ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 52, No. 7, pp. 601-610, 2025. 7
https://doi.org/10.5626/JOK.2025.52.7.601

신약 후보 물질의 ADMET 속성 예측을 위한 사전학습 모델 기반의 일반화 성능 향상 기법

(A Pretrained Model-Based Approach to Improve Generalization Performance for ADMET Prediction of Drug Candidates)

김 윤 주 †　　박 상 현 ††
(Yoonju Kim)　(Sanghyun Park)

요 약 신약 개발 과정에서 ADMET(흡수, 분포, 대사, 배설, 독성) 속성의 정확한 예측은 임상 시험 실패율을 낮추고 개발 비용을 절감하는 데 중요한 역할을 한다. 본 연구에서는 그래프 트랜스포머 기반의 분자 임베딩과 사전 학습된 UniMol 모델 기반의 임베딩을 결합하여 신약 후보 물질의 ADMET 예측 성능을 높이는 방법을 제안한다. 제안된 모델은 분자의 그래프 구조에서 결합 유형 정보를 반영하여 보다 화학적으로 정교한 표현을 생성하며, UniMol의 사전 학습된 3D 임베딩을 활용하여 분자의 공간적 특성을 효과적으로 학습한다. 이를 통해 데이터 부족 문제를 보완하고, 모델의 일반화 성능을 향상시킬 수 있도록 설계하였다. 본 연구에서는 총 10개의 ADMET 속성을 대상으로 예측 실험을 수행하였다. 실험 결과, 제안된 모델은 기존 방법들보다 우수한 예측 성능을 보였으며, 원자의 결합 정보와 3D 구조를 효과적으로 통합함으로써 ADMET 속성 예측의 정확도를 향상시킬 수 있음을 확인하였다.

키워드: ADMET 예측, 사전 학습 모델, 그래프 트랜스포머, 분자 임베딩, 신약 개발

Abstract Accurate prediction of Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) properties plays an important role in reducing clinical trial failure rates and lowering drug development costs. In this study, we propose a novel method to improve ADMET prediction performance for drug candidate compounds by integrating molecular embeddings from a graph transformer model with pretrained embeddings from a UniMol model. The proposed model can capture bond type information from molecular graph structures, generating chemically refined representations, while leveraging UniMol's pretrained 3D embeddings to effectively learn spatial molecular characteristics. Through this, the model is designed to address the problem of data scarcity and enhance the generalization performance. In this study, we conducted prediction experiments on 10 ADMET properties. The experiment results demonstrated that our proposed model outperformed existing methods and that the prediction accuracy for ADMET properties could be improved by effectively integrating atomic bond information and 3D structures.

Keywords: ADMET prediction, pre-trained model, graph transformer, molecular embedding, drug discovery

· 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 : 숙명여자대학교 소프트웨어융합전공 학생
yoonju49@gmail.com
†† 종신회원 : 연세대학교 컴퓨터과학과 교수(Yonsei Univ.)
sanghyun@yonsei.ac.kr
(Corresponding author임)

논문접수 : 2025년 2월 28일
(Received 28 February 2025)
논문수정 : 2025년 4월 30일
(Revised 30 April 2025)
심사완료 : 2025년 5월 6일
(Accepted 6 May 2025)

정보과학회논문지 제52권 제7호(2025. 7)

1. 서 론

신약 개발은 질병의 원인 단백질(Target)을 표적으로 삼아 효과적인 치료제를 찾고 개발하는 과정이다. 일반적으로 유효물질 발굴(Hit Identification), 선도물질 도출(Hit to Lead), 선도물질 최적화(Lead Optimization), 임상 시험(Clinical)의 단계를 거치며, 각 단계마다 신중한 평가와 최적화가 요구된다. 이 과정에는 막대한 시간과 비용이 투입됨에도 불구하고 최종적으로 임상 시험을 통과하는 신약의 비율은 극히 낮다. 임상 시험에서 실패하는 주요 원인 중 하나는 예상하지 못한 약물의 독성 때문인데, 이는 약동학적(Pharmacokinetic, PK) 특성에 대한 고려가 충분히 이루어지지 못했기 때문이다[1].

이러한 실패를 줄이기 위해 신약 후보 물질의 ADMET (흡수, 분포, 대사, 배설, 독성) 특성을 조기에 평가하는 것에 대한 필요성이 높아지고 있다. ADMET 특성은 약물이 체내에서 어떻게 작용하는지를 결정하는 핵심 요소이며, 신약의 효능과 안전성에 직접적인 영향을 미친다. 예를 들어, 흡수(Absorption) 단계에서 경구 투여된 약물이 장에서 얼마나 효과적으로 흡수되는지(HIA, OB), 분포(Distribution) 과정에서 혈장 단백질 결합율(PPB)이나 혈액-뇌 장벽(BBB) 투과성이 어느 정도인지, 대사 (Metabolism) 단계에서 간 효소(CYP450 계열)에 의해 어떻게 분해되는지, 배설(Excretion) 과정에서 반감기(Half-life)와 체내 제거율(Clearance)이 어떠한지를 고려해야 한다. 또한, 독성(Toxicity) 분석을 통해 간독성, 심독성, 호흡기 독성 등 예기치 않은 부작용 가능성을 사전에 평가하는 것이 중요하다.

머신러닝 및 딥러닝 기반의 In Silico 예측 모델은 기존의 실험적 방법보다 비용과 시간이 절감되며, 더 많은 후보 물질을 평가할 수 있는 장점이 있다. 이러한 접근 방식은 신약 개발의 초기 단계에서 불필요한 후보 물질을 배제하고, 보다 유망한 화합물에 집중할 수 있도록 하여 신약 개발 성공률을 높이는 데 기여할 수 있다.

기존 ADMET 예측 연구는 신약 개발 과정에서 유의미한 성과를 내고 있으나, 여전히 몇 가지 한계점이 존재한다. 가장 대표적인 문제는 데이터 부족 문제로 인한 일반화 성능 저하 문제이다. 현실적으로 신약 개발 과정에서 다양한 화합물의 ADMET 속성을 실험적으로 확보하는 것은 매우 어렵고 비용이 많이 든다. 따라서 머신러닝 모델을 훈련하는 데 필요한 데이터가 부족한 경우가 많으며, 이는 모델의 일반화 성능을 저하시킬 가능성이 크다.

또한, 기존 분자 표현 방식에는 한계가 존재한다. 전통적인 QSAR(Quantitative Structure - Activity Relationship) 모델은 분자의 구조 정보를 ECFPs, MACCS와 같은 고정된 1차원 벡터로 표현하며, 이는 분자의 복잡한 입체적 구조나 원자 간 상호작용을 충분히 반영하지 못한다. 그래프 신경망(Graph Neural Networks, GNNs) 기반 모델은 이러한 한계를 일부 극복했으나, 여전히 대부분의 그래프 신경망 모델은 분자를 단순히 노드(원자)와 엣지(결합)의 2D 그래프로만 표현하며, 결합의 유형이나 분자의 3차원 구조에 대한 정보는 고려하지 못하는 경우가 많다. 이로 인해 예측 모델의 표현력 한계가 존재하게 된다.

따라서 본 연구는 이러한 문제점을 극복하기 위해 분자 임베딩 기법의 고도화 전략을 제안하며, 이를 위해 다음과 같은 두 가지 방법을 적용하였다.

첫째, 결합 유형 정보를 포함한 분자 그래프 임베딩 기법을 도입하였다. 기존 그래프 신경망 기반 모델들은 원자 간 결합 유무만을 반영하는 단순한 그래프 구조를 사용하여, 분자의 실제 화학적 특성을 충분히 반영하지 못하는 한계가 있었다. 본 연구에서는 결합의 유형까지 반영한 분자 그래프를 구성하고, 이를 효과적으로 임베딩하기 위해 그래프 트랜스포머(Graph Transformer)[2] 구조를 활용하였다. 이를 통해 원자 간의 복잡한 상호작용을 더 정밀하게 표현할 수 있도록 하였다.

둘째, 분자의 3차원 입체 구조 정보를 반영하기 위해, 대규모 화합물 구조 데이터를 기반으로 사전 학습된 UniMol[3] 모델을 도입하였다. UniMol은 분자의 공간적 배치와 구조적 패턴을 학습하여 입체적인 화학 정보를 포착할 수 있도록 설계된 모델로, 본 연구에서는 이를 전이 학습(Transfer Learning) 방식으로 적용하여 데이터 부족 문제를 완화하고 예측 성능을 향상시켰다.

최종적으로, 위의 두 임베딩 결과를 결합하여 최종 분자 표현을 구성하였다. 이를 통해 각 기법의 장점을 통합하고, ADMET 속성 예측의 정밀도를 높이는 데 기여하고자 하였다. 이를 통해 본 연구는 기존의 분자 표현 방식이 갖는 한계를 극복하고, 신약 개발 초기 단계에서 보다 정확하고 신뢰성 있는 ADMET 속성 예측을 가능하게 하는 것을 목표로 한다.

2. 관련 연구

2.1 ADMET 예측

ADMET 특성 예측을 위한 전통적인 접근 방식은 QSAR 모델에 기반하며, 분자 지문(Molecular Fingerprints) 및 분자 서술자(Descriptors)를 벡터화하여 머신러닝 알고리즘에 입력하는 방식이 일반적이었다. MACCS Keys, ECFPs 등과 같은 지문 기반 표현은 예측 모델 개발의 초기 입력으로 활용되었으나, 분자의 입체적 구조나 복잡한 상호작용을 반영하는 데에는 한계가 있었다. 이를 보완하기 위해, 분자를 그래프로 표현하고 원자 및 결합

간 관계를 학습하는 그래프 신경망 기반의 예측 모델들이 제안되었다. MPNN, GCN, GAT, ChemProp[4] 등이 대표적이며, 분자의 구조적 정보 학습에 있어 일정 수준의 성능 향상을 보여주었다.

그래프 신경망 기반 모델의 발전에도 불구하고, 대부분의 모델은 여전히 결합의 유무만을 반영한 2D 그래프 구조에 의존하고 있으며, 결합 유형이나 분자의 3D 입체 구조 정보는 포함되지 않거나 부정확하게 반영되는 경우가 많다. 또한, 라벨이 포함된 ADMET 관련 데이터가 제한적이기 때문에, 소규모 데이터셋에서의 일반화 성능 확보 또한 주요 과제로 남아 있다.

본 연구는 이러한 기존 연구의 흐름을 계승하면서도, (1) 그래프 트랜스포머를 기반으로 한 결합 유형 정보의 명시적 반영, (2) 3D 구조 임베딩 기반 전이 학습을 통해 표현력과 일반화 성능을 동시에 강화하고자 한다.

2.2 전이 학습

전이 학습(Transfer Learning)은 소량의 데이터만으로도 좋은 성능을 낼 수 있도록, 사전 학습된 모델의 지식을 새로운 작업에 적용하는 기법이다. 신약 개발에서는 대규모 화합물 데이터셋을 활용한 사전 학습을 통해 일반적인 분자 표현을 학습한 후, ADMET 속성 예측과 같은 다운스트림 작업(Downstream-task)에서 미세조정(Fine-tuning)하는 방식이 많이 사용된다.

전이 학습을 활용한 대표적인 딥러닝 기반 분자 임베딩 모델로는 ChemBERTa[5], MolBERT[6], UniMol 등이 있으며, 이들은 트랜스포머 구조를 기반으로 대규모 화합물 데이터에서 분자 구조 정보를 학습한다. ChemBERTa와 MolBERT는 BERT를 기반으로 SMILES 표현을 입력으로 받아 자연어 처리(NLP) 방식으로 분자 표현을 학습하는 방식이며, UniMol은 대규모 비지도 학습 데이터를 활용하여 사전 학습된 트랜스포머 모델로, 3D 분자 구조 정보를 활용함으로써 기존의 2D 표현 방식이 간과했던 입체적 특성을 학습할 수 있다. 이는 신약 개발에서 분자의 실제 작용 기전을 더욱 정밀하게 분석하는 데 도움을 줄 수 있다.

본 연구에서는 전이 학습의 장점을 활용하여, 사전 학습된 UniMol 모델을 통해 분자의 3D 구조 정보를 효과적으로 학습함으로써, 기존 모델이 고려하지 못했던 입체적 분자 구조와의 결합 특성을 보다 정교하게 반영할 수 있도록 한다. 이를 그래프 트랜스포머 기반의 분자 임베딩과 결합하여 실제 분자 구조를 보다 잘 반영하는 분자 표현을 생성하고자 한다.

3. 제안 방법

3.1 UniMol Embedding

UniMol 모델에 대한 학습을 위해, 우선 분자를 원자 단위로 나누어 각 원자 별 3차원 좌표와 원자 유형(atom type)을 추출한다. 이후, 두 입력 정보는 그림 1 (A)와 같이 각각의 인코더를 통해 임베딩되며, 이렇게 생성된 원자 표현(Atom Representation)과 쌍 표현(Pair Representation)은 UniMol의 백본 트랜스포머 아키텍처에 입력된다. 이후, 분자의 3D 공간적 관계를 효과적으로 반영하기 위해서 어텐션 메커니즘에 Pair Representation d_{ij}^{l}을 추가한다. 기존의 scaled dot-product attention은 Query 벡터와 Key 벡터 간의 내적을 기반으로 어텐션 스코어를 계산하지만, UniMol에서는 bias 항을 통해 3D 구조 정보를 추가적으로 반영한다. 이에 따라, 트랜스포머 레이어 l에서의 노드 i와 노드 j 간의 어텐션 연산은 수식 (1)과 같이 정의된다.

$$Attention(Q_{i,}^{l} K_{j,}^{l} V_{j,}^{l}) = Softmax\left(\frac{Q_{i}^{l}(K_{j}^{l})^{T}}{\sqrt{d}} + d_{ij}^{l}\right) \quad (1)$$

Pair Representation d_{ij}^{l}의 초기값은 원자 간의 거리 값(pairwise atomic distance)으로 설정된다. 이를 통해 SE(3)-불변성(SE(3)-invariance)을 유지하면서, 3D 공간적 구조를 학습할 수 있도록 설계하였다. SE(3)-불변성은 회전(Rotation), 이동(Translation), 반사(Reflection)과 같은 변환이 적용되더라도 동일한 상대적 거리를 유지하는 성질을 의미한다. 이를 통해, 모델은 원자 간의 상대적인 거리 정보를 학습하여, 분자의 위치나 방향이 달라져도 일관된 표현을 유지할 수 있도록 한다.

또한, Pair Representation d_{ij}^{l}은 각 트랜스포머 레이어를 거치면서 점진적으로 업데이트되며, 이를 통해 모델이 원자 간의 장거리 의존성과 복잡한 상호작용을 효과적으로 학습할 수 있도록 도와준다.

뿐만 아니라, 모델의 강건성을 보장해주기 위해 원자 좌표에 랜덤 노이즈를 추가한 후, 이를 원래 상태로 복원하도록 학습이 진행된다. 이를 위해 전체 원자 좌표의 15%에 대해 [-1,1] 범위에서 균등 분포(uniform distribution)로 샘플링된 랜덤 노이즈를 추가하였다. 복원 학습 과정에서는 다음의 두 가지 접근법을 활용하였다.

첫 번째로, Pair-Distance Prediction Head는 원자 간 상대적 거리 관계를 예측하는 방식으로 학습된다. 즉, 분자가 회전하거나 이동하더라도 원자 간 상대적인 거리 값은 변하지 않으므로, 모델이 이러한 특징을 학습함으로써 공간적인 불변성을 확보할 수 있다. 두 번째로, SE(3)-Equivariant Head는 노이즈가 추가된 원자의 위치를 원래 상태로 복원하는 학습을 수행하며, 이는 SE(3)-Equivariance 관점에서 분자의 공간적 구조를 학습하게 된다. SE(3)-Equivariance는 모델이 SE(3) 변환을 받았을 때, 동일한 변환을 반영한 출력을 생성하는 성질을 의미한다. 이 과정을 통해 모델이 원자의 개

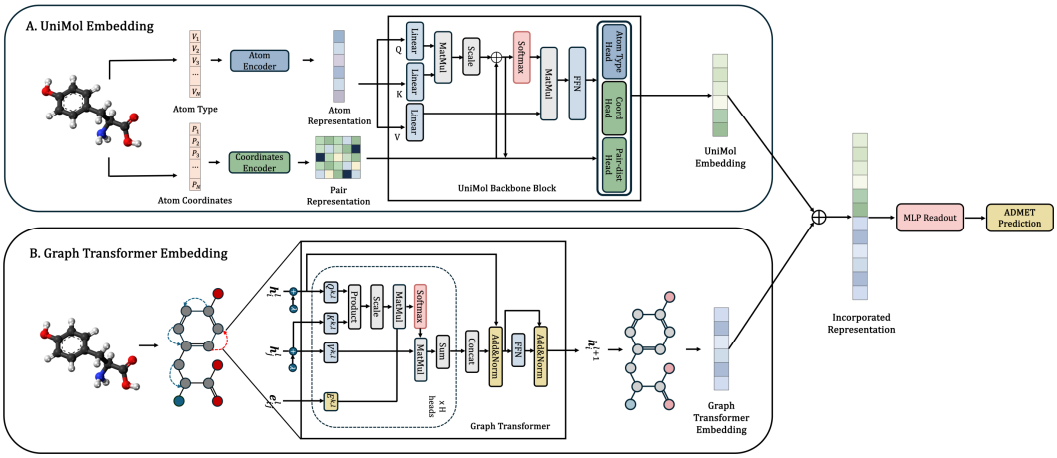


그림 1 전체 모델 구조

Fig. 1 Overview of Proposed Method

별적인 좌표를 직접 복원함으로써 보다 정교한 3D 공간적 특성을 학습할 수 있도록 하였다. 이러한 훈련 과정을 통해 얻어진 사전 학습 모델[3]을 활용하여 훈련 데이터에 대한 UniMol 임베딩 벡터를 생성할 수 있다.

3.2 분자 모델링

본 연구에서는 분자의 구조적 특징을 더욱 정밀하게 반영하기 위해 결합 유형 정보를 포함한 그래프 기반 분자 임베딩 기법을 적용하였다. 기존의 그래프 신경망 기반 분자 모델링은 원자 간 결합의 존재 여부만 고려하여 결합의 물리화학적 차이를 충분히 반영하지 못하는 한계가 있었다. 그러나 실제 분자 구조에서는 단일, 이중, 삼중 결합에 따라 화학적 특성이 크게 달라진다. 따라서 본 연구에서는 분자 그래프에 결합 유형 정보를 포함시켜 보다 정밀한 분자 표현을 학습하도록 설계하였다.

이를 위해, RDKit을 활용하여 분자의 SMILES 코드로부터 3D 구조(Conformer)를 생성하였으며, ETKDG (Experimental-Torsion Knowledge Distance Geometry) 알고리즘을 적용하였다. ETKDG 알고리즘은 실험적으로 관찰된 결합각 분포를 반영하여 보다 현실적인 분자 구조를 만들며, 거리 기하학(distance geometry) 방법을 이용해 저에너지 구조를 예측하는 방식으로 동작한다. 이렇게 생성된 3D Conformer를 기반으로 원자 간 결합 정보를 그래프 엣지의 속성으로 포함하여 분자 임베딩을 수행하였다. 각 원자는 해당 원자 번호를 기반으로 벡터화되며, 결합 정보는 단순한 연결 정보뿐만 아니라 결합의 종류를 나타내는 원-핫(One-Hot) 벡터 형태로 저장된다. 이를 통해 기존의 원자 간 단순 연결 정보를 사용하는 방식보다 더 정밀한 분자 표현이 가능해졌다.

3.3 그래프 트랜스포머

신약 후보 물질의 ADMET 특성을 예측하기 위한 모델 구조 중, 그림 1 (B)는 그래프 트랜스포머를 통해 결합 유형 정보를 반영한 분자 임베딩을 생성하는 과정을 보여준다. 그래프 트랜스포머는 어텐션 메커니즘을 통해 각 원자는 다른 모든 원자와의 관계를 학습하며, 특정 원자가 분자의 물리화학적 성질에서 중요한 역할을 한다면 해당 원자와의 연결에 높은 가중치를 부여하여 정보가 집중적으로 전달될 수 있도록 한다. 이를 통해 모델은 단순히 인접한 원자들과의 관계만 학습하는 것이 아니라, 멀리 떨어져 있지만 분자의 기능적 역할에서 중요한 원자 간의 관계를 강화할 수 있다. 예를 들어, 특정 약리학적 성질을 결정하는 기능기(Functional Group)들이 분자 내에서 떨어져 있는 경우, 그래프 트랜스포머는 이들 간의 관계를 인식하고, 관련 정보가 효율적으로 전달될 수 있도록 가중치를 조정한다.

또한, 본 연구에서는 단순히 노드 간의 연결 여부만 고려하는 기존 방식과 달리, [3.2 분자 모델링]에서 얻어진 결합 종류에 대한 정보를 명시적으로 반영하였다. 또한, 분자 내 원자 간의 결합 정보가 중요한 역할을 한다는 점을 반영하기 위해 단순한 메시지 전달 방식에서 벗어나 결합 특성을 모델링할 수 있는 어텐션 스코어 조정 기법을 도입하였다. 이를 위해 수식 (2)와 같이, 첫 번째 레이어에서의 노드 i와 노드 j에 대한 특징 벡터 h_i^l, h_j^l 를 각각 k번째 head의 Query 행렬 $Q^{k,l}$와 Key 행렬 $K^{k,l}$을 통해 변환하고, 이들의 내적을 이용해 초기 어텐션 스코어를 계산하였다. 이후, 노드 간 결합 유형 정보를 포함하는 엣지 특징 $e_{i,j}^l$에 가중 행렬 $E^{k,l}$을 적

용하여 결합 정보를 반영한 최종 어텐션 스코어 $w_{ij}^{k,l}$를 도출하였다. 정규화된 어텐션 스코어는 이웃 노드들의 정보를 가중합할 때 사용되며, 이를 통해 업데이트된 노드 표현 h_i^{l+1}을 얻는다.

$$h_i^{l+1} = O_h^l \|_{k=1}^{H} \left(\sum_j w_{ij}^{k,l} V^{k,l} h_j^l \right)$$

$$, where\ w_{ij}^{k,l} = Softmax\left(\left(\frac{Q^{k,l} h_i^l \cdot K^{k,l} h_i^l}{\sqrt{d_k}} \right) \cdot E^{k,l} e_{ij}^l \right) \quad (2)$$

본 논문에서 제안하는 ADMET 예측 모델은 다음과 같은 단계로 구성된다. 먼저, 분자의 그래프 구조를 입력으로 받아 그래프 트랜스포머 레이어를 통과하면서 원자 간의 관계뿐만 아니라 결합 유형 정보를 함께 고려하여 분자 임베딩을 생성한다. 이를 통해, 기존 방식에서 고려되지 못했던 화학적 결합 정보를 모델 학습 과정에 효과적으로 반영할 수 있도록 하였다. 이후, 앞서 전이 학습을 통해 얻어진 UniMol 임베딩과 결합하여 생성된 벡터를 다층 퍼세트론 (MLP) 기반의 예측 모델에 입력하여, ADMET 속성을 예측하는 구조를 갖는다.

4. 실험 및 결과

4.1 데이터셋 및 평가지표

ADMET 속성 예측을 위한 모델을 설계하였으며, 총 10개의 속성을 대상으로 실험을 진행하였다. 이 중 8개는 분류(Classification) task, 2개는 회귀(Regression) task로 구성되어 있으며, 각 Endpoint는 서로 다른 논문 및 데이터 소스에서 수집된, 정답 레이블이 포함된 데이터를 기반으로 실험되었다. 이에 따라, 각 Endpoint별로 사용된 약물 유사 화합물의 구성은 상이하다.

데이터셋은 총 8개의 선행 연구로부터 수집되었으며, 2개의 흡수(Absorption)[7,8], 1개의 분포(Distribution)[9,10], 3개의 대사(Metabolism)[11], 1개의 배설(Excretion)[12], 2개의 독성(Toxicity)[13], 그리고 1개의 ADMET 관련 물리화학적 특성[13]을 포함한다. 각 Endpoint task에서 사용된 데이터의 구체적인 크기는 표 1의 Dataset 항목에 제시되어 있다.

본 연구에서는 제안한 모델을 포함한 모든 비교 모델에 대해, 10개의 tasks에 걸쳐 동일한 데이터 전처리, 학습 방식, 평가 파이프라인을 적용하여 실험의 일관성과 성능 비교의 공정성을 확보하였다. 각 모델은 동일한 하이퍼파라미터(batch size=64, learning rate=0.0005, optimizer=AdamW)를 사용하였으며, 동일한 데이터 분할 및 랜덤 시드 설정(torch.manual_seed, numpy.random.seed)을 통해 실험 간 편차를 최소화하였다. 전체 데이터셋은 8:1:1 비율로 학습(Train), 검증(Validation), 테스트(Test) 세 부분으로 분할하였으며, 동일한 분할 결과를 모든 task에 일관되게 적용하여 실험의 공정성과 재현성을 확보하였다. 또한, 원자의 수가 1개 이하인 분자는 제거하여 학습 안정성을 확보하였다.

모델의 학습 과정에서 Classification task는 데이터 불균형 문제를 완화하기 위해 가중치가 적용된 Cross-Entropy loss 를 적용하고, 성능 평가를 위해 AUC(Area Under the Receiver Operating Characteristic Curve)를 활용하였다. Regression task에서는 MAE(Mean Absolute Error)를 적용하여 학습을 진행하였으며, 모델의 성능을 평가하기 위해 결정계수(R^2)를 통해 모델 성능을 평가하였다.

4.2 성능 비교

비교 실험을 위해 총 5개의 그래프 기반 베이스라인 모델(ST-GCN, ST-MGA, MT-GCN[14], MGA[15], MTGL-ADMET[16])을 선정하였다. 이들 모델은 모두 분자의 구조적 특성을 그래프 형태로 표현하고, 그래프 신경망 계열의 아키텍처를 활용하여 ADMET 속성을 예측하는 방식이다. ST-GCN과 MT-GCN은 GCN을 기반으로 하며, ST-MGA와 MGA는 R-GCN(Relational GCN)에 기반하여 결합 정보를 보다 정교하게 반영한다. MTGL-ADMET은 추가적인 ADMET Endpoint정보를 보조적으로 활용함으로써 예측 성능 향상을 달성한 모델이다. 본 연구 또한 그래프 기반의 분자 표현을 활용하여 ADMET 속성을 예측하므로, 이와 같은 모델들을 비교 대상으로 선정하였다.

표 1은 10개의 ADMET Endpoints에 대해 기존 베이스라인 모델들과 제안 모델의 성능을 비교한 결과를 나타낸다. Average는 각 모델이 수행한 10개 tasks의 산술 평균으로 계산하였다. 본 연구에서 사용된 모든 평가지표(AUC, 결정계수)는 공통적으로 0에서 1 사이의 값을 가지며, 값이 클수록 모델의 예측 성능이 우수함을 의미한다. 이러한 특성으로 인해 각 지표의 평균값을 활용하여 모델의 전반적인 성능 경향을 파악할 수 있는 참고 지표로 활용하였다. 실제로 해당 평균 값에서도 제안 모델이 가장 높은 성능을 기록하였으며, 이는 제안된 방법이 다양한 ADMET 예측 과제 전반에서 일관되게 우수한 성능을 보였음을 시사한다. 특히, 제안 모델은 10개의 예측 항목 중 8개에서 최고 성능을 기록하였으며, 나머지 2개 항목에서도 두 번째로 높은 성능을 달성하였다.

뿐만 아니라, 본 논문에서 제안하는 모델의 궁극적인 목표는 ADMET 분야의 부족한 데이터셋에서도 높은 일반화 성능을 확보하는 것이다. Half-life와 Hepatotoxicity 속성 예측의 경우, 각각 1323개, 1313개의 제한된 데이터셋을 사용했음에도 불구하고 기존 모델 대비 우수한

표 1 10개의 ADMET Endpoints에 대한 베이스라인 모델과 제안 모델의 성능 비교
Table 1 Performance Comparison Between the Baseline Model and the Proposed Model on 10 ADMET Endpoints

No.	ADMET Endpoints	Dataset (Positive/Negative)	Metric	ST-GCN	ST-MGA	MT-GCN	MGA	MTGL-ADMET	Ours
1	HIA	734 (632/102)	AUC	0.916	0.972	0.899	0.911	0.981	**0.985**
2	p-gp substrates	888 (548/340)	AUC	0.775	0.755	0.733	0.719	0.801	**0.802**
3	CYP1A2 inhibitor	9748 (3582/6166)	AUC	0.932	0.931	0.914	0.941	**0.952**	0.942
4	CYP2D6 inhibitor	10682 (1401/9281)	AUC	0.848	0.841	0.839	0.877	0.869	**0.887**
5	CYP3A4 inhibitor	11369 (3586/7783)	AUC	0.892	0.915	0.865	0.875	0.916	**0.925**
6	Half life	1323 (538/785)	AUC	0.725	0.708	0.688	0.707	0.727	**0.762**
7	Hepatotoxicity	1313 (782/531)	AUC	0.653	0.669	0.612	0.713	0.701	**0.812**
8	Respiratory-tox	1399 (844/555)	AUC	0.842	0.872	0.810	0.828	0.859	**0.877**
9	PPB	1830 (Regression)	R^2	0.599	0.585	0.589	0.568	0.626	**0.636**
10	ESOL	1128 (Regression)	R^2	0.814	0.896	0.824	0.866	**0.931**	0.923
	Average			0.7996	0.8144	0.7773	0.8005	0.8363	**0.8551**

성능을 달성하였음을 확인할 수 있다. 이는 소규모 데이터셋에서도 강한 일반화 성능을 확보할 수 있음을 입증할 수 있는 결과로, 제안 모델이 데이터 제한이 존재하는 ADMET 예측 문제에 효과적으로 적용될 수 있음을 보여준다.

4.3 Abaltion Study

본 연구에서는 그래프 트랜스포머를 활용하여 원자 간 결합 정보를 반영한 그래프 구조를 학습하였다. 또한, 데이터 부족 문제를 해결하고 보다 정밀한 분자 표현을 생성하기 위해, 사전 학습된 UniMol 모델의 3D 구조 정보를 반영한 분자 임베딩 벡터를 추출하여 결합하는 방법을 제안하였다.

제안 모델의 유효성을 검증하기 위해 Ablation Study를 수행하였으며, 실험은 세 가지 설정으로 진행되었다. 첫 번째 실험에서는 노드의 결합 유형 정보를 제거한 그래프 임베딩을 사용하였고, 두 번째 실험에서는 UniMol 임베딩을 제거하였으며, 세 번째 실험에서는 결합 유형 정보와 UniMol 임베딩을 모두 제거하였다. Ablation Study는 제안한 모델이 기존 베이스라인 대비 가장 높은 성능을 나타낸 ADMET Endpoints를 대상으로 수행되었다. 모든 실험은 표 1과 동일한 평가지표를 사용하였으며, Classification task에는 AUC, Regression task에는 결정계수(R^2)를 적용하여 모델 성능을 평가하였다.

표 2 분자 임베딩 방식에 따른 Ablation Study
Table 2 Ablation Study Based on Molecular Embedding Methods

ADMET Endpoints	Metric	Settings		
		w/o Bond-type	w/o UniMol	w/o Both
HIA	AUC	0.9692	0.5000	0.7877
p-gp-substrates	AUC	0.7016	0.5417	0.4792
CYP2D6 inhibitor	AUC	0.82193	0.7455	0.7149
CYP3A4 inhibitor	AUC	0.88728	0.8452	0.8348
half-life	AUC	0.6843	0.5214	0.6391
Hepatotoxicity	AUC	0.7135	0.5000	0.6391
Respiratory-tox	AUC	0.7452	0.4889	0.6046
PPB	R^2	0.3417	-2.86973	-2.8125

실험 결과는 표 2와 같으며, 모든 Ablation 설정에서 제안 모델보다 성능이 저하되는 것을 확인할 수 있다. 특히, p-gp-substrates Endpoint의 경우, 제안 모델을 통해서 0.802의 AUC 성능을 기록한 반면, 결합 유형 정보를 제거한 경우(w/o Bond-type)는 0.7016로 약 12.52%, UniMol 임베딩을 제외한 경우(w/o UniMol)는 0.5417로, 약 32.45%, 두 정보를 모두 제거한 경우(w/o Both)는 0.4792로 약 40.25% 성능이 감소하였다. 이는 결합 유형 정보를 추가함으로써 화합물의 물리화학적 특성을 더욱 정교하게 학습할 수 있으며, UniMol 임베딩을 통해 분자의 3D 구조 정보를 효과적으로 반영할 수 있음을 반증하는 결과이다.

추가적으로, Ablation Study에서 사용된 세 가지 설정과 제안 모델을 비교하기 위해 t-SNE 시각화를 수행하였다. 그림 2는 두 가지 ADMET 속성에 대한 t-SNE시각화 결과를 보여주며, 각각 (a) CYP2D6 (Classification task)와 (b) PPB (Regression task)에 대해 t-SNE 분석을 수행한 결과이다.

t-SNE 결과를 살펴보면, 임베딩의 일부 요소가 제거된 경우 데이터 분포가 상대적으로 불분명하고, 클래스 간 구분이 모호한 경향을 보였다. 반면, 제안 모델에서는 데이터가 더욱 명확하게 군집화되는 것을 확인할 수 있다.

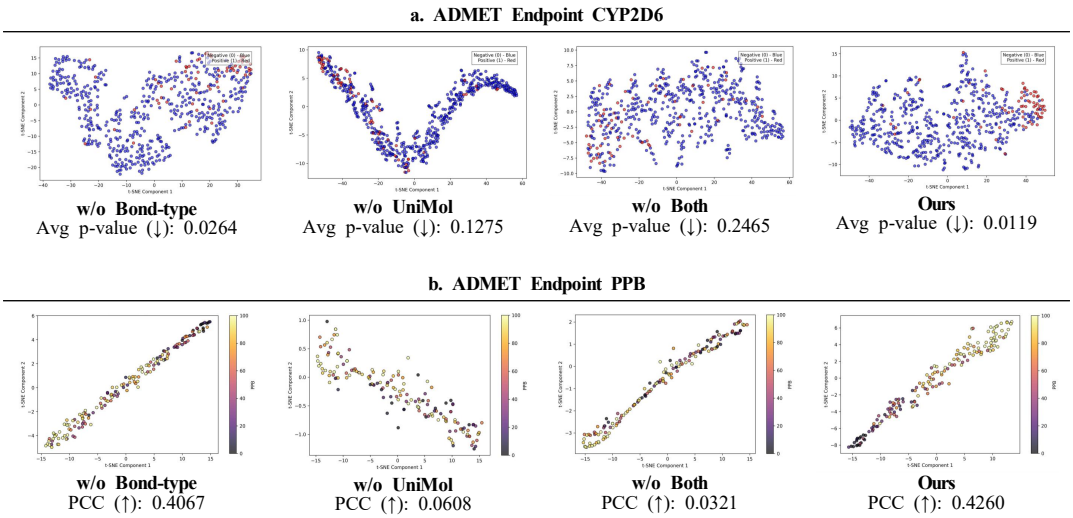


그림 2 분자 임베딩 방식에 따른 특징 표현의 t-SNE 분석. 임베딩을 통한 표현력 정량화를 위해, a. CYP2D6 (Classification)에서는 t-SNE 각 축에 대해 t-test를 수행하여 평균 p-value(Avg p-value)를 계산하였으며, b. PPB (Regression)에서는 t-SNE 거리 차이와 라벨 차이 간 피어슨 상관계수(PCC)를 활용하였다

Fig. 2 t-SNE Analysis of Feature Representations Based on Molecular Embedding Methods. Representation ability is quantified via embeddings. a. For CYP2D6 (Classification), t-tests were performed on each t-SNE axis and averaged (Avg p-value); b. For PPB (Regression), Pearson correlation coefficient(PCC) between t-SNE distances and label differences was utilized

t-SNE를 이용하여 각 모델 설정에서의 임베딩 군집도를 시각적으로 비교함과 동시에, 이를 정량적으로 평가하여 모델 간의 성능 차이를 보다 명확히 분석하고자 하였다. 이를 위해 Classification task와 Regression task 각각에 대해, 임베딩의 군집 정도를 정량화하는 서로 다른 방법을 적용하였다.

Classification task (CYP2D6)에서는, 클래스 간 임베딩 분포 차이를 정량적으로 평가하기 위해 t-test를 수행하였다. t-SNE 시각화로 얻은 2차원 임베딩 공간의 각 축에 대해 독립적으로 t-test를 적용하고, 축별 p-value를 평균내어 평균 p-value(Avg p-value)로 사용하였다. 이 평균 p-value는 클래스 간 분포 차이의 통계적 유의성을 나타내며, 값이 낮을수록 두 클래스 간 구분이 명확함을 의미한다. 분석 결과, 제안 모델의 평균 p-value는 0.0119로, 각각 0.0264, 0.1275, 0.2465의 평균 p-value를 나타낸 다른 설정들과 비교해 더욱 뚜렷한 클래스 구분을 형성함을 확인할 수 있다.

Regression task (PPB)에서는, 임베딩이 연속적인 라벨 정보를 반영하는 정도를 평가하였다. 이를 위해 두 샘플 간의 실제 라벨 값 차이와 t-SNE 공간 상 거리 차이를 각각 x축과 y축으로 하여 산점도를 구성하고, 이 두 변수 간의 피어슨 상관계수(PCC)를 계산 하였다. 상관계수가 클수록 거리 차이가 실제 라벨 차이에 더 잘 대응함을 의미한다. 제안 모델의 PCC 값은 0.4260으로, 각각 0.4067, 0.0608, 0.0321의 PCC 값을 가지는 다른 설정 대비 연속적인 속성 차이를 보다 정밀하게 반영하고 있음을 정량적으로 입증하였다.

4.4 Hyperparameter sensitivity

본 연구에서는 그래프 임베딩과 UniMol 임베딩을 결합하여 분자 예측 성능을 향상시키고자 하였다. 그 과정에서 각 임베딩의 상대적 중요도를 분석하기 위해, 두 임베딩을 가중합 형태로 결합하여 성능을 비교하였다. 구체적으로, 그래프 트랜스포머를 통해 얻은 그래프 임베딩 h_g와 사전 학습된 UniMol 모델로부터 추출한 3D 구조 기반 임베딩 h_u를 결합하여 최종 분자 표현 $h_{integrated}$을 구성하였다. 이때 결합은 가중치 $a \in [0.1,0.9]$를 사전 설정하여 선형 결합하는 방식으로 구성하였으며, 수식 (3)과 같이 표현된다.

$$h_{integrated} = a \cdot h_g + (1-a) \cdot h_u \quad (3)$$

실험에서는 가중치 값을 0.1부터 0.9까지 0.1 단위로 변화시키며, 각 가중치에 따른 속성별 모델 성능 변화율의 평균을 계산하고 비교하였다. 여기서 성능 변화율은 가중합을 적용하지 않았을 때의 성능을 기준으로, 가중

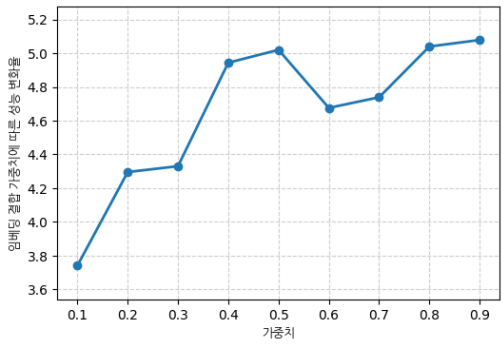


그림 3 두 임베딩의 결합 가중치에 따른 성능 변화율
Fig. 3 Performance Change Rate According to Combination Weights of Two Embeddings

합 적용 후 성능이 얼마나 변화했는지를 백분율(%)로 나타낸 값이다. 일부 속성에서는 성능이 감소하는 경우도 존재하므로, 평균을 계산하기에 앞서 전체 변화율에 동일한 값을 더하여 최소 변화율이 0이 되도록 정규화하는 과정을 수행하였다. 이를 통해 음수의 변화율이 평균 값에 미치는 영향을 보정하고, 다양한 가중치 설정에서의 상대적인 성능 변화를 보다 명확하게 비교할 수 있도록 하였다.

그림 3에 제시된 결과에서 확인할 수 있듯이, 가중합 방식에 따라 모델 성능에 차이가 존재하며, 특히 그래프 임베딩의 가중치가 증가할수록 모델 성능이 향상되는 경향을 보였다. 이는 분자 예측 성능을 결정하는 데 있어 원자 간 결합 정보를 반영하는 그래프 임베딩이 UniMol 임베딩보다 더 중요한 역할을 수행함을 의미한다.

4.5 어텐션 스코어 기반의 모델 해석력 분석

모델의 해석 가능성을 평가하기 위하여, 예측 과정에서 주목하는 화학적, 구조적 특성을 분석하고자 하였다. 구체적으로, 학습 데이터셋을 기반으로 주요 서브구조(Substructure)를 식별하고, 이 서브구조를 포함하는 테스트 데이터셋 분자에 대해 모델의 어텐션 분포를 시각화하였다. 주요 서브구조는 학습 데이터셋에서 실제 라벨 기준 하위 10%에 해당하는 샘플들로부터 공통적으로 나타나는 구조적 패턴을 추출하여 정의하였으며, 해당 구조가 화학적 특성 예측에서 중요한 역할을 할 것이라는 가정 하에 분석을 수행하였다.

그림 4는 ESOL Endpoint에 대한 실험 결과로, 학습 데이터셋으로부터 식별한 공통 서브구조를 포함하는 테스트 데이터셋 분자들에 대해 모델의 어텐션 분포를 시각화한 결과를 제시한다. 주요 서브구조는 SMARTS (SMILES Arbitrary Target Specification) 표현 방식으로 추출되었으며, 이 실험에서는 [#6]-,:[#6]-,:[#6]-,:[#6] 구조가 공통 서브구조로 도출되었다. 어텐션 스코어는 SMILES 문자열을 기반으로 생성한 2차원 분자 구조 위에 매핑하였으며, 각 부분의 어텐션 값에 비례하여 색농도가 변하도록 시각화 하였다. 특히, 모델이 높은 중요도를 부여한 영역에 집중하기 위해, 일정 수준 이상의 어텐션 스코어를 갖는 부분만을 시각화 대상으로 선정하였다. 그림 4의 (a), (b)는 어텐션 스코어가 0.7 이상인 부분만을, (c), (d)는 0.8 이상인 부분만을 표시하여 모델의 어텐션 분포 특성을 보다 명확히 비교할 수 있도록 구성하였다.

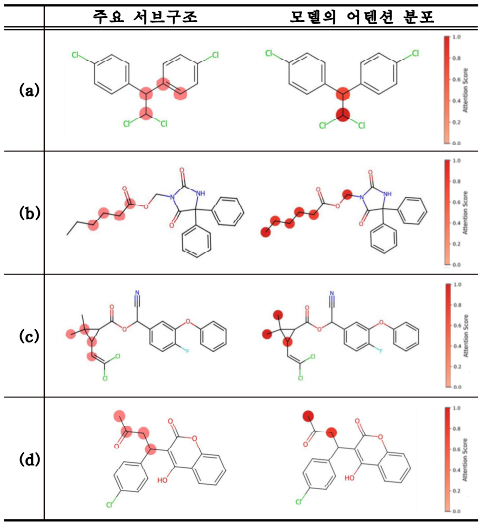


그림 4 주요 서브구조와 모델의 어텐션 분포 시각화
Fig. 4 Visualization of Key Substructure and Model Attention Distribution

테스트 데이터셋에서 해당 서브구조를 포함하는 샘플을 대상으로 수행한 어텐션 가중치 분포와 주요 서브구조를 비교하였을 때, 모델이 주요 서브구조에 대해 상대적으로 높은 가중치를 부여하는 경향을 보였다. 이는 모델이 예측 과정에서 중요할 것으로 판단되는 구조적 정보를 실제로 주목하고 있음을 간접적으로 시사하며, 모델의 결정 과정을 해석할 수 있는 가능성을 기대할 수 있게 한다.

5. 결 론

본 논문에서는 신약 개발 과정에서 중요한 ADMET (흡수, 분포, 대사, 배설, 독성) 속성을 보다 정밀하게 예측하기 위해, 그래프 트랜스포머 기반 모델을 설계하였다. 기존 QSAR 및 GNN 기반 모델이 갖는 분자 표현

의 한계를 극복하고자, 결합 유형 정보를 포함한 그래프 임베딩과 사전 학습된 UniMol 모델에서 추출한 3D 구조 정보를 결합하는 방법을 도입하였다.

먼저, 그래프 임베딩 과정에서 단순한 원자 간 결합 여부뿐만 아니라 결합 유형을 반영하여 보다 정교한 분자 표현을 학습하도록 설계하였다. 이를 기반으로, 그래프 트랜스포머의 셀프 어텐션 메커니즘을 활용하여 원자 간 장거리 상호작용을 효과적으로 학습하고, 엣지 정보를 통해 결합의 화학적 특성을 정밀하게 반영하였다. 또한, 사전 학습된 UniMol 임베딩을 적용하여 분자의 입체적 구조 정보를 학습함으로써 데이터 부족 문제를 보완하고, 모델의 일반화 성능을 향상시켰다.

실험에서는 10개의 ADMET 속성 예측(8개의 Classification task, 2개의 Regression task)을 수행하였으며 실험 결과, 결합 유형을 포함한 그래프 임베딩과 UniMol 임베딩을 함께 활용한 제안 모델이 기존 방식보다 ADMET 예측 성능에서 우수한 결과를 보였다. 이는 제안 모델이 결합 정보와 3D 구조를 효과적으로 통합함으로써, 신약 후보 물질의 약리학적 특성을 보다 정확하게 평가할 수 있음을 의미한다.

그러나, 본 연구는 여전히 제한된 수의 ADMET 라벨 데이터를 기반으로 모델을 학습하였다. UniMol과 같은 사전 학습된 분자 표현 모델을 활용하여 이러한 한계를 일부 보완하였으나, 다양한 화학적 특성을 포괄하기에는 데이터의 양과 다양성 측면에서 여전히 제약이 존재한다. 이에 따라, 향후에는 정답 정보가 없는 화합물 데이터를 추가로 활용하여 분자 표현 학습의 범위를 확장할 필요가 있다. 예를 들어, 대조 학습(Contrastive Learning) 기반의 자기 지도 학습(Self-supervised Learning) 기법이나 슈도 라벨링(Pseudo Labeling) 기반의 준지도 학습(Semi-supervised Learning) 기법을 적용하면, 제한된 라벨 환경에서도 보다 일반화된 분자 표현을 효과적으로 학습할 수 있을 것으로 기대된다.

궁극적으로, 대규모 데이터셋을 효과적으로 활용할 수 있는 학습 기법에 대한 연구는 필수적이며, 이러한 보완을 통해 제안한 모델은 실제 신약 개발 과정에서도 더욱 정밀하고 신뢰성 있는 ADMET 예측 도구로 발전할 수 있을 것이다.

References

[1] Waring, M.J., Arrowsmith, J., Leach, A.R., et al., “An analysis of the attrition of drug candidates from four major pharmaceutical companies,” *Nat. Rev. Drug Discov.,* Vol. 14, No. 7, pp. 475-486, 2015.

[2] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar, “Chemberta: Large-scale self-supervised pretraining for molecular property prediction,” arXiv preprint arXiv:2010.09885, 2020.

[3] Zhou G, Gao Z, Ding Q, et al., “Uni-Mol: a universal 3D molecular representation learning framework,” *In 2023 11th International Conference on Learning Representation (ICRL),* 2023.

[4] Heid E, Greenman KP, Chung Y, et al., “Chemprop: A machine learning package for chemical property prediction,” *Journal of Chemical Information and Modeling,* Vol. 64, No. 1, pp. 9-17, 2024.

[5] Vijay Prakash Dwivedi, Xavier Bresson, “A Generalization of Transformer Networks to Graphs,” *AAAI Workshop*, 2020.

[6] Li, J., Jiang, X., “Mol-BERT: An effective molecular representation with BERT for molecular property prediction,” *Wirel Commun. Mobile Comput 2021,* pp. 1-7, 2021.

[7] Yang, M., Chen, J., Xu, L., et al., “A novel adaptive ensemble classification framework for ADME prediction,” *RSC Adv.* Vol. 8, No. 21, pp. 11661-11683, 2018.

[8] Wang, X., Liu, M., Zhang, L., et al., “Optimizing pharmacokinetic property prediction based on integrated datasets and a deep learning approach,” *J. Chem. Inf. Model,* Vol. 60, No. 10, pp. 4603-4613, 2020.

[9] Wang, N.-N., Deng, Z.-K., Huang, C., et al., “ADME properties evaluation in drug discovery: prediction of plasma protein binding using NSGA-II combining PLS and consensus modeling,” *Chemom. Intell. Lab. Syst,* Vol. 170, pp. 84-95, 2017.

[10] Alsenan, S., Al-Turaiki, I., Hafez, A., “A deep learning approach to predict blood-brain barrier permeability,” *PeerJ Comput. Sci.* 7:e515, 2021.

[11] Wu, Z., Jiang, D., Wang, J., et al., “Mining toxicity information from large amounts of toxicity data,” *J. Med. Chem*, Vol. 64, No. 10, pp. 6924-6936, 2021.

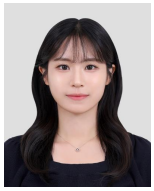
[12] Lombardo, F., Berellini, G., Obach, R.S., “Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 1352 drug compounds,” *Drug Metab. Dispos,* Vol. 46, No. 11, pp. 1466, 2018.

[13] Delaney, J.S., “ESOL: estimating aqueous solubility directly from molecular structure,” *J. Chem. Inf. Comput. Sci,* Vol. 44, No. 3, pp. 1000-1005, 2004.

[14] Montanari, F., Kuhnke, L., Ter Laak, A., et al., “Modeling physico-chemical ADMET endpoints with multitask graph convolutional networks,” *Molecules,* Vol. 25, No. 1, 44, 2019.

[15] Xiong, G., Wu, Z., Yi, J., et al., “ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties,” *Nucleic Acids Res.,* Vol. 49, No. W1, pp. W5-W14, 2021.

[16] Du B-X, Xu Y, Yiu S-M. et al., “MTGL-ADMET: A Novel Multi-task Graph Learning Framework for ADMET Prediction Enhanced by Status-Theory and Maximum Flow,” *International Conference on Research in Computational Molecular Biology,* pp. 85-103, 2023.



김 윤 주

2025년 숙명여자대학교 소프트웨어융합전공(학사). 관심분야는 AI 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 Center PostDoctoral 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&기계학습