

AI 기반 바이오마커 발굴과 약물 반응 예측: 정밀의료를 앞당기는 데이터 기반 신약개발

서울대학교 컴퓨터공학부 / (주)아이젠드릭 김선 교수

(주)아이젠드릭 안수균 선임연구원

(주)아이젠드릭 최정욱 연구원

(주)아이젠드릭 조창연 책임연구원

들어가며

같은 병명으로 진단받았다고 해서 모두 같은 환자는 아니다. 어떤 환자는 병이 빠르게 진행하고, 어떤 환자는 오랫동안 안정적으로 지낸다. 같은 치료를 받아도 누구에게는 효과가 있지만 누구에게는 그렇지 않다. 바이오마커는 이러한 차이를 환자의 분자적 특징으로 설명하고, 실제 진단과 치료에 활용하기 위한 지표다.

바이오마커의 목적은 여러 종류가 있다. 어떤 바이오마커는 환자가 특정 질환이나 아형에 속하는지를 구분하고, 어떤 바이오마커는 앞으로의 질병 경과나 예후를 알려준다. 신약개발에서는 여기에 더해 특정 환자에게 어떤 치료가 더 효과적일지를 알려주는 치료 반응 예측 바이오마커가 중요하다. 진단, 예후, 치료 반응 예측 바이오마커는 서로 연결돼 있지만, 실제로 답하려는 질문과 검증 방법은 구분된다 [1].

질환을 구분하는 바이오마커는 “이 환자는 어떤 상태에 있는가”를 묻는다. 반면 치료 반응 예측 바이오마커는 “이 환자에게 어떤 치료가 더 적합한가”를 묻는다. 환자를 잘 구분하는 유전자가 치료 선택에도 도움이 될 수 있지만, 두 가지 기능이 항상 함께 나타나는 것은 아니다.

오늘날에는 한 명의 환자에서도 수천, 수만 개의 유전자와 다양한 오믹스 정보를 측정할 수 있다. 따라서 질환군에서 통계적으로 차이가 나는 유전자를 찾는 일은 과거보다 훨씬 쉬워졌다. 어려운 문제는 그 다음부터 시작된다. 수많은 후보 가운데 어떤 것이 질병 기전과 가까운지, 다른 환자에서도 재현되는지, 그리고 많은 시간과 비용을 들여 실제로 검증할 가치가 있는지를 판단해야 한다.

AI 기반 바이오마커 발굴이 기여할 수 있는 지점도 여기에 있다. 제한된 실험 자원을 고려해 어떤 후보를 다음 실험으로 가져갈 것인지 잘 판단하고, 실험에서 얻은 정보를 다음 후보 선정에 다시 활용하는 능력이 점점 중요해지고 있다.

변이에서 실제로 작동하는 생물학을 읽어내기

바이오마커라고 하면 DNA 변이를 먼저 떠올리기 쉽다. DNA 변이는 질병의 원인을 추적하고 환자군을 구분하는 중요한 정보다. 다만 치료와 연결하려면 그 변이가 환자의 세포에서 어떤 결과를 만들어내는지도 함께 볼 필요가 있다.

HER2 가 좋은 예다. 일부 유방암에서는 ERBB2 유전자 증폭과 그에 따른 HER2 단백질 과발현이 실제 치료 선택의 중요한 근거가 된다. DNA 수준의 변화가 RNA 와 단백질, 세포 신호를 거쳐 질병 상태로 이어지고, 그 상태를 표적하는 약물이 존재하기 때문에 임상적으로 유용한 바이오마커로 자리 잡았다. 실제로 HER2 단백질 과발현이나 ERBB2 유전자 증폭을 확인하는 검사는 여러 HER2 표적치료제의 환자 선별에 사용된다 [2]. 이 사례는 유전자 변이를 발견한 이후에 무엇을 봐야 하는지를 잘 보여준다. 유전적 변화가 실제로 유전자 발현이나 단백질의 양과 활성을 바꾸는지, 어느 조직과 세포에서 그 변화가 나타나는지, 그리고 그 상태가 질병 표현형과 어떻게 연결되는지를 함께 살펴야 한다.

DNA 메틸화도 비슷하다 [3]. 같은 유전자 주변이라도 어느 조절 영역의 메틸화가 변했는지에 따라 의미가 달라질 수 있고, 서로 다른 세포가 섞인 조직에서는 동일한 측정값도 다른 생물학을 반영할 수 있다. RNA 와 단백질 역시 서로 다른 정보를 준다. RNA 발현량이 변해도 단백질의 양이나 활성은 유지될 수 있고, 반대로 단백질 수준의 중요한 변화가 RNA 에서는 뚜렷하게 드러나지 않을 수도 있다. 유전변이, DNA 메틸화, 전사체, 단백질체는 질병의 서로 다른 분자 층을 보여준다. 이들을 함께 볼 때 한 층에서 관찰된 변화가 실제 질병 과정과 어떻게 연결되는지를 더 잘 해석할 수 있다. AI 역시 어느 한 층에서 가장 크게 변한 유전자를 고르는 데 그치기보다, 여러 분자 층의 변화가 환자의 질병 상태와 어떤 관계를 이루는지를 찾는 방향으로 발전할 필요가 있다.

좋은 바이오마커 후보를 보는 세 가지 방향

좋은 바이오마커 후보를 하나의 점수만으로 정의하기는 어렵다. 환자군과 대조군 사이에서 발현 차이가 가장 큰 유전자가 반드시 질병 기전의 중심에 있다고 볼 수 없으며, 가장 쉽게 측정되는 후보가 반드시 치료 선택에 유용한 것도 아니다. 예를 들어, 일부 암에서는 특정 유전적 변화와 그 변화가 만드는 분자 상태, 그리고 이를 표적하는 약물을 비교적 직접적으로 연결할 수 있다. 반면 하나의 지배적인 변이로 설명되지 않는 복합질환에서는 질병 신호가 여러 유전자와 세포, 환경 요인에 분산되어 있다. 천식은 같은 진단 안에서도 염증의 형태, 관련 세포, 중증도와 치료 반응이 크게 다른 대표적인 이질적 질환이다.

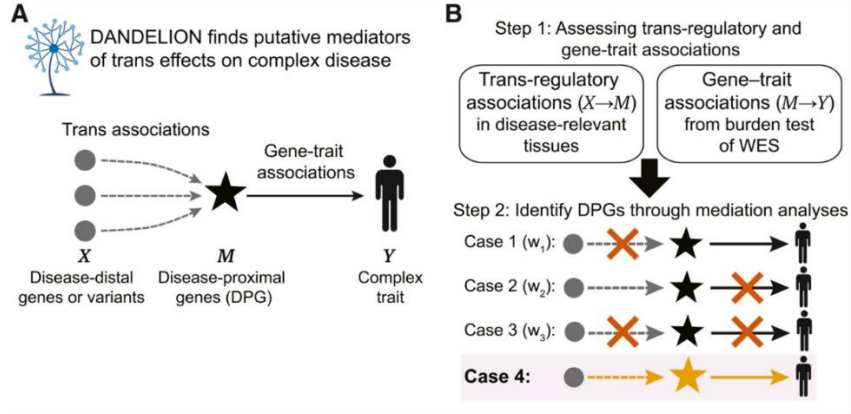
최근 Cell 에 발표된 DANDELION [4]과 필자가 대표로 있는 A 사에서 개발한 GOAT [5] 및 MIL²Het 은 모두 천식을 중요한 연구 또는 평가 대상으로 삼았지만, 서로 다른 방향으로 접근한다. DANDELION 은 질병 기전에 가까운 유전자를 우선순위화하고, GOAT 는 멀티오믹스 정보와 유전자 관계를 이용해 환자 아형을 설명하며, MIL²Het 은 단일세포 수준의 이질성을 환자 수준의 질병 상태와 연결한다. 세 접근을 함께 보면 좋은 바이오마커 후보를 만들기 위해 어떤 종류의 근거가 필요한지를 살펴볼 수 있다 (그림 1).

방향 1. 질병 기전에 가까운 유전자를 찾다: DANDELION

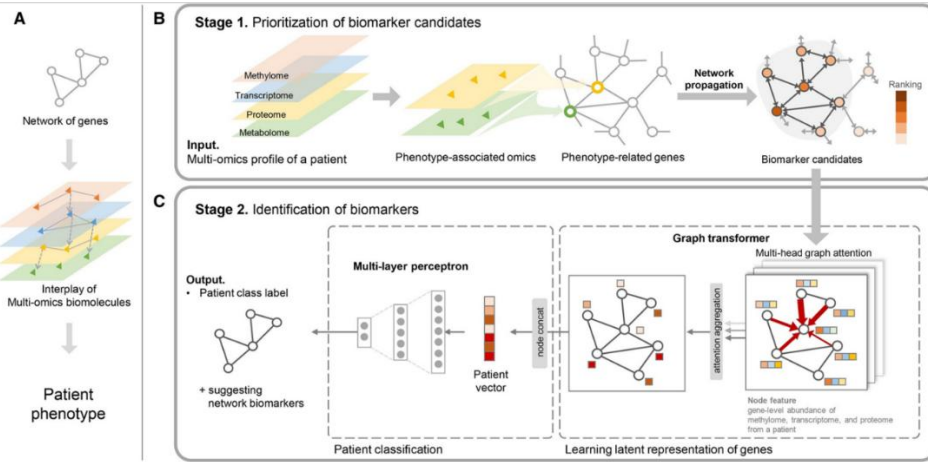
유전변이의 원거리 조절 효과까지 추적해 실제 질병 기전에 가까운 유전자(disease-proximal gene)를 우선순위화한다. 후보는 CRISPR 와 동물모델을 통해 기능적으로 검증했다.

DANDELION 은 질환과 연관된 유전적 신호 가운데 실제 질병 기전에 더 가까운 유전자를 찾는 연구다. 유전적 변이가 인접한 유전자에만 영향을 미치는 것이 아니라, 유전자 조절망을 통해 멀리 떨어진 유전자의

1. DANDELION



2. GOAT



3. MIL²Het

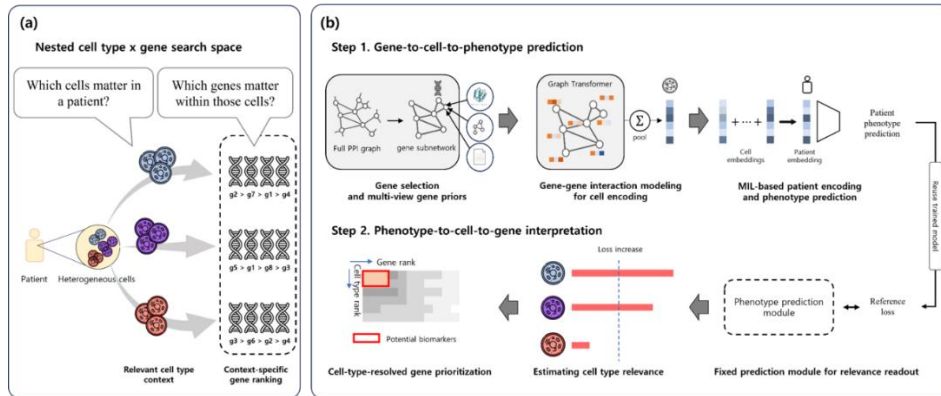


그림 1. AI 기반 바이오마커 후보 발굴을 위한 세 가지 모델링 전략.

상단부터 DANDELION [4], GOAT [5], MIL²Het의 개념적 구조를 나타낸다. DANDELION은 유전변이의 원거리 조절 효과를 추적해 질병 기전에 가까운 disease-proximal gene을 우선순위화하고, GOAT는 멀티오믹스 정보와 유전자 상호작용 네트워크를 그래프 AI로 함께 학습해 환자 아형을 설명하는 바이오마커 후보를 발굴한다. MIL²Het은 단일세포 데이터를 유전자-세포-환자 계층으로 학습하여 세포 수준의 이질성을 환자 수준의 질병 상태와 연결한다. 세 접근은 각각 질병 기전 근접성, 유전자 네트워크 맥락, 세포 수준 이질성이라는 서로 다른 근거를 이용해 바이오마커 후보를 우선순위화한다.

발현에도 영향을 미칠 수 있다는 점을 이용한다. DANDELION 은 이러한 원거리 조절 효과와 전장 엑솜 시퀀스에서 얻은 유전자 수준의 정보를 결합해 질병 근접 유전자, 즉 disease-proximal gene 을 우선순위화한다 [4].

엄밀히 말하면 DANDELION 은 임상적인 치료 반응 바이오마커를 직접 개발한 연구라기보다, 질병을 실제로 구동할 가능성이 높은 유전자와 치료표적 후보를 찾는 방법이다. 그러나 어떤 유전자가 질환과 통계적으로 연관돼 있다는 사실을 넘어, 그 유전자가 질병 과정에 기능적으로 관여하는지를 확인했다는 점에서 바이오마커 후보 발굴에도 중요한 시사점을 준다.

이 연구에서 특히 주목할 부분은 계산 결과 이후의 검증이다. 연구진은 천식에서 찾은 후보를 사람의 기도 상피세포와 T 세포에서 CRISPR 로 검증하고, 생화학적 분석과 알레르기성 천식 동물 모델까지 연결했다. 그 결과 SLC27A3 와 SCD 를 포함한 후보가 천식 관련 세포 표현형과 염증, 기도 재형성에 영향을 줄 수 있다는 근거를 제시했다 [4].

DANDELION 은 질환과 통계적으로 관련된 신호를 실제 질병 기전에 가까운 후보로 좁혀가는 과정을 보여준다. 바이오마커 후보를 고를 때 연관성의 크기뿐 아니라, 그 후보가 어떤 생물학적 과정을 통해 질병으로 이어지는지를 함께 고려해야 한다는 점을 보여주는 사례다.

방향 2. 발현량 너머의 유전자 관계를 읽다: GOAT

멀티오믹스 정보와 유전자 상호작용 네트워크를 그래프 AI 로 함께 학습해, 개별 발현량 차이만으로는 드러나지 않는 유전자와 유전자 관계까지 바이오마커 후보로 발굴한다.

A 사에서 개발한 멀티오믹스 바이오마커 발굴 기술인 GOAT [5]는 국내 COREA 천식 코호트의 환자 300 명에게서 얻은 메틸롬, 전사체, 단백질체, 대사체 정보를 이용해 호산구성 천식과 비호산구성 천식을 구분하고, 두 아형을 설명할 수 있는 유전자 수준의 바이오마커를 찾았다 [5].

GOAT 의 특징은 유전자 하나의 발현량만 보는 것이 아니라, 유전자들이 서로 맺고 있는 관계를 함께 분석한다는 점이다. 단백질체와 대사체에서 질환 아형과 관련된 신호를 찾고, 이를 유전자 상호작용 네트워크에 연결해 후보를 좁힌다. 이후 환자별 멀티오믹스 측정값과 유전자 관계망을 그래프 신경망으로 학습해 환자 아형을 구분하는 데 중요한 유전자와 유전자 관계를 찾는다.

CTNNB 과 JUN 이 대표적인 사례다. 두 유전자는 개별 전사체 발현량만 비교했을 때 두 천식 아형 사이에서 뚜렷한 차이를 보이지 않았다. 그러나 다른 유전자들과의 관계와 조절 네트워크를 함께 고려하자 호산구성 천식의 병태생리를 설명하는 중요한 위치에 있는 것으로 나타났다 [5]. 즉, 같은 발현량이라도 어떤 생물학적 관계 안에서 해석하느냐에 따라 얻을 수 있는 정보가 달라질 수 있다는 뜻이다. 바이오마커 후보를 찾을 때 변화의 크기와 함께, 그 변화가 유전자 네트워크에서 어떤 역할을 하는지도 살펴봐야 한다.

방향 3. 환자 안의 세포 이질성을 읽다: MIL²Het

단일세포 데이터를 유전자→세포→환자 계층으로 학습해 환자의 질병 상태와 치료 반응을 예측하고, 다시 역으로 어떤 세포 유형의 어떤 유전자 프로그램이 중요한지 우선순위화한다.

한 환자의 조직은 하나의 평균적인 전사체만으로 설명하기 어렵다. 조직 안에는 서로 다른 종류와 상태의 세포가 함께 존재하며, 중요한 질병 신호가 특정 세포 집단에만 집중될 수도 있다. 이런 신호는 조직 전체의 평균을 계산하는 과정에서 약해지거나 사라질 수 있다.

A사에서 개발한 MIL²Het은 이러한 단일세포 이질성을 환자 수준의 질병 상태와 연결하는 바이오마커 발굴 기술이다. 한 환자에게서 측정한 여러 단일세포를 개별 관측치로 풀어놓지 않고, 하나의 환자 안에 존재하는 세포 집단으로 해석한다. 이를 통해 환자 수준의 질병 상태를 예측하는 동시에, 그 예측에 어떤 세포 유형과 유전자 신호가 중요했는지를 추적한다.

MIL²Het은 특정 유전자가 중요하다는 사실에서 멈추지 않고, 그 유전자가 어느 세포에서 중요하며 그 세포의 상태가 환자 전체의 질병 상태와 어떻게 연결되는지를 모델링한다. 이는 조직 평균에서는 보이지 않던 환자별 세포 구성과 세포 상태를 바이오마커로 활용할 가능성을 연다.

DANDELION, GOAT, MIL²Het은 서로 다른 종류의 근거를 제공한다. 세 연구를 함께 보면 바이오마커 후보를 평가할 때 질병 기전에 얼마나 가까운지, 환자 수준의 표현형을 얼마나 잘 설명하는지, 어느 세포에서 작동하는지를 함께 물어야 한다는 점이 드러난다. 치료 선택까지 연결하려면 여기에 한 가지 질문이 더해진다. 바로 그 바이오마커의 상태에 따라 약물의 효과가 실제로 달라지는가이다.

바이오마커 검증에서 환자 수준 약물 반응 예측으로

앞서 살펴보았듯이 다중오믹스 데이터와 인공지능의 발전은 이러한 환자 간 차이를 설명할 수 있는 바이오마커 후보를 보다 정교하게 발굴할 수 있는 기반을 제공하고 있다. 발견된 분자적 특징이 실제 치료 선택으로 이어지기 위해서는 특정 바이오마커를 가진 환자가 어떤 치료에 반응할 것인지를 확인해야 한다. 이러한 관점에서 인공지능 기반 약물 반응 예측(drug response prediction)은 바이오마커 발굴과 정밀의료의 실제 치료 전략을 연결하는 중요한 기술이라 할 수 있다.

한편 약물 반응을 검증하기 위한 실험 기술 역시 지속적으로 발전하고 있다. 환자 유래 오가노이드(patient-derived organoid, PDO)는 종양의 3차원 구조와 이질성을 보다 충실하게 반영할 수 있으며, 실제 환자의 치료 반응을 재현하기 위한 전임상 모델로 활용되고 있다 [6]. 환자 유래 이종이식 모델(patient-derived xenograft, PDX) 역시 생체 내 환경에서 다양한 후보 치료제의 반응을 평가하고 임상적 약효와의 연관성을 탐색할 수 있는 중요한 모델이다 [7]. 이러한 환자 유래 모델은 정밀医료를 위한 중요한 검증 수단이며 앞으로도 그 활용이 확대될 것으로 기대된다. 다만 환자별 모델을 구축하여 다양한 바이오마커와 후보 약물의 조합을 모두 실험적으로 탐색하기에는 시간, 비용 및 처리량 측면에서 현실적인 제약이 있다.

따라서 계산 모델과 환자 유래 실험 모델은 서로를 대체하기보다, AI 를 통해 넓은 환자-약물 탐색 공간에서 유망한 후보를 먼저 선별하고 이를 PDO 나 PDX 에서 정밀하게 검증하는 상호보완적인 관계로 활용될 필요가 있다.

세포 수준 예측에서 환자 수준 약물 반응 예측으로

AI 기반 약물 반응 예측의 발전에는 GDSC(Genomics of Drug Sensitivity in Cancer), CCLE(Cancer Cell Line Encyclopedia)와 같은 대규모 암 세포주 데이터가 중요한 역할을 해왔다 [8,9]. 세포주에서는 다양한 약물을 동일한 조건에서 처리하고 농도별 반응을 반복적으로 측정할 수 있기 때문에, 유전자 발현이나 돌연변이와 같은 분자적 특징과 약물 감수성 사이의 관계를 학습하기 위한 풍부한 데이터를 확보할 수 있다. 최근에는 LINCS L1000 과 같이 약물 처리에 따른 유전자 발현 변화를 대규모로 측정한 perturbation 데이터까지 활용되면서, 단순한 약물 감수성을 넘어 약물이 세포의 분자 상태를 어떠한 방향으로 변화시키는지까지 모델링할 수 있는 기반이 마련되고 있다 [10].

그러나 정밀의료에서 궁극적으로 필요한 질문은 “이 세포주가 이 약물에 반응할 것인가”가 아니라 “이 환자가 이 치료에 반응할 것인가”이다.

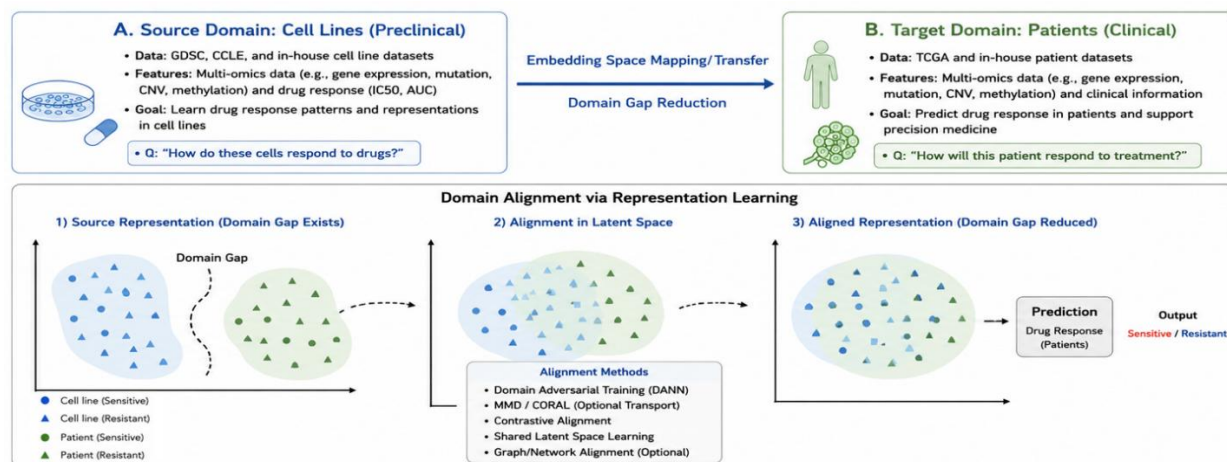


그림 2. 세포주에서 환자로의 약물 반응 지식 이전을 위한 임베딩 공간 정렬 개념도.

대규모 약물 반응 데이터가 축적된 세포주는 풍부한 학습 정보를 제공하는 source domain 인 반면, 실제 환자 데이터는 치료 반응 레이블이 제한적이고 종양 미세환경, 세포 구성 및 유전자 조절 상태 등의 차이로 인해 세포주와 서로 다른 데이터 분포를 형성한다. 이러한 전임상-환자 간 domain gap 을 극복하기 위해 세포주와 환자의 분자적 표현을 공통 임베딩 공간으로 매핑하고 정렬함으로써, 세포주에서 학습한 약물 반응 지식을 환자 수준으로 이전할 수 있다. 정렬된 임베딩 공간에서는 세포주와 환자의 생물학적으로 대응되는 표현 간 차이를 줄이면서 약물 반응과 관련된 정보를 보존하여, 제한된 환자 데이터에서도 환자 수준의 치료 반응을 예측하는 것을 목표로 한다.

이 두 문제 사이에는 여전히 상당한 간극이 존재한다. 세포주에서는 수많은 약물을 반복적으로 시험할 수 있지만, 실제 환자에게 다양한 치료제를 체계적으로 투여하여 각각의 반응을 확인하는 것은 불가능하다. 또한 세포주와 실제 종양은 종양 미세환경, 세포 구성, 유전자 조절 상태 등에서 차이를 보이며, 실제 종양 자체도 다양한 세포 집단이 공존하는 이질적인 생물학적 시스템이다.

결국 환자 수준 약물 반응 예측에서 중요한 것은 단순히 더 복잡한 인공지능 모델을 개발하는 데 있지 않다. 보다 본질적인 목표는 전임상 실험에서 축적된 풍부한 약물 반응 지식을 실제 환자의 생물학적 맥락으로 이전하고, 환자마다 다른 치료 반응을 예측하는 것이다 [11-14] (그림 2). 이하에서는 이러한 전임상-임상 간 간극을 줄이고 세포 수준의 약물 반응 지식을 환자 수준으로 확장하기 위한 세 가지 연구 방향을, A 사에서 개발한 모델들을 사례로 살펴본다.

방향 1. 전임상 데이터를 환자의 언어로 번역하기

초기의 인공지능 기반 약물 반응 예측은 주로 세포주의 유전자 발현과 약물의 특성을 이용하여 약물 감수성을 예측하는 문제로 발전해 왔다. 그러나 세포주에서 학습한 모델을 실제 환자에게 그대로 적용하기에는 전임상 모델과 환자 사이의 생물학적 차이, 즉 domain gap 이 존재한다.

최근에는 이러한 차이를 명시적으로 고려하면서 세포 수준에서 축적된 약물 반응 지식을 환자에게 이전하려는 연구가 이루어지고 있다 [11-14]. 예를 들어 THERAPI(그림 3a)는 하나의 세포주를 특정 환자의 대리 모델로 간주하는 대신, 여러 세포주의 정보를 환자의 분자적 특성에 따라 조합하여 종양의 이질성을 표현하고, 세포 및 유전자 수준에서 학습한 약물 반응 지식을 환자 수준의 치료 반응 예측으로 이전한다 [15].

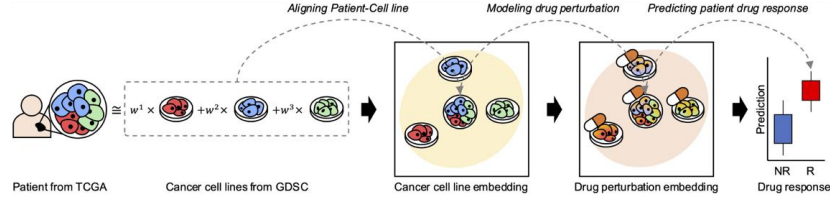
여기서 중요한 것은 세포주에서 관찰된 반응을 환자에게 그대로 적용하는 것이 아니라, 전임상 데이터에 축적된 약물 반응 지식을 환자의 생물학적 특성에 맞게 번역(Cell-to-patient transfer)한다는 관점이다. 이는 제한적인 환자 데이터를 풍부한 전임상 데이터로 보완함으로써 환자 수준 약물 반응 예측의 가능성을 확장한다.

방향 2. 환자의 현재 상태에서 약물에 의한 변화까지

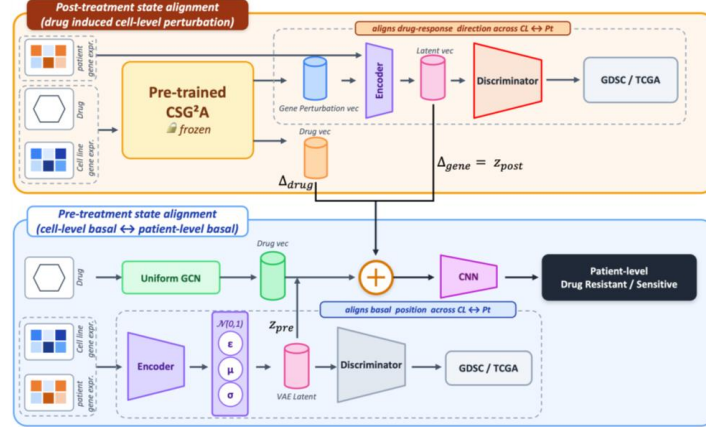
환자의 치료 전 기저 상태만으로 약물 반응을 완전히 예측하기 어렵다. 치료 효과는 환자가 치료 전에 어떤 상태에 있었는지뿐 아니라, 약물을 투여했을 때 그 상태가 어떻게 변화하는지도 고려되어야 하기 때문이다.

이러한 관점에서 세포주의 대규모 섭동(perturbation) 데이터는 실제 환자에서 관찰하기 어려운 치료 후 분자 상태에 대한 정보를 보완할 수 있다. 다양한 세포와 약물에서 측정된 전사체 변화를 이용하면 약물이 생물학적 시스템을 어떠한 방향으로 변화시키는지에 대한 일반적인 패턴을 학습할 수 있다. 예를 들어 CSG²A 는 LINCS L1000 에서 약물에 의해 유도되는 condition-specific gene-gene interaction 을 학습한

(a) THERAPI: Tumor Heterogeneity-aware Embedding for Response Adaptation and Patient Inference



(b) PANDA: Patient-level Adversarial Network for Dual Alignment of pre- and post-treatment transcriptomic states



(c) PREDIKTOR: Personalized drug Response Estimation via Dual Integration of Knowledge graph and Transferable Omics-driven Representations

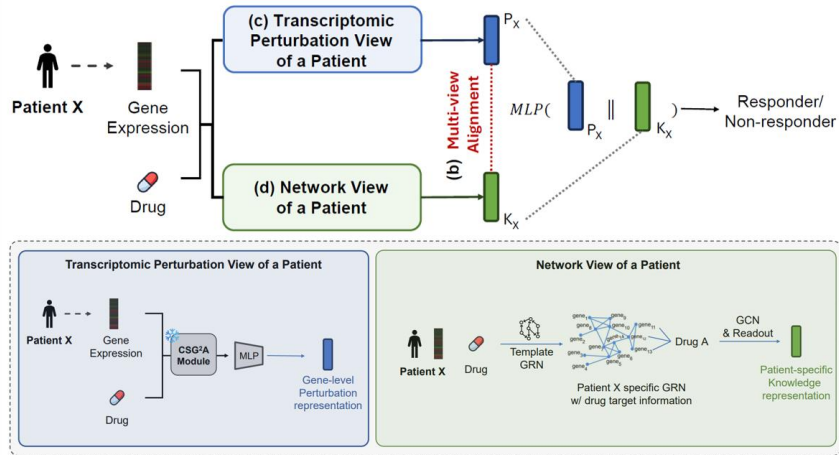


그림 3. 환자 수준 약물 반응 예측을 위한 전임상-환자 정보 통합 및 모델링 전략.

상단부터 각각 THERAPI [15], PANDA [17], PREDIKTOR [18]의 개념적 구조를 나타낸다. THERAPI 는 여러 암 세포주의 정보를 환자 특성에 맞게 통합하여 세포 수준에서 학습된 약물 반응 지식을 환자 수준으로 이전한다. PANDA 는 세포주와 환자의 치료 전 상태뿐 아니라 약물 처리에 의해 유도되는 전사체 상태 변화까지 정렬하여 환자 수준의 약물 반응 예측에 활용한다. PREDIKTOR 는 환자의 전사체 기반 perturbation view 와 환자 특이적 유전자 조절 네트워크 및 drug-target 관계로 구성된 network view 를 함께 학습하여, 약물의 작용을 환자별 분자적 맥락에서 해석하고 치료 반응을 예측한다. 세 모델은 각각 전임상 지식의 환자 이전, 약물 유도 상태 변화의 정렬, 환자 특이적 분자 네트워크의 통합이라는 서로 다른 관점에서 환자 수준 약물 반응 예측을 고도화한다.

뒤 이를 GDSC 의 세포 수준 약물 반응 예측으로 이전함으로써, gene-level perturbation 과 cell-level drug response 를 연결할 수 있음을 보여주었다 [16].

이러한 섭동 모델링을 환자 수준으로 확장한 PANDA(그림 3b)는 세포주와 환자의 치료 전 전사체 상태를 정렬하는 동시에, 약물 처리에 의해 유도되는 변화된 상태까지 환자 수준으로 정렬하여 활용한다. 핵심은 약물이 현재의 환자 상태를 어떠한 방향으로 변화시킬 가능성이 있는지를 치료 반응 예측에 반영하는 것이다 [17].

이에 따라 약물 반응 예측의 질문도 단순히 “이 환자가 이 약물에 반응할 것인가”에서, “이 환자의 분자 상태에 이 약물을 적용하면 어떠한 변화가 발생하며 그 변화가 치료 효과와 어떻게 연결되는가”라는 보다 동적인 문제로 확장되고 있다.

방향 3. 유전자 발현에서 환자 특이적 분자 네트워크로

약물 반응을 보다 환자 특이적으로 설명하기 위해서는 유전자 발현량뿐 아니라 유전자들이 어떠한 관계를 형성하고 있는지도 고려할 필요가 있다. 생물학적 시스템에서 유전자는 독립적으로 작동하지 않으며, 동일한 수준으로 발현된 유전자도 주변 유전자와의 조절 관계나 전체 네트워크에서의 위치에 따라 서로 다른 기능적 의미를 가질 수 있다. 따라서 최근에는 환자의 분자적 특성을 단순한 유전자 발현 벡터로 표현하는 것을 넘어 환자별 분자 네트워크 자체를 모델링하려는 접근으로 연구가 확장되고 있다.

PREDIKTOR(그림 3c)는 환자의 치료 전 전사체로부터 patient-specific gene regulatory network 를 구성하고 여기에 drug-target 관계를 연결함으로써, 특정 약물이 각 환자의 분자 시스템에서 어떠한 맥락으로 작용할 수 있는지를 표현한다. 이를 약물에 의해 유도되는 전사체 변화 정보와 함께 활용하여 치료 반응을 예측한다 [16,18]. 핵심은 약물의 작용을 환자마다 다른 molecular network context 안에서 해석한다는 점이다. 한쪽에서는 “이 환자의 분자 시스템이 어떻게 구성되어 있는가”를 보고, 다른 한쪽에서는 “이 약물이 그 시스템에 어떠한 변화를 유도할 수 있는가”를 보면서 두 정보를 하나의 치료 반응 예측으로 연결한다.

이러한 접근은 약물 반응 예측을 단순한 통계적 상관관계에서 보다 환자 특이적이고 기전적인 예측으로 확장한다. 또한 예측에 기여한 유전자와 경로, 약물 표적을 함께 분석할 수 있다는 점에서 치료 반응을 설명하는 새로운 바이오마커와 작용기전 가설을 발굴 및 검증하는 데에도 활용될 수 있다.

AI 기반 약물 반응 예측은 단순히 세포주에서 학습한 반응을 환자에게 적용하는 기술에서 벗어나, 전임상 데이터로부터 환자의 이질성을 표현하고, 약물에 의한 상태 변화를 반영하며, 환자마다 다른 분자 네트워크의 맥락까지 고려하는 방향으로 발전하고 있다. 이처럼 서로 다른 접근법은 환자의 치료 반응을 설명하는 다양한 정보를 제공하지만, 높은 환자 이질성과 제한된 임상 데이터로 인해 하나의 모델만으로

모든 상황에서 안정적인 예측 성능을 확보하기는 쉽지 않다. 이에 실제 활용 단계에서는 서로 다른 모델의 예측 정보를 앙상블하여 예측의 안정성과 일반화 성능을 높이는 통합 전략이 필요하다. 필자가 대표로 있는 A 사에서도 이러한 전략을 활용하여 보다 안정적인 환자 수준의 약물 반응 예측을 제공하는 DrugVLAB™ 서비스를 운영하고 있다 [19].

마치며: 검증 결과가 다시 모델을 가르친다

데이터 규모와 모델 성능만으로 좋은 바이오마커가 자동으로 만들어지는 것은 아니다. AI 가 후보를 제시한 뒤에는 어떤 후보를 어떤 실험으로 검증할지 결정하고, 그 결과를 다시 다음 후보 선별에 활용하는 과정이 필요하다. 결국 중요한 것은 더 많은 후보를 제시하는 능력뿐 아니라, 실험을 통해 새롭게 얻은 정보를 다음 의사결정으로 효과적으로 반영하는 것이다.

이러한 관점에서 Doctor-in-the-Loop 는 환자 데이터, AI 예측, 전문가의 실험 선택과 검증 결과를 하나의 폐쇄형 순환으로 연결한다. 초기 환자 데이터를 바탕으로 AI 가 약물 반응을 예측하고 바이오마커 후보를 우선순위화하면, 전문가는 필요한 검증을 설계하고, 새롭게 생성된 실험적 근거는 다시 모델에 반영되어 다음 라운드의 후보 선별을 개선한다 (그림 4).

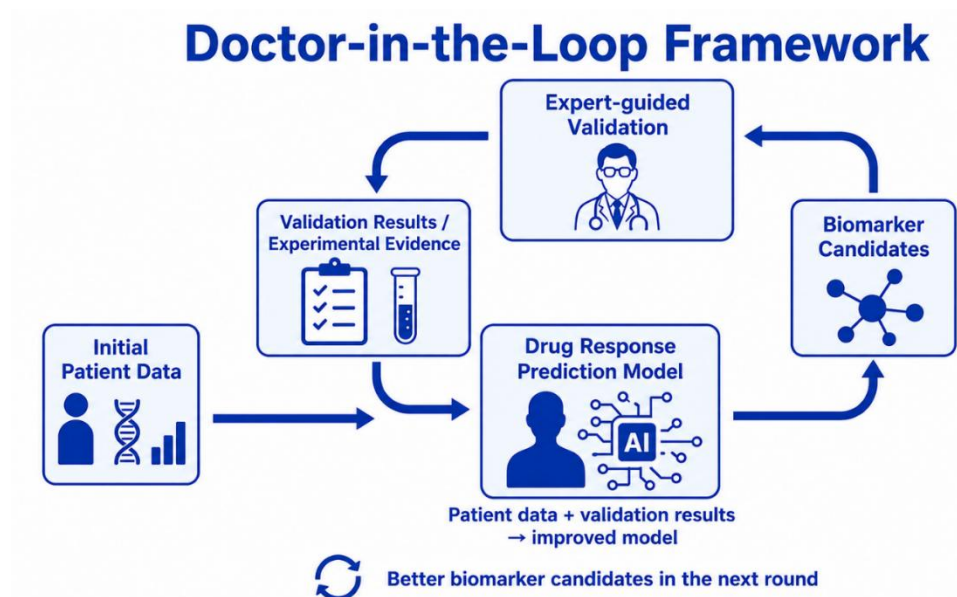


그림 4. Doctor-in-the-Loop 바이오마커 발굴 프레임워크.

초기 환자 데이터를 바탕으로 약물 반응 예측 모델이 바이오마커 후보를 선별하고, 전문가 주도의 검증에서 생성된 실험 결과가 다시 모델에 반영된다. 다음 라운드의 모델은 초기 환자 데이터와 축적된 검증 결과를 함께 활용함으로써 더 나은 바이오마커 후보를 선별한다.

여기서 Doctor 는 임상 의 한 사람만을 의미하지 않는다. 질환 생물학자, 약리학자, 실험 연구자와 중개연구 전문가를 포함하며, 이들의 역할은 AI 가 제시한 후보를 단순히 승인하는 데 있지 않다. 질병 기전, 관련 세포, 측정 및 조절 가능성, 안전성, 실험 비용 등을 함께 고려해 어떤 실험이 현재의 불확실성을 가장 효과적으로 줄일 수 있는지 판단하는 것이 핵심이다.

이 과정에서는 효과가 확인된 양성 결과뿐 아니라 효과가 없었던 음성 결과, 특정 조건에서만 나타난 효과와 예상과 다른 결과 역시 중요한 데이터가 된다. 검증 대상, 세포 유형, 약물 농도와 측정 시점 등의 조건과 함께 이러한 결과를 구조화하면 모델은 후보의 의미와 우선순위를 지속적으로 갱신할 수 있다. 이는 정보가치가 높은 다음 실험을 선택하고 새로운 결과로 모델을 업데이트하는 능동학습(active learning)의 개념과도 맞닿아 있다 [20].

이를 실제 신약개발 과정에 구현하기 위해서는 치료 정보와 임상 결과가 연결된 환자 데이터를 지속적으로 축적하고, 성공과 실패를 모두 포함한 실험 결과를 재사용 가능한 형태로 구조화하며, AI 예측-전문가 주도 검증-모델 갱신이 반복되는 검증 파이프라인을 구축할 필요가 있다. 특히 개별 연구에서 생성된 결과가 일회성으로 소모되지 않고 다음 모델과 다음 실험 설계에 다시 활용될 수 있는 데이터 체계를 만드는 것이 중요하다.

이러한 기반이 마련된다면 AI 는 단순히 후보를 많이 제시하는 예측 도구를 넘어, 어떤 후보를 먼저 검증하고 어떤 환자군에서 어떤 치료를 우선적으로 탐색할 것인지를 함께 설계하는 의사결정 파트너로 발전할 수 있다. 환자 데이터, AI, 전문가의 판단과 실험적 검증이 지속적으로 연결되는 이러한 순환 구조가 정착될 때, 제한된 검증 자원으로 더 빠르고 확실하게 유망한 바이오마커와 치료 전략에 도달하고, 궁극적으로 데이터 기반 신약개발이 정밀의료로 이어지는 속도를 앞당길 수 있을 것이다.

References

1. FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource. NCBI Bookshelf.
2. National Cancer Institute. Breast Cancer Treatment (PDQ®); U.S. Food and Drug Administration. List of FDA-Authorized Companion Diagnostic Devices.
3. Lee D, Koo B, Kim S, Byun JM, Hong J, Shin D-Y, Sun C-H, Kim J, Song J-J, Jaiswal S, Yoon S-S, Kim S, Koh Y. Increased local DNA methylation disorder in AMLs with DNMT3A-destabilizing variants and its clinical implication. *Nature Communications*. 2025;16:560. doi:10.1038/s41467-024-55691-z.
4. Salamone IM, Tian P, Qi Z, et al. Trans-regulatory gene mapping prioritizes disease drivers in asthma. *Cell*. 2026;189(17):5250–5265.e11. doi:10.1016/j.cell.2026.07.034.
5. Jeong D, Koo B, Oh M, Kim TB, Kim S. GOAT: Gene-level biomarker discovery from multi-Omics data using graph ATtention neural network for eosinophilic asthma subtype. *Bioinformatics*. 2023;39(10):btad582. doi:10.1093/bioinformatics/btad582.
6. Vlachogiannis G, Hedayat S, Vatsiou A, et al. Patient-derived organoids model treatment response of metastatic gastrointestinal cancers. *Science*. 2018;359(6378):920–926. doi:10.1126/science.aao2774.
7. Gao H, Korn JM, Ferretti S, et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nature Medicine*. 2015;21:1318–1325. doi:10.1038/nm.3954.
8. Garnett MJ, Edelman EJ, Heidorn SJ, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*. 2012;483:570–575. doi:10.1038/nature11005.
9. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;483:603–607. doi:10.1038/nature11003.
10. Subramanian A, Narayan R, Corsello SM, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*. 2017;171(6):1437–1452.e17. doi:10.1016/j.cell.2017.10.049.
11. He D, Liu Q, Wu Y, Xie L. A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening. *Nature Machine Intelligence*. 2022;4:879–892. doi:10.1038/s42256-022-00541-0.
12. Sharifi-Noghabi H, Harjandi PA, Zolotareva O, Collins CC, Ester M. Out-of-distribution generalization from labelled and unlabelled gene expression data for drug response prediction. *Nature Machine Intelligence*. 2021;3:962–972. doi:10.1038/s42256-021-00408-w.
13. Kim J, Park SH, Lee H. PANCDR: precise medicine prediction using an adversarial network for cancer drug response. *Briefings in Bioinformatics*. 2024;25(2):bbae088. doi:10.1093/bib/bbae088.
14. Li K, Shen B, Feng F, et al. A disentangled generative model for improved drug response prediction in patients via sample synthesis. *Journal of Pharmaceutical Analysis*. 2025;15(6):101128. doi:10.1016/j.jpha.2024.101128.
15. Sung I, Bang D, Kim S, Lee S. Deep learning for predicting patient drug response by transferring gene-level and cell-level knowledge to tumors. *npj Precision Oncology*. 2026;10:310. doi:10.1038/s41698-026-01491-1.

16. Bang D, Koo B, Kim S. Transfer learning of condition-specific perturbation in gene interactions improves drug response prediction. *Bioinformatics*. 2024;40(Suppl 1):i130–i139. doi:10.1093/bioinformatics/btae249.
17. An S, Cho C, Kim S. Patient-level drug response prediction through dual adversarial alignment of pre- and post-drug treatment transcriptomic states. *isbAI 2026*. 2026 Aug 21.
18. Bang D, An S, Sung I, Yun I, Kim S, Lee S. Predicting Therapeutic Outcome via Aligning Patient-Specific Knowledge Graph and Gene-Level Perturbation Representations. *BioKDD 2026*; arXiv:2607.04557.
19. AIGENDRUG. DrugVLAB™: Integrated modular AI workspaces for biomarker discovery, drug candidate generation, response prediction, and toxicity evaluation. Available at: <https://drugvlab.com/>
20. Tosh C, Tec M, White JB, et al. A Bayesian active learning platform for scalable combination drug screens. *Nature Communications*. 2025;16:156. doi:10.1038/s41467-024-55287-7.