

□ 연구 개요

○ 배경

- 약물 개발 과정에서 안전성과 효능을 사전에 예측하기 위한 ADMET 데이터는 신약 개발 분야의 핵심 데이터로써, 이들 데이터에서 유의미한 지식을 추출 및 해석의 중요성 증가[1-3]

○ 관련 연구

- Zhang et al은 분자를 SMILES·그래프·지문 등에서 통합하여, 단일모달보다 더 정확한 분자 특성 예측을 가능하게 하는 멀티뷰 분자 표현 학습 모델 제시.[4].
- Wang et al 은 SMILE 시퀀스, 분자그래프, 3D 기하구조를 통합한 멀티모달 방식 기반의 데이터를 활용한 ADMET 예측 모델 제시 [5].

○ 기존 연구의 한계점 및 개선 사항

- 기존 연구들은 데이터 모달의 통합 하기 위한 연구가 대부분 진행되었을뿐, 데이터 컬럼 자체에서 가지고 있는 멀티 분포에 대한 고려가 미흡

○ 연구 목표

- ADMET 컬럼이 지니는 다중 분포성과 실험적 변동성을 반영한 분포 기반(multi distribution-aware) 의미 추출 및 예측 모델을 설계하는 것을 목표

□ 주요 설계

○ 수도 코드 (Pesudo code)

Input : $X_{new} \in R^{n \times p}$ # 새 샘플 테이블

$\Theta_{MM} = (G, D, C, ENC)$

Output : $Z \in R^{n \times d}$ # 단일 연속형 representation

Procedure :

$Z = []$

가우시안 연속형 \rightarrow Direct embedding

for each f in G :

$(\mu, \sigma) = G[f]$

$Z.append((X_{new}[f] - \mu) / \sigma)$

비가우시안 연속형 \rightarrow 이산화

for each f in D :

$X_{new}[f+ "_disc"] = DISCRETIZE(X_{new}[f], D[f])$

명목형(원래 + 이산화) \rightarrow 임베딩

for each f in C :

$Z.append(APPLY_ENCODER(ENC[f], X_{new}[f]))$

return CONCAT(Z)

□ 2단계 개발 일정

구분	내용	1	2	3	4	5	6	7	8	9	10	11	12
1	Multi-modal Converter 구현(Toxicity)												
2	Multi-modal Converter 소스코드 검증												
3	Multi-modal Converter 생산												
4	Multi-modal Converter 공개												

□ 참고문헌

1. Integrating artificial intelligence into small-molecule drug discovery. Nat. Rev. Drug Discov. 24, 123-145 (2025).

2. Bridging data and drug development: machine learning advances in ADMET prediction. Drug Discov. Today 30, 104487 (2025).

3. Artificial intelligence in small-molecule drug discovery: a critical review of methods, applications, and real-world outcomes. Pharmaceutics 18, 1271 (2025)

4. MvMRL: a multi-view molecular representation learning method for molecular property prediction. Brief. Bioinform. 25, bbae298 (2024).

5. Multi-Modal Representation Learning for Molecular Property Prediction: Sequence, Graph, Geometry. arXiv (2024).