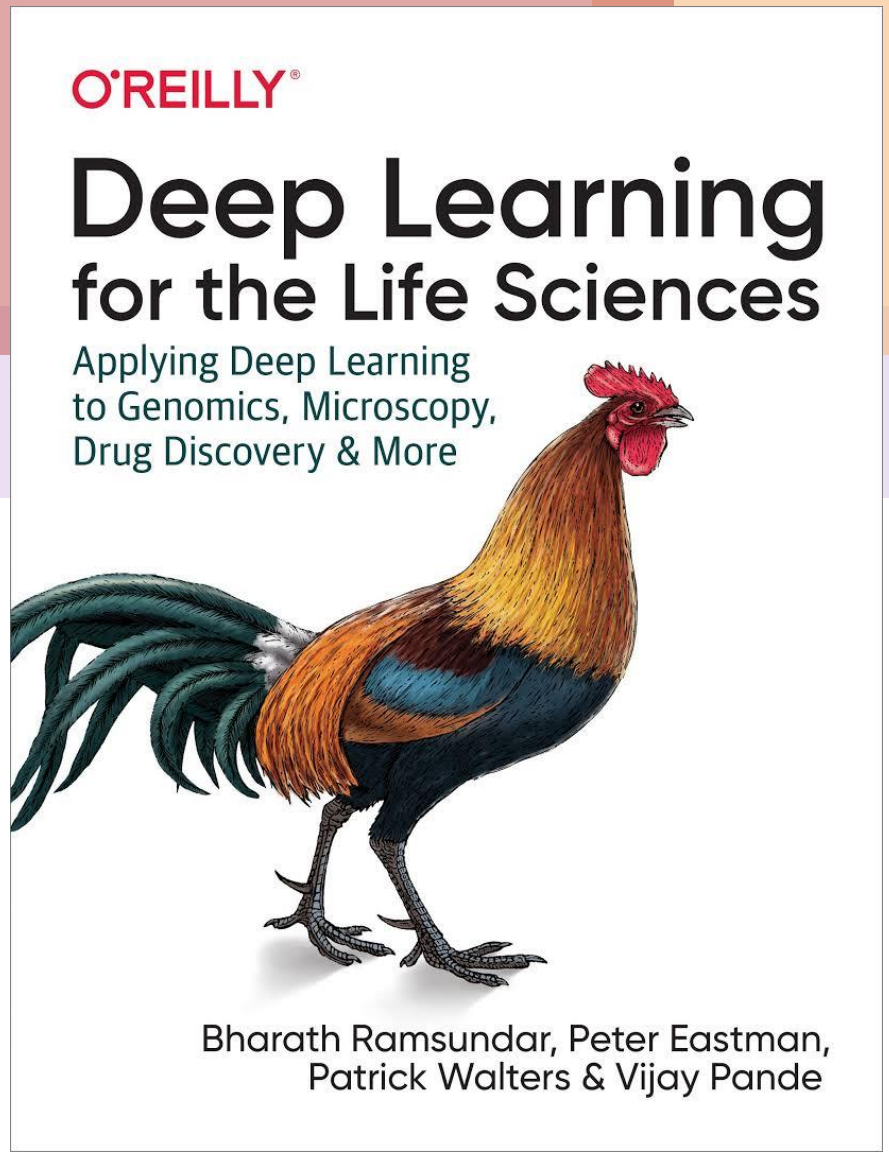


라 가 영

Deep Learning for the Life Sciences

Chapter 4. Machine Learning for Molecules

Introduction

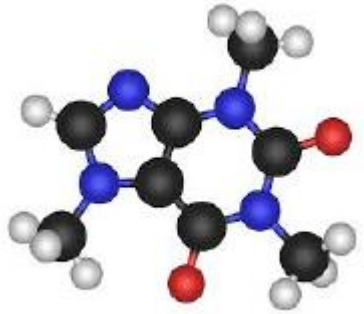
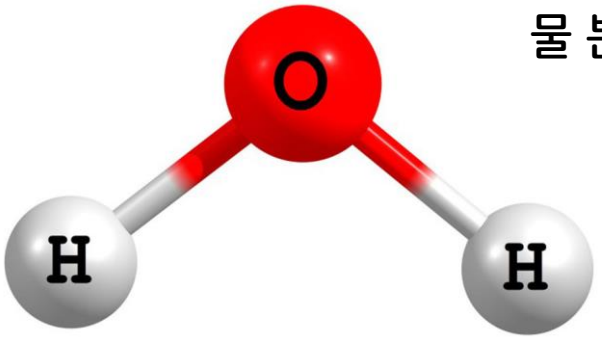


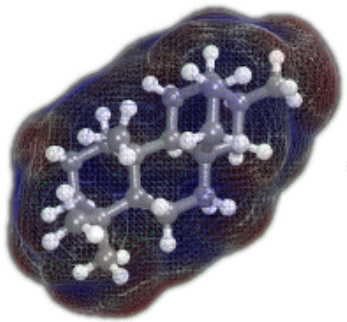
Figure 4-2. A simple representation of a caffeine molecule as a “ball-and-stick” diagram. Atoms are represented as colored balls (black is carbon, red is oxygen, blue is nitrogen, white is hydrogen) joined by sticks which represent chemical bonds.

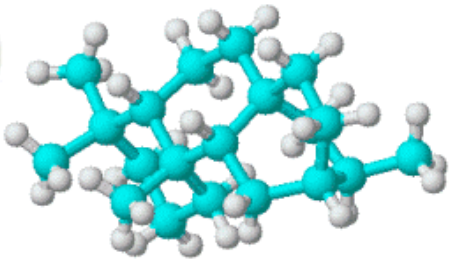
- 이 장에서는 molecular data를 이용하여 기계 학습을 수행하는 데 필요한 기본 사항에 대해 알아보고자 함
  - Data preparation -> 복잡한 molecular data를 벡터로 변형시킴 (molecular featurization)
    - Chemical descriptor vectors, LD graph representations, MD electrostatic grid representations, orbital basis function representations, …
  - Modeling
    - Fully connected networks, …

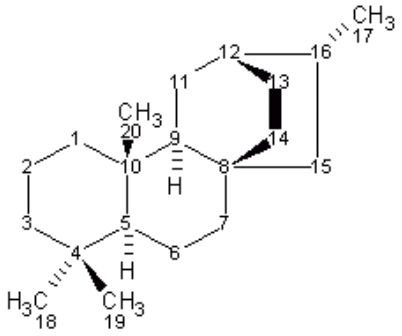
What Is a Molecule?



물 분자







테르페노이드 분자

- 물리적 힘에 의해 결합된 원자그룹
- 화학 반응에 참여 할 수 있는 화합물의 가장 작은 기본 단위
- 분자 내의 원자는 화학 결합으로 서로 연결되어 서로 붙잡고 서로의 움직임을 제한함
- 단지 몇 개의 원자부터 수천 개의 원자에 이르는 거대한 범위의 크기임

Mass Spectroscopy

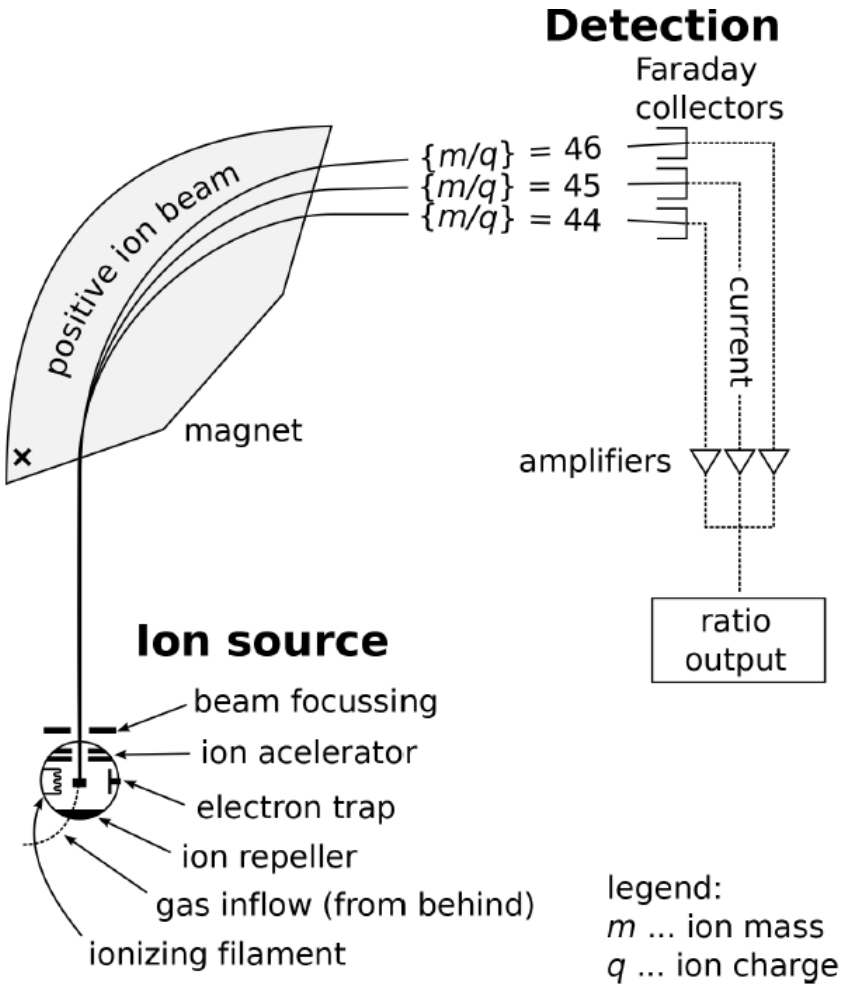
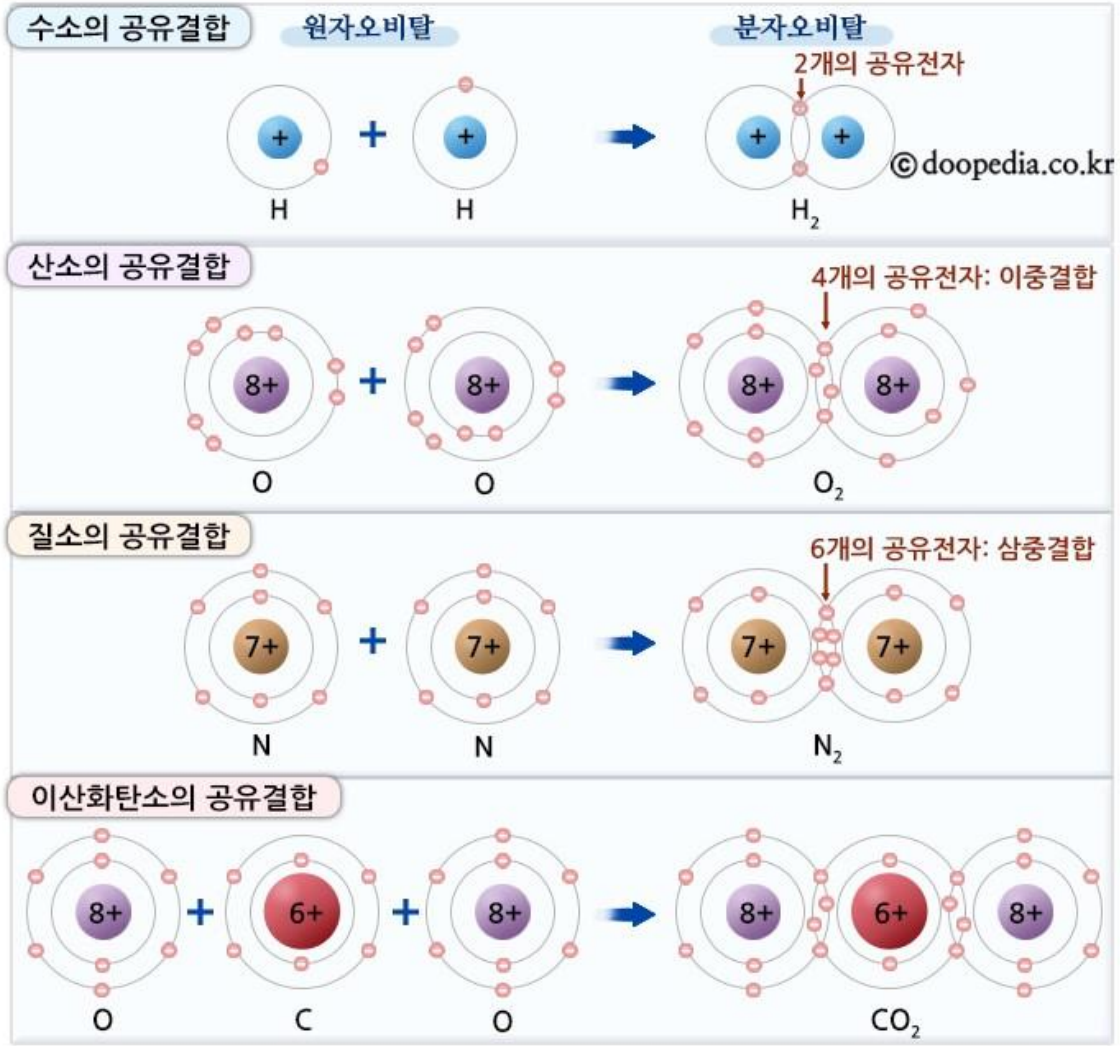


Figure 4-1. A simple schematic of a mass spectrometer. (Source: Wikimedia.)

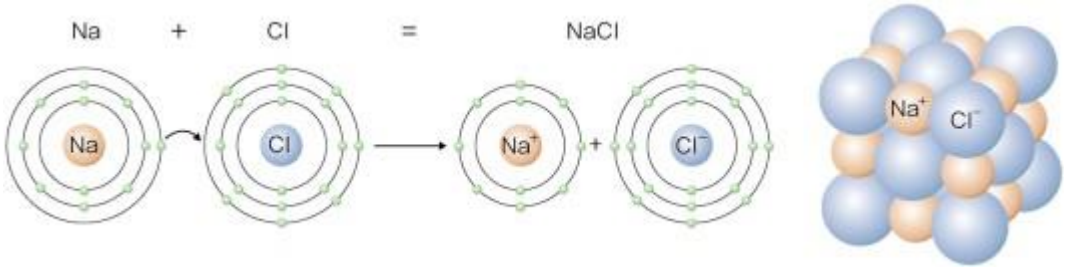
- 분자를 확인하는 가장 보편적인 방법
- 전자를 샘플에 충돌시킴 -> 분자가 이온화됨(전자를 잃거나 얻음) -> 이온화된 이온의 질량/전하비에 의해서 분리 검출함(질량 스펙트럼 이용)
- 화합물의 분자량, 원소조성 및 구조 등에 관한 정보 뿐만 아니라 구조의 추정이나 화합물의 동정 또한 가능

What Are Molecular Bonds?



- Covalent bonds(공유결합)
  - 화학결합의 하나로 2개의 원자가 서로 전자를 방출하여 전자쌍을 형성하고 이를 공유함으로써 생기는 결합
  - 대부분의 유기 화합물이나 무기화합물에서 볼 수 있음
  - -> 서로 주기 싫어하는 원자 사이의 타협
  - 도식적으로 나타낼 때 -> H:H, Cl:Cl, O::O, N:::N, 혹은, H-H, O=O

What Are Molecular Bonds?



소금이 물에 녹는 이온화 반응
나트륨은 전자 하나를 잃음으로써 염소는 전자 하나를 얻음으로써 옥테트 룰을 만족하는 이온이 된다.
각각 양이온과 음이온이 된 나트륨과 염소는 이온결합에 의해 서로 강하게 잡아당긴다.

- Noncovalent bonds(비공유결합)
  - 전자가 떨어지거나 수소가 떨어져 나가면서 양으로 혹은 음으로 이온화되어 서로 잡아 당기는 결합
  - Ex) 소금이 물에 녹는 이온화 반응
    - 나트륨은 최외각 전자껍질에 있는 전자 하나를 포기함
    - 염소는 최외각 전자껍질에 전자 하나를 채움 => 옥테르 룰 완성

Molecular Graphs

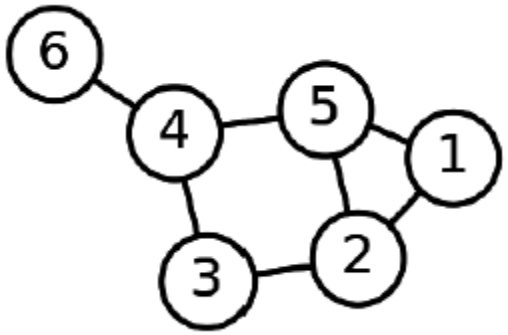


Figure 4-5. An example of a mathematical graph with six nodes connected by edges. (Source: Wikimedia.)

- 그래프는 node가 edge로 연결된 수학적 데이터 구조임
- 그래프는 컴퓨터의 네트워크 연결 구조부터 이미지를 구성하는 픽셀, 영화에서의 배우 간 관계도까지 모든 것을 설명할 수 있음

Molecular Graphs

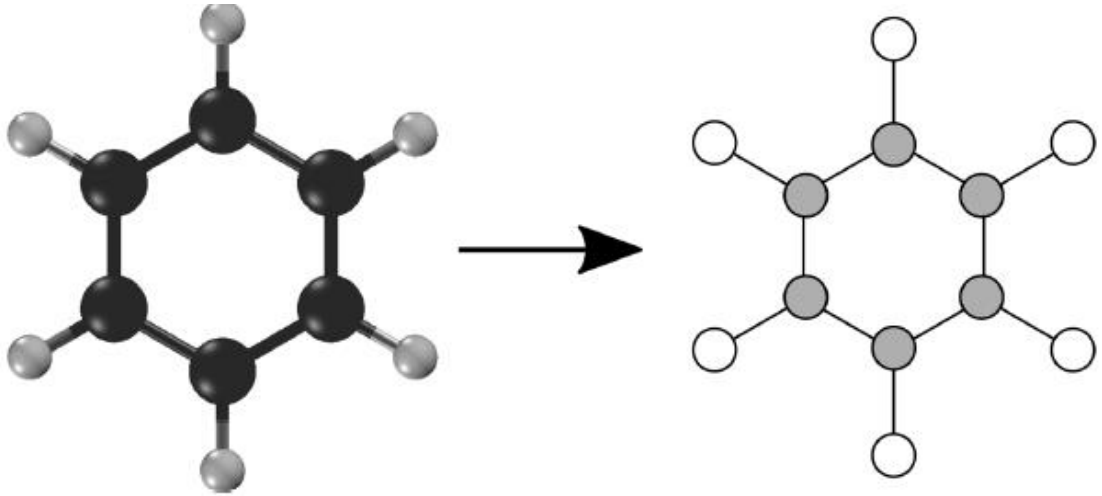


Figure 4-6. An example of converting a benzene molecule into a molecular graph. Note that atoms are converted into nodes and chemical bonds into edges.

- 분자 구조도 그래프로 볼 수 있음 -> 원자는 node로, 화학 결합은 edge로 표현 할 수 있음
- 이 단원의 나머지 부분에서는 분자 구조를 그래프로 변환하여 분자를 분석하고 예측하는 방법에 대해 다룸

Molecular Conformations

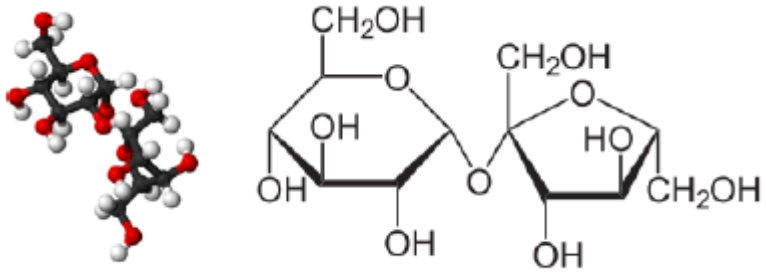


Figure 4-7. Sucrose, represented as a MD conformation and a LD chemical structure. (Adapted from Wikimedia images (Wikimedia and Wikipedia.)

- Molecular conformation (분자 구조): 원자들이 한 공간에서 상대적으로 서로에 대해 위치하는지
  - 두 개의 원자가 공유결합 하게 될 경우, 원자 사이의 거리가 고정되어 형태가 제한됨
  - 3개 혹은 4개 이상의 원자가 결합한 경우에도 각도가 고정되어 형태가 제한됨
  - 그 외의 경우에는 원자들이 서로에 대해 상대적으로 유동적임
  - 전부는 아니지만 종종 공유 결합된 원자 사이에 결합 축 회전이 가능함
- => 원자들은 다양한 형태 구성이 가능함

Molecular Conformations

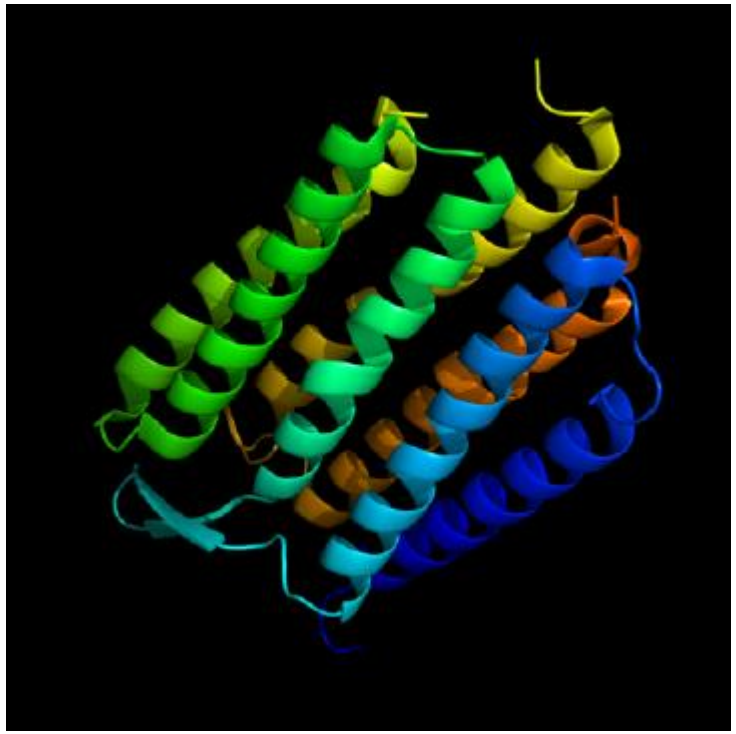


Figure 4-8. A conformation of bacteriorhodopsin (used to capture light energy) rendered in MD. Protein conformations are particularly complex, with multiple MD geometric motifs, and serve as a good reminder that molecules have geometry in addition to their chemical formulas. (Source: Wikimedia.)

- 분자 구조가 커짐에 따라, 가능한 구조의 수는 기하급수적으로 증가함
- 단백질과 같은 macromolecule의 경우, 가능한 구조를 계산하기 위해 시뮬레이션이 필요함

Chirality of Molecules

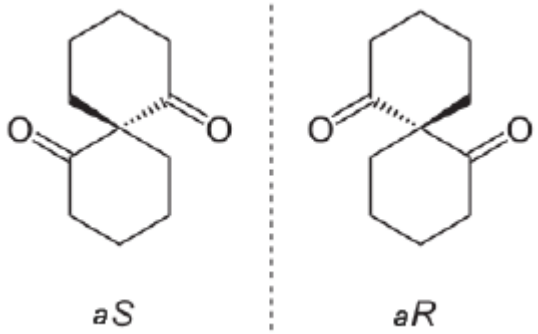
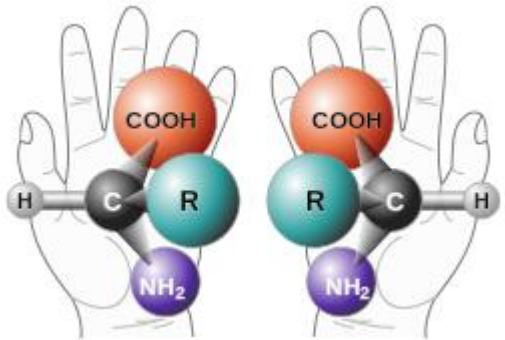


Figure 4-9. Axial chirality of a spiro compound (a compound made up of two or more rings joined together). Note that the two chiral variants are respectively denoted as “R” and “S.” This convention is widespread in the chemistry literature.



- 거울상 이성질체
- 거울에 비춘 듯 대칭으로 생겼는데, 절대 겹쳐지지 않음
- 화학적으로나 물리적으로 큰 차이가 없으나, 광학적인 성질이 다르게 나타남

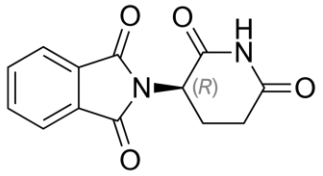
Chirality of Molecules

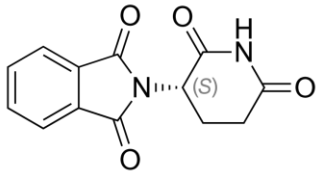
- Chirality는 실험자나 계산화학자 모두에게 어려움을 주는 매우 중요한 요소임

- 실험적으로, 화학 반응으로는 구분할 수 없으며, 물리적 특성이 동일 하므로 실험으로도 구분이 어려움
- 계산적으로도, 거울상 이성질체는 동일한 분자 그래프를 지니기 때문에 분자 그래프를 사용한 모델로는 이를 구분할 수 없음

- 심지어 같은 화합물이라도, 거울상 이성질체는 서로 다른 단백질에 결합하여 신체에 매우 다른 영향을 줄 수 있음 (ex. 부작용, …)

Chirality of Molecules





- Thalidomide
  - 1950 ~ 1960년대까지 임산부들의 입덧 방지용으로 판매된 약
  - Thalidomide의 R형은 입덧 치료에 효과적이었지만, S형은 태아 기형을 발생하는 등 심각한 선천적 결함을 유발함
  - 심지어 thalidomide는 신체 내에서 두 가지 형태가 상호 변환 가능함

Featurizing a Molecule

- 분자 데이터를 활용하여 기계학습을 수행하기 위해서는, 데이터를 벡터로 변환하는 작업이 필요함
- 이번 단원에서는 DeepChem의 서브 모듈인 dc.feat을 활용하여 데이터를 벡터로 변환하는 방법을 알아보고자 함

SMILES Strings and RDKit

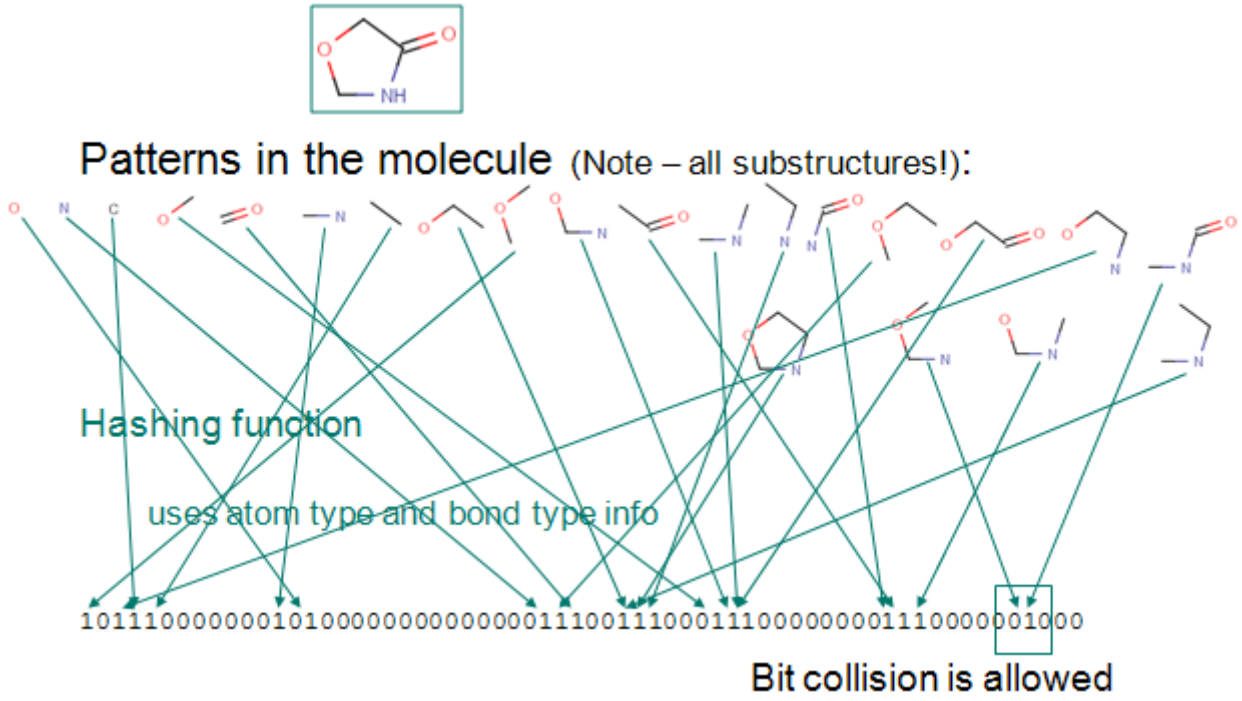
- SMILES는 분자를 텍스트로 나타내는 방법 중 하나
- DeepChem에서 input으로 사용함

- “Simplified Molecular-Input Line-Entry System”

- ex) OCCcKc(C)[n+](csK)CcLcnc(C)ncLN -> 비타민B1

- DeepChem에서는 RDKit을 활용하여 데이터를 다른 형식으로 변환하기도 함

Extended-Connectivity Fingerprints(ECFPs)



- ECFP를 활용하면 임의의 크기의 분자를 고정 길이의 벡터로 변환이 가능
- 서로 다른 크기를 가진 분자를 동일한 모델로 분석하는 것이 가능
- 두 분자간의 비교가 쉬우며 계산이 빠름

Extended-Connectivity Fingerprints(ECFPs)

```
smiles = ['C1CCCCC1', 'O1CCOCC1'] # cyclohexane and dioxane
mols = [Chem.MolFromSmiles(smile) for smile in smiles]
feat = dc.feat.CircularFingerprint(size=1024)
arr = feat.featurize(mols)
# arr is a 2-by-1024 array containing the fingerprints for
# the two molecules
```

- dc.feat.CircularFingerprint 클래스를 활용

- 단점
  - 일부 정보가 유실될 수 있음
  - 서로 다른 2개의 분자가 같은 형식으로 변환 될 수 있음
  - 같은 형식으로 변환 될 경우, 서로 다른 분자라는 걸 구분하기 어려움

Molecular Descriptors

```
feat = dc.feat.RDKitDescriptors()
arr = feat.featurize(mols)
# arr is a 2-by-111 array containing properties of the
# two molecules
```

- 화합물의 물리화학적 특성을 나타내는 표현자
- 화합물 구조를 수학적으로 처리 하여 분석에 사용
- 분자의 일반적인 특성에 의존하는 것보다 예측에 잘 작동함
- dc.feat.RDKitDescriptors() 를 활용하여 descriptors 계산 가능

Graph Convolutions

- CNN이 raw image를 입력 받아 차원을 축소시키는 것과 마찬가지로 graph convolution은 고차원의 데이터를 축소시킬 수 있음
- 원소 수, 전하 및 하이브리드 상태 같은 고차원적인 화학성질을 convolution layer를 통해 차원을 축소 시킴

- DeepChem을 이용하여 분석 가능
  - Graph convolution -> GraphConvModel
  - Weave models -> WeaveModel
  - Message passing neural network -> MPNModel
  - Deep tensor neural networks -> DTNNModel

Training a Model to Predict Solubility

```
tasks, datasets, transformers = dc.molnet.load_delaney(featurizer='GraphConv')
train_dataset, valid_dataset, test_dataset = datasets
```

- Solubility(수용성)은 얼마나 물에 잘 녹는지에 대한 척도
- 수용성이 높을 수록 체내 흡수율이 올라가기 때문에 수용성을 높이기 위해 많은 과학자들이 노력하고 있음
- Molnet에서 데이터를 로드함
- 해당 코드에서 featurizer = 'GraphConv' 옵션은 로드한 데이터를 graph convolution model에 사용할 것이므로 SMILES 형식으로 데이터를 변환해 주는 옵션

Training a Model to Predict Solubility

한 가지(solubility)만 예측하기 때문

Overfitting을 줄이기 위해
각각의 layer의 무작위한 20% output을 0으로

```
model = GraphConvModel(n_tasks=1, mode='regression', dropout=0.2)
model.fit(train_dataset, nb_epoch=100)

metric = dc.metrics.Metric(dc.metrics.pearson_r2_score)
print(model.evaluate(train_dataset, [metric], transformers))
print(model.evaluate(test_dataset, [metric], transformers))
```

Pearson correlation coefficient metrics 사용

Training 결과 ->
Training set = 0.95
Test set = 0.83

Training a Model to Predict Solubility

5개의 서로 다른 화합물 SMILES 코드를 이용하여
구축한 모델로 수용성을 예측해보자!

```
smiles = ['COC(C)(C)CCCC(C)CC=CC(C)=CC(=O)OC(C)C',
          'CCOC(=O)CC',
          'CSc1nc(NC(C)C)nc(NC(C)C)n1',
          'CC(C#C)N(C)C(=O)Nc1ccc(Cl)cc1',
          'Cc1cc2ccccc2cc1C']

from rdkit import Chem
mols = [Chem.MolFromSmiles(s) for s in smiles]
featurizer = dc.feat.ConvMolFeaturizer()
x = featurizer.featurize(mols)

predicted_solubility = model.predict_on_batch(x)
```

구축한 모델로 예측

RDKit를 사용하여 SMILES 코드를 변환시킴

MoleculeNet

- 기계학습에 사용할 수 있는 대용량 데이터가 모여 있는 저장소

- Low-level -> physical properties
- High-level -> information about interaction with a human body
  - Tox21 -> toxicity dataset
  - Delaney -> solubility dataset
  - Side effect data set

- http://moleculenet.ai/

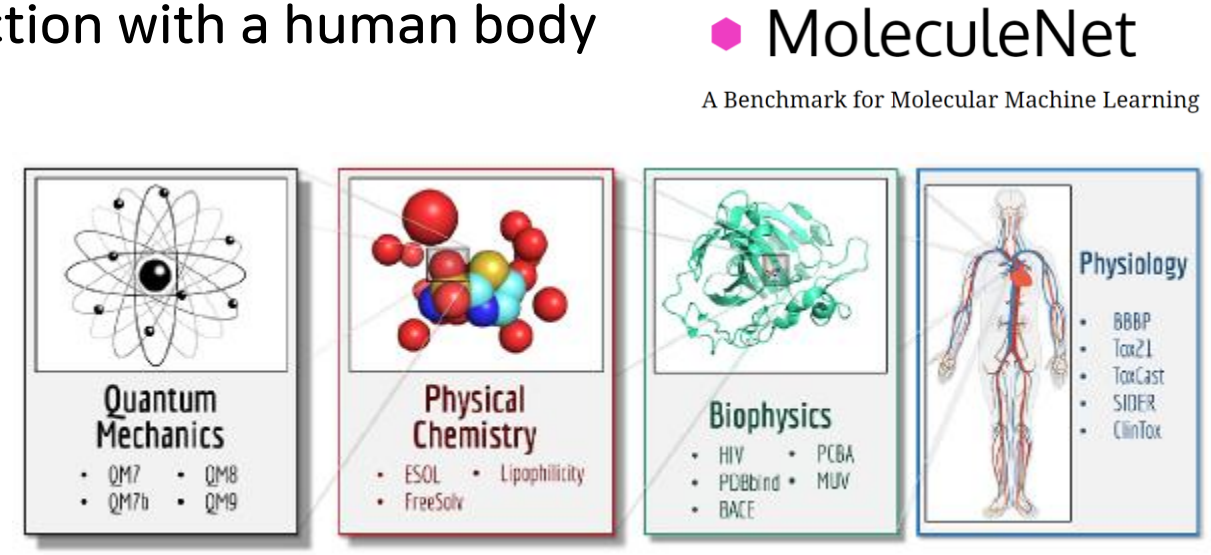


Figure 4-10. MoleculeNet hosts many different datasets from different molecular sciences. Scientists find it useful to predict quantum, physical chemistry, biophysical, and physiological characteristics of molecules.

SMARTS String

SMARTS Atomic Primitives

Symbol	Symbol name	Atomic property requirements	Default
*	wildcard	any atom	(no default)
a	aromatic	aromatic	(no default)
A	aliphatic	aliphatic	(no default)
D<n>	degree	<n> explicit connections	exactly one
H<n>	total-H-count	<n> attached hydrogens	exactly one[1]
h<n>	implicit-H-count	<n> implicit hydrogens	at least one
R<n>	ring membership	in <n> SSSR rings	any ring atom
r<n>	ring size	in smallest SSSR ring of size <n>	any ring atom[2]
v<n>	valence	total bond order <n>	exactly one[2]
X<n>	connectivity	<n> total connections	exactly one[2]
x<n>	ring connectivity	<n> total ring connections	at least one[2]
- <n>	negative charge	-<n> charge	-1 charge (-- is -2, etc)
+<n>	positive charge	+<n> formal charge	+1 charge (++ is +2, etc)
#n	atomic number	atomic number <n>	(no default)[2]
@	chirality	anticlockwise	anticlockwise, default class[2]
@@	chirality	clockwise	clockwise, default class[2]
@<c><n>	chirality	chiral class <c> chirality <n>	(nodefault)
@<c><n>?	chiral or unspec	chirality <c><n> or unspecified	(no default)
<n>	atomic mass	explicit atomic mass	unspecified mass

- 쿼리를 이용하여 SMILE 데이터를 다루는 언어
  - 분자 사이의 유사성을 비교할 때
  - 주어진 분자 데이터를 시각화 할 때
  - 시각화 된 데이터에서 특정 부분을 highlight하고 싶을 때

SMARTS String

실습을 위해 4개의 화합물 구조를 가져옴

```
from rdkit import Chem
from rdkit.Chem.Draw import MolsToGridImage

smiles_list = ["CCCCC","CCOCC","CCNCC","CCSCC"]
mol_list = [Chem.MolFromSmiles(x) for x in smiles_list]
```

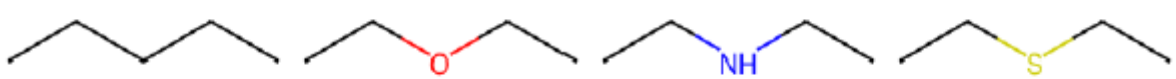


Figure 4-11. Chemical structures generated from SMILES

SMARTS String

'CCC'로 구성된 부분을 highlight

```
query = Chem.MolFromSmarts("CCC")
match_list = [mol.GetSubstructMatch(query) for mol in
mol_list]
MolsToGridImage(mols=mol_list, molsPerRow=4,
highlightAtomLists=match_list)
```

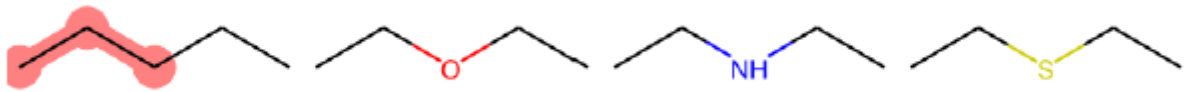


Figure 4-12. Molecules matching the SMARTS expression "CCC."

SMARTS String

‘C*C’로 구성된 부분을 highlight

```
query = Chem.MolFromSmarts("C*C")
match_list = [mol.GetSubstructMatch(query) for mol in
mol_list]
MolsToGridImage(mols=mol_list, molsPerRow=4,
highlightAtomLists=match_list)
```

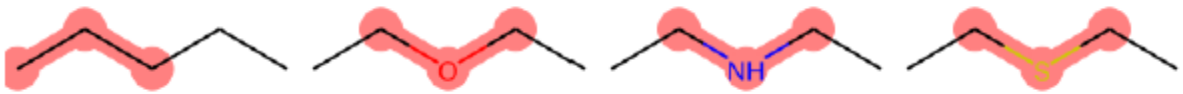


Figure 4-13. Molecules matching the SMARTS expression “C*C”.

SMARTS String

‘C[C,N,O]C’로 구성된 부분을 highlight

```
query = Chem.MolFromSmarts("C[C,N,O]C")
match_list = [mol.GetSubstructMatch(query) for mol in
mol_list]
MolsToGridImage(mols=mol_list, molsPerRow=4,
highlightAtomLists=match_list)
```

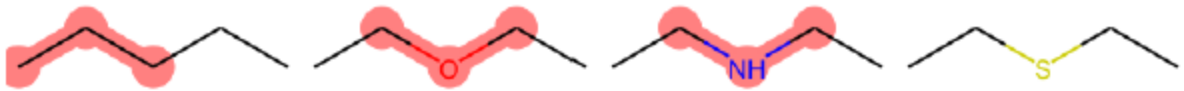


Figure 4-14. Molecules matching the SMARTS expression “C[C,N,O]C”.

- The End -

더 자세한 분석은 11단원에서!