

약물의 구조 및 LLM 기반 임베딩을 통한 멀티모달 딥러닝 모델 기반 약물들 간의 상호작용 예측 모델

임송¹, 고윤희²

¹한국의외국어대학교 바이오메디컬공학부 학부생

²한국의외국어대학교 바이오메디컬공학부 교수

song.im@hufs.ac.kr, younko@hufs.ac.kr

Multi-Modal Deep Learning Model for Drug-Drug Interaction Prediction Using Structural and LLM-Based Embeddings

Song Im¹, Younhee Ko¹

¹Dept. of Biomedical Engineering, Hankuk University of Foreign Studies

요 약

약물 간 상호작용(Drug-Drug Interaction, DDI)은 병용 약물 사용 시 발생할 수 있는 예기치 못한 약리학적 반응으로, 환자의 안전과 치료 효과에 중대한 영향을 미친다. 본 연구에서는 약물 설명 텍스트, 화학 구조, 생물학적 네트워크를 통합한 멀티모달 모델을 통한 약물 표현을 기반으로 79 가지 DDI 유형을 예측하는 모델을 제안한다.

다양한 이질 데이터를 통합한 멀티모달 딥러닝 모델을 구성하여 다중 분류 모델을 학습했고, 기존 모델 대비 예측 모델의 성능 향상을 확인했다. 제안된 방법은 약물의 의미적, 화학적, 생물학적 특징을 함께 반영함으로써 잠재적 DDI의 사전 예측 가능성을 향상시키며, 새로운 약물 조합에 대한 DDI 탐색에도 기여할 수 있다.

1. 서론

의약품의 병용 처방이 증가함에 따라, DDI에 대한 정확한 예측은 환자의 안전성과 치료 효과 확보에 있어 중요한 과제이다. 특히, 임상에서 예기치 못한 DDI로 인해 심각한 문제가 발생하기에, 이에 대한 연구는 임상분야에 있어서도 매우 중요한 문제다. 기존에는 실제 환자에게 처방 후 나타나는 부작용이 보고되는 방식의 사후 대응이 주를 이루었지만, 최근 기계학습 기반의 사전 예측 모델을 통해 DDI를 미리 예측하고 이를 통해, 실제 처방에 활용하고자 하는 시도가 활발하다. 기존에는 약물 구조나 단백질 정보 등 단일 modality에 기반한 예측이 많았지만, 이는 약물의 의미적 특성과 작용 기전을 충분히 반영하지 못하는 한계가 있다.

본 연구에서는 약물 설명 텍스트로부터 깊이 있는 의미 정보를 추출하기 위해 일반 BERT 모델을 PubMed와 PMC의 생의학 논문 데이터로 사전학습하여 도메인 특화된 언어모델 BioBERT를 활용했다. BioBERT

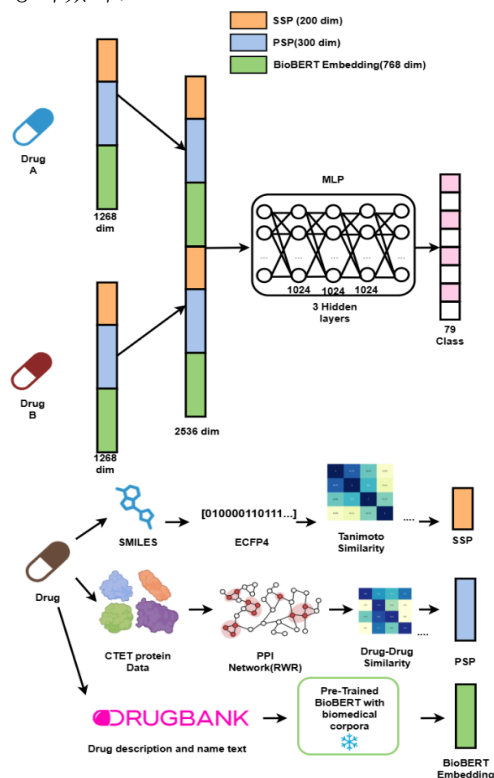
는 단순 임베딩보다 관련 문헌에 표현된 약물의 기능 및 특성등을 효과적으로 임베딩 함으로써 기존의 약물들의 화학적 구조나 타겟 유전자들만으로 표현할 수 없는 특징들을 효과적으로 표현해 낼 수 있다. 또한, 화학적 구조 기반의 tanimoto 유사도(Structural Similarity Profile, SSP)와 PPI 네트워크 기반 유사도(Protein Similarity Profile, PSP)[1]를 함께 고려하여, 이질적 데이터의 통합을 통해 79개의 알려진 DDI를 예측해내는 딥러닝 모델을 제안했고, 그 성능을 평가하여 실제 LLM 모델을 통한 임베딩이 DDI 예측을 향상시킴을 확인했다.

2. 방법

본 연구에서 사용된 DDI 데이터는 DrugBank[2]에서 추출된 86 개 중, 샘플 수가 적은 7 개를 제외한 총 79 개를 사용하였다. 각 DDI type은 Drug A, B 간 상호작용의 구체적 기전이나 결과를 문장 형태로 기술하며, 단순한 이진 상호작용 유무가 아닌 작용 기전과 방향성까지 반영된 정성적 라벨이다.

약물의 의미적 특성을 반영하기 위해, DrugBank의 약물 설명에 약물명을 결합한 텍스트 총 9,377개 약물에 대해 텍스트를 수집한 뒤, 특수문자 제거 등의 전처리를 수행했다. 이를 BioBERT 모델에 입력해 약물 단위 임베딩 벡터를 생성했고, 입력 크기가 512 토큰을 초과하는 경우는 제외했다. 최종적으로 9,294개의 약물에 대한 (1, 768)차원의 임베딩 벡터를 확보했다.

약물 간 화학 구조 유사도를 계산하기 위해, 약물별 SMILES 문자열을 기반으로 RDKit을 사용해 ECFP4를 생성했다. SMILES 정보는 주로 DrugBank에서 수집했고, 누락된 약물은 PubChem을 통해 보완했다. 총 13,082개 약물을 이용해 모든 약물 쌍에 대해 tanimoto 유사도를 계산했다. 본 연구에서는 이를 PCA를 통해 200차원으로 축소하여 모델 입력으로 활용하였다.

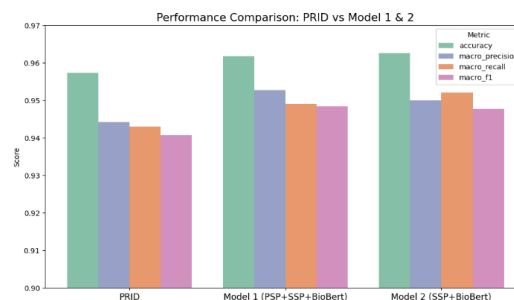


(그림 1) 전체적 모델 구성도.

PSP는 PRID 연구에서 제안한 방식과 동일하게 구성했다. 약물과 연관된 CTET 단백질(carriers, transporters, enzymes, targets)을 STRING Database 기반 PPI 네트워크 상에 매핑하고, 이를 시드 노드로 설정하여 Random Walk with Restart(RWR) 알고리즘을 [3] 수행했다. 각 약물에 대해 계산된 RWR 전과 결과를 바탕으로 PSP 벡터를 생성해, 간접적인 생물학적 연관성을 반영한 표현으로 활용했다. 이를 PCA를 통해 300차원으로 축소하여 입력으로 사용했다. 세 가지 modality로부터 얻어진 벡터는 약물 단위로 결합되며, DDI 예측을 위해 3 개의 은닉층과 ReLU, Dropout을 포함한 MLP 기반 다중 분류 모델이 최종

적으로 79 개의 DDI class에 대해 softmax 출력 값을 생성한다. 여러 modality의 조합으로 학습한 모델의 성능을 비교했다.

3. 결과



(그림 2) PRID 모델과 2 가지 조합의 모델들에 대한 지표 비교 그래프

두 가지 조합의 모델들은 PRID 모델 대비 정확도를 포함한 지표들에서 좋은 결과를 보였다. 이는 약물 간 의미적 정보를 반영한 표현이 모델의 예측 성능을 높이는 데 효과적임을 시사한다.

4. 결론 및 향후 연구 방향

본 연구에서는 DDI 예측을 위해 사전학습된 모델 BioBERT로 약물들의 관련 논문 기반 임베딩정보를 얻었고, 이를 약물의 화학적 구조 유사도와 약물의 CTET 단백질간의 간접적 관계를 PPI 모델에 기반한 유사도를 함께 사용했다. 사전학습된 모델을 통한 약물의 특성을 반영함으로써 실제 DDI 예측 능력이 향상됨을 확인했다. 추후 DDI의 형태가 기술된 문장을 순수 PubMed 문헌만으로 사전학습된 언어모델인 PubMedBERT를 추가 학습(fine-tuning)하는 방안을 고려하고 있다. 이를 통해 도메인 특화 임베딩 표현을 확보하고, 보다 정밀한 약물 간 의미 정보 추출이 가능할 것으로 기대된다.

5. 사사

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2023R1A2C1007756).

참고문헌

- [1] Seo, J., Jung, H., Ko, Y. PRID: Prediction Model Using RWR for Interactions between Drugs. *Pharmaceutics*, 15(10), 2469, 2023.
- [2] Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018, 46, D1074–D1082.
- [3] Valdeolivas, A.; Tichit, L.; Navarro, C.; Perrin, S.; Odelin, G.; Levy, N.; Cau, P.; Remy, E.; Baudot, A. Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics* 2019, 35, 497–505.