



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년12월27일
(11) 등록번호 10-2618357
(24) 등록일자 2023년12월21일

(51) 국제특허분류(Int. Cl.)
G16B 40/20 (2019.01) G06F 40/279 (2020.01)
G06F 40/40 (2020.01) G06N 3/096 (2023.01)
G16B 15/30 (2019.01) G16B 30/10 (2019.01)
G16B 5/00 (2019.01) G16B 50/00 (2019.01)
G16C 20/70 (2019.01) G16C 20/90 (2019.01)
(52) CPC특허분류
G16B 40/20 (2019.02)
G06F 40/279 (2020.01)
(21) 출원번호 10-2022-0184891
(22) 출원일자 2022년12월26일
심사청구일자 2022년12월26일

(73) 특허권자
충남대학교산학협력단
대전광역시 유성구 대학로 99 (궁동, 충남대학교)
(72) 발명자
강현석
대전광역시 유성구 와룡로136번길 97, 303동 180
6호(봉산동)
구성우
대전광역시 유성구 온천북로33번길 22-4, 806호(봉명동, 신영 미소랑 도시형생활주택)
(뒷면에 계속)
(74) 대리인
특허법인다움

전체 청구항 수 : 총 6 항

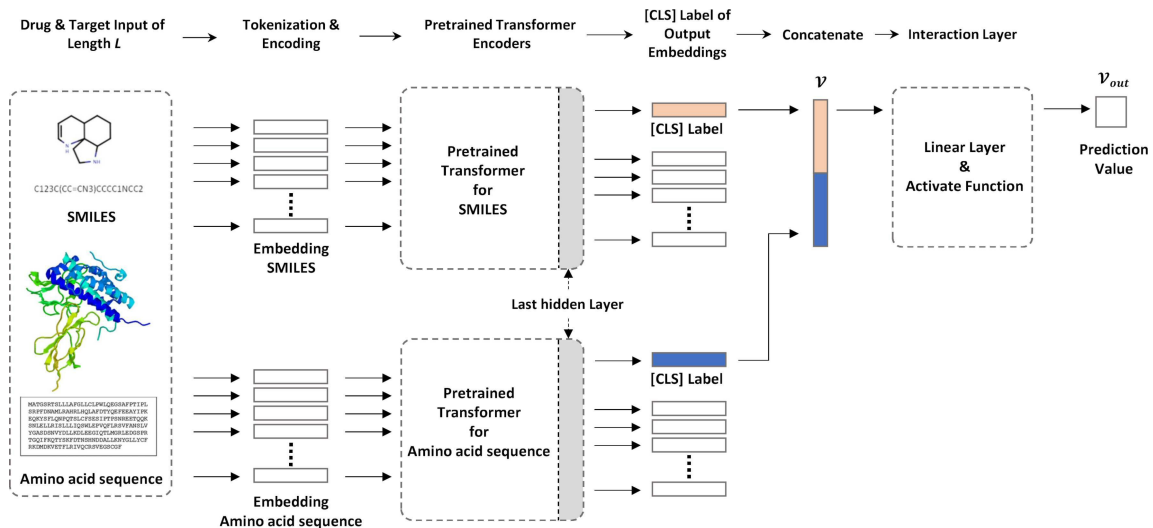
심사관 : 성경아

(54) 발명의 명칭 자연어 처리 모델을 이용한 약물-표적 상호작용 예측 방법

(57) 요약

본 발명은 약물-표적 상호작용 예측 방법에 관한 것이다. 본 발명은 방대한 데이터로 사전 학습된 자연어 처리 모델을 이용하여 약물과 단백질 간의 상호작용에 대한 복합적인 패턴을 학습할 수 있고, 실험이나 시뮬레이션으로 파악하기 전에 약물과 단백질이 상호작용을 하는지 여부를 예측할 수 있다. 이에 어떤 약물-단백질 쌍부터 실험이나 시뮬레이션을 진행할지 우선순위를 정할 수 있다. 또한, 본 발명은 약물 및 단백질의 물리화학적 구조에 대한 구체적인 정보가 없더라도, 통합 데이터로 학습되는 인공 신경망에 서열 정보만을 입력하여 약물-단백질 상호작용 여부를 예측할 수 있다. 이에 기존의 시뮬레이션 방식보다 훨씬 빠르고 정확하게 선도물질을 찾아낼 수 있고, 효율적으로 신약 개발을 가속화할 수 있다.

대표도 - 도1



(52) CPC특허분류

G06F 40/40 (2020.01)

G06N 3/096 (2023.01)

G16B 15/30 (2019.02)

G16B 30/10 (2019.02)

G16B 5/00 (2019.02)

G16B 50/00 (2019.02)

G16C 20/70 (2019.02)

G16C 20/90 (2019.02)

(72) 발명자

이현정

대전광역시 중구 대종로 713, 1동 904호(중촌동,
금호아파트)

윤휘열

대전광역시 유성구 죽동로 321, 104동 2103호(
죽동, 금성백조예미지)

정상근

대전광역시 서구 둔산북로 160, 102동 1502호(둔산
동, 한마루아파트)

채정우

대전광역시 유성구 봉명로 94, 701동 1302호(봉명
동, 도안신도시7단지예미지백조의호수)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711152822
 과제번호 2020-0-01441-003
 부처명 과학기술정보통신부
 과제관리(전문)기관명 정보통신기획평가원
 연구사업명 인공지능융합선도프로젝트
 연구과제명 인공지능융합연구센터지원(충남대학교)
 기여율 1/4
 과제수행기관명 충남대학교 산학협력단
 연구기간 2020.04.01 ~ 2022.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1711174038
 과제번호 00155857
 부처명 과학기술정보통신부
 과제관리(전문)기관명 정보통신기획평가원
 연구사업명 인공지능융합혁신인재양성
 연구과제명 인공지능융합혁신인재양성(충남대학교)
 기여율 1/4
 과제수행기관명 충남대학교 산학협력단
 연구기간 2022.07.01 ~ 2025.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1711154253
 과제번호 2022R1A2C1010929
 부처명 과학기술정보통신부
 과제관리(전문)기관명 한국연구재단
 연구사업명 중견연구(개인연구)
 연구과제명 전주기 신약개발 통합 인공지능 시스템 개발
 기여율 1/4
 과제수행기관명 충남대학교 산학협력단
 연구기간 2022.03.01 ~ 2026.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호 1711174955
 과제번호 2022R1A5A7085156
 부처명 과학기술정보통신부
 과제관리(전문)기관명 한국연구재단
 연구사업명 선도연구센터(CRC)지원사업
 연구과제명 라이프사이클 기반 시니어헬스 융합연구센터
 기여율 1/4
 과제수행기관명 충남대학교 산학협력단
 연구기간 2022.09.01 ~ 2029.02.28

명세서

청구범위

청구항 1

(1) 약물 및 타겟 단백질의 각각의 서열을 사전 학습된 자연어 처리 모델에 입력하고, 출력 임베딩을 출력하는 단계;

(2) 상기 자연어 처리 모델의 출력 임베딩 중 일부 토큰을 연결하여 벡터를 생성하는 단계; 및

(3) 상기 벡터를 인공 신경망 모델에 입력하고, 약물-표적 상호작용 확률을 출력하는 단계를 포함하고,

상기 인공 신경망 모델은,

서로 다른 데이터셋을 통합하여 생성하는 통합 데이터셋의 통합 학습 데이터 및 통합 검증 데이터로 학습하며,

상기 통합 데이터셋은,

서로 다른 데이터셋의 학습 데이터를 각각 병합(Merge) 및 혼합(Shuffle)하여 생성하는 통합 학습 데이터;

서로 다른 데이터셋의 검증 데이터를 각각 병합(Merge)하여 생성하는 통합 검증 데이터; 및

서로 다른 데이터셋의 시험 데이터를 모두 포함하는 통합 시험 데이터를 포함하고,

상기 (1) 단계는,

(1-1) 약물의 화합물 서열 및 타겟 단백질의 아미노산 서열을 입력 받는 단계; 및

(1-2) 상기 화합물 서열 및 상기 아미노산 서열을 토큰화하여 상기 화합물 서열의 토큰을 화합물 자연어 처리 모델에, 상기 아미노산 서열의 토큰을 단백질 자연어 처리 모델에 입력하고, 각각 하나 이상의 토큰을 포함하는 출력 임베딩을 출력하는 단계를 포함하며,

상기 (2) 단계는,

상기 각각의 출력 임베딩 중 서열에 대한 요약 정보를 가지는 각각의 첫 번째 토큰만을 연결하여 특징 벡터를 생성하는 것을 특징으로 하는, 자연어 처리 모델을 이용한 약물-표적 상호작용 예측 방법.

청구항 2

삭제

청구항 3

청구항 1에 있어서,

상기 화합물 자연어 처리 모델 및 상기 단백질 자연어 처리 모델은 BERT 계열 모델인 것을 특징으로 하는, 자연어 처리 모델을 이용한 약물-표적 상호작용 예측 방법.

청구항 4

삭제

청구항 5

삭제

청구항 6

삭제

청구항 7

삭제

청구항 8

청구항 1에 있어서,

상기 (3) 단계는,

(3-1) 벡터를 서로 다른 활성화 함수를 가지는 복수의 층을 포함하는 인공 신경망 모델에 입력하고, 출력 벡터를 출력하는 단계; 및

(3-2) 상기 출력 벡터에 활성화 함수를 적용하여 0 내지 1 사이의 실수값을 출력하는 단계를 포함하는 것을 특징으로 하는, 자연어 처리 모델을 이용한 약물-표적 상호작용 예측 방법.

청구항 9

청구항 8에 있어서,

상기 (3-1) 단계의 인공 신경망 모델은,

다음의 수학식 1 내지 수학식 3이 연산되는 은닉층을 포함하는 것을 특징으로 하는, 자연어 처리 모델을 이용한 약물-표적 상호작용 예측 방법.

[수학식 1]

$$v_1 = \text{ReLU}(\text{FCL}_1(v))$$

[수학식 2]

$$v_2 = \tanh(\text{FCL}_2(v_1))$$

[수학식 3]

$$v_{\text{out}} = \text{FCL}_3(v_2)$$

여기서, v , v_1 , v_2 , 및 v_{out} 은 벡터, FCL은 Fully Connected Layer(FCL) 연산, 그리고 ReLU 및 tanh는 활성화 함수임.

청구항 10

청구항 8에 있어서,

상기 (3-1) 단계는,

상기 인공 신경망의 임의의 노드를 미리 정해진 확률로 드롭하는 것을 특징으로 하는, 자연어 처리 모델을 이용한 약물-표적 상호작용 예측 방법.

청구항 11

명령어를 로드하는 메모리; 및

상기 명령어를 실행하는 프로세서를 포함하고,

상기 명령어는,

상기 프로세서가 청구항 1, 청구항 3, 청구항 8 내지 청구항 10 중 어느 한 청구항의 단계를 수행하도록 하는 것을 특징으로 하는, 자연어 처리 모델을 이용한 약물-표적 상호작용 예측 장치.

발명의 설명

기술 분야

본 발명은 약물-표적 상호작용 예측 방법에 관한 것이다.

[0001]

배경 기술

- [0002] 일반적으로 신약을 시판하기까지 약 10년 이상의 기간과 약 2조원 이상의 개발 비용이 투입된다고 알려져 있다. 신약 개발의 첫 단계는 발견 단계(Discovery)로서, 약 5천 내지 1만개의 후보 물질을 탐색하여 대상 질환의 치료에 효과를 보일 수 있는 유효 물질을 선별하는 단계이다.
- [0004] 발견 단계에서는 다종의 약물과 특정한 단백질들 간의 상호작용 여부를 실험으로 판정한다. 이 과정에서 단백질과 상호작용을 하는 약물들이 치료 약물의 후보로서 선도물질(Lead compound)이 된다. 이러한 발견 단계에만 약 5년 정도의 기간이 소요되고 있다. 따라서 선도물질을 빠르고 정확하게 찾아내는 것이 신약 개발의 첫 단계라고 할 수 있다.
- [0006] 선도물질을 찾아내기 위해 약물과 단백질 간의 상호작용을 물리화학적으로 시뮬레이션하는 기술이 도입되었으나, 시뮬레이션 연산에 굉장히 긴 소요되는 단점이 있었다. 더군다나, 이러한 시뮬레이션을 위해서는 약물과 단백질의 물리화학적 구조를 모델링하는 과정이 필수적이기 때문에, 구조에 대한 정보가 없는 경우 활용하기에는 한계점이 있었다.

선행기술문헌

특허문헌

- [0007] (특허문헌 0001) 대한민국 등록특허공보 제10-2284532호 '분자 활성도 예측 방법 및 이를 위한 장치'

발명의 내용

해결하려는 과제

- [0008] 물리화학적 구조에 대한 시뮬레이션 없이도, 약물-표적 간의 상호작용을 예측하여 선도물질을 빠르고 정확하게 찾을 수 있는 방법을 제안한다.

과제의 해결 수단

- [0009] 본 발명에 따른 자연어 처리 모델을 이용한 약물-표적 상호작용 예측 방법은 (1) 약물 및 타겟 단백질의 각각의 서열을 사전 학습된 자연어 처리 모델에 입력하고, 출력 임베딩을 출력하는 단계; (2) 자연어 처리 모델의 출력 임베딩 중에서 일부 토큰 임베딩을 연결하여 벡터를 생성하는 단계; 및 (3) 벡터를 인공 신경망 모델에 입력하고, 약물-표적 상호작용 확률을 출력하는 단계를 포함한다.

발명의 효과

- [0010] 본 발명은 방대한 데이터로 사전 학습된 자연어 처리 모델을 이용하여 약물과 단백질 간의 상호작용에 대한 복합적인 패턴을 학습할 수 있고, 실험이나 시뮬레이션으로 파악하기 전에 약물과 단백질이 상호작용을 하는지 여부를 예측할 수 있다. 이에 어떤 약물-단백질 쌍부터 실험이나 시뮬레이션을 진행할지 우선순위를 정할 수 있다.
- [0012] 또한, 본 발명은 약물 및 단백질의 물리화학적 구조에 대한 구체적인 정보가 없더라도, 통합 데이터로 학습되는 인공 신경망에 서열 정보만을 입력하여 약물-단백질 상호작용 여부를 예측할 수 있다. 이에 기존의 시뮬레이션 방식보다 훨씬 빠르고 정확하게 선도물질을 찾아낼 수 있고, 효율적으로 신약 개발을 가속화할 수 있다.
- [0014] 여기에 직접적으로 기재되지 않은 효과라도, 발명의 설명에 의해 예상되거나 기대되는 효과는 발명의 효과에 기재된 것으로 이해되어야 한다.

도면의 간단한 설명

- [0015] 도 1은 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 전체적인 아이디어를 나타내는 도면이다.
- 도 2는 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 구성을 나타내는 도면이다.
- 도 3은 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 (1) 단계의 구성을 나타내는 도면이다.

도 4는 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법에서 화합물 자연어 처리 모델의 학습 방법을 나타낸 도면이다.

도 5는 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법에서 단백질 자연어 처리 모델의 학습 방법을 나타낸 도면이다.

도 6은 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 학습 데이터를 나타내는 도면이다.

도 7은 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 통합 학습 데이터를 나타내는 도면이다.

도 8은 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 (2) 단계 및 (3) 단계에서 연산되는 수학적식을 나타내는 도면이다.

도 9는 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 비교를 위한 구성을 나타내는 도면이다.

도 10은 데이터셋에 따른 모델의 성능 비교 결과를 나타내는 도면이다.

도 11 및 도 12는 모델의 평가 지표 중 ROC-AUC에 대한 그래프를 나타내는 도면이다.

도 13은 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 FP 모델과 MolTrans 모델의 성능을 비교하는 도면이다.

도 14는 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 FP 모델과 MolTrans 모델의 성능을 비교하는 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0016] 본 발명은 발명의 설명에서 기재하는 실시예로 한정되지 않고 다양하게 구현될 수 있다. 발명의 설명에서 사용하는 용어는 실시예를 설명하기 위한 것으로, 본 발명을 한정하려는 의도로 사용하는 것이 아니다. 발명의 설명에서 '포함하다'의 용어는 기재된 특징이 조합된 구성이 존재함을 지정하려는 것이다. 그러므로 하나 또는 그 이상의 다른 특징이나 이들을 조합한 구성의 존재 또는 부가 가능성을 배제하지 않는 것으로 이해되어야 한다. 이하에서는 본 발명이 속한 분야의 기술자에게 명백한 내용으로서 공개된 구성이나 기능에 대한 상세한 설명은 생략한다.
- [0018] 본 발명의 실시예에 따른 자연어 처리 모델을 이용한 약물-표적 상호작용 예측 방법(이하에서는 '약물-표적 상호작용 예측 방법'이라 함)은 기존의 물리화학적 시뮬레이션으로 상호작용을 예측했던 방법과 달리, 사전 학습된 자연어 처리 모델을 이용하여 서열 정보만으로 약물-표적 상호작용을 예측하는 방법에 관한 것이다.
- [0020] 도 1은 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 전체적인 아이디어를 나타내는 도면이다. 도 1을 참조하면, 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법은 약물에 대한 자연어 처리 모델 및 단백질에 대한 자연어 처리 모델을 각각 구성한다. 그리고 각각의 자연어 처리 모델에서 출력하는 약물 및 단백질의 특성 정보를 결합하여 벡터를 생성한다. 인공 신경망 모델에 이 벡터를 입력하고, 약물 및 단백질 간의 상호작용 확률을 출력한다. 본 발명에 따르면, 여러 가지 약물과 단백질 쌍을 입력하여 확률값이 높게 나오는 쌍들을 판별할 수 있고, 이러한 쌍들을 DTI(Drug Target Interaction) 실험의 후보로 결정하여 신약 개발을 가속화할 수 있다.
- [0022] 도 2는 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 구성을 나타내는 도면이다. 도 2를 참조하면, 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법은 (1) 약물 및 타겟 단백질의 각각의 서열을 사전 학습된 자연어 처리 모델에 입력하고, 출력 임베딩을 출력하는 단계(S100), (2) 자연어 처리 모델의 출력 임베딩 중 일부 토큰을 연결하여 벡터를 생성하는 단계(S200), 및 (3) 벡터를 인공 신경망 모델에 입력하고, 약물-표적 상호작용 확률을 출력하는 단계(S300)를 포함한다. 각각의 단계는 컴퓨팅 디바이스에 의해 수행될 수 있으며, 각각의 단계에 대한 설명에서 주어가 생략될 수 있다. 여기서, 컴퓨팅 디바이스는 데이터의 연산 및 저장을 수행하여 데이터 처리가 가능한 디바이스이다.
- [0024] (1) 단계(S100)에서는 약물 및 타겟 단백질의 각각의 서열을 사전 학습된 자연어 처리 모델에 입력하고, 출력 임베딩을 출력한다. 본 발명에서는 약물 및 타겟 단백질의 각각의 서열을 임베딩하는 과정에서 사전 학습된 자연어 처리 모델을 이용한다. 여기서, 임베딩(Embedding)은 텍스트를 벡터화하는 것으로서, 컴퓨팅 디바이스가 텍스트를 효율적으로 처리할 수 있도록 하는 전처리 과정이다. 이러한 임베딩 과정을 통해 텍스트 데이터를 활용하는 인공지능 모델의 성능을 더욱 향상시킬 수 있다. 특히, 데이터가 충분히 많은 경우, 임베딩이 인공지능

모델의 성능에 큰 영향을 미치므로, 데이터의 특성을 고려한 임베딩 과정이 매우 중요하다.

- [0026] 이하에서는 자연어 처리 모델로서 BERT(Bidirectional Encoder Representations from Transformers) 계열 모델을 사용하는 실시예를 기준으로 설명한다. 그러나 본 발명에서 자연어 처리 모델이 BERT로 한정되는 것은 아니며, 다양한 자연어 처리 모델을 이용할 수 있다. 실시예에 따라서, 본 발명에서는 사전 학습된 자연어 처리 모델을 이용하여 약물-표적 상호작용 확률을 출력하는 모델로 파인 튜닝(Fine tuning)할 수 있다.
- [0028] 도 3은 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 (1) 단계의 구성을 나타내는 도면이다. 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법에서 (1) 단계(S100)는 세부 단계를 포함할 수 있다. 도 1 및 도 3을 참조하면, (1-1) 단계(S110)에서는 약물의 화합물 서열 및 타겟 단백질의 아미노산 서열(Amino Acid Sequence)을 입력 받는다. 이때, 화합물 서열은 SMILES 포맷의 서열이고, 아미노산 서열은 FASTA 포맷의 서열이다. SMILES(Simplified Molecular-Input Line-Entry System)는 화합물의 구조를, FASTA는 단백질의 구조를 표현하는 텍스트 포맷이다. 따라서 약물 및 타겟 단백질을 각각 SMILES 포맷 및 FASTA 포맷으로 인코딩하면, 그 구조를 문자열로 표현할 수 있다.
- [0030] (1-2) 단계(S120)에서는 화합물 서열 및 아미노산 서열을 토큰화(Tokenization)한다. 임베딩을 위해 먼저 텍스트 데이터를 전처리해야 하는데, 데이터를 토큰(Token) 단위로 나누는 작업을 토큰화라고 한다. 본 발명에서는 SMILES 또는 FASTA 포맷으로 변환된 문자열 서열을 토큰화하므로, 단어 토큰화(Word Tokenization)를 수행한다. 이때, 토큰화를 수행하는 토큰나이저는 하나의 단어를 더 작은 단위의 여러 서브 워드들로 분리하는 서브워드 토큰나이저일 수 있다.
- [0032] (1-2) 단계(S120)에서는 각각의 토큰을 자연어 처리 모델에 입력하고, 하나 이상의 토큰을 포함하는 출력 임베딩을 출력한다. (1-2) 단계(S120)는 문자열을 토큰화한 후, 사전 학습된 자연어 처리 모델을 이용하여 임베딩하는 단계이다. 다만, 본 발명에서는 보다 정확한 확률 예측을 위해 약물과 타겟 단백질별로 서로 다른 자연어 처리 모델을 사용한다. 즉, 본 발명에서는 화합물 서열을 처리하는 화합물 자연어 처리 모델과 아미노산 서열을 처리하는 단백질 자연어 처리 모델을 모두 이용한다.
- [0034] 화합물 자연어 처리 모델과 단백질 자연어 처리 모델은 BERT 계열 모델일 수 있다. BERT 계열 모델은 전체 문장의 문맥을 반영하여 단어를 임베딩할 수 있다. BERT 계열 모델의 입력 임베딩은 임베딩 레이어(Embedding Layer)를 지난 임베딩 벡터들이고, BERT의 연산을 거친 후의 출력 임베딩은 전체 문장의 문맥을 반영한 임베딩 벡터들이 된다. 이때, 출력 임베딩이 포함하는 각각의 벡터를 토큰이라 할 수 있다.
- [0036] 도 4는 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법에서 화합물 자연어 처리 모델의 학습 방법을 나타낸 도면이다. 그리고 도 5는 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법에서 단백질 자연어 처리 모델의 학습 방법을 나타낸 도면이다. BERT 계열 모델은 사전 훈련을 위해서 인공 신경망의 입력으로 들어가는 입력 텍스트의 일부 단어를 랜덤으로 마스킹(Masking)한다. 그리고 인공 신경망에게 이 가려진 단어들(Masked words)을 예측하도록 한다. 도 4를 참조하면, 화합물 자연어 처리 모델은 합성 가능한 화합물을 SMILES 포맷의 서열로 입력 받고, 일부분을 삭제한 변형된 화합물 SMILES 포맷을 화합물 자연어 처리 모델에 입력한다. 화합물 자연어 처리 모델은 삭제된 부분에 가장 확률이 높은 원자나 원자간 결합을 SMILES 포맷으로 출력한다. 이를 정확하게 복구할 때까지 화합물 자연어 처리 모델의 가중치를 변경하며 학습한다. 도 5를 참조하면, 단백질 자연어 처리 모델 또한 자연계에 존재하는 아미노산을 FASTA 포맷의 서열로 입력 받고, 일부분을 삭제한 변형된 단백질 아미노산 FASTA 포맷을 단백질 자연어 처리 모델에 입력한다. 단백질 자연어 처리 모델은 삭제된 부분에 가장 확률이 높은 아미노산을 예측하여 변형된 단백질 아미노산 서열을 복구한다. 이를 정확하게 복구할 때까지 단백질 자연어 처리 모델의 가중치를 변경하며 학습한다.
- [0038] 실시예에 따라서, (1-2) 단계(S120)에서는 화합물 자연어 처리 모델에 ChemBERTa를 사용하고, 단백질 자연어 처리 모델이 ProtBERT를 사용할 수 있다. ChemBERTa 모델은 PubChem 데이터셋을 이용하여 사전 학습된 BERT 모델이다. 그리고 ProtBERT는 거대한 아미노산 데이터를 포함하는 UniREF 및 BFD 데이터셋을 이용하여 사전 학습된 BERT 모델이다. 이때, 각각의 자연어 처리 모델에 입력되는 화합물 서열의 최대 길이는 510이고, 아미노산 서열의 최대 길이는 545개일 수 있다. 이는 전체 데이터셋의 95%를 차지한다. 서열의 최대 길이를 미리 정해진 값으로 제한함으로써 컴퓨팅 리소스를 보다 효율적으로 사용할 수 있다.
- [0040] 다시 도 1을 참조하면, (2) 단계(S200)에서는 자연어 처리 모델의 출력 임베딩 중 일부 토큰을 연결하여 벡터를 생성한다. 이때, (2) 단계(S200)에서는 각각의 출력 임베딩 중 각각의 첫 번째 토큰을 연결하여 벡터를 생성한다. 도 1에서는 CLS 토큰이 별도의 색으로 표시되어 있고, 이 CLS 토큰을 연결하여 벡터 v 를 생성한다. CLS

토큰은 BERT 모델의 출력 임베딩 중 첫 번째 토큰 임베딩에 해당한다. 본 발명에서는 출력 임베딩 중 특정 단어와 대응하는 토큰이 아닌, 문장의 첫 번째 토큰으로 삽입/출력되는 CLS 토큰을 벡터를 생성하는데 사용한다. 각각의 CLS 토큰은 토큰화된 화합물 및 단백질의 서열 전체를 참고하여 생성되는 벡터이므로, 서열에 대한 요약 정보를 가진다. 따라서 이러한 CLS 토큰을 이용하여 약물과 타겟 단백질의 구조적 정보를 상호작용 확률을 예측하는데 활용할 수 있다. 즉, 모델의 성능을 대폭 향상시킬 수 있다.

[0042] (3) 단계(S300)에서는 벡터를 인공 신경망 모델에 입력하고, 약물-표적 상호작용 확률을 출력한다. 여기서, 인공 신경망은 심층학습(Deep Learning) 모델일 수 있고, 완전 연결 신경망(Fully Connected Network; FCN)이거나 적어도 완전 연결 레이어(Fully Connected Layer; FCL)를 가진 신경망일 수 있다. 인공 신경망의 활성화 함수는 ReLU 함수 및 tanh 함수 중 하나 이상을 사용할 수 있다. 그리고, 인공 신경망의 손실 함수(Loss Function)는 평균 제곱 오차(Mean Square Error; MSE) 또는 이상값(Outlier)에 덜 민감한 Smooth L1 함수를 사용할 수 있다.

[0044] 도 6은 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 학습 데이터를 나타내는 도면이다. 그리고 도 7은 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 통합 학습 데이터를 나타내는 도면이다. 도 6를 참조하면, 실시예에 따라서, 본 발명은 BIOSNAP, DAVIS, 및 BindingDB 데이터셋을 인공 신경망 모델의 학습, 검증, 및 시험 데이터로 사용될 수 있다. 이때, (3) 단계(S300)에서 인공 신경망 모델은 서로 다른 데이터셋을 통합하여 생성하는 통합 데이터셋을 학습 데이터로 이용할 수 있다. 이때, 통합 데이터셋은 서로 다른 데이터셋의 학습 데이터를 병합 및 혼합하여 생성하는 통합 학습 데이터 및 서로 다른 데이터셋의 검증 데이터를 병합하여 생성하는 통합 검증 데이터를 포함한다. 도 7을 참조하면, 서로 다른 데이터셋인 BIOSNAP의 학습 데이터, DAVIS의 학습 데이터, 및 BindingDB의 학습 데이터를 병합 및 혼합하는 통합 학습 데이터를 인공 신경망의 학습 데이터로 사용한다. 그리고 BIOSNAP의 검증 데이터, DAVIS의 검증 데이터, 및 BindingDB의 검증 데이터를 병합하는 통합 검증 데이터를 인공 신경망의 검증 데이터로 사용한다. 따라서 본 발명에서는 하나의 데이터셋에 의한 과적합(Overfitting)을 해소하고, 보다 높은 범용성과 예측 성능을 가지는 모델을 학습할 수 있다. 한편, 시험 데이터는 병합하지 않고 BIOSNAP의 시험 데이터, DAVIS의 시험 데이터, 및 BindingDB의 시험 데이터를 각각 사용하는데, 이는 보다 객관적으로 성능을 평가하고 다른 모델들과 비교하기 위함이다.

[0046] 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법에서 (3) 단계(S300)는 세부 단계를 포함할 수 있다. (3-1) 단계(S310)에서는 벡터를 서로 다른 활성화 함수를 가지는 복수의 층을 포함하는 인공 신경망 모델에 입력하고, 출력 벡터를 출력한다. (3-2) 단계(S320)에서는 출력 벡터에 활성화 함수를 적용하여 0 내지 1 사이의 실수값을 출력한다. 여기서, (3-2) 단계(S320)의 활성화 함수는 Sigmoid 함수일 수 있다. (3) 단계(S300)에서는 세부 단계를 통해, 인공 신경망이 다차원의 벡터를 입력 받아 0 내지 1 사이의 값을 출력한다. 그리고 이 값이 자연어 처리 모델에 입력된 약물 및 타겟 단백질 쌍의 상호작용 확률에 해당한다.

[0048] (3-1) 단계(S310)의 인공 신경망 모델은 다음의 수학적 1 내지 수학적 3이 연산되는 은닉층을 포함한다.

수학적 1

[0050]
$$v_1 = \text{ReLU}(FCL_1(v))$$

수학적 2

[0051]
$$v_2 = \tanh(FCL_2(v_1))$$

수학적 3

[0052]
$$v_{out} = FCL_3(v_2)$$

[0053] 여기서, v , v_1 , v_2 , 및 v_{out} 은 벡터, FCL은 Fully Connected Layer(FCL) 연산, 그리고 ReLU 및 tanh는 활성화 함수

수입.

- [0055] FCL은 한 레이어의 모든 노드가 다음 레이어의 모든 노드와 연결된 상태를 가진다. 본 발명의 인공 신경망은 여러 FCL을 적층한 모델일 수 있다. 첫 번째 레이어에서는 벡터 v 를 입력 받고, ReLU 함수로 노드를 활성화하여 벡터 v_1 을 출력한다. 두 번째 레이어에서는 벡터 v_1 을 입력 받고, tanh 함수로 노드를 활성화하여 v_2 를 출력한다. 마지막으로, 세 번째 레이어에서는 벡터 v_2 를 입력 받고, 차원이 감소된 벡터 v_{out} 을 출력한다. 즉, 입력 벡터가 여러 FCL을 통과하면서 서로 다른 활성화 함수로 활성화되고, 최종적으로 차원이 감소된 벡터로 출력된다.
- [0057] 도 8은 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 (2) 단계 및 (3) 단계에서 연산되는 수학적식을 나타내는 도면이다. (2) 단계(S200) 및 (3) 단계(S300)에서 이루어지는 연산은 도 8과 같다. 화합물 특성 벡터(Compound Feature Vector)는 화합물 자연어 처리 모델에서 출력하는 화합물 특성 토큰이며, 단백질 특성 벡터(Protein Feature Vector)는 단백질 자연어 처리 모델에서 출력하는 단백질 특성 토큰이다. 이러한 2개의 토큰을 결합 특성 벡터(Binding Feature Vector)로 연결한 후, 수학적식 1 내지 수학적식 3을 연산한다. 이때, 여러 FCL로 이루어진 레이어 전체를 인터랙션 레이어(Interaction Layer)라 한다. 그리고 마지막 인터랙션 레이어를 통과한 출력값에 Sigmoid 함수를 연산하면 최종적으로 약물-표적 상호작용 확률이 연산된다.
- [0059] 실시예에 따라서, (3-1) 단계(S310)는 인공 신경망의 임의의 노드를 미리 정해진 확률로 드롭할 수 있다. 이는 드롭아웃(Dropout)을 구현한 것으로, 학습 데이터에 대한 과적합을 완화시킬 수 있다. 이러한 드롭아웃은 활성화 함수를 가지는 은닉층에서 이루어질 수 있다.
- [0061] 도 9는 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 비교를 위한 구성을 나타내는 도면이다. 도 10은 데이터셋에 따른 모델의 성능 비교 결과를 나타내는 도면이다. 그리고 도 11 및 도 12는 모델의 평가 지표 중 ROC-AUC에 대한 그래프를 나타내는 도면이다. 도 9를 참조하면, NP 모델은 사전 학습된 자연어 처리 모델을 사용하지 않고, CP는 ChemBERTa 모델만을 사용하며, PP는 ProtBERT 모델만을 사용하고, FP는 ChemBERTa 및 ProtBERT 모델을 모두 사용한 모델이다. 데이터셋에서 Integration은 통합 학습 데이터셋으로 학습된 모델을 말한다. 도 10을 참조하면, 데이터셋별로 모델과 그에 따른 평가 지표를 확인할 수 있다. 평가 지표는 ROC-AUC(Receiver Operating Characteristic Under the Curve), PR-AUC(Precision-Recall-Area under the Curve), 민감도(Sensitivity), 특이도(Specificity), CI(Concordance Index), 및 r_m^2 계수이다.
- [0063] ROC-AUC 및 PR-AUC는 다양한 임계값에서 모델의 성능을 측정한 지표이며, 값이 높을수록 상호작용이 있는 약물-타겟 단백질 쌍(DTI가 있는 쌍)을 배제하지 않을 가능성이 높아진다. PR-AUC는 값이 높을수록 DTI가 없는 쌍일 가능성이 적다. 민감도(Sensitivity)는 DTI가 있는 쌍을 상호작용이 있다고 판단하는 정도이고, 특이도(Specificity)는 DTI가 없는 쌍을 상호작용이 없다고 판단하는 정도이다. DTI 실험은 약물 및 타겟 단백질 쌍을 스크리닝하기 위한 것이므로, 민감도가 중요한 지표에 해당한다. CI 및 r_m^2 계수는 연속형 변수의 출력값을 평가하기 위한 것으로, CI는 1에 가까울수록 정확도가 높고, 0.5에 가까울수록 무작위로 예측한다고 평가한다. r_m^2 계수는 회귀 기반 QSAR(Quantitative Structure-Activity Relationship) 모델의 검증에 사용하는 지표이다.
- [0065] 도 10에서 ROC-AUC, PR-AUC, 민감도, 및 특이도는 5번의 무작위 실행에 걸쳐 측정되었다. 비교 대상 중 MolTrans(Molecular Interaction Transformer) 모델은 레이블이 지정되지 않은 대량의 생물 의학 데이터를 학습하여 DTI를 예측하는 모델 중 하나이다. 도 10을 참조하면, ChemBERTa 및 ProtBERT 모델을 모두 사용한 모델인 FP 모델과 통합 데이터셋을 사용하여 학습한 INT 모델의 조합(FP-Model-INT)의 평가 지표가 전체적으로 높은 것을 확인할 수 있다. 특히, FP-Model-INT 모델이 다른 모델보다 ROC-AUC, PR-AUC, 및 특이도(Specificity)가 높게 나타난다.
- [0067] 도 11을 참조하면, ChemBERTa 및 ProtBERT 모델을 모두 사용한 모델인 FP 모델이 다른 모델들에 비해 가장 빠른 학습 시간에서 가장 높은 ROC-AUC를 나타내고 있다. 또한, 도 12를 참조하면, FP 모델이 모든 학습 데이터 비율에서 가장 높은 ROC-AUC를 나타내고 있다.
- [0069] 도 13은 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 FP 모델과 MolTrans 모델의 성능을 비교하는 도면이다. 도 13은 DAVIS 및 BindingDB를 각각 학습 데이터로 이용한 FP 모델과 MolTrans 모델의 성능을 비교한 결과이다. (A) 및 (B)는 FP 모델의 결과이고, (C) 및 (D)는 MolTrans 모델의 결과이다. Measurement는 pK_d

이며, 아래의 수식 4에 의해 계산된다.

수식 4

$$D+T=DT$$

$$K_d=[D][T]/[DT]$$

$$pK_d=-\log K_d$$

여기서, D는 약물, T는 타겟 단백질, DT는 약물-타겟 복합체, [x]는 x의 평형에서의 몰농도, 그리고 K_d 는 해리 상수임.

pK_d 는 수용액 속에서 화합물의 이온화 상태를 정량적으로 나타내는 값으로, 값이 높을수록 약물-표적 단백질 간의 친화도가 높다. 도 12를 참조하면, FP 모델에서는 선형성이 분명하게 나타나나, MolTrans 모델에서는 그 분포가 2가지로 나뉘고 있다. 따라서 FP 모델이 더 나은 예측 분포를 가진다.

도 14는 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 FP 모델과 MolTrans 모델의 성능을 비교하는 도면이다. 도 14는 잘 알려진 13개의 약물 및 단백질 쌍의 상호작용을 통해 FP 모델과 MolTrans 모델을 비교한다. 도 14에서는 상호작용이 있는 약물 및 단백질 쌍에 대한 비교이므로, 평균적인 값(BP)이 높을수록 모델의 성능이 우수하다. 도 14를 참조하면, 학습 데이터는 BindingDB 단독(BDB)보다 통합 데이터셋(INT)을 사용하는 경우가 평균적인 값(BP)이 높다. 또한, 사전 학습된 자연어 처리 모델을 사용하지 않는 모델(NP)보다 FP 모델의 평균적인 값(BP)이 높다.

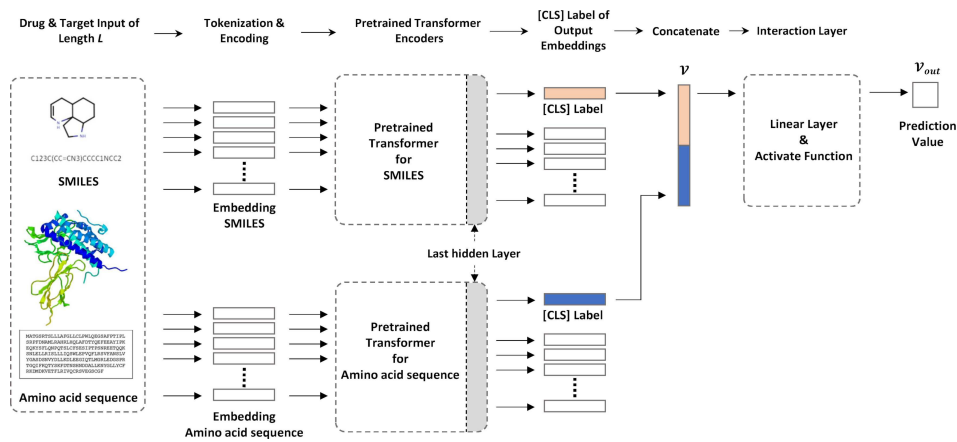
이와 같이, 본 발명의 실시예에 따라 서로 다른 사전 학습된 자연어 처리 모델을 파인 튜닝(Fine tuning)하고, 서로 다른 데이터셋을 조합한 통합 데이터셋으로 학습하는 모델이 다른 모델들보다 약물-표적 상호작용 예측 성능이 높은 것을 알 수 있다. 서로 다른 사전 학습된 자연어 처리 모델을 이용하여 서열을 임베딩하고, 이를 파인 튜닝(Fine tuning)하여 약물-표적 상호작용 확률을 출력하는 구성은 발명자가 반복적인 시뮬레이션을 통해 도출한 구성으로서, 방대한 자연어 데이터를 기반으로 약물-표적 상호작용 확률을 예측할 수 있도록 하는 구성에 해당한다.

실시예에 따라서, 자연어 처리 모델을 이용한 약물-표적 상호작용 예측 장치(이하에서는 '약물-표적 상호작용 예측 장치'라 함)는 명령어를 저장하는 메모리 및 명령어를 실행하는 프로세서를 포함한다. 여기서, 명령어는 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 각각의 단계를 포함한다. 또는, 실시예에 따라서, 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 각각의 단계에서 수행하는 기능은 각각의 구성으로 변환되고, 이러한 구성을 포함하는 본 발명의 실시예에 따른 약물-표적 상호작용 예측 장치로 구현될 수 있다. 각각의 구성은 전술한 진단 방법의 단계에 대한 설명과 중복되므로, 그 자세한 설명은 생략한다.

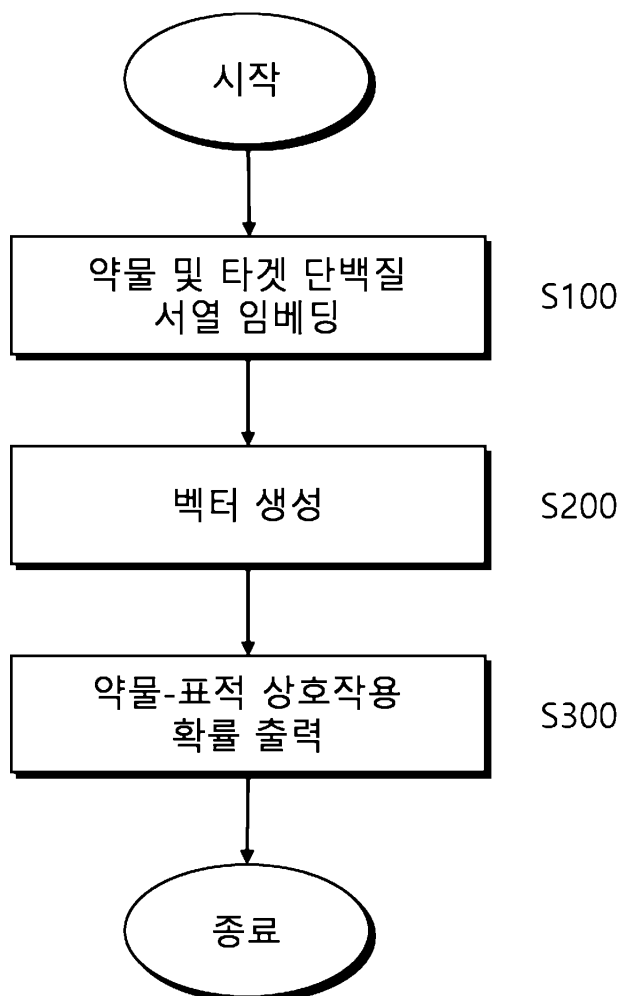
발명의 설명에 기재된 내용은 예시에 불과하며, 본 발명은 이 분야의 기술자에 의하여 다양하게 변형되어 실시될 수 있다. 따라서 본 발명의 보호범위는 설명된 실시예의 기재와 표현으로 제한되지 않는다.

도면

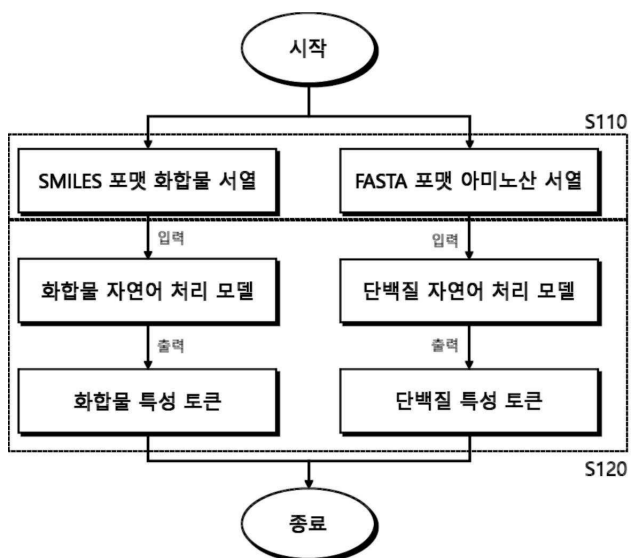
도면1



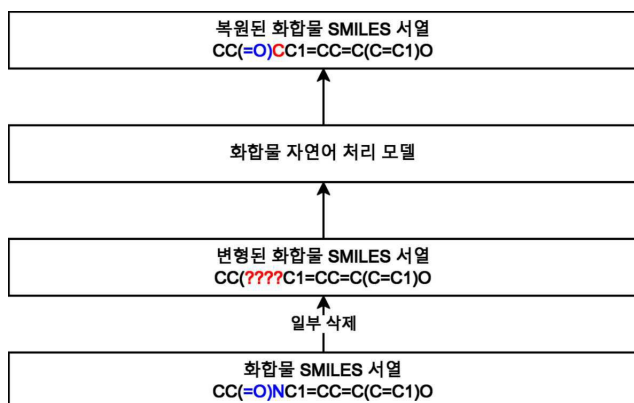
도면2



도면3



도면4



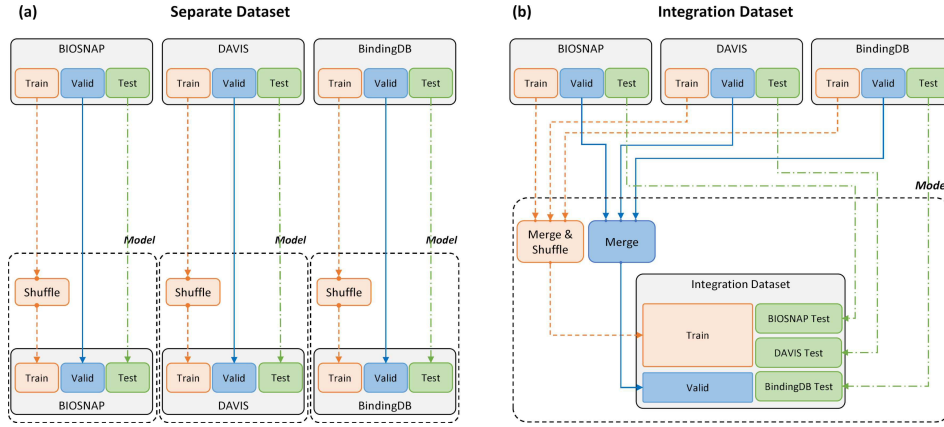
도면5



도면6

Dataset	Drug	Proteins	Interactions
BIOSNAP	4510	2181	27,482 (13,741/13,741)
DAVIS	68	379	11,103 (1506/9597)
BindingDB	10,665	1413	32,601 (9166/23,435)

도면7



도면8

B : Binding Feature Vector
 P : Protein Feature Vector
 C : Compound Feature Vector

$$\begin{aligned}
 B &= P \parallel C = [p_1, p_2, p_3, \dots, p_{d_p}] \parallel [c_1, c_2, c_3, \dots, c_{d_c}] \\
 &\downarrow \\
 B &= Y_1 = [p_1, p_2, p_3, \dots, p_{d_p}, c_1, c_2, c_3, \dots, c_{d_c}] \\
 &\downarrow \\
 v_{out} &= \text{interaction layer}(v_{in}) \leftarrow \begin{aligned} v_1 &= \text{ReLU}(FCL_1(v_{in})) \\ v_2 &= \tanh(FCL_2(v_1)) \\ v_{out} &= FCL_3(v_2) \end{aligned} \\
 &\downarrow \\
 Y_1 &= [p_1, p_2, p_3, \dots, p_{d_p}, c_1, c_2, c_3, \dots, c_{d_c}] \\
 &\downarrow \\
 Y_2 &= \text{interaction layer}(Y_1) \\
 &\downarrow \\
 Y_3 &= \text{interaction layer}(Y_2) \\
 &\downarrow \\
 &\dots \\
 &\downarrow \\
 Y_n &= \text{interaction layer}(Y_{n-1}) \\
 p &= \text{sigmoid}(Y_n) \quad p \in \mathbb{R}^1, 0 \leq p \leq 1
 \end{aligned}$$

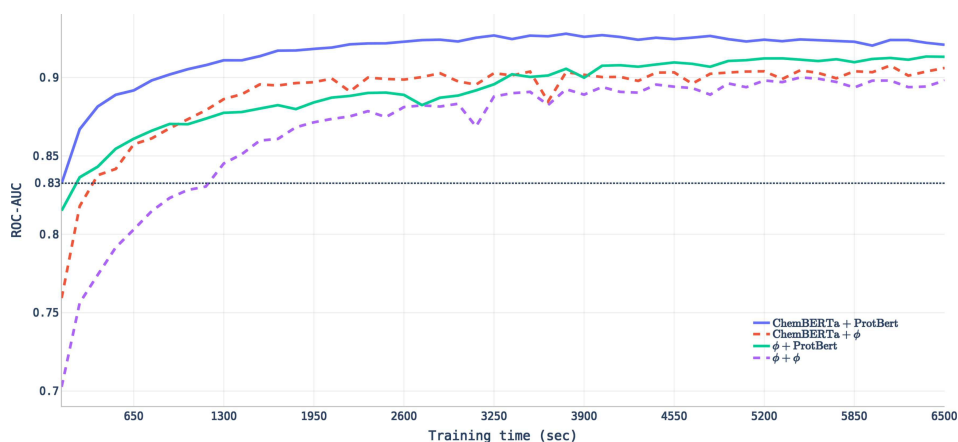
도면9

Dataset	BIOSNAP (BS)	DAVIS (DV)	BindingDB (BDB)	Integration (INT)
Non-pretrained (NP) ($\phi + \phi$)	NP-Model-BS	NP-Model-DV	NP-Model-BDB	NP-Model-INT
ChemBERTa-pretrained (CP) (ChemBERTa + ϕ)	CP-Model-BS	CP-Model-DV	CP-Model-BDB	CP-Model-INT
ProtBERT-pretrained (PP) (ϕ + ProtBERT)	PP-Model-BS	PP-Model-DV	PP-Model-BDB	PP-Model-INT
Full-pretrained (FP) (ChemBERTa + ProtBERT)	FP-Model-BS	FP-Model-DV	FP-Model-BDB	FP-Model-INT

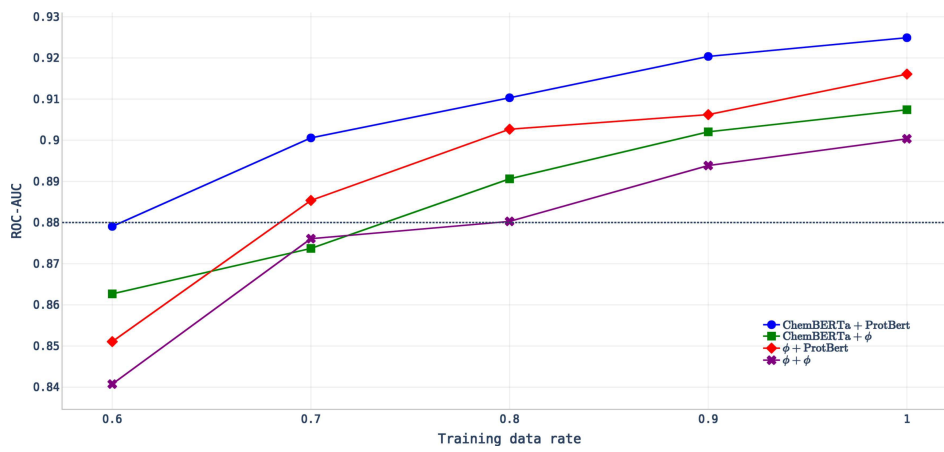
도면10

Method	ROC-AUC	PR-AUC	Sensitivity	Specificity	CI	r_m^2
Dataset 1. BIOSNAP						
MolTrans	0.895 $\pm$ 0.002	0.901 $\pm$ 0.004	0.775 $\pm$ 0.032	0.851 $\pm$ 0.014	0.889	0.449
NP-Model-BS	0.882 $\pm$ 0.004	0.871 $\pm$ 0.015	0.779 $\pm$ 0.020	0.850 $\pm$ 0.012	0.895	0.428
CP-Model-BS	0.881 $\pm$ 0.009	0.859 $\pm$ 0.017	0.811 $\pm$ 0.018	0.835 $\pm$ 0.008	0.891	0.406
PP-Model-BS	0.893 $\pm$ 0.003	0.874 $\pm$ 0.006	0.803 $\pm$ 0.033	0.851 $\pm$ 0.019	0.896	0.425
FP-Model-BS	0.914 $\pm$ 0.006	0.900 $\pm$ 0.007	0.862 $\pm$ 0.025	0.847 $\pm$ 0.007	0.913	0.467
NP-Model-INT	0.877 $\pm$ 0.007	0.860 $\pm$ 0.010	0.785 $\pm$ 0.007	0.842 $\pm$ 0.008	0.897	0.421
CP-Model-INT	0.875 $\pm$ 0.006	0.851 $\pm$ 0.009	0.775 $\pm$ 0.023	0.844 $\pm$ 0.016	0.885	0.401
PP-Model-INT	0.895 $\pm$ 0.003	0.880 $\pm$ 0.008	0.802 $\pm$ 0.018	0.852 $\pm$ 0.009	0.896	0.435
FP-Model-INT	0.910 $\pm$ 0.012	0.897 $\pm$ 0.014	0.830 $\pm$ 0.029	0.863 $\pm$ 0.011	0.911	0.447
Dataset 2. DAVIS						
MolTrans	0.907 $\pm$ 0.002	0.404 $\pm$ 0.016	0.800 $\pm$ 0.022	0.876 $\pm$ 0.013	0.903	0.156
NP-Model-DV	0.870 $\pm$ 0.003	0.283 $\pm$ 0.005	0.738 $\pm$ 0.030	0.871 $\pm$ 0.026	0.875	0.118
CP-Model-DV	0.882 $\pm$ 0.006	0.250 $\pm$ 0.023	0.744 $\pm$ 0.021	0.888 $\pm$ 0.019	0.878	0.117
PP-Model-DV	0.866 $\pm$ 0.003	0.263 $\pm$ 0.007	0.747 $\pm$ 0.020	0.856 $\pm$ 0.012	0.864	0.115
FP-Model-DV	0.920 $\pm$ 0.002	0.395 $\pm$ 0.007	0.824 $\pm$ 0.026	0.889 $\pm$ 0.015	0.917	0.167
NP-Model-INT	0.899 $\pm$ 0.008	0.322 $\pm$ 0.030	0.814 $\pm$ 0.039	0.857 $\pm$ 0.028	0.892	0.141
CP-Model-INT	0.904 $\pm$ 0.011	0.351 $\pm$ 0.035	0.814 $\pm$ 0.030	0.859 $\pm$ 0.020	0.917	0.169
PP-Model-INT	0.923 $\pm$ 0.005	0.417 $\pm$ 0.028	0.844 $\pm$ 0.017	0.876 $\pm$ 0.021	0.916	0.162
FP-Model-INT	0.942 $\pm$ 0.005	0.517 $\pm$ 0.017	0.903 $\pm$ 0.017	0.866 $\pm$ 0.015	0.940	0.201
Dataset 3. BindingDB						
MolTrans	0.914 $\pm$ 0.001	0.622 $\pm$ 0.007	0.797 $\pm$ 0.005	0.896 $\pm$ 0.007	0.899	0.267
NP-Model-BDB	0.891 $\pm$ 0.005	0.515 $\pm$ 0.014	0.774 $\pm$ 0.012	0.897 $\pm$ 0.013	0.899	0.309
CP-Model-BDB	0.914 $\pm$ 0.003	0.585 $\pm$ 0.021	0.803 $\pm$ 0.011	0.904 $\pm$ 0.010	0.907	0.320
PP-Model-BDB	0.897 $\pm$ 0.003	0.557 $\pm$ 0.013	0.775 $\pm$ 0.019	0.900 $\pm$ 0.009	0.913	0.324
FP-Model-BDB	0.922 $\pm$ 0.001	0.623 $\pm$ 0.010	0.814 $\pm$ 0.025	0.916 $\pm$ 0.016	0.927	0.365
NP-Model-INT	0.904 $\pm$ 0.001	0.574 $\pm$ 0.008	0.766 $\pm$ 0.015	0.910 $\pm$ 0.015	0.907	0.315
CP-Model-INT	0.909 $\pm$ 0.005	0.600 $\pm$ 0.019	0.787 $\pm$ 0.008	0.907 $\pm$ 0.008	0.918	0.330
PP-Model-INT	0.918 $\pm$ 0.001	0.607 $\pm$ 0.012	0.787 $\pm$ 0.014	0.920 $\pm$ 0.010	0.916	0.344
FP-Model-INT	0.926 $\pm$ 0.001	0.639 $\pm$ 0.018	0.802 $\pm$ 0.022	0.928 $\pm$ 0.013	0.926	0.362

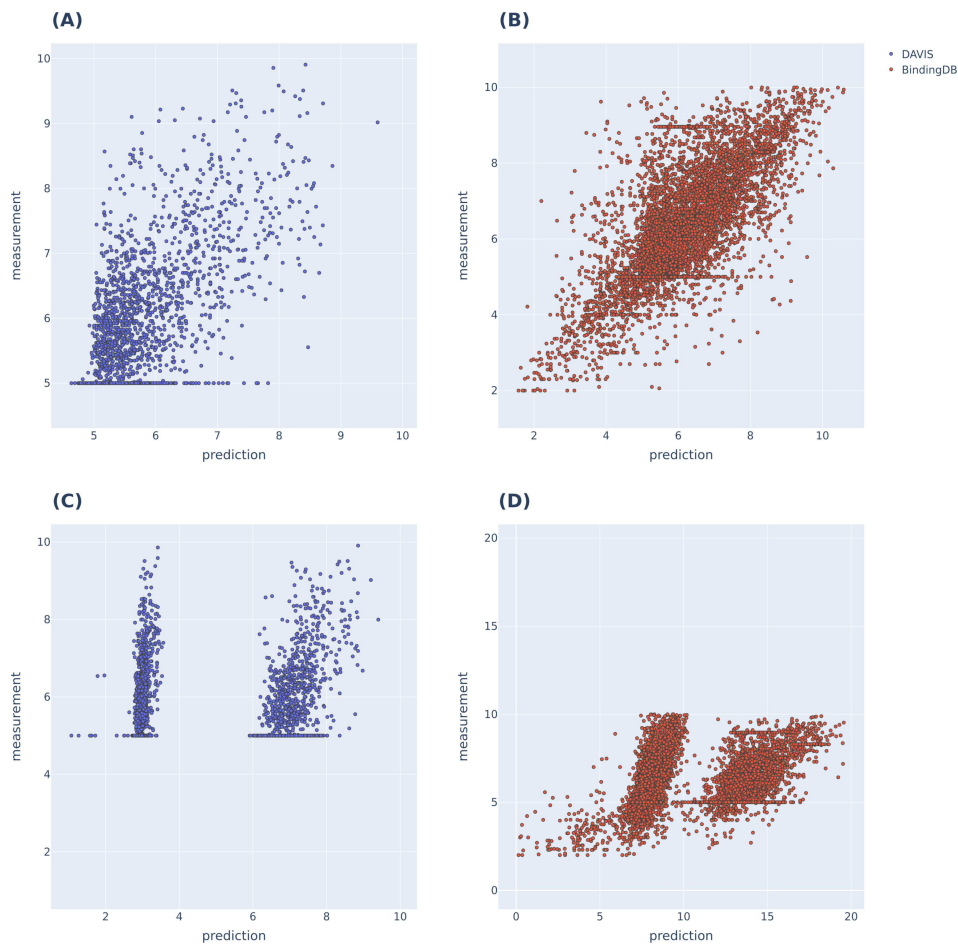
도면11



도면12



도면13



도면14

CYP Subtype (Targets)	Drugs	MolTrans	NP-Model- BDB	FP-Model-BDB	NP-Model-INT	FP-Model-INT
1A2	Phenacetin	0.391	0.745	0.774	0.767	0.764
1A2	7-Ethoxyresorufin	0.255	0.747	0.786	0.762	0.784
2B6	Efavirenz	0.755	0.728	0.755	0.687	0.749
2B6	Bupropion	0.524	0.727	0.757	0.712	0.733
3A4	Midazolam	0.352	0.707	0.694	0.719	0.721
3A4	Testosterone	0.125	0.741	0.699	0.714	0.662
2C8	Paclitaxel	0.823	0.698	0.676	0.722	0.732
2C8	Amodiaquine	0.631	0.694	0.712	0.748	0.759
2C9	S-Warfarin	0.546	0.725	0.655	0.701	0.705
2C9	Diclofenac	0.556	0.645	0.758	0.726	0.711
2C19	S-Mephenytoin	0.313	0.738	0.742	0.740	0.744
2D6	Bufuralol	0.475	0.355	0.749	0.756	0.699
2D6	Dextromethorphan	-	0.490	0.740	0.750	0.740
Average BP		0.479	0.687	0.730	0.731	0.731