The Korean Journal of Applied Statistics
2024, Vol. 37, No. 3, 355–380

DOI: https://doi.org/10.5351/KJAS.2024.37.3.355
Print ISSN1225-066X / Online ISSN 2383-5818

Toxicity prediction of chemicals using OECD test guideline data with graph-based deep learning models

Daehwan Hwang[a], Changwon Lim[1,a]

[a]Department of Applied Statistics, Chung-Ang University

Abstract

In this paper, we compare the performance of graph-based deep learning models using OECD test guideline (TG) data. OECD TG are a unique tool for assessing the potential effects of chemicals on health and environment. but many guidelines include animal testing. Animal testing is time-consuming and expensive, and has ethical issues, so methods to find or minimize alternatives are being studied. Deep learning is used in various fields using chemicals including toxicity prediciton, and research on graph-based models is particularly active. Our goal is to compare the performance of graph-based deep learning models on OECD TG data to find the best performance model on there. We collected the results of OECD TG from the website eChemportal.org operated by the OECD, and chemicals that were impossible or inappropriate to learn were removed through pre-processing. The toxicity prediction performance of five graph-based models was compared using the collected OECD TG data and MoleculeNet data, a benchmark dataset for predicting chemical properties.

Keywords: toxicity prediction, graph neural network, OECD test guideline, deep learning

1. 서론

과학기술의 발전으로 신약과 신물질과 같은 화학물질들이 개발되고 있다. 그러나 많은 화학물질에는 긍정적 효과 외에도 유해성이 존재할 수 있다. 따라서 국제기구인 경제협력개발기구(organisation for economic co-operation and development; OECD)는 유해성에 대한 검증방법을 화학물질시험 가이드라인(test guidelines; TG) 4장에 제시하고 있다. 이 가이드라인에는 동물실험에 대한 내용이 포함되어 있는데, 동물실험은 많은 비용과 시간이 소요될뿐만 아니라 윤리적으로도 문제가 제기되고 있다. 예를 들면 TG422 실험은 63일의 실험기간이 소요되고, 최소 20마리의 실험 동물이 필요하며, 실험 종료 후 출생된 새끼를 포함한 모든 동물을 해부하여 검사한다 (OECD, 2016a). 따라서 국제기구 및 국가기관들은 동물 실험을 하지 않고, 화학물질의 유해성을 평가하기 위한 대안을 모색하고 있다.

OECD는 이를 위해 2008년에 QSAR Toolbox를 개발하였다. 이것은 화학물질의 유해성 평가를 위한 의사결정을 돕는 소프트웨어로서 화학물질을 범주로 그룹화하고, 화학물질의 유해성을 평가하는데 필요한 독성 데이터의 공백을 메우는데 사용된다 (https://qsartoolbox.org/). 또한 미국국립보건원(national institutes of health; NIH), 미국 환경보호국 (U.S. environmental protection agency; EPA), 미국 식품의약국(U.S. food and drug administration; FDA) 등 다양한 미국정부기관이 화학물질의 잠재적 유해성을 신속하고 효율적으로 테스트할 수 있는 독성평가방법을 개발하는 것을 목표로 하는 Tox21 프로그램을 추진중에 있다 (National Research

This research was supported by the Chung-Ang University Research Scholarship Grants in 2021.
[1]Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, DongJak-Gu, Seoul 06974, Korea. E-mail: clim@cau.ac.kr

Published 30 June 2024 / journal homepage: http://kjas.or.kr

Council, 2007). 이 프로그램에는 동물실험을 시험관 내 세포실험으로 대체하는 것이 포함되며, 2014년에 12가지 독성에 대한 7,000여개의 화학물질의 데이터셋을 활용해 독성을 예측하는 챌린지를 개최되었다. 이 챌린지를 통해 머신러닝과 딥러닝이 화학물질의 독성 예측에 사용될 수 있음이 확인되었다 (Mayr 등, 2016). 이후 많은 연구자들이 독성을 포함한 화학물질의 다양한 특성 예측이 가능한 머신러닝, 딥러닝 모델 개발에 관한 연구를 수행하였다 (Coley 등, 2017; Schutt 등, 2017; Gilmer 등, 2017; Rong 등, 2020; Zhang 등, 2021).

딥러닝에서는 화학물질의 특징을 1열의 벡터로 표현하는 분자지문 또는 화학물질을 이루는 원자와 원자사이의 결합에 기반한 graph로 만들어 활용할 수 있다. Connor 등 (2017)은 화학물질의 옥탄올 용해성과 독성 등 다양한 특성을 예측하는 task에 서포트 벡터 머신과 convolutional neural network (CNN)을 사용할 때, 분자의 구조를 분석해 얻어진 특징을 담고있는 분자지문보다, 각 노드, 엣지별로 특징을 추출한 벡터를 신경망에 통과시킨 후 1개의 벡터로 합치는 graph를 사용하는 것이 더 효과적임을 보였다. Schutt 등 (2017)은 양자역학의 원자화 에너지 예측을 위해 원자번호와 각 원자간 거리행렬을 이용한 deep tensor neural networks을 제안하였다. 또한 Gilmer 등 (2017)은 원자화 에너지 예측을 위해 화학물질의 노드와 엣지의 특성을 추출해 같이 활용하는 구조인 edge neural network (enn)가 유용함을 보였다. Wu 등 (2018)은 생리학, 물리화학 등 다양한 화학물질의 특성에 대한 데이터셋들을 정리한 MoleculeNet 데이터베이스를 구축하였고, 각 데이터별 성능평가지표와 평가하는 방법 등을 정리하여 전통적인 머신러닝 기법 및 딥러닝 기법들의 성능을 비교하였다. 이후 개발된 모델들은 MoleculeNet의 평가데이터셋과 방법을 적용하여 비교한 경우가 많다.

최근에도 화학물질의 특성을 예측하기 위한 Graph기반 딥러닝 모델을 개발하는 연구가 활발히 이루어지고 있다. Rong 등 (2020)은 신약개발을 위한 화학물질의 특성 예측을 위해 graph transformer를 원자와 결합의 각 관점에서 분석하도록 병렬로 구성하였으며, 인접 원자와 결합의 정보를 활용하는 contextual property prediction을 포함한 자기지도학습을 통해 state-of-the-art (SOTA)를 달성한 모델인 Graph Representation frOm self-superVised mEssage passing tRansformer (GROVER)를 제안하였다. Zhang 등 (2021)은 화학물질의 특성 예측을 위해 분자의 topology information에 기반한 자기지도학습을 하여 성능을 향상시킨 Motif-based Graph Self-Supervised Learning for molecular property prediction (MGSSL) 모델을 제안하였다. 이 모델에는 인코더로 Graph Isomorphism Network (GIN), Graph Convolutional Network (GCN), Graph SAmple and aggreGatE (GraphSAGE) 등 다양한 Graph Neural Network (GNN) 계열 모델이 사용될 수 있음을 보였다. GCN은 Graph기반 데이터에서 semi-supervised 방식에서 노드를 분류하는데 합성곱 연산을 적용한 모델이다 (Kipf와 Welling, 2016). GraphSAGE는 단백질 구조와 소셜 네트워크와 같은 대규모 graph에서 노드를 분류하는데 임의의 K개의 노드를 샘플링하여 정보를 종합하는 모델이다 (Hamilton 등, 2017). GIN은 graph의 동형여부를 확인하기 위해 각 노드의 고유의 레이블 지정하고, 정보를 종합한 결과가 같은지 비교하는 Weisfeiler-Lehman test의 방법을 신경망으로 구현하여 단백질 구조와 소셜 네트워크에서 노드와 graph를 분류할 수 있는 모델이다 (Xu 등, 2018).

본 논문에서 사용할 OECD TG 데이터는 화학물질의 독성을 어떻게 검증해야되는지에 대한 구체적인 가이드라인을 따라 시행된 실험의 결과이다. 예를 들면, TG423의 경우 1회만 투여하고, 14일의 관찰 기간을 둘 것, 농도 당 3마리 이상을 사용할 것, 독에 더 민감한 암컷을 사용할 것, 사육시설은 25 ± 3°C을 유지할 것 등 세심하고도 구체적인 실험방법을 제시하고 있다. 이렇게 얻어진 OECD TG 데이터를 사용하여 화학물질의 독성을 예측하는 연구들이 있다. Luechtefeld 등 (2018)은 동물실험을 통해 확인 가능한 급성경구독성 등 19가지 독성을 예측하기 위해 분자지문에서 유사도를 이용하는 머신러닝 모델을 개발하였으며, OECD TG 실험결과를 따르는 데이터셋을 연구에 활용하였다. Bae 등 (2021)은 유전 독성 예측 머신러닝 모델에 적용할 분자지문에서의 클래스 불균형 해소 기법 연구에 OECD TG 471 데이터를 수집하여 활용하였고, Silva 등 (2021)은 안구 관련 독성을 예측하기 위해 4가지 분자지문을 동시에 활용하는 머신러닝 모델 연구에 TG405에 기반한 데이터를 활용하였다.

본 연구에서는 OECD TG 데이터를 사용하여 graph기반 딥러닝 모델들을 학습하고 그 모델들의 화학물

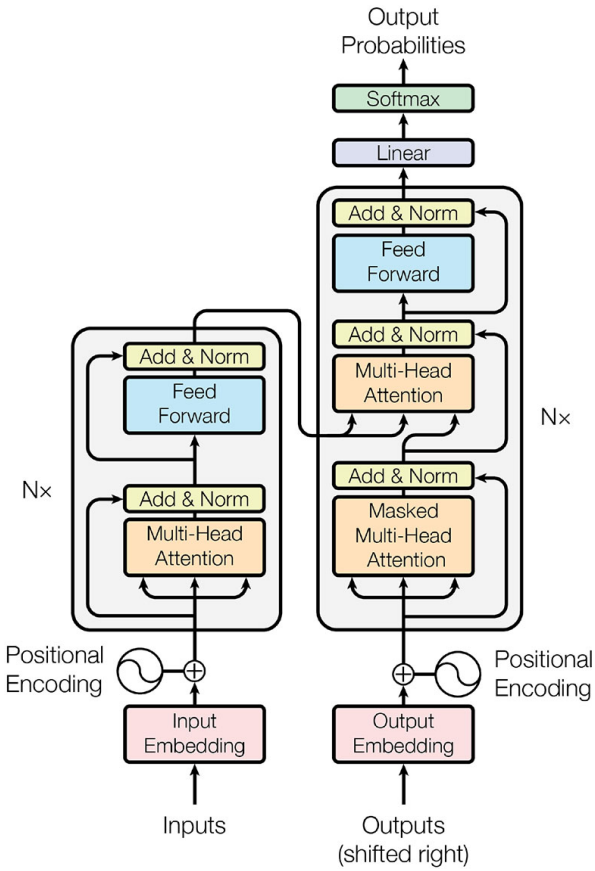


Figure 1: *Architecture of transformer (Vaswani et al., 2017).*

질의 독성 예측 성능을 비교해보고자 한다. 먼저 OECD TG를 따른 실험결과를 제공하는 eChemportal.org로부터 데이터를 수집한다. 둘째, 결합이 없어 graph로 표현이 불가능한 이온상태의 화학물질과 같이 사용이 불가능한 화학물질을 제거한다. 셋째, 3회 서로 다른 seed값에서 데이터를 분할하여 학습을 하고, 성능지표의 평균과 표준편차를 구한다. 화학물질의 특성 예측 분야에서 분류 성능 평가 지표로 ROC-AUC만을 사용하지만, 우리는 독성이 있는 화학물질을 예측하는 것이 중요하기 때문에 F1점수, 재현율, 정밀도와 정확도를 함께 보며, 주요 지표는 ROC-AUC과 F1점수로써 성능을 비교하였다. 이를 통해 OECD TG 데이터에 적합한 모델을 찾고자 한다.

2. 선행 및 관련 연구

2.1. Transformer

Vaswani 등 (2017)은 기계번역 모델로 attention을 병렬로 작동하는 multi-head attention 메커니즘을 활용한 트랜스포머를 제안하였다. Figure 1은 트랜스포머 구조이다. Input 문장을 임베딩 벡터로 만들어주는 첫 레이어 이후, 문장 내 각 단어들의 위치를 벡터로 생성해주는 positional encoding을 거친 후에 attention block을 통과한다. 여기서 attention이란 번역과 같이 입력된 sequence로부터 다른 sequence를 출력하는 sequence-to-

sequence (Seq2Seq) 모델들에서 사용하는 방법으로 문장에서 서로 다른 단어간 어떤 것이 중요한지, 주의를 기울여야할지 점수를 매기는 방법이다 (Bahdanau 등, 2015).

Transformer에서는 Seq2Seq모델들처럼 문장 내 단어를 순차적으로 넣지 않고, 한 번에 넣으며, 문장 내의 단어들간의 attention을 모델이 스스로 생성하는 self-attention을 사용한다. Attention 값은 scaled dot-product attention layer를 통해 계산한다. 단어에서 문장간의 관계를 분석하기 위해 입력값을 query, key, value로 구성하고, 모든 key와 query들을 점곱을 하고, value로부터 가중치를 얻어 softmax함수를 적용시켜 계산을 한다. Attention 점수의 식은 다음과 같다.

$$\text{softmax}(X) = e^X / \sum_{n=1}^{N} e^{x_n},$$
$$\text{Attention}(Q, K, V) = \text{softmax}\left(QK^T / \sqrt{d_k}\right) V. \tag{2.1}$$

Softmax는 입력값을 각 클래스별 예측 확률의 합을 1이 되도록하는 멀티-클래스 분류에 사용되는 함수이다. 여기서 N은 문장 내 단어의 개수이며, X는 input 행렬이며, x_n은 n번째 입력값이다. Q, K, V는 각각 query, key, value의 행렬이다. d_k는 query와 key의 차원수이다.

Attention block에는 multi-head attention과 feed forward network와 잔차연결이 있다. Multi-head attention은 여러개의 scaled dot-product attention layer들을 쌓고, 동시에 병렬로 작동을 시킨다. 이로 인해 여러개의 head로 동시에 다양한 기준을 갖고 분석이 가능하다. n개의 attention layer로 구성된 multi-head attention을 통과했을때의 식은 다음과 같다.

$$\text{Multihead}\,(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_n) W^O,$$
$$\text{where head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right). \tag{2.2}$$

W_i^Q, W_i^K, W_i^V는 query, key, value별 i 번째 head의 은닉층의 가중치이다. W^O는 Multi-Head Attention의 마지막 은닉층의 가중치이다.

Figure 1의 우측에 있는 masked multi-head attention 앞에 masked가 붙은 이유는 디코더는 t시점의 단어를 예측하는데 t이후의 정보를 사용 못하도록 가리기 때문이다. Feed forward network는 완전연결층 2개와 활성화함수로 이루어져 있다. Multi-head attention의 결과로 나온 head별 attention값을 분석하는 네트워크다. Block의 중간과 끝에 있는 add & norm에서 add는 깊은 네트워크로 인한 경사도 소실과 같은 문제를 예방하는 잔차 연결 값을 더하는 과정을 의미하며, norm은 layer normalization으로 일반화 성능을 향상시킨다. 이 transformer 구조는 기계번역에 사용된 이후에, 오디오 분류, 이미지 분류 등 다양한 분야에 적용된다.

2.2. MoleculeNet

Wu 등 (2018)은 화학물질의 특성 예측 모델간 비교를 위해 데이터셋과 평가방법을 제안하였다. 데이터는 화학물질의 다양한 특성들을 고려하여 양자역학, 물리화학, 생물리학, 생리학 4가지 분야에 대한 17가지의 공개된 데이터를 수집하였으며, 각 데이터별 성능 평가 지표와 분할 방법을 제안한다.

데이터셋들은 FDA의 데이터베이스, 화학물질에 대한 데이터 및 머신러닝 라이브러리인 DeepChem과 같은 다양한 DB들로부터 수집되었다. 양자역학 데이터셋으로 QM7, QM7b, QM8, QM9가 있으며, 원자화 에너지, 상태 변동시 에너지 준위와 진동자 세기 등의 양자역학에 사용되는 특성을 포함하는 회귀 데이터셋이다. 물리화학 데이터는 ESOL, Lipophilicity, FreeSolv로 화학물질이 물에 녹는 정도인 수용성, 수화작용시 방출되는 에너지 양과 같은 물리화학적 특성을 포함하는 회귀 데이터셋들이다. 생물리학은 PCBA, MUV, HIV, PDBbind, BACE로 에이즈 바이러스나 치매에 영향을 끼치는 BACE-1 단백질을 화학물질이 억제할 수 있는지 여부와 같은 생물리학적 특성을 포함한 데이터셋이다. 생리학은 BBBP, Tox21, ToxCast, SIDER, Clintox로 독성 및 부작용에 대한 분류 데이터셋이다.

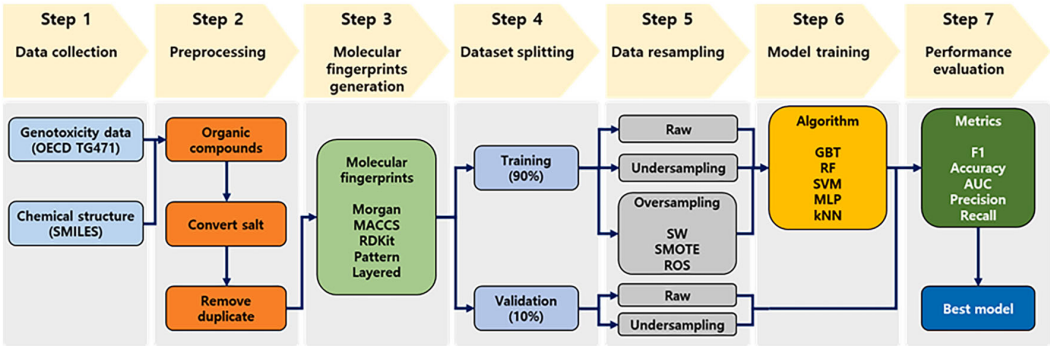


Figure 2: *Bae's workflow (Bae et al., 2021).*

분할 방법은 일반적으로 사용되는 random split뿐만이 아닌 화학물질의 scaffold를 기반으로 분할하는 방식도 활용하였다. Scaffold는 벤젠과 같은 고리구조 중 화학물질의 특성과 관련있는 구조를 의미하며, 약물의 subgraph 분류를 위해 만들어졌다 (Bemis와 Murcko, 1999). Scaffold를 기반으로 분할하는 방법은 화학물질의 특성을 고려한 방법으로 학습에 사용된 scaffold와 검증, 테스트시 사용되는 scaffold가 다르기 때문에 모델이 얼마나 잘 학습했는지 확인할 수 있는 challenge한 분할 방법이다. 평가시 시드값을 달리함으로써 데이터셋이 다르게 분할이 되도록 하여 3회 학습을 한 결과의 평균과 표준편차를 각 모델의 점수로 하였다. 지표는 분류에서는 ROC-AUC와 PRC-AUC를, 회귀에서는 RMSE와 MAE를 사용했다.

모델은 로지스틱 회귀분석과 같은 일반적인 방법 8가지와 graph구조로 만들어 노드(원자)와 엣지(결합)의 특성을 추출해 같이 활용한 edge neural network와 같은 graph 기반 모델 6종류를 선정했다. 특징 추출방법은 일반 모델들은 각 원자를 기준으로 인접 원자가 무엇인지에 따라 벡터를 생성하는 extended-connectivity fingerprint (ECFP)방법을 사용하였으며, graph 기반 모델들은 각 모델들이 사용한 방법을 그대로 사용했다. 비교 결과 양자역학, 물리화학 회귀데이터셋에서는 graph 기반 모델이 전반적으로 성능이 높았다. 분류는 생물리학은 전통적인 방법들이 성능이 높았고, 생리학에서는 두 방법이 비슷하나 graph 기반 방법이 약간 더 높은 성능을 보였다.

2.3. OECD TG 데이터를 사용한 연구

Bae 등 (2021)은 화학물질의 독성 데이터셋의 클래스 불균형비가 심한 경우가 많으며, 이런 경우 학습에 부정적 영향을 주기 때문에 분자지문에서 특징을 추출하여 독성을 예측하는 머신러닝 모델에 효과적인 클래스 불균형 해소 기법을 비교하였으며, 이를 위해 OECD TG 471 유전독성 데이터를 수집하였다.

Figure 2는 실험 workflow이다. 맨 먼저 OECD TG를 따른 실험결과를 제공하는 eChemPortal로부터 TG471에 해당하는 데이터를 수집하였으며, 분자지문 추출을 위해 미국 환경보호국(U.S. EPA)의 CompTox 화학 대시보드(https://comptox.epa.gov/dashboard)를 통해 simplified molecular input line entry system (SMILES)식으로 변환한다. SMILES식이란 분자구조를 결합 순서와 형태를 알 수 있도록 표기하는 방법이며, 5.1장에서 자세하게 설명하겠다.

2단계는 데이터 전처리로 3가지를 시행하였다. 첫째, 기본 골격이 탄소인 유기화합물이 아닌 경우 제거한다. 둘째, 이온상태인 염의 경우 이온상태가 아닌 적절한 화학물질로 변환한다. 셋째, 중복 화학물질을 제거하고, 만일 실험 결과가 하나는 양성, 하나는 음성과 같이 상이할 경우 제거한다. 전처리 결과 총 4,171개 데이터가 생성되었으며, 250개의 양성데이터와 3,921개의 음성데이터로 약 1 : 16에 달하는 불균형비를 가진 데이터셋이 생성되었다.

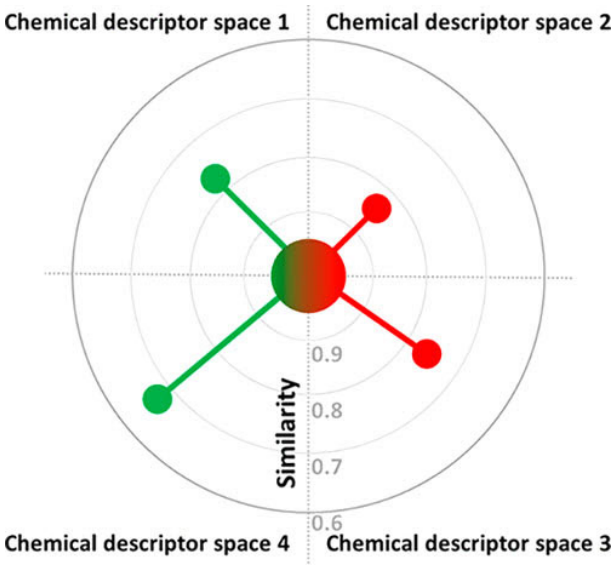


Figure 3: *MuDRA's descriptor space (Alves et al., 2018).*

3단계는 분자지문은 화학분야 라이브러리인 RDKit을 이용하여 생성하였다. 분자지문 생성방법은 각 원자를 기준으로 인접 수준별 원자와 결합이 어떠한지 원형태를 기준으로 분석하여 벡터를 생성하는 Morgan, 화학물질 내에 특정 구조가 포함되는지 여부에 대한 molecular access system (MACCS)과 RDKit의 pattern 방법, 화학물질의 원자들의 원자번호와 결합들이 어떤 결합인지 등의 특징을 추출하여 벡터로 생성하는 RDKit 분자지문, layered 방법을 사용하였다 (Landrum, 2019).

4단계는 학습 데이터와 검증 데이터를 random split을 활룡해 9 : 1비율로 나누어 분할한다.

5단계는 클래스 불균형 해소기법으로 up-sampling기법은 3가지를 사용했다. *K*-인접 방법으로 유사한 데이터를 생성하는 synthetic minority over-sampling technique (SMOTE)를 분자지문에 적용하는 방법, 임의의 데이터를 더 뽑는 random oversampling 방법, 불균형비에 따라 가중치를 주는 sample weight 방법을 사용했으며, down-sampling은 임의의 데이터를 제외하는 random undersampling을 사용하였다.

6단계 모델학습시 모델은 random forest (RF), support vector machine (SVM), k-nearest neighborhood (K-NN), gradient boosting tree (GBT), multi-layer perceptron (MLP)을 활용하였다.

마지막으로 성능지표는 ROC-AUC, 정확도, 정밀도, 재현율, F1점수를 활용하였으며, 주요 지표는 F1점수이다. 5-fold cross-validation을 통해 하이퍼파라미터를 탐색하고, F1점수가 가장 높은 결과를 모델별 최고성능으로 활용하였다. 연구 결과 SMOTE를 이용한 그레디언트 부스팅 트리 방법이 가장 높은 성능을 발휘하였다.

Silva 등 (2021)은 화학물질의 안구 자극이나 부식과 같은 유해성을 확인하는 동물실험인 OECD TG 405의 대안으로 Figure 3와 같이 4가지 분자지문을 서로 다른 축으로 하여 유사도를 비교하는 K-NN기반 모델인 multi-descriptor read across (MuDRA)의 사용을 제안한다. MuDRA 모델은 Alves 등 (2018)이 제안한 모델이다. 데이터는 유럽화학청(European chemical agency; ECHA)가 각종 실험문서들로부터 수집한 동물실험 결과 DB를 활용하였으며, 전처리는 4가지를 시행하였다. 첫째, OECD TG 405의 절차를 따르지 않은 경우 제거한다. 둘째, QSAR모델에 의한 예측 결과 데이터는 제거한다. 셋째, 무기화합물과 혼합물, 염의 경우 제거한다. 넷째, 중복 데이터와 동일한 화학물질의 결과가 서로 다른 경우 제거한다. 전처리 결과 총 3,547개 데이터가 생성되었으며, 2,401개는 독성이 없으며, 937개는 자극을 유발하고, 209개는 부식을 유발하는 데이터가 생성되었다.

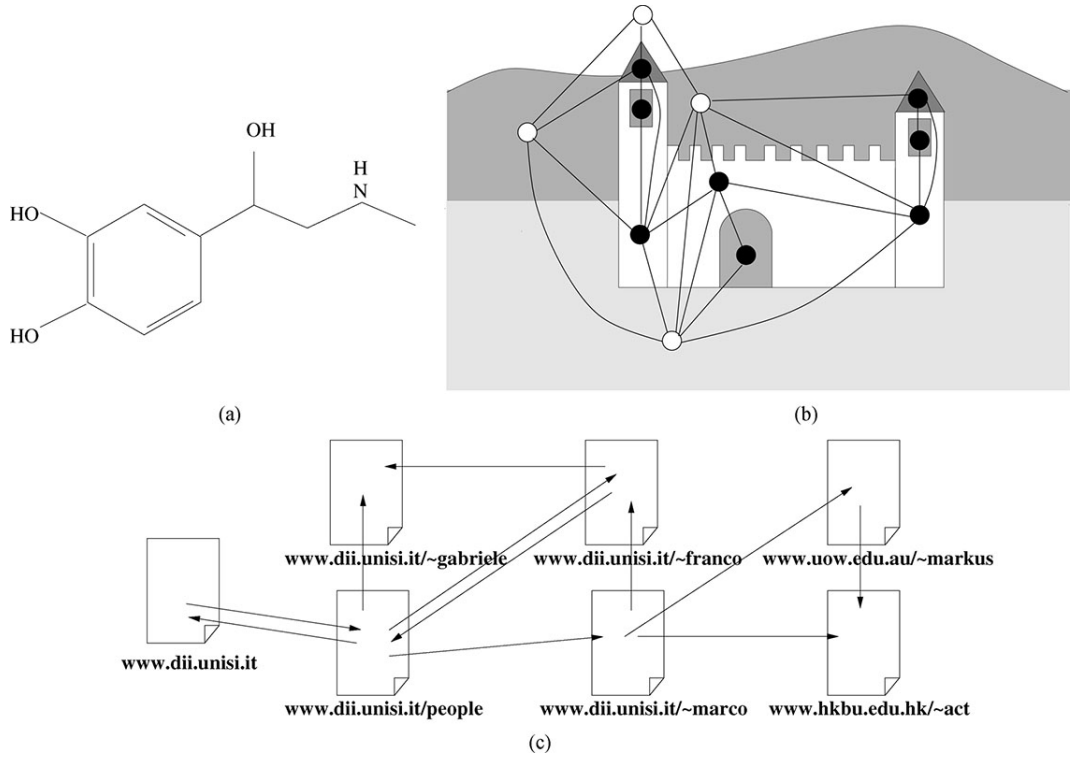


Figure 4: *Examples where the information can be represented by graphs (Scarselli et al., 2009).*

실험 모델은 4종류의 분자지문을 input으로 활용한 RF와 MuDRA 모델이다. 분자지문 4가지는 skelsphere, morgan, dragon, MACCS다. Skelsphere 분자지문과 morgan 분자지문 2가지는 동일한 원형 방법으로 각 원자를 기준으로 인접 수준별 원자와 결합이 어떠한지 원형태를 기준으로 분석하여 벡터를 생성하되 skelsphere는 1024 byte로, morgan은 2048 byte로 생성한다. Dragon 분자지문은 분자 내 원자들의 원자번호와 결합들이 어떤 결합인지와 같은 특징을 추출하여 벡터로 생성하는 방법이다. MACCS는 분자 내 특정 구조가 포함되는지 여부에 대한 분자지문 생성방법이다. 성능은 MuDRA 모델이 분자지문을 1개씩 사용한 RF모델들의 성능보다 약 10~20% 더 높은 성능을 보였다.

3. 모델

3.1. Graph convolutional networks (GCN)

Graph neural network (GNN)은 graph로 표현되는 정보를 처리하는 모델을 말한다. Graph란 각 지점으로 연결관계를 가질 수 있는 node와 각 지점간의 관계를 나타내는 edge로 구성되는 정보를 의미한다. 처음 GNN을 제안한 Scarselli 등 (2008)이 설명한 graph가 무엇인지 task를 따라 알아보겠다. Figure 4(a)는 화학물질의 유전독연변이 유발 여부 분류에 화학물질의 분자구조를 graph로 표현하였다. 각 원자를 node로, 결합을 edge로써 표현하였으며, graph인 화학물질을 분류하므로 graph-focused task이다. Figure 4(b)의 이미지 내 객체 분류로 배경과 성을 분류하는 것이다. 검정색 node는 객체가 포함되는 부분을 의미하며, 흰색 node는 배경 부분이다. edge는 이미지간 인접함을 의미한다. 이미지란 graph에서 각 node를 분류하므로 node-focused task이다.

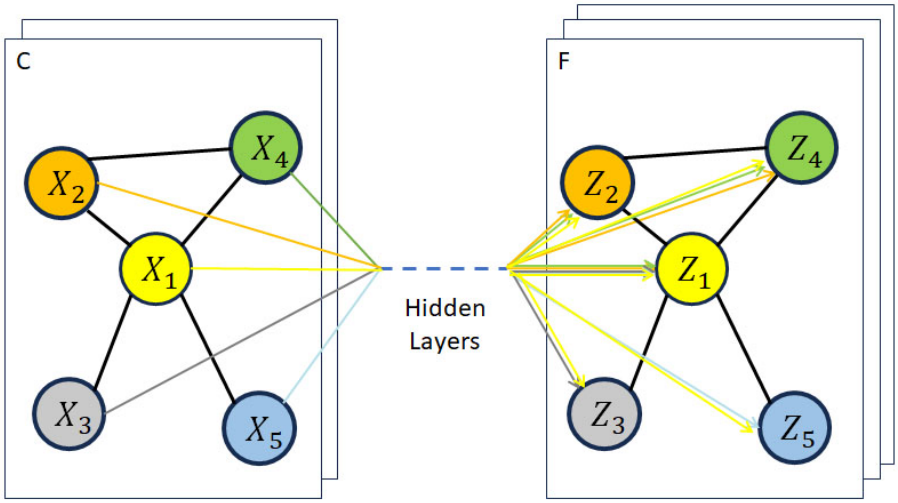
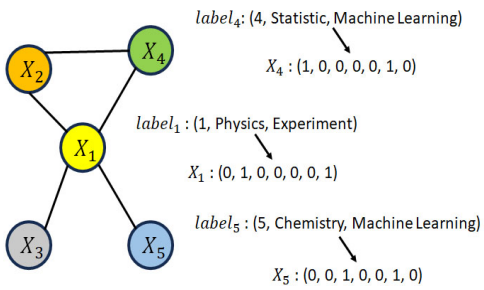


Figure 5: *Schematic depiction of GCN with C input channels and F feature maps in the output layer (Kipf and Welling, 2017).*



(a) Feature matrix

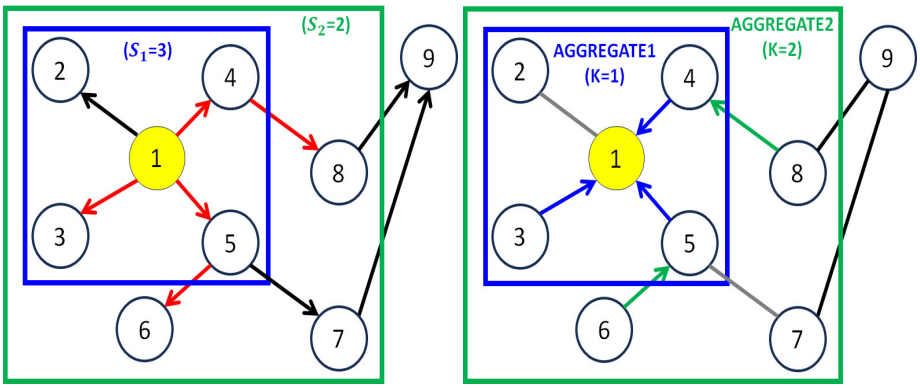
node	1	2	3	4	5
1	1	1	1	1	1
2	1	1	0	1	0
3	1	0	1	0	0
4	1	1	0	1	0
5	1	0	0	0	1

(b) Adjacency matrix

Figure 6: *Example of input of GCN in citation network.*

Figure 4(c)는 각 웹페이지의 주제를 분류하는 것으로 각 사이트 도메인이 node이며, 다른 사이트의 도메인으로 갈 수 있는 하이퍼링크가 edge가 될 수 있다. 웹페이지 네트워크에서 각 웹페이지의 도메인을 분류하는 것이므로 node-focused task이다. 관계를 활용하여 분석할 수 있는 GNN모델은 이외에도 다양한 task에 사용될 수 있어 많은 연구자들이 관심을 갖고 개발중에 있다. 대표적인 모델로 graph convolutional network (GCN)에 대해 알아보겠다.

Kipf와 Welling (2017)은 Figure 5와 같이 이미지에서 사용되는 합성곱 연산방식을 이용하는 방법을 graph에 적용함으로써 인용 네트워크(graph)에서 문서(node)의 분류 작업이 가능한 구조를 제안하였다. GCN은 그래프의 총 N개의 노드들의 각 C개의 feature에 대한 $N \times F$ 크기의 feature 행렬 X와 각 노드끼리 인접하는지 여부를 0과 1로 나타내는 $N \times N$의 인접 행렬 A를 입력으로 받는다. 그리고 각 노드별 F개의 feature map을 output으로 만드는 $Z = f(X, A)$형태의 모형이며, 인접 노드들의 은닉상태를 종합한 평균을 활용해 node



(a) Sample neighborhood (b) Aggregate feature information from neighbors

Figure 7: *GraphSAGE's sample and aggregate approach.*

들의 은닉상태를 업데이트한다. Figure 6(a)는 feature를 추출하는 과정의 예시다. 여기선 분야와 방법에 원-핫 인코더를 적용한 결과로 벡터를 생성한다고 가정한다. 그리고 Figure 6(b)는 인접행렬로, 자기 자신에게는 1이 할당되며, 2의 경우 1과 4가 인접해있으니 (1,2), (2,1), (2,4), (4,2)에 각각 1이 할당된다. l레이어를 통과한 은닉상태의 공식은 다음과 같다.

$$H^{(l)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l-1)}W^{(l)}\right) \tag{3.1}$$

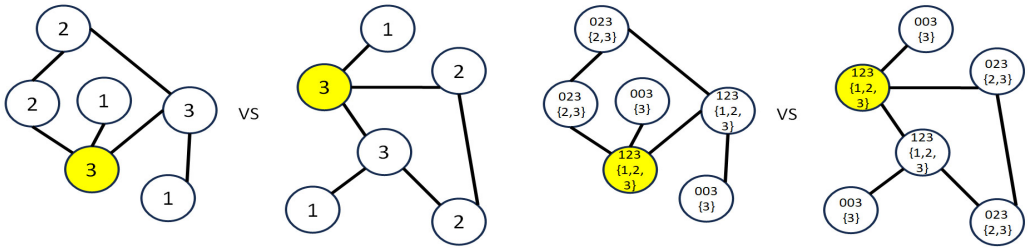
이 식에서 I_N은 graph의 node의 개수와 크기가 같은 단위행렬로, $\tilde{A} = A + I_N$인 방향성이 없는 $N \times N$의 graph G의 인접행렬이다. 그리고 $\tilde{D}_{ii} = \sum_j^N \tilde{A}_{ij}, i, j = 1, 2, \ldots, N$으로 $N \times N$인 행렬이며, $H^{(l)}$은 N개의 노드들이 l번째 레이어를 통과한 은닉상태이며, $H^{(0)}$는 노드의 input인 X와 같다. $\sigma()$은 ReLU와 같은 활성화 함수를 의미한다. d_l을 l번째 레이어의 output 크기라고 할 때, $W^{(l)}$는 l번째 레이어의 가중치 행렬로 $d_{l-1} \times d_l$의 행렬이다. $W^{(1)}$은 input_feature_size$\times d_1$크기의 행렬이다. 예시를 통해 알아보자. 만일 Figure 5가 멀티클래스 분류를 위해 2개의 GCN레이어로 구성되어 있고, graph의 node의 개수가 N개라면 식은 아래와 같이 될 것이다.

$$H^{(2)} = f(X, A) = \text{softmax}\left(\hat{A}\sigma\left(\hat{A}XW^{(1)}\right)W^{(2)}\right), \tag{3.2}$$

여기서 $\hat{A} = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$이다. $W^{(0)}$은 $C \times d_1$의 행렬이며, $W^{(1)}$은 $d_1 \times F$의 행렬이다. 그리고 $H^{(2)}$는 $N \times F$의 행렬이 된다.

3.2. Graph sample and aggregate (GraphSAGE)

Hamilton 등 (2018)은 단백질 구조 분석과 소셜네트워크와 같은 대규모의 graph에서의 node 분류에 있어 인접 노드들의 임베딩된 은닉상태를 종합하는 방법으로 모든 노드의 은닉상태를 종합하지 않고, 일부를 활용하는 GraphSAGE를 제안하였다. 모든 노드의 은닉상태를 사용하지 않음으로써 학습때 경험하지 못했던 데이터에서도 유용한 임베딩 정보를 생성할 수 있다. 이후 K-인접한 노드들 중 고정된 수만큼의 노드를 선택하고, 노드들의 은닉상태를 종합하고, 가중치 행렬에 곱한 결과를 이용한다. 이때 종합시 max-pooling을 이용한다.



(a) 1. Labeling node (b) 2. Aggregate neighborhood node's label and compare

Figure 8: *Weisfiler-Lehman test example in isomorphic case. The both graph look like difference but same result.*

모델의 계산식은 다음과 같다.

$$
\begin{aligned}
h_v^{(0)} &= x_v, \forall v \in V \\
\text{AGGREGATE}_{(k)}^{pool} &= \max\left(\left\{\sigma\left(W_{pool} h_u^{(k-1)} + b\right), \forall u \in \mathcal{N}(v)\right\}\right), \\
h_{N(v)}^{(k)} &= \text{AGGREGATE}_{(k)}^{pool}\left(\left\{h_u^{(k-1)}, \forall u \in \mathcal{N}(v)\right\}\right), \\
h_v^{(k)} &= \sigma\left(W^k \cdot \text{CONCAT}\left(h_v^{(k-1)}, h_{N(v)}^{(k)}\right)\right), \\
h_v &= h_v^{(k)} / \|h_v^{(k)}\|_2, \forall v \in V.
\end{aligned} \tag{3.3}
$$

x_v는 v번 노드의 input이며, $\mathcal{N}(v)$는 v노드를 기준으로 임의로 선택된 노드 집합이며, $h_v^{(k)}(k = 1, 2, \ldots K)$는 v 노드의 k-인접까지 종합한 은닉상태이며, W^k는 k-인접 상태에서의 완전연결층의 가중치 행렬이며, σ는 ReLU와 같은 활성화 함수이며, AGGREGATE$_{(k)}$는 k-인접에서 종합함수이다. CONCAT은 행을 유지한채 두 행렬을 이어붙이는 것을 의미하며, $N \times M$행렬 A와 $N \times K$행렬 B를 CONCAT하면 $N \times M + K$의 행렬이 만들어진다. h_v는 v노드의 최종 은닉상태이다.

Figure 7의 예시를 통해 알아보자. 1번 노드를 기준으로 인접 노드를 임의 지정하고, 종합하는 과정을 설명하겠다. Figure 7(a)에서 S_k은 k-인접 표본으로 하이퍼파라미터 값이며, 각각 인접단계마다 지정한다. 여기에선 1-인접에선 3,4,5가 지정되었고, 2-인접에서는 6과 8이 지정되었다. Figure 7(b)는 인접 노드의 은닉상태 종합단계이다. 각 노드는 1-인접 노드의 은닉상태를 종합하고 자신의 은닉상태를 업데이트 한다. $h_1^{(1)}$는 x_3, x_4, x_5을, $h_5^{(1)}$의 경우 x_1, x_6을 AGGREGATE한 다음 자신의 최초 은닉상태와 이어붙인 후 1번째 레이어의 가중치를 곱하고, 활성화 함수를 적용한다. 이후 $h_1^{(2)}$를 위해 $h_3^{(1)}, h_4^{(1)}, h_5^{(1)}$을 AGGREGATE한 후 $h_1^{(1)}$와 이어붙인다.

3.3. Graph isomorphism network (GIN)

Xu 등 (2018)은 단백질 구조 분석에 있어서 어떠한 단백질인지, 소셜 네트워크의 댓글이 어떤 종류의 댓글인지 등과 같은 graph 분류와 node 분류를 하는데 graph의 동형여부를 시험하는 방법인 Weisfiler-Lehman (WL) test를 GNN구조에 적용하여 제안하였다. WL test는 각 노드를 고유의 레이블로 생성하고, 인접 노드들의 레이블을 집계하여 이진벡터화 한다. 이진벡터의 결과가 동일하면 동형이다. 이는 GNN모델들이 인접 노드들의 메시지를 종합하여 자신을 업데이트하는 방법과 유사하다.

WL test를 동형인 graph에서 시행한 예시를 Figure 8을 활용해 설명하겠다. Figure 8(a)에서 좌우의 graph는 90도를 회전시킨 것이며, 노란색 노드 옆의 1을 밖에 두었냐 안에 두었냐만 차이가 있다. 구조적으로는 같은 것이나 이를 한눈에 같은 것이라고 알기는 어렵다. 각 노드의 label은 인접 노드의 개수로 지정했다. 그리고

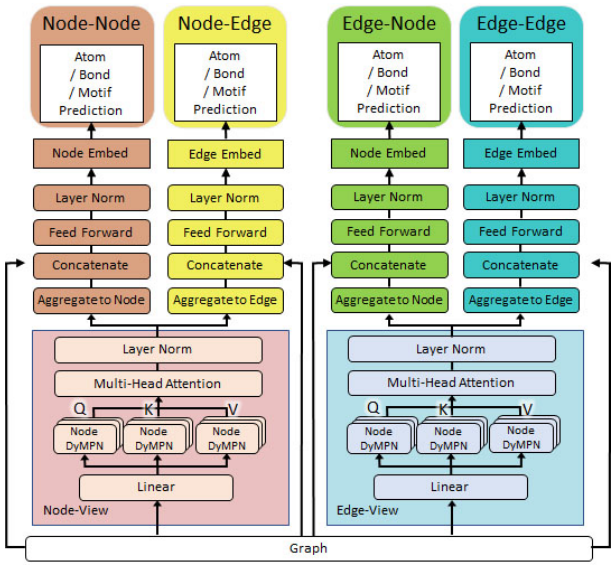


Figure 9: *GROVER's dual-GTransformer architecture.*

Figure 8(b)에서 인접 노드의 레이블을 종합하여 길이 3의 벡터로 생성하였다. 이때 인접노드가 3개 이하인 경우 0을 할당시켰다. 이 벡터들을 합치면 둘 다 298로 동일하다. 아니더라도 노드들의 레이블과 인접 노드의 레이블을 비교해도 동일한 것을 볼 수 있을 것이다.

GIN은 MLP을 통해 WL test를 구현하였다. 모델의 은닉상태의 식은 다음과 같다.

$$h_G = \text{CONCAT}\left(\text{READOUT}\left(\left\{h_v^{(k)} \mid v \in G\right\}\right) \mid k = 0, 1, \ldots, K\right),$$
$$h_v^{(k)} = \text{MLP}^{(k)}\left(\left(1 + \epsilon^{(k)}\right) \cdot h_v^{(k-1)} + \sum_{u \in \mathcal{N}(v)} h_u^{(k-1)}\right), \tag{3.4}$$

여기서 CONCAT은 행을 유지한채 두 행렬을 이어붙이는 것을 의미하며, $h_v^{(k)}$는 v노드의 k-인접 은닉상태이며, $h_v^{(0)} = x_v$이며, $\epsilon^{(k)}$은 k-인접에서의 학습 가능한 파라미터 또는 고정된 스칼라값이다. $N(v)$는 v번 노드의 1-인접한 노드들이다. $MLP^{(k)}$는 k-인접에서의 MLP의 가중치행렬이며, MLP는 완전연결층 2개로 구성된다. READOUT은 단백질 데이터에서는 합-풀링 방식을, 소셜 네트워크 데이터에서는 평균-풀링 방식을 적용한다. 풀링이란, feature 벡터에서 모든 값을 더하지 않고, column을 기준으로 연산을 적용하는 것을 의미한다. 예를 들어, 길이 3인 feature 벡터 3개를 쌓은 [(1,2,3), (2,3,4), (4,5,6)]에서 평균풀링을 하면 [(1+2+4)/3, (2+3+5)/3, (3+4+6)/3]이 결과가 된다.

3.4. Graph representation from selfsupervised message passing transformer(GROVER)

Rong 등 (2020)는 약물개발을 위해 화학물질의 다양한 특성을 예측하기 위하여 자기지도학습을 한 graph트랜스포머 모델을 공개하였다. Figure 9는 GROVER의 구조이다. 일반적인 트랜스포머 구조와 다르게 node와 edge에 대해 각각을 기준으로 분석하는 2개의 트랜스포머가 병렬로 구성되어 있다. 화학물질의 특성예측시 일반적으로 원자(노드)를 중심으로 정보를 종합하는 방법을 사용하는데, 노드(원자)와 엣지(결합)의 두가지 관점을 동시에 분석하는 것으로써 분자의 특성을 모델에 적용시켰다고 볼 수 있다. 모델구조 중 dynamic message passing network (DyMPN)는 인접 노드들의 고정된 i-hop 대신 에포크마다 정규분포나 균등분포로

부터 i값을 임의로 추출하여 보다 더 향상된 일반화 성능을 보여준다. 정규분포는 평균은 3, 표준편차는 1이 기본값이며, 균등분포는 0 ~ 6이 기본값이다.

pretext task는 화학물질 내 임의의 원자와 결합쌍을 15% 가리고 예측하는 contextual property prediction 과 화학물질 내에 벤젠고리와 같이 자주 나오고, 주요 역할을 하는 motif 85가지가 몇개씩 포함되는지 여부를 맞추는 것이다. contextual property란 1-인접 원자까지의 원자와 결합의 정보다. 예를 들어, 임의의 원자가 C일 때, 주변에 C가 단일결합 되어있고, O가 이중결합 되어있다면, C_C-SINGLE1_O-DOUBLE1이 된다. 이러한 1-인접한 원자까지의 원자, 결합쌍을 가리고 맞추는 것이다.

pretext task를 위한 계산식들에 대해 알아보자. 먼저 v번 node의 은닉상태 h_v 계산식은 다음과 같다.

$$
\begin{aligned}
h_v^{(0)} &= W_{Linear} x_v, \\
h_v^{(i)} &= \sigma\left(W_{DyMPN}^{(i)} \sum_{u \in N(v)} h_u^{(i-1)} + b^{(i)} \right), \\
\text{Attention}(Q, K, V) &= \text{softmax}\left(QK^T / \sqrt{d} \right) V, \\
\text{head}_j &= \text{Attention}\left(QW_j^Q, KW_j^K, VW_j^V \right), \\
h_v &= \text{CONCAT}\left(x_v, \text{head}_1, \ldots, \text{head}_J \right) W_{Feed}.
\end{aligned} \tag{3.5}
$$

W_{Linear}는 Figure 9의 linear레이어의 가중치이며, x_v는 v번 node의 input이다. $h_v^{(i)}$는 v번 node의 i-인접에서의 은닉상태이며, $N(v)$는 v노드의 1-인접 노드 집합을 의미하며, $W_{DyMPN}^{(i)}$와 $b^{(i)}$는 DyMPN의 i-인접에서의 가중치와 bias이다. σ는 활성화함수 PReLU이다. Q, K, V는 각 DyMPN의 output이다. head_j는 j번째 head값이며, W_j^Q는 j번째 head의 query에 대한 가중치이다. h_v는 v번 노드가 feed forward network를 통과한 node의 최종 은닉상태이며, CONCAT은 행을 유지한채 두 행렬을 이어붙이는 것을 의미하며, W_{Feed}는 feed forward network의 가중치이다.

이를 통해 구해진 h_v를 이용하여 가려진 v번째 node가 C개의 contextual property 사전에서 c번째에 해당될 확률 $p_{v,c}$예측할 수 있으며, node의 은닉상태들로 graph의 은닉상태 h_G를 만드는 READOUT 레이어를 통해 graph에 85개 중 m번째 motif가 몇 개 포함되어있는지 여부인 q_m을 예측할 수 있다. 그리고 이 관점별로 예측결과가 다른 것에 대한 패널티 또한 loss로 적용시켜서 서로 같은 결과를 낼 수 있도록 학습을 한다. 관점별 차이에 불일치 계수를 곱하여 loss로 만들어 손실함수에 더한다. 손실함수에 대한 식은 다음과 같다.

$$
\begin{aligned}
\hat{p}_{v,c} &= \text{softmax}(W_{contextual} h_v), \\
h_G &= \text{READOUT}(\{h_v \mid v \in G\}), \\
\hat{q}_m &= W_{motif} h_G, \\
L_{contextual} &= -\sum_{v=1}^{N} \sum_{c=1}^{C} p_{v,c} \log(\hat{p}_{v,c}), \\
L_{motif} &= \| q_m - \hat{q}_m \|_2, \\
L_{disagreement} &= \eta_{dist} \Bigg[\sum_{v=1}^{N} \sum_{c=1}^{C} \{ (\hat{p}_{v,c,node-node} - \hat{p}_{v,c,node-edge})^2 + (\hat{p}_{v,c,edge-node} - \hat{p}_{v,c,edge-edge})^2 \} \\
&\quad + \sum_{m=1}^{85} \{ (\hat{q}_{m,node-node} - \hat{q}_{m,node-edge})^2 + (\hat{q}_{m,edge-node} - \hat{q}_{m,edge-edge})^2 \} \Bigg] \\
L &= L_{contextual} + L_{motif} + L_{disagreement},
\end{aligned} \tag{3.6}
$$

Table 1: Atom and bond feature list of GROVER

Type	Feature name	Description
Atom	Atom number	Atomic number
	Formal charge	Integer electronic charge assigned to atom
	Number of bonds	Number of bonds the atom is involved in
	Chirality	Whether this atom has mirror image like rotations, translations
	Number of H	Number of bonded hydrogen atoms
	Atomic mass	Mass of the atom, divided by 100
	Aromaticity	Whether this atom is part of an aromatic system
	Hybridization	Whether include sp, sp2, sp3, sp3d, or sp3d2
Bond	Bond type	Single, double, triple, or aromatic
	Stereo	Whether the bond is isomer (none, any, E/Z or cis/trans)
	In ring	Whether the bond is part of a ring
	Conjugated	Whether the bond is conjugated

여기서 $W_{contextual}$과 W_{motif}는 각 task의 마지막 레이어의 가중치를 의미하며, READOUT은 평균-풀링을 시행한다. y_v와 y_m은 각 task의 label을 의미하며, N는 임의로 가려진 노드의 개수를 의미한다. η_{dist}는 불일치계수로 0 ~ 1 사이의 값을 하이퍼파라미터로써 지정한다. $p_{v,node-node}$는 node-node관점에서의 v번째 node의 예측값이며, $q_{m,node-edge}$는 node-edge에서의 m번째 motif에 대한 예측 값이다.

GROVER는 또한 다양한 원자/결합의 feature를 추출하여 활용했다. 일반적으로 원자 번호나 결합의 형태에 1,2가지를 feature로 사용하는 경우가 많으나, 해당 모델에서는 보다 다양한 feature를 사용함으로써 더 다양한 원자/결합의 특성을 학습할 수 있던 것으로 생각된다. Table 1은 GROVER에서 사용한 원자와 결합의 feature 리스트이며, 원자의 질량을 제외한 feature들은 one-hot 인코더를 적용한 후 이어붙여서 벡터로 만든다. 결합의 feature는 방향성을 고려하여 2개를 생성한다. 1번 원자에서 2번 원자로 가는 결합(e_{12})의 경우 1번 원자의 feature 벡터에 결합으로부터 나온 feature 벡터를 이어붙인다. 반대 방향의 e_{21}은 2번 원자의 feature 벡터에 결합으로부터 나오는 특징을 이어붙여서 만든다.

3.5. Motif-based graph self-supervised learning (MGSSL)

Zhang 등 (2021)은 분자의 위상학적 정보를 이용한 사전학습 방법을 제시하였다. Figure 10에서와 같이 MGSSL은 GIN구조를 여러층 쌓아서 인코더로 활용하다. 인코더는 GCN, GraphSAGE 등 다른 GNN계열 모델로 변경할 수 있다. 그리고 pretext task를 위한 GRU가 디코더 역할을 한다. pretext task는 분자구조를 고리 구조와 고리 옆에 홀로 붙은 측쇄 구조 등을 중심으로 분할될 수 있는 subgraph인 motif단위로 분할하여 트리구조를 만들고, 이 트리구조를 처음부터 끝까지 다음 motif가 존재할지와 어떠한 motif인지를 맞추는 것이다. 이 분자의 위상학적 정보를 가리고 맞추는 것은 분자의 핵심 구조인 motif를 기반으로 분자를 해석할 수 있음을 의미하며, 분자의 구조를 이해하는데 도움이 되어 사전학습시 AUC가 최소 5% 향상되었다. MGSSL에서는 원자의 feature는 원자 번호와 비대칭성을 사용했으며, 결합의 feature는 결합형태와 방향을 사용했다.

MGSSL의 사전학습에는 motif의 은닉상태가 필요하다. 화학물질이 총 I개의 motif로 나뉘어진다고 할 때, i번째 motif의 은닉상태 x_i는 motif에 해당되는 모든 노드들의 은닉상태 h_v를 합침으로써 만들어진다.

$$x_i = \sum_{v \in motif(i)} h_v \tag{3.7}$$

motif(i)는 i번째 motif의 원자 집합을 의미한다. 그리고 t시점에서 i 번째 motif에서 j번째 motif에 대한 예측을

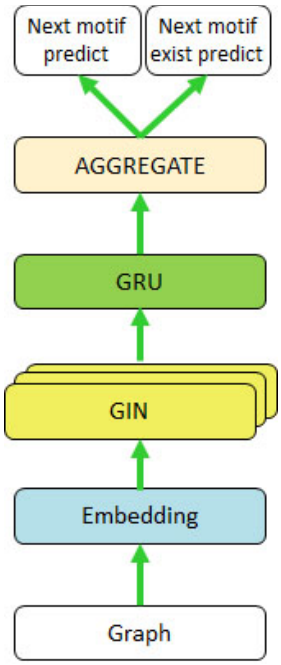


Figure 10: *MGSSL's architecture.*

하는데 사용되는 은닉상태 $h_{i,j}$의 계산에는 GRU가 사용되며, 식은 다음과 같다.

$$
\begin{aligned}
s_{i,j} &= \sum_{k,i\in\hat{\varepsilon}_t,\ k\neq j} h_{k,i}, \\
r_{k,i} &= \sigma(W^r x_i + U^r h_{k,i} + b^r), \\
z_{i,j} &= \sigma(W^z x_i + U^z s_{i,j} + b^z), \\
\tilde{h}_{i,j} &= \tanh\left(W^h x_i + U^h \sum_{k,i\in\hat{\varepsilon}_t} r_{k,i} \odot h_{k,i}\right), \\
h_{i,j} &= (1 - z_{i,j}) \odot s_{i,j} + z_{i,j} \odot \tilde{h}_{i,j},
\end{aligned}
\tag{3.8}
$$

여기서 $\hat{\varepsilon}_t$는 t시점에서의 예측된 motif들의 집합이며, k는 t시점에서 motif i의 인접 motif들 중 j를 제외한 인접 motif들과 같다. $s_{i,j}$는 t시점에서의 i 번째 motif의 j번째 motif를 제외한 나머지 motif들의 은닉상태의 합이며, $r_{k,i}$는 GRU에서 t-1시점까지의 정보를 적절하게 잊게하는 reset gate를 통과한 은닉상태이며, $z_{i,j}$는 GRU에서 t-1시점과 t시점의 정보를 활용해 update를 하는 update gate를 통과한 은닉상태이다. W^z, W^r, W^h은 각각 GRU에서 reset gate와 update gate과 GRU의 마지막 은닉층에서의 현재 은닉상태에 대한 가중치이며, U^z, U^r, U^h은 GRU의 reset gate와 update gate과 마지막 은닉층에서의 이전 시점의 은닉상태에 대한 가중치이다. $\odot$은 아다마르 곱으로 원소간 곱을 의미한다. $\tilde{h}_{i,j}$는 t시점에서의 candidate 은닉상태이다.

앞에서 나온 은닉상태들을 활용하여 t시점에서 다음 motif의 존재 확률인 p_t를 예측한다. 예측결과 motif가 존재할 경우 j번째 motif가 크기가 M인 motif사전에서 m번째 motif일 확률인 $q_{j,m}$를 예측한다. 이 예측

Table 2: MoleculeNet datasets

Category	Dataset	Number of task	Number of data
Classification	BACE	1	1,513
	BBBP	1	2,039
	Clintox	2	1,478
	SIDER	27	1,427
	Tox21	12	7,831
	ToxCast	617	8,576
Regression	ESOL	1	1,128
	FreeSolv	1	642
	Lipo	1	4,200
	QM7	1	6,800
	QM8	12	21,786

결과와 실제 결과간에 크로스 엔트로피 함수를 적용하여 loss를 계산한다. 식은 다음과 같다.

$$\begin{aligned}
\hat{p}_t &= \text{sigmoid}\left(U_{exist} \cdot \tanh\left(W_{exist-1} x_i + W_{exist-2} \sum_{k, i \in \hat{\epsilon}_t} h_{k,i}\right)\right), \\
\hat{q}_{j,m} &= \text{softmax}(U_{motif} \cdot \tanh(W_{motif} h_{i,j})), \\
L_{exist}(p_t, \hat{p}_t) &= -p_t \log(\hat{p}_t), \\
L_{motif}(q_j, \hat{q}_j) &= -\sum_{m=1}^{M} q_{j,m} \log(\hat{q}_{j,m}), \\
L &= \sum_{t=1}^{T} L_{exist}(p_t, \hat{p}_t) + \sum_{j=1}^{J} L_{motif}(q_j, \hat{q}_j).
\end{aligned} \tag{3.9}$$

U_{exist}, $W_{exist-1}$, $W_{exist-2}$는 다음 motif가 존재하는지 예측하는 레이어들의 가중치다. U_{motif}와 W_{motif}는 다음 motif가 어떤것인지 예측하는 레이어들의 가중치다.

4. 데이터셋

우리가 성능비교에 사용한 데이터는 MoleculeNet에서 제공하는 데이터셋 11개와 OECD TG 데이터셋 6개이다. 기존 연구에서도 MoleculeNet을 사용했으나 AUC만을 지표로 삼았다. 우리는 여기에 정확도, 정밀도, 재현율, F1 점수를 추가하여 다시 비교해보았다. 그리고 OECD TG 데이터는 OECD가 유럽화학물질청(ECHA)과 협력하여 만든 eChemPortal.org에서 공개된 데이터를 활용하여 구축한 데이터셋이다.

4.1. MoleculeNet datasets

본 연구에 사용된 MoleculeNet 데이터셋은 Table 2와 같이 분류 6개와 회귀 5개로 총 11개이다. 분류 데이터셋은 독성과 관련이 있는 데이터 위주이다. 다운로드는 https://moleculenet.org/datasets-1에서 가능하다.

먼저 분류 데이터셋들에 대해 알아보겠다. BACE 데이터셋은 총 1,513개의 분자로 구성되어 있으며, 타겟변수는 사람에게 치매와 관련이 있는 것으로 알려진 BACE-1(β-secretase) 단백질 효소의 억제제로써 효과가 있는지에 대한 이진분류 데이터셋입니다. 데이터들은 약 10년 동안 과학적 문건으로 보고된 실험값이다.

BBBP 데이터셋은 총 2,039개의 분자로 구성되어 있으며, 타겟변수는 뇌의 세포외액과 혈액을 분리하는 막을 투과할 수 있는지에 대한 이진분류 데이터셋이다. 이 막은 대부분의 약물, 호르몬, 신경 전달 물질들을

차단한다. 그러나 이를 투과할 수 있는 경우 중추신경계에 작용할 수 있는 신경계 부작용을 일으킬 수 있는 물질이다.

Clintox 데이터셋은 총 1,478개 약물 화합물에 대한 타겟변수가 2개인 이진분류 데이터셋이다. 1번째는 임상 실험 결과 독성 유무이며, 2번째는 FDA 승인 여부에 대한 레이블이다. 1번과 2번은 서로 음의 상관관계를 가진다.

SIDER데이터셋은 총 1,427개의 시중에 판매되고 있는 약물의 부작용에 대한 이진분류 데이터셋이다. DeepChem에서 수집되었으며, 심장 질환과 안구 질환 등 27가지의 약의 부작용에 분류된 데이터셋이다.

Tox21은 7,831개의 화학물질의 12개의 독성에 대한 이진분류 데이터셋이다. 2014 Tox21 데이터 챌린지에서 사용된 데이터베이스이며, 핵수용체 작용 7가지와 스트레스 반응 5가지에 대해 다룬다. 핵수용체는 인슐린 감수성 수용체 페릭소솜, 호르몬 균형 유지 수용체 아로마타제 등이 포함되며, 스트레스 반응은 비정상 세포를 발견 및 파괴하는 p53단백질과 항산화 반응 세포 등이 포함된다.

ToxCast데이터셋은 총 8,576개의 분자와 617개의 독성에 대한 이진분류 데이터셋이다. Tox21과 동일하게 미 정부기관 주도하에 생성된 데이터베이스이며, Tox21의 12가지 독성을 포함하며, 기타 다양한 과학적 실험 결과를 포함한다. 그러나 label에 null값이 매우 많은 데이터셋이다.

다음으로 회귀 데이터셋들에 대해 알아보겠다. ESOL은 1,128개의 화학물질의 로그 수용성 능력(log(mol / liter))에 대한 회귀 데이터셋이다.

Free solvation database (FreeSolv)는 642개의 화학물질에 대한 이온 1몰이 수화작용할 때 방출되는 에너지의 양(kcal/mol)에 대한 회귀 데이터셋이다. 이는 화학물질 및 단백질의 특성을 예측하는 SAMPL 블라인드 예측 챌린지의 데이터셋의 서브셋이다.

Lipophilicity (Lipo)는 분자의 옥탄올이 풍부한 상태에서의 분자의 농도 대비 물이 풍부한 상태에서의 분자의 농도의 분포 계수를 D라 할때, 용매의 농도가 pH7.4일때의 로그 D값 (logD at pH 7.4)을 타겟변수로 하는 회귀 데이터셋이다. 이 데이터셋은 의약품 분야 화학물질 DB인 ChEMBL 데이터베이스에서 추출한 데이터셋이며, 4,200개의 화학물질이 포함된다.

QM7은 양자역학 분야에서 화학물질의 모든 원자를 완전 분리하는데 필요한 원자화 에너지를 계산한 결과값이 타겟변수인 회귀 데이터셋이다. 단위는 kcal / mol이며, GDB-13 데이터베이스의 부분데이터셋으로 6,800개의 화학물질을 포함한다.

QM8은 원자들이 S0 → S1, S2로 전이시 에너지 준위와 진동자 세기에 대한 4가지 정보를 양자역학분야의 3가지 계산방법을 따라 각각 계산한 12가지 결과값에 대한 회귀 데이터셋이다. 이는 GDB-17 데이터베이스의 부분데이터셋이며, 21,786개의 화학물질이 포함된다.

4.2. OECD TG 데이터셋

OECD TG는 화학물질이 인간의 건강과 환경에 미치는 잠재적 영향을 평가하기 위한 실험방법에 대한 안내서이다. 국제적으로 안전성 검사를 위한 표준 방법으로 받아들여지고 있으며, 이 test guideline은 정부기관, 산업현장, 학술적인 기준이 된다 (OECD, 1994). 우리는 Bae 등 (2021)과 같이 eChemPortal에서 OECD TG 데이터를 수집하였다. 수집한 카테고리는 급성 독성, 생식 및 발달 독성, 유전 독성 3종류의 데이터를 수집하였다. 수집된 데이터의 화학물질에 대한 ID가 미국 화학회(American chemical society)에서 구분하기 위해 만든 chemical abstracts service (CAS) number로 되어 있어 SMILES식으로 변경하기 위한 데이터 전처리를 하였다.

1) Chemical identifier resolver (CIR) 파이썬 라이브러리를 활용하여 각 화학물질의 CAS number를 SMILES식으로 변환
2) 결합이 없어 graph로 표현이 불가능한 이온 상태의 물질 제거

Table 3: OECD TG dataset

Category	Endpoint	Name	Number of data	etc
Acute oral toxcity	LD50	TG423	1,577	Multi-class
Toxicity to reproduction	Developmental toxicity	TG414	1,301	
	Developmental toxicity	TG422	908	
	Reproductive toxicity	TG422_repro	1,149	
Genotoxicity	Genotoxicity	TG471	3,563	
	Genotoxicity	TG473	2,175	
	Genotoxicity	TG474	1,286	

3) Graph 정보 추출이 불가능한 SMILES식을 제거하였다.

데이터 전처리 결과는 Table 3와 같다. 이어서 각 독성 분야와 데이터셋의 세부내용을 알아보겠다.

4.2.1. 급성 경구 독성(TG423)

TG423은 섭취시 사망에 이르게 하는 치사량에 대한 실험결과 데이터셋이다. 독성 경구 시험은 종말점이 LD50이며, 실험동물에게 물질을 1회 투여 후 50%의 동물이 사망하게 되는 단위 체중 당 화학물질의 중량(mg/kg)을 나타낸다. 관찰기간은 14일이다.

투약 용량은 국제 조화 화학물질 및 혼합물 분류 체계 (globally harmonized system; GHS)의 분류기준 4개 등급 중 3번째 용량을 시작으로 LD50에 포함되는지 여부에 따라 다음 실험 용량을 정하기 때문에 멀티클래스로 바꾸었다. 분류기준은 class 1은 5mg/kg, class 2는 50mg/kg, class 3은 300mg/kg, class 4는 2,000mg/kg이다 (OECD, 2002).

4.2.2. 생식 및 발달 독성(TG414, 422)

생식기능, 능력 또는 태아 발생 발육에 유해한 영향을 주는지 등 생식과 발달에 대한 독성 실험결과 데이터셋이다. 실험동물은 주로 랫드, 마우스를 이용한다. TG414는 착상일로부터 분만 하루전까지 투여한 후, 모동물의 체중, 임신성립여부, 황체수 및 착상수를 조사하고, 태아는 생사여부, (사망시)사망시기, 체중, 성별, 변형, 기형, 골격과 연부조직의 변화 등의 이상여부를 확인한다 (OECD, 2018).

TG422는 반복 투여 및 생식 독성 시험으로 63일간 매일 반복 투여함으로써 체내에 화학물질이 누적될 경우 체중, 음식/물 소비, 발정 주기, 갑상선 호르몬, 후대들의 전반적인 상태에 영향을 미치는지 관찰한다. TG422는 일반적인 이상여부와 생식 이상여부 두가지를 동시에 실험한다. TG422는 일반적인 이상여부를, TG422_repro는 생식 이상여부를 기록한 결과로 나누었다 (OECD, 2016a).

4.2.3. 유전 독성(TG471, 473, 474)

DNA, 세포, 염색체에 영향을 주는지 독성을 갖고 있는지에 대한 실험결과 데이터셋이다. TG471은 박테리아 복귀 돌연변이 시험 또는 Bruce Ames에 의해 개발되어 Ames test라고 불리는 실험 결과이다. 필수 아미노산을 합성하지 못하는 상태의 살모넬라 티피뮤리움과 대장균 최소 5개 균주를 이용하여 필수 아미노산을 합성할 수 있는 균주로 전환되게 하는 복귀 돌연변이를 일으키는 물질인지 여부에 대한 실험이다 (OECD, 2020).

TG473는 시험관 내 포유류 염색체이상시험으로 배양된 포유류 동물의 체세포에서 구조적 염색체 이상을 일으키는 인자를 확인한다. 세포는 햄스터의 난소 또는 폐의 조직 또는 사람 또는 기타 포유류 말초 혈액 림프구 등을 사용한다 (OECD, 2016b).

TG474는 포유류 적혈구 소핵 실험으로 쥐와 같은 포유류 동물의 골수 또는 말초 혈액 세포에서 수집된

Table 4: Hyperparameter list

Hyperparameter	Value	Model
Dropout rate	0 ~ 0.5	All
Learning rate	0.0001 ~ 0.002	MGSSL, GIN, GCN, GraphSAGE
Embedding dense	300 ~ 1300	GIN, GCN, GraphSAGE
Embedding learning rate	0.5 ~ 1.5	GIN, GCN, GraphSAGE
Learning rate	0.0001 ~ 0.0007	GROVER
Dense	300 ~ 1300	GROVER
Dense layer number	2 5	GROVER
Dist coefficient rate	0.05, 0.1, 0.15	GROVER
Bond dropout rate	0, 0.2, 0.4, 0.6	GROVER
Attention out size	4, 8	GROVER

적혈구를 주로 사용한다. 화학물질 투여 후 염색체 또는 적혈구에 손상이 있었는지 여부를 확인 한 결과이다 (OECD, 2016c).

5. 방법

5.1. 화학물질의 특징 추출

화학물질의 특징추출 방법에 대해 알아보겠다. 분자구조는 원자와 결합으로 이루어져 있으나 일반적 표기방법은 각 원자가 몇개 이루어져있는지만 표기한다. 아세트아미노펜을 예시로 설명하자면 분자식은 C8H9NO2이다. 그러나 이런 표기법은 원자의 개수는 같아도 다른 구조나 성질을 가진 이성질체가 존재할 수도 있다. 그렇기 때문에 기계학습시 사용되는 분자표기법 분자간 구분이 명확해야한다. 기계학습에 주로 사용되는 표기법은 원자 순서와 결합 상태를 기준으로 표기해주는 SMILES식이다. 이 SMILES식으로부터 구조를 인식하여 정보를 추출해 분자지문과 graph로 만들기위해 화학분야 라이브러리인 RDKit을 활용할 수 있다 (Landrum, 2012).

SMILES식은 분자를 각 원자와 결합상태에 따라 순서대로 표기를 하는 방법이다. 수소와 단일결합은 보통 생략하고, 각 원자는 기호나 번호로, 결합은 = (이중결합), #(3중결합) 등 고유의 상태별로 표기를 한다. 아세트아미노펜의 일반적인 표기법은 C8H9NO2이며, SMILES식은 CC(= O)NC1 = CC = C(C = C1)O이다.

분자지문은 분자내에 특정 분자 구조가 있는지를 벡터로써 표현하는 하거나 분자구조가 어떤 순서로 구성되어있는지에 따라 분석하여 이진벡터로 표현하는 방법 등 다양한 방법을 사용할 수 있다. 여기서는 구조기반 방법과 원형 방법에 대해 설명하겠다. 구조기반 방법은 분자구조를 기반으로 분자 구조에 어떤 원자 또는 구조가 포함되는지를 벡터로써 표현한다. 대표적으로는 MACCS가 있다 (Durant 등, 2002). 원형 방법은 분자 구조가 어떤 순서로 구성되어있는지를 경로를 원형 단위로 분석하여 벡터로 표현한다. 원형 방법은 대표적으로 ECFP가 있다 (Rogers와 Hahn, 2010).

GNN에서는 Smiles식으로부터 각 원자(또는 결합)별로 원자번호나 결합형태와 같은 feature를 추출하여 각 node와 edge의 feature로 사용한다. 예를 들면, 앞의 아세트아미노펜의 첫번째 원자는 탄소이다. 탄소의 원자번호는 12로, 원자번호는 주기율표 기준 118까지 있어 이를 원-핫 인코딩을 통해 벡터로 표현할 수 있다. 그리고 다음 원자는 탄소이며, 단일 결합으로 되어있다. 결합에 대해서 인접행렬도 만들어지지만, 결합 종류가 3종류이니 이 또한 원-핫 인코딩을 통해 벡터로 표현할 수도 있으며, 결합이 방향성 고리에 포함되는지 여부 등 다양한 정보를 이어서 같이 벡터에 포함시킬 수 있다. 이렇게 각 원자와 결합에 대한 feature를 추출하여 사용하는 방법이 분자지문으로 추출하는 방법보다 높은 성능을 발휘한다 (Devenaud 등, 2015).

Table 5: Atom/bond feature list

Atom features	Size	Description	Model
Atom type	100	Atom number like periodic table	All
Chirality	5	Like mirror, atom has same position in other	All model
Formal charge	5	Integer electronic charge assigned to atom	Except MGSSL
Number of bonds	6	Number of bonds the atom is involved in	Except MGSSL
Number of H	5	Number of bonded hydrogen	Except MGSSL
Atomic mass	1	Mass of atom and divided by 100	Except MGSSL
Aromaticity	1	If this atom is part of aromatic	Except MGSSL
Hybridization	5	sp, sp2, sp3, sp3d, or sp3d2	Except MGSSL
Bond features	**Size**	**Description**	**Model**
Bond type	4	Single, double, triple, or aromatic	All model
Stereo	6	Isomer of bond like any, E/Z or cis/trans	Except MGSSL
In ring	1	Whether the bond is part of a ring	Except MGSSL
Conjugated	1	Whether the bond is conjugated	Except MGSSL
Bond direction	1	Direction of bond	Only MGSSL

5.2. 실험 방법

우리는 MoleculeNet의 실험방법을 준용하여 실험했다. 데이터 분할시 train:valid:test 비율을 8 : 1 : 1로 지정하였으며, test시 valid loss가 가장 낮은 체크포인트의 가중치를 활용하여 성능을 측정했다. 성능 지표는 100 에포크씩 서로 다른 seed 값을 적용하여 실험을 3번 반복하여 나온 성능지표의 평균과 표준편차를 활용한다. seed값을 달리하여 3번 실험한 이유는 seed 값을 바꾸면 데이터 분할 결과가 달라지기 때문에 일반화성능을 확인하는데 용이하기 때문이다. 데이터분할은 분자의 핵심구조인 scaffold를 기반으로 분할하는 방법을 사용하였다. 이 방법은 분자의 핵심구조를 기준으로 데이터셋을 분할하는 방법이다. 모든 분자의 핵심구조에 대한 학습을 하지 않고, 일부를 남겨 학습되지 않은 핵심구조에 대한 성능을 검증, 평가함으로써 모델의 일반화 성능을 확인하는데 도움이 되는 challenge한 방법이다 (Wu 등, 2018). 지표는 AUC에 추가로 정확도, 재현율, 정밀도, F1점수까지 활용할 것이다. 주요 지표는 AUC와 F1점수이며, 나머지 정확도, 재현율, 정밀도는 부가지표이다.

모델은 앞서 소개한 GROVER, MGSSL, GIN, GCN, GraphSAGE 5종을 활용할 것이다. 옵티마이저는 Adam을 사용하였다. 하이퍼파라미터 튜닝은 랜덤서치를 실시하였다. 하이퍼파라미터 리스트는 Table 4에 있다. GROVER와 기타 모델간 학습률이 다른 이유는 GROVER는 상대적으로 모델규모가 커서 학습률을 크게하면 쉽게 발산하기 때문이다. 그리고 분자의 graph 생성시 feature는 사전학습 모델은 원본 모델의 feature 추출방법을 이용하였으며, GIN, GCN, GraphSAGE는 GROVER와 동일한 feature 추출방법을 활용하였다. MGSSL은 노드 엣지별로 2종류만 사용한 반면 GROVER는 더 다양하여 높은 성능이 발휘될 것이 기대되었기 때문이다. 사용된 feature는 Table 5와 같다.

6. 결과

MoleculeNet과 OECD TG 데이터를 활용하여 5가지 모델을 학습한 결과이다. 분류에서의 성능지표는 MoleculeNet에서 제시한 AUC 뿐만 아니라 정확도, 정밀도, 재현율, F1점수를 포함한 5가지를 확인하였다. 회귀에서는 MoleculeNet에서 제시한 ESOL, FreeSolv, Lipo는 RMSE를, QM7과 QM8은 MAE를 적용했다.

Table 6: MoleculeNet classification dataset results

Dataset	Model	AUC	Accuracy	Recall	Precision	F1
BACE	GROVER	**0.894(0.029)**	**0.818(0.025)**	**0.840(0.063)**	0.778(0.039)	**0.807(0.046)**
	MGSSL	0.844(0.013)	0.785(0.019)	0.774(0.006)	**0.815(0.030)**	0.794(0.015)
	GIN	0.783(0.026)	0.711(0.023)	0.749(0.078)	0.720(0.011)	0.732(0.037)
	GCN	0.835(0.016)	0.752(0.006)	0.770(0.073)	0.771(0.034)	0.767(0.023)
	GraphSAGE	0.800(0.022)	0.730(0.030)	0.720(0.041)	0.763(0.040)	0.740(0.029)
BBBP	GROVER	**0.940(0.019)**	**0.896(0.030)**	**0.954(0.012)**	**0.912(0.035)**	**0.932(0.017)**
	MGSSL	0.689(0.026)	0.621(0.030)	0.806(0.013)	0.608(0.024)	0.693(0.020)
	GIN	0.677(0.013)	0.605(0.016)	0.840(0.031)	0.589(0.009)	0.692(0.016)
	GCN	0.665(0.019)	0.603(0.007)	0.765(0.027)	0.598(0.010)	0.671(0.005)
	GraphSAGE	0.682(0.009)	0.637(0.008)	0.750(0.015)	0.633(0.010)	0.686(0.005)
Clintox	GROVER	**0.944(0.021)**	0.940(0.007)	0.567(0.094)	0.639(0.235)	0.581(0.134)
	MGSSL	0.742(0.029)	0.917(0.010)	0.536(0.061)	0.634(0.127)	0.566(0.083)
	GIN	0.920(0.024)	0.948(0.013)	0.642(0.072)	0.817(0.122)	0.694(0.092)
	GCN	0.933(0.009)	0.934(0.011)	0.742(0.024)	0.767(0.060)	0.750(0.032)
	GraphSAGE	0.935(0.006)	**0.960(0.008)**	**0.762(0.027)**	**0.868(0.025)**	**0.803(0.020)**
SIDER	GROVER	**0.658(0.023)**	0.766(0.006)	**0.685(0.012)**	**0.655(0.021)**	**0.655(0.010)**
	MGSSL	0.608(0.013)	0.754(0.007)	0.591(0.030)	0.594(0.014)	0.583(0.014)
	GIN	0.632(0.007)	0.740(0.004)	0.544(0.013)	0.593(0.018)	0.554(0.011)
	GCN	0.638(0.005)	**0.768(0.009)**	0.613(0.033)	0.596(0.005)	0.594(0.021)
	GraphSAGE	0.615(0.010)	0.751(0.004)	0.572(0.007)	0.593(0.009)	0.574(0.007)
Tox21	GROVER	**0.831(0.025)**	**0.912(0.009)**	**0.271(0.053)**	**0.661(0.063)**	**0.368(0.064)**
	MGSSL	0.743(0.004)	0.905(0.001)	0.130(0.010)	0.516(0.022)	0.185(0.004)
	GIN	0.754(0.007)	0.906(0.001)	0.145(0.015)	0.494(0.039)	0.204(0.015)
	GCN	0.755(0.004)	0.908(0.001)	0.192(0.011)	0.589(0.015)	0.261(0.014)
	GraphSAGE	0.752(0.008)	0.906(0.002)	0.200(0.012)	0.547(0.014)	0.271(0.016)
ToxCast	GROVER	**0.737(0.010)**	**0.838(0.005)**	**0.266(0.016)**	**0.429(0.029)**	**0.299(0.016)**
	MGSSL	0.607(0.007)	0.800(0.002)	0.143(0.004)	0.235(0.021)	0.144(0.002)
	GIN	0.631(0.013)	0.798(0.003)	0.172(0.026)	0.307(0.009)	0.190(0.027)
	GCN	0.647(0.009)	0.804(0.001)	0.179(0.011)	0.302(0.019)	0.190(0.009)
	GraphSAGE	0.646(0.008)	0.801(0.002)	0.166(0.028)	0.281(0.021)	0.180(0.024)

6.1. MoleculeNet 분류 데이터셋 결과

Table 6는 MoleculeNet 벤치마크 분류 데이터셋에서의 학습결과로 GROVER가 전반적으로 가장 좋은 성능을 발휘하였다. 특히 BBBP에서는 AUC와 F1점수 모두 다른 모델 대비 30%가량 높은 성능의 큰 차이를 보였다. 그리고 Tox21과 ToxCast에서는 AUC는 약 10%이상 높았으며, F1점수는 30%이상 더 높은 경향을 보인다. 그러나 Clintox에서는 AUC만 높고, 나머지 지표에서는 GraphSAGE가 가장 높은 성능을 발휘하였다. 2번째 성능을 발휘한 모델을 비교하자면 BACE와 BBBP에서는 MGSSL이 AUC과 F1에서 2번째로 높은 성능이었다. SIDER는 GCN이 2번째로 높은 성능이었으며, Tox21에서는 GCN과 GraphSAGE가 비슷하게 2,3위를 하였다. ToxCast는 GCN이 2번째로 높은 성능을 보였다. 그러나 모든 모델이 Clintox, Tox21과 ToxCast에서는 양성 예측 지표인 F1점수가 AUC의 보다 비교적 낮은 것을 볼 수 있다. 이는 해당 데이터셋들의 불균형비가 심하기 때문으로 보인다.

Table 7: OECD testguideline dataset results

Dataset	Model	AUC	Accuracy	Recall	Precision	F1
TG414	GROVER	**0.796(0.037)**	**0.715(0.054)**	**0.804(0.071)**	**0.729(0.089)**	**0.757(0.038)**
	MGSSL	0.614(0.025)	0.588(0.019)	0.582(0.120)	0.632(0.012)	0.600(0.058)
	GIN	0.598(0.009)	0.573(0.027)	0.531(0.096)	0.623(0.011)	0.569(0.064)
	GCN	0.604(0.026)	0.557(0.023)	0.521(0.105)	0.608(0.019)	0.555(0.059)
	GraphSAGE	0.656(0.021)	0.588(0.045)	0.427(0.106)	0.692(0.036)	0.522(0.084)
TG422	GROVER	**0.917(0.083)**	**0.899(0.041)**	**0.725(0.112)**	**0.899(0.057)**	**0.800(0.092)**
	MGSSL	0.626(0.025)	0.678(0.032)	0.300(0.072)	0.530(0.073)	0.376(0.068)
	GIN	0.594(0.035)	0.663(0.014)	0.122(0.031)	0.477(0.093)	0.191(0.038)
	GCN	0.645(0.012)	0.689(0.005)	0.100(0.0)	0.700(0.071)	0.175(0.002)
	GraphSAGE	0.630(0.016)	0.681(0.016)	0.267(0.054)	0.539(0.054)	0.353(0.049)
TG422_repro	GROVER	0.536(0.083)	0.592(0.004)	0.202(0.055)	0.311(0.028)	0.242(0.049)
	MGSSL	0.584(0.044)	0.588(0.029)	0.135(0.030)	0.361(0.060)	0.191(0.025)
	GIN	0.562(0.037)	0.629(0.008)	**0.294(0.041)**	0.485(0.021)	**0.365(0.038)**
	GCN	**0.626(0.014)**	**0.655(0.015)**	0.230(0.063)	**0.561(0.046)**	0.323(0.072)
	GraphSAGE	0.467(0.029)	0.577(0.015)	0.175(0.030)	0.343(0.041)	0.231(0.034)
TG423	GROVER	**0.829(0.014)**	**0.841(0.003)**	**0.399(0.026)**	**0.454(0.057)**	**0.400(0.025)**
	MGSSL	0.732(0.036)	0.774(0.003)	0.337(0.005)	0.329(0.049)	0.310(0.014)
	GIN	0.753(0.026)	0.762(0.012)	0.331(0.005)	0.412(0.026)	0.309(0.002)
	GCN	0.789(0.009)	0.766(0.010)	0.341(0.009)	0.362(0.029)	0.324(0.013)
	GraphSAGE	0.737(0.043)	0.762(0.018)	0.383(0.079)	0.325(0.065)	0.337(0.057)
TG471	GROVER	**0.964(0.025)**	**0.951(0.015)**	**0.721(0.073)**	**0.947(0.037)**	**0.816(0.038)**
	MGSSL	0.753(0.019)	0.882(0.004)	0.194(0.026)	0.773(0.118)	0.306(0.022)
	GIN	0.822(0.014)	0.907(0.004)	0.333(0.017)	0.923(0.029)	0.490(0.022)
	GCN	0.804(0.017)	0.873(0.009)	0.326(0.010)	0.557(0.070)	0.409(0.011)
	GraphSAGE	0.780(0.017)	0.877(0.008)	0.375(0.0)	0.567(0.050)	0.451(0.016)
TG473	GROVER	**0.807(0.023)**	**0.891(0.013)**	0.184(0.008)	**0.755(0.035)**	0.296(0.009)
	MGSSL	0.673(0.015)	0.867(0.015)	0.173(0.019)	0.370(0.094)	0.230(0.011)
	GIN	0.740(0.040)	0.876(0.000)	0.253(0.068)	0.427(0.020)	0.314(0.062)
	GCN	0.769(0.016)	0.876(0.010)	0.240(0.033)	0.447(0.078)	0.307(0.019)
	GraphSAGE	0.755(0.003)	0.888(0.009)	**0.280(0.033)**	0.528(0.070)	**0.365(0.043)**
TG474	GROVER	**0.780(0.120)**	**0.946(0.019)**	0.222(0.157)	0.093(0.069)	0.130(0.094)
	MGSSL	0.722(0.024)	0.907(0.004)	0.133(0.047)	0.306(0.039)	0.182(0.048)
	GIN	0.769(0.022)	0.910(0.010)	0.267(0.047)	**0.387(0.088)**	**0.315(0.059)**
	GCN	0.633(0.040)	0.871(0.004)	**0.300(0.0)**	0.237(0.009)	0.265(0.006)
	GraphSAGE	0.640(0.027)	0.907(0.006)	0.200(0.0)	0.340(0.047)	0.251(0.013)

6.2. OECD TG데이터셋 결과

Table 7의 OECD TG 데이터셋에서의 결과를 보면, 생식 및 발달 독성인 TG414에서는 GROVER가 타 모델 성능 대비 약 20% 가량 더 높은 성능을 발휘했으며, TG422에서는 AUC는 40%가량 더 높았고, F1점수는 2배 이상 높았다. 그러나 TG422_repro에서는 GCN이 좋은 성능을 발휘했다. 급성 독성 데이터셋이자 multi-class 데이터셋인 TG423에서는 GROVER가 전반적으로 가장 좋은 성능을 발휘했다. 성능지표는 특정 클래스를 양성, 나머지 클래스를 음성으로하여 계산하는 one-vs-rest 방식으로 계산하였으며, 클래스들의 평균을 결과로 활용하였다. 가장 다음으로는 AUC를 기준으로 GCN이 2번째, F1을 기준으로는 GraphSAGE가 2번째였다.

유전독성에서도 GROVER가 전반적으로 가장 좋은 성능을 발휘하였다. 특히 TG471에서는 다음 모델들보다 AUC는 약 18%이상 더 높았으며, F1점수는 60%이상 더 높았다.

TG473에서는 AUC는 GROVER가 가장 높았지만, F1점수는 GraphSAGE가 가장 높았다. TG474에서는 AUC는 GROVER가 가장 높았지만, F1점수는 GIN이 가장 높고, GROVER가 가장 낮았다. 특히 AUC와 정확도를 제외한 성능의 표준편차가 크다. 이유는 TG474는 클래스 불균형비가 1 : 19로 양성 데이터가 너무 적어서 test셋에 양성 데이터가 4 ~ 7개로 매우 적게 들어갔다. 학습시 양성데이터가 적어서 대부분을 음성으로 예측하기 쉽고, 1개를 맞추냐 못맞추냐에 따라 표준편차가 매우 크기 때문에 이러한 특이한 결과가 나온 것으로 보인다. TG474에서는 GIN이 가장 좋은 성능을 발휘했다고 본다.

6.3. MoleculeNet 회귀 데이터셋 결과

Table 8의 MoleculeNet 회귀 데이터셋 결과에서는 GROVER가 큰 차이로 가장 좋은 성능을 발휘하였다. 2위는 GraphSAGE가 3회, GCN이 2회를 하였다.

실험결과 가장 최신 SOTA 모델인 GROVER가 대부분의 데이터셋에서 가장 좋은 성능을 발휘하였다. 그러나 Clintox, TG422의 생식 데이터셋, TG474 4개에서는 가장 좋은 성능을 발휘하진 않았다. 이는 이진 분류 데이터셋이 간단한 모형에서도 좋은 성능을 발휘할 수 있는 것과 데이터의 불균형비가 원인으로 보인다. 특히 TG474의 경우 불균형비는 19 : 1이며, 3회동안의 평균 confusion matrix값은 true positive 0.67, false positive 1.33 false negative 5.67 true negative 122.33이었다.

MGSSL과 GIN, GCN, GraphSAGE들은 성능이 전반적으로 유사하다만 그래도 GraphSAGE와 GCN이 상대적으로 성능이 높은 것으로 보인다. MoleculeNet 분류 데이터셋에서 GraphSAGE가 Clintox에서 AUC는 2위, 나머지는 1위를 하였으며, GCN은 Clintox, Tox21, Toxcast에서 2위의 성능을 보였다. MGSSL은 BACE와 BBBP에서는 괜찮은 성능을 보였지만, multi-task인 나머지 데이터셋들에서는 저조한 성능을 보인 경우가 많다. 회귀에서는 GraphSAGE가 3회 2위, GCN이 2회 2위를 하였다. OECD TG데이터에서는 GIN이 좋은 결과를 많이 보였다. TG471에선 거의 2위를 하였고, TG474에서는 종합적으로 볼 때 좋은 결과를 보여줬다.

MGSSL과 GIN은 성능이 비슷한 것이 흥미로운 점이다. MGSSL은 사전학습을 했으나, 원자와 결합의 feature를 각각 2가지만 사용한 반면 GIN은 사전학습은 안하되 GROVER에서 사용한 보다 다양한 feature를 사용하였다. 사전학습 여부와 feature 변경이 동일한 수준의 성능 향상을 보여준다고 볼 수 있을 것이다.

7. 결론

본 연구에서는 화학물질을 graph로 표현하여 화학물질의 특성을 예측할 수 있는 5가지 모델로 실제 동물실험 결과 데이터인 OECD TG 데이터에서 5개의 다양한 지표를 통해 성능을 비교해보았다. 기존의 MoleculeNet은 다양한 화학물질의 특성의 일반적인 성능을 비교해볼 수 있었지만, 화학물질의 독성 여부에 대해서는 OECD TG가 기준이 되기 때문에 OECD TG의 3가지 독성에 대한 8종류의 데이터를 수집하여 독성 예측 성능을 비교하였다. 결과 SOTA모델인 GROVER가 대부분의 데이터셋에서 가장 좋은 성능을 발휘했지만, MoleculeNet의 Clintox에서는 AUC는 높으나 실제 양성 예측 성능은 낮았으며, OECD TG에서는 8개 데이터셋 중 5개에서는 가장 좋은 성능이었으나 나머지 3개는 다른 모델이 가장 좋은 성능을 발휘하였다. 실험 결과를 통해 또 GIN과 MGSSL의 성능이 비슷함을 통해 사전학습만큼이나 유용한 feature 선정이 중요한 것을 알 수 있었다.

또 연구에서 결과에 중요한 영향을 미쳤던 불균형에 대한 해결이 필요함을 볼 수 있었다. 이로 인해 양성 데이터를 거의 학습하지 못하여 F1점수가 많이 낮음을 볼 수 있었다. 그렇기 때문에 TG474의 경우 양성 예측 성능이 많이 안좋았음을 볼 수 있다. 이번 연구에서는 아쉽게도 불균형 해소기법을 다루지 못하고 원본 모델의 성능만을 갖고 비교를 하였다. 향후에는 화학물질의 특성을 예측하는 graph 기반 모델에 적용하기 좋은 불균형 해소기법을 적용하여 성능을 향상시킬 수 있는 연구를 수행하고자 한다.

Table 8: MoleculeNet regression dataset results

Dataset	Model	RMSE/MAE
ESOL	GROVER	**0.831**(0.120)
	MGSSL	1.118(0.048)
	GIN	1.138(0.018)
	GCN	1.159(0.057)
	GraphSAGE	1.102(0.056)
FreeSolv	GROVER	**1.544**(0.397)
	MGSSL	2.855(0.321)
	GIN	2.822(0.318)
	GCN	2.244(0.085)
	GraphSAGE	2.618(0.117)
Lipo	GROVER	**0.560**(0.024)
	MGSSL	0.761(0.003)
	GIN	0.753(0.036)
	GCN	0.723(0.006)
	GraphSAGE	0.730(0.018)
QM7	GROVER	**72.6**(3.8)
	MGSSL	95.246(1.350)
	GIN	108.830(14.410)
	GCN	91.954(2.914)
	GraphSAGE	83.683(5.686)
QM8	GROVER	**0.0125**(0.002)
	MGSSL	0.0350(0.0002)
	GIN	0.0357(0.001)
	GCN	0.0355(0.0007)
	GraphSAGE	0.0348(0.001)

References

Alves VM, Golbraikh A, Capuzzi SJ *et al.* (2018). Multi-descriptor read across (MuDRA): A simple and transparent approach for developing accurate quantitative structure–activity relationship models, *Journal of Chemical Information and Modeling*, **58**, 1214–1223.

Bae SY, Lee J, Jeong J, Lim C, and Choi J (2021). Effective data-balancing methods for class-imbalanced genotoxicity datasets using machine learning algorithms and molecular fingerprints, *Computational Toxicology*, **20**, 100178.

Bahdanau D, Cho K, and Bengio Y (2014). Neural machine translation by jointly learning to align and translate, Available from: arXiv preprint arXiv:1409.0473

Bemis GW and Murcko MA (1999). Properties of known drugs. 2. side chains, *Journal of Medicinal Chemistry*, **42**, 5095–5099.

Coley CW, Barzilay R, Green WH, Jaakkola TS, and Jensen KF (2017). Convolutional embedding of attributed molecular graphs for physical property prediction, *Journal of Chemical Information and Modeling*, **57**, 1757–1772.

Durant JL, Leland BA, Henry DR, and Nourse JG (2002). Reoptimization of MDL keys for use in drug discovery, *Journal of Chemical Information and Computer Sciences*, **42**, 1273–1280.

Duvenaud DK, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru-Guzik A, and Adams RP (2015). Convolutional networks on graphs for learning molecular fingerprints, *Advances in Neural Information Processing Systems*, **28**, 2224–2232.

Gilmer J, Schoenholz SS, Riley PF, Vinyals O, and Dahl GE (2017). Neural message passing for quantum chemistry, *International Conference on Machine Learning*, **70**, 1263–1272. PMLR.

Hamilton W, Ying Z, and Leskovec J (2017). Inductive representation learning on large graphs, *Advances in Neural Information Processing Systems*, **30**, 1024–1034.

Kipf TN and Welling M (2016). Semi-supervised classification with graph convolutional networks, Available from: arXiv preprint arXiv:1609.02907

OECD (1994). OECD Guidelines for the Testing of Chemicals. OECD, Available from: https://www.oecd-ilibrary. org/environment/oecd-guidelines-for-the-testing-of-chemicals_72d77764-en

OECD (2002). Test No. 423: Acute Oral toxicity - Acute Toxic Class Method, OECD Guidelines for the Testing of Chemicals, Section 4, OECD Publishing, Paris, Available from: https://doi. org/10.1787/9789264071001-en

OECD (2016a). Test No. 422: Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test, OECD Guidelines for the Testing of Chemicals, Section 4, OECD Publishing, Paris, https://doi.org/10.1787/9789264264403-en

OECD (2016b). Test No. 473: In Vitro Mammalian Chromosomal Aberration Test, OECD Guidelines for the Testing of Chemicals, Section 4, OECD Publishing, Paris, Available from: https://doi.org/10.1787/97892642646 49-en

OECD (2016c). Test No. 474: Mammalian Erythrocyte Micronucleus Test, OECD Guidelines for the Testing of Chemicals, Section 4, OECD Publishing, Paris, Available from: https://doi.org/10.1787/9789264264762-en

OECD (2018). Test No. 414: Prenatal Developmental Toxicity Study, OECD Guidelines for the Testing of Chemicals, Section 4, OECD Publishing, Paris, Available from: https://doi.org/10.1787/9789264070820-en

OECD (2020). Test No. 471: Bacterial Reverse Mutation Test, OECD Guidelines for the Testing of Chemicals, Section 4, OECD Publishing, Paris, Available from: https://doi.org/10.1787/9789264071247-en

Landrum G (2012). Fingerprints in the RDKit, Available from: https://www.rdkit.org/UGM/2012/Landrum_RDKit_UGM.Fingerprints.Final.pptx.pdf

Landrum G (2019). Rdkit documentation, Available from: https://buildmedia.readthedocs.org/media/pdf/rdkit/latest/rdkit.pdf

Luechtefeld T, Marsh D, Rowlands C, and Hartung T (2018). Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility, *Toxicological Sciences*, **165**, 198–212.

Mayr A, Klambauer G, Unterthiner T, and Hochreiter S (2016). DeepTox: Toxicity prediction using deep learning, *Frontiers in Environmental Science*, **3**, 80.

National Research Council (2007). *Toxicity Testing in the 21st Century: A Vision and a Strategy*, National Academies Press, Washington D.C.

Rogers D and Hahn M (2010). Extended-connectivity fingerprints, *Journal of Chemical Information and Modeling*, **50**, 742–754.

Rong Y, Bian Y, Xu T, Xie W, Wei Y, Huang W, and Huang J (2020). Self-supervised graph transformer on large-scale molecular data, *Advances in Neural Information Processing Systems*, **33**, 12559–12571.

Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, and Monfardini G (2008). The graph neural network model,

IEEE Transactions on Neural Networks, **20**, 61–80.

Schütt K, Kindermans PJ, Sauceda Felix HE, Chmiela S, Tkatchenko A, and Müller KR (2017). Schnet: A continuous-filter convolutional neural network for modeling quantum interactions, *Advances in Neural Information Processing Systems*, 30.

Silva AC, Borba JV, Alves VM *et al.* (2021). Novel computational models offer alternatives to animal testing for assessing eye irritation and corrosion potential of chemicals, *Artificial Intelligence in the Life Sciences*, **1**, 100028.

Vaswani A, Shazeer N, Parmar N *et al.* (2017). Attention is all you need, *Advances in Neural Information Processing Systems*, **30**, 5998–6008.

Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswingd K, and Pande V (2018). MoleculeNet: A benchmark for molecular machine learning, *Chemical Science*, **9**, 513–530.

Xu K, Hu W, Leskovec J, and Jegelka S (2018). How powerful are graph neural networks?, Available from: arXiv preprint arXiv:1810.00826

Zhang Z, Liu Q, Wang H, Lu C, and Lee CK (2021). Motif-based graph self-supervised learning for molecular property prediction, *Advances in Neural Information Processing Systems*, **34**, 15870–15882.

Received January 17, 2024; Revised January 26, 2024; Accepted January 27, 2024

OECD TG데이터를 이용한 그래프 기반 딥러닝 모델 분자 특성 예측

황대환[a], 임창원[1,a]

[a]중앙대학교 응용통계학과

요 약

본 연구에서는 OECD test guideline 데이터를 이용하여 graph기반 딥러닝 모델들의 성능을 비교하고자 한다. OECD TG는 화학물질들이 인체와 환경에 미칠 잠재적 영향에 대해 시험하는 방법이며, 많은 실험이 동물실험을 통해 독성을 확인한다. 동물실험은 많은 시간과 비용이 들며, 윤리적 이슈가 있어 대안을 찾거나 최소화하는 방법들이 연구되고 있다. 딥러닝은 화학물질을 활용하는 다양한 분야에서 사용되고 있으며, 독성예측 분야에도 사용되고 있으며, 특히 graph 기반 모델에 대한 연구가 활발하다. 우리의 목표는 OECD TG 데이터에 대한 graph기반 딥러닝 모델들의 성능을 비교하여 가장 성능이 좋은 모델을 찾는 것이다. 우리는 OECD에서 운영하는 웹사이트 eChemportal.org에서 OECD TG를 따른 결과를 수집하였으며, 전처리 과정을 통해 학습이 불가능하거나 부적절한 화학물질은 제거하였다. 수집된 OECD TG데이터와 화학물질 특성 예측 성능의 벤치마크 데이터셋인 MoleculeNet 데이터를 활용하여 5개의 graph기반 모델들의 독성 예측 성능을 비교하였다.

주요용어: 독성 예측, 그래프 신경망, OECD TG, 딥러닝

이 논문은 2021년도 중앙대학교 연구장학기금 지원에 의한 것임

[1]교신저자: (06974) 서울시 동작구 흑석로 84, 중앙대학교 응용통계학과. E-mail: clim@cau.ac.kr