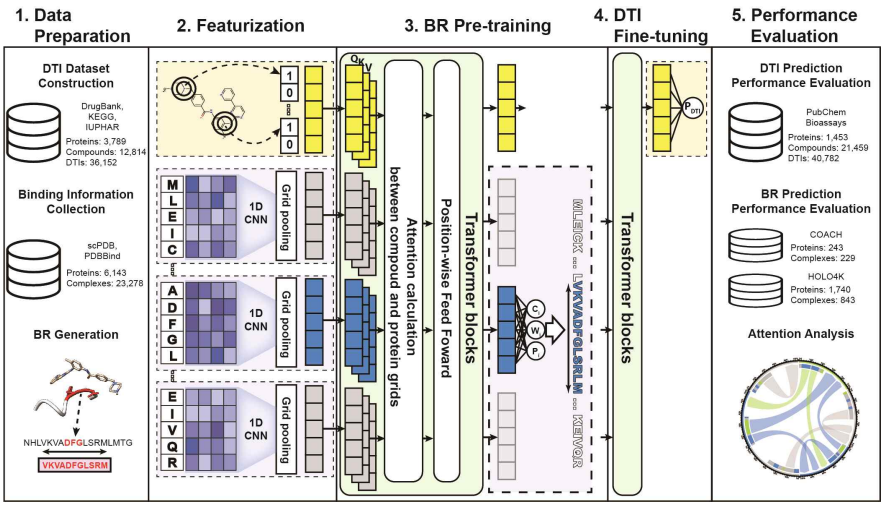
# 신약 후보물질 탐색 인공지능 모델 개발

- 단백질 3차원 구조 정보 없이 신약 개발이 가능한 인공지능 기술 개발


▲ 왼쪽부터 남호정 교수, 이인구 석박통합과정생

천문학적 시간과 비용을 필요로 하는 신약개발 산업은 인공지능 기술을 활용하여 혁명적 변화를 이끌 수 있는 산업으로 주목받고 있다. 인공지능을 이용하여 신약 후보물질 탐색 시간을 단축함으로써 결과적으로 신약개발에 소요되는 기간과 비용을 획기적으로 줄일 수 있다.

지스트(광주과학기술원, 총장 김기선) 전기전자컴퓨터공학부 남호정 교수 연구팀은 단백질 서열 기반으로 약물과 표적 단백질의 결합지역 및 상호작용을 예측(Highlights on Target Sequence, HoTS) 하는 인공지능 기술을 개발했다.




▲ HoTS 모델 개요. HoTS 모델의 학습 데이터셋, 모델 구조, 평가 및 분석 방법을 종합적으로 보여주고 있다.

신약개발의 초기 단계인 후보 물질 발굴단계는 표적 단백질에 활성을 보이는 화합물을 찾아내는 단계로써, 수만・수십만 개의 화합물로부터 표적 단백질에 활성을 보이는 화합물을 찾아야 하는 힘겨운 과정이다.

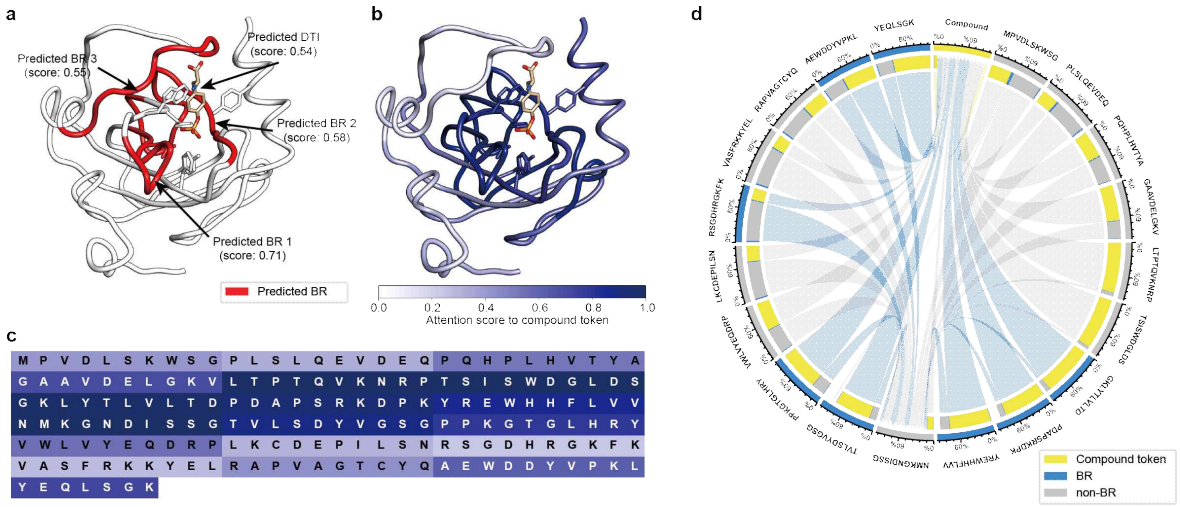
이러한 상황을 해결하기 위하여 다양한 약물-표적 단백질 상호작용 예측 인공지능 모델들이 개발됐지만, 좋은 예측 성능에도 불구하고 예측 결과에 대한 설명력이 부족하였기 때문에 실제 신약 개발에서 적극적인 도입이 꺼려져 왔다.

그러나, 이번에 연구팀이 개발한 모델인 HoTS는 약물-표적 단백질이 결합하는 부분을 사전학습한 후 예측하게 함으로써, 높은 예측 정확도와 함께 약물-표적 단백질 상호작용 예측의 근거도 함께 제시하여 신약개발 연구자들에게 보다 신뢰할 수 있는 유효화합물 예측 결과를 제시해 줄 수 있게 되었다.

본 연구는 대규모의 단백질 3차원 구조 데이터베이스로부터 화합물과의 결합지역을 추출하여 CNN(Convolutional Neural Network)과 트랜스포머(Transformer) 기반의 딥러닝 모델로 단백질 서열상의 결합지역을 예측할 수 있도록 학습되었다.

결합지역을 학습한 후, 해당 학습을 기반으로 하여 더 많은 트랜스포머 계층을 통해 약물-표적 단백질 상호작용을 예측할 수 있으며, 그 결과딥러닝 모델이 결합지역과 함께 약물-표적 상호작용을 예측할 수 있게 되었다.

결과적으로 HoTS 모델은 다른 딥러닝 모델들보다 더 높은 예측력을 보여주었으며, 결합지역 예측도 단백질 서열 정보만을 사용함에도 불구하고 3차원 구조 기반의 타 예측 모델과 비슷한 수준의 성능을 확인하였다.




▲ HoTS의 결합지역 예측과 트랜스포머의 Attention 분포. HoTS의 트랜스포머가 단백질의 결합지역을 중점적으로 고려하고 있음을 보여준다.

남호정 교수는 "본 연구성과는 신약 개발 단계 중 유효화합물 발굴의 효율성을 크게 높여주는 기술이며, 무엇보다 3차원 구조 정보가 없는 신규 표적 단백질에 대한 신약 개발의 가능성을 열어줬다는데 의의가 있다"면서 "향후 해당 모델을 통해 약 개발 단계에서의 빠르고 효율적인 유효화합물 발굴이 가능할 수 있을 것으로 기대된다"고 말했다.

지스트 남호정 교수팀이 수행한 이번 연구는 '설명가능 인공지능 기반 약물 후보의 독성 및 부작용 예측 시스템 개발'(한국연구재단 중견연구자지원사업), '지스트-전남대학교병원 공동연구과제', 'GRI(GIST 연구원) 생명노화연구소' 사업의 지원을 받아 수행되었으며, 'Journal of Cheminformatics'에 2월 8일자 온라인 게재되었다.

## 논문의 주요 내용

**1. 논문명, 저자정보**

- 저널명 : Journal of Cheminformatics IF 5.514 (20년 기준)
- 논문명 : Sequence-based prediction of protein binding regions and drug-target interactions
- 저자 정보 : 이인구 (제1저자, 전기전자컴퓨터공학부), 남호정 (교신저자, 전기전자컴퓨터공학부, AI대학원)

## 용 어 설 명

○ 약물-표적 상호작용: 약물은 세포의 표적 단백질과 결합함과 동시에 표적 단백질의 생물학적 작용을 제어함으로써 약물의 효과를 보이게 된다. 이러한 약물과 표적 단백질이 결합을 약물-표적 상호작용이라고 하며 이러한 결합을 확인하는 것이 약물개발의 첫 중요 단계이다.

○ Convolutional Neural Networks (CNN): CNN은 이미지 분석에 주요 사용되는 딥러닝 기법 중 하나, 인공신경망이 공간정보를 효율적으로 처리할 수 있도록 설계되었다. 특히 단백질 서열에 1D-CNN을 적용하였을 경우 단백질의 주요 모티프들을 잘 인지할 수 있다고 알려져 있다.

○ 트랜스포머 (Transformer): 트랜스포머는 자연어처리에 주로 사용되는 딥러닝 기법 중 하나이며, 문장같은 서열 등의 데이터를 처리할 때 각 단어 간의 연관성을 self-attention (자기주의 기법)을 통하여 모델링 한다. 따라서 여러 트랜스포머 계층을 거친 신경망 모델의 결과물은 문맥적인 의미를 더 잘 파악할 수 있게 된다.