

# Chemical Functional Group Finding in Inter-Molecular Graph

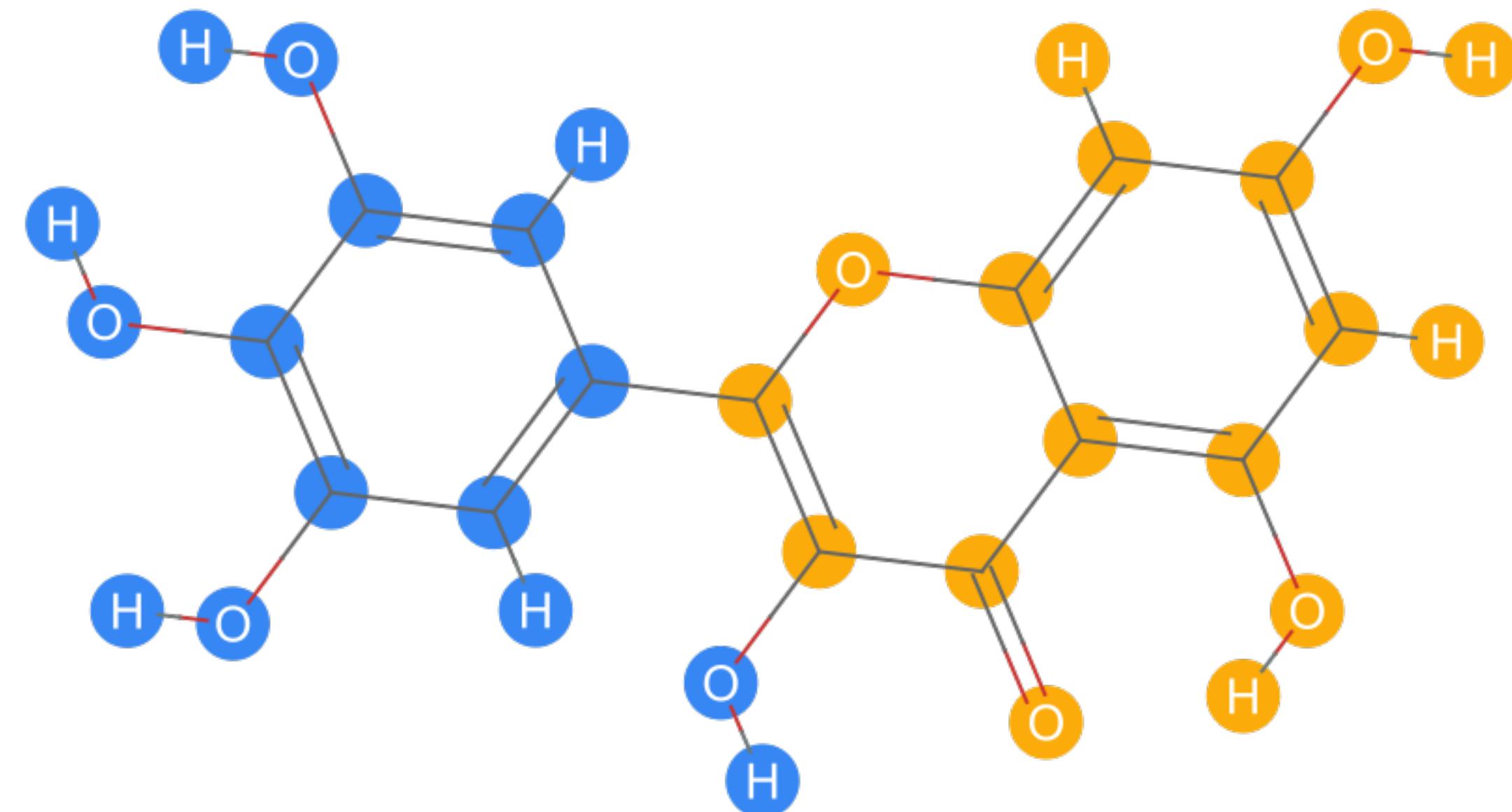
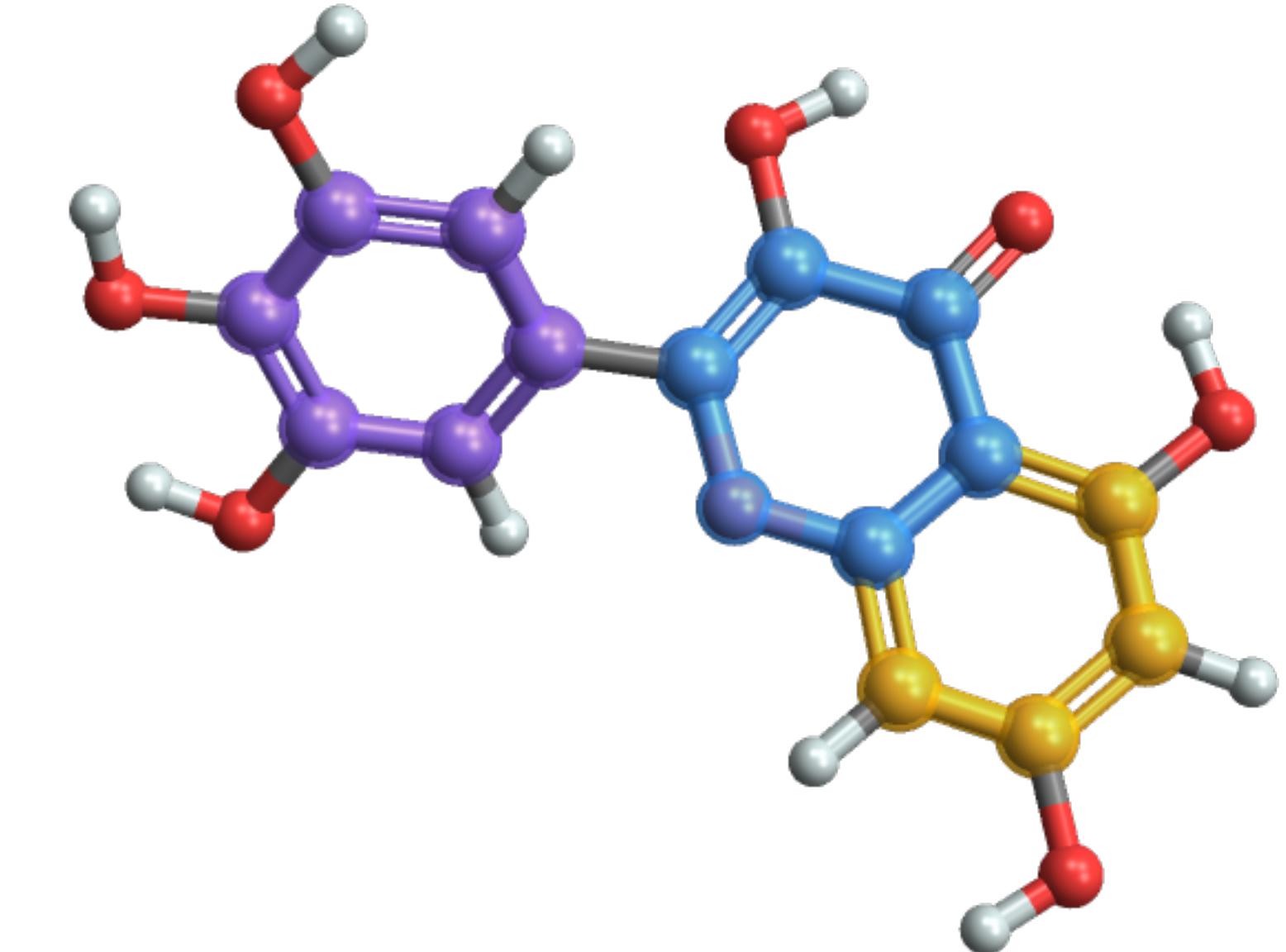
Graph Mining Project Presentation  
{ Interactive Graph Mining 강의 }

성균관대학교 소프트웨어학과 김산  
2023. 11. 15. 수요일.

# Motivation

Before this work,

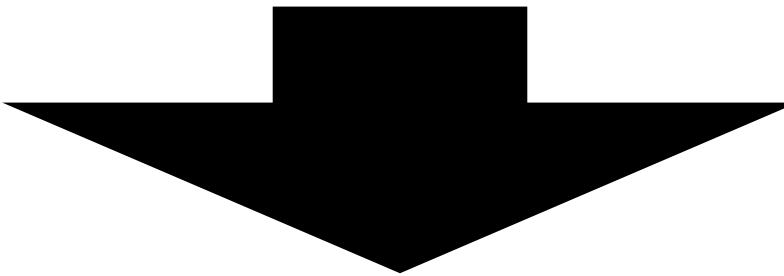
- Molecular graph
  - RDKit
  - LGB
- Molecular substructure relationship
  - RDKit
  - ...?



# Purpose

Expected results are,

- Inter-molecular Graph(IMG) 구축



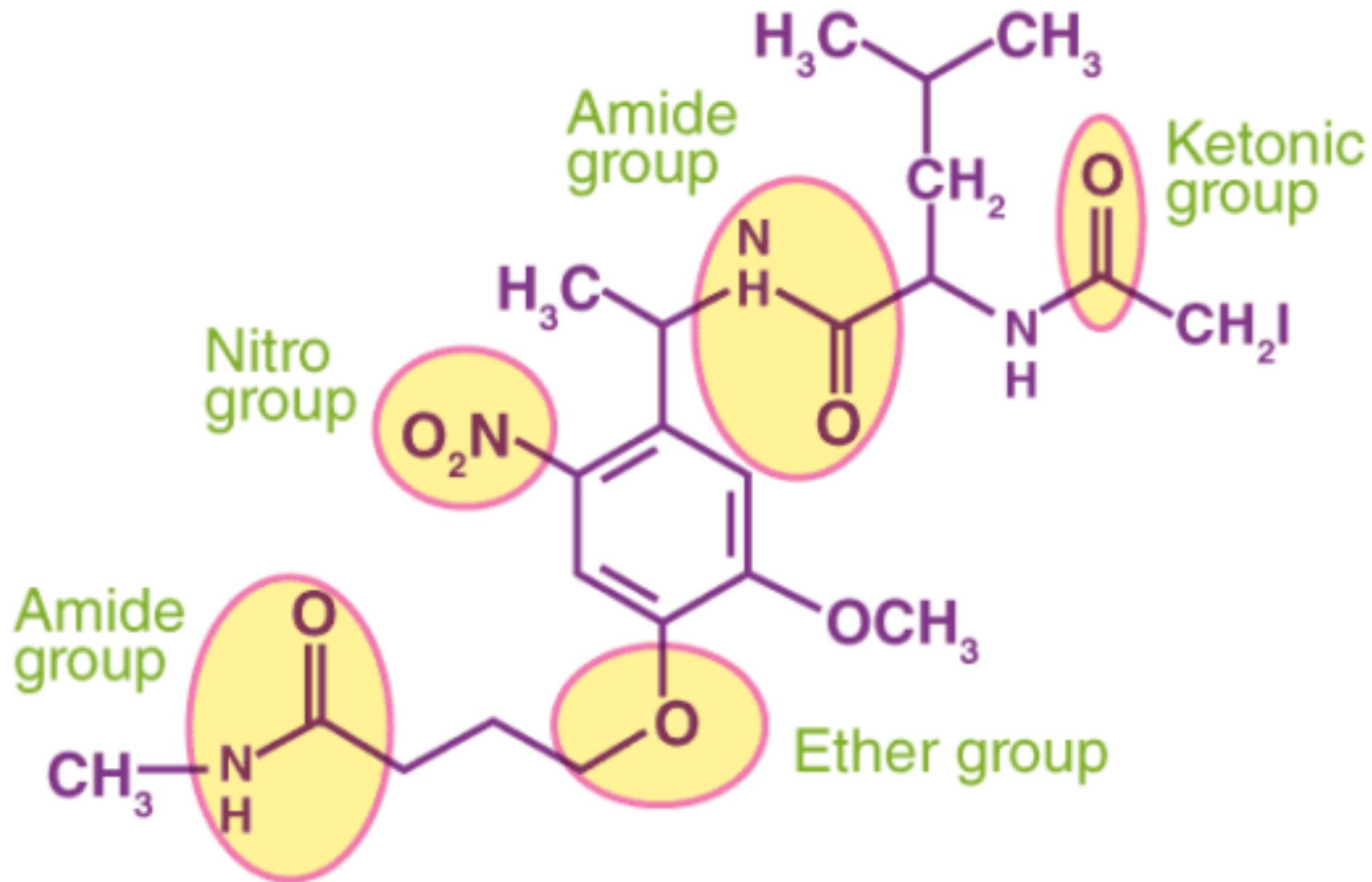
- IMG의 성질 파악
  - Basic graph analysis
  - Centrality analysis
- Path 분석
- 정보 엔트로피 분석

새로운 functional group 찾기

# Purpose

## Reveal undefined functional groups

- 작용기 확인하기
  - Functional group은 화학적으로 안정하면서도
    - 자주 나타나기 때문에 molecular graph의 edge feature를 수집해 발견할 수 있음
- 작용기를 분류하기 :: **Inter-molecular feature**
  - 기존에 알려진 작용기들과 비슷한 빈도-중요도로 등장하면서
  - 특별한 기능으로 명명되지 않았던 작용기를 그래프마이닝을 통해 발굴하기!

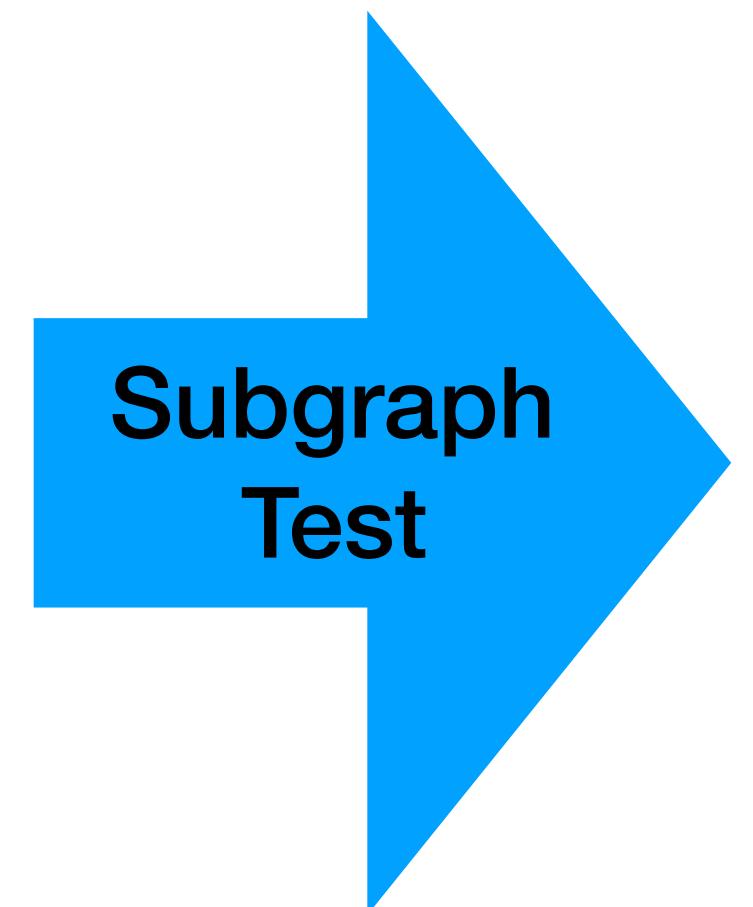


# Dataset

QM9 dataset → **Inter-molecular graph**

- **Quantum Machines 9 (QM9) dataset**

- Nodes
  - Atoms
- Edges
  - Molecular bonds  
(Valence Shell Electron Pair Repulsion)
- # of graphs
  - 7165 graphs(= molecules)
  - MAX 23 nodes(= atoms) per graph
    - MAX 7 heavy atoms per graph



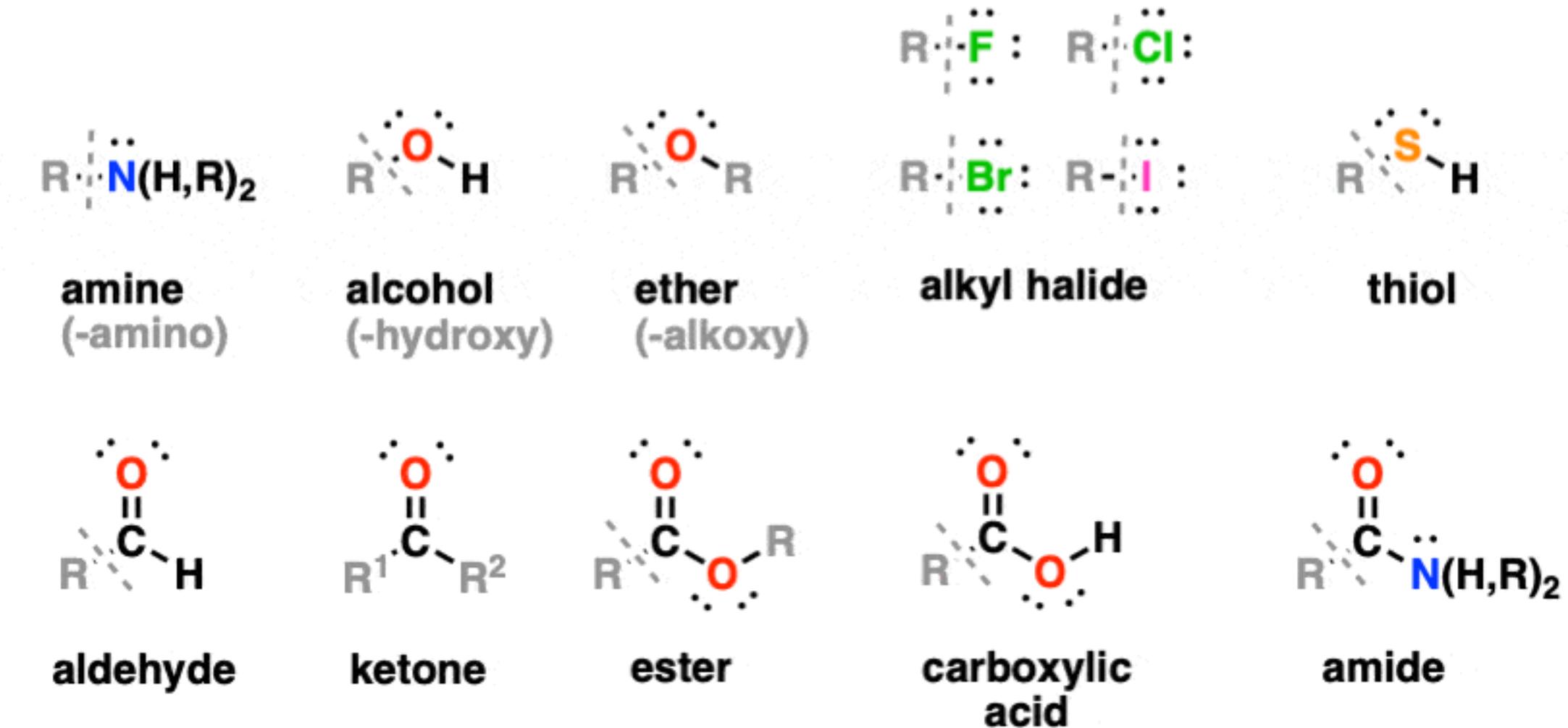
- **Inter-molecular graph**

- Nodes
  - Molecules
- Edges
  - **Subgraph relations ( $\Delta_{atom} = 1$ )**
- 1 large graph
  - 7165 nodes(= molecules)

# Experiments

## Path and Information Entropy

- 경로 길이 3의 경로들을 수집
  - 경로들은 한 문자에서 다른 문자가 되는 경로임
  - 따라서 서로 다른 두 문자간의 차이를 알 수 있음
  - 이때 문자간의 차이를 typical하냐, atypical하냐를 기준으로 통계적으로 분류할 수 있음



⇒ 새로운 functional group 찾기

# Results

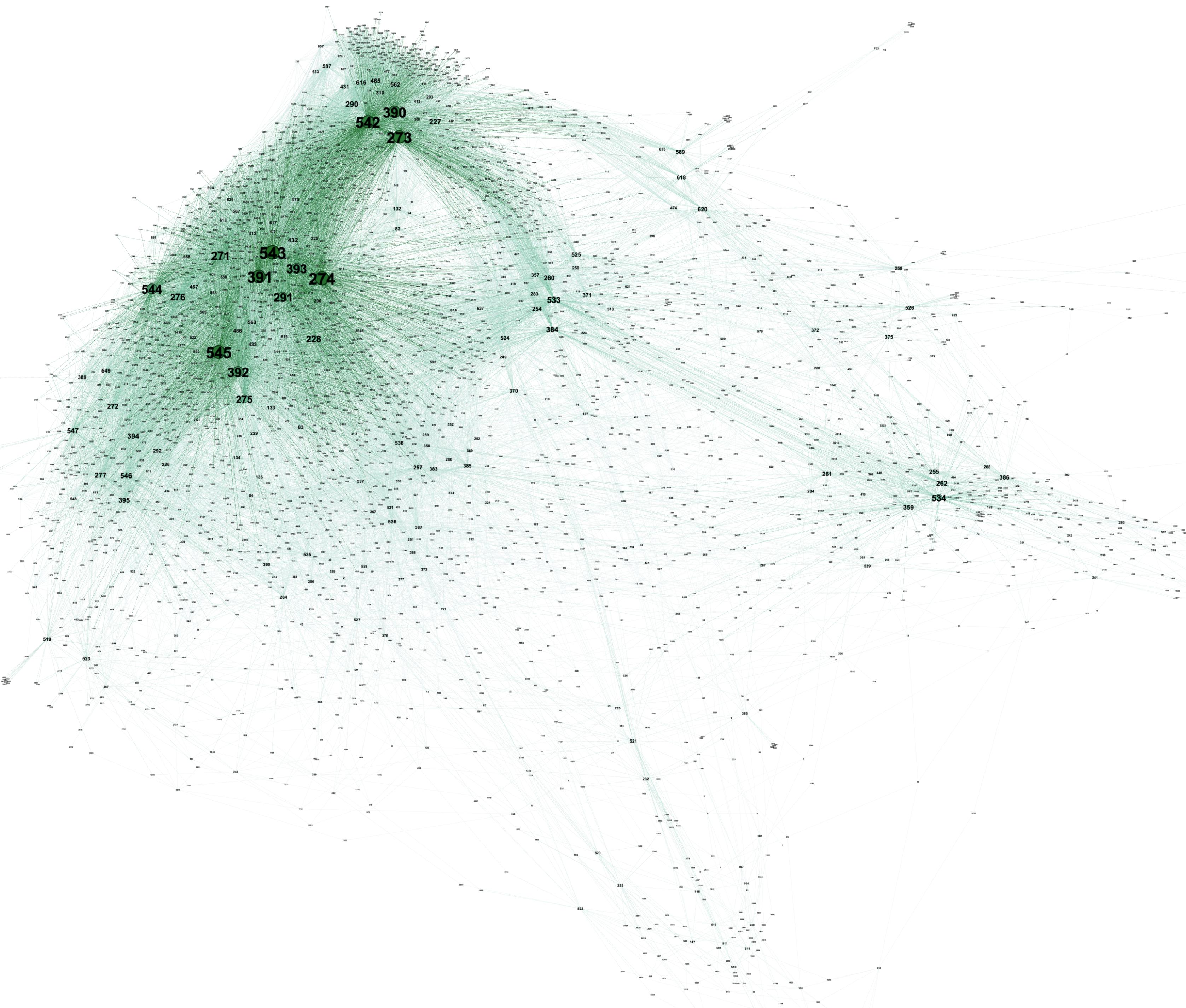
## Graph properties

- Node Count • 3869
- Edge Count • 23470
- Edge Density • 0.00156
- Avg. Degree • 12.132
- Avg. In Degree • 6.0661
- Avg. Out Degree • 6.0662

# Results

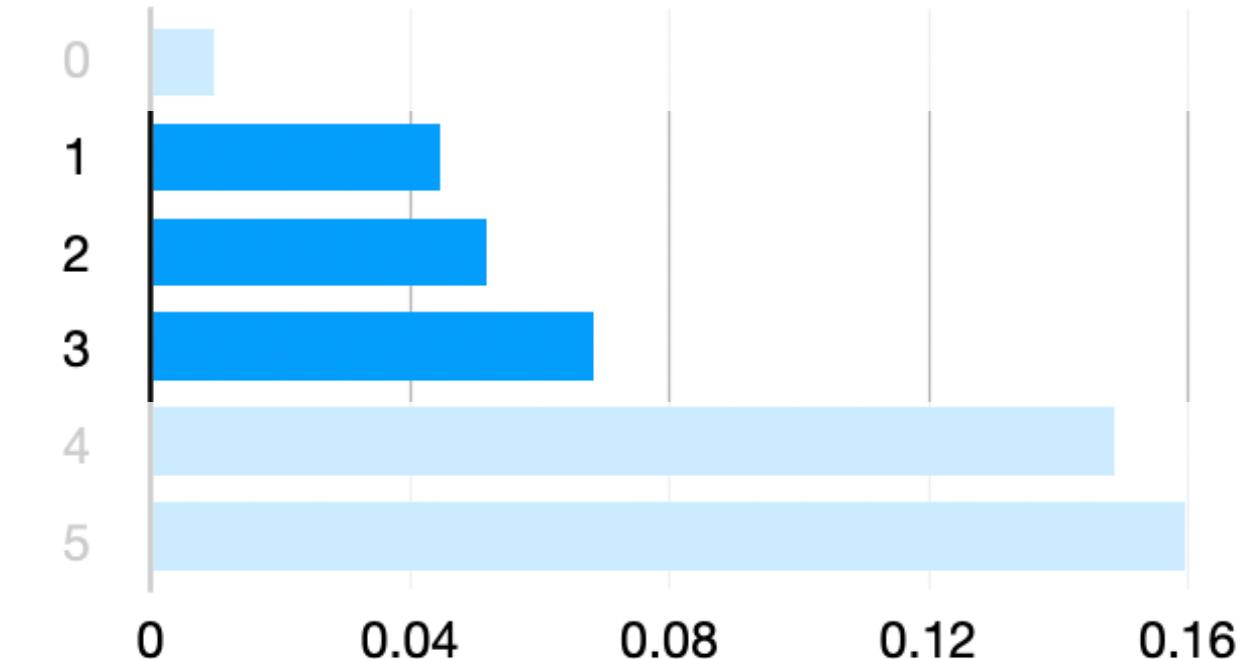
# Visualization

- 비슷한 문자들이 주요 백본 위치에 등장함
  - Typical한 문자구조의 문자들이 dense 한 커뮤니티를 이루고
  - 이후에 더 작은 문자들이 작은 커뮤니티를 이룸



# Result

## Path and Entropy → Functional Groups



Molecular Graph	Name	Information Entropy	물리/화학적 의미를 가지고 있음 0, 4, 5는 그렇지 않음
$\begin{array}{c} \text{O} \\ \parallel \\ \text{R}-\text{C}-\text{R}' \end{array}$	Ketones	0.0445	지용성과 수용성을 동시에 가질 수 있음, 반응성 높임
$\begin{array}{c} \text{O} \\ \parallel \\ \text{R}-\text{C}-\text{OH} \end{array}$	Carboxylic Acids	0.0518	산성, 신맛, 수용성 부여
$\begin{array}{c} \text{O} \\ \parallel \\ \text{R}-\text{C}-\text{O}-\text{R}' \end{array}$	Esters	0.0682	좋은 향이 남, 섬유로 가공할 수 있는 가능성

# Conclusion

## 결론

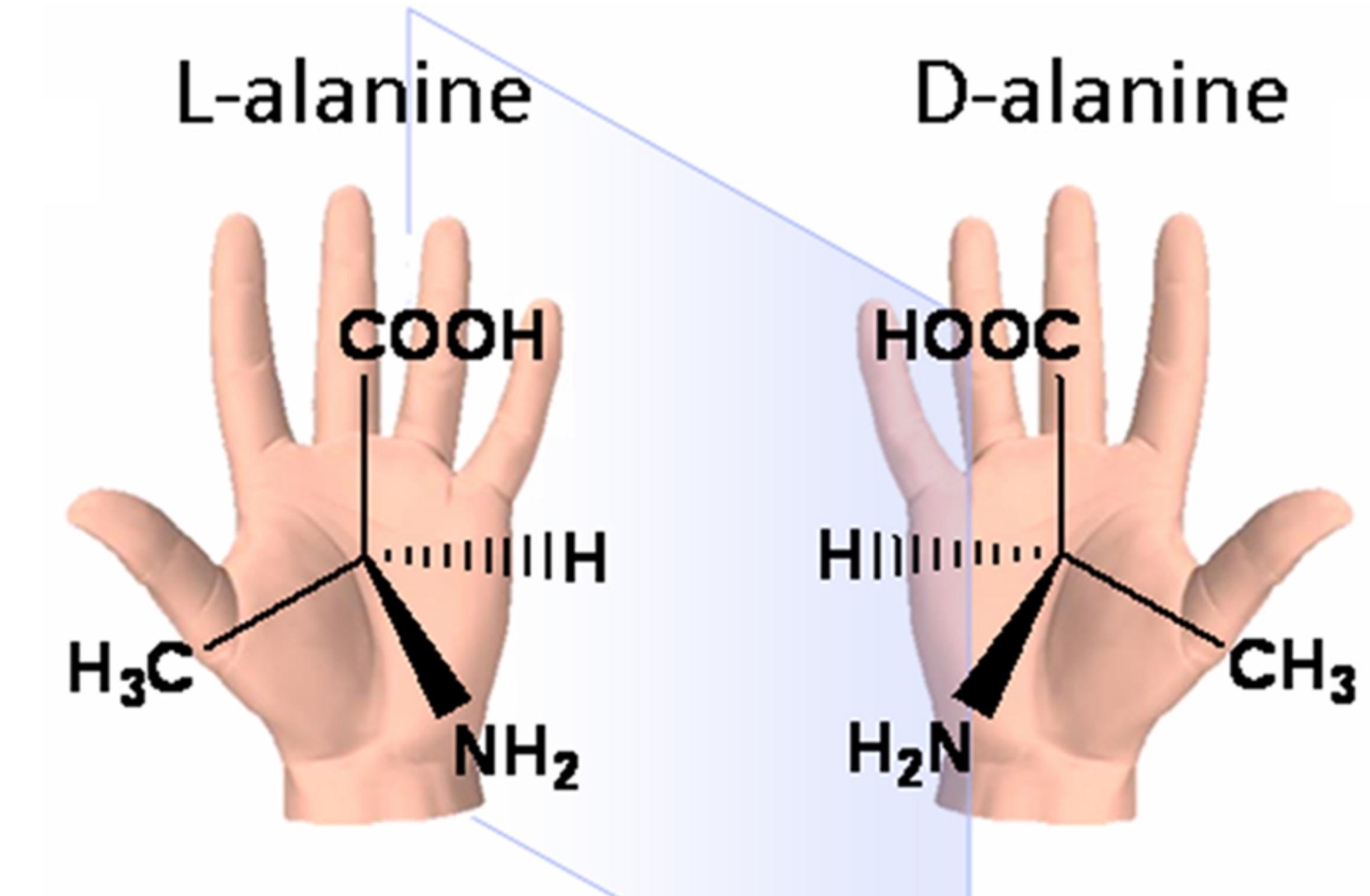
- **Inter-molecular graph**를 적절히 구현하는데 성공함
- **Inter-molecular graph**를 시각화하여
  - 유사한 문자들이 위치적으로 비슷한 지점에 등장함을 확인함
- **Inter-molecular graph**에서 centrality 분석을 통해
  - 유의미한 문자들이 높은 centrality를 보여줌을 확인함
- **Inter-molecular graph**에서 path 분석을 통해
  - **information entropy**가 비슷한 문자구조가 작용기로 나타남을 확인함

# Debation

## Limitations

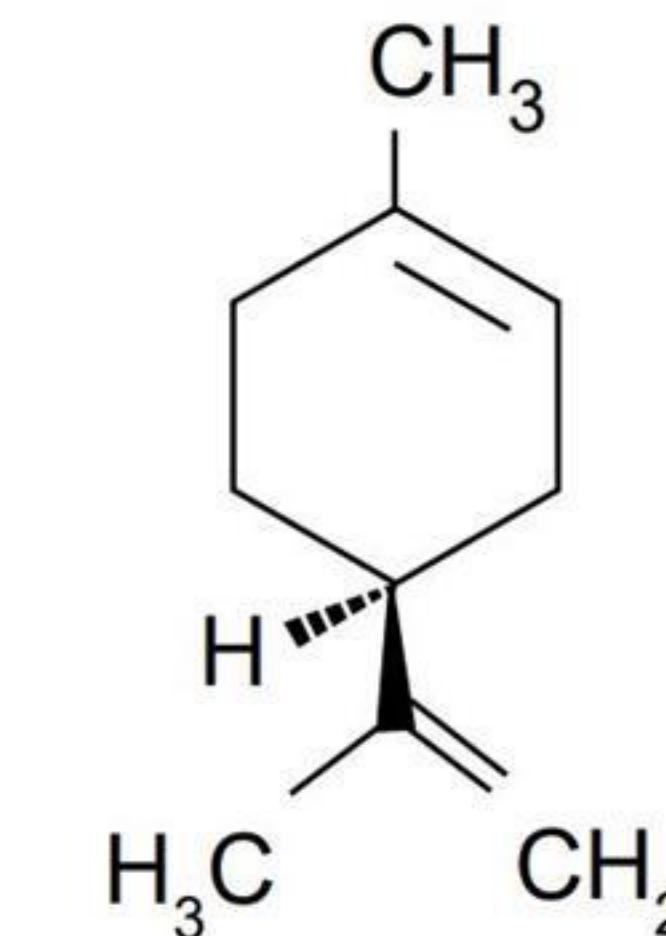
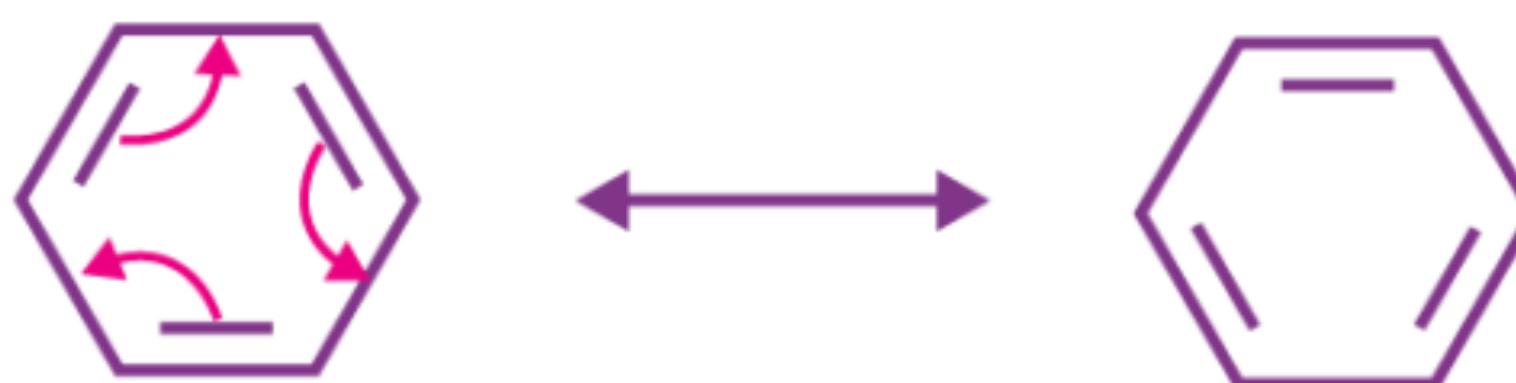
- **Chirality (SE3 group invariance)**

- 물리적 graph에서 일반적인 문제

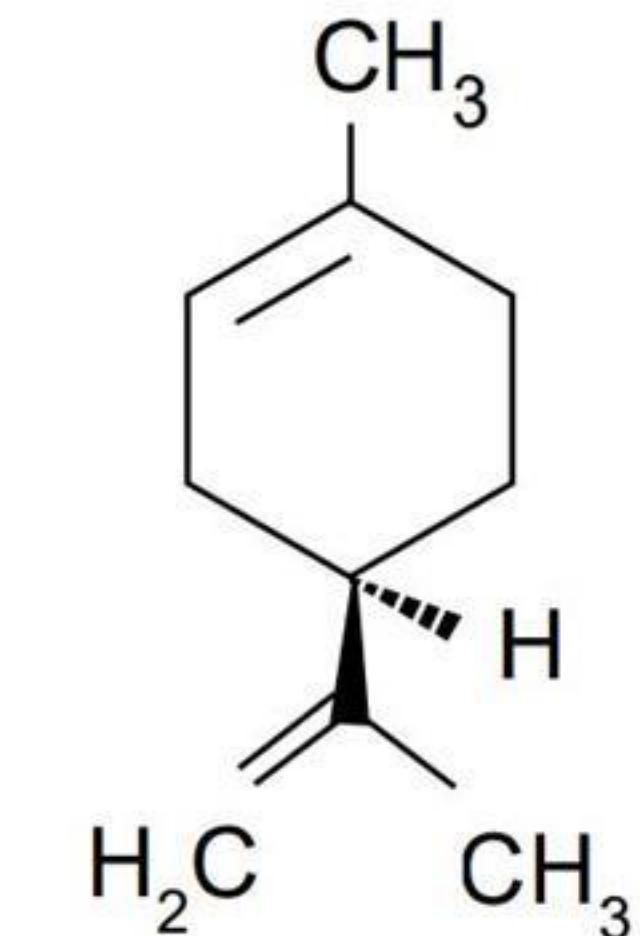


- Resonance

- 화학 graph에서의 특별한 문제



(R) - (+) - Limonene



(S) - (-) - Limonene

# Further Investigation

## Generative model

- 유사한 Information entropy를 가지는 문자들을 Typical set으로 구성할 수 있음
- Typical set과 Atypical set을 구분하여 학습시키는
  - MoE 구조를 이용해 inference 성능을 높일 수 있음
- 또한 주어진 문자 데이터만 이용하는 것이 아니라,
  - Entropy에 따라 적절한 구조를 추가적으로 제안함으로써
  - 데이터 편중 문제를 극복하거나, 데이터 augmentation 효과를 낼 수 있음

# Thanks

성균관대학교 소프트웨어학과 김산  
인터랙티브 그래프マイ닝 강의 - マイニングプロジェクト 결과발표  
2023년 11월 15일 수요일

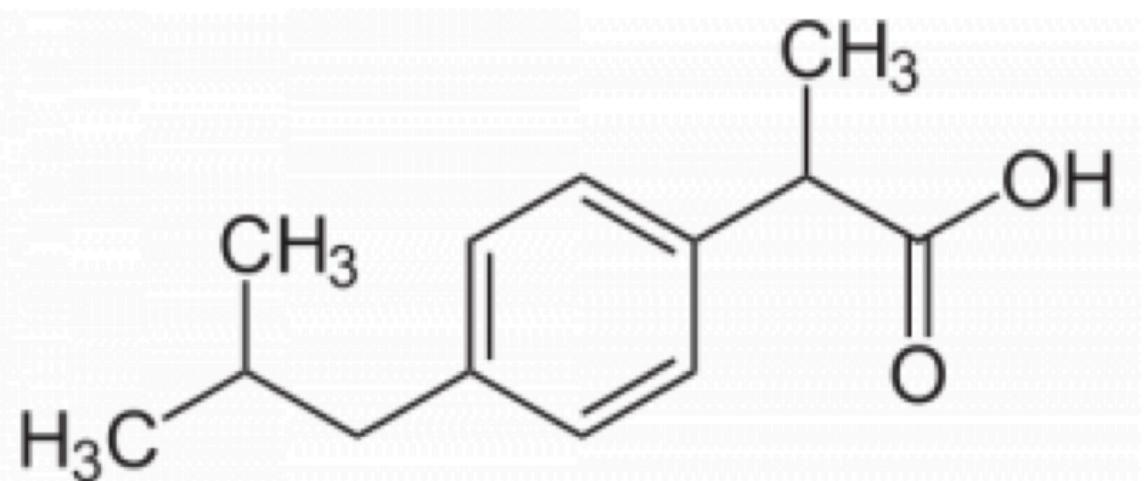
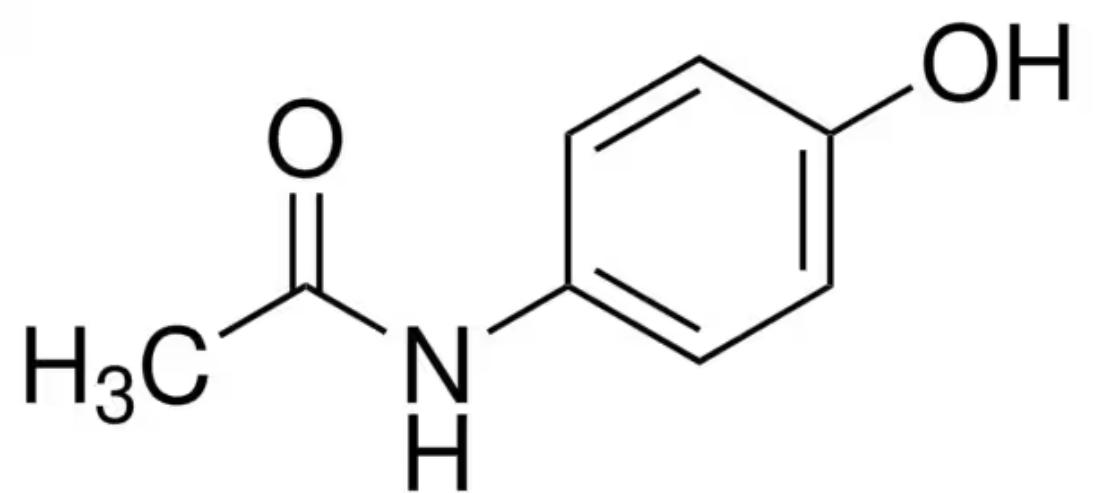
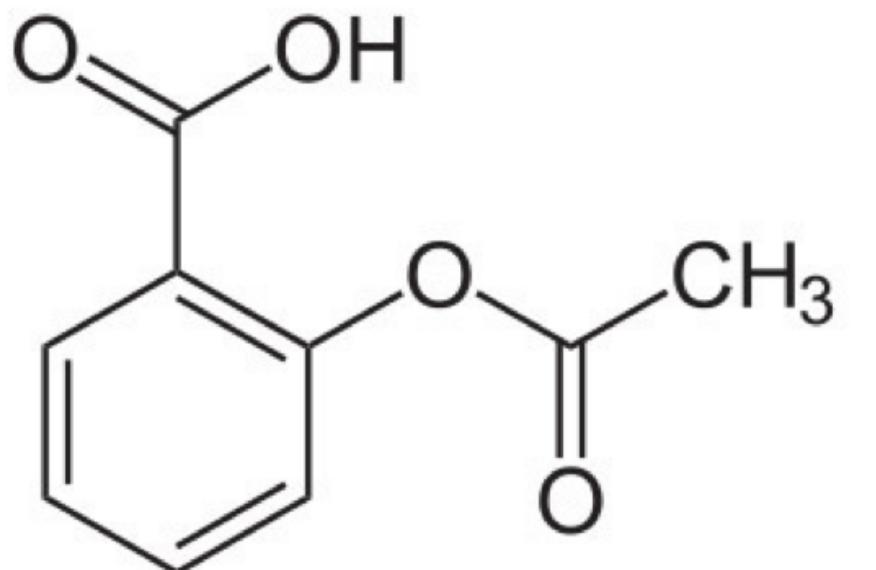
# Motivation

## Obstacles were,

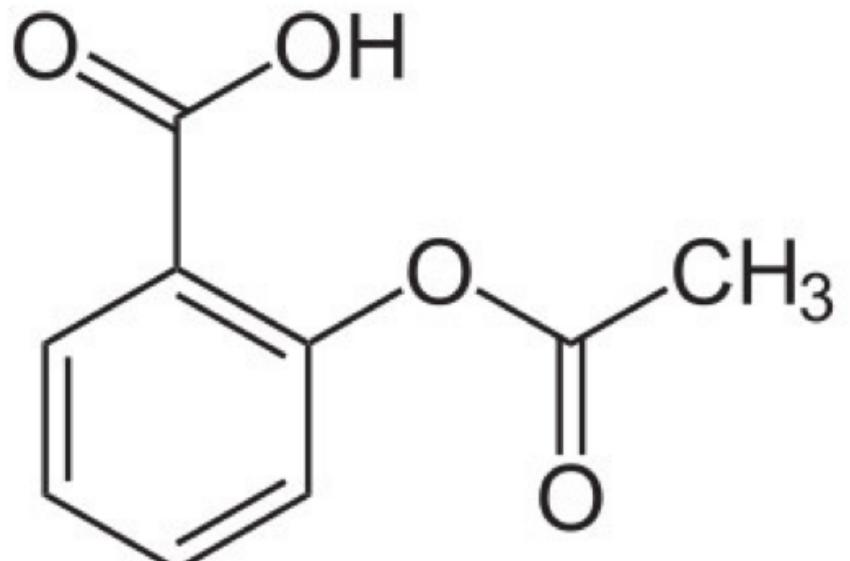
- Density of molecular subgraph relationships
- DeleteSubstructure 함수의 구조적 문제

# Introduction

## Pharmaceutics :: 제약

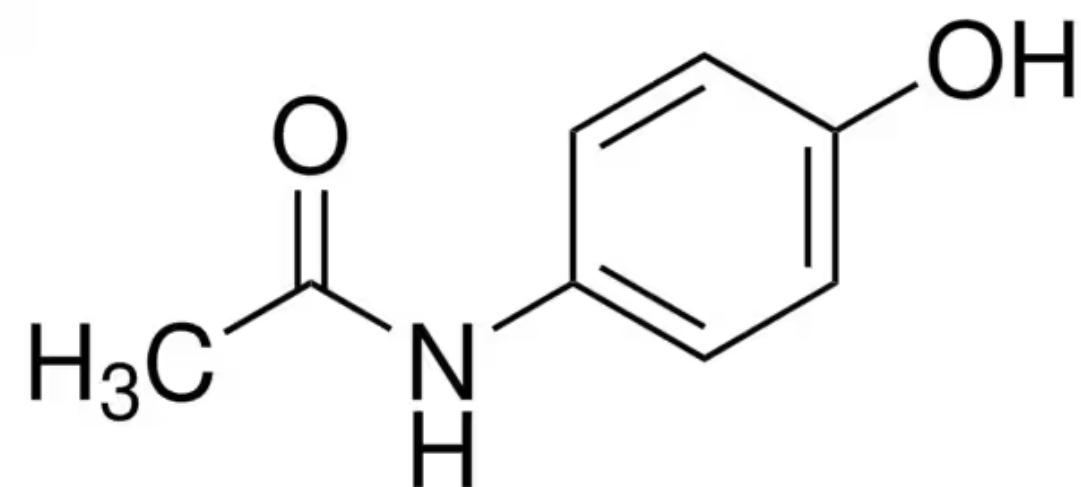


# Introduction Pharmaceutics :: 제약



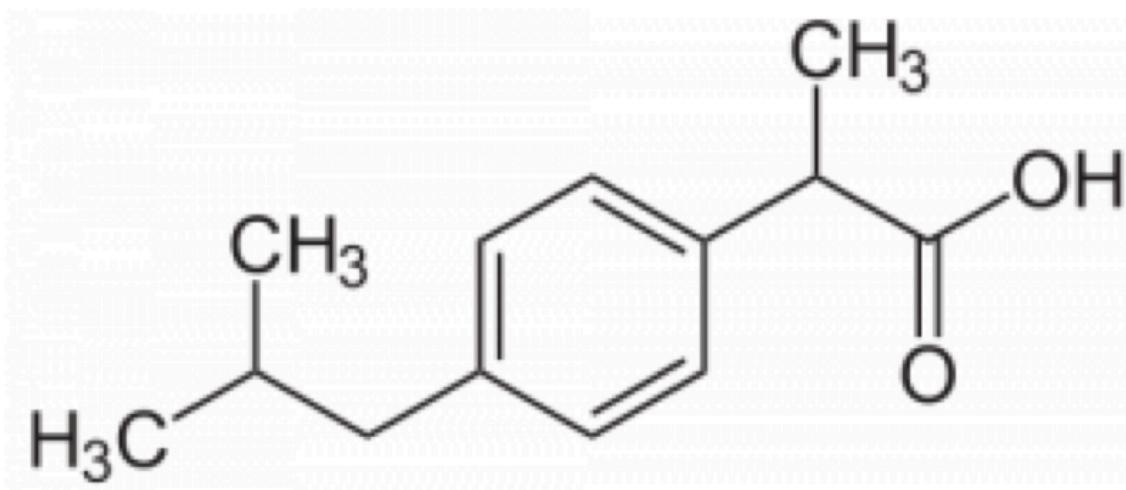
Acetylsalicylic acid

아세틸살리실산



Acetaminophen

아세트아미노펜



Ibuprofen

이부프로펜

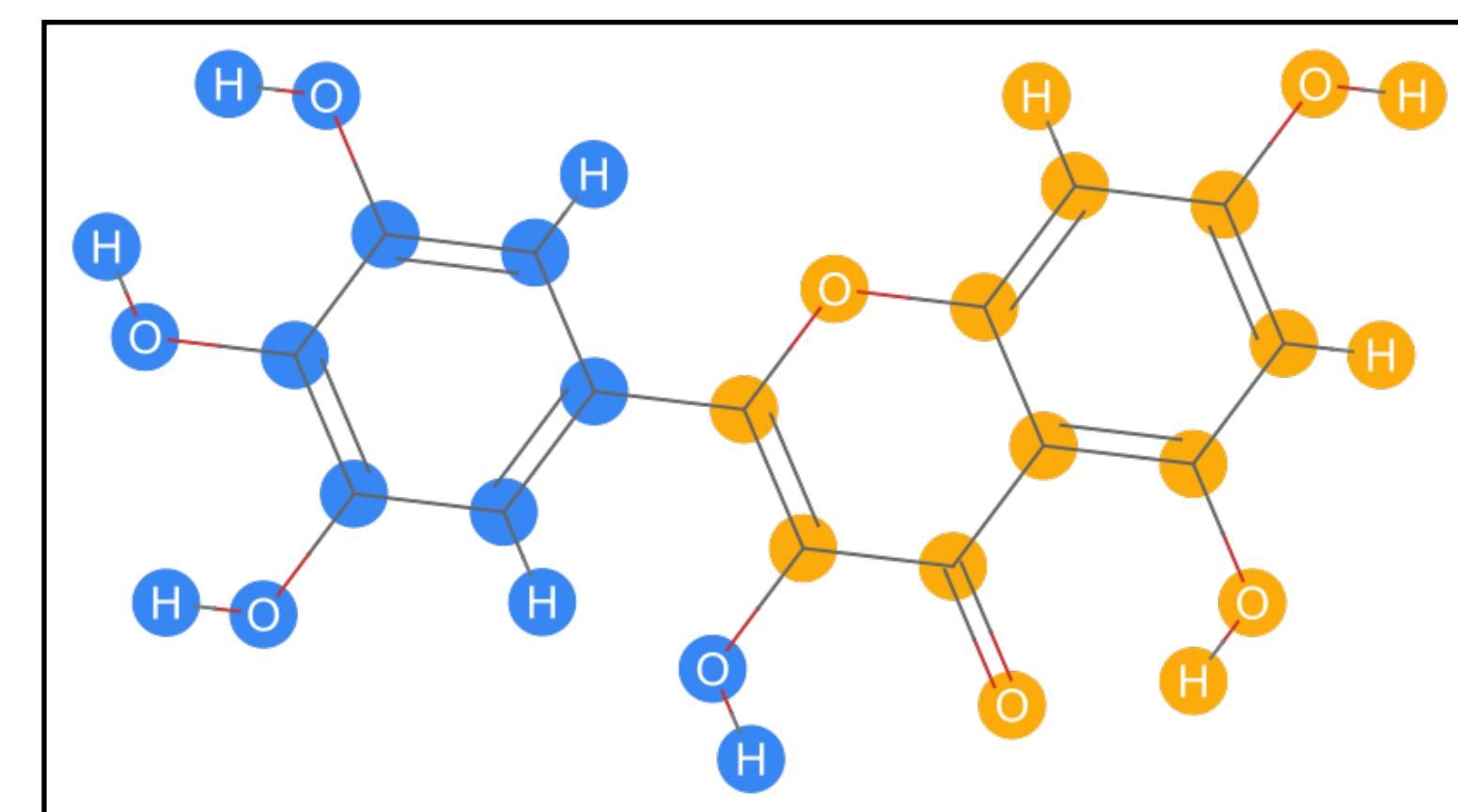


# **Research Objective**

**In this work,**

# Motivation

어떻게 chemical graphlet들을 찾을 수 있을까?



- **(Intra) Molecular graph**

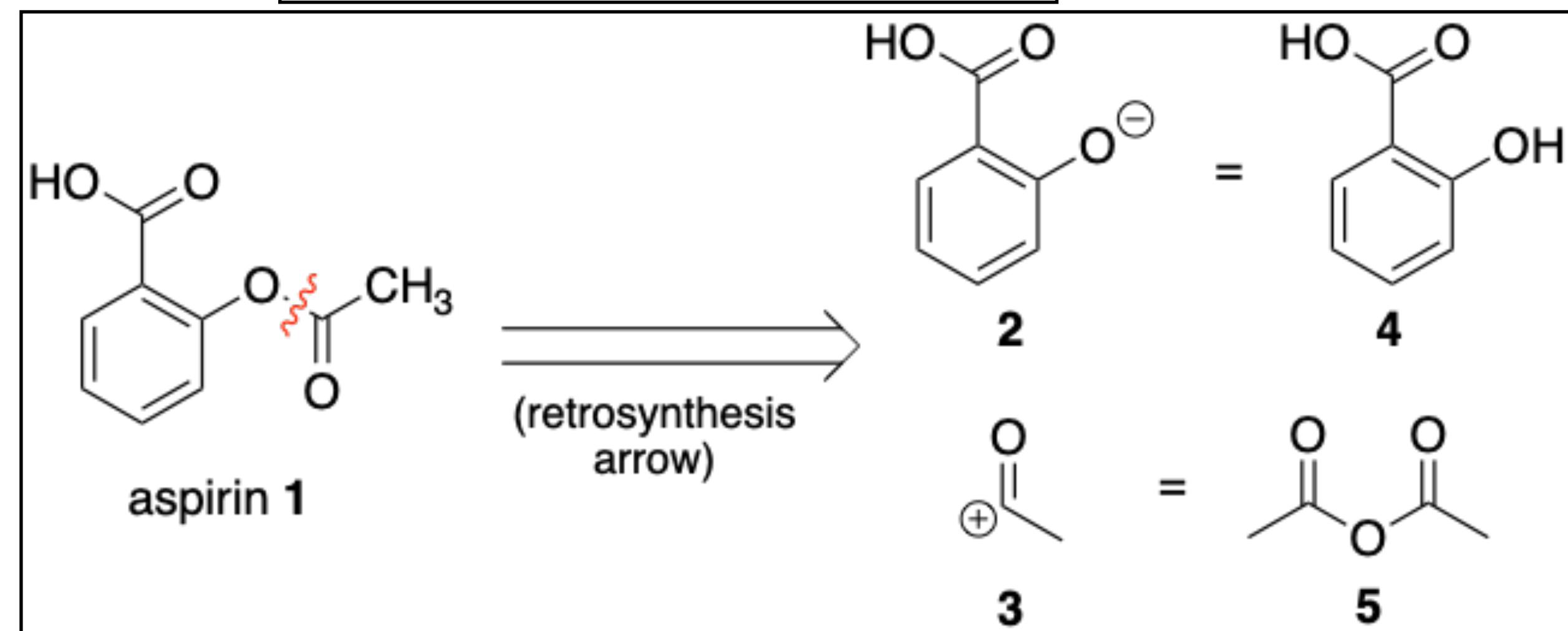
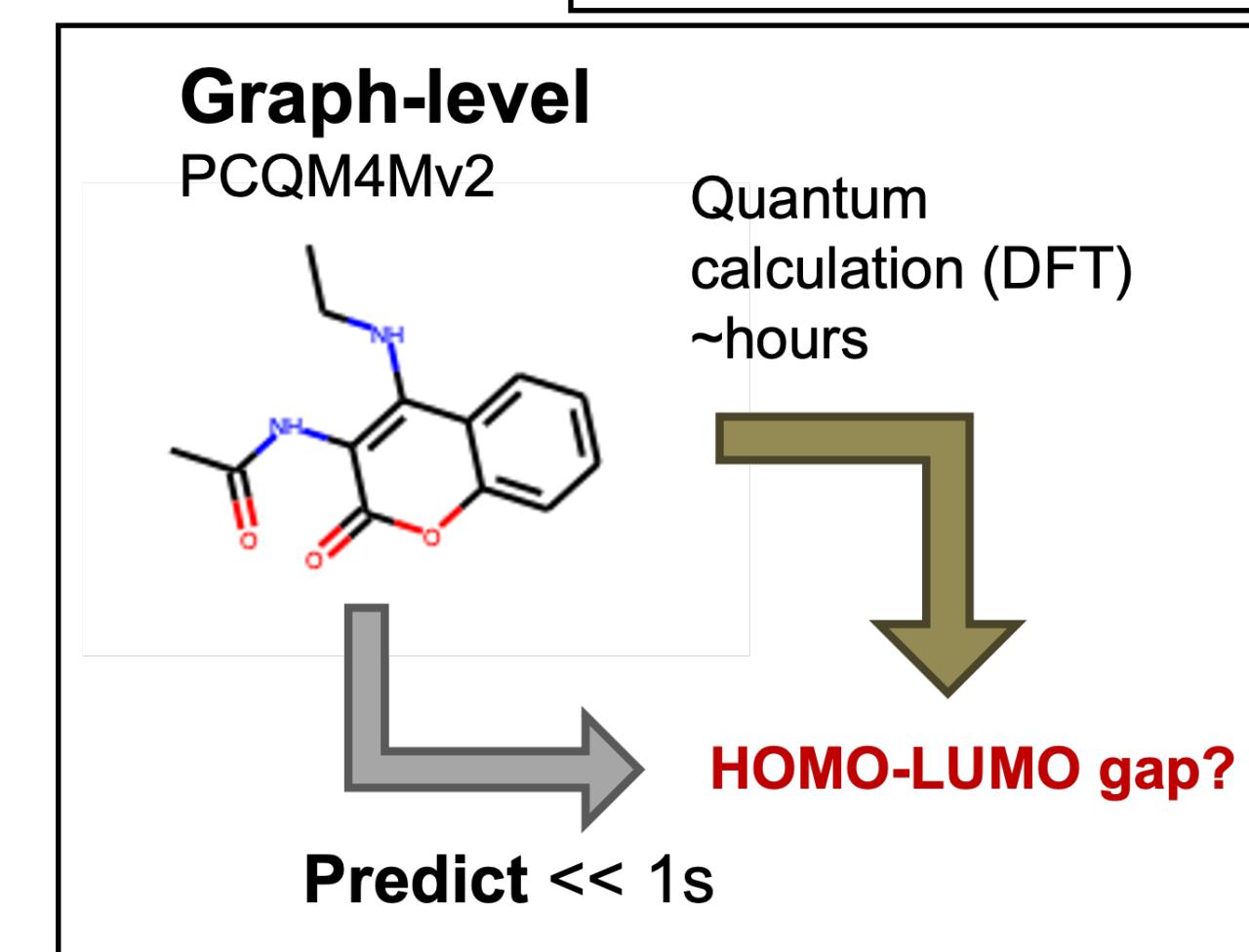
- Node = atom + atomic feature
- Edge = bond + bond feature

- **작용기 (Functional group)**

- Graphlets, Inter-molecular feature

- **역합성 (Retrosynthesis)**

- 제약, 화학공학



# Dataset

QM7 dataset → **Inter-molecular graph**

- **Quantum Machines 9 (QM9) dataset**
  - Nodes
    - Atoms
  - Edges
    - Molecular bonds  
(Valence Shell Electron Pair Repulsion theory based)
  - # of graphs
    - 7165 graphs(= molecules)
    - MAX 23 nodes(= atoms) per graph
      - MAX 7 heavy atoms per graph
- **Inter-molecular graph**
  - Nodes
    - Molecules
  - Edges
    - **Subgraph relations ( $\Delta_{atom} = 1$ )**
  - 1 large graph
  - 7165 nodes(= molecules)

# Method

## Sample pathways from Eigen-central nodes

- **Inter-molecular graph**의 Eigenvector-centrality 계산
  - 주요 backbone 문자구조 indentify
  - 모든 path를 계산하는건 computational/physical 문제가 있기 때문에 중심노드 선정
- **Path** 정보 수집
  - Eigenvector-center로 부터 커지는 방향으로 원자들을 붙여나가면서 지나가는 edge 정보 수집
- **Path dataset** 분석
  - Hierarchical clustering을 통해 자주 등장하는 path를 hierarchical 분류

# Expected Result & Conclusion

## Hierarchical clustering of functional groups

- 예상 결과
  - Inter-molecular graph를 정의하고 구축
  - Chemical path를 hierarchical clustering
    - 기존에 알려진 작용기와 비교 분석
- 의의
  - Inter-molecular feature를 graph mining 기법으로 발견하려는 시도
  - Path를 역으로 밟아 retrosynthesis software에 응용 가능
  - Learning project에 사용 가능

