

# 단백질-리간드 결합 부위 및 친화도 예측 모델 연구



Team 으쌔으쌔

정보컴퓨터공학부

202255512 김다현

202255552 박주은

202255565 안수현

지도교수: 송길태

# 목차

1. 요구조건 및 제약 사항 분석에 대한 수정사항
  - a. 기존 요구조건
  - b. 요구조건 수정사항
  - c. 기존 제약 사항 및 개선 방안
2. 설계 상세화 및 변경 내역
3. 갱신된 과제 추진 계획
4. 구성원별 진척도
5. 보고 시점까지의 과제 수행 내용 및 중간 결과

## 1. 요구조건 및 제약 사항 분석에 대한 수정사항

### a. 기존 요구조건

단백질과 리간드 간의 결합 부위(binding site)와 결합 친화도(binding affinity)를 예측할 수 있는 딥러닝 기반의 예측 모델을 개발하며 이를 쉽게 활용할 수 있도록 웹페이지 형태의 인터페이스로 구현한다.

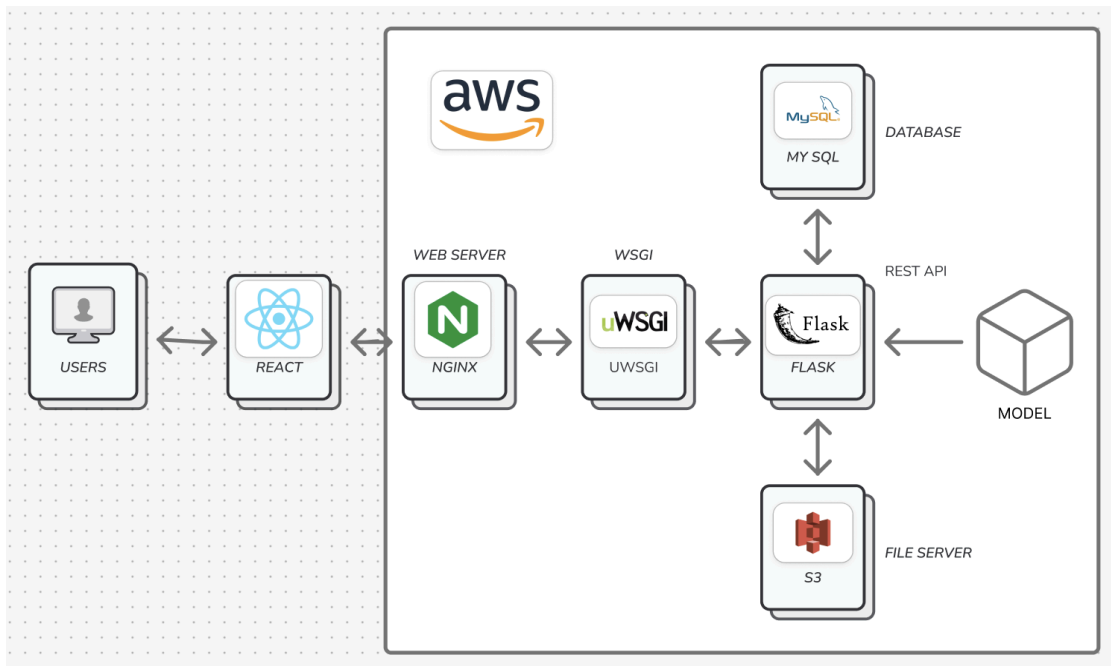
### b. 요구조건 수정사항

모델의 예측 속도 및 시스템 응답 시간에 대한 논의를 바탕으로, 모델 개발 과정에서 해당 성능 지표를 고려하여 향상시키기로 하였다.

### c. 기존 제약 사항 및 개선 방안

착수 보고서에 작성한 현실적 제약 사항에 대한 개선 가능성에 대해 논의하였다. 생체 환경 요소를 고려하려고 했으나 pH, 이온 농도 등의 요소는 모든 데이터셋 파일에 포함되어 있지는 않아 모델에 직접 반영이 어려울 것으로 판단하였다. 반면 물리화학적 요인의 경우 초기에는 반영이 어려울 것으로 예상되었으나, 해당 특성들이 결합 친화도에 중요한 영향을 미친다는 점을 고려하여, 학습 시 단백질과 리간드의 물리화학적 특성을 함께 입력에 포함하여 반영하기로 하였다. 또한, 복합체 수준의 전역적 특성을 학습에 포함시키는 방안도 고려했으나, 사용자의 데이터 입력 방식이 단백질 시퀀스와 리간드 정보로 제한되므로 학습과 테스트 환경 간 불일치가 발생할 수 있어, 각각의 개체 단독으로부터 얻을 수 있는 특성만을 학습에 활용하기로 하였다.

## 2. 설계 상세화 및 변경 내역



### ★ 서비스 구조

#### a. 사용자 접속 및 로그인

- 사용자는 AWS EC2에 배포된 웹 애플리케이션에 접속
- 회원가입 및 로그인 후 서비스 이용

#### b. 단백질 & 리간드 입력

- 사용자는 웹 페이지 인터페이스를 통해 단백질 파일과 리간드 파일을 업로드

#### c. 예측 요청

- 사용자가 입력한 데이터가 Axios를 통해 REST API로 전달
- 백엔드 서버(Flask)는 데이터를 받아
  - Biopython, RDKit, DSSP로 전처리
  - PyTorch 기반 예측 모델에 입력

#### d. 예측 수행

- Protein Encoder (ESM2), Ligand Encoder (Morgan FP)를 통해 피쳐 추출
- CNN + Cross Attention 기반 모델로
  - 결합 부위 (Binding Site) 예측
  - 결합 친화도 (Binding Affinity) 예측

#### e. 예측 결과 반환 및 시각화

- 예측 결과를 JSON 형태로 프론트엔드로 반환
- 프론트엔드에서
  - NGL Viewer로 결합 부위를 3D 시각화
  - 결합 친화도 값을 숫자로 표시

#### f. 결과값 저장

- 예측 결과는 데이터베이스에 저장
  - 사용자 ID
  - 입력한 단백질, 리간드 정보
  - 결합 부위 및 친화도 예측 결과


#### g. 저장된 결과 조회

- 사용자는 예측 결과를 웹 인터페이스에서 조회 가능
  - 본인이 저장한 결과를 목록으로 확인
  - 결과 상세 페이지에서 시각화 및 친화도 재확인

사용자의 인터페이스 예시는 아래와 같다.


- 사용자는 파일을 드롭하여 단백질 및 리간드를 입력한다.
- 그 후, 단백질 및 리간드의 3d 구조를 확인할 수 있다 . (NGL viewer 사용)

- 입력된 단백질 파일의 **sequence**를 확인할 수 있고, 그 중 **binding site**에 해당하는 부분은 **highlight** 표시가 되어있다. 표시에 마우스를 가져다 대면 **binding affinity**를 확인할 수 있다.




**PLANET - X**

**PROTEIN**




클릭 혹은 파일을 이곳에 드롭하세요.  
파일당 최대 3MB

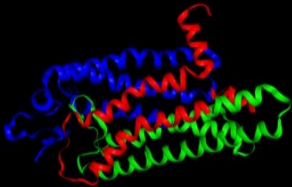
**LIGAND**

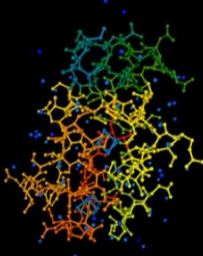


클릭 혹은 파일을 이곳에 드롭하세요.  
파일당 최대 3MB

**RUN**







**BINDING SITE**

98%

**Chain A**

```

STAGKYIKCKAAVLWEEKKPPFIEEEVAPPKAHEVRIMVATGTCRSDDIVVSGTVTP 60
LPVLAGAAGIVESIGEGVTVVRPGDKVIFLTPQCGKCRVCKHPEGNFCLKND SMPR 120
GTMQDGTSRFTCRGKPIIHIFIGTSFSQYTVYDEISVAKIDAASPLEKVCFIGGFSIGY 180
GSAVKYAKVTQGSTCAVEGLGGGLSVIMECKAAGAARIIGVDINKDFAKAKEVGATEC 240
VNPDYKKPIQEVLTESNGGVDFSFEVIGRLDMVTALSCCQAYGVSVVGVPPDSQN 300
LSMNPILLSGRTWKGATGGFKSKDSVPKLYADFMAKKALDPLITHVLPFEKINEGFD 360
LRSGESITITLF
                    
```

**Chain B**

```

STAGKYIKCKAAVLWEEKKPPFIEEEVAPPKAHEVRIMVATGTCRSDDIVVSGTVTP 60
LPVLAGAAGIVESIGEGVTVVRPGDKVIFLTPQCGKCRVCKHPEGNFCLKND SMPR 120
GTMQDGTSRFTCRGKPIIHIFIGTSFSQYTVYDEISVAKIDAASPLEKVCFIGGFSIGY 180
GSAVKYAKVTQGSTCAVEGLGGGLSVIMECKAAGAARIIGVDINKDFAKAKEVGATEC 240
VNPDYKKPIQEVLTESNGGVDFSFEVIGRLDMVTALSCCQAYGVSVVGVPPDSQN 300
LSMNPILLSGRTWKGATGGFKSKDSVPKLYADFMAKKALDPLITHVLPFEKINEGFD 360
LRSGESITITLF
                    
```

**Chain D**

```


STAGKYIKCKAAVLWEEKKPPFIEEEVAPPKAHEVRIMVATGTCRSDDIVVSGTVTP 60
LPVLAGAAGIVESIGEGVTVVRPGDKVIFLTPQCGKCRVCKHPEGNFCLKND SMPR 120
GTMQDGTSRFTCRGKPIIHIFIGTSFSQYTVYDEISVAKIDAASPLEKVCFIGGFSIGY 180
GSAVKYAKVTQGSTCAVEGLGGGLSVIMECKAAGAARIIGVDINKDFAKAKEVGATEC 240
VNPDYKKPIQEVLTESNGGVDFSFEVIGRLDMVTALSCCQAYGVSVVGVPPDSQN 300
LSMNPILLSGRTWKGATGGFKSKDSVPKLYADFMAKKALDPLITHVLPFEKINEGFD 360
LRSGESITITLF
                    
```

**Chain C**

```

STAGKYIKCKAAVLWEEKKPPFIEEEVAPPKAHEVRIMVATGTCRSDDIVVSGTVTP 60
LPVLAGAAGIVESIGEGVTVVRPGDKVIFLTPQCGKCRVCKHPEGNFCLKND SMPR 120
GTMQDGTSRFTCRGKPIIHIFIGTSFSQYTVYDEISVAKIDAASPLEKVCFIGGFSIGY 180
GSAVKYAKVTQGSTCAVEGLGGGLSVIMECKAAGAARIIGVDINKDFAKAKEVGATEC 240
VNPDYKKPIQEVLTESNGGVDFSFEVIGRLDMVTALSCCQAYGVSVVGVPPDSQN 300
LSMNPILLSGRTWKGATGGFKSKDSVPKLYADFMAKKALDPLITHVLPFEKINEGFD 360
LRSGESITITLF
                    
```

**BINDING AFFINITY**



## ★ 모델 구조

### Binding site 예측 모델

- 사용자에게 단백질과 리간드 파일을 입력으로 받아, 단백질과 리간드를 기반으로 **binding site**를 예측하게 된다.
- 처음의 계획은 트랜스포머 기반 화합물과 표적 단백질의 결합 부위 예측 모델 개발이었다. 이를 기반으로 관련 논문들을 공부하던 도중, 단백질의 **sequence**를 입력으로 받아 결합 부위를 예측하는 모델의 대부분이 **CNN**과 **attention**을 함께 활용하는 것을 알게되었다. 아무래도 **motif**와 같은 근거리의 특징을 파악하기에는 **CNN**을 활용하는 것이 맞으나, **CNN**만을 활용하면 원거리의 관계를 파악하기 어렵기 때문이다. 그래서 이번 프로젝트에 본 팀 역시 **CNN**에 **attention**을 함께 활용한다.
- 모델을 학습시키기 전, 단백질 **sequence**를 3D 구조 정보를 반영한 **sequence representation**으로 바꾸고 싶었다. 이를 위해 사용할 모델의 후보로는 **Protbert**와 **ESM-2**가 있었다. 성능 비교 결과, **ESM-2**의 속도가 더 느리기는 하나, 더 풍부한 **contextual feature**를 보유하고 있으며, 더 정교한 패턴 반영 가능하여 **ESM-2**를 사용하기로 결정하였다.

### Binding affinity 예측 모델

- 사용자로부터 입력받은 단백질, 리간드 파일과 **Binding site** 예측 모델로부터 도출된 **binding site** 정보를 기반으로 **binding affinity**를 예측한다.
- 초기 설계 단계에서는 단백질과 리간드의 물리화학적 특성 반영 여부에 대해 고민이 있었으나 해당 특성들이 결합 친화도에 중요 요인으로 작용한다는 점을 고려하여 이를 입력 임베딩에 반영하기로 하였다. 이에 따라 각 **residue**에 대해 극성, 비극성, 산성, 염기성 및 클러스터 정보를 포함할 수 있게끔 임베딩을 진행한다.
- 임베딩된 입력값은 **CNN**을 통해 로컬 구조적 특성과 문맥에 민감한 표현을 효과적으로 학습하며, 이후 **cross-attention** 메커니즘을 통해 단백질과 리간드 간의 상호작용 정보를 통합적으로 반영할 수 있도록 한다. 최종적으로 이 정보를 기반으로 결합 친화도를 회귀 형태로 예측한다.

- 또한 단백질 및 리간드의 전역적인 물리화학 특성(LogP, GRAVY 등) 역시 예측에 기여할 수 있는 요소로 고려되어, 회귀 단계에서 추가적인 피처로 활용될 수 있도록 설계의 유연성을 확보하였다.

### 3. 갱신된 과제 추진 계획

7월		8월				9월		
3	4	1	2	3	4	1	2	3
기술 분석								
임베딩 적용								
모델 아키텍처 구현								
		모델 성능 개선 및 최적화						
	웹 인터페이스 개발 및 연동							
							최종보고서	

### 4. 구성원별 진척도

공통	<ul style="list-style-type: none"> <li>- 논문 분석</li> <li>- 모델 구조 상세화</li> </ul>
김다현	<ul style="list-style-type: none"> <li>- COACH420 데이터 전처리</li> <li>- HOLO4K 데이터 전처리</li> </ul>
박주은	<ul style="list-style-type: none"> <li>- scPDB 데이터 전처리</li> </ul>
안수현	<ul style="list-style-type: none"> <li>- PDBbind 데이터 전처리</li> <li>- binding site 예측 모델 임베딩</li> </ul>



## 5. 보고 시점까지의 과제 수행 내용 및 중간 결과

현재, 모델을 학습시키기 위한 **training data**와 모델의 성능을 검증하기 위한 **test data**에 대한 전처리를 완료했다.

우선, **training data**로는 실험을 기반으로 하여 검증된 표준 벤치마크로 사용되는 **PDBbind**, **scPDB**를 사용하였다. **training data**의 단백질 **sequence** 길이는 1,000자로 제한하였다. 너무 긴 길이의 **sequence**를 학습하는 과정에서 모델의 성능이 떨어질 것을 우려한 결정이다.

리간드 **sequence**를 **SMILES**로 변환하는 과정에서, 처음에는 **rdkit**를 사용하려고 했다. 그러나, **kekulization**을 하는 과정에서 **rdkit**는 명확하고 확실한 경우를 따져 **kekulization**이 수행되지 않는 경우가 많았다. 이렇게 되면, 학습 과정에 입력하는 데이터의 형태가 달라지는 문제가 생겨 고민한 결과 **obabel**을 **SMILES**로 사용하면 어떨지에 대한 논의가 나왔다.

**obabel**은 **rdkit**보다 더 관대한 방식으로 **kekulization**을 수행하여, 결합이 애매하거나 부정확한 경우도 어쨌든 그럴듯한 **kekule**구조로 바꿔준다. 만약 **rdkit**를 사용한다면 **kekulization**이 적용되지 않은 데이터와 적용이 된 데이터는 형식이 다르기 때문에, 이에 대한 처리에 대해 **kekulization**이 되지 않는 데이터는 제외시킬까도 고민하였다. 그러나 그렇게 되면 학습시킬 수 있는 **data**의 수가 급격히 감소할 것으로 예상되어 **obabel**을 사용하기로 했다.

**test data**로는 **COACH420**과 **HOLO4K**를 사용하였으며, 복합체 리스트를 기반으로 **UCSF Chimera**를 이용해 구조를 불러오고, 리간드 코드 및 결합 부위 정보를 함께 로딩하는 과정을 거쳤다. 이후, **test data** 역시 단백질 서열 길이가 1,000자를 초과하는 샘플을 제거하였고, 리간드의 **SMILES** 길이가 160자를 초과하는 샘플도 제거하며, 각 단백질에 대해 **UniProt ID** 매핑을 수행하여 **UniProt**과 매핑되지 않는 단백질을 제거하였다. 여기서 **Uniprot**은 믿을 수 있는 단백질 서열 **DB**이다.

그 뒤, **binding site** 예측 모델을 학습 시키기 전 단백질 데이터의 임베딩까지 진행하였다. **ESM-2** 모델을 사용하여, 단백질 **sequence**를 구조 정도를 반영한 **sequence representation**으로 변환하여 3D구조가 아니라 **sequence**를 활용하더라도 충분히 **binding site**를 잘 예측할 수 있도록 하였다.