 용매 정보는 학습 데이터 어디에도 들어 있지 않았기 때문입니다.

그러니 해자는 코드가 아닙니다. 알고리즘은 논문과 함께 공개되고, 가중치는 몇 달이면 따라잡힙니다. 남는 것은 두 가지입니다. 하나는 오래 쌓아 온 실험 데이터입니다. 실패한 실험까지 형식을 맞춰 기록해 둔 데이터베이스는 돈으로 사기 어렵습니다. 다른 하나는 질문을 고르는 감각입니다. 어떤 병의 어떤 표적을 왜 지금 건드려야 하는지 아는 사람은 모델이 뱉은 만 개의 후보 중에서 열 개를 집어냅니다. 그 열 개를 고르는 손이 연구의 속도를 정합니다.

인공지능은 하루에 수만 개의 가설을 만듭니다. 그중 무엇을 실험대에 올릴지, 어디서 멈출지, 환자에게 줄 것인지를 정하는 자리는 여전히 비어 있지 않습니다. 유럽의 감사 로그 규정이 요구하는 항목 중에 이런 것이 있습니다. 결과를 검증한 자연인의 신원. 조항의

문장은 건조하지만 뜻은 분명합니다. 기계가 낸 판단 옆에 사람 이름을 적으라는 것입니다.

실험 노트의 마지막 줄에는 확인자 서명란이 있습니다. 오늘도 어느 실험실에서 누군가 그 칸에 이름을 적고 날짜를 씁니다. 3차 반복, 음성, 재설계. 볼펜 자국이 종이 뒷장까지 배어납니다. 그 칸에는 모델 이름이 들어가지 않습니다.

### 참고자료


[1] Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature, 533(7604), 452-454. https://doi.org/10.1038/533452a

[2] Ioannidis, J. P. A. (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

[3] Chen, Z., et al. (2026). Fair splits flip the leaderboard: CHANRG reveals limited generalization in RNA secondary-structure prediction. arXiv preprint arXiv:2603.22330. https://arxiv.org/abs/2603.22330

[4] Souza, H., & Mehta, P. (2026). Parameter-free representations outperform single-cell foundation models on downstream benchmarks. arXiv preprint arXiv:2602.16696. https://arxiv.org/abs/2602.16696

[5] Cheng, B., & Zhang, J. (2026). The Mechanistic Invariance Test: Genomic Language Models Fail to Learn Positional Regulatory Logic. arXiv preprint arXiv:2604.06549. https://arxiv.org/abs/2604.06549

[6] Balanov, A., et al. (2025). Structure from Noise: Confirmation Bias in Particle Picking in Structural Biology. arXiv preprint arXiv:2507.03951. https://arxiv.org/abs/2507.03951

[7] Al-Huraibi, M. S., et al. (2026). A multiverse-consensus pipeline for reproducible feature selection in untargeted LC-MS metabolomics. arXiv preprint arXiv:2607.17345. https://arxiv.org/abs/2607.17345

[8] Muñoz, A. F., et al. (2026). JUMP-lite: Compact, reproducible benchmarking of cell representations. arXiv preprint arXiv:2608.07632. https://arxiv.org/abs/2608.07632

[9] Saran, S., et al. (2026). Uncertainty-Aware Deep Learning for Genomics Applications: Insights from an Empirical Study. arXiv preprint arXiv:2608.11054. https://arxiv.org/abs/2608.11054

[10] Ebner, A., et al. (2025). Measuring AI Progress in Drug Discovery: A Reproducible Leaderboard for the Tox21 Challenge. arXiv preprint arXiv:2511.14744. https://arxiv.org/abs/2511.14744

[11] Shoeibi, M., et al. (2026). Learning Stable Predictors from Weak Supervision under Distribution Shift. arXiv preprint arXiv:2604.05002. https://arxiv.org/abs/2604.05002

[12] Jayatunga, M. K. P., Ayers, M., Bruens, L., Jayanth, D., & Meier, C. (2024). How successful are AI-discovered drugs in clinical trials? A first analysis and emerging lessons. Drug Discovery Today, 29(6), 104009. https://doi.org/10.1016/j.drudis.2024.104009

[13] Kova□evi□, L., et al. (2025). No Foundations without Foundations -- Why semi-mechanistic models are essential for regulatory biology. arXiv preprint arXiv:2501.19178. https://arxiv.org/abs/2501.19178

# 인공지능이 여는 생명과학의 새 시대

ISBN |

저 자 | 김경진

펴낸이 | 김경진

펴낸곳 | 김경진 변호사 출판사

출판사등록 | 2025. 3. 10. (제2025-000015호)

주 소 | 서울특별시 동대문구 전농로 91, 백일빌딩 304호

전 화 | 02-6338-1905

이메일 | kimkj008@gmail.com

가격 : 20,000원



참고) 이 책의 일부 내용과 편집 과정에는 인공지능 도구의 도움을 받았습니다.

이 책을 잘 읽으셨으면 그리고 새로운 가치있는 지식을 얻으셨다고 판단되시면
농협 302-1096-0948-81 (예금주 김경진) 에 자발적 후원 부탁드립니다.