ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 52, No. 5, pp. 415-423, 2025. 5
https://doi.org/10.5626/JOK.2025.52.5.415

단백질 모달리티 정렬을 활용한 단백질-리간드 결합 친화도 예측

(Protein-Ligand Binding Affinity Prediction Using Protein Modality Alignment)

이 승 용 †　　박 상 현 ††
(Seungyong Lee)　(Sanghyun Park)

요 약 신약 후보물질 도출을 위해 타겟 단백질과 높은 결합 친화도를 지닌 분자를 식별하는 것은 많은 자원과 시간이 필요하다. 딥러닝을 활용한 단백질-리간드 결합 친화도 예측 모델은 기존 실험적 접근법의 자원 및 시간 소요 문제를 효율적으로 해결할 수 있는 대안으로 주목받고 있다. 기존 방법론은 리간드의 2D 그래프 정보만을 사용하여 3차원 공간에서의 입체적 상호작용을 충분히 모델링하지 못하는 근본적 한계를 지닌다. 또한 단백질을 모델링할 때 서열 및 구조, 표면 정보를 사용하지만, 이들의 연결 관계를 모델에 주입하지 못하는 문제가 존재한다. 본 연구는 단백질의 다중 모달리티(서열, 구조, 표면) 정보의 서열 기준 정렬과 리간드 정보의 SE(3)-invariant 그래프 신경망을 통합한 새로운 딥러닝 프레임워크를 제안한다. 제안된 모델은 기존 베이스라인 모델들보다 뛰어난 성능을 보였으며, Ablation study를 통해 단백질의 모달리티 정렬과 3D 리간드 정보의 중요성을 입증하였다.

키워드: 단백질-리간드 결합 친화도, DTA, 단백질 모델링, SE(3)-invariant 그래프 신경망

Abstract Identifying molecules with high binding affinity to a target protein for drug candidate discovery requires significant resources and time. Deep learning-based protein-ligand binding affinity prediction research plays a crucial role in addressing this challenge. Existing studies have utilized protein sequence and structural information along with ligand 2D structures. However, they have limitations in fully capturing complex interactions. Additionally, while sequence, structure, and surface information are used for protein modeling, previous approaches have struggled to incorporate their dependent relationships into the model. In this paper, we proposed a model that could inject these dependencies by aligning protein sequence, structure, and surface information based on sequence data. Furthermore, our model leverages both 2D structure of the ligand and its 3D representation using an SE(3)-invariant graph neural network. The proposed model outperformed existing baseline models. An ablation study demonstrated the importance of aligning different protein modalities and incorporating both 2D and 3D ligand information.

Keywords: protein-ligand binding affinity, DTA, protein modeling, SE(3)-invariant graph neural networks

· 이 논문은 2025년도 정부(과기정통부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2023-00229822)
· 이 논문은 2025년도 국토교통부의 스마트시티 혁신인재육성사업으로 지원을 받아 수행된 연구임

† 학생회원 : 홍익대학교 컴퓨터공학과 학생
bluearth4587@gmail.com
†† 종신회원 : 연세대학교 컴퓨터과학과 교수(Yonsei Univ.)
sanghyun@yonsei.ac.kr
(Corresponding author임)

논문접수 : 2025년 2월 28일
(Received 28 February 2025)
논문수정 : 2025년 3월 25일
(Revised 25 March 2025)
심사완료 : 2025년 3월 27일
(Accepted 27 March 2025)

정보과학회논문지 제52권 제5호(2025. 5)

1. 서 론

신약이 될 수 있는 후보 물질을 도출하기 위해서는 일반적으로 대규모 화합물 라이브러리에서 타겟 단백질에 잘 결합할 수 있는 분자를 식별하는 과정을 따른다. 이 과정에서 결합 친화도를 계산이 필요하며, 결합 친화도는 단백질과 리간드 간 결합 강도를 나타내는 지표이다. 일반적으로 도킹, 분자동역학 시뮬레이션과 같은 복잡한 물리화학적 계산이나 고속 스크리닝 같은 실험을 통해 타겟 단백질과의 결합 친화도를 계산하고, 이를 통해 특정 분자가 타겟 단백질과 얼마나 안정적으로 결합할 수 있는지를 검증한다. 이 과정은 많은 자원과 시간을 요구하기 때문에 적합한 후보 물질을 도출하는 데 장애물이 되며, 효율적인 신약 개발에 걸림돌이 될 수 있다[1].

이 문제를 해결하기 위해 딥러닝을 활용한 타겟 단백질과 리간드(저분자 후보물질)의 결합 친화도 예측 연구가 활발히 수행되고 있다. 이러한 연구들은 크게 서열이나 그래프 같은 1, 2차원으로 표현된 결합 정보를 활용하는 방식과 단백질-리간드 3차원 결합 구조를 활용하여 예측하는 방식으로 분류된다[1]. 그러나 단백질과 다른 생체분자 사이의 결합에 있어서 서열이나 구조 정보는 단백질의 기능적, 화학적 역할을 고려할 수 없기 때문에 복합체 자체의 다양한 정보를 담기에 한계가 있다. 일반적으로 단백질 분자 표면이 단백질 구조의 화학적 및 기하학적 특징을 포함한 고수준 표현을 담고 있기 때문에 리간드와의 상호작용을 합리적으로 포착하기 위해서는 단백질 표면 정보가 필수적이다[2]. 그 예시로 최근에는 단백질의 표면 정보를 효율적으로 모델링하기 위해, MaSIF[2]와 dMaSIF[3]에서 각각 삼각 메쉬와 포인트 클라우드 형태로 단백질 표면을 표현한 여러 연구가 수행되었다. 또한 이러한 선행 연구에 힘입어 단백질-리간드의 결합 친화도 예측 연구에도 단백질 표면 정보를 활용하려는 시도가 있었다[4].

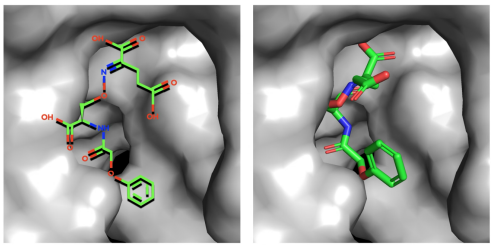

그림 1 3차원 단백질 구조에 대한 2D(왼쪽)와 3D(오른쪽) 리간드 구조 시각화 (PDB ID: 2JCH)

Fig. 1 Visualization of ligand structures in 2D and 3D for 3D protein structures

그러나 기존 접근법들은 리간드 표현에 있어 2차원 분자 구조에 의존함으로써, 3차원 공간에서의 입체적 상호작용 정보가 소실되는 근본적 한계를 지니고 있다. 단백질을 3차원으로 표현했을 때, 리간드를 3차원이 아닌 2차원으로 표현할 경우, 그림 1과 같이 각 요소의 데이터 형식의 차이로 인해 두 요소(단백질과 리간드)의 구조적 상호작용을 명시적으로 포착할 수 없는 문제가 발생한다. 그러므로 단백질을 3차원으로 표현할 때, 리간드를 3차원으로 표현하여 모델링해야 단백질-리간드의 관계를 합리적으로 추론할 수 있게 된다. 우리는 이 문제를 해결하기 위해 리간드의 분자 구조의 2D, 3D 형태를 모두 활용하여 결합 친화도를 예측할 수 있는 모델을 제안한다. 이 모델은 단백질의 서열, 구조, 표면 정보에 대해 리간드의 2D, 3D 형태의 데이터를 모델링하여 두 구조 간의 결합 친화도를 예측한다. 우리는 3차원 공간에서 특정 객체의 회전과 이동에 불변하도록 정보를 추출할 수 있는 SE(3)-invariant GNN을 활용하여 리간드의 3D 구조를 효율적으로 학습할 수 있도록 설계했다.

또한 기존 연구들은 1차원으로 표현되는 단백질의 서열 정보와, 3차원으로 표현되는 단백질의 구조 및 표면 정보를 모두 사용하였으나, 각 모달리티를 효율적으로 정렬하지 못하는 한계가 있었다. 단백질의 서열, 구조, 표면 정보는 독립적으로 존재하지 않고, 그림 2와 같이 연결 관계를 갖고 있다. 때문에 세 모달리티(서열, 구조, 표면)의 통합된 임베딩을 구하기 전에, 각 모달리티 간의 연결 관계를 모델에 주입하여 단백질의 세 모달리티를 정렬할 필요가 있다. 이를 통해 모델이 단백질의 다양한 표현법을 효과적으로 이해하도록 유도할 수 있다. 단백질의 세 모달리티를 정렬할 때, 단백질의 서열 정보를 기준으로 구조 및 표면 정보를 정렬함으로써 연결 관계를 반영한다. 이후 트랜스포머 아키텍처를 거쳐 각 모달리티의 이질성을 고려하여 통합할 수 있다. 이 같은 과정을 거쳐 단백질을 잘 모델링할 수 있는 세 모달리티의 통합된 임베딩을 구한다.

본 논문의 핵심 기여도를 요약하면 다음과 같다:

- 첫째, 서열 및 구조, 표면 같은 다양한 모달리티를 이용해 단백질을 모델링할 때, 단백질의 서열 정보를 기준으로 단백질의 구조, 표면 모달리티를 정렬함으로써 각 모달리티 간의 연결 관계를 모델에 반영했다.
- 둘째, 단백질의 표면 정보를 인코더로 한 트랜스포머 아키텍처를 구축해, 단백질의 서열 및 구조 정보와의 상호 관계를 이해하는 방식을 제안하였다.
- 셋째, 리간드의 3D 그래프 정보를 추출해, 단백질-리간드 결합 친화도를 예측할 때 단백질의 3D 정보인 구조 및 표면 정보와 호환시키는 접근법을 제안하였다.
- 넷째, 베이스라인 모델들과의 예측 성능 비교 실험과, 단백질 모달리티 정렬 및 리간드 3D 그래프 정보의 중

요성을 분석하기 위한 Ablation study를 통해 우리는 제안하는 모델의 우수함을 입증하였다.

2. 관련 연구

2.1 단백질 서열 기반 예측 모델

단백질은 아미노산의 펩타이드 결합으로 이루어져 있기 때문에, 아미노산 서열을 나열하는 것으로 해당 단백질을 표현할 수 있다. DeepDTA[5]는 오직 단백질의 서열과 약물의 SMILES 정보만으로 분자 그래프를 모델링하고, 1D CNN을 적용해 결합 친화도를 예측하는 모델이다. 하지만 단백질의 3D 구조 정보를 무시하고 서열만을 이용함에 따라, 실제 단백질-리간드 결합 시에 발생하는 공간적 제약을 학습하지 못하는 근본적 한계가 있다.

2.2 단백질 구조 기반 예측 모델

구조 기반 예측 모델에서는 일반적으로 단백질-리간드 복합체를 3D 공간에서의 Pixel이라 볼 수 있는 3D grid 형태로 나타내고 3D CNN을 적용하는 방식을 사용하거나, 복합체의 원자를 노드로 결합 여부를 엣지로 하는 그래프 형태로 나타내고 GNN을 적용하는 방식을 사용한다.

Pafnucy[6]는 단백질-리간드 복합체를 3D grid로 표현하고, 3D CNN을 활용해 결합 친화도를 예측하는 모델이다. 서열 기반 모델과 달리 공간 구조를 학습할 수 있지만, 3D grid의 해상도에 따라 높은 계산 비용이 발생하기 때문에 확장성에 제약이 있다. 또한 원자 회전과 변환에 민감하기 때문에 동일 분자일지라도 예측 오차가 증가할 수 있는 문제가 있다.

CurvAGN[7]은 곡률 기반 적응형 GNN을 사용해 결합 친화도를 예측하는 모델이다. GNN의 성능을 향상시키기 위해 단백질-리간드 복합체를 기하학적 관점에서 분석할 필요성을 강조하며, 이를 위해 복합체의 표현에 곡률 정보와 적응형 그래프 Attention 메커니즘을 반영하여, 단백질 표면의 국소적인 굴곡 정도에 따라 리간드 원자와의 결합 가능성을 다르게 평가한다.

2.3 단백질 표면 기반 예측 모델

단백질의 표면은 외부 생체분자와 직접적으로 상호작용하는 곳이기 때문에, 단백질 표면의 기하학적 및 화학적 특성이 두 생체분자 사이의 상호작용 결과를 도출하는 데 중요한 역할을 한다. MaSIF[2]는 단백질 표면을 삼각형 메쉬 형태로 표현하여, 결합 친화도를 예측하는 모델이다. Geodesic convolution을 적용해 단백질 표면의 기하학적 곡률과 화학적 특성을 학습할 수 있다. dMaSIF[3]는 MaSIF와 달리 단백질 표면을 삼각형 메쉬 형태가 아닌 포인트 클라우드 형태로 나타내 결합 친화도를 예측하는 모델이다. 샘플링 알고리즘으로 원자의 표면을 잘 나타낼 수 있는 포인트 클라우드를 생성하고, Quasi-geodesic convolution을 적용해 MaSIF에 비해 더 빠른 속도로 결과를 출력할 수 있다.

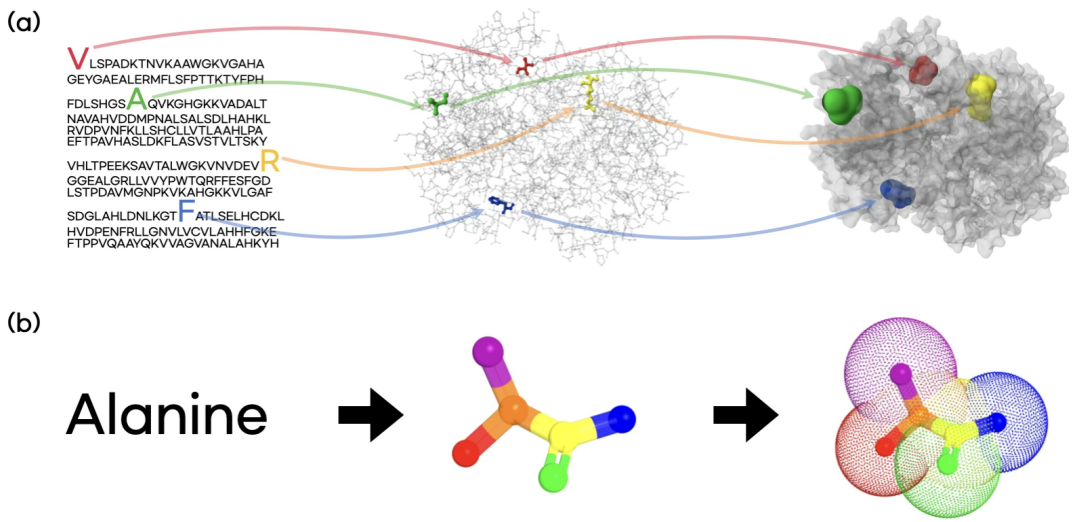

그림 2 단백질의 세 가지 모달리티 별 관계 모식도 (서열, 구조, 표면)

Fig. 2 Illustration of the relationship among three modalities of protein (sequence, structure, surface)

3. 단백질 모델링

3.1 단백질 정보 추출

그림 2의 (a)와 같이 단백질은 아미노산 서열, 아미노산을 구성하는 원자의 3차원 좌표, 메쉬 및 포인트 클라우드 형태의 표면으로 표현된다. 아미노산 서열(왼쪽)에서 Valine, Alanine, Arginine, Phenylalanine은 각각 같은 색으로 표시된 원자 구조(중앙)로 매핑된다. 마찬가지로 각 원자는 같은 색으로 표시된 표면 Point cloud(오른쪽)와 매핑된다. 3D 구조로 나타낸 단백질 표현은 각 원자들이 단백질 내에서 어떤 위치에 배치되고, 어떤 결합 각도로 Edge를 구성하고 있는지를 포함한다. 3D 표면으로 나타낸 단백질 표현은 해당 단백질이 외부 생체분자와 직접적으로 상호작용하는 부위이기 때문에, 단백질-리간드 결합과 관련한 중요한 정보를 갖고 있다. 이처럼 단백질은 같은 아미노산 서열로 구성되어 있을지라도, 여러 모달리티를 통해 표현함으로써 다양한 측면의 공간 구조에 대한 풍부한 정보를 가진 모델링이 가능하다.

단백질 서열 정보는 MLM(Masked Language Modeling) 방식으로 대규모 단백질 서열 데이터를 사전 학습한, ProtBert 모델[8]을 활용하여 인코딩한다. ProtBert는 단백질 서열을 아미노산 단위로 토큰화하여 학습한 모델이다.

단백질 구조 정보는 단백질을 구성하는 Residue의 연결 정보로 구축된 그래프 데이터를 통해 추출된다. 그래프의 적 정보를 정교하게 모델링 하는 것에 특화된 GVP 모델[9]을 채택하여 단백질의 구조 정보를 인코딩한다.

단백질 표면 정보는 dMaSIF[3]에서 제안된 샘플링 알고리즘을 사용하여 화학적 및 기하학적 특징을 추출한다. 원자 클라우드를 입력으로 받아 단백질 표면을 나타낼 수 있는 포인트 클라우드를 생성한다. 이중 리간드 중심에서 가장 가까운 512개의 표면 포인트만을 선택한다. 각 표면 포인트에 원자 정보를 주입하기 위해 16개의 가장 가까운 원자 중심의 위치를 찾아, 원자 유형에 대한 인덱스를 가지는 One-hot vector로 표현한다. 이후 각 표면 포인트에 대해 원자 타입 와 거리 정보 를 결합한 벡터를 생성하여 MLP의 입력으로 사용한다. 또한, 평균 곡률과 가우스 곡률을 계산하여 기하학적 특징으로 사용한다. 표면 곡률을 잘 반영하기 위해 Quasi-geodesic convolution을 적용하여 각 포인트의 법선, 접선 벡터를 정렬한다.

3.2 단백질 모달리티 정렬 모듈

그림 2의 (b)는 단백질의 세 가지 모달리티(서열, 구조, 표면) 간의 연결 관계를 표현하고 있다. 아미노산 서열(왼쪽)은 여러 원자가 결합된 형태이다. 단백질 구조 표현(중앙)은 원자 간의 결합 형태와 각도 등에 대한 정보를 갖고 있는데, 이는 하나의 아미노산 서열에 여러 개의 원자가 매핑된 것이다. 마찬가지로 각 원자의 주변에는 샘플링 알고리즘을 통해 생성된 포인트 클라우드가 형성되어 있는데, 이를 단백질 표면 표현(오른쪽)으로 나타내면 하나의 원자에 n개의 포인트가 매핑된 것이다. 이처럼 단백질 서열, 구조, 표면 모달리티의 연결 관계를 정의할 수 있다.

본 논문에서 제시하는 단백질 모달리티 정렬 모듈은 앞서 언급한 각 모달리티의 연결 관계를 모델에 반영한다. 서열 $S=\{s_1,\ldots,s_N\}$, 구조 $A=\{a_1,\ldots,a_M\}$, 표면 $P=\{p_1,\ldots,p_K\}$ 이 주어질 때, 단백질 모달리티 정렬 모듈은 각 s_i에 대응하는 a_j 및 a_j에 대응하는 p_k를 정렬한다. 그림 3을 보면, 아미노산 s_1에 매핑되어 있는 원자 a_1, a_2가 존재하고 각 원자에 매핑되어 있는 포인트 p_1, p_2, p_3 및 p_4, p_5, p_6가 존재한다. 기존 방식(왼쪽)에서는 통합된 단백질 임베딩을 구하기 위해 각 모달리티를 독립적으로 연산한다. 하지만 본 논문에서 제안하는 접근법(오른쪽)은 앞서 언급한 방식대로 아미노산 서열을 기준으로 그에 대응하는 원자 및 표면 모달리티의 임베딩을 각각 평균 풀링하는 방식으로 모달리티를 우선 정렬한다. 기존의 모델은 각 원자와 포인트가 어떤 서열에 속하는지 파악할 수 없다. 즉, 각 모달리티 간의 연관성에 대한 정보가 없기 때문에, 다양한 모달리티의 정보들을 종합적으로 고려하지 못하고 개별적으로 사용하는 문제점이 있었다. 위와 같은 정렬 방식을 사용하면, 모델이 아미노산 서열을 기준으로 각 모달리티의 연결 관계를 구축하여 통합된 단백질 임베딩을 구할 수 있다.

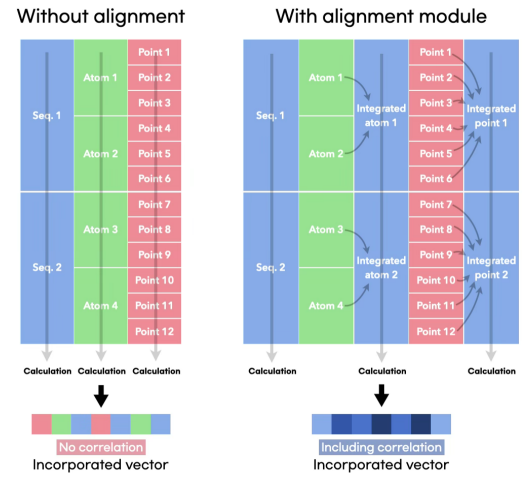

그림 3 기존 접근법(왼쪽)과 본 논문의 단백질 모달리티 정렬 모듈(오른쪽)

Fig. 3 Existing approach (left) and our protein modality alignment module (right)

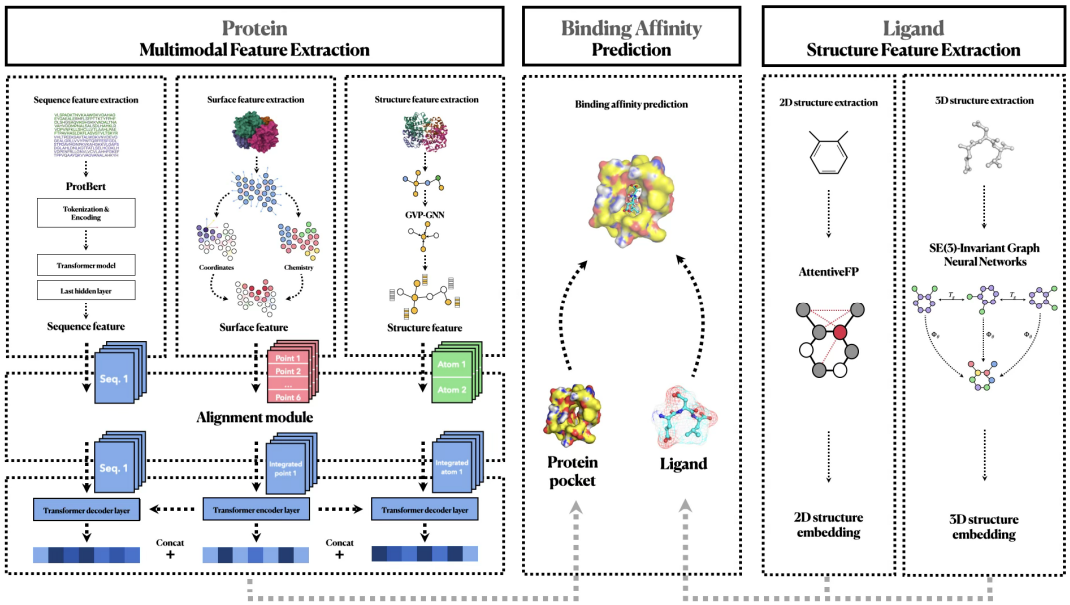

그림 4 본 논문에서 제시하는 프레임워크 모식도
Fig. 4 Illustration of the framework presented in this paper

3.3 트랜스포머 아키텍처

최근 트랜스포머 기반 연구들은 주로 단일 모달리티에만 트랜스포머를 적용하며, 다중 모달리티 기반 연구의 경우 트랜스포머를 적용하지 않고 단순히 각 모달리티의 임베딩을 연결하는 방식을 사용한다. 예를 들어, CAPLA[10]는 Cross-attention을 기반으로 단백질 전역 서열, 포켓 서열, 리간드 SMILES 문자열을 통합하는 방식을 사용한다. 하지만 3D 구조 정보나 표면 정보를 사용하지 않는 한계가 있다. 또한 HaPPy[11]는 단백질 및 리간드 서열 정보와 단백질 구조 정보를 포함하는 다중 모달리티를 사용하였으나, 서로 다른 모달리티 간의 이질성을 무시하여 단순히 연결(Concat)했기 때문에 각 모달리티 간의 관계성을 잃어버릴 수 있다. 따라서 최종적인 단백질 임베딩을 구하기 위해서, 단백질 모달리티 정렬 모듈을 통해 출력된 각 모달리티의 특징은 트랜스포머 아키텍처를 거쳐야 한다.

트랜스포머는 Self-attention 및 Cross-attention을 활용하여 서로 다른 모달리티 간의 관계를 학습할 수 있다. 우선 단백질 표면 정보를 트랜스포머 인코더에 입력하여, Self-attention을 통해 표면 정보를 학습한다. 이후 서열 및 구조 정보를 각각 트랜스포머 디코더에 입력으로 주어 인코더의 표면 정보와 Cross-attnetion을 수행한다. 위와 같이 트랜스포머 아키텍처를 이용해 서열 및 구조 정보와 표면 임베딩을 전역적으로 연결할 수 있다. 이후 구해진 표면, 서열, 구조 임베딩에 대해 각각 평균 풀링을 적용한다. 이런 방식을 통해, 각 모달리티의 서로 다른 이질성을 반영하면서도 효과적으로 통합된 임베딩을 구할 수 있다.

4. 리간드 모델링

4.1 리간드 2D 정보 추출

리간드 2D 정보는 리간드 구조의 위상학적 형상을 담고 있는 데이터 타입이며, 이를 모델링 하기 위해 노드를 원자 타입으로, 엣지를 원자 간의 공유결합으로 설정된 그래프를 구축하였다. 리간드를 표현하는 해당 그래프의 2D 구조는 어텐션 메커니즘을 통해 원자 간의 복잡한 상호작용을 학습할 수 있는 AttentiveFP[12]를 활용하여 인코딩 되었다.

4.2 리간드 3D 정보 추출

그림 4의 오른쪽에 표현된 SE(3)-invariant GNN은 평행 이동과 회전에 대해 불변성을 가진다. 여기서 SE(3)는 유클리드 공간에서의 평행 이동, 회전을 포함하는 특수 유클리드 그룹을 의미한다. SE(3)-invariance 특성을 가지지 않은 기존의 일반적인 GNN은, 리간드 분자를 이루고 있는 원자 사이의 거리 정보 $\| x_i - x_j \|^2$를 연산에 사용하지 않는다. 때문에 리간드 분자가 유클

리드 공간에서 평행 이동 및 회전할 경우 완전히 다른 분자로 고려될 수 있는데, 이는 모델에 충분한 귀납적 편향을 주입하지 못해 일반화 능력을 떨어뜨리는 원인이 될 수 있다. 그에 반해 불변성 그래프 신경망은 입력에 회전과 이동을 의미하는 SE(3)-transformation T_g를 적용해도 출력이 동일하다: $f(g(x)) = f(x)$. 리간드 3D 정보는 3차원 공간 상의 기하학적 정보를 담고 있는 데이터 타입이며, 이는 SE(3)-invariant GNN[13]을 활용하여 인코딩 될 수 있다. 모델을 구성하는 레이어 l에서의 노드 임베딩을 h_i^l, 엣지 e_{ij}에 대한 속성을 a_{ij}라고 할 때 SE(3)-invariant GNN에서 메시지 값과 노드 임베딩 값은 수식 1~3과 같이 구할 수 있다. 여기서 φ_e와 φ_h는 각각 다중 레이어 퍼셉트론(MLP)으로 근사화되는 에지 연산과 노드 연산을 의미한다. 리간드 그래프에 SE(3)-invariant를 부여한 이유는, 리간드 분자의 결합 친화도가 분자의 절대적인 위치나 방향에 의존하지 않기 때문이다. 즉, 리간드가 단백질에 결합하는 능력은 공간에서의 배치나 방향과 무관하게 상대적인 기하학적 관계에 의해 결정된다. 이러한 특성을 모델링 하기 위해 SE(3)-invariant 그래프 신경망을 적용하였다.

$$m_{ij} = \varphi_e(h_i^l, h_j^l, \| x_i^l - x_j^l \|^2, a_{ij}) \quad (1)$$

$$m_i = \sum_{j \in N(i)} m_{ij} \quad (2)$$

$$h_i^{l+1} = \varphi_h(h_i^l, m_i) \quad (3)$$

5. 실험 및 성능 평가

5.1 데이터셋 및 학습 시간

PDBbind 데이터셋(2016 버전)은 Protein Data Bank의 생체분자 복합체와 실험적으로 측정된 결합 친화도 데이터가 포함되어 있으며, 단백질-리간드 결합 친화도 예측 과제에 자주 사용된다. 훈련 세트로 11163개를, 검증 세트로 1000개를, 평가 세트로 285개의 데이터를 사용하였다.

학습은 단일 NVIDIA RTX 3090 GPU에서 수행되었으며, 약 12000개의 단백질-리간드 쌍으로 구성된 데이터셋을 대상으로 약 16시간이 소요되었다. 이때 배치 크기는 8, Adam 옵티마이저, 학습률은 0.0001로 설정되었다. 학습 중 GPU 메모리 사용량은 최대 21GB로, NVIDIA RTX 3090의 24GB 메모리 내에서 충분히 처리 가능했다. 추론 시에는 단백질-리간드 쌍당 약 300ms가 소요되어, 대규모 가상 스크리닝을 수행하기에 적합하다.

표 1 베이스라인 모델들과 비교한 성능 평가 결과
Table 1 Performance evaluation results compared to baseline models

Model	RMSE	MAE	SD	R ↑
DeepDTA(2018)[5]	1.443	1.148	1.445	0.749
Pafnucy(2018)[6]	1.418	1.129	1.375	0.775
CurvAGN(2023)[7]	1.217	**0.930**	1.191	0.830
dMaSIF(2021)[3]	1.324	1.067	1.277	0.809
DPLA(2021)[14]	1.255	0.972	1.248	0.820
HaPPy(2023)[11]	1.228	0.936	1.221	0.827
MFE(2024)[4]	1.245	0.988	1.218	0.828
Ours	**1.215**	0.967	**1.160**	**0.845**

우리 모델의 전반적인 성능을 평가하기 위해, 단백질 결합 친화도 예측 모델로 유명한 베이스라인들과 비교했다. 서열 기반 방법을 사용하는 DeepDTA[5], 구조 기반 방법을 사용하는 Pafnucy[6], CurvAGN[7], 표면 기반 방법을 사용하는 dMaSIF[3], 다중 모달리티를 사용하는 DPLA[14], HaPPY[11], MFE[4]가 있다. 모델의 성능을 평가하기 위해 평균 제곱근 오차(RMSE), 평균 절대 오차(MAE), 표준 편차(SD), 피어슨 상관 계수(R)를 사용했다.

5.2 결합 친화도 예측 실험

표 1은 우리 모델과 베이스라인 모델들의 성능을 평가한 것이다. MAE를 제외한 모든 지표에서 우리 모델이 가장 높은 성능을 보였으며, CurvAGN 모델이 두 번째로 높은 성능을 보였다. DeepDTA는 단백질 서열 정보에만 의존하여 분자 그래프를 모델링 하였고, 공간 정보를 무시했기 때문에 가장 낮은 성능을 보였다[5]. CurvAGN은 곡률 기반 적응형 GNN과 그래프 어텐션 메커니즘을 적용하였기 때문에, 단백질의 기하학적 정보를 합리적으로 추출하여 좋은 성능을 보이는 것으로 해석할 수 있다[7]. dMaSIF는 단백질 표면의 화학적 및 기하학적 특징을 강조하고 Quasi-geodesic convolution을 사용해 표면 지문을 학습했지만, 단백질 표면 정보만 사용할 경우 국소적 특징에 편향되기 때문에 단일 모달리티의 한계로 높은 성능을 보이지 못했다[3]. MFE는 서열, 구조, 표면 정보에서 추출한 다중 모달리티를 정렬하였으나, 단백질 모달리티 정렬 모듈이 부재하며 리간드 그래프를 2D 구조로만 구현한 점이 한계로 작용했다[4]. 우리 모델의 경우 단백질 모달리티 정렬 모듈을 통해 연결 관계를 명시했고, 리간드 구조를 단백질 구조와 동일한 3D 구조로 구현하여 높은 성능을 보일 수 있었다.

표 2 리간드 3D 그래프 유무, 트랜스포머 인코더 종류, 단백질 모달리티 정렬 여부에 따른 성능 평가 결과
Table 2 Performance evaluation results according to presence or absence of ligand 3D graph, type of transformer encoder, and protein modality alignment

	Ligand 3D graph		Transformer encoder			Alignment module		
	w/o	**w/**	Sequence	**Surface**	Seq.+Surf.	w/o	w/	**Combined**
RMSE	1.245	**1.215**	1.220	**1.215**	1.271	1.229	**1.211**	1.215
MAE	0.988	**0.967**	**0.947**	0.967	1.010	0.970	**0.946**	0.967
SD	1.218	**1.160**	1.196	**1.160**	1.239	1.220	1.188	**1.160**
R ↑	0.828	**0.845**	0.835	**0.845**	0.822	0.828	0.838	**0.845**

5.3 Ablation Study

표 2의 왼쪽(Ligand 3D graph)은 리간드 3D 그래프 구조를 사용한 경우, 사용하지 않은 경우를 비교한 Ablation study 결과이다. 리간드 3D 구조 정보를 함께 주입한 경우에 더 좋은 성능을 보였는데, 이는 단백질-리간드 결합 친화도를 예측할 때 단백질 정보뿐만 아니라 리간드 정보의 깊이 있는 분석이 필요하다는 것을 시사한다.

표 2의 중앙(Tranformer encoder)은 트랜스포머 인코더 종류를 각각 서열, 표면 정보로 한 경우와 두 결과를 합친 경우를 비교한 Ablation study 결과이다. 표면 정보를 인코더로 한 경우가 서열 정보를 인코더로 한 경우보다 성능이 좋다. 이는 서열 정보를 기준으로 모달리티를 정렬했다 하더라도, 리간드와 직접적으로 상호작용하는 표면 정보를 중심으로 전역적인 단백질 서열 및 구조 정보를 통합하는 것이 단백질을 모델링하기에 더 유리하기 때문이다. 두 인코더에서 출력된 결과를 각각 합치는 방식은 모델에 너무 많은 정보를 유입하기 때문에, 과적합을 유발하여 예측 능력이 떨어지는 것을 확인할 수 있다.

표 2의 오른쪽(Alignment module)은 단백질 모달리티 정렬 모듈을 사용하지 않은 경우와 사용한 경우, 두 결과를 합친 경우를 비교한 Ablation study 결과이다. 정렬 모듈을 통해 단백질 모달리티를 서열 기준으로 정렬한 쪽이 정렬하지 않은 쪽보다 전반적으로 성능이 좋다. 이는 단백질 모델링 과정에서 각 모달리티를 정렬하여 연결 관계를 주입할 필요성을 시사한다. 정렬하지 않은 임베딩과 정렬한 임베딩을 Summation 연산을 통해 합친 경우가 가장 성능이 우수한 것은, 단백질을 아미노산 범위만을 고려하는 것보다 단백질 전체 범위를 고려한 임베딩을 함께 사용하는 것이 단백질을 더 잘 표현한다는 것을 시사한다.

5.4 시각화

그림 5는 단백질 모달리티 정렬 모듈을 적용하기 이전(왼쪽)과 이후(오른쪽)의 PCA 차원 축소 결과를 시각화한 것이다. 단백질 모달리티 정렬을 하지 않은 경우 단백질 구조 정보(검정)와 표면 정보(회색) 사이에 패턴이 반영되지 않고 따로 분포되어 있다. 이는 두 정보가 독립적으로 존재하며, 연결 관계가 반영되지 않은 상태임을 의미한다. 단백질 모달리티를 정렬한 경우 단백질 구조 정보(파랑)와 표면 정보(빨강)가 군집을 형성하며 서로 짝을 이루는 형태로 배치되어 있다. 이는 두 정보 사이의 연결 관계가 반영되었음을 의미한다. 차원 축소 시각화 결과는 본 논문의 단백질 모달리티 정렬 모듈이 단백질 서열 정보를 기준으로 구조 및 표면 정보를 효과적으로 정렬한 것을 보여준다. 이를 통해 단백질의 다중 모달리티의 연결 관계를 모델에 주입할 수 있으며, 다중 모달리티를 이용하여 단백질을 모델링할 때 정렬 모듈의 중요성을 확인할 수 있다.

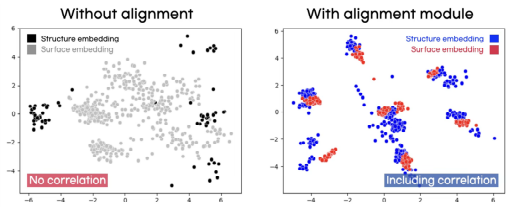

그림 5 단백질 모달리티 정렬 모듈 유무에 따른 PCA 차원 축소 시각화 결과

Fig. 5 PCA dimensionality reduction visualization results with and without protein modality alignment module

6. 결 론

본 연구는 단백질-리간드 결합 친화도 예측을 위해, 단백질 모달리티 정렬 모듈 및 리간드 정보의 SE(3)-invariant 그래프 신경망 학습을 통합한 딥러닝 아키텍처를 제안했다. 단백질 모달리티 정렬 모듈은 단백질의 구조 및 표면 정보를 서열을 기준으로 정렬함으로써 단백질 다중 모달리티의 연결 관계를 모델에 주입했다. 또한 리간드의 2D 및 3D 구조를 함께 활용해 리간드의 3차원 공간 구조 정보를 파악할 수 있게끔 유도하였다. 본 논문의 접근법은 베이스라인 모델과의 비교와 Ablation study 결과를 통해, 단백질-리간드 복합체를 효과적으로 모델링할 수 있음을 실험적으로 입증하였다. 일반적으로 신약 개발에 있어서 잠재적 리드를 식별하는 가상 스크리닝 단계나, 리드 화합물의 구조를 수정하는 리드 최적화 단계에 오랜 시간이 소요된다. 제안된 모델을 통해 단백질-리간드 복합체의 결합 친화도를 예측하면, 기존의 방식과 비교하여 더 빠르고 효율적으로 잠재적 리드를 선별하고 리드 화합물의 구조를 최적화할 수 있다. 이를 통해 시간과 자원을 절약하면서도 신뢰할 수 있는 약물 후보를 신속히 도출할 가능성을 높일 수 있다. 하지만 단백질의 구조 정보가 여전히 하나의 알파카본 원자만을 갖고 있는 것은, 단백질의 구조를 정확히 명시하는 데 한계로 작용할 수 있다. 이를 해결하기 위해 단백질의 구조 정보가 다양한 원자를 같이 반영하면서도, 연산의 양을 적절한 수준으로 조절할 수 있는 작업을 향후 연구에서 진행하고자 한다.

References

[1] Y. Wang, Q. Jiao, J. Wang, X. Cai, W. Zhao, and X. Cui, "Prediction of protein-ligand binding affinity with deep learning," *Computational and Structural Biotechnology Journal*, Vol. 21, pp. 5796-5806, 2023.

[2] P. Gainza, F. Sverrisson, F. Monti, E. Rodola, D. Boscaini, M. M. Bronstein, and B. E. Correia, "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning," *Nature Methods*, Vol. 17, pp. 184-192, 2020.

[3] F. Sverrisson, J. Feydy, B. E. Correia, and M. M. Bronstein, "Fast End-to-End Learning on Protein Surfaces," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 15272-15281, 2021.

[4] S. Xu, L. Shen, M. Zhang, C. Jiang, X. Zhang, Y. Xu, J. Liu, and X. Liu, "Surface-based multimodal protein-ligand binding affinity prediction," *Bioinformatics,* Vol. 40, No. 7, btae413, Jul. 2024.

[5] H. Ozturk, A. Ozgur, and E. Ozkirimli, "DeepDTA: deep drug-target binding affinity prediction," *Bioinformatics,* Vol. 34, No. 17, pp. i821-i829, Sep. 2018.

[6] M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, and P. Siedlecki, "Development and evaluation of a deep learning model for protein-ligand binding affinity prediction," *Bioinformatics,* Vol. 34, No. 21, pp. 3666-3674, Nov. 2018.

[7] J. Wu, H. Chen, M. Cheng, and H. Xiong, "CurvAGN: Curvature-based Adaptive Graph Neural Networks for Predicting Protein-Ligand Binding Affinity," *BMC Bioinformatics,* Vol. 24, article 378, Oct. 2023.

[8] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, and L. Jones, "ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 44, No. 10, pp. 7112-7127, Oct. 2022.

[9] R. Zheng, Z. Huang, and L. Deng, "Large-scale predicting protein functions through heterogeneous feature fusion," *Briefings in Bioinformatics,* Vol. 24, Issue 4, bbad243, July 2023.

[10] Z. Jin, T. Wu, T. Chen, D. Pan, X. Wang, J. Xie, L. Quan, Q. Lyu, "CAPLA: improved prediction of protein-ligand binding affinity by a deep learning approach based on a cross-attention mechanism," *Bioinformatics,* Vol. 39, No. 2, btad049, Feb. 2023.

[11] X. Zhang, Y. Li, J. Wang, G. Xu, Y. Gu, "A multi-perspective model for protein-ligand-binding affinity prediction," *Interdisciplinary Sciences: Computational Life Sciences,* Vol. 15, pp. 696-709, Dec. 2023.

[12] Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, Z. Li, X. Luo, K. Chen, H. Jiang, and M. Zheng, "Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism," *Journal of Medicinal Chemistry*, Vol. 63, No. 16, pp. 8749-8760, Aug. 2020.

[13] V. G. Satorras, E. Hoogeboom, and M. Welling, "E(n) Equivariant Graph Neural Networks," *Proc. of the 38th International Conference on Machine Learning,* PMLR, Vol. 139, pp. 9323-9332, 2021.

[14] W. Wang, B. Sun, D. Liu, X. Wang, and H. Zhang, "DPLA: Prediction of protein-ligand binding affinity by integrating multi-level information," *Proc. of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* IEEE, pp. 3428-3434, 2021.

이 승 용

2025년 홍익대학교 컴퓨터공학과 졸업 (학사). 관심분야는 AI 기반 신약 개발

박 상 현

1989년 서울대학교 컴퓨터공학과 (학사) 1991년 서울대학교 대학원 컴퓨터공학과 (공학석사). 2001년 UCLA 대학원 컴퓨터과학과(공학박사). 1991년~1996년 대우통신 연구원. 2001년~2002년 IBMT. J. Watson Research CenterPostDoctoral Fellow. 2002년~2003년 포항공과대학교 컴퓨터 공학과 조교수. 2003년~2006년 연세대학교 컴퓨터과학과 조교수 2006년~2011년 연세대학교 컴퓨터과학과 부교수. 2011년~현재 연세대학교 컴퓨터과학과 교수. 관심분야는 데이터베이스, 데이터 마이닝, 바이오인포매틱스, 빅데이터 마이닝 & 기계 학습