

Chemical Functional Group Finding in Inter-Molecular Graph

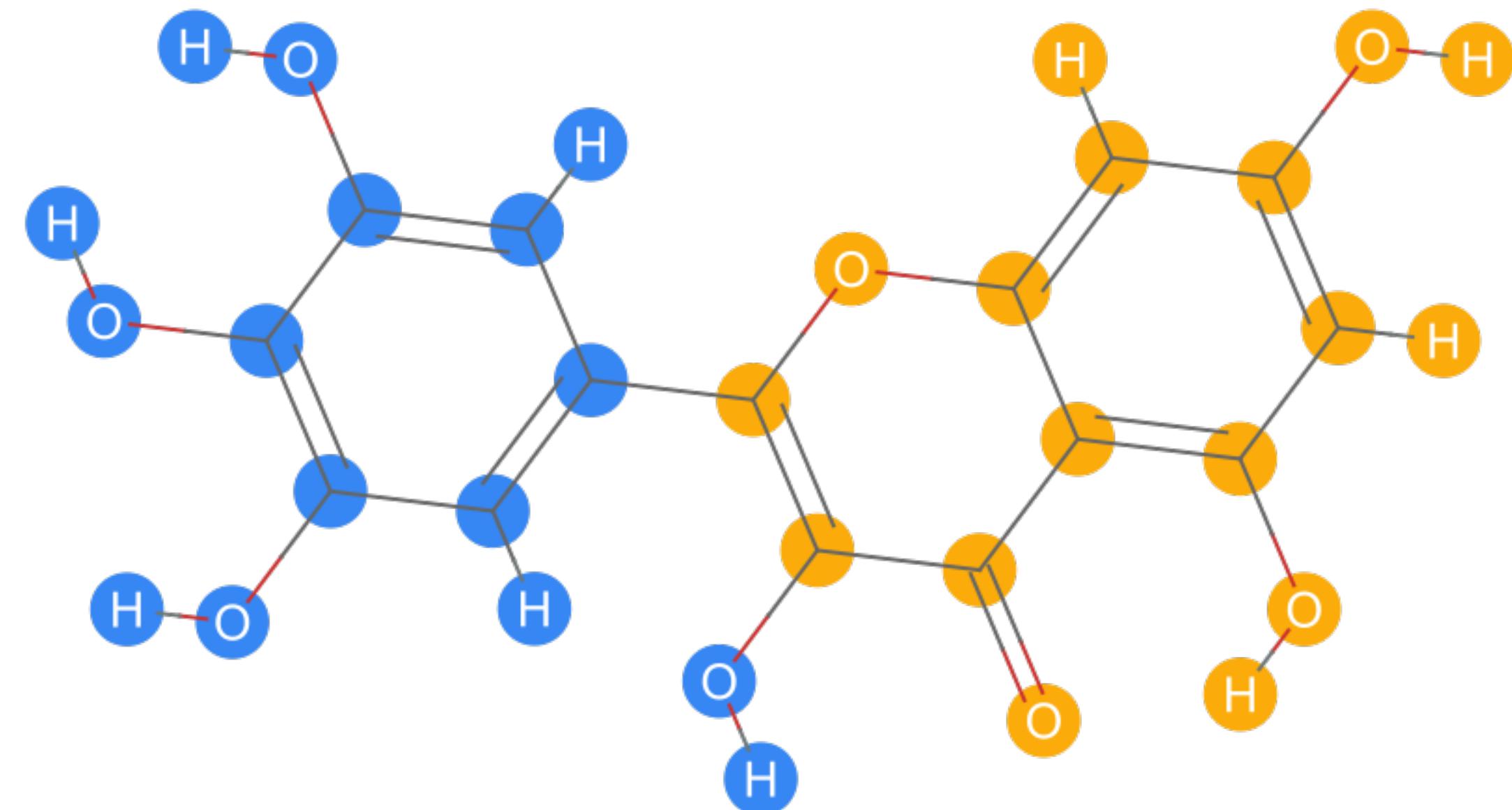
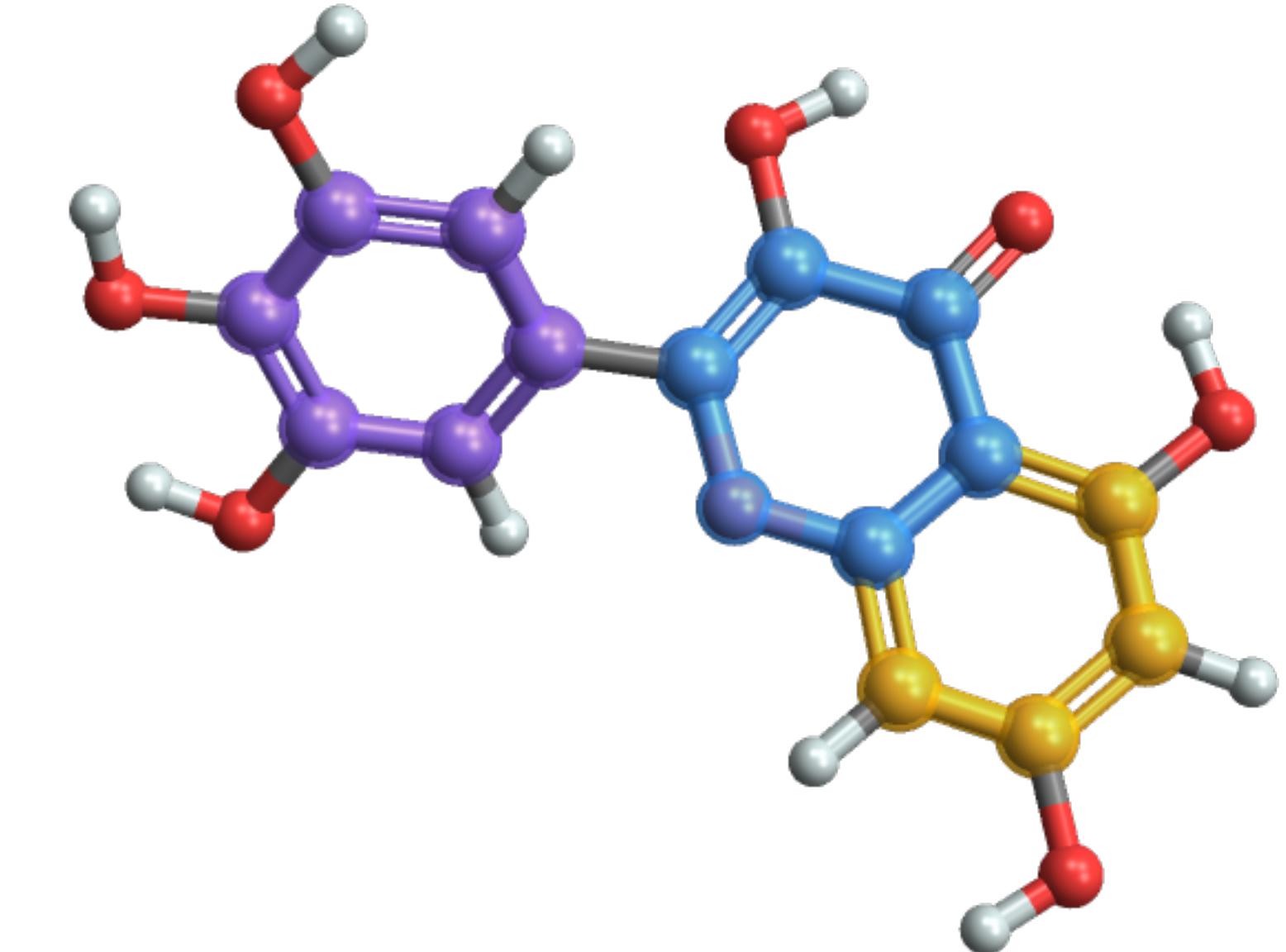
Graph Mining Project Presentation
{ Interactive Graph Mining 강의 }

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** 구축



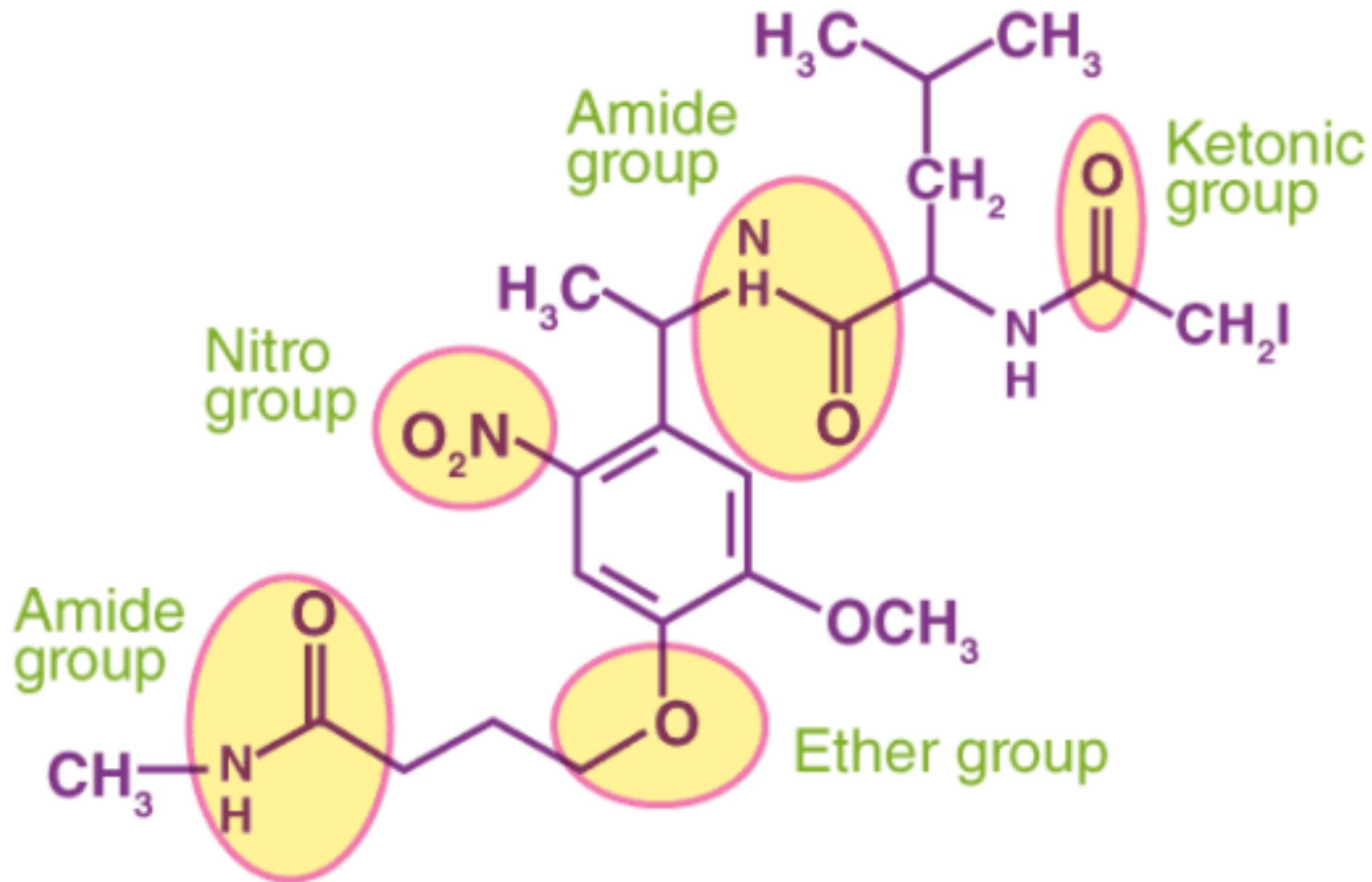
- **Inter-molecular Graph**의 성질 파악
 - Basic graph analysis
 - Centrality analysis
 - Path 분석
 - 정보 엔트로피 분석

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!

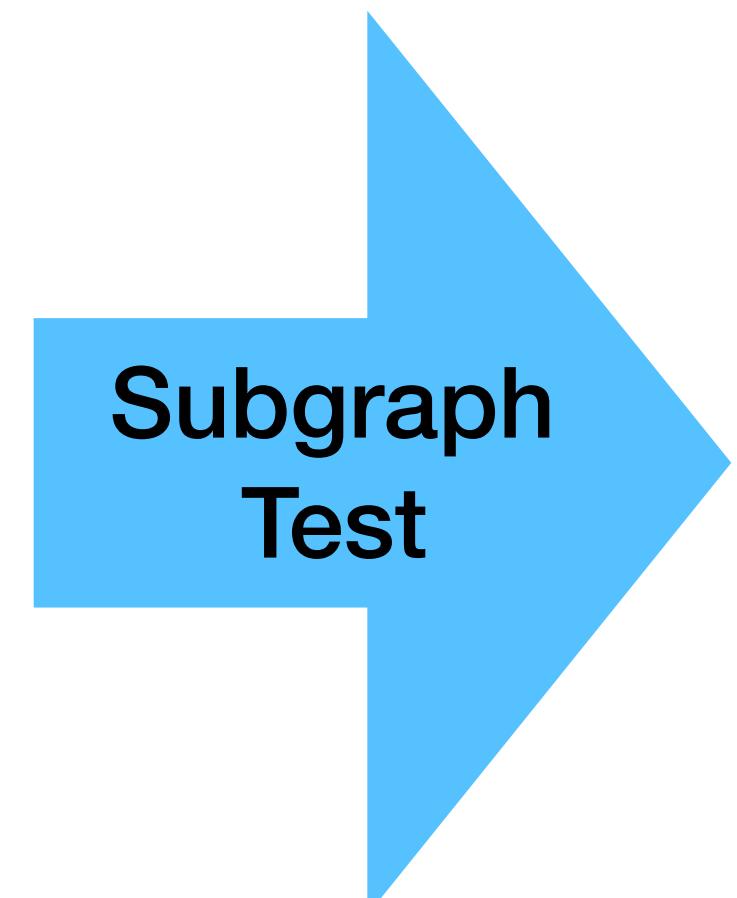


Dataset

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

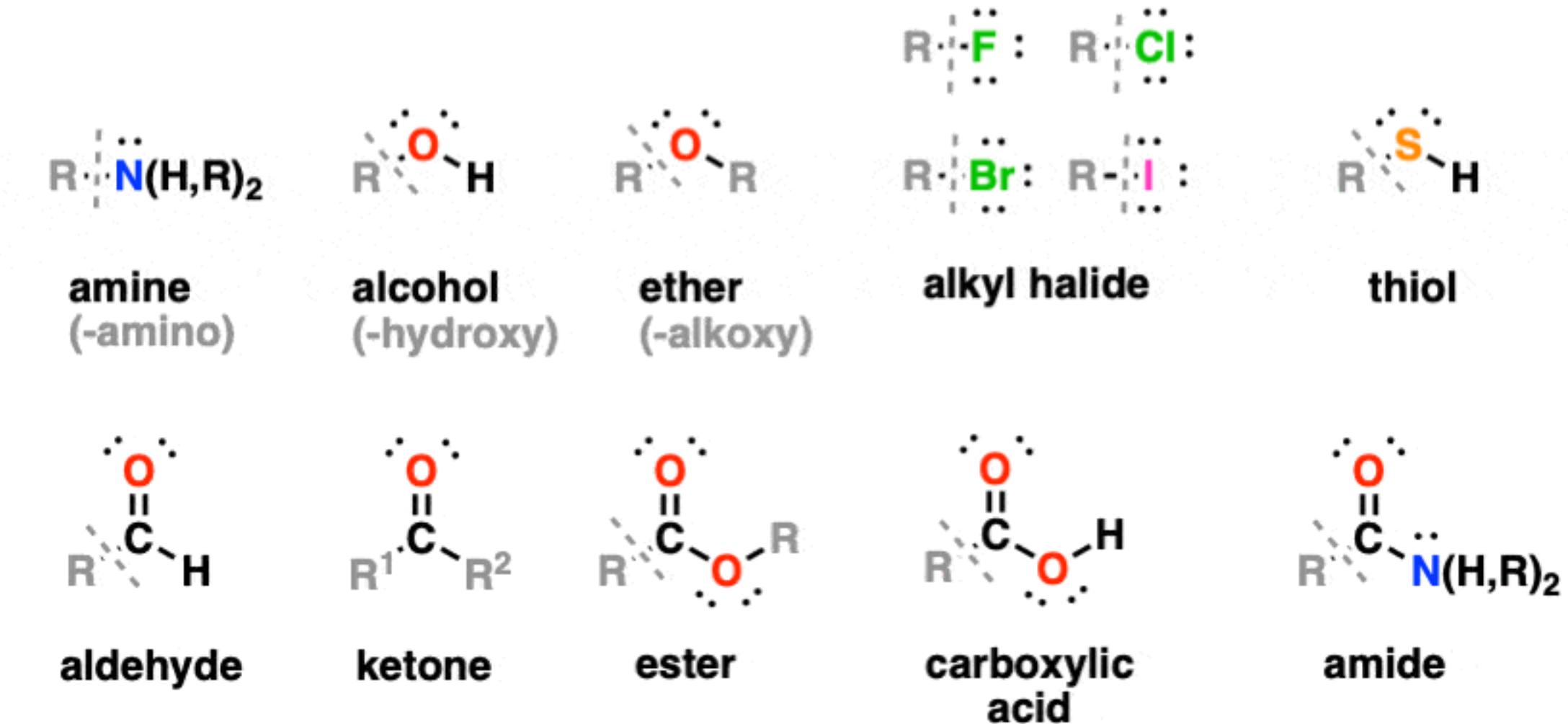


- **Inter-molecular graph**

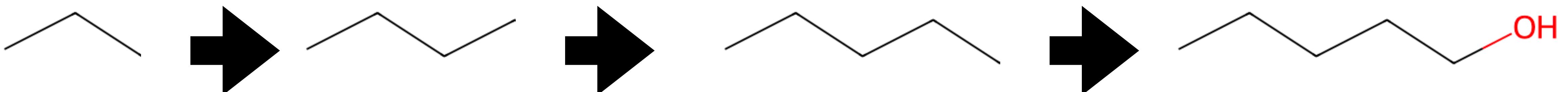
- Nodes
 - Molecules
- Edges
- **Subgraph relations
(Δ atom = 1)**
- 1 large graph

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

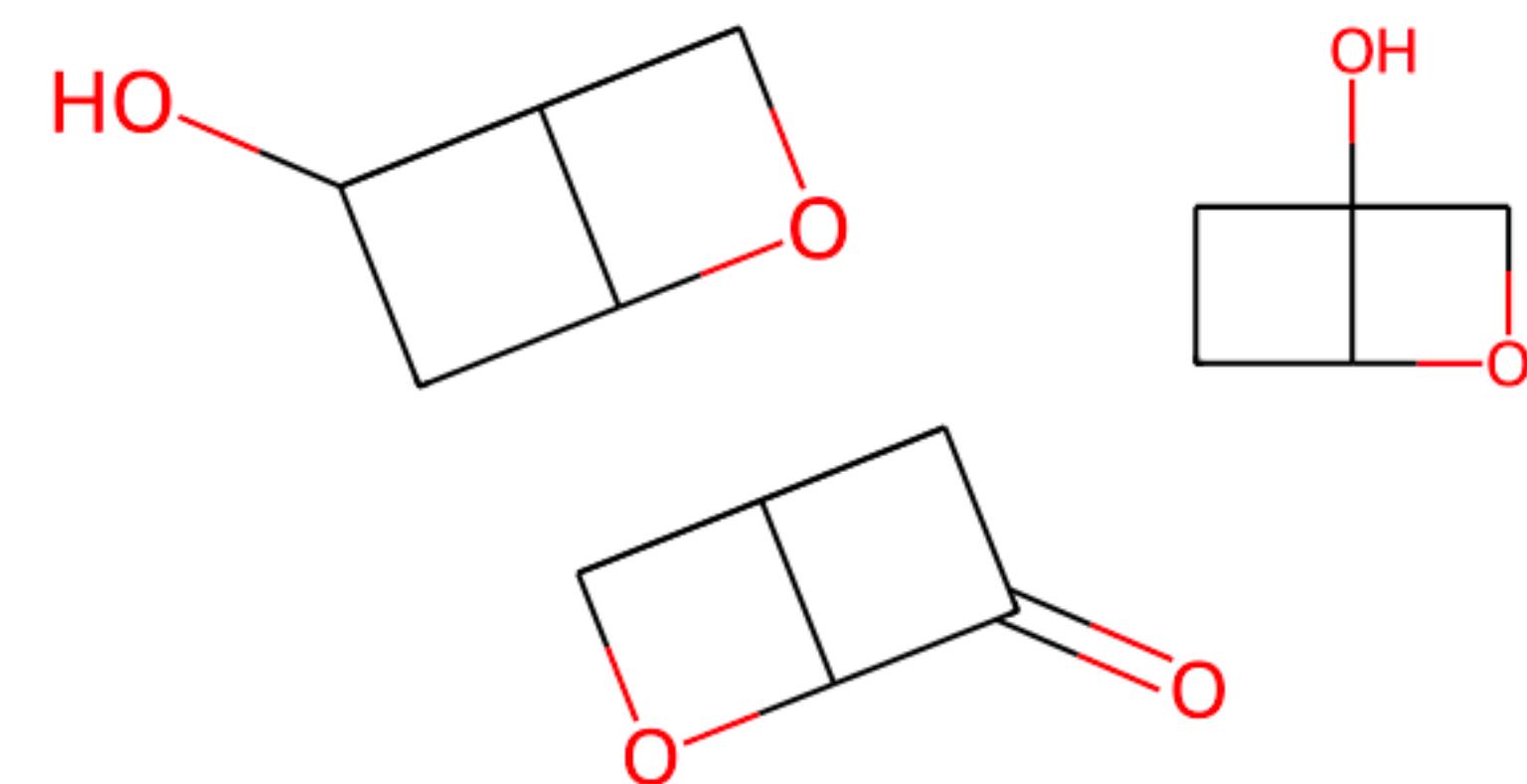
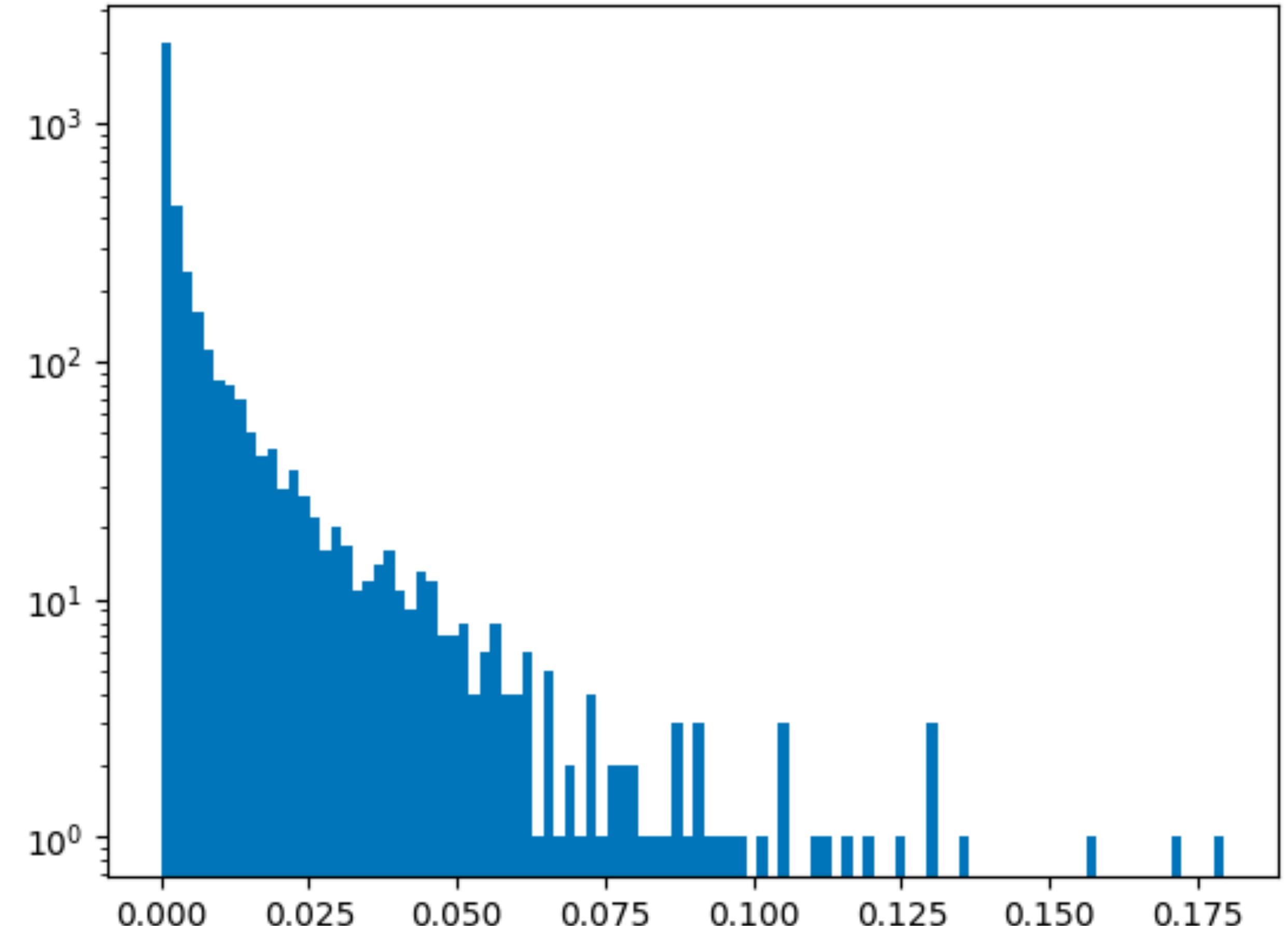
• Node Count	3869
• Edge Count	23470
• Edge Density	0.00156
• Avg. Degree	12.132
• Avg. In Degree	6.0661
• Avg. Out Degree	6.0662

*원자 갯수 차이가 1개만 나도록 제한
⇒ 적절한 density를 가지고 있는 것으로 보임

Results

Eigenvector centrality analysis

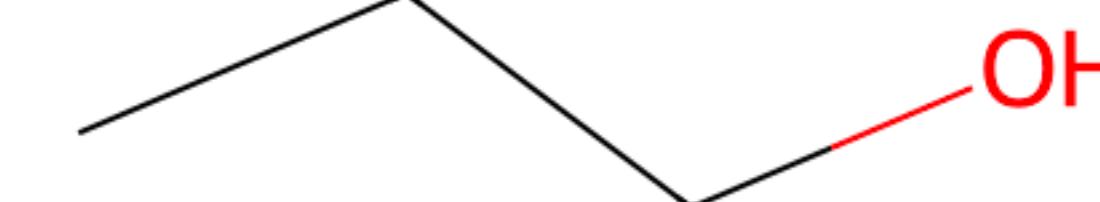
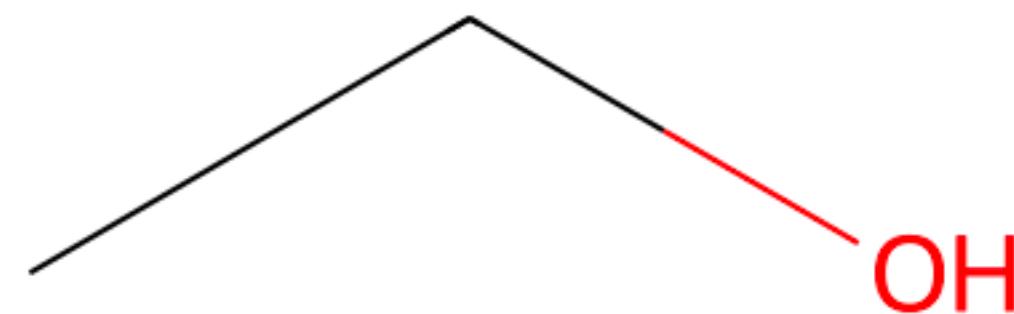
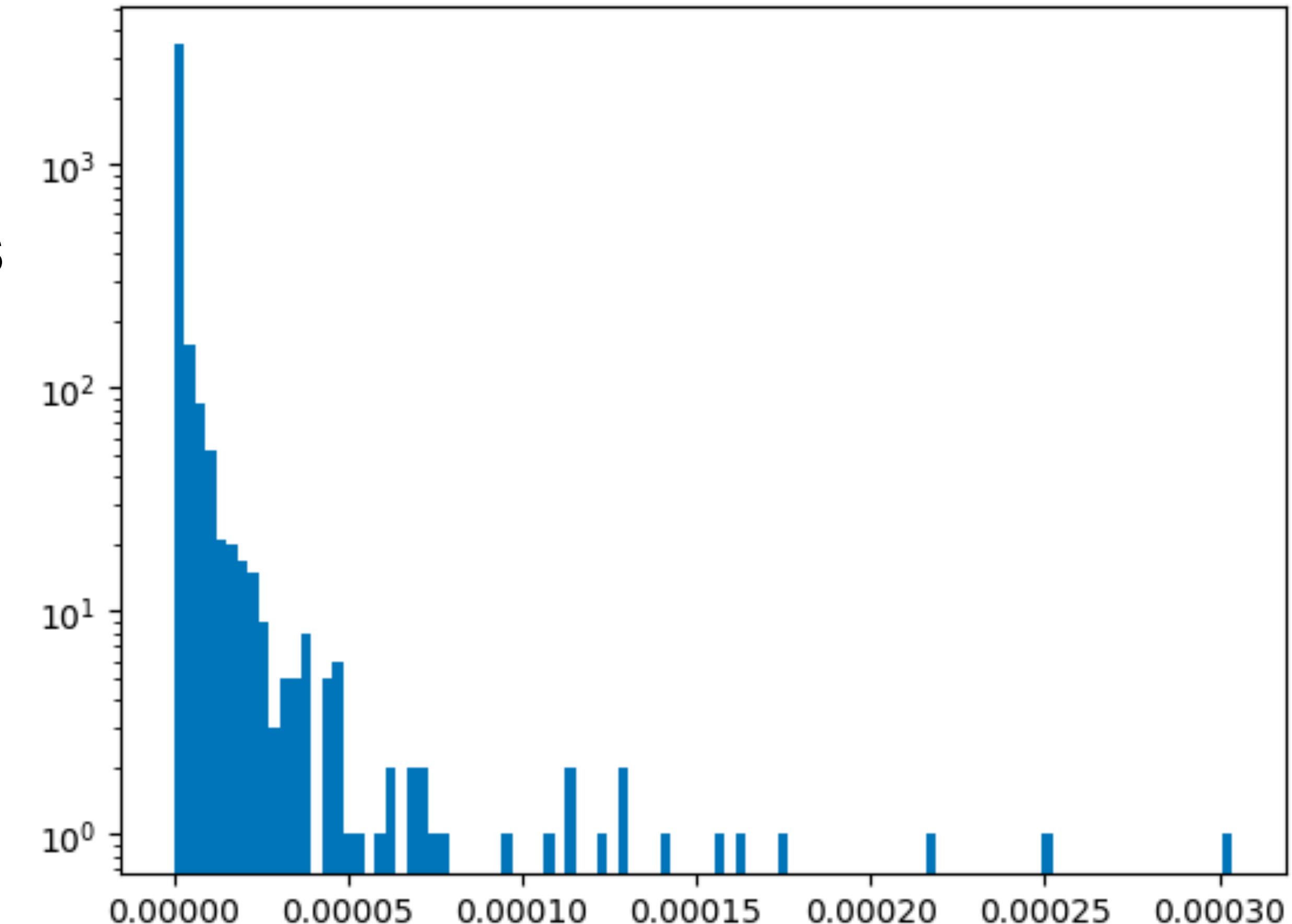
- 중요도가 높지 않은 분자들이 등장함
- Asymmetric graph에서 Eigen decomposition을 하다 보니 생긴 문제로 보임



Results

Betweenness centrality analysis

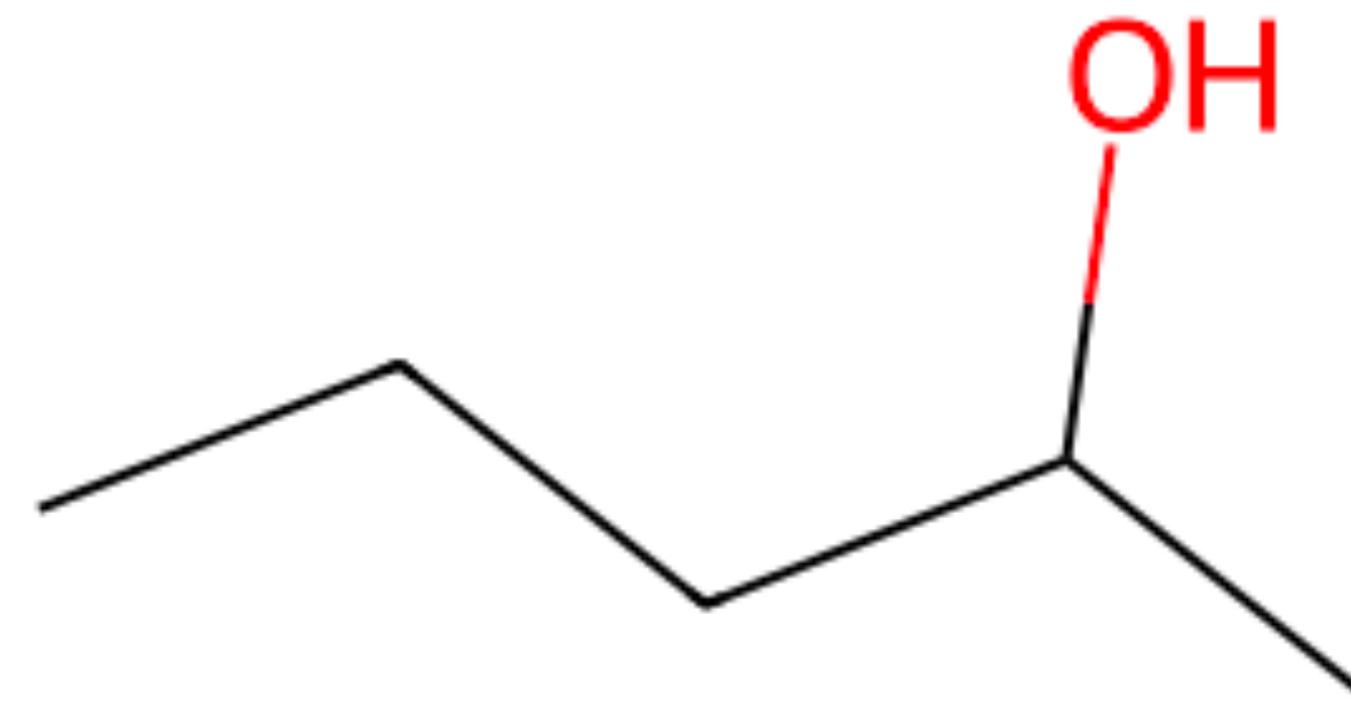
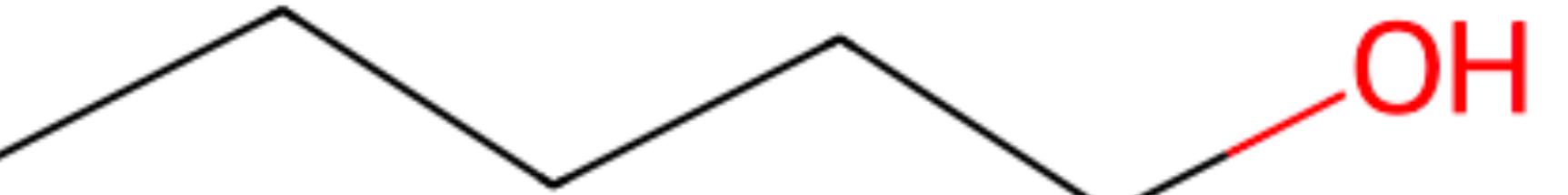
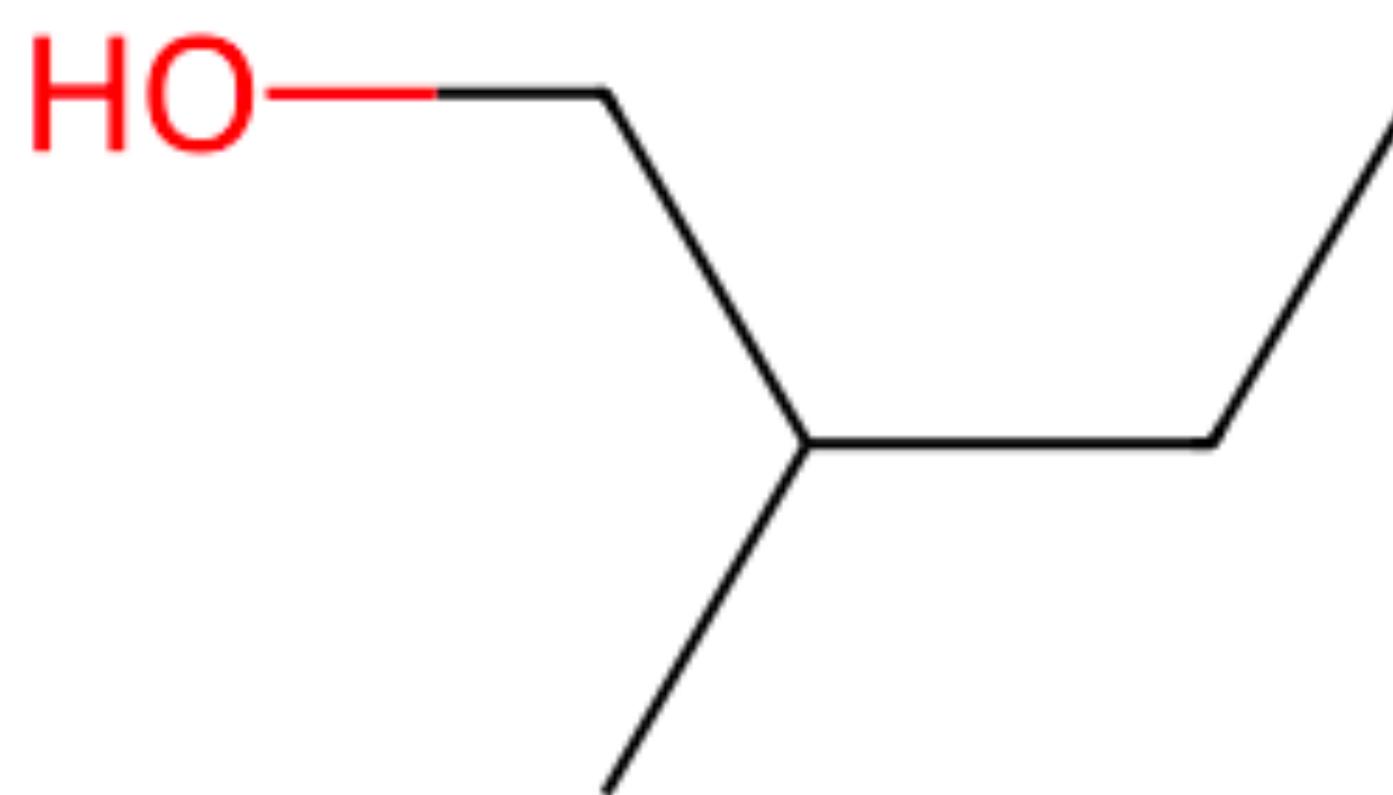
- 비교적 단순한 문자들이 등장함
- 자식 노드를 많이 가지고 있는 문자들을 확인할 수 있었음



Results

Pagerank Score analysis

- 화학적으로 유의미한 분자들이 많이 등장 함
- 대칭 graph로 변환해 분석한 덕분



Results Visualization

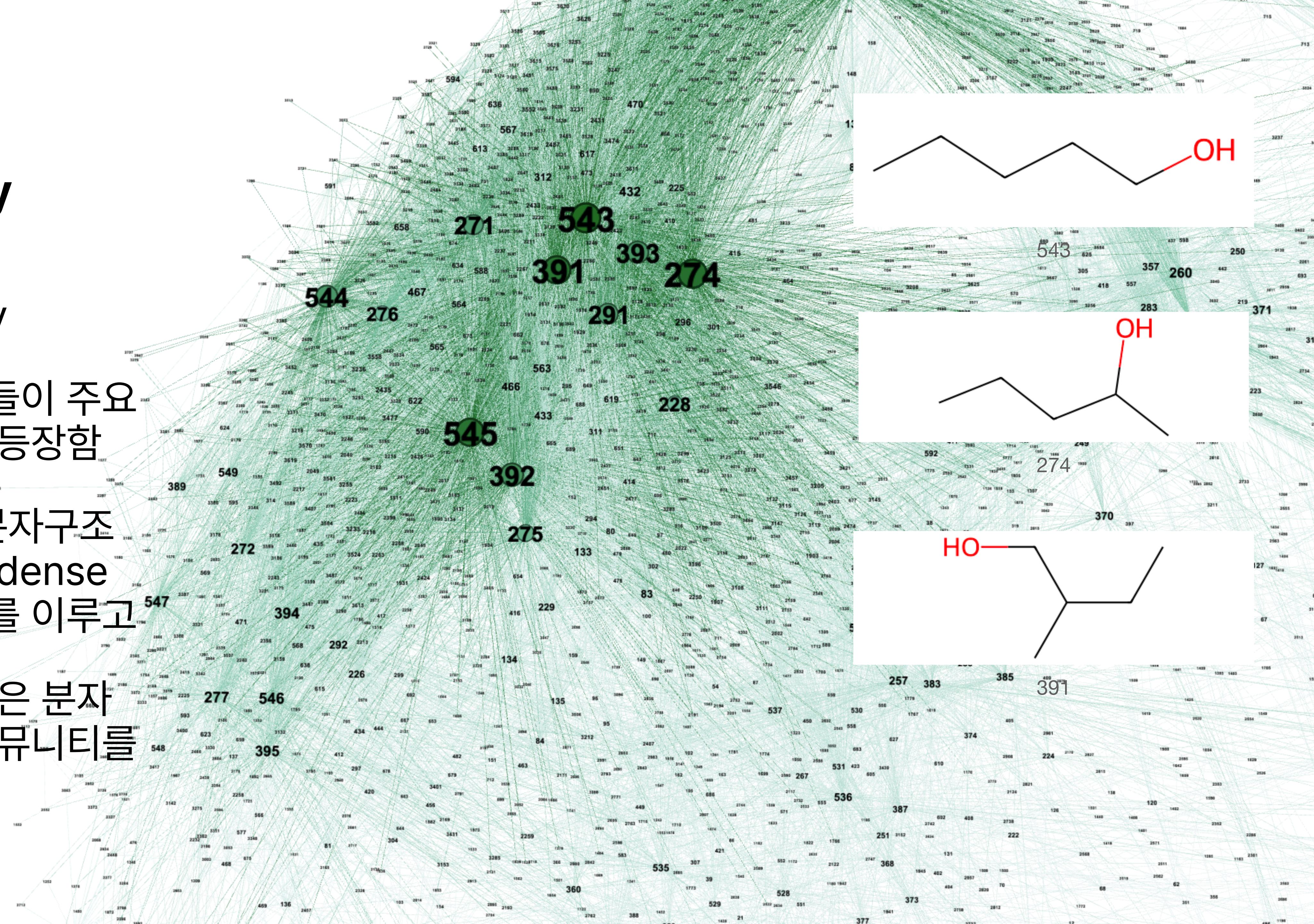
- 일부 문자들이 dense한 커뮤니티를 이루고
- 이후에 더 작은 문자들이 작은 가지들로 뻗어나가는 구조



Results

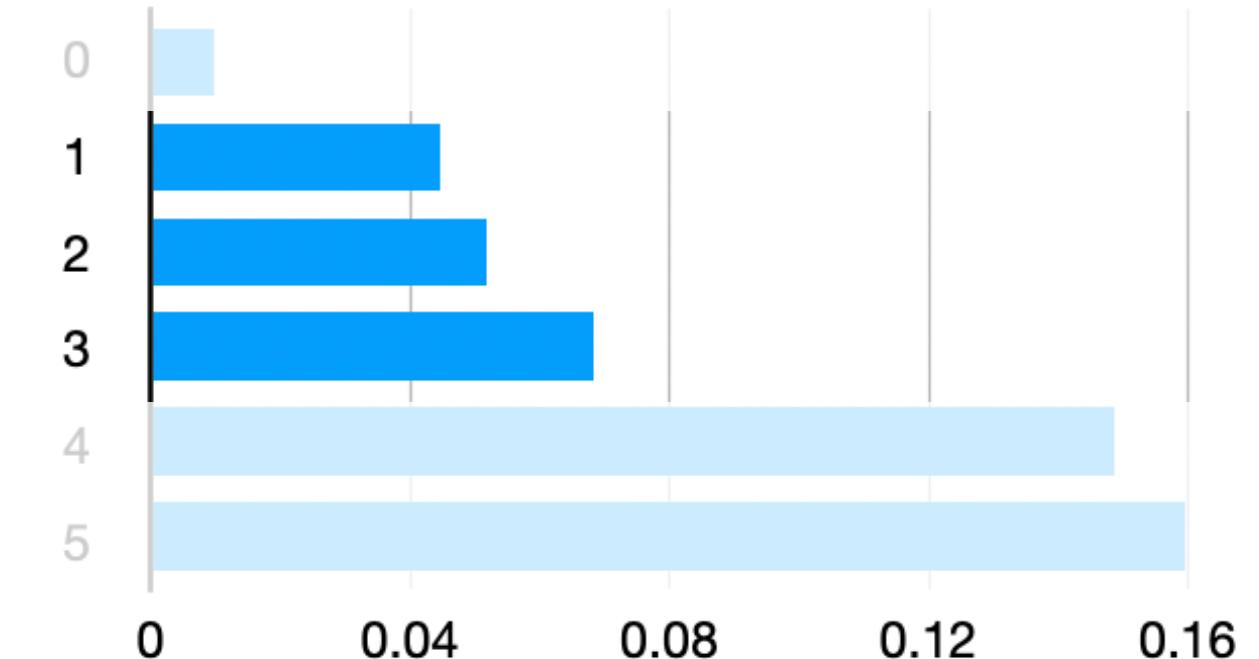
Community

- Community
- 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

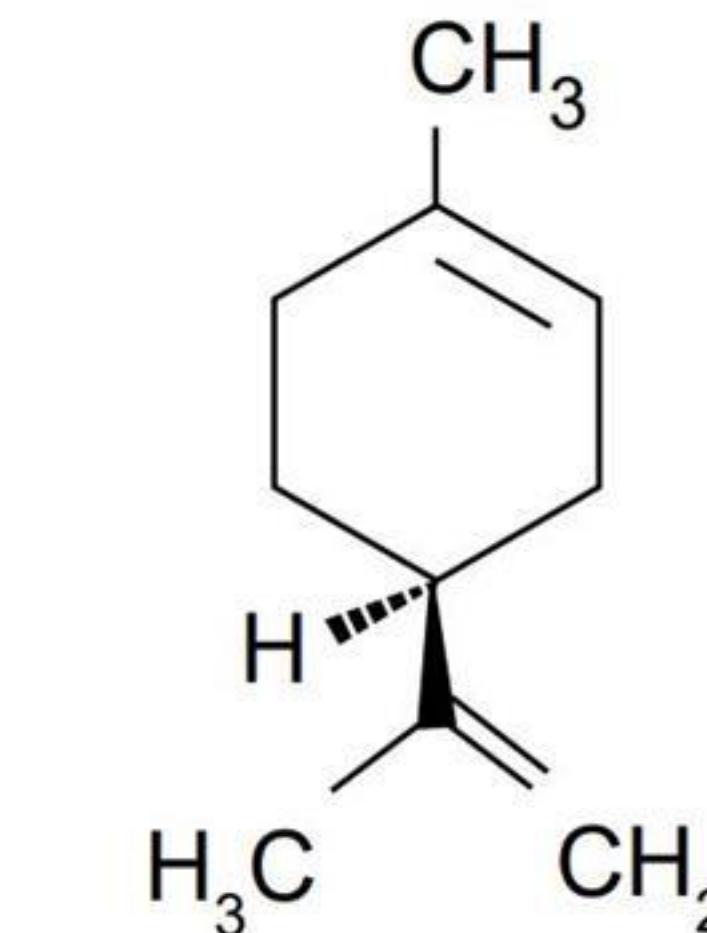
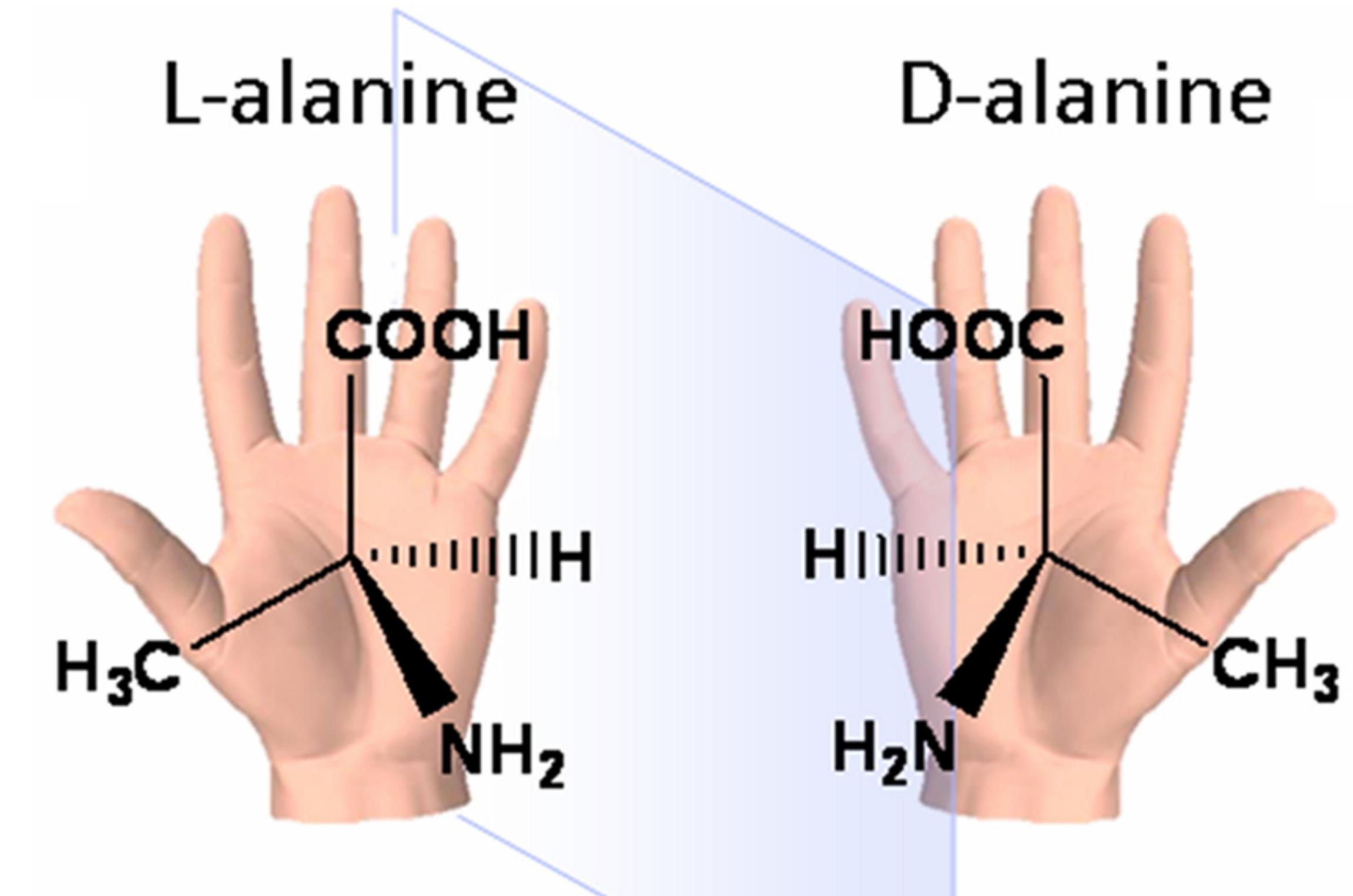
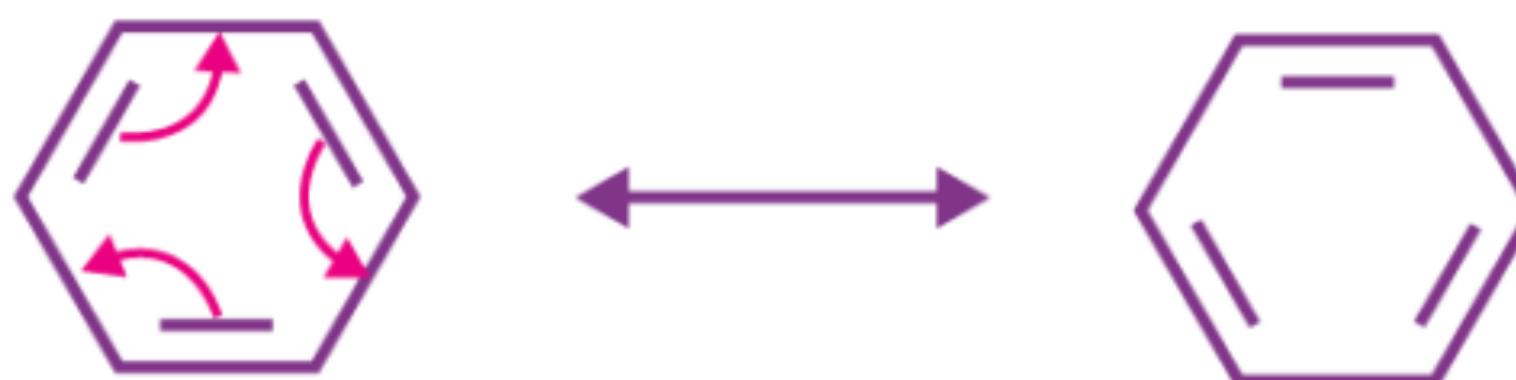
결론

- **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 entropy가 비슷한 문자구조가 작용기로 나타남을 확인함

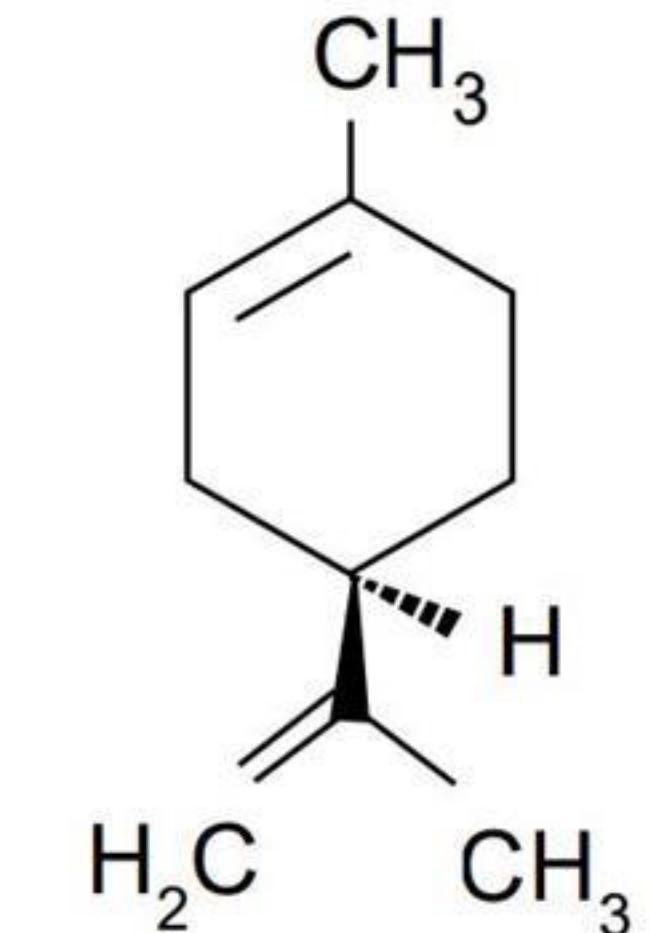
Debation

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