Neuro-Ligand Discovery Pipeline

뇌질환 표적 단백질 기반 리간드 선별, 생성 파이프라인

작성자: Dayeon Youm

작성일: 2026-01-12

버전: v1.0

Guideline

한 줄 요약	선정한 뇌질환 관련 표적 단백질에 대해 1 만 개 규모의 리간드 라이브러리를 물성 규칙으로 1 차 선별 후, 머신러닝 모델로 상위 후보(Top 10)를 도출한 뒤, 구조 기반 결합 예측과 3D 모델링으로 검증한다.
문제 정의	광범위한 후보 화합물 중 실제 약물 가능성이 높은 리간드를 효율적으로 찾기 어렵다(탐색 공간 큼, 실험 비용이나 시간 높음).
핵심 접근	Lipinski Rule of Five 등 물리/화학적 규칙 기반 필터링(약 10,000→2,000) + 학습 데이터(5,000–10,000) 기반 예측 모델 + 도킹/구조 기반 검증(Boltz-2 등) + 생성 모델(Boltzgen 등) 연계
주요 산출물	데이터셋/코드 repository, 후보 리간드 Top 10 목록, 모델 성능 리포트, 최종 선정 리간드에 대한 도킹 결과 및 단백질-리간드 복합체 3D 시각화, 최종 보고서.

목차

1. 프로젝트 개요

1.1 배경 및 필요성

뇌질환(e.g. 알츠하이머, 파킨슨병, 뇌전증 등)은 치료제 개발 난이도가 높고 후보 물질 탐색에 막대한 비용과 시간이 소요된다. 본 프로젝트는 컴퓨팅 기반 선별(in silico screening)로 탐색 공간을 축소하여, 약물 후보 리간드를 빠르게 좁히고 구조 기반 근거를 확보하는 것을 목표로 한다.

1.2 목표

- 뇌질환 1 종과 이에 대한 표적 단백질 1 종을 선정하고(선정 근거 포함), 표적의 3D 구조를 확보한다(PDB/예측 구조).
- 약 10,000 개 규모 리간드 라이브러리를 준비한 뒤, Lipinski Rule of Five 등 물성 규칙으로 1 차 필터링하여 후보를 약 2,000 개로 축소한다.
- 학습 데이터셋(약 5,000–10,000 개)을 구축하여 결합/활성 예측 머신러닝 모델을 학습하고, 후보 리간드를 스코어링하여 상위 Top 10 을 선정한다.
- 구조 기반 결합 예측(도킹 또는 동등한 결합 예측 도구: Boltz-2 등)으로 Top 10 후보를 검증하고, 단백질-리간드 복합체를 3D 로 시각화한다.
- 필요 시 생성 모델(예: Boltzgen)을 통해 신규 리간드를 생성하고 동일 파이프라인으로 평가한다.

1.3 범위 및 제외 범위

본 제안서는 계산 기반 후보 발굴 및 검증(모델링/도킹/시각화)까지를 범위로 하며, 실제 합성 및 in vitro/in vivo 실험은 수행하지 않는다. 단, 공개 데이터 및 문헌 기반 근거(기존 약물/리간드 정보) 수집은 포함한다.

2. 대상(뇌질환·표적 단백질) 선정 계획

스케치 기준으로 프로젝트의 첫 단계는 '뇌질환 종류 정하기'이며, 질환과 표적 단백질은 아래 기준으로 선정한다.

2.1 선정 기준

기준	설명
과학적 타당성	질환 병태생리와 표적 단백질의 연관성이 문헌으로 뒷받침될 것.

데이터 접근성	단백질 구조(PDB 또는 예측 구조)와 리간드/활성 데이터(ChEMBL 등) 접근이 가능할 것.
계산 가능성	단백질 크기·구조, 결합부위 정보가 도킹/스코어링에 적합할 것.
프로젝트 적합성	학부 수준에서 구현 가능한 범위(시간/자원) 내에서 수행 가능할 것.

2.2 후보 예시(선택지)

아래는 예시이며, 실제 선정은 팀/지도교수의 요구사항과 데이터 접근성 검토 후 확정한다.

- 알츠하이머병: AChE, BACE1, Tau 관련 효소/단백질 등
- 파킨슨병: MAO-B, LRRK2 등
- 뇌전증/신경흥분: GABA 수용체 관련 표적 등

3. 데이터 및 자원 계획

3.1 리간드 라이브러리(약 10,000 개)

공개 화합물 데이터베이스에서 SMILES/SDF 형태로 수집한다. 중복 제거, 이성질체 정리, 표준화(tautomer/염 형태) 과정을 포함한다.

3.2 학습 데이터셋(약 5,000–10,000 개)

머신러닝 모델 학습을 위해 표적 단백질에 대해 결합/활성 레이블이 존재하는 화합물을 수집한다. 활성 지표(IC50/Ki/EC50 등)를 공통 스케일로 정규화하고, 임계값 기반 이진 분류 또는 회귀 문제로 정의한다.

3.3 단백질 구조

PDB 에 공개된 구조가 존재하면 이를 우선 사용하며, 없을 경우 예측 구조(예: AlphaFold 등)를 활용한다. 결합부위는 공결정 리간드, 예측 포켓 탐색, 또는 문헌 기반 정보를 사용한다.

3.4 개발 환경

- 언어/라이브러리: Python 3.x, RDKit(화학 표현/지문), scikit-learn(기본 ML), PyTorch(딥러닝 선택), pandas/numpy
- 실험 관리: GitHub 리포지토리, requirements.txt/conda 환경, 실험 로그(MLflow 또는 간단한 CSV)
- 도킹/구조 도구: Boltz-2(또는 AutoDock Vina 등 동등 도구), PyMOL/ChimeraX(3D 시각화)

4. 방법론 및 파이프라인 설계

스케치의 핵심 흐름은 '단백질 + 리간드' 조합을 대상으로, 물성 규칙으로 1 차 축소(10,000→2,000) 후 머신러닝으로 추가 선별하여 Top 10 을 뽑고, 도킹/3D 로 검증하는 것이다.

4.1 물성(물리·화학) 규칙 기반 1 차 필터링

Lipinski Rule of Five 를 중심으로 경구(oral) 약물 가능성을 평가한다. 스케치에는 '입으로 먹는 약 규칙 4 가지'로 표현되어 있으며, 실제 적용 시에는 아래 지표를 기본으로 한다.

- 분자량(MW) <= 500
- logP <= 5
- 수소 결합 공여체(HBD) <= 5
- 수소 결합 수용체(HBA) <= 10
- (선택) 회전 가능한 결합 수, TPSA 등 추가 규칙(Veber) 적용 가능
- (선택) PAINS/반응성 경보, 독성 경보 등 추가 필터

필터링 목표는 약 10,000 개 라이브러리를 약 2,000 개 수준으로 축소하는 것이며(스케치의 10,000→2,000), 이 결과를 '테스트 분자'(후속 모델/도킹 입력)로 사용한다.

4.2 특징 표현 및 데이터 전처리

화합물은 다음 중 하나 이상의 방식으로 수치화한다: (1) Morgan/ECFP 지문, (2) 물성/구조 descriptor, (3) 그래프(GNN 입력). 학습 데이터는 중복·누락을 정리하고, 스플릿(Train/Valid/Test) 시 scaffold split 등 누설 방지 전략을 적용한다.

4.3 머신러닝 모델 설계 및 학습

베이스라인부터 시작해 성능을 비교한다. 데이터 규모(5,000–10,000 개)에 따라 아래 조합을 권장한다.

- 베이스라인: Logistic Regression, Random Forest, XGBoost/LightGBM

- 딥러닝(선택): MLP(지문 입력), Graph Neural Network(GCN/GAT), 분자 Transformer(사전학습 모델 활용 가능)

학습 후 후보(2,000 개)에 대해 예측 스코어를 계산하고, 상위 후보를 랭킹한다. 최종적으로 스케치와 같이 '줄어드는 리간드 분자 10 개 선정'을 산출한다.

4.4 구조 기반 결합 예측(도킹) 및 3D 시각화

선정된 Top 10 후보에 대해 단백질 결합부위에 대한 도킹을 수행하여 결합 포즈와 점수를 산출한다. 스케치의 예시처럼 'Boltz-2: 있는거 붙이기'는 기존 후보를 도킹하여 결합 가능성을 확인하는 단계로 해석한다.

도킹 결과는 (1) 결합 포즈의 타당성(수소결합/소수성 상호작용), (2) 도킹 점수, (3) 결합부위 잔기와의 상호작용 요약으로 정리하고, PyMOL/ChimeraX 등을 이용해 복합체를 3D 로 시각화한다(스케치의 '3D 로 가동').

4.5 생성 모델 기반 신규 리간드 생성(선택)

스케치의 'Boltzgen: 리간드 만들기' 단계는 신규 리간드 생성 모델을 통해 후보 공간을 확장하는 선택 기능이다. 생성된 후보도 동일한 물성 필터링→ML 스코어링→도킹 검증을 반복하여 유망 후보를 추가 확보한다.

5. 성능 평가 및 검증 계획

5.1 모델 성능 지표

- 분류 문제: ROC-AUC, PR-AUC, F1, 정밀도/재현율(Top-k 후보 선택에 중요)
- 회귀 문제: RMSE/MAE, Pearson/Spearman 상관
- 리간드 랭킹 품질: Enrichment Factor(EF), BEDROC 등(가능 시)

5.2 검증 전략

- 데이터 분할: 무작위 split 과 더불어 scaffold split 을 검토하여 일반화 성능을 평가
- 상위 후보 검증: 도킹 점수/포즈 기반 재랭킹, 기존 알려진 리간드와의 비교
- 재현성: 랜덤 시드 고정, 실험 설정/버전 관리, 결과 로그 저장

6. 일정 및 마일스톤

아래 일정은 8 주 기준 예시이며, 학사 일정에 맞춰 조정한다.

기간	주요 작업	산출물
1 주차	뇌질환/표적 단백질 확정, 구조 확보 및 결합부위 정의	질환·표적 선정 문서, 구조 파일 준비
2 주차	리간드 라이브러리(10,000) 수집/표준화, 물성 계산 파이프라인 구축	정제된 라이브러리, 필터링 스크립트
3 주차	물성 규칙 기반 1 차 필터링(10,000→~2,000) 및 품질 점검	후보 2,000 리스트, 통계 리포트
4 주차	학습 데이터(5,000–10,000) 수집/정제, 특징 생성(지문/descriptor)	학습용 데이터셋 v1
5 주차	베이스라인 모델 학습/평가, 모델 선택 및 튜닝	모델 성능표, 최종 모델 후보
6 주차	후보 2,000 스코어링 및 Top 10 선정	Top 10 후보 리스트/근거
7 주차	Top 10 도킹 및 상호작용 분석, 3D 시각화	도킹 결과, 3D 이미지/파일
8 주차	(선택) 생성 모델로 신규 리간드 생성 및 추가 검증, 최종 보고서 작성	최종 산출물 패키지

7. 산출물 정의

- 코드 리포지토리: 데이터 수집/정제, 물성 계산 및 필터링, 학습/평가, 후보 랭킹, 도킹 자동화 스크립트

- 데이터: (1) 정제된 리간드 라이브러리, (2) 학습 데이터셋, (3) 후보/결과 CSV
- 모델: 학습된 모델 파일 및 재현 가능한 학습 설정
- 결과: Top 10 후보 리스트(물성/예측/도킹 요약), 3D 복합체 시각화 결과
- 문서: 프로젝트 제안서(본 문서), 최종 결과 보고서, 발표 자료(선택)

8. 리스크 및 대응

리스크	대응 방안
활성 데이터 부족/품질 문제	표적을 데이터가 풍부한 후보로 재선정, 유사 표적 데이터로 보강, 라벨 기준을 명확화
모델 과적합/일반화 실패	scaffold split, 단순 모델 베이스라인 유지, 특징/하이퍼파라미터 튜닝 제한
도킹 결과 신뢰도	도킹은 정성적 근거로 사용하고, 기존 알려진 리간드 재도킹으로 도구 검증
연산 자원 제한	모델 복잡도 단계적 증가, 배치 처리, 후보 수 제한(Top-k), 필요 시 클라우드/학교 자원 활용

9. 기대효과

본 프로젝트는 물성 규칙(약물유사성)과 데이터 기반 예측(ML), 구조 기반 근거(도킹/3D)를 연결한 종단(end-to-end) 파이프라인을 구현한다. 이를 통해 (1) 후보 탐색 공간 축소, (2) 후보 선정의 정량적 근거 제시, (3) 재현 가능한 연구 코드/데이터 구축이라는 학습·연구 성과를 기대할 수 있다.

참고자료

- Lipinski CA et al. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.
- RDKit: Open-source cheminformatics.
- Protein Data Bank (PDB).
- ChEMBL: a large-scale bioactivity database.

부록 A. 초기 아이디어 스케치