【서지사항】

【서류명】 특허출원서

【참조번호】 P20240520KR

【출원구분】 특허출원

【출원인】

【명칭】 연세대학교 산학협력단

【특허고객번호】 2-2005-009509-9

【대리인】

【명칭】 특허법인 지담

【대리인번호】 9-2018-100261-1

【지정된변리사】 김재흥

【포괄위임등록번호】 2023-020263-8

【발명의 국문명칭】 SE(3)-불변성 및 물리 지식 기반 네트워크 모델을 이용한 단백질-리간드 결합 친화도 예측 시스템, 방법 및 프로그램

【발명의 영문명칭】 SYSTEM, METHOD AND PROGRAM FOR PREDICTING PROTEIN-LIGAND BINDING AFFINITY USING SE(3)-INVARIANT AND PHYSICS-BASED NETWORK MODEL

【발명자】

【성명】 박상현

【성명의 영문표기】 PARK, Sang Hyun

【주민등록번호】 670101-1XXXXXX

【우편번호】 03626

【주소】 서울특별시 서대문구 세무서8길, 101동 1502호

【발명자】

【성명】 최승연

【성명의 영문표기】 CHOI, Seung Yeon

【주민등록번호】 970828-1XXXXXX

【우편번호】 07213

【주소】 서울특별시 영등포구 당산로45길 7-3, 101동 808호

【발명자】

【성명】 서상민

【성명의 영문표기】 SEO, Sang Min

【주민등록번호】 930507-1XXXXXX

【우편번호】 63272

【주소】 제주특별자치도 제주시 고마로 44, 801호

【출원언어】 국어

【심사청구】 청구

【공지예외적용대상증명서류의 내용】

【공개형태】 논문발표

【공개일자】 2024.10.19

【이 발명을 지원한 국가연구개발사업】

【과제고유번호】 1711198526

【과제번호】 00229822

【부처명】 과학기술정보통신부

【과제관리(전문)기관명】 한국연구재단

【연구사업명】 원천기술개발사업

【연구과제명】 [통합이지바로/주관] 난치성 질환 극복을 위한 인공지능 기반의 다중 약물 적응증 최적화 플랫폼 개발 및 혁신신약 발굴 (1/2단계)(1/2)

【과제수행기관명】 연세대학교

【연구기간】 2024.01.01 ~ 2024.12.31

【취지】 위와 같이 특허청장에게 제출합니다.

대리인 특허법인 지담 (서명 또는 인)

【수수료】

【출원료】	0	면	46,000 원
【가산출원료】	65	면	0 원
【우선권주장료】	0	건	0 원
【심사청구료】	15	항	931,000 원

【합계】 977,000원

【감면사유】 전담조직(50%감면)[1]

【감면후 수수료】 488,500 원

【첨부서류】 1.공지예외적용대상(신규성상실의예외, 출원시의특례)규정을 적용받기 위한 증명서류_1통

1 : 공지예외적용대상(신규성상실의예외,_출원시의특례)규정을_적용받기_위한_증명

서류

PDF 파일 첨부

【발명의 설명】

【발명의 명칭】

SE(3)-불변성 및 물리 지식 기반 네트워크 모델을 이용한 단백질-리간드 결합 친화도 예측 시스템, 방법 및 프로그램{SYSTEM, METHOD AND PROGRAM FOR PREDICTING PROTEIN-LIGAND BINDING AFFINITY USING SE(3)-INVARIANT AND PHYSICS-BASED NETWORK MODEL}

【기술분야】

【0001】본 발명은 단백질-리간드 결합 친화도를 예측하기 위한 시스템, 방법 및 프로그램에 관한 것으로, 기하학적 및 물리화학적 원리에 기반한 네트워크 모델을 활용하여 단백질 및 리간드 간의 결합 친화도를 예측하는 기술에 관한 것이다.

【발명의 배경이 되는 기술】

【0003】단백질-리간드 결합 친화도(Binding Affinity, BA) 예측은 약물 스크리닝에서 필수적인 과정으로, 수많은 분자 구조 중에서 약물 후보 물질을 선택하는데 매우 중요한 역할을 한다. 최근 머신 러닝과 딥러닝의 발전으로 다양한 단백질-리간드 결합 친화도 예측 방법론이 제안되었다. 특히, AlphaFold와 같은 단백질의 3차원 구조를 모델링하는 연구는 단백질-리간드 복합체의 3D 구조를 기반으로 한 결합 친화도 예측의 가능성을 크게 향상시켰다. 이러한 접근 방식은 크게 ML 기

반, CNN 기반, GNN 기반 방법으로 분류될 수 있다.

【0004】ML 기반 방법은 단백질과 리간드 원자 간의 상호작용을 바탕으로 미리 정의된 규칙을 통해 서포트 벡터 회귀(SVR) 및 랜덤 포레스트와 같은 모델을 사용하여 결합 친화도를 예측한다. 그러나 이 방법은 원자 간의 공간적 상관관계를 충분히 반영하지 못한다는 한계가 있다.

【0005】CNN(Convolutional Neural Network) 기반 방법은 복합체를 3D 그리드로 변환하여 3D CNN 모델을 통해 결합 친화도를 예측하지만, 그리드에서 빈 공간이 발생해 계산 비효율성과 메모리 낭비를 초래할 수 있다. 또한, 거리 인식 및 회전 불변성을 고려하지 않아 예측 성능이 불안정할 수 있다.

【0006】GNN(Graph Neural Network) 기반 방법은 단백질과 리간드의 원자를 그래프의 노드로 정의하고, 특정 거리 내 원자 쌍을 엣지로 연결하여 GNN 모델로 처리한다. 일부 연구에서는 3D 좌표를 직접 입력하여 결합 친화도를 예측하기도 하지만, 이러한 방법은 회전 및 평행 이동에 민감하거나 훈련 중 보지 못한 기하학적 구성을 처리하는 데 어려움이 있다.

【0007】한편, 딥러닝 모델은 컴퓨터 비전 및 자연어 처리와 같은 다양한 분야에서 큰 성공을 거두고 있지만, 여전히 해석 가능한 정보를 추출하는데 어려움을 겪고 있으며, 특히 물리적 일관성을 유지하지 못하는 예측이나 불가능한 값을 예측하는 경우가 있다. 이는 물리화학과 같이 데이터 수집이 어려운 분야에서 더욱 두드러진다. 이를 해결하기 위해, 물리 정보 신경망(Physics-Informed Neural Network; PINN)이 활발히 연구되고 있다.

【0008】PINN은 네트워크 모델에 물리적 법칙과 도메인 지식을 통합하여 학습 과정에서 귀납적 편향(Inductive Bias)으로 적용할 수 있도록 한다. 이를 통해, 모델이 학습 및 추론 과정에서 이러한 법칙을 내재적으로 만족하도록 하여, 노이즈가 포함된 데이터셋에서도 모델의 강건성을 확보하고, 일반화 성능을 향상시킬 수 있다.

【0009】그러나 기존 PINN 모델은 단백질과 리간드 간의 연결 정보만을 고려하여 두 구조 간의 기하학적 정보를 충분히 모델링하지 못하는 한계가 있다. 이러한 문제들로 인해, 기존 모델들은 독립적인 데이터셋에서의 예측 성능이 떨어지며, 약물 개발 과정에서의 실용성이 제한될 수 있다. 예를 들면, 새로운 단백질-리간드 복합체에 대해 정확한 결합 친화도 예측이 어렵기 때문에, 약물 개발 초기 단계에서 시간과 비용이 증가할 수 있다.

【0010】따라서, 기하학적 변환에 불변한 성능을 제공하고, 단백질-리간드 복합체 간의 상호작용을 보다 정확하게 반영할 수 있는 네트워크 모델 개발이 필요하다.

【선행기술문헌】

【특허문헌】

【0012】**(특허문헌 0001)** 한국 특허등록공보 제10-2617957호

【발명의 내용】

【해결하고자 하는 과제】

【0013】 본 발명이 이루고자 하는 기술적 과제는 상기한 문제점을 해결하기 위해 SE(3)-불변성 및 물리 지식 기반 네트워크 모델을 이용한 단백질-리간드 결합 친화도 예측 시스템, 방법 및 그 프로그램을 제공하는 것이다.

【0014】 구체적으로, 단백질-리간드 복합체의 기하학적 변환에 불변한 단백질-리간드 결합 친화도 예측 시스템, 방법 및 그 프로그램을 제공하는 것이다.

【0015】 또한, 본 발명은 물리화학적 귀납적 편향을 통합하여 결합 자유 에너지가 최소화되는 지점에서 결합이 이루어지는 물리화학적 원칙을 반영한 단백질-리간드 결합 친화도 예측 시스템, 방법 및 그 프로그램을 제공하는 것이다.

【0016】 또한, 본 발명은 단백질-리간드 간 결합 친화도를 예측함에 있어 해석 가능성과 높은 일반화 성능을 보장하며, 실제 약물 개발 과정에서 가상 스크리닝에 실질적으로 적용 가능한 시스템, 방법 및 그 프로그램을 제공하는 것이다.

【0017】 본 발명이 이루고자 하는 기술적 과제는 이상에서 언급한 기술적 과제로 제한되지 않으며, 언급되지 않은 또 다른 기술적 과제들은 아래의 기재로부터 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

【과제의 해결 수단】

【0019】상기 기술적 과제를 달성하기 위하여, 본 발명의 일실시예는 단백질-리간드 간 결합 친화도 예측 시스템에 있어서, 단백질-리간드 복합체의 원자 정보 및 3차원 좌표 데이터를 수신하는 수신부; 상기 원자 정보 및 상기 3차원 좌표 데이터를 기초로 하여 전처리 작업을 수행하여 그래프 구조를 생성하는 그래프 전처리 모듈; 3차원 공간에서 물질 또는 시스템이 회전 또는 평행 이동에 대해 상기 결합 친화도가 불변하도록 제1 귀납적 편향을 기초로 하여 상기 그래프 구조를 임베딩 레이어 및 은닉 표현으로 불변성 변환하는 SE(3)-불변성 변환 모듈; 상기 임베딩 레이어 및 상기 은닉 표현을 입력으로 하여 상기 결합 친화도를 예측하는 물리 지식 기반 네트워크 모델; 및 상기 예측된 결합 친화도를 출력하는 출력부; 를 포함하는 단백질-리간드 간 결합 친화도 예측 시스템을 제공한다.

【0020】본 발명의 실시예에 있어서, 상기 물리 지식 기반 네트워크 모델은, (a) 상기 은닉 표현을 이용하여 단백질-리간드 상호작용을 계산하는 단계; (b) 상기 단백질-리간드 상호작용을 이용하여 상기 결합 친화도를 예측하는 단계; 및 (c) 단백질-리간드 복합체의 결합 자유 에너지가 최소화되는 지점에서 결합 상태가 형성된다는 제2 귀납적 편향을 기초로 하여 손실 함수를 계산하는 단계; 를 통해 상기 손실 함수가 최소가 되도록 학습되는 것일 수 있다.

【0021】또한, 상기 그래프 전처리 모듈은, 상기 원자 정보에 기반하여 각 원자를 노드로 정의하여 각 노드에 원자의 물리화학적 특징을 포함시키고, 상기 3차원 좌표를 포함한 각 원자의 위치 행렬을 생성하고, KNN(K-Nearest Neighbors) 알고리즘을 사용하여 원자 간의 거리를 기준으로 엣지를 정의하여 이웃 원자와의

상호작용을 엣지로 연결하여 나타내는 것일 수 있다.

【0022】 또한, 상기 SE(3)-불변성 변환 모듈은 아래 수학식 1, 2 및3을 통해 상기 물리화학적 특징을 인코딩하여 임베딩 레이어 및 은닉 표현으로 변환하고, 아래 수학식 4, 5, 6, 7 및 8을 통해 상기 은닉 표현을 업데이트하는 것일 수 있다.

【0023】 또한, 상기 (a) 단계는, 상기 은닉 표현을 이용하여 단백질-리간드 상호작용 매트릭스를 아래 수학식 9를 통해 계산하는 것일 수 있다.

【0024】 또한, 상기 (b) 단계는, 아래 수학식 10을 통해 계산되는 원자쌍별 반 데르 발스 상호작용 에너지의 합을 기초로 상기 결합 친화도를 예측하는 것일 수 있다.

【0025】 또한, 상기 (c) 단계는, 아래 수학식 12를 통해 계산된 L_d 및 아래 수학식 13을 통해 계산된 L_p의 합으로 된 손실 함수를 계산하는 것일 수 있다.

【0026】 상기 기술적 과제를 달성하기 위하여, 본 발명의 다른 실시예는 단백질-리간드 간 결합 친화도 예측 방법에 있어서, (A) 단백질-리간드 복합체의 원자 정보 및 3차원 좌표 데이터를 수신하는 단계; (B) 상기 원자 정보 및 상기 3차원 좌표 데이터를 기초로 하여 전처리 작업을 수행하여 그래프 구조를 생성하는 그래프 전처리 단계; (C) 3차원 공간에서 물질 또는 시스템이 회전 또는 평행 이동을 하더라도 상기 결합 친화도가 변하지 않는 제1 귀납적 편향을 기초로 하여 상기 그래프 구조를 임베딩 레이어 및 은닉 표현으로 변환하는 SE(3)-불변성 변환 단계; 및 (D) 상기 임베딩 레이어 및 상기 은닉 표현을 입력으로 하여 물리 지식 기반 네

트워크 모델을 이용하여 상기 결합 친화도를 예측하는 단계; 를 포함하는 단백질-리간드 간 결합 친화도 예측 방법을 제공한다.

【0027】 본 발명의 실시예에 있어서, 상기 물리 지식 기반 네트워크 모델은 (E) 상기 은닉 표현을 이용하여 단백질-리간드 상호작용을 계산하는 단계; (F) 상기 단백질-리간드 상호작용을 이용하여 상기 결합 친화도를 예측하는 단계; 및 (G) 단백질-리간드 복합체의 결합 자유 에너지가 최소화되는 지점에서 결합 상태가 형성된다는 제2 귀납적 편향을 기초로 하여 손실 함수를 계산하는 단계;를 통해 상기 손실 함수가 최소가 되도록 학습되는 것일 수 있다.

【0028】 또한, 상기 (B) 단계는, 상기 원자 정보에 기반하여 각 원자를 노드로 정의하여 각 노드에 원자의 물리화학적 특징을 포함시키고, 상기 3차원 좌표를 포함한 각 원자의 위치 행렬을 생성하고, KNN(K-Nearest Neighbors) 알고리즘을 사용하여 원자 간의 거리를 기준으로 엣지를 정의하여 이웃 원자와의 상호작용을 엣지로 연결하여 나타내는 것일 수 있다.

【0029】 또한, 상기 (C) 단계는 아래 수학식 14, 15 및 16을 통해 상기 물리화학적 특징을 인코딩하여 임베딩 레이어 및 은닉 표현으로 변환하고, 아래 수학식 17, 18, 19, 20 및 21을 통해 상기 은닉 표현을 업데이트하는 것일 수 있다.

【0030】 또한, 상기 (E) 단계는 상기 은닉 표현을 이용하여 단백질-리간드 상호작용 매트릭스를 아래 수학식 22를 통해 계산하는 것일 수 있다.

【0031】또한, 상기 (F) 단계는, 아래 수학식 22를 통해 계산되는 원자쌍별 반 데르 발스 상호작용 에너지의 합을 기초로 상기 결합 친화도를 예측하는 것일 수 있다.

【0032】또한, 상기 (G) 단계는, 아래 수학식 24를 통해 계산된 L_d 및 아래 수학식 25를 통해 계산된 L_p의 합의 합으로 된 손실 함수를 계산하는 것일 수 있다.

【0033】상기 기술적 과제를 달성하기 위하여, 본 발명의 또 다른 실시예는 상기 단백질-리간드 간 결합 친화도 예측 방법을 실행하는 프로그램이 기록된 컴퓨터가 판독 가능한 기록 매체를 제공한다.

【발명의 효과】

【0035】본 발명의 실시예에 따르면, 본 발명은 SE(3)-불변성 및 물리 지식 기반 네트워크 모델을 이용한 단백질-리간드 결합 친화도 예측 시스템을 제공할 수 있다.

【0036】또한, 본 발명은 여러 귀납적 편향을 도입하여 제한된 데이터로부터 우수한 일반화 성능을 달성할 수 있다.

【0037】또한, 본 발명은 결합 친화도가 3차원 공간에서의 회전 및 평행 이동에 관계없이 일정하게 유지된다는 기하학적 귀납적 편향과 결합이 최소 결합 자유 에너지에서 발생한다는 물리화학적 귀납적 편향을 통합하여 네트워크 모델에 반

영할 수 있다.

【0038】 또한, 본 발명은 2 개의 벤치마크 세트를 통해 엄격한 검증을 수행한 결과, 단백질-리간드 간 결합 친화도 예측 성능의 우수성, 실제 약물 개발 과정에서의 가상 스크리닝에서의 높은 실용성 및 해석 가능성을 시각화하여 예측값의 높은 신뢰성을 제공할 수 있다.

【0039】 본 발명의 효과는 상기한 효과로 한정되는 것은 아니며, 본 발명의 설명 또는 청구범위에 기재된 발명의 구성으로부터 추론 가능한 모든 효과를 포함하는 것으로 이해되어야 한다.

【도면의 간단한 설명】

【0041】 도 1은 본 발명의 실시예에 따른 SE(3) 불변성 및 물리 지식 기반 딥러닝 모델을 이용한 단백질-리간드 결합 친화도 예측 시스템 및 그 방법에 관한 개요도이다.

도 2는 본 발명의 실시예에 따른 단백질-리간드 결합 친화도 예측 시스템의 구성을 도시한 블록도이다.

도 3은 본 발명의 실시예에 따른 물리 지식 기반 네트워크 모델의 학습 과정 및 결합 친화도 예측 방법을 보여주는 참고도이다.

도 4는 본 발명의 일 실시예에 따른 SPIN 모델의 각 귀납적 편향 제거에 따른 성능 비교 결과를 보여주는 참고도이다.

도 5는 본 발명의 일 실시예에 따른 SPIN 모델을 이용한 가상 스크리닝 실험에서 순위 결정력을 비교한 결과를 보여주는 참고도이다.

도 6은 본 발명의 일 실시예에 따른 SPIN 모델을 이용한 단백질-리간드 상호작용 분석 및 해석 가능성을 검증하는 참고도이다.

도 7 은 본 발명의 실시예에 따른 단백질-리간드 결합 친화도 예측 방법을 보여주는 흐름도이다.

도 8 은 본 발명의 실시예에 따른 단백질-리간드 결합 친화도 예측 시스템 및 그 방법을 구현하는 컴퓨팅 장치를 도시한다.

【발명을 실시하기 위한 구체적인 내용】

【0042】이하에서는 첨부한 도면을 참조하여 본 발명을 설명하기로 한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 따라서 여기에서 설명하는 실시예로 한정되는 것은 아니다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

【0043】명세서 전체에서, 어떤 부분이 다른 부분과 "연결(접속, 접촉, 결합)"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 부재를 사이에 두고 "간접적으로 연결"되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 구비할 수 있

다는 것을 의미한다.

【0044】본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

【0045】본 명세서에서, "모듈"은 하드웨어, 소프트웨어 또는 펌웨어로 구성된 유닛을 포함하며, 예컨대 로직, 논리 블록, 부품, 또는 회로 등의 용어와 상호 호환적으로 사용될 수 있다. 모듈은 일체로 구성된 부품 또는 하나 또는 그 이상의 기능을 수행하는 최소 단위 또는 그 일부가 될 수 있다. 예컨대 모듈은 ASIC(application-specific integrated circuit)으로 구성될 수 있다.

【0047】이하 첨부된 도면을 참조하여 본 발명의 실시예를 상세히 설명하기로 한다.

【0048】도 1은 본 발명의 실시예에 따른 SE(3) 불변성 및 물리 지식 기반 딥러닝 모델을 이용한 단백질-리간드 결합 친화도 예측 시스템 및 그 방법에 관한 개요도이다.

【0049】본 발명은 두 가지의 귀납적 편향(inductive bias)을 정의할 수 있다.

【0050】도 1의 (A)에서 보여주는 제1 귀납적 편향은, 단백질-리간드 복합체의 결합 친화도는 SE(3) 변환(예컨대, 회전 및 평행 이동)에도 불변한다는 기본 원리를 기초로 할 수 있다. 이러한 편향은 일부 예측 모델들이 3차원 좌표만 입력으로 받는 것과 비교하여 복합체의 공간 정보를 더 효율적으로 모델링할 수 있도록 할 수 있다.

【0051】여기서, 결합 친화도(Binding Affinity)란, 질병과 관련된 표적 단백질에 리간드가 결합하는 친화도를 의미하며, 결합 친화도가 높을수록 리간드가 표적 단백질에 더 강하게 결합한다.

【0052】또한, SE(3)는 Special Euclidean group in 3 dimensions의 약자로, 3차원 공간에서 물체의 회전과 평행 이동을 설명하는 군(group)을 의미한다. 예를 들어, 단백질-리간드 복합체에서 SE(3)-불변성을 적용하면, 복합체의 공간적 위치나 방향에 상관없이 일관된 결합 친화도를 예측할 수 있다. 즉, 단백질-리간드 결합 예측 모델에서 SE(3)-불변성은 복합체의 기하학적 변형에 영향을 받지 않고 일관된 결과를 제공하는 것을 의미한다.

【0053】도 1의 (B)에 나타낸 제2 귀납적 편향은 모든 가능한 기하학적 구성을 고려했을 때 단백질-리간드 복합체의 결합 자유 에너지(Binding Free Energy)가 최소화되는 지점에서 결합 상태(bound state)가 형성된다 물리화학적 원리를 기초

로 할 수 있다.

【0054】 즉, 제1 및 제2 귀납적 편향은 단백질-리간드 복합체의 맥락에서 설명될 수 있는 도메인 지식을 나타내며, 딥러닝 모델의 훈련 및 테스트 과정 모두에 동일하게 적용할 수 있는 법칙이다.

【0055】 따라서, 본 발명은 SPIN이라는 그래프 변환 모듈을 제안하며, 이 모델은 회전과 평행 이동에 불변한 SE(3)-불변 원칙과 결합 자유 에너지가 최소화된다는 물리화학적 원칙을 각각 딥러닝 모델에 기하학적 및 물리화학적 귀납적 편향으로 통합하여 높은 일반화 성능을 달성할 수 있다.

【0057】 도 2는 본 발명의 실시예에 따른 단백질-리간드 결합 친화도 예측 시스템의 구성을 도시한 블록도이다.

【0058】 단백질-리간드 결합 친화도 예측 시스템(100)은 단백질-리간드 복합체 원자 정보 및 3차원 좌표 수신부(110), 그래프 전처리 모듈(120), SE(3)-불변성 변환 모듈(130), 물리 지식 기반 네트워크 모델(140), 및 결합 친화도 출력부(150)를 포함할 수 있다.

【0059】 단백질-리간드 복합체 원자 정보 및 3차원 좌표 수신부(110)는 단백질-리간드 복합체의 원자 정보 및 3차원 좌표 데이터를 수신할 수 있다. 예를 들면, 단백질-리간드 복합체의 입체 구조 정보를 입력받을 수 있으며, 이러한 좌표 정보는 결합 친화도를 계산하기 위한 기본적인 입력 데이터로 사용될 수 있다. 좌

표 데이터는 실험적 방법(예컨대, X-Ray 결정 구조)이나 예측 모델(예컨대, AlphaFold)을 통해 획득할 수 있다.

【0061】 그래프 전처리 모듈(120)은 상기 원자 정보 및 상기 3차원 좌표 데이터를 기초로 하여 전처리 작업을 수행하여 그래프 구조를 생성할 수 있다.

【0062】 구체적으로, 그래프 전처리 모듈(120)은 상기 원자 정보의 각 원자를 노드로 정의하고, 각 노드에 원자의 다양한 물리화학적 특징을 포함시킬 수 있다. 예를 들면, 단백질 원자 특징(V_P)은 원자 유형, 아미노산 유형, 그리고 원자가 백본 원자인지 여부와 같은 정보를 포함할 수 있고, 리간드 원자 특징(V_L)은 원자 유형, 혼성화 상태, 형식 전하, 차수, 그리고 방향족 원자 여부와 같은 정보를 포함할 수 있다. 만약, N개의 단백질 원자와 M개의 리간드 원자가 포함된 경우, 이들의 위치 행렬은 X로 정의되어 좌표 정보를 포함하게 될 수 있다.

【0063】 또한, 그래프 전처리 모듈(120)은 KNN(K-Nearest Neighbors) 그래프 알고리즘을 사용하여 상기 위치 행렬에 포함된 좌표 정보를 기초로 하여 원자 간 거리를 기준으로 엣지(e_{ij})를 정의할 수 있다. 즉, 이웃 원자와의 상호작용을 엣지로 연결하여 나타낼 수 있다. 각 엣지(e_{ij})는 4차원 원-핫 벡터(one-hot vector)로 표현되며, 이를 통해 단백질 원자 간의 연결, 리간드 원자 간의 연결, 단백질-리간드 간의 연결, 리간드-단백질 간의 연결과 같은 연결 정보를 나타낼 수 있다.

【0064】따라서, 그래프 전처리 모듈(120)은 단백질-리간드 복합체의 원자 정보와 3차원 좌표 정보를 기반으로 노드와 엣지 정보를 종합하는 그래프 구조를 생성할 수 있다. 생성된 그래프 구조는 네트워크 모델의 입력으로 사용되며, 이를 통해 단백질-리간드 복합체의 구조적 정보를 반영하여 결합 친화도를 예측할 수 있다.

【0066】SE(3)-불변성 변환 모듈(130)은 3차원 공간에서 물질 또는 시스템이 회전 또는 평행이동을 하더라도 단백질-리간드 간 결합 친화도가 변하지 않는 제1 귀납적 편향을 기초로 하여 상기 그래프 구조를 임베딩 레이어 및 초기 은닉 표현으로 변환할 수 있다.

【0067】구체적으로, SE(3)-불변성 변환 모듈(130)은 상기 단백질-리간드 복합체를 구성하는 각 원자들의 물리화학적 특징(V)을 개별적으로 인코딩하여 임베딩 레이어 및 초기 은닉 표현으로 변환할 수 있다. 이 과정은 아래와 같은 수학식 1, 2 및 3 을 통해 이루어질 수 있다.

【0068】**【수학식 1】**

$$h^0=(h_P^0 \parallel h_L^0)$$

【0069】【수학식 2】

$$h_P^0 = Linear(V_P) \in R^{N \times D_E}$$

【0070】【수학식 3】

$$h_L^0 = Linear(V_L) \in R^{N \times D_E}$$

【0071】여기서 h^0는 h_P^0와 h_L^0 결합(concatenating)이고, h^0는 초기 은닉 표현을 의미한다. Linear()는 선형 변환 함수를 의미한다. R은 실수 공간을 나타내며, 해당 벡터나 행렬이 실수로 이루어져 있다는 것을 의미한다. N과 M은 각각 단백질과 리간드의 원자 개수를 의미한다. D_E는 각 원자가 가지는 물리화학적 특징의 차원수를 의미한다. 차원수는 각 원자가 몇 개의 특징(예컨대, 원자 유형, 전하 등)을 가지는지에 따라 결정된다.

【0072】이후 각 노드의 초기 은닉 표현을 업데이트 하기 위해 변환된 임베딩 레이어를 아래 수학식 4와 같이 정의할 수 있다.

【0073】【수학식 4】

$$h_i^{(l+1)} = h_i^l + \sum_{j \in \mathbf{v}, i \neq j} f_h(||x_i - x_j||, h_i^l, h_j^l, e_{ij}; \theta_h)$$

【0074】여기서, h_i^{l+1}은 i번째 노드의 l+1번째 임베딩 레이어에서의 은닉 표현을 의미한다. 이 값은 이전 l 번째 임베딩 레이어에서의 은닉 표현을 업데이트한 결과이다. h_i^l은 i번째 노드의 l번째 임베딩 레이어에서의 은닉 표현이며, 이 값은 해당 노드의 현재 상태를 나타낼 수 있다. $\|x_i - x_j\|^2$은 원자 i와 j 사이의 유클리드 거리를 의미한다. e_{ij}는 KNN(K-Nearest Neighbors) 그래프 알고리즘을 사용하여 원자 i와 j 사이의 거리를 기준으로 정의된 엣지이다. θ_h는 네트워크 모델의 학습 가능한 파라미터를 의미한다. 업데이트 함수 $f_h()$은 이웃 노드로부터 정보를 집계한 후 아래 수학식 5, 6, 7, 8을 이용하여 어텐션 연산을 통해 노드 상태를 업데이트 할 수 있는 메시지를 계산할 수 있다.

【0075】【수학식 5】

$$f_h = Attention(q_i, k_j, v_j) \cdot Linear(r_{ij})$$

【0076】【수학식 6】

$$q_i = Linear(h_i^0)$$

【0077】【수학식 7】

$$k_j = Linear([r_{ij} \| e_{ij} \| h_i \| h_j])$$

【0078】【수학식 8】

$$v_j = Linear([r_{ij} \| e_{ij} \| h_i \| h_j])$$

【0079】여기서, q_i, k_j, v_j는 어텐션(Attention) 연산을 위한 쿼리(query), 키(key), 밸류(value) 행렬을 의미한다. r_{ij}는 방사 기저 함수를 이용하여 0Å에서 10Å사이의 20개 중심에 위치한 거리 임베딩으로 정의된다. 최종 원자의 은닉 표현 h^L은 이웃 원자들의 공간 정보를 집계한 노트 상태가 되며, 단백질과 리간드 원자 간의 관계를 명시적으로 고려할 수 있다.

【0080】본 발명의 실시예에 따르면, 여기서 사용되는 쿼리(query), 키(key), 밸류(value) 임베딩은 레이어 정규화와 ReLU 활성화를 포함한 2층 MLP를 통

해 얻을 수 있다. SE(3)-그래프 변환 모듈은 16개의 레이어를 가질 수 있고, 은닉 차원과 헤드의 개수는 각각 128과 9로 설정될 수 있다. 또한 각 레이어의 활성화 함수로 swish 함수가 사용될 수 있다.

【0082】 종합하면, SE(3)-불변성 변환 모듈(130)은 단백질-리간드 복합체의 원자 정보와 3차원 좌표 정보를 각각 인코딩하여, 상기 수학식 1 내지 3에서 설명한 것처럼 임베딩 레이어로 변환하고, 이를 결합하여 초기 은닉 표현을 생성할 수 있다. 나아가, 상기 수학식 4에서 설명된 것과 같이, 회전이나 평행 이동에 영향을 받지 않는 유클리드 거리를 사용하여 기하학적 변환에도 불구하고 동일한 값을 유지하도록 보장할 수 있다. 구체적으로, 상기 수학식 5 내지 8에서 설명된 어텐션 메커니즘 기반으로 노드 간의 상호작용을 계산하고 이를 통해 노드(원자)의 상태를 업데이트할 수 있다. 이때, r_{ij}는 두 원자 간의 거리 정보를 반영하는데, 이 값은 방사 기저 함수로 처리되어 회전과 평행 이동에 대해 불변성을 유지할 수 있다.

【0083】 즉, 본 발명은 이러한 어텐션 연산을 통해 모델은 이웃 노드(원자)들의 공간적 관계를 효과적으로 반영하면서도 SE(3)-불변성을 유지할 수 있다. 이러한 SE(3)-불변성에 대한 증명은 도3 및 수학식 11을 참고하여 후술하도록 한다.

【0085】 물리 지식 기반 네트워크 모델(140)은 상기 임베딩 레이어 및 상기 은닉 표현을 입력으로 하여 단백질-리간드 간 결합 친화도를 예측할 수 있다.

【0086】상기 물리 지식 기반 네트워크 모델은 GNN(Graph Neural Network) 기반 방법 등으로 구현될 수 있다.

【0087】또한, 상기 물리 지식 기반 네트워크 모델은 단백질-리간드 복합체의 결합 자유 에너지가 최소화되는 지점에서 결합 상태가 형성된다는 제2 귀납적 편향을 기초로 하는 물리 지식 기반 네트워크 모델이다.

【0088】상기 제2 귀납적 편향을 기초로 하는 물리 지식 기반 네트워크 모델의 학습 과정 및 결합 친화도 예측 과정은 도 3을 참고하여 후술하도록 한다.

【0090】결합 친화도 출력부(150)는 상기 물리 지식 기반 네트워크 모델을 이용한 결합 친화도 예측값을 출력할 수 있다.

【0091】상기 결합 친화도 출력부는 결합 친화도를 사용자에게 제공하거나, 약물 개발 과정에서 활용될 수 있도록 데이터를 외부 시스템으로 전송하는 역할을 할 수 있다.

【0092】또한, 상기 결합 친화도 출력부는 예측된 결합 친화도를 화면에 시각적으로 표시하거나, 출력된 값을 외부 시스템으로 전송하여 약물 후보군 중 결합력이 높은 물질을 식별하는데 사용될 수 있다.

【0093】또한, 상기 결합 친화도 출력부는 실시간으로 결합 친화도를 분석할 수 있으며, 해당 데이터를 바탕으로 추가적인 피드백을 제공하거나, 예측값의 신뢰성을 확인하기 위해 해석 가능성 분석을 수행할 수도 있다.

【0094】도 3은 본 발명의 실시예에 따른 물리 지식 기반 네트워크 모델의 학습 과정 및 결합 친화도 예측 방법을 보여주는 참고도이다.

【0095】물리 지식 기반 네트워크 모델의 학습 과정은 단백질-리간드 복합체 원자 정보 및 3차원 좌표 준비 단계(S110), 그래프 전처리 단계(S120), SE(3)-불변성 변환 단계(S130), 단백질-리간드 상호작용 매트릭스 계산 단계(S140), 결합 친화도 예측 단계(S150) 및 물리화학적 귀납적 편향을 기초로 한 손실함수 계산 단계(S150)를 포함할 수 있다.

【0096】단백질-리간드 복합체 원자 정보 및 3차원 준비 단계(S110)에서, 단백질-리간드 복합체의 원자 정보 및 3차원 좌표를 수신하여 준비할 수 있다. 본 발명의 실시예에 따르면, 단백질-리간드 복합체의 구조와 그에 따른 결합 친화도 정보를 제공하는 대규모 데이터베이스인 PDBbind로부터 단백질-리간드 복합체의 원자 정보 및 3차원 좌표 정보, 각 단백질-리간드 복합체에 대한 실험적으로 얻어진 결합 친화도 등을 제공받을 수 있다. 상기 PDBbind v2020 데이터셋은 총 19,443개의 샘플로 구성되어 있다.

【0097】그래프 전처리 단계(S120)에서, 상기 원자 정보 및 상기 3차원 좌표 데이터가 전처리되어 그래프로 생성될 수 있다. 본 발명의 실시예에 따르면, 단백질-리간드 복합체를 구성하는 원자들은 노드로 정의되며, 원자 간의 연결은 엣지로 정의되어 전체 단백질-리간드 구조를 하나의 그래프 구조로 나타낼 수 있다.

【0098】SE(3)-불변성 변환 단계(S130)에서, 상기 노드와 상기 엣지의 물리화학적 특징들이 SE(3)-그래프 변환 모듈을 임베딩 레이어 및 은닉 표현으로 변환되고, 어텐션 함수를 통해 업데이트될 수 있다.

【0099】단백질-리간드 상호작용 매트릭스 계산 단계(S140)에서, 상기 은닉 표현을 이용하여 단백질-리간드 상호작용 매트릭스 H가 아래 수학식 9를 이용하여 계산될 수 있다.

【0100】**【수학식 9】**

$$H=(h_M^L \cdot h_P^L)^T$$

【0101】여기서, h_M^L은 리간드를 구성하는 원자의 최종 은닉 표현을 의미하며, h_P^L은 단백질을 구성하는 원자의 최종 은닉 표현을 의미한다.

【0102】결합 친화도 예측 단계(S150)에서, 단백질-리간드 결합 친화도는 단백질과 리간드 간의 원자 쌍별 반 데르 발스(Van der Waals; VDW) 상호작용 에너지의 합을 기초로 예측될 수 있다. 본 발명의 실시예에 따르면, 반 데르 발스 상호작용 에너지의 합은 Lennard-Jones 잠재 함수에 기초한 아래 수학식 10을 이용하여 계산할 수 있다.

【0103】【수학식 10】

$$E^{VDW}=\sum_{i,j}C_{ij}[(\frac{u_{ij}+H_{ij}}{||x_i-x_j||})^{12}-2(\frac{u_{ij}+H_{ij}}{||x_i-x_j||})^{6}]$$

【0104】여기서, E^{VDW}는 반 데르 발스 상호작용 에너지의 합을 의미하고, C_{ij}는 i번째 리간드 원자와 j번째 단백질 원자 간의 상호작용 강도 또는 가중치(상수)를 의미한다. u_{ij}는 i번째 리간드 원자와 j번째 단백질 원자의 반 데르 발스 반경을 의미한다. H는 단백질-리간드 상호작용 매트릭스를 의미한다.

【0105】각 원자의 반 데르 발스 반경은 X-score 파라미터에서 얻을 수 있다. 2는 Lennard-Jones 잠재 함수의 식에서 인력과 반발력 사이의 균형을 맞추기 위한 상수를 나타내며, 12와 6은 각각 원자 간의 강한 반발력과 약한 인력을 나타내는 항을 의미한다.

【0106】따라서, E^{VDW}는 SE(3)-불변 변환 모듈로부터 파라미터화된 은닉 표현을 이용하여 계산된 단백질-리간드 상호작용 매트릭스 H와 단백질-리간드 복합체의 물리화학적 정보를 사용하여 계산될 수 있다.

【0107】이때, 결합 친화도 예측을 위한 전반적인 프레임워크가 3D 객체의 회전 및 평행 이동에 대해 불변인 점을 증명하기 위해, Tg(x)=Rx+b로 정의하면, 아래 수학식 11이 도출된다. 여기서 R은 3X3 회전 행렬이고, b는 3차원 평행 이동 벡

터를 의미한다.

【0108】 【수학식 11】

$$\sum_{i,j}[(\frac{u_{ij}+H_{ij}}{||x_i-x_j||^2})^{12}-(\frac{u_{ij}+H_{ij}}{||x_i-x_j||^2})^{6}]$$

【0109】
$$=\sum_{i,j}[(\frac{u_{ij}+H_{ij}}{T_g||x_i-x_j||^2})^{12}-(\frac{u_{ij}+H_{ij}}{T_g||x_i-x_j||^2})^{6}]$$

【0110】
$$=\sum_{i,j}[(\frac{u_{ij}+H_{ij}}{||(Rx_i+b)-(Rx_j+b)||^2})^{12}-(\frac{u_{ij}+H_{ij}}{||(Rx_i+b)-(Rx_j+b)||^2})^{6}]$$

【0111】
$$=\sum_{i,j}[(\frac{u_{ij}+H_{ij}}{||(x_i-x_j)^TR^TR(x_i-x_j)||^2})^{12}-(\frac{u_{ij}+H_{ij}}{||(x_i-x_j)^TR^TR(x_i-x_j)||^2})^{6}]$$

【0112】
$$=\sum_{i,j}[(\frac{u_{ij}+H_{ij}}{||(x_i-x_j)^TI(x_i-x_j)||^2})^{12}-(\frac{u_{ij}+H_{ij}}{||(x_i-x_j)^TI(x_i-x_j)||^2})^{6}]$$

【0113】
$$=\sum_{i,j}[(\frac{u_{ij}+H_{ij}}{||x_i-x_j||^2})^{12}-(\frac{u_{ij}+H_{ij}}{||x_i-x_j||^2})^{6}]$$

【0115】물리화학적 귀납적 편향을 기초로 한 손실함수 계산 단계(S160)에서, 물리 지식 기반 네트워크 모델은 아래 수학식 12와 같은 오차 최소화 및 아래 수학식 13과 같은 물리 법칙을 모두 만족하는 방향으로 손실함수를 계산하여 손실함수를 최소화하는 방향으로 최적화될 수 있다.

【0116】【수학식 12】

$$L_d = \sum_N (y - \hat{y})^2, where, \hat{y} = \sigma \cdot E^{VDW}$$

【0117】【수학식 13】

$$L_P = \sum_N \sum_{i,j} \left[\frac{\partial c_{ij} \left(\frac{r_i + M_{ij}}{||x_i - x_j||} \right)^{12} - 2 \left(\frac{M_{ij}}{||x_i - x_j||} \right)^6}{\partial ||x_i - x_j||} \right]^2$$

【0118】여기서, y는 실제 실험 데이터를 기반으로 한 결합 친화도를 의미하며, $\hat{y}$는 예측된 결합 친화도를 의미한다. σ는 학습 가능한 파라미터를 의미한다. r_i는 리간드 원자의 반지름을 의미한다. M은 단백질-리간드 상호작용 매트릭스를 의미한다.

【0119】상기 수학식 13은 단백질-리간드 간 예측된 결합 친화도가 물리 법칙을 만족하도록 하는 수학식을 적용하는 것으로, 리간드와 단백질 원자 사이의 거리에 대한 결합 자유 에너지의 미분값이 0이 되도록 보장한다. 이는 실험적으로 규

명된 구조로 모델링된 복합체가 해당 반응 좌표에서 결합 자유 에너지가 최소화된 지점에 위치해 있다는 사전 지식을 반영한 것이다. 따라서, 상기 물리 지식 기반 네트워크 모델의 전체 손실함수(L)은 L_d와 L_p의 합으로 정의될 수 있다.

【0120】 즉, 상기 물리 지식 기반 네트워크 모델은 L_d 및 L_p의 합인 손실함수가 최소가 되도록 하여 반복학습될 수 있다. 이를 통해 본 발명은 결합 에너지를 최소화시키는 물리화학적 귀납적 편향을 기초로 하여 단백질-리간드 간 결합 친화도를 예측할 수 있다.

【0121】 따라서, 본 발명은 단백질과 리간드 간의 결합 자유 에너지가 최소화되는 지점에서 결합이 형성된다는 물리화학적 원칙을 반영하여 보다 현실적으로 단백질-리간드 간 결합 에너지를 예측할 수 있다. 또한, SE(3)-불변성과 같은 기하학적 귀납적 편향을 통합하여 효율적으로 결합 에너지 예측을 수행할 수 있다. 결국, 본 발명은 결합 친화도(Binding Affinity, BA)를 보다 정확하게 제공할 수 있다.

【0123】 표 1은 본 발명의 일 실시예 따른 SPIN 모델과 기존의 기준 모델들과 두 가지 벤치마크 데이터셋에서 비교한 결과를 나타낸다.

【0124】 PDBbind v2020 데이터셋으로 훈련된 본 발명의 일 실시예와 비교 모델들의 성능을 평가하기 위해 CASF-2016 및 CSAR-HiQ 벤치마크 세트를 사용하였다. CASF-2016 및 CSAR-HiQ 벤치마크 세트는 PDBbind와 유사하게 단백질-리간드 복합체

의 3D 결정 구조와 그들의 결합 친화도를 제공한다. CASF-2016 및 CSAR-HiQ 벤치마크 세트는 각각 285개와 343개의 샘플로 구성되어 있다.

【0125】본 발명의 실시예는 두 벤치마크 데이터셋에서 모든 네 가지 지표(RMSE, MAE, SD, R)에서 최고의 성능을 보여준다. RMSE(Root Mean Square Error)는 평균 제곱근 오차이고, MAE(Mean Absolute Erroe)는 평균 절대 오차이며, SD(Standard Deviation)는 표준 편차, R(Correlation Coefficient)은 상관 계수이다.

【0126】Pafnucy와 같은 CNN 기반 방법은 복합체의 회전 및 평행 이동에 대해 불변성을 가지지 않기 대문에 벤치마크 데이터셋에서 일반화 성능이 부족하고, SGCN은 GNN 기반 방법으로 공간 구조를 활용하지만, 좌표를 직접 입력하여 예측하기 때문에 CNN 기반 모델과 유사한 문제를 겪으며, 회전 및 평행 이동에 민감하여 마찬가지로 일반화 성능이 부족하다. Grpahtrans, NL-GCN, PIGNet, CMPNN 모델은 3차원 공간의 기하학적 특성을 모델링하지 않고 연결성 정보에만 의존하기 때문에 결합 친화도 예측에 적합하지 않다. SIGN과 GIANT 모델은 각도 정보를 고려하여 단백질과 리간드의 공간 정보를 더 잘 포착하기 위해 다양한 모듈을 포함하고 있지만, 이들은 훈련 데이터에서 얻은 미리 정의된 규칙에 기반한 방법을 따르기 때문에 훈련 데이터에 없는 다양한 복합 구조를 처리하는 유연성이 부족하다.

【0127】특히, 본 발명의 일 실시예인 SPIN 모델은 모든 비교 모델들에서 일반화에 어려움을 겪었던 CSAR 데이터셋에서 30% 이상의 성능 개선을 보여준다. 이는 귀납적 편향을 예측 모델에 도입하는 것이 보지 못한 데이터셋에 일반화 성능

향상에 중요한 역할을 한다는 것을 시사한다.

【0128】흥미로운 점은 PIGNet 모델도 물리화학적 법칙을 예측 모델의 귀납적 편향으로 정의한 모델로서, 다른 비교 모델들에 비해 CSAR 세트에서 상대적으로 높은 성능을 달성했다. 이는 다양한 복합체에 대한 결합 친화도를 예측하는 응용 측면에서, 물리화학적 정보를 예측 모델의 귀납적 편향으로 정의하는 것의 중요성을 재확인시켜줄 수 있다. 그러나 PIGNet 모델은 복합 구조의 공간 정보를 명확하게 모델링하지 못했기 때문에 CASF-2016에서 열등한 성능을 보였으며, 이는 결합 친화도를 예측하기 위한 모델로서 적합하지 않음을 나타낸다.

【0129】**【표 1】**

방법	CASF-2016 set				CSAR-HiQ set			
	RMSE (↓)	MAE (↓)	SD (↓)	R (↑)	RMSE (↓)	MAE (↓)	SD (↓)	R (↑)
LR	1.675 (0.000)	1.358 (0.000)	1.612 (0.000)	0.671 (0.000)	2.071 (0.000)	1.622 (0.000)	1.973 (0.000)	0.652 (0.000)
SVR	1.555 (0.000)	1.264 (0.000)	1.493 (0.000)	0.727 (0.000)	1.995 (0.000)	1.553 (0.000)	1.911 (0.000)	0.679 (0.000)
RF-Score	1.446 (0.000)	1.161 (0.000)	1.409 (0.000)	0.789 (0.033)	1.939 (0.103)	1.562 (0.094)	1.885 (0.071)	0.686 (0.027)
Pafnucy	1.585 (0.013)	1.284 (0.021)	1.507 (0.013)	0.744 (0.027)	1.947 (0.065)	1.562 (0.094)	1.885 (0.071)	0.686 (0.027)
OnionNet	1.457 (0.031)	1.198 (0.035)	1.382 (0.007)	0.727 (0.019)	1.902 (0.050)	1.472 (0.067)	1.881 (0.077)	0.686 (0.027)
SGCN	1.583 (0.033)	1.250 (0.036)	1.528 (0.012)	0.688 (0.011)	1.902 (0.033)	1.472 (0.067)	1.881 (0.077)	0.686 (0.027)
GraphTrans	1.536 (0.013)	1.198 (0.034)	1.482 (0.020)	0.720 (0.019)	1.871 (0.028)	1.517 (0.071)	1.831 (0.028)	0.716 (0.011)
NL-GCN	1.516 (0.019)	1.198 (0.034)	1.462 (0.015)	0.720 (0.019)	1.857 (0.011)	1.481 (0.052)	1.817 (0.028)	0.716 (0.011)
GNN-DTI	1.492 (0.025)	1.133 (0.026)	1.432 (0.010)	0.752 (0.009)	1.532 (0.000)	1.472 (0.077)	1.742 (0.066)	0.701 (0.035)
PIGNet	1.482 (0.000)	1.133 (0.026)	1.432 (0.010)	0.772 (0.013)	1.742 (0.000)	1.427 (0.010)	1.832 (0.028)	0.671 (0.000)
MAT	1.457 (0.037)	1.134 (0.037)	1.432 (0.020)	0.727 (0.009)	1.742 (0.000)	1.427 (0.010)	1.832 (0.028)	0.671 (0.000)
CMPNN	1.307 (0.028)	1.100 (0.029)	1.322 (0.019)	0.790 (0.029)	1.742 (0.000)	1.362 (0.035)	1.742 (0.066)	0.731 (0.035)

SIGN	1.316 (0.031)	1.080 (0.029)	1.323 (0.019)	0.777 (0.029)	1.753 (0.032)	1.362 (0.035)	1.742 (0.066)	0.742 (0.021)
GIANT	1.269 (0.028)	0.999 (0.018)	1.293 (0.011)	0.800 (0.019)	1.743 (0.021)	1.342 (0.018)	1.598 (0.059)	0.742 (0.021)
CAPLA	1.389 (0.018)	1.109 (0.021)	1.230 (0.025)	0.775 (0.021)	1.742 (0.021)	1.362 (0.035)	1.742 (0.066)	0.731 (0.035)
SPIN (본 발명 실시예)	**1.258 (0.013)**	**0.996 (0.021)**	**1.229 (0.014)**	**0.826 (0.027)**	**1.238 (0.022)**	**0.999 (0.034)**	**1.270 (0.022)**	**0.800 (0.017)**

【0131】 도 4는 본 발명의 일 실시예에 따른 SPIN 모델의 각 귀납적 편향 제거에 따른 성능 비교 결과를 보여주는 참고도이다.

【0132】 본 발명의 일 실시예에 따르면, SPIN 모델의 각 구성 요소의 중요성을 검증하기 위해 성능 분석 연구(Ablation study)를 진행하였다. SPIN(w/o G)은 SE(3)-불변성을 만족하는 기하학적 귀납적 편향을 제거한 모델이며, SPIN(w/o P)는 결합 자유 에너지 최소화에 관한 물리화학적 귀납적 편향을 제거한 모델이다. 마지막으로, SPIN(w/o GP)는 두 가지 종류의 귀납적 편향을 모두 제거한 모델이다.

【0133】 두 가지 벤치마크 데이터셋에서 완전한 SPIN 모델을 포함한 네 가지 변형 모델의 성능을 측정한 결과를 도4에 도시하였다. 완전한 모델의 성능이 네 가지 지표(RMSE, MAE, SD, R)에서 모두 우수하게 나타났으므로, 기하학적 및 물리화학적 귀납적 편향 모두 결합 친화도를 예측하는 데 중요한 역할을 한다는 것을 확인할 수 있다. 특히, SPIN(w/o G)과 SPIN(w/o P)의 성능을 비교했을 때, 단백질-리간드 복합체의 회전 및 평행 이동에 대해 결합 친화도가 불변하는 조건을 만족하는 기하학적 귀납적 편향이 예측 성능에 가장 중요한 구성 요소임을 확인할 수 있다.

흥미로운 점은, 물리화학적 귀납적 편향이 복합체의 기하학적 정보를 합리적으로 모델링하는 조건이 충족되었을 때 더 나은 시너지를 발휘한다는 것이다. 이 결과는 결합 자유 에너지 계산이 3차원 공간에서 복합체의 위치를 기반으로 이루어진다는 점을 고려했을 때, 두 가지 종류의 귀납적 편향 간의 상호작용에 대한 중요한 통찰을 제공할 수 있다.

【0135】 도 5는 본 발명의 일 실시예에 따른 SPIN 모델을 이용한 가상 스크리닝 실험에서 순위 결정력을 비교한 결과를 보여주는 참고도이다.

【0136】 단백질-리간드 복합체의 결합 친화도를 예측하는 것과 더불어, 특정 표적 단백질에 대한 리간드들의 결합 강도를 기반으로 정확한 순위를 매기는 것은 약물 개발에서 매우 중요하다. 이는 가장 유망한 약물 후보 물질을 올바르게 목록화함으로써 약물 개발 과정을 더욱 효율적으로 만들 수 있다. 이를 위해, CASF-2016 벤치마크 세트에 대한 기존 연구의 방법론을 채택하여 SPIN 모델의 실용성을 검증하였다.

【0137】 구체적으로, 순위 결정력을 측정하는데, 이는 결합 친화도를 기준으로 특정 표적 단백질의 알려진 리간드들을 정확하게 순위화하는 능력을 의미한다. 이 방법을 통해 SPIN이 결합 친화도를 예측하는 데 있어 얼마나 효과적인지 엄밀하게 평가할 수 있다.

【0138】CASF-2016 세트는 57개의 단백질 클러스터로 구성되어 있으며, 각 클러스터는 동일한 단백질에 결합하지만 결합 친화도가 크게 다른 다섯 개의 복합체를 포함한다. 각 클러스터는 리간드의 결합 친화도에 기반하여 미리 정의된 순위를 가지고 있다. 우리는 각 클러스터에서 예측 모델이 이러한 순위를 얼마나 정확하게 추론하는지 Spearman 순위 상관 계수를 사용하여 측정하였다. 모든 클러스터의 평균값이 이 실험에서 순위 결정력으로 정의된다.

【0139】비교 모델로서 실제 도킹 프로토콜에서 사용하는 여러 스코어링 함수를 선택하였다. 도 5에 나타난 바와 같이, SPIN은 다른 도킹 프로그램의 스코어링 함수보다 우수한 순위 결정력을 보여주는 것을 확인할 수 있다. 본 발명은 실제 약물 개발 시나리오에서 후보 화합물의 우선순위를 효과적으로 결정할 수 있음을 나타낸다.

【0141】도 6은 본 발명의 일 실시예에 따른 SPIN 모델을 이용한 단백질-리간드 상호작용 분석 및 해석 가능성을 검증하는 참고도이다.

【0142】예측 모델의 출력 결과가 약물 개발 과정에서 신뢰할 수 있는 방식으로 적극적으로 사용될 수 있는지를 확인하기 위해서는 예측 모델의 해석 가능성을 분석하는 것이 중요하다. 실제 약물 개발에서 예측 성능과는 별개로, 예측된 결합 친화도의 근거가 모호할 경우 결과 검증이 어려워지며, 이로 인해 리드 최적화와 같은 후속 과정의 효율적인 개발이 저해될 수 있다. 이를 평가하기 위해 SPIN

프레임워크 내에서 단백질과 리간드 원자 간의 상호작용을 포함하는 쌍별 상호작용 매트릭스 H를 사용하였다. 단백질-리간드 상호작용 매트릭스 Hij는 단백질 원자 i와 리간드 원자 j 사이의 결합 에너지를 나타내며, 값이 낮을수록 두 하위 구조 간의 상호작용이 강하다는 것을 의미한다.

【0143】훈련된 예측 모델의 출력 상호작용 매트릭스에서 에너지 값이 가장 낮은 10%에 해당하는 단백질의 아미노산을 추출하였다. SPIN의 해석 가능성을 입증하기 위해, 도 6의 (A)에 표시된 3bu1(PDB ID) 단백질-리간드 복합체에 대해 강한 상호작용을 하는 아미노산을 시각화하고 분석을 집중하였다. 이 추출된 아미노산들이 실제로 분자 간 상호작용에 관여하고 있는지 확인하기 위해, Discovery Studio의 분자 간 상호작용 프로파일러의 결과와 비교하였다. SPIN 상호작용 매트릭스에서 에너지 값이 가장 낮은 10%에 해당하는 아미노산은 21.A TYR, 37.A VAL, 51.A TYR, 94.A ASP, 96.A VAL, 105.A TRP로 확인되었다. 이러한 잔기들은 SPIN이 예측 모델링을 수행하는 동안 단백질-리간드 상호작용에서 중요한 역할을 하는 것으로 추정된다.

【0144】분석 결과, 이러한 아미노산들이 Discovery Studio 프로파일링 결과와 정확하게 일치함이 확인되었다. 이 일치는 SPIN이 결합 친화도를 정확하게 예측할 뿐만 아니라 이러한 예측을 뒷받침하는 생물학적으로 관련된 상호작용을 신뢰성 있게 식별할 수 있음을 강하게 시사한다. 생물학적 시스템에서 복잡한 상호작용을 구별하고 합리적으로 설명할 수 있는 이 능력은 SPIN의 해석 가능성을 강조하며, 생화학 연구 분야에서 예측 모델링에 유용하게 적용될 수 있음을 확인시켜 준다.

이 검증을 통해 SPIN이 예측과 해석을 동시에 제공하여 계산적 약물 발견 분야에 효과적으로 기여할 수 있는 잠재력을 보여준다.

【0146】 따라서, 본 발명은 SPIN과 같은 SE(3)-불변성 및 물리 지식 기반 네트워크 모델을 이용한 단백질-리간드 결합 친화도 예측 시스템을 제공할 수 있다.

【0147】 또한, 본 발명은 여러 귀납적 편향을 도입하여 제한된 데이터로부터 우수한 일반화 성능을 달성할 수 있다.

【0148】 또한, 본 발명은 결합 친화도가 3차원 공간에서의 회전 및 평행 이동에 관계없이 일정하게 유지된다는 기하학적 귀납적 편향과 결합이 최소 결합 자유 에너지에서 발생한다는 물리화학적 귀납적 편향을 통합하여 네트워크 모델에 반영하고 있다.

【0149】 또한, 2 개의 벤치마크 세트를 통해 엄격한 검증을 수행한 결과, 본 발명의 단백질-리간드 간 결합 친화도 예측 성능의 우수성, 실제 약물 개발 과정에서의 가상 스크리닝에서의 실용성 및 해석 가능성을 시각화하여 예측값의 신뢰성을 확인할 수 있다.

【0151】 도 7은 본 발명의 실시예에 따른 단백질-리간드 결합 친화도 예측 방법을 보여주는 흐름도이다.

【0152】단백질-리간드 결합 친화도 예측 방법은 단백질-리간드 복합체 원자 정보 및 3차원 좌표 수신 단계(S210), 그래프 전처리 단계(S220), SE(3)-불변성 변환 단계(S230), 물리 지식 기반 네트워크 모델을 이용한 결합 친화도 예측 단계(S240)를 포함할 수 있다.

【0153】단백질-리간드 복합체 원자 정보 및 3차원 좌표 수신 단계(S210)에서, 단백질-리간드 결합 친화도 예측 시스템은 단백질-리간드 복합체의 원자 정보 및 3차원 좌표 데이터를 수신할 수 있다. 예를 들면, 단백질-리간드 복합체의 입체 구조 정보를 입력받을 수 있으며, 이러한 좌표 정보는 결합 친화도를 계산하기 위한 기본적인 입력 데이터로 사용될 수 있다. 좌표 데이터는 실험적 방법(예컨대, X-Ray 결정 구조)이나 예측 모델(예컨대, AlphaFold)을 통해 획득할 수 있다.

【0155】그래프 전처리 단계(S220)에서, 단백질-리간드 결합 친화도 예측 시스템은 상기 원자 정보 및 상기 3차원 좌표 데이터를 기초로 하여 전처리 작업을 수행하여 그래프 구조를 생성할 수 있다.

【0156】구체적으로, 그래프 전처리 단계(S220)에서는, 상기 원자 정보의 각 원자가 노드로 정의되고, 각 노드에 원자의 다양한 물리화학적 특징이 포함될 수 있다. 예를 들면, 단백질 원자 특징(V_P)은 원자 유형, 아미노산 유형, 그리고 원자가 백본 원자인지 여부와 같은 정보를 포함할 수 있고, 리간드 원자 특징(V_L)은 원자 유형, 혼성화 상태, 형식 전하, 차수, 그리고 방향족 원자 여부와 같은 정보를

포함할 수 있다. 만약, N개의 단백질 원자와 M개의 리간드 원자가 포함된 경우, 이들의 위치 행렬은 X로 정의되어 좌표 정보를 포함하게 될 수 있다.

【0157】 또한, 그래프 전처리 단계(S220)에서는, KNN(K-Nearest Neighbors) 그래프 알고리즘을 사용하여 상기 위치 행렬에 포함된 좌표 정보를 기초로 하여 원자 간 거리를 기준으로 엣지(e_{ij})가 정의될 수 있다. 즉, 이웃 원자와의 상호작용을 엣지로 연결하여 나타낼 수 있다. 각 엣지(e_{ij})는 4차원 원-핫 벡터(one-hot vector)로 표현되며, 이를 통해 단백질 원자 간의 연결, 리간드 원자 간의 연결, 단백질-리간드 간의 연결, 리간드-단백질 간의 연결과 같은 연결 정보를 나타낼 수 있다.

【0158】 따라서, 그래프 전처리 단계(S220)에서는, 단백질-리간드 복합체의 원자 정보와 3차원 좌표 정보를 기반으로 노드와 엣지 정보를 종합하는 그래프 구조가 생성될 수 있다. 생성된 그래프 구조는 네트워크 모델의 입력으로 사용되며, 이를 통해 단백질-리간드 복합체의 구조적 정보를 반영하여 결합 친화도를 예측할 수 있다.

【0160】 SE(3)-불변성 변환 단계(S230)에서, 단백질-리간드 결합 친화도 예측 시스템은 3차원 공간에서 물질 또는 시스템이 회전 또는 평행이동을 하더라도 단백질-리간드 간 결합 친화도가 변하지 않는 제1 귀납적 편향을 기초로 하여 상기 그래프 구조를 임베딩 레이어 및 초기 은닉 표현으로 변환할 수 있다.

【0161】구체적으로, SE(3)-불변성 변환 단계(S230)에서는 상기 단백질-리간드 복합체를 구성하는 각 원자들의 물리화학적 특징(V)이 개별적으로 인코딩되어 임베딩 레이어 및 초기 은닉 표현으로 변환될 수 있다. 이 과정은 아래와 같은 수학식 14, 15 및 16 을 통해 이루어질 수 있다.

【0162】**【수학식 14】**

$$h^0=(h_P^0 \parallel h_L^0)$$

【0163】**【수학식 15】**

$$h_P^0=Linear(V_P)\in R^{N\times D_E}$$

【0164】**【수학식 16】**

$$h_L^0=Linear(V_L)\in R^{N\times D_E}$$

【0165】여기서 h^0는 h_P^0와 h_L^0 결합(concatenating)이고, h^0는 초기 은닉 표현을 의미한다. Linear()는 선형 변환 함수를 의미한다. R은 실수 공간을 나타내며, 해당 벡터나 행렬이 실수로 이루어져 있다는 것을 의미한다. N과 M은 각각 단백질과 리간드의 원자 개수를 의미한다. D_E는 각 원자가 가지는 물리화학적 특

징의 차원수를 의미한다. 차원수는 각 원자가 몇 개의 특징(예컨대, 원자 유형, 전하 등)을 가지는지에 따라 결정된다.

【0166】 이후 각 노드의 초기 은닉 표현을 업데이트 하기 위해 변환된 임베딩 레이어를 아래 수학식 17과 같이 정의할 수 있다.

【0167】 **【수학식 17】**

$$h_i^{(l+1)} = h_i^l + \sum_{j \in v, i \neq j} f_h(||x_i - x_j||, h_i^l, h_j^l, e_{ij}; \theta_h)$$

【0168】 여기서, h_i^{l+1}은 i번째 노드의 l+1번째 임베딩 레이어에서의 은닉 표현을 의미한다. 이 값은 이전 l 번째 임베딩 레이어에서의 은닉 표현을 업데이트한 결과이다. h_i^l은 i번째 노드의 l번째 임베딩 레이어에서의 은닉 표현이며, 이 값은 해당 노드의 현재 상태를 나타낼 수 있다. $\|x_i - x_j\|^2$은 원자 i와 j 사이의 유클리드 거리를 의미한다. e_{ij}는 KNN(K-Nearest Neighbors) 그래프 알고리즘을 사용하여 원자 i와 j 사이의 거리를 기준으로 정의된 엣지이다. θ_h는 네트워크 모델의 학습 가능한 파라미터를 의미한다. 업데이트 함수 $f_h()$은 이웃 노드로부터 정보를 집계한 후 아래 수학식 18, 19, 20, 21을 이용하여 어텐션 연산을 통해 노드 상태를 업데이트할 수 있는 메시지를 계산할 수 있다.

【0169】【수학식 18】

$$f_h = Attention(q_i, k_j, v_j) \cdot Linear(r_{ij})$$

【0170】【수학식 19】

$$q_i = Linear(h_i^0)$$

【0171】【수학식 20】

$$k_j = Linear([r_{ij} \| e_{ij} \| h_i \| h_j])$$

【0172】【수학식 21】

$$v_j = Linear([r_{ij} \| e_{ij} \| h_i \| h_j])$$

【0173】여기서, q_i, k_j, v_i는 어텐션(Attention) 연산을 위한 쿼리(query), 키(key), 밸류(value) 행렬을 의미한다. r_{ij}는 방사 기저 함수를 이용하여 0Å에서 10Å사이의 20개 중심에 위치한 거리 임베딩으로 정의된다. 최종 원자의 은닉 표현

h^L은 이웃 원자들의 공간 정보를 집계한 노트 상태가 되며, 단백질과 리간드 원자 간의 관계를 명시적으로 고려할 수 있다.

【0174】본 발명의 실시예에 따르면, 여기서 사용되는 쿼리(query), 키(key), 밸류(value) 임베딩은 레이어 정규화와 ReLU 활성화를 포함한 2층 MLP를 통해 얻을 수 있다. SE(3)-그래프 변환 모듈은 16개의 레이어를 가질 수 있고, 은닉 차원과 헤드의 개수는 각각 128과 9로 설정될 수 있다. 또한 각 레이어의 활성화 함수로 swish 함수가 사용될 수 있다.

【0175】종합하면, SE(3)-불변성 변환 단계(S230)에서는 단백질-리간드 복합체의 원자 정보와 3차원 좌표 정보가 각각 인코딩되어, 상기 수학식 14 내지 16에서 설명한 것처럼 임베딩 레이어로 변환되고, 이를 결합하여 초기 은닉 표현이 생성될 수 있다. 나아가, 상기 수학식 17에서 설명된 것과 같이, 회전이나 평행 이동에 영향을 받지 않는 유클리드 거리를 사용하여 기하학적 변환에도 불구하고 동일한 값을 유지하도록 보장할 수 있다. 구체적으로, 상기 수학식 18 내지 21에서 설명된 어텐션 메커니즘 기반으로 노드 간의 상호작용을 계산하고 이를 통해 노드(원자)의 상태를 업데이트할 수 있다. 이때, r_{ij}는 두 원자 간의 거리 정보를 반영하는데, 이 값은 방사 기저 함수로 처리되어 회전과 평행 이동에 대해 불변성을 유지할 수 있다.

【0176】즉, 본 발명은 이러한 어텐션 연산을 통해 모델은 이웃 노드(원자)들의 공간적 관계를 효과적으로 반영하면서도 SE(3)-불변성을 유지할 수 있다.

【0178】물리 지식 기반 네트워크 모델을 이용한 결합 친화도 예측 단계(S240)에서, 단백질 리간드 결합 친화도 예측 시스템은 상기 임베딩 레이어 및 상기 은닉 표현을 입력으로 하여 물리 지식 기반 네트워크 모델을 이용하여 단백질-리간드 간 결합 친화도를 예측할 수 있다.

【0179】상기 물리 지식 기반 네트워크 모델은 단백질-리간드 복합체의 결합 자유 에너지가 최소화되는 지점에서 결합 상태가 형성된다는 제2 귀납적 편향을 기초로 하는 물리 지식 기반 네트워크 모델이다.

【0180】상기 물리 지식 기반 네트워크 모델은 상기 은닉 표현을 이용하여 아래 수학식 22를 통해 단백질-리간드 상호작용 매트릭스 H를 계산할 수 있다.

【0181】**【수학식 22】**

$$H=(h_M^L \cdot h_P^L)^T$$

【0182】여기서, h_M^L은 리간드를 구성하는 원자의 최종 은닉 표현을 의미하며, h_P^L은 단백질을 구성하는 원자의 최종 은닉 표현을 의미한다.

【0183】한편, 단백질-리간드 결합 친화도는 단백질과 리간드 간의 원자 쌍별 반 데르 발스(Van der Waals; VDW) 상호작용 에너지의 합을 기초로 예측될 수 있다. 본 발명의 실시예에 따르면, 반 데르 발스 상호작용 에너지의 합은 Lennard-Jones 잠재 함수에 기초한 아래의 수학식 23을 이용하여 계산할 수 있다.

【0184】【수학식 23】

$$E^{VDW}=\sum_{i,j}C_{ij}[(\frac{u_{ij}+H_{ij}}{||x_i-x_j||})^{12}-2(\frac{u_{ij}+H_{ij}}{||x_i-x_j||})^{6}]$$

【0185】여기서, E^{VDW}는 반 데르 발스 상호작용 에너지의 합을 의미하고, C_{ij}는 i번째 리간드 원자와 j번째 단백질 원자 간의 상호작용 강도 또는 가중치(상수)를 의미한다. u_{ij}는 i번째 리간드 원자와 j번째 단백질 원자의 반 데르 발스 반경을 의미한다. H는 단백질-리간드 상호작용 매트릭스를 의미한다.

【0186】각 원자의 반 데르 발스 반경은 X-score 파라미터에서 얻을 수 있다. 2는 Lennard-Jones 잠재 함수의 식에서 인력과 반발력 사이의 균형을 맞추기 위한 상수를 나타내며, 12와 6은 각각 원자 간의 강한 반발력과 약한 인력을 나타내는 항을 의미한다.

【0187】따라서, E^{VDW}는 SE(3)-불변 변환 모듈로부터 파라미터화된 은닉 표현을 이용하여 계산된 단백질-리간드 상호작용 매트릭스 H와 단백질-리간드 복합체의 물리화학적 정보를 사용하여 계산될 수 있다.

【0188】상기 물리 지식 기반 네트워크 모델은 아래 수학식 24 와 같은 오차 최소화 및 아래 수학식 25와 같은 물리 법칙을 모두 만족하는 방향으로 손실함수를 계산하여 손실함수를 최소화하는 방향으로 최적화될 수 있다.

【0189】【수학식 24】

$$L_d = \sum_{N}(y-\hat{y})^2, where, \hat{y} = \sigma \cdot E^{VDW}$$

【0190】【수학식 25】

$$L_P = \sum_{N}\sum_{i,j}\left[\frac{\partial c_{ij}\left(\frac{r_i + M_{ij}}{||x_i - x_j||}\right)^{12} - 2\left(\frac{M_{ij}}{||x_i - x_j||}\right)^{6}}{\partial ||x_i - x_j||}\right]^2$$

【0191】여기서, y는 실제 실험 데이터를 기반으로 한 결합 친화도를 의미하며, $\hat{y}$는 예측된 결합 친화도를 의미한다. σ는 학습 가능한 파라미터를 의미한다. r_i는 리간드 원자의 반지름을 의미한다. M은 단백질-리간드 상호작용 매트릭스를 의미한다.

【0192】상기 수학식 25는 단백질-리간드 간 예측된 결합 친화도가 물리 법칙을 만족하도록 하는 수학식을 적용하는 것으로, 리간드와 단백질 원자 사이의 거리에 대한 결합 자유 에너지의 미분값이 0이 되도록 보장한다. 이는 실험적으로 규명된 구조로 모델링된 복합체가 해당 반응 좌표에서 결합 자유 에너지가 최소화된 지점에 위치해 있다는 사전 지식을 반영한 것이다. 따라서, 상기 물리 지식 기반 네트워크 모델의 전체 손실함수(L)은 L_d와 L_p의 합으로 정의될 수 있다.

【0193】즉, 상기 물리 지식 기반 네트워크 모델은 L_d 및 L_p의 합인 손실함수가 최소가 되도록 하여 반복학습될 수 있다. 이를 통해 본 발명은 결합 에너지를 최소화시키는 물리화학적 귀납적 편향을 기초로 하여 단백질-리간드 간 결합 친화도를 예측할 수 있다.

【0194】따라서, 본 발명은 단백질과 리간드 간의 결합 자유 에너지가 최소화되는 지점에서 결합이 형성된다는 물리화학적 원칙을 반영하여 보다 현실적으로 단백질-리간드 간 결합 에너지를 예측할 수 있다. 또한, SE(3)-불변성과 같은 기하학적 귀납적 편향을 통합하여 효율적으로 결합 에너지 예측을 수행할 수 있다. 결국, 본 발명은 결합 친화도(Binding Affinity, BA)를 보다 정확하게 제공할 수 있다.

【0195】또한, 본 발명은 여러 귀납적 편향을 도입하여 제한된 데이터로부터 우수한 일반화 성능을 달성할 수 있다.

【0196】또한, 본 발명은 결합 친화도가 3차원 공간에서의 회전 및 평행 이동에 관계없이 일정하게 유지된다는 기하학적 귀납적 편향과 결합이 최소 결합 자유 에너지에서 발생한다는 물리화학적 귀납적 편향을 통합하여 네트워크 모델에 반영하고 있다.

【0197】또한, 2 개의 벤치마크 세트를 통해 엄격한 검증을 수행한 결과, 본 발명의 단백질-리간드 간 결합 친화도 예측 성능의 우수성, 실제 약물 개발 과정에서의 가상 스크리닝에서의 실용성 및 해석 가능성을 시각화하여 예측값의 신뢰성을

확인할 수 있다.

【0199】도 8은 본 발명의 실시예에 따른 단백질-리간드 결합 친화도 예측 시스템 및 그 방법을 구현하는 컴퓨팅 장치를 도시한다.

【0200】도 1 내지 도7에 의해 설명된 본 발명의 실시예는 적어도 하나의 프로세서에 의해 동작하는 컴퓨팅 장치(200)로 구현될 수 있다.

【0201】컴퓨팅 장치(200)는 프로세서(210), 메모리(220), 스토리지(230), 통신 인터페이스(240), 시스템 인터커넥트(250) 및 디스플레이(260)를 포함할 수 있다.

【0202】프로세서(210)은, CPU(Central Processing Unit), MPU(Micro Processor Unit), MCU(Micro Controller Unit), GPU(Graphic Processing Unit) 및 APU(Application Processing Unit)을 포함한다.

【0203】메모리(220)는 프로세서(210)와 상호작용하여 프로그램이 효율적으로 실행될 수 있도록 데이터를 저장하고 필요한 정보에 빠르게 접근할 수 있도록 하는 기능을 수행한다. 메모리(220)는 레지스터, 캐시 메모리, 주 메모리, 읽기 전용 메모리, 가상 메모리, 비휘발성 메모리 중 적어도 하나를 포함한다.

【0204】스토리지(230)는 데이터를 영구적으로 저장하고 관리하는 역할을 한다. 스토리지는 컴퓨팅 시스템이 꺼지거나 재부팅된 후에도 데이터를 보존하며, 운영 체제, 애플리케이션, 사용자 파일 등을 저장하는 데 사용된다.

스토리지(230)은, 하드 디스크 드라이브(HDD), 솔리드 스테이트 드라이브(SSD), 광학 디스크, 네트워크 스토리지 및 클라우드 스토리지 중 적어도 하나를 포함한다.

【0205】 통신 인터페이스(240)는 컴퓨팅 시스템 내부 및 외부의 다양한 장치들 간에 데이터를 주고받기 위한 경로를 제공한다. 통신 인터페이스(240)는 USB(Universal Serial Bus), PCIe(Peripheral Component Interconnect Express), SATA(Serial ATA), Ethernet, Wi-Fi, Thunderbolt 및 HDMI(High-Definition Multimedia Interface) 중 적어도 하나의 통신 방식을 지원할 수 있다.

【0206】 시스템 인터커넥트(250)는 컴퓨팅 시스템 내부에서 다양한 구성 요소들 간의 데이터와 신호를 주고받는 역할을 한다. 시스템 인터커넥트(250)는, 버스(Bus), 포인트-투-포인트(Point-to-Point), 크로스바 스위치(Crossbar Switch), 네트워크-온-칩(Network-on-Chip, NoC) 중 적어도 하나의 방식을 지원할 수 있다.

【0207】 디스플레이(260)는 컴퓨팅 시스템의 출력 장치로서, 사용자에게 시각적인 정보를 제공하는 기능을 수행한다.

【0208】 전술한 구성에 의하여, 본 발명의 실시예에 따른 프로그램은, 프로세서(210)에 의해 실행되는 명령어들에 기초하여 실행되며, 메모리(220) 또는 스토리지(230)에 저장될 수 있다.

【0210】 전술한 본 발명의 실시예에 따른 방법은 다양한 컴퓨터 구성요소를 통하여 실행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능한 기록매

체에 기록될 수 있다. 컴퓨터 판독 가능한 기록매체는 프로그램 명령어, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 컴퓨터 판독 가능한 기록매체에 기록되는 프로그램 명령은 본 발명의 실시예를 위하여 특별히 설계되고 구성된 것이거나, 컴퓨터 소프트웨어 분야의 통상의 기술자에게 공지되어 사용가능한 것일 수 있다. 컴퓨터 판독 가능한 기록매체는 하드디스크, 플로피디스크, 자기테이프 등의 자기기록 매체, CD-ROM, DVD 등의 광기록 매체, 플롭티컬디스크 등의 자기-광 매체, ROM, RAM, 플래시 메모리 등과 같이, 프로그램 명령을 저장하고 수행하도록 구성된 하드웨어를 포함한다. 프로그램 명령은, 컴파일러에 의해 만들어지는 기계어 코드, 인터프리터를 사용하여 컴퓨터에서 실행될 수 있는 고급언어 코드를 포함한다. 하드웨어는 본 발명에 따른 방법을 처리하기 위하여 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있고, 그 역도 마찬가지이다.

【0211】 본 발명의 실시예에 따른 방법은 프로그램 명령 형태로 전자장치에서 실행될 수 있다. 전자장치는 스마트폰이나 스마트패드 등의 휴대용 통신 장치, 컴퓨터 장치, 휴대용 멀티미디어 장치, 휴대용 의료 기기, 카메라, 웨어러블 장치, 가전 장치를 포함한다.

【0212】 본 발명의 실시예에 따른 방법은 컴퓨터 프로그램 제품에 포함되어 제공될 수 있다. 컴퓨터 프로그램 제품은 상품으로서 판매자 및 구매자 간에 거래될 수 있다. 컴퓨터 프로그램 제품은 기기로 읽을 수 있는 기록매체의 형태로, 또는 어플리케이션 스토어를 통해 온라인으로 배포될 수 있다. 온라인 배포의

경우에, 컴퓨터 프로그램 제품의 적어도 일부는 제조사의 서버, 어플리케이션 스토어의 서버, 또는 중계 서버의 메모리와 같은 저장 매체에 적어도 일시 저장되거나, 임시적으로 생성될 수 있다.

【0213】 본 발명의 실시예에 따른 구성요소, 예컨대 모듈 또는 프로그램 각각은 단수 또는 복수의 서브 구성요소로 구성될 수 있으며, 이러한 서브 구성요소들 중 일부 서브 구성요소가 생략되거나, 또는 다른 서브 구성요소가 더 포함될 수 있다. 일부 구성요소들(모듈 또는 프로그램)은 하나의 개체로 통합되어, 통합되기 이전의 각각의 해당 구성요소에 의해 수행되는 기능을 동일 또는 유사하게 수행할 수 있다. 본 발명의 실시예에 따른 모듈, 프로그램 또는 다른 구성요소에 의해 수행되는 동작들은 순차적, 병렬적, 반복적 또는 휴리스틱하게 실행되거나, 적어도 일부 동작이 다른 순서로 실행되거나, 생략되거나, 또는 다른 동작이 추가될 수 있다.

【0214】 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.

【0215】본 발명의 범위는 후술하는 청구범위에 의하여 나타내어지며, 청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

【부호의 설명】

【0217】100: 단백질-리간드 결합 친화도 예측 시스템

110: 단백질-리간드 복합체 원자 정보 및 3차원 좌표 수신부

120: 그래프 전처리 모듈

130: SE(3)-불변성 변환 모듈

140: 물리 지식 기반 네트워크 모델을 이용한 결합 친화도 예측부

【청구범위】

【청구항 1】

단백질-리간드 간 결합 친화도 예측 시스템에 있어서,

단백질-리간드 복합체의 원자 정보 및 3차원 좌표 데이터를 수신하는 수신부;

상기 원자 정보 및 상기 3차원 좌표 데이터를 기초로 하여 전처리 작업을 수행하여 그래프 구조를 생성하는 그래프 전처리 모듈;

3차원 공간에서 물질 또는 시스템이 회전 또는 평행 이동에 대해 상기 결합 친화도가 불변하도록 제1 귀납적 편향을 기초로 하여 상기 그래프 구조를 임베딩 레이어 및 은닉 표현으로 불변성 변환하는 SE(3)-불변성 변환 모듈;

상기 임베딩 레이어 및 상기 은닉 표현을 입력으로 하여 상기 결합 친화도를 예측하는 물리 지식 기반 네트워크 모델; 및

상기 예측된 결합 친화도를 출력하는 출력부; 를 포함하는

단백질-리간드 간 결합 친화도 예측 시스템.

【청구항 2】

제1항에 있어서,

상기 물리 지식 기반 네트워크 모델은,

(a) 상기 은닉 표현을 이용하여 단백질-리간드 상호작용을 계산하는 단계;

(b) 상기 단백질-리간드 상호작용을 이용하여 상기 결합 친화도를 예측하는 단계; 및

(c) 단백질-리간드 복합체의 결합 자유 에너지가 최소화되는 지점에서 결합 상태가 형성된다는 제2 귀납적 편향을 기초로 하여 손실 함수를 계산하는 단계;

를 포함하여 상기 손실 함수가 최소가 되도록 학습되는 것인

단백질-리간드 간 결합 친화도 예측 시스템.

【청구항 3】

제2항에 있어서

상기 그래프 전처리 모듈은,

상기 원자 정보에 기반하여 각 원자를 노드로 정의하여 각 노드에 원자의 물리화학적 특징을 포함시키고,

상기 3차원 좌표를 포함한 각 원자의 위치 행렬을 생성하고,

KNN(K-Nearest Neighbors) 알고리즘을 사용하여 원자 간의 거리를 기준으로 엣지를 정의하여 이웃 원자와의 상호작용을 엣지로 연결하여 나타내는 것인

단백질-리간드 간 결합 친화도 예측 시스템.

【청구항 4】

제3항에 있어서

상기 SE(3)-불변성 변환 모듈은

아래 수학식 1, 2 및 3을 통해 상기 물리화학적 특징을 인코딩하여 임베딩 레이어 및 은닉 표현으로 변환하고,

아래 수학식 4, 5, 6, 7 및 8을 통해 상기 은닉 표현을 업데이트하는 것인

단백질-리간드 간 결합 친화도 예측 시스템.

[수학식 1]

$$h^0=(h_P^0 \| h_L^0)$$

[수학식 2]

$$h_P^0=Linear(V_P)\in R^{N\times D_E}$$

[수학식 3]

$$h_L^0=Linear(V_L)\in R^{N\times D_E}$$

여기서 h^0는 h_P^0와 h_L^0 결합(concatenating)이고, h^0는 초기 은닉 표현을 의미한다. Linear()는 선형 변환 함수를 의미한다. R은 실수 공간을 나타내며, 해당 벡터나 행렬이 실수로 이루어져 있다는 것을 의미한다. N과 M은 각각 단백질과 리간드의 원자 개수를 의미한다. D_E는 각 원자가 가지는 물리화학적 특징의 차원수를 의미한다.

[수학식 4]

$$h_i^{(l+1)} = h_i^l + \sum_{j \in \nu, i \neq j} f_h(||x_i - x_j||, h_i^l, h_j^l, e_{ij}; \theta_h)$$

여기서, h_i^{l+1}은 i번째 노드의 l+1번째 임베딩 레이어에서의 은닉 표현을 의미한다. h_i^l은 i번째 노드의 l번째 임베딩 레이어에서의 은닉 표현을 의미한다. $\|x_i - x_j\|^2$은 원자 i와 j 사이의 유클리드 거리를 의미한다. e_{ij}는 KNN(K-Nearest Neighbors) 그래프 알고리즘을 사용하여 원자 i와 j 사이의 거리를 기준으로 정의된 엣지를 의미한다. θ_h는 네트워크 모델의 학습 가능한 파라미터를 의미한다.

[수학식 5]

$$f_h = Attention(q_i, k_j, v_j) \cdot Linear(r_{ij})$$

[수학식 6]

$$q_i = Linear(h_i^0)$$

[수학식 7]

$$k_j = Linear([r_{ij} \| e_{ij} \| h_i \| h_j])$$

[수학식 8]

$$v_j = Linear([r_{ij} \| e_{ij} \| h_i \| h_j])$$

여기서, q_i, k_i, v_i는 어텐션(Attention) 연산을 위한 쿼리(query), 키(key), 밸류(value) 행렬을 의미한다.

【청구항 5】

제4항에 있어서,

상기 (a) 단계는,

상기 은닉 표현을 이용하여 단백질-리간드 상호작용 매트릭스를 아래 수학식 9를 통해 계산하는 것인

단백질-리간드 간 결합 친화도 예측 시스템.

[수학식 9]

$$H=(h_M^L \cdot h_P^L)^T$$

여기서, h_M^L은 리간드를 구성하는 원자의 최종 은닉 표현을 의미하며, h_P^L은 단백질을 구성하는 원자의 최종 은닉 표현을 의미한다.

【청구항 6】

제5항에 있어서

상기 (b) 단계는,

아래 수학식 10을 통해 계산되는 원자쌍별 반 데르 발스 상호작용 에너지의

합을 기초로 상기 결합 친화도를 예측하는 것인

단백질-리간드 간 결합 친화도 예측 시스템.

[수학식 10]

$$E^{VDW}=\sum_{i,j}C_{ij}[(\frac{u_{ij}+H_{ij}}{||x_i-x_j||})^{12}-2(\frac{u_{ij}+H_{ij}}{||x_i-x_j||})^{6}]$$

여기서, E^{VDW}는 반 데르 발스 상호작용 에너지의 합을 의미하고, C_{ij}는 i번째 리간드 원자와 j번째 단백질 원자 간의 상호작용 강도 또는 가중치(상수)를 의미한다. u_{ij}는 i번째 리간드 원자와 j번째 단백질 원자의 반 데르 발스 반경을 의미한다.

【청구항 7】

제6항에 있어서,

상기 (c) 단계는,

아래 수학식 12를 통해 계산된 L_d 및 아래 수학식 13을 통해 계산된 L_p의 합으로 된 손실 함수를 계산하는 것인,

단백질-리간드 간 결합 친화도 예측 시스템.

[수학식 12]

$$L_d = \sum_{N}(y-\hat{y})^2, where, \hat{y} = \sigma \cdot E^{VDW}$$

[수학식 13]

$$L_P = \sum_{N}\sum_{i,j}\left[\frac{\partial c_{ij}\left(\frac{r_i + M_{ij}}{||x_i - x_j||}\right)^{12} - 2\left(\frac{M_{ij}}{||x_i - x_j||}\right)^{6}}{\partial ||x_i - x_j||}\right]^2$$

여기서, y는 실제 실험 데이터를 기반으로 한 결합 친화도를 의미하며, $\hat{y}$는 예측된 결합 친화도를 의미한다. σ는 학습 가능한 파라미터를 의미한다. r_i는 리간드 원자의 반지름을 의미한다. M은 단백질-리간드 상호작용 매트릭스를 의미한다.

【청구항 8】

단백질-리간드 간 결합 친화도 예측 방법에 있어서,

(A) 단백질-리간드 복합체의 원자 정보 및 3차원 좌표 데이터를 수신하는 단계;

(B) 상기 원자 정보 및 상기 3차원 좌표 데이터를 기초로 하여 전처리 작업을 수행하여 그래프 구조를 생성하는 그래프 전처리 단계;

(C) 3차원 공간에서 물질 또는 시스템이 회전 또는 평행 이동에 대해 상기 결합 친화도가 불변하도록 제1 귀납적 편향을 기초로 하여 상기 그래프 구조를 임베딩 레이어 및 은닉 표현으로 불변성 변환하는 SE(3)-불변성 변환 단계; 및

(D) 상기 임베딩 레이어 및 상기 은닉 표현을 입력으로 하여 물리 지식 기반 네트워크 모델을 이용하여 상기 결합 친화도를 예측하는 단계; 를 포함하는

단백질-리간드 간 결합 친화도 예측 방법.

【청구항 9】

제8항에 있어서,

상기 물리 지식 기반 네트워크 모델은,

(E) 상기 은닉 표현을 이용하여 단백질-리간드 상호작용을 계산하는 단계;

(F) 상기 단백질-리간드 상호작용을 이용하여 상기 결합 친화도를 예측하는 단계; 및

(G) 단백질-리간드 복합체의 결합 자유 에너지가 최소화되는 지점에서 결합 상태가 형성된다는 제2 귀납적 편향을 기초로 하여 손실 함수를 계산하는 단계;

를 포함하여 상기 손실 함수가 최소가 되도록 학습되는 것인

단백질-리간드 간 결합 친화도 예측 방법.

【청구항 10】

제9항에 있어서,

상기 (B) 단계는,

상기 원자 정보에 기반하여 각 원자를 노드로 정의하여 각 노드에 원자의 물

리화학적 특징을 포함시키고,

상기 3차원 좌표를 포함한 각 원자의 위치 행렬을 생성하고,

KNN(K-Nearest Neighbors) 알고리즘을 사용하여 원자 간의 거리를 기준으로 엣지를 정의하여 이웃 원자와의 상호작용을 엣지로 연결하여 나타내는 것인

단백질-리간드 간 결합 친화도 예측 방법.

【청구항 11】

제10항에 있어서,

상기 (C) 단계는

아래 수학식 14, 15 및 16을 통해 상기 물리화학적 특징을 인코딩하여 임베딩 레이어 및 은닉 표현으로 변환하고,

아래 수학식 17, 18, 19, 20 및 21을 통해 상기 은닉 표현을 업데이트하는 것인

단백질-리간드 간 결합 친화도 예측 방법.

[수학식 14]

$$h^0=(h_P^0 \parallel h_L^0)$$

[수학식 15]

$$h_P^0=Linear(V_P)\in R^{N\times D_E}$$

[수학식 16]

$$h_L^0 = Linear(V_L) \in R^{N \times D_E}$$

여기서 h^0는 h_p^0와 h_r^0 결합(concatenating)이고, h^0는 초기 은닉 표현을 의미한다. Linear()는 선형 변환 함수를 의미한다. R은 실수 공간을 나타내며, 해당 벡터나 행렬이 실수로 이루어져 있다는 것을 의미한다. N과 M은 각각 단백질과 리간드의 원자 개수를 의미한다. D_E는 각 원자가 가지는 물리화학적 특징의 차원수를 의미한다.

[수학식 17]

$$h_i^{(l+1)} = h_i^l + \sum_{j \in \nu, i \neq j} f_h(||x_i - x_j||, h_i^l, h_j^l, e_{ij}; \theta_h)$$

여기서, h_i^{l+1}은 i번째 노드의 l+1번째 임베딩 레이어에서의 은닉 표현을 의미한다. h_i^l은 i번째 노드의 l번째 임베딩 레이어에서의 은닉 표현을 의미한다. $\|x_i - x_j\|^2$은 원자 i와 j 사이의 유클리드 거리를 의미한다. e_{ij}는 KNN(K-Nearest Neighbors) 그래프 알고리즘을 사용하여 원자 i와 j 사이의 거리를 기준으로 정의된 엣지를 의미한다. θ_h는 네트워크 모델의 학습 가능한 파라미터를 의미한다.

[수학식 18]

$$f_h = Attention(q_i, k_j, v_j) \cdot Linear(r_{ij})$$

[수학식 19]

$$q_i = Linear(h_i^0)$$

[수학식 20]

$$k_j = Linear([r_{ij} \| e_{ij} \| h_i \| h_j])$$

[수학식 21]

$$v_j = Linear([r_{ij} \| e_{ij} \| h_i \| h_j])$$

여기서, q_i, k_j, v_j는 어텐션(Attention) 연산을 위한 쿼리(query), 키(key), 밸류(value) 행렬을 의미한다.

【청구항 12】

제11항에 있어서,

상기 (E) 단계는,

상기 은닉 표현을 이용하여 단백질-리간드 상호작용 매트릭스를 아래 수학식 22를 통해 계산하는 것인

단백질-리간드 간 결합 친화도 예측 방법.

[수학식 22]

$$H=(h_M^L \cdot h_P^L)^T$$

여기서, h_M^L은 리간드를 구성하는 원자의 최종 은닉 표현을 의미하며, h_P^L은 단백질을 구성하는 원자의 최종 은닉 표현을 의미한다.

【청구항 13】

제12항에 있어서,

상기 (F) 단계는,

아래 수학식 23을 통해 계산되는 원자쌍별 반 데르 발스 상호작용 에너지의 합을 기초로 상기 결합 친화도를 예측하는 것인

단백질-리간드 간 결합 친화도 예측 방법.

[수학식 23]

$$E^{VDW}=\sum_{i,j} C_{ij}\left[\left(\frac{u_{ij}+H_{ij}}{||x_i-x_j||}\right)^{12}-2\left(\frac{u_{ij}+H_{ij}}{||x_i-x_j||}\right)^{6}\right]$$

여기서, E^{VDW}는 반 데르 발스 상호작용 에너지의 합을 의미하고, C_{ij}는 i번째 리간드 원자와 j번째 단백질 원자 간의 상호작용 강도 또는 가중치(상수)를 의미한다. u_{ij}는 i번째 리간드 원자와 j번째 단백질 원자의 반 데르 발스 반경을 의미

한다.

【청구항 14】

제13항에 있어서,

상기 (G) 단계는,

아래 수학식 24를 통해 계산된 L_d 및 아래 수학식 25를 통해 계산된 L_p의 합의 합으로 된 손실 함수를 계산하는 것인,

단백질-리간드 간 결합 친화도 예측 방법.

[수학식 24]

$$L_d = \sum_N (y - \hat{y})^2, where, \hat{y} = \sigma \cdot E^{VDW}$$

[수학식 25]

$$L_P = \sum_N \sum_{i,j} [\frac{\partial c_{ij} (\frac{r_i + M_{ij}}{||x_i - x_j||})^{12} - 2 (\frac{M_{ij}}{||x_i - x_j||})^6}{\partial ||x_i - x_j||}]^2$$

여기서, y는 실제 실험 데이터를 기반으로 한 결합 친화도를 의미하며, $\hat{y}$는 예측된 결합 친화도를 의미한다. σ는 학습 가능한 파라미터를 의미한다. r_i는 리간드 원자의 반지름을 의미한다. M은 단백질-리간드 상호작용 매트릭스를 의미한다.

【청구항 15】

컴퓨터가 판독 가능한 기록 매체에 있어서,

제8항 내지 제14항 중 어느 하나의 항에 따른 단백질-리간드 간 결합 친화도 예측 방법을 실행하는 프로그램이 기록된 컴퓨터가 판독 가능한 기록 매체.

【요약서】

【요약】

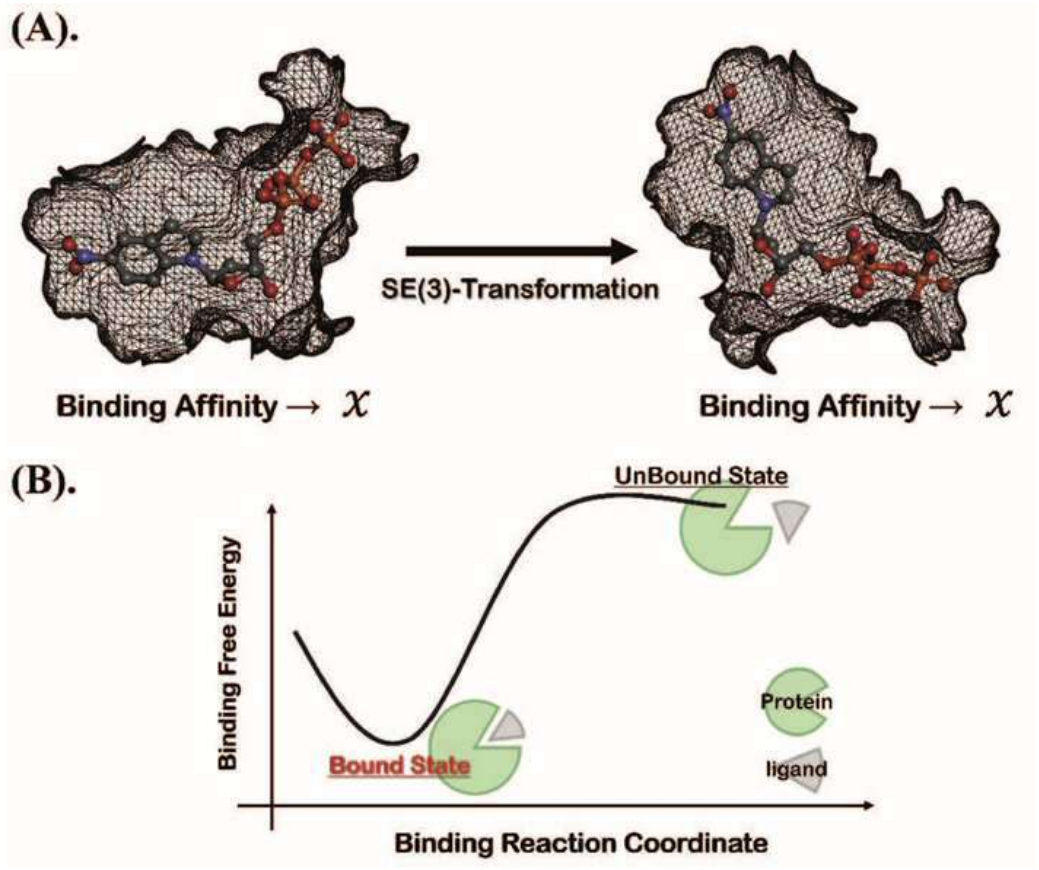
본 발명의 일실시예는 단백질-리간드 간 결합 친화도 예측 시스템에 있어서, 단백질-리간드 복합체의 원자 정보 및 3차원 좌표 데이터를 수신하는 수신부; 상기 원자 정보 및 상기 3차원 좌표 데이터를 기초로 하여 전처리 작업을 수행하여 그래프 구조를 생성하는 그래프 전처리 모듈; 3차원 공간에서 물질 또는 시스템이 회전 또는 평행 이동에 대해 상기 결합 친화도가 불변하도록 제1 귀납적 편향을 기초로 하여 상기 그래프 구조를 임베딩 레이어 및 은닉 표현으로 불변성 변환하는 SE(3)-불변성 변환 모듈; 상기 임베딩 레이어 및 상기 은닉 표현을 입력으로 하여 상기 결합 친화도를 예측하는 물리 지식 기반 네트워크 모델; 및 상기 예측된 결합 친화도를 출력하는 출력부; 를 포함하는 단백질-리간드 간 결합 친화도 예측 시스템을 제공한다.

【대표도】

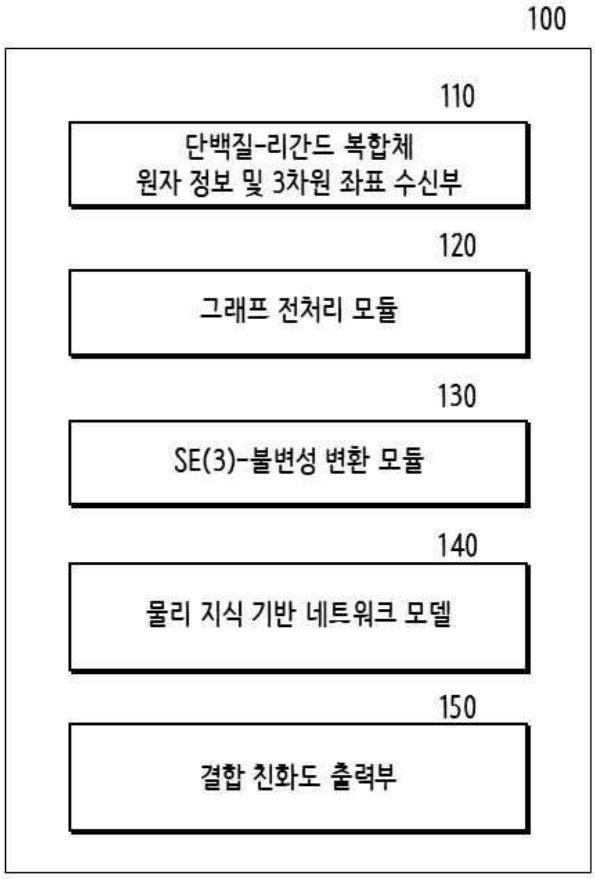
도 3

【도면】

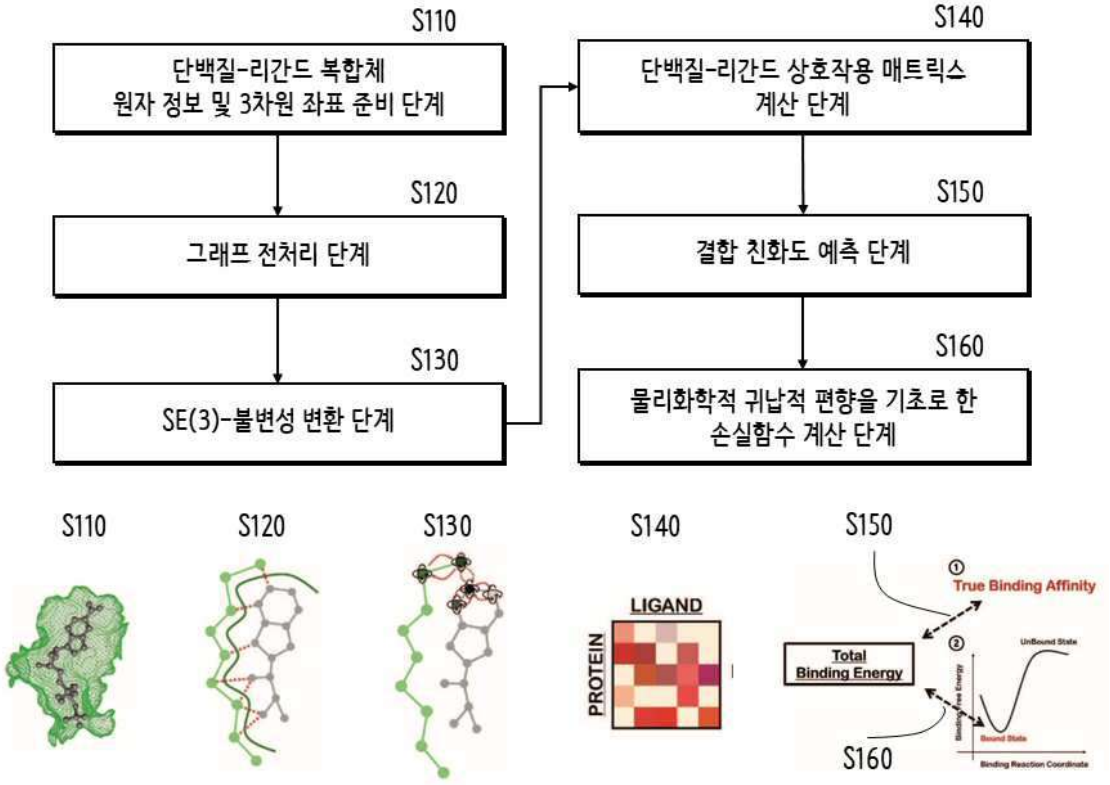
【도 1】


(A).
SE(3)-Transformation
Binding Affinity → X
Binding Affinity → X
(B).
UnBound State
Binding Free Energy
Protein
ligand
Bound State
Binding Reaction Coordinate

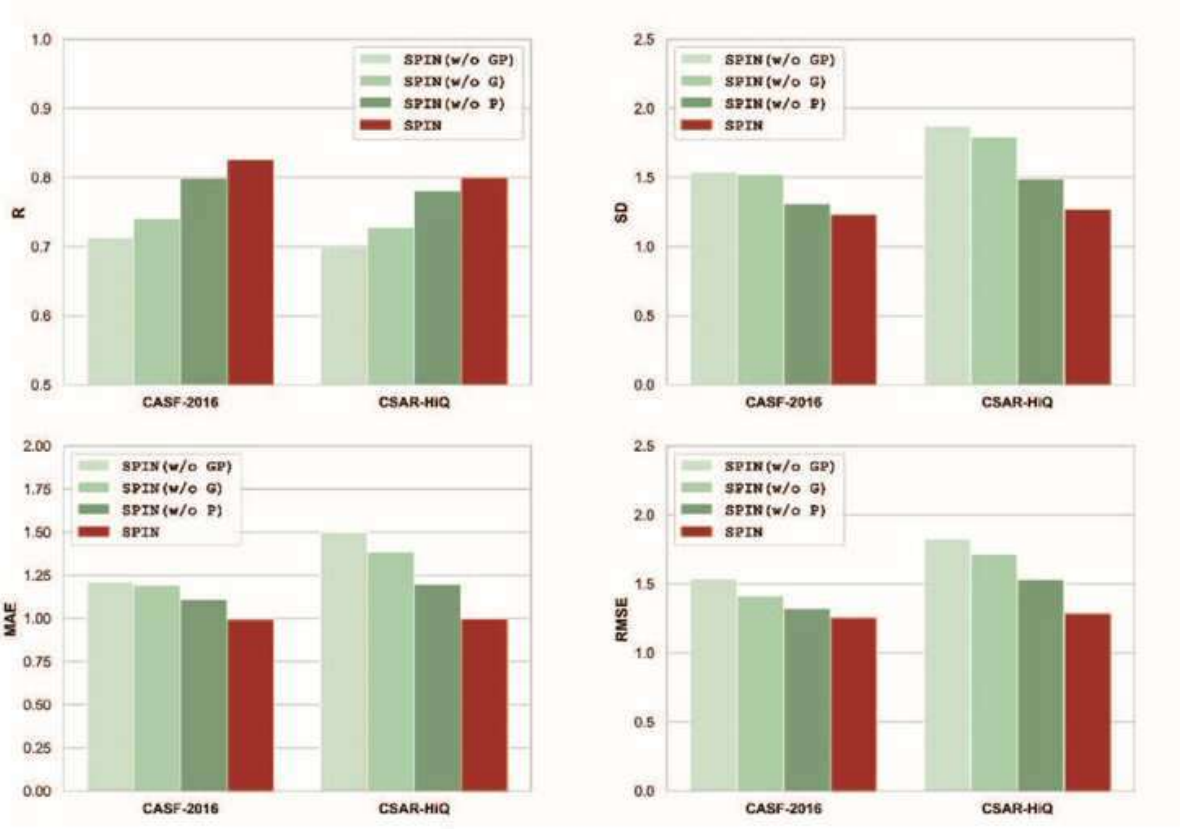
【도 2】


100
110
단백질-리간드 복합체
원자 정보 및 3차원 좌표 수신부
120
그래프 전처리 모듈
130
SE(3)-불변성 변환 모듈
140
물리 지식 기반 네트워크 모델
150
결합 친화도 출력부

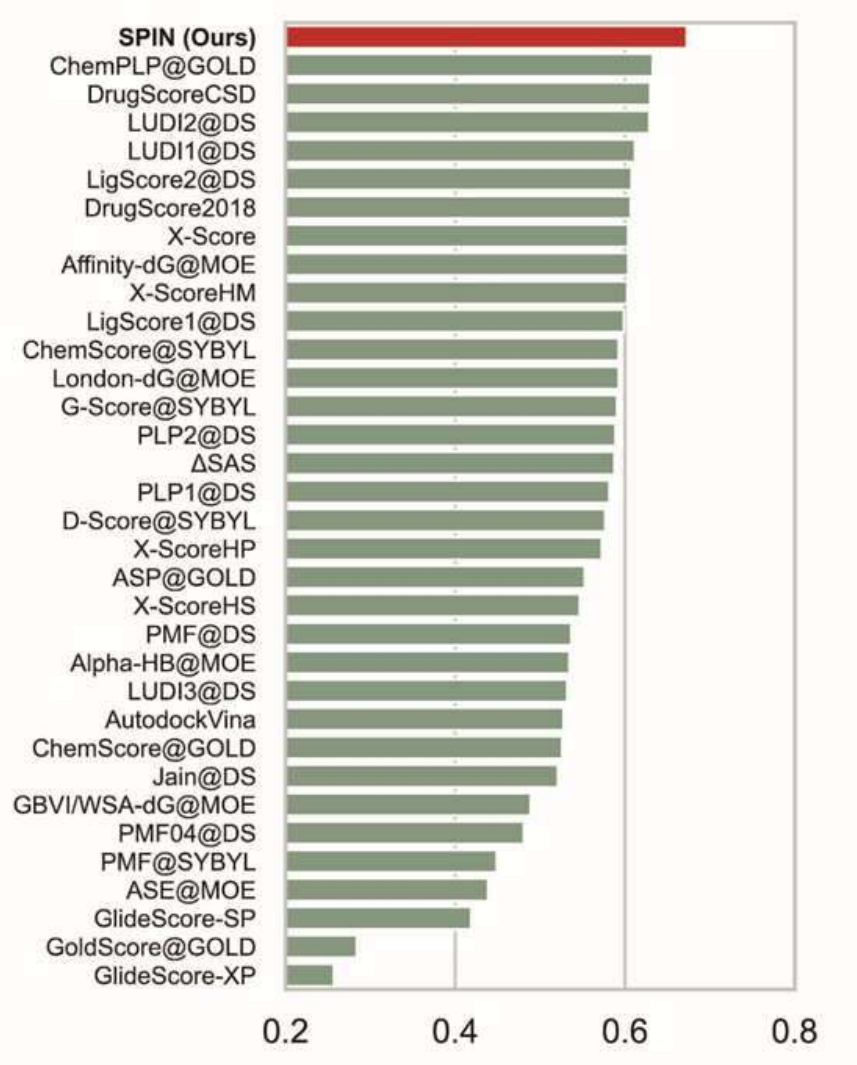
【도 3】


S110
단백질-리간드 복합체
원자 정보 및 3차원 좌표 준비 단계
S120
그래프 전처리 단계
S130
SE(3)-불변성 변환 단계
S140
단백질-리간드 상호작용 매트릭스
계산 단계
S150
결합 친화도 예측 단계
S160
물리화학적 귀납적 편향을 기초로 한
손실함수 계산 단계
S110
S120
S130
S140
LIGAND
PROTEIN
S150
True Binding Affinity
Total
Binding Energy
S160

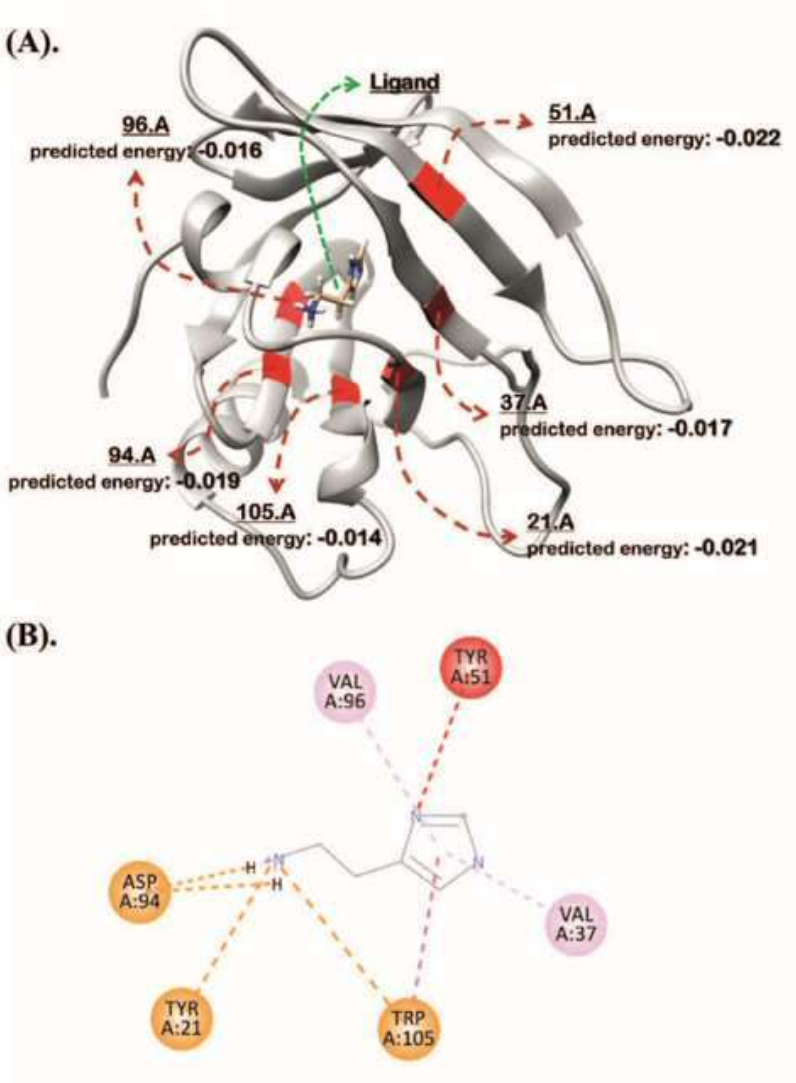
【도 4】


R
SD
MAE
RMSE
SPIN(w/o GP)
SPIN(w/o G)
SPIN(w/o P)
SPIN
CASF-2016
CSAR-HiQ

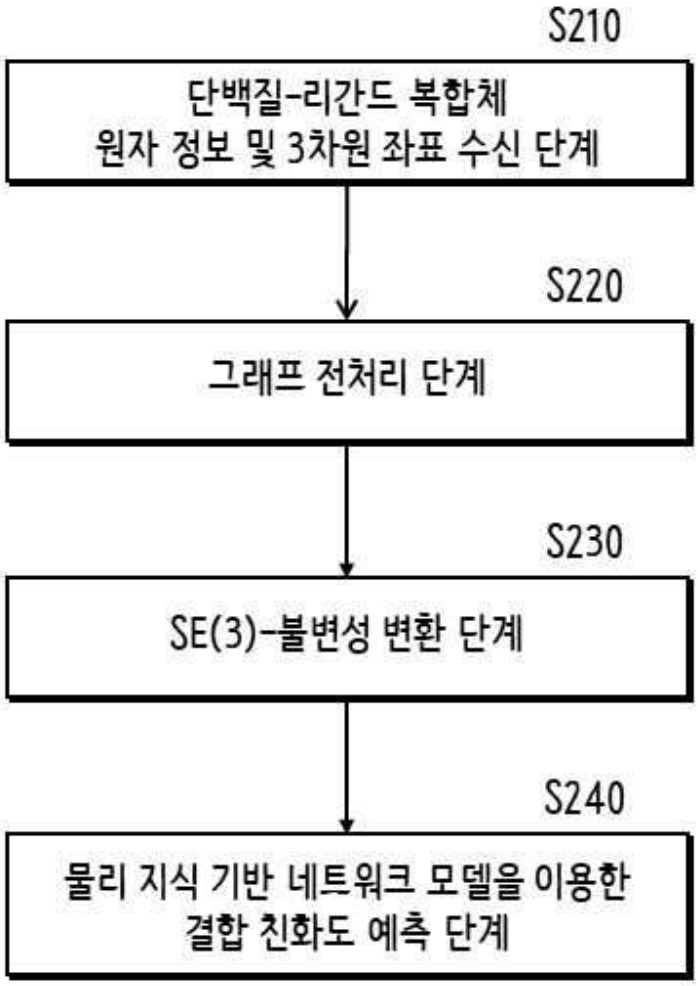
【도 5】


SPIN (Ours)
ChemPLP@GOLD
DrugScoreCSD
LUDI2@DS
LUDI1@DS
LigScore2@DS
DrugScore2018
X-Score
Affinity-dG@MOE
X-ScoreHM
LigScore1@DS
ChemScore@SYBYL
London-dG@MOE
G-Score@SYBYL
PLP2@DS
ΔSAS
PLP1@DS
D-Score@SYBYL
X-ScoreHP
ASP@GOLD
X-ScoreHS
PMF@DS
Alpha-HB@MOE
LUDI3@DS
AutodockVina
ChemScore@GOLD
Jain@DS
GBVI/WSA-dG@MOE
PMF04@DS
PMF@SYBYL
ASE@MOE
GlideScore-SP
GoldScore@GOLD
GlideScore-XP
0.2
0.4
0.6
0.8

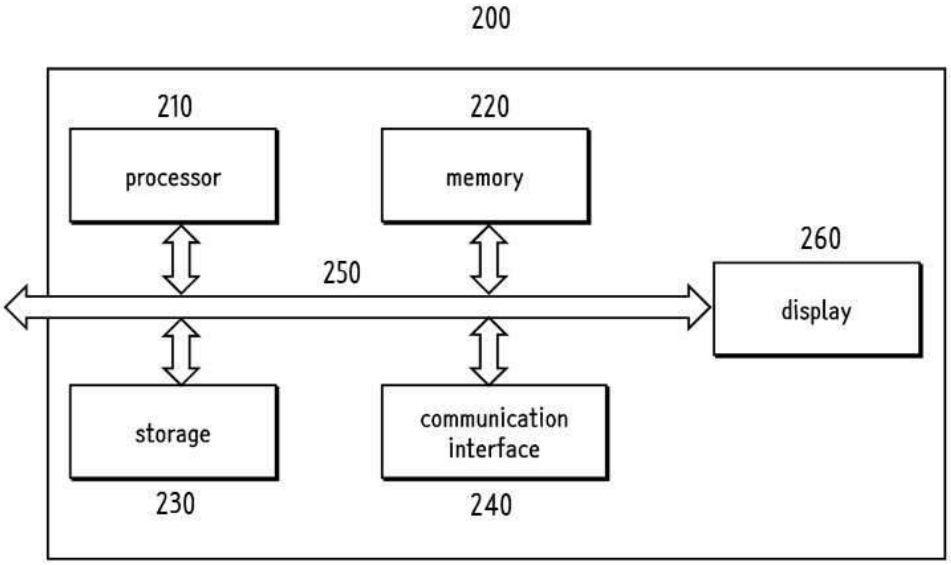
【도 6】


(A).
Ligand
96.A
predicted energy: -0.016
51.A
predicted energy: -0.022
37.A
predicted energy: -0.017
94.A
predicted energy: -0.019
105.A
predicted energy: -0.014
21.A
predicted energy: -0.021
(B).
VAL A:96
TYR A:51
ASP A:94
VAL A:37
TYR A:21
TRP A:105

【도 7】


S210
단백질-리간드 복합체
원자 정보 및 3차원 좌표 수신 단계
S220
그래프 전처리 단계
S230
SE(3)-불변성 변환 단계
S240
물리 지식 기반 네트워크 모델을 이용한
결합 친화도 예측 단계

【도 8】


200
210
processor
220
memory
260
250
display
storage
communication interface
230
240