공학석사 학위논문

Attention 과 fingerprint 을 활용한 분자 특성 예측을 위한 그래프 뉴럴 네트워크

2022 년 8 월

서울대학교 대학원
컴퓨터공학부
정재헌

Attention 과 fingerprint 을 활용한 분자 특성 예측을 위한 그래프 뉴럴 네트워크

지도 교수 문봉기

이 논문을 공학석사 학위논문으로 제출함

2022 년 8 월

서울대학교 대학원

컴퓨터공학부

정재헌

정재헌의 공학석사 학위논문을 인준함

2022 년 8 월

위 원 장	권 태 경	(인)
부위원장	문 봉 기	(인)
위 원	이 영 기	(인)

초 록

신약 개발에서 머신러닝은 분자 생성(molecule generation), 분자 특성 예측(molecular property prediction) 등 여러 분야에서 활용되어 왔다. 특히 분자 특성 예측은 drug discovery 분야에서 중요한 역할을 한다. 신약에 필요한 특성을 갖을 것이라고 예상되는 약물을 잘 선정하고 그렇지 못할 것이라고 예상되는 약물은 잘 걸러 냄으로써 더 빠르고 저렴한 방법으로 전체 프로세스를 크게 가속화할 수 있다. 현재 신약 개발 평균 비용이 약 28 억 달러로 추정된다는 점을 고려할 때, 분자 특성 예측을 통해 시행착오를 줄이는 일이 얼마나 중요한지 체감할 수 있다.

전통적으로는 subgraph 을 분석한 fingerprint, 물리적 규칙에 기반한 hand-engineered features, DFT(Density Functional Theory) 등을 통해 분자 특성을 예측해왔다. 하지만 최근 AI 가 다양한 분야에서 눈부신 발전을 했다. 분자 특성 예측에서도 딥러닝(Deep learning)을 통해 분자 특성을 예측하려는 많은 시도들이 있었고 기존의 방식들보다 좋은 결과를 내왔다. 특히 많은 연구들이 Graph neural network 기반의 모델들이 효과적임을 보여왔다. 분자는 원자와 원자 간의 결합으로 이루어지는데, 그것은 그래프에서 vertex 와 edge 로 잘 표현할 수 있기 때문이다.

본 연구는 기존의 분자 특성 예측 전략을 분석하고 새로운 전략을 제안한다. 자연어처리 등 다양한 분야에서 활발하게 쓰이고 있는 attention 에 관하여, molecular property prediction

task 에 적합한 attention 을 분석하고 graph neural network 에 적용한다. 그리고 일반적으로 작용기가 분자의 특성과 연관성이 있다는 것을 활용하기 위해 fingerprint 을 사용한다. 본 연구에서 fingerprint 종류에 대해 분석하고 graph neural network 에 결합하여 사용하며 실험을 통해 비교 분석한다.

다양한 데이터셋을 활용한 실험을 통해서 본 연구에서 제안한 분석 전략이 baseline 모델들과 비교했을 때 competitive 한 결과를 보이는 것을 확인할 수 있었다.

주요어 : 분자 특성 예측, Deep learning, Drug discovery, Graph neural network, Attention, Fingerprint
학 번 : 2020-25719

목 차

제 1 장 서 론 1
제 1 절 연구의 배경 1
제 2 절 연구의 내용 3

제 2 장 관련 연구 4
제 1 절 SMILES 4
1.1 SMILES 의 규칙 4
1.2 SMILES 의 단점 6
제 2 절 Graph neural netwrok 7
2.1 Graph 7
2.2 Graph convolutional network 7
2.3 Attention 9
제 3 절 Fingerprint 10

제 3 장 모델 14
제 1 절 GCN with Attention 14
1.1 Feature matrix 14
1.2 Attention 15
제 2 절 Molecular fingerprint 15
제 3 절 모델 15

제 4 장 실험 및 평가 17
제 1 절 Datasaet 17
제 2 절 실험 환경 17
제 3 절 실험 결과 18
3.1 GCN block with attention 18
3.2 Attention 18
3.3 fingerprint 19
3.4 Baseline 과 비교 20

제 5 장 결론 22
참고문헌 23
Abstract 25
Acknowledgements 27

표 목차

[표 1] 사용한 데이터셋 17
[표 2] 실험 환경 17
[표 3] QM9 Dataset 에 대한 성능 비교 20
[표 4] CEP, ZINC dataset 에 대한 성능 비교 21

그림 목차

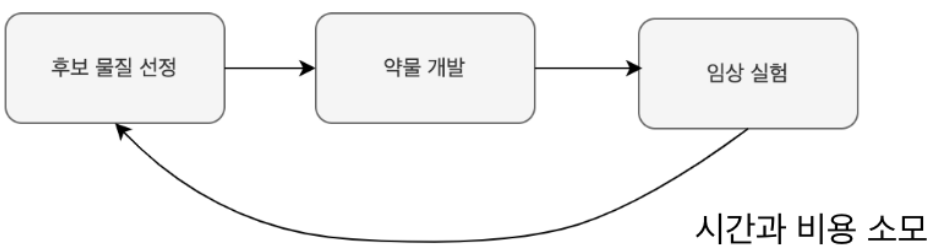
[그림 1] Molecular property prediction 1
[그림 2] 약물 개발 과정 1
[그림 3] GNN 활용 분야 3
[그림 4] Dioxane 분자의 SMILES 표기법 4
[그림 5] Pyridine 분자의 SMILES 표기법 5
[그림 6] Fluoroform 분자의 SMILES 표기법 5
[그림 7] SMILES 만드는 과정 6
[그림 8] 그래프의 표현 7
[그림 9] Adjacency metrix(NxN), Feature matrix(NxF) 8
[그림 10] graph 에서 attention 이 적용되는 과정 10
[그림 11] 하이드록시기, 카복시기 10
[그림 12] Structural fingerprint 12
[그림 13] Hashed fingerprint 12
[그림 14] Path-based fingerprint 에서 subgraph 집합을 구하는 예시 13
[그림 15] Circular fingerprint 의 substructure 예시 13
[그림 16] 모델의 전체적인 구조 16
[그림 17] GCN with attention block 수에 따른 성능 비교 18
[그림 18] Attention 에 따른 성능 비교 19
[그림 19] Fingerprint 종류에 따른 성능 비교 20

제 1 장 서 론

제 1 절 연구의 배경

신약을 개발하는 과정은 크게 봤을 때, 후보 물질 선정, 약물 발굴 및 구조 최적화, 임상 실험으로 나눌 수 있다. 첫번째 과정인 후보 물질 선정은 개발하고자 하는 약물의 특성을 갖을 것이라고 예측되는 약물들을 선정하고 그렇지 못할 것이라고 예측되는 약물들을 screening 하는 과정이다. 이 과정은 매우 중요하다. 원하는 특성을 갖을 것이라고 기대했던 물질을 선정해서 개발한 후에 임상 실험에서 원하는 효과를 거두지 못하거나 생각지 못했던 side effect 을 관찰하는 경우, 다시 개발 초기 단계로 돌아가야 하기 때문이다. 문제가 생길 때마다 다시 처음으로 돌아가는 과정이 반복되면서 시간적 경제적 손실이 막대하게 발생하게 된다. 그렇기 때문에 신약을 개발하는 과정에서 시간을 단축시키고 비용을 절감할 수 있도록 후보 물질을 효과적으로 선정하고 screening 하는 일은 매우 중요하다.

[그림 1] Molecular property prediction



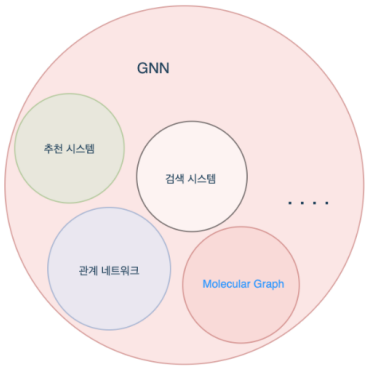
[그림 2] 약물 개발 과정

후보 물질을 효과적으로 선정하고 screening 하기 위해서는 실험을 하지 않고도 화학 구조를 통해 분자의 특성을 예측할 수 있어야 한다. 이를 Molecular property prediction 이라고 하며 drug discovery 분야에서 기본 과제로 여겨져 왔다. 전통적으로는 subgraph 을 분석한 fingerprint, 물리적 규칙에 기반한 hand-engineered features, DFT(Density Functional Theory) 등을 통해 분자 특성을 예측해왔다[1, 2, 3]. 최근에는 AI 의 발전으로 Deep learning 을 활용한 Molecular property prediction method 들이 연구되어 왔고 좋은 성과를 보여왔다.

분자 데이터를 머신러닝 모델의 입력으로 사용할 때 다양한 형태로 사용될 수 있는데, 그 중 graph 형태로 표현하여 입력으로 사용하는 연구들이 가장 좋은 성과를 보여왔다[4, 5, 6, 7, 8]. 분자는 원자와 원자 간의 결합으로 구성되는데 이는 각각 컴퓨터 공학 자료구조에서 graph 의 vertex 와 edge 에 대응시킬 수 있기 때문에 분자를 가장 잘 표현할 수 있기 때문이다.

본 연구는 기존의 분자 특성 예측 전략을 분석하고 새로운 전략을 제안한다. 특히 다양한 분야에서 state-of-the-art 을 달성하고 있는 attention 이 그래프 데이터에서도 쓰이는데, molecular graph prediction task 에 적합한 attention 을 분석해 본 연구의 graph neural network 에 적용한다.

그래프 뉴럴 네트워크가 사용되는 분야는 다양하다. 그 중 하나가 Molecular property prediction 이다. 본 연구에서는 다양한 분야에서 사용되는 GNN 과 더불어 Molecular property prediction 에만 추가적으로 사용될 수 있는 feature 사용한다. 유기화학에서 분자들의 특징적인 화학 반응을 담당하는 부분을 뜻하는 작용기가 일반적으로 분자의 특성과 관련이 크다는 점을 이용하기 위해 fingerprint 을 graph neural network 와 함께 사용한다. 본 연구에서는 fingerprint 종류에 대해 분석하고 실험에서 비교한다.



[그림 3] GNN 활용 분야

제 2 절 연구의 내용

본 논문의 구성은 다음과 같다. 2 장에서는 관련 연구인 SMILES, Graph neural network, Attention, Fingerprint 에 대해서 소개한다. 3 장에서는 본 연구를 통해 제안하는 실질적인 모델 구조에 대해서 소개한다. 4 장에서는 다양한 Dataset 에서의 실험을 통해 해당 모델과 baseline 모델과의 비교, 분석 및 평가를 제시한다. 마지막으로 5 장은 각각 결론을 기술한다.

제 2 장 관련 연구

제 1 절 SMILES

SMILES(Simplified Molecular Input Line Entry System)[9]은 WLN, ROSDAL 등과 마찬가지로 분자의 구조를 문자열로 나타내는 방법 중 하나로 1986 년에 도입된 방법이다. SMILES 는 쉽게 분자의 구조를 문자열로 나타낼 수 있어 널리 쓰이고 있다. 사실상 분자를 다루는 모든 dataset 에서 분자를 표현하기 위해 SMILES 을 활용하여 표기하고 있다. 이것은 우리가 사용해야 하는 데이터의 1 차적인 형태가 SMILES 라는 것을 의미한다.

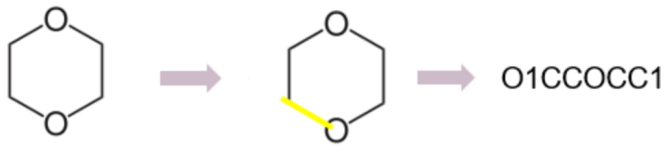
1.1 SMILES 규칙

SMILES 의 크게 5 가지 구성요소가 있다.

1. 원자(atom) - 표준 원소 기호(C, O, N, CL 등)로 나타낸다. 수소 원자는 생략한다.

2. 결합(bond) - 기본적으로 이웃한 원자는 인접해서 쓴다. 단일결합(생략), 2 중결합(“=”), 3 중 결합(“#”)으로 나타낸다.

3. 고리(ring) - 원자들의 결합이 cycle 을 있음을 나타낸다. 고리는 임의로 한 지점의 결합을 끊고, 해당 끊긴 부분의 원자 두개의 번호를 표시하는 방식으로 표기한다. 다음은 Dioxane 의 예시이다.

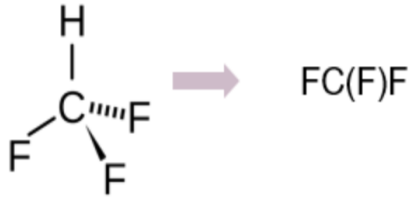


[그림 4] Dioxane 분자의 SMILES 표기법

3. 방향족(aromaticity ring) – 탄소화합물이 평면의 고리 형태로 결합하여 안정적인 구조를 가지는 aromatic ring 을 포함하고 있는 것을 말한다. 이곳에 포함된 원자는 소문자로 표시한다.

[그림 5] Pyridine 분자의 SMILES 표기법

4. 가지(branch) – 괄호 “()”로 표현한다. 2 번에서 기본적으로 이웃한 원자는 인접해서 쓴다고 했지만 한 원자에 2 개 이상의 원자가 결합되어 있는 경우 해당 규칙을 만족할 수가 없다. 따라서 “가지”라는 개념을 사용하고 괄호 “()”로 표현한다. 이 때, 괄호 안에 포함된 첫번째 원자와 괄호가 끝나고 나오는 첫번째 원자가 같은 원자에 연결되어 있다. 예를 들어, fluoroform 분자는 다음과 같이 표현할 수 있다.

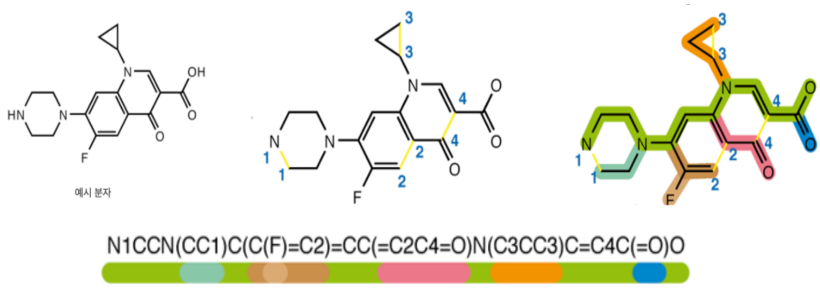


[그림 6] Fluoroform 분자의 SMILES 표기법

위 규칙을 바탕으로 다음과 같이 분자로부터 SMILES 를 만들 수 있다.

1. 수소 원자 제거

2. Ring 과 Aromatic ring 에 포함된 결합을 랜덤으로 하나씩 제거하고 고리마다 번호를 매긴다.
3. 특정 원자를 시작 원자로한 DFS(Depth First Search)을 통해 SMILES string 을 만들어낸다.



[그림 7] SMILES 만드는 과정

1.2 SMILES 단점

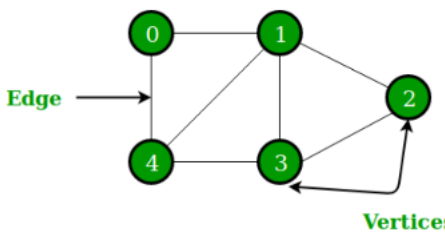
SMILES 는 분자를 string 로 간단하게 표현했음에도 많은 단점을 갖고 있다. SMILES 을 만드는 법에서 알 수 있듯이 한 분자에서 여러 SMILES string 이 추출될 수 있다. 심지어 한 분자를 표현하는 다양한 SMILES 는 전혀 다른 string 이 될 수 있고 전혀 다른 분자에서 나온 SMILES 가 비슷한 string 이 될 수도 있다. 자연어처리 모델에서 영감을 받아 SMILES string 자체를 입력으로 property 을 예측하려는 여러 시도[10, 11]들이 있었는데, 생각보다 좋지 못한 결과를 얻었던 이유가 이것이다. 자연어처리 모델은 문자들의 혹은 단어들의 전후 관계나 문법적인 통계를 학습하는데 위에서 말한 단점으로 학습이 힘들기 때문이다. 그렇기 때문에 본 연구에서는 여러 Dataset 에 있는 SMILES 을 Graph 로 변환하여 입력으로 사용한다.

제 2 절 Graph neural network for molecular graph

Graph neural network[12]는 Graph 을 직접적인 입력으로 하는 대표적인 모델이다. Node level, edge level 그리고 graph level 의 분석이 가능한데 본 논문에서는 분자 하나를 하나의 그래프로 대응하기 때문에 graph level 의 GNN 만을 다룬다. 본 논문을 포함한 대부분의 관련 연구들이 다양한 Graph neural network 중에 가장 널리 쓰이고 있는 모델인 Graph convolutional network 기반의 모델을 사용한다. 자연어처리를 비롯한 대부분의 AI 분야에서 어떤 것에 더 집중할 것인지에 관한 개념인 attention 이 활발하게 쓰이고 좋은 결과를 보이고 있다. 본 논문에서도 attention 을 분석하고 attention 을 적용한다.

2.1 Graph

Graph 란 Node 와 Edge 로 정의되는 자료구조이다. Edge 의 방향성 여부와 weight 존재 여부에 따라 각각 directed/undirected 와 weighted/un-weighted 로 구분된다. 본 연구에서는 Molecular graph 을 다루기 위해 undirected, un-weighted 그래프를 사용한다.

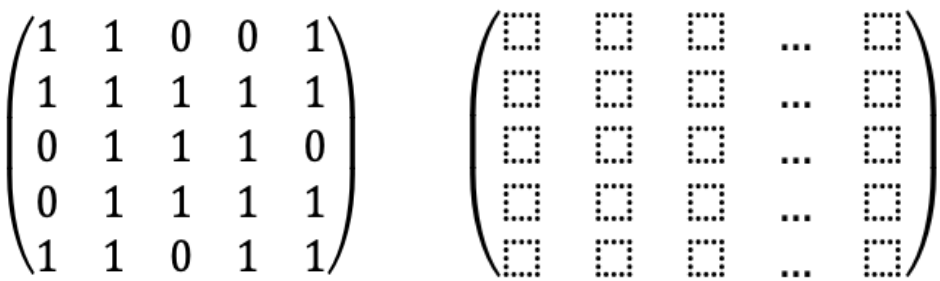


[그림 8] 그래프의 표현

2.2 Graph convolutional network for molecular graph

GCN(Graph convolutional network)[13]는 Graph network 에 CNN[14]에서의 Convolution 개념을 적용한 것이다. GCN 에서 Graph 는 |V(G)| x |V(G)| 크기의 Adjacency matrix A 와 노드의|V(G)| x | F | (F 는 node feature 의 차원)로 표현된다. 여기서 노드는 atom 이고 node feature 는 atom 이 갖는 특성들의

vector 이다. [그림 5]의 예시의 경우 [그림 6]에서 왼쪽 matrix 가 adjacency matrix 가 된다. [그림 6]에서 오른쪽 matrix 는 feature matrix 이며 shape 은 (노드 수 x feature 수)가 된다.



[그림 9] Adjacency metrix(NxN), Feature matrix(NxF)

GCN 은 기본적으로 node states 을 업데이트 하는 방식으로 진행된다. Node states 의 초기값은 Feature Matrix 이다. 위의 예시에서, atom 4 는 atom 0, 1 그리고 3, 모두 3 개의 adjacent atoms 을 갖는다. GCN 에서 atom 4 의 (l-th) node state 는 다음과 같이 구한다.

$$H_4^{(l+1)} = \sigma(H_0^{(l)} W^{(l)} + H_1^{(l)} W^{(l)} + H_3^{(l)} W^{(l)} + H_4^{(l)} W^{(l)})$$

$H^{(l)}$는 l-th node state 을 의미하며 σ, $W^{(l)}$는 각각 activation function, l-th layer 의 convolutional weights 을 의미한다. 한 atom 의 node state 을 업데이트 할 때, 모든 atom 의 state 에서 정보를 가져오는 것이 아닌 그 atom 과 연결된 atom 들만으로부터 정보를 가져온다. 이것을 Adjacency matrix 을 활용해 다음과 같이 한번에 표현할 수 있다.

$$H^{(l+1)} = \sigma(AH^{(l)}W^{(l)})$$

2.3 Attention

Attention[15]은 자연어 처리 분야에서 처음 사용되었고 비전이나 graph 데이터에 대해서도 활발하게 사용되고 있는 기법이다. Graph 에 처음 attention 을 도입한 연구는 GAT(Graph attention network)[16]인데 이는 GCN 기반의 모델에 attention 기법을 적용한 것이다. GAT 에서 attention 을 사용한 목적은 GCN 에서는 인접한 원자에서 정보를 가져올 때 모두 같은 비중으로 정보를 가져오게 되는데 이런 문제점을 해결하기 위해 GAT 에서 attention 이 도입되었다. Attention 을 적용한 (l-th) node state 을 update 하는 equation 은 다음과 같다.

$$H_4^{(l+1)} = \sigma(\alpha_{40} H_0^{(l)} W^{(l)} + \alpha_{41} H_1^{(l)} W^{(l)} + \alpha_{43} H_3^{(l)} W^{(l)} + \alpha_{44} H_4^{(l)} W^{(l)})$$

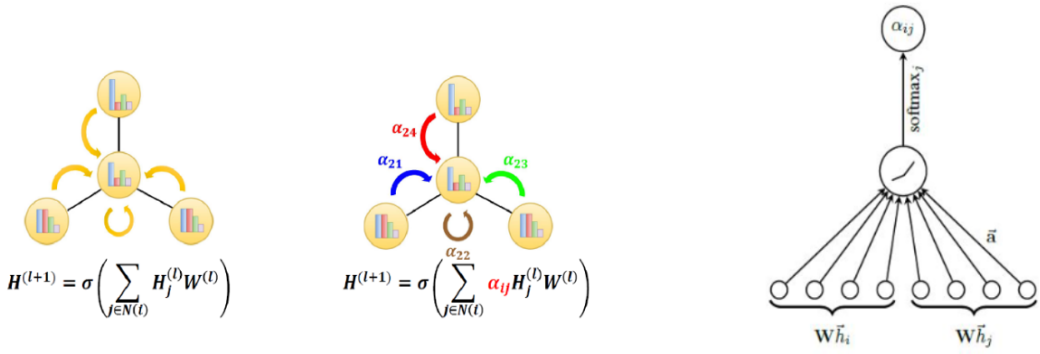
$\alpha_{ij}^{(l)}$는 l번째 layer 에서 i 번째 atom 에 대한 j 번째 atom 의 중요도를 의미하는 attention coefficient 이다. 중요도를 의미하도록 $\sum_{k \in N(i)} \alpha_{ik}$ 의 합은 1 이 되도록 하는 경우가 많다. Attention coefficient 을 구하는 방식은 다양하며 general 한 표현 방식은 다음과 같다.

$$\alpha_{ij}^{(l)} = f(H_i^{(l)} W^{(l)}, H_j^{(l)} W^{(l)})$$

f 는 임의의 과정을 뜻하는데, GAT 연구에서는 다음과 같이 coefficient 을 구한다.

$$\alpha_{ij} = \frac{e_{ij}}{\sum_{k \in N(i)} e_{ik}} = \frac{\sigma(MLP[H_i W, H_j W])}{\sum_{k \in N(i)} \sigma(MLP[H_i W, H_k W])}$$

[,]는 두 행렬의 concatenation 을 의미한다. 앞서 설명했듯이 $\sum_{k \in N(i)} \alpha_{ik}$ 의 합이 1 이 되도록 하기 위해 Softmax 함수를 사용한다.

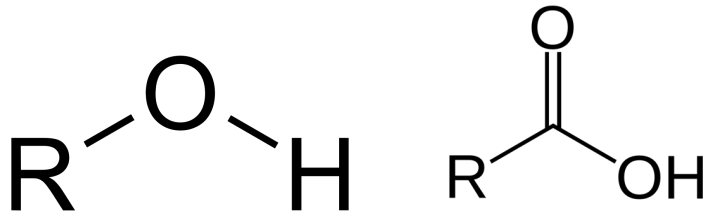


[그림 10] graph 에서 attention 이 적용되는 과정

제 3 절 Molecular fingerprint

Molecular fingerprint 는 분자를 substrucutre 로 나누어 분자의 구조를 encoding 하는 방식이다. Molecular fingerprint 는 처음에는 chemical database 에서 substructure searching 을 위해 개발되었다. 이후에 Molecular fingerprint 을 비교함으로써 두 분자의 구조적 유사도를 측정할 수 있음을 이용해 chemical analysis 에도 사용되기 시작했다[17, 18, 19, 20].

Fingerprint 을 GNN 과 함께 사용하는 이유는 다음과 같다. 분자 특성에는 화학에서 작용기라고 부르는 분자의 특정 substructure 가 큰 영향을 끼치는 경우가 많다. 예를 들면, 카복시기(carboxyl group : -COOH)와 하이드록시기(hydroxy group : -OH) 등이 있다. 카복시기는 탄소와 수소로 이루어진 작용기의 하나인데 카복시기를 갖는 유기물은 녹는점과 끓는점이 높고 물에 대한 용해도가 큰 성질을 갖는다. 또한 하이드록시기는 수소 결합이 가능한 것이 특징이며 물과 친화성을 띠기 때문에 물에 녹기 쉽다.



[그림 11] 하이드록시기, 카복시기

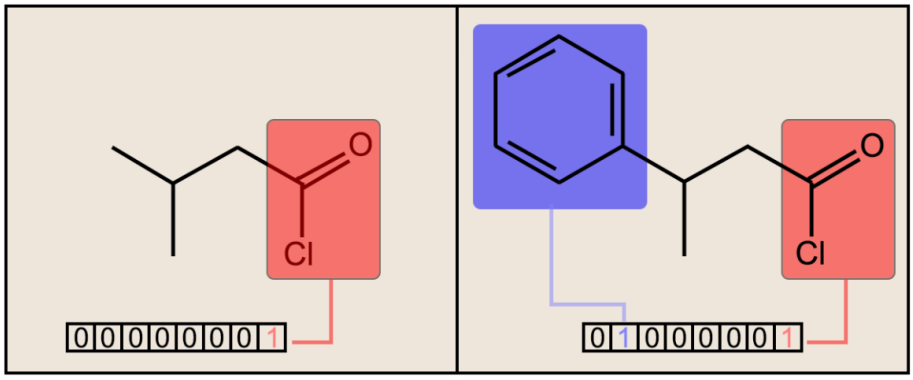
이처럼 특성에 대해 특정 작용기가 결정적인 역할을 하는 경우가 있다. 그리고 이런 작용기는 GNN 을 통해서도 학습될 수 있지만 그것은 간접적이며 fingerprint 에서 더욱 직접적으로 잘 표현될 수 있다. 따라서 GNN 을 통해 분자의 전체적인 구조를 학습하고 fingerprint 을 통해 작용기의 존재 여부를 보다 직접적으로 학습할 수 있길 기대하기 때문이다.

Fingerprint 만드는 다양한 알고리즘이 존재하며 크게 structural fingerprint 와 hashed fingerprint 로 나누어 진다. Structural fingerprint 는 분자에서 [그림 8]과 같이 분자의 특성에 중요하다고 여겨지는 미리 정의된 특정 구조들의 존재 여부를 벡터로 표현한다. 각 bit 로 표현되는 특성은 비트가 1 일 경우 분자에 존재함을 뜻한다. Hashed fingerprint 는 미리 정의된 substructure 을 사용하지 않고 분자를 직접 분해한다. 특정 알고리즘에 따라 분자를 가능한 조각의 집합으로 쪼갠다. 그리고 그 집합의 각 조각을 hash 함수를 통해 [그림] 처럼 fixed-vector 의 위치로 대응시켜 표시한다. 널리 쓰이는 알고리즘으로는 path-based fingerprint 와 circular fingerprint 가 있다. Hashed fingerprint 는 특정 확률로 비트가 충돌될 가능성이 존재하는 단점을 갖고 있다. 따라서 vector 의 크기를 적절하게 설정하는 것이 중요하다.

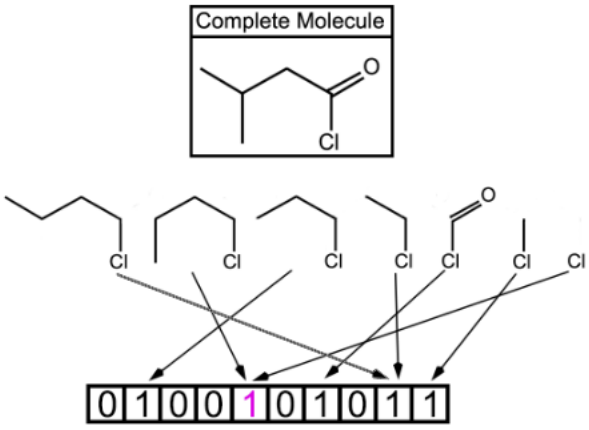
Path-based fingerprint 는 [그림 9]와 같이 path 기준으로 threshold path 이하의 모든 집합을 구한 후 hash 함수를 통해 벡터로 표현한다. [그림 9]는 OC=CN 의 path-based fingerprint 을 구하기 위한 substructure 의 집합의 예시이다. 마지막으로 circular fingerprint 는 [그림 10]과 같이 iteration 별로 직경을 넓혀가며 특정 iteration 에 도달하면 종료한다. Circular fingerprint 는 path-based fingerprint 의 간소화 버전이라고 생각할 수 있다. 하지만 path-based fingerprint 가 circular fingerprint 의 subgraph 의 수보다 많기 때문에 해시 함수를 통해 벡터를 만들 때 충돌을 일으킬 확률이 높다는 단점도 갖고 있다.

이렇게 Molecular graph 의 substructure 로 쪼개어 vector 로 표현하게 되면 consine similarity, tanimoto similarity[21]등으로 분자 간 유사도를 측정할 수 있다. 사실 분자의 유사도에는 정해진 metric 이 없기 때문에 이것을 명확히 유사도라고 할 순 없지만 동일한 subgraph 을 많이 포함할수록 비슷한 분자이며

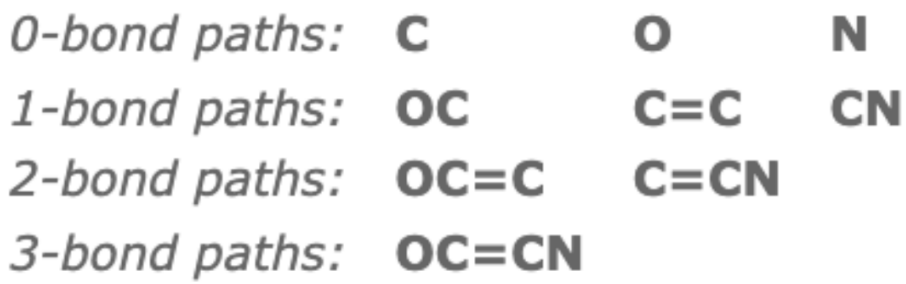
비슷한 성질을 갖을 것이고 가정하에 널리 쓰이고 있다. Fingerprint 추출의 구현 알고리즘은 MACCS keys, Daylight-like, ECFP[22, 23](각각 structural fingerprint, path-based fingerprint, circular fingerprint 의 대표적인 구현 알고리즘이다) 등 다양하며 본 연구에서의 구현은 python rdkit[24] 오픈소스를 사용했다.



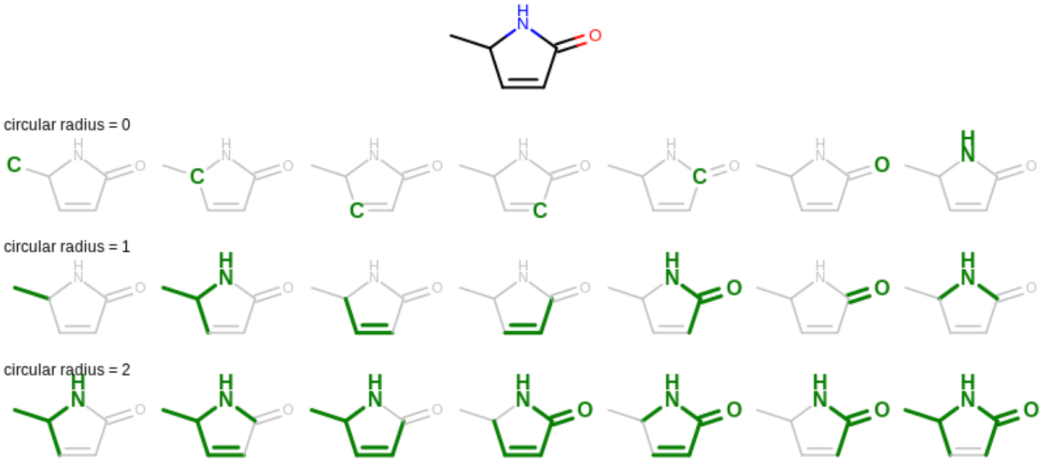
[그림 12] Structural fingerprint



[그림 13] Hashed fingerprint



[그림 14] Path-based fingerprint 에서 substructure 집합을 구하는 예시



[그림 15] Circular fingerprint 의 substructure 예시

제 3 장 모델

제 1 절 GCN with attention

1.1 Feature matrix

앞서 설명했듯이 GCN 계열의 모델에서 molecular graph 을 입력으로 사용하기 위해서는 adjacency matrix 와 원자의 feature matrix 가 필요하다. 본 연구에서는 5 가지 feature 을 사용했다.

1. 원자 종류
2. 연결되어 있는 원자 수
3. 연결되어 있는 수소의 수
4. 원자가(valence)
5. 방향족성 여부

원자의 종류는 일반적으로 분자에 존재하는 40 가지 원소들을 포함한다. 원자에 연결되어 있는 원자 수, 수소의 수, 원자가는 각각 0 부터 5, 0 부터 4, 0 부터 5 까지의 값을 갖는다. 모든 feature 는 one-hot encoding 방식을 사용했으며 각각 40, 6, 5, 6, 1 크기로 총 58 크기의 feature vector 을 사용했다.
모델에 입력으로 사용하기 위해서는 matrix 의 shape 이 고정되어야 한다. 따라서 atom 의 수를 고정해야 한다. 만일 분자의 atom 의 수가 고정된 값보다 커지면 데이터의 손실이 발생한다. 그리고 atom 의 수가 고정된 값보다 적은 경우에는 0 으로 padding 을 해준다. 따라서 데이터셋을 잘 파악하고 적절한 atom 의 수를 잘 설정해 주어야한다. 본 연구에서는 데이터셋에 존재하는 분자 중에 가장 큰 분자를 최대치로 설정했다.

1.2 Attention

본 논문에서는 GAT 에서 사용했던 attention 사용하지 않고 [25]에서 제안했고 분자 특성 예측에서 좋은 성능을 보였던 다음의 attention 을 사용했다.

$$\alpha_{ij}^{(l)} = \sigma((H_i^{(l)} W^{(l)}) C^{(l)} (H_j^{(l)} W^{(l)})^T)$$

4 장에서 진행한 실험에서 attention 에 따른 결과를 비교 분석하며 최종적으로 [25]연구에서 제안한 attention 을 채택하여 사용했다. 또한 multi-head attention 로 구현되었다.

제 2 절 Molecular fingerpint

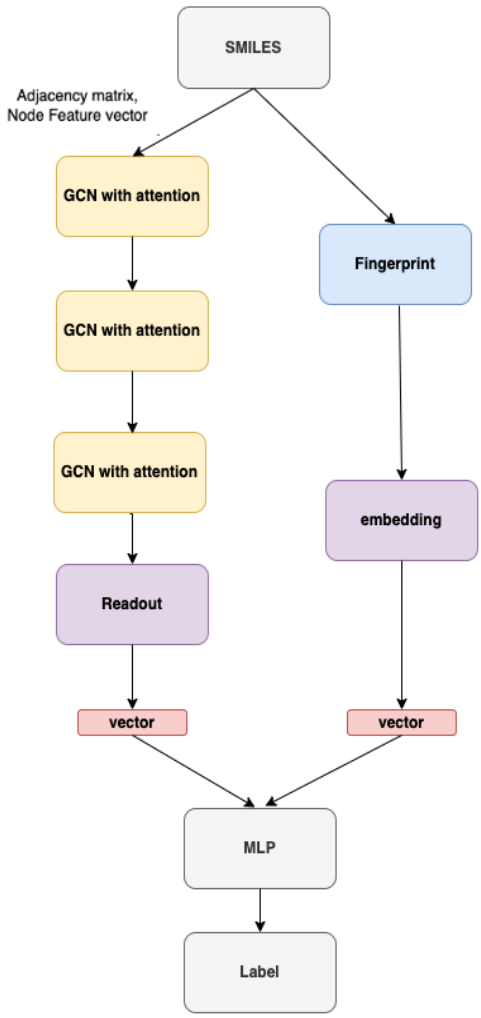
본 연구에서는 3 가지의 대표적인 fingerprint 을 구현하고 분석했다. 3 가지 fingerprint 는 MACCS keys, Daylight-like, ECFP 로 각각 structural fingerprint, path-based fingerprint, circular fingerprint 의 대표 알고리즘이다.

MACCS keys 의 경우 166 크기의 벡터가 추출된다. Daylight-like fingerprint 와 ECFP fingerprint 의 경우 벡터의 크기는 사용자가 설정해줄 수 있으며 본 연구에서는 1024 를 사용했다. Fingerprint 종류에 따른 실험 결과를 4 장 실험 및 평가에서 보인다.

제 3 절 모델

전체적인 모델의 구조는 [그림 11]과 같다. 모델은 Molecular graph 을 활용하는 부분과 fingerprint 을 활용하는 부분으로 이루어진다. SMILES 로 구성된 dataset 을 adjacency matrix 와 feature vector 로 변환한다. 그리고 Attention 을 적용한 GCN block 과 Readout 을 통해 64 크기의 벡터로 인코딩한다. 동시에 SMILES 을 fingerprint 로 변환 후 embedding layer 을 거쳐 32 크기의 벡터로 인코딩한다. 마지막으로 두 벡터를 MLP 을 통해

label 을 예측한다. Loss function 으로는 mae(mean absolute error)을 사용했고 adam optimizer 을 사용했다.



[그림 16] 모델의 전체적인 구조

제 4 장 실험 및 평가

실험은 총 4 가지를 진행했다. 각 실험은 GCN with attention block 의 수에 따른 성능, attention 에 따른 성능 비교, fingerprint 종류에 따른 성능 비교, Graph baseline 들과의 비교로 구성된다.

제 1 절 데이터셋

데이터셋은 관련 연구에서 가장 많이 사용되었던 ZINC[26], QM9[27] 그리고 CEP[28] dataset 을 사용했다. ZINC 데이터셋에서 100K 크기의 데이터를 랜덤으로 추출해서 logP, TPSA, SAS 를 예측하는 비교 실험을 진행했다. QM9 데이터셋의 경우 50K 크기의 데이터를 랜덤으로 추출해 Mu(Dipole moment), alpha(Isotropic polarizability), homo(Energy of Highest occupied molecular orbital), G(Free energy at 298.15 K), H(Enthalpy at 298.15 K)을 예측하는 비교 실험을 진행했다. 마지막으로 CEP 데이터셋에서 PVE 를 예측하는 실험을 진행했다.

Dataset	ZINC	QM9	CEP
Size	230M(100K)	133K(50K)	30K
Label	LogP, TPSA, SAS	Mu, alpha, homo, G, H	PVE

[표 1] 사용한 데이터셋

제 2 절 실험환경

모든 실험은 다음과 같은 환경에서 진행되었다.

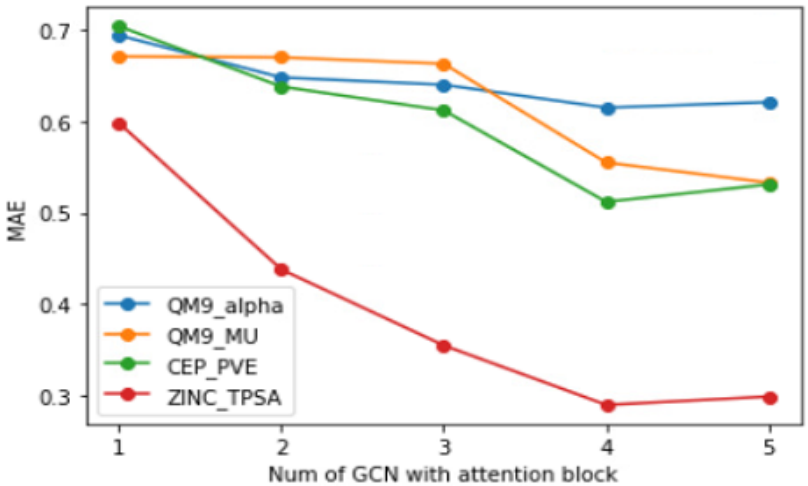
Os	Ubuntu 20.04 LTS
CPU	AMD Ryzen 9 5900X 12-Core Processor
GPU	GeForce RTX 3060
Memory	32G

[표 2] 실험 환경

제 3 절 실험 결과

3.1. GCN with attention block

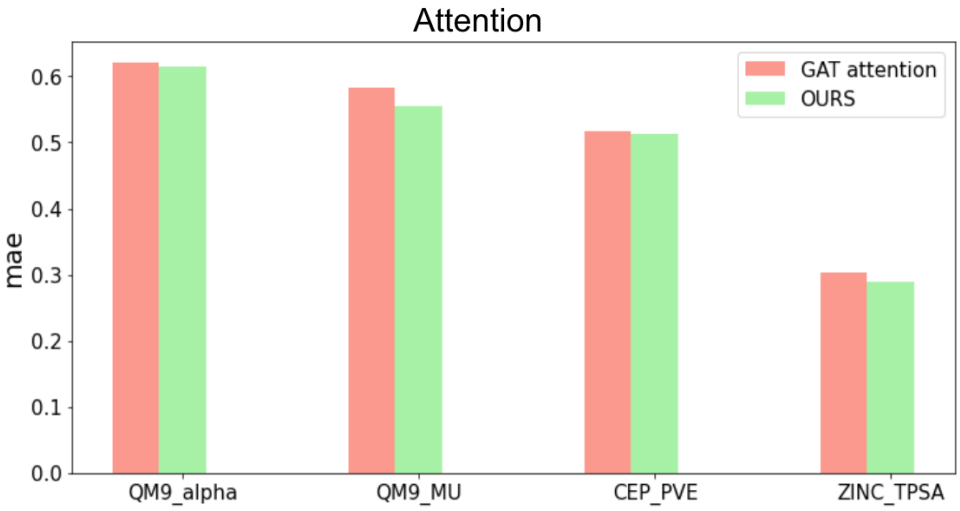
첫번째로 GCN with attention block 의 수를 1 부터 5 까지 바꿔가면서 실험을 진행했다. QM9 dataset 에서는 alpha 과 mu, CEP 에서 PVE, ZINC 에서 TPSA 로 진행했다. [그림 12]에서 볼 수 있듯이, GCN with attention block 이 늘어날수록 대체적으로 성능이 좋아졌다. 하지만 4 개를 넘어가면서 큰 변화가 없고 오히려 성능이 저하되는 경향을 보였다. 따라서 baseline 과의 비교에서는 4 개의 block 을 사용했다.



[그림 17] GCN with attention block 수에 따른 성능 비교

3.2 Attention

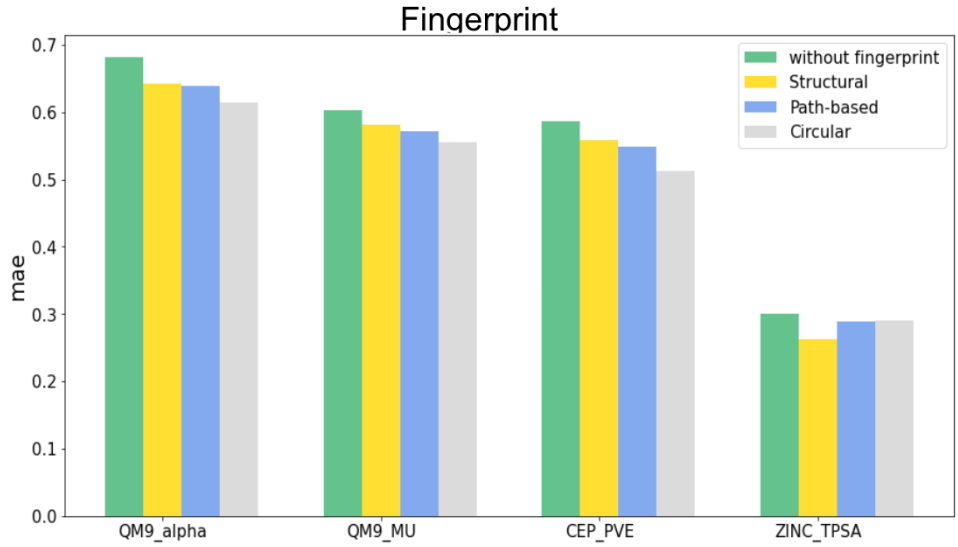
GAT 에서 도입했던 attention coefficient 와 본 연구에서 사용한 attention coefficient 을 사용해서 비교했다. [그림 13]에서 확인할 수 있듯이, 본 연구에서 사용한 attention coefficient 이 실험에서 좋거나 비슷한 성능을 내는 것으로 확인할 수 있었다.



[그림 18] Attention 에 따른 성능 비교

3.3 Fingerprint

각각 structural fingerprint, path-based fingerprint, circular fingerprint 의 대표적인 구현 알고리즘인 MACCS keys, Daylight-like, ECFP 을 모델에 적용해 실험했다. 그 성능 비교는 [그림 14]와 같다. Fingerprint 을 사용했을 때 사용하지 않았을 때에 비해 약 10%의 성능 향상이 있었다. Fingerprint 종류에 따른 실험 결과에서는 ECFP 가 가장 좋은 성능을 냈고 structural fingerprint 가 가장 성능이 낮았다. MACCS 는 predefined feature 들을 사용하는데 그 feature 만을 사용하는 것보다 더 다양한 feature 들을 사용하는 것이 성능에 도움이 됨을 알 수 있다. 단, ZINC_TPSA 실험에서는 Structural fingerprint 가 가장 좋은 성능을 보였다. 이를 고려할 때, structural fingerprint 에서 미리 정의된 substructure 와 예측하고자 하는 분자 특성이 깊은 연관이 있는 경우 가장 강력한 효과를 낼 수 있다고 해석할 수 있다. 그리고 Daylight-like 보다 ECFP4 가 나은 결과를 내는 것을 보았을 때, 많은 subgraph 을 고려하는 것보다 적절한 threshold 을 두는 것이 molecular property prediction 에 효과적임을 알 수 있다.



[그림 19] Fingerprint 종류에 따른 성능 비교

3.4 Baselines 과 비교

본 연구에서는 총 6 가지 baseline 모델과 성능(mae)을 비교한다. Baseline 로 가장 널리 쓰이고 있는 GCN 과 GAT. GCN 과 GAT 에 각각 gate 의 개념을 추가한 gated-GCN 과 gated-GAT, Graph 데이터에서 RNN 기반 모델의 baseline 로 사용되는 GGNN, 자연어처리 기반 모델인 SMILES-BERT 로 구성된다. 모든 실험은 learning rate 0.001, 500 epoch 로 진행됐다. 모든 실험은 5 번 반복하여 평균을 측정되었다.

	QM9_alpha	QM9_mu	QM9_homo	QM9_G	QM9_H
GCN	0.740	0.701	0.114	0.056	0.078
GAT	0.701	0.620	**0.041**	0.049	0.059
Gated-GCN	0.739	0.579	0.068	**0.039**	0.063
Gated-GAT	0.688	0.564	0.058	0.089	0.066
GGNN	0.721	0.611	0.055	0.071	0.063
SMILES-BERT	0.708	0.634	0.099	0.092	0.089
Our method	**0.615**	**0.555**	0.054	0.058	**0.049**

[표 3] QM9 Dataset 에 대한 성능 비교

	CEP_PVE	ZINC_TPSA	ZINC_lopP	ZINC_SAS
GCN	0.690	0.394	0.072	0.059
GAT	0.598	0.294	0.066	0.049
Gated-GCN	0.610	0.324	**0.038**	0.051
Gated-GAT	0.580	**0.274**	0.052	0.079
GGNN	0.707	0.421	0.071	0.045
SMILES-BERT	0.682	0.446	0.050	0.049
Our method	**0.512**	0.290	0.048	**0.038**

[표 4] CEP, ZINC Dataset 에 대한 성능 비교

본 연구의 모델이 총 9 가지의 실험 중 alpha, mu, H, PVE, SAS 을 예측하는 실험에서 가장 좋은 성능을 보였다. 위 비교 모델 중 GAT 모델이 본 연구의 모델과 fingerprint 을 제외하면 가장 유사하며 직접적인 비교를 할 수 있는데 가장 좋은 결과를 내지 못한 4 가지 실험 중 2 가지 실험에서 GAT 보다 좋은 성능을 보였다.

HOMO 와 G 와 같이 특정 property 에서 실험 결과가 좋지 못했는데 그 이유는 다음과 같이 추측할 수 있다. 모델들의 가장 큰 차이점은 fingerprint 사용 유무이며 앞서 설명한 것과 같이 본 모델에서 fingerprint 을 사용함으로써 기대하는 바는 작용기의 존재 여부를 더 직접적으로 학습하는 것이다. 대부분의 분자의 특성이 작용기에 영향을 많이 받기 때문에 Fingerprint 을 GNN 과 함께 사용함으로써 모델이 더 예측을 잘하는 경향을 보였지만 어떤 소수의 특성들은 작용기와 관련이 적기 때문에 fingerprint 을 함께 사용했을 때 성능이 낮아졌다고 추측할 수 있다. 실제로 HOMO 와 G 는 특정 상황에서의 에너지와 관련된 특성이며 작용기보다는 분자의 질량이나 전체적인 구조에 영향을 받을 가능성이 클 수 있다는 chemical 분야의 연구자의 의견이 있었다.

제 5 장 결론 및 향후 연구

본 논문에서는 효과적인 분자 특성 예측을 위한 그래프 뉴럴 네트워크를 제안한다. SMILES 로 구성된 데이터셋을 Graph 와 fingerprint 로 나누어 표현한 후에 분석했다. 본 연구에서 사용된 모델은 Graph convolutional network 기반의 모델에 attention 을 적용한 모델이며 Attention coefficient 는 분자 특성 예측에 특화된 방식으로 정의했다. 그리고 작용기를 표현하기 위해 추가적으로 fingerprint 을 사용했으며 GNN 과 결합하여 분자 특성을 예측했다.

분자 특성 예측 연구에서 공통적으로 쓰이는 dataset 과 baseline 을 활용해서 실험을 진행했다. Molecular property prediction 에 적합한 attention 과 fingerprint 을 비교 분석했고 이를 적용한 모델과 Baseline 을 비교했을 때, 9 가지의 실험에서 5 개의 실험에서 가장 좋은 성능을 냈고 남은 모든 실험에서도 competitive 한. 성능을 보였다.

한 가지 아쉬운 점은, fingerprint 활용을 위해 GCN with attention 을 통해 나온 벡터와 MLP 을 통해 단순히 결합하는 형태로 분석을 했다는 점이다. Fingerprint 는 subgraph 들의 집합이라고 할 수 있는데, 이를 GCN with attention 과 조금 더 유기적으로 결합할 수 있는 아이디어가 있으면 더 좋은 모델이 될 것이라고 생각하며 후속 연구를 진행할 예정이다.

참고 문헌

[1] Shoichet BK. Virtual screening of chemical libraries. Nature. 2004
[2] Zeeshan Ahmad, Tian Xie, Chinmay Maheshwari, Jeffrey C. Grossman, and Venkatasubramanian Viswanathan *ACS Central Science* 2018
[3] Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik *ACS Central Science* 2018
[4] Oliver Wieder, Stefan Kohlbacher, Mélaine Kuenemann, Arthur Garon, Pierre Ducrot, Thomas Seidel, Thierry Langer,Volume 37, 2020, ISSN 1740-6749
[5] Gabriel A. Pinheiro, Johnatan Mucelini, Marinalva D. Soares, Ronaldo C. Prati, Juarez L. F. Da Silva, and Marcos G. Quiles . *The Journal of Physical Chemistry A* 2020
[6] C. Lu, Q. Liu, C. Wang, Z. Huang, P. Lin, and L. He, "Molecular Property Prediction: A Multilevel Quantum Interactions Modeling Perspective", *AAAI*, vol. 33, no. 01, pp. 1052-1060, Jul. 2019.
[7] Dillard L. Self-Supervised Learning for Molecular Property Prediction. ChemRxiv. Cambridge: Cambridge Open Engage; 2021
[8] Guo, Zhichun and Zhang, Chuxu and Yu, Wenhao and Herr, John and Wiest, Olaf and Jiang, Meng and Chawla, Nitesh V; WWW '21, April 19–23, 2021, Ljubljana, Slovenia
[9] David Weininger; *Journal of Chemical Information and Computer Sciences* 1988 *28* (1), 31-36
[10] Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, Junzhou Huang; ACM-BCB '19, September 7–10, 2019, Niagara Falls, NY, USA.
[11] Xiaoyu Zhang, Sheng Wang, Feiyun Zhu, Zheng Xu, Yuhong Wang, and Junzhou Huang. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. ACM, 404–413.
[12] Zhang, S., Tong, H., Xu, J. *et al.* Graph convolutional networks: a comprehensive review. *Comput Soc Netw* 6, 11(2019)
[13] Thomas N. Kipf, Max Welling, Semi-Supervised Classification with Graph Convolutional Networks, ICLR 2017
[14] Yamashita, R., Nishio, M., Do, R.K.G. *et al.* Convolutional neural networks: an overview and application in radiology. *Insights Imaging* 9, 611–629 (2018)
[15] *Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.*

[16] Veličković, Petar and Cucurull, Guillem and Casanova, Arantxa and Romero, Adriana and Liò, Pietro and Bengio, Yoshua. Graph Attention Networks. ICLR 2018
[17] Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods. 2015 Jan;71:58-63. doi: 10.1016/j.ymeth.2014.08.005. Epub 2014 Aug 15.
[18] Raymond JW, Willett P. Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. J Comput Aided Mol Des. 2002 Jan;16(1):59-71. doi: 10.1023/a:1016387816342. PMID: 12197666.
[19] Willett P. Similarity-based virtual screening using 2D fingerprints. Drug Discov Today. 2006 Dec;11(23-24):1046-53. doi: 10.1016/j.drudis.2006.10.005. Epub 2006 Oct 20. PMID: 17129822.
[20] Eckert H, Bojorath J: Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. Drug Discov Today 2007, 12:225-233.
[21] Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations. *Cheminform* 7, 20 (2015)
[22] Daylight Fingerprints. https://www.daylight.com/meetings/summerschool01/course/basics/fp.html. Accessed October 2019
[23] Rogers D, Hahn M: Extended-Connectivity Fingerprints. J Chem Inf Model 2010, 50:742-754.
[24] RDKit. https://www.rdkit.org/. Accessed October 2019.
[25] Ryu S, Lim J, Hong SH, Kim WY. Deeply learning molecular structure-property relationships using attention- and gate-augmented graph convolutional network; 2018.
[26] J. J. Irwin and B. K. Shoichet, Journal of chemical information and modeling, 2005, 45, 177–182.
[27] Fu, Tianfan, 2022, "qm9", https://doi.org/10.7910/DVN/8ZZZ6J, Harvard Dataverse, V3
[28] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway and A. Aspuru-Guzik, The Journal of Physical Chemistry Letters, 2011, 2, 2241–2251.

Abstract

Graph Neural Network for Prediction of Molecular Properties Using Attention and Fingerprint

Jaeheon Jung

Computer science and Engineering

The Graduate School

Seoul National University

In the development of new drugs, machine learning has been utilized in several fields, including molecular generation and molecular property prediction. In particular, molecular property prediction plays an important role in the field of drug discovery. By selecting the drugs that are expected to have the necessary properties of the new drug and filtering the drugs that are expected to not, the entire process can be greatly accelerated in a faster and cheaper way. Considering that the average cost of new drug development is currently estimated to be about $2.8 billion, we can feel how important it is to reduce trial and error through molecular characteristic prediction.

Traditionally, molecular properties have been predicted through fingerprints that analyzed subgraphs, hand-engineered features based on physical rules, and DFT (Density Functional Theory). However, AI has recently made remarkable progress in various fields. In molecular characteristic prediction, there have been many attempts to predict molecular characteristics through deep learning and have produced better results than conventional methods. In particular, many studies have shown that Graph neural network-based models are effective. Molecules are made up of bonds between atoms because they can be well expressed in vertex and edge on a graph.

This study analyzes existing molecular characteristic prediction strategies and proposes new strategies. As for the attention that is actively used in various fields such as natural language processing, the attention suitable for the molecular property prediction task is analyzed and applied to the graph natural network. And we use fingerprint to take advantage of the fact that functional groups are generally related to the properties of molecules. In this study, the type of fingerprint is analyzed and used in combination with a graph natural network, and compared and analyzed through experiments.
Through experiments using various datasets, it is observed that the analysis strategy proposed in this study showed competitive results compared to baseline models.

Keywords : molecular property prediction, deep learning, drug discovery, graph neural network, attention, fingerprint
Student Number : 2020-25719

Acknowledgements

논문을 작성하기까지 도움을 주신 많은 분들께 감사의 말씀을 전합니다. 먼저, 지도 교수님이신 문봉기 교수님께 감사의 말씀을 드립니다. 논문을 작성하는 과정에서 교수님의 섬세하고 날카로운 조언과 격려가 있어 끝까지 마무리할 수 있었습니다. 부족한 학생이었지만, 앞으로 최선을 다해 부끄럽지 않은 제자가 되도록 노력하겠습니다.

항상 친절하게 대해준 DBS 연구실 분들에게 감사의 인사를 전합니다. 연구실에서 소통하며 많은 것을 배웠고 의지했습니다. 본 연구는 DDI 과제에서 연구했던 것들을 기반으로 시작되었는데, DDI 과제를 하는 동안 많은 도움을 주신 지현씨와 상하씨에게 특히 감사 인사를 전하고 싶습니다. 뒤늦게 합류한 과제였지만, 모르는 것이 있고 어려움이 있을 때마다 도와주셔서 큰 도움이 됐습니다.

또한, 학교 적응에 큰 도움을 주었고 석사 기간 동안 많은 의지가 되었던 CTA 연구실의 장지훈과 CSL 연구실 조규진, 늘 저를 응원해주고 함께 즐거운 시간을 보내주는 정컴, 짱구들 그리고 부족한 저를 믿어주고 성장할 수 있게 해준 바비디 팀원들에 감사의 인사를 전하고 싶습니다.

마지막으로, 항상 부족한 저를 최고라고 믿어주고 묵묵히 응원해주는 가족들에게 감사의 인사를 전합니다. 한없이 우울할 때도 가족이 있어 중심을 잡고 극복할 수 있었습니다. 앞으로도 오랫동안 함께 행복한 시간 보냅시다.