Molecule Dataset

Presented by: Hyungeun Lee

Last Updated: October 12th, 2022

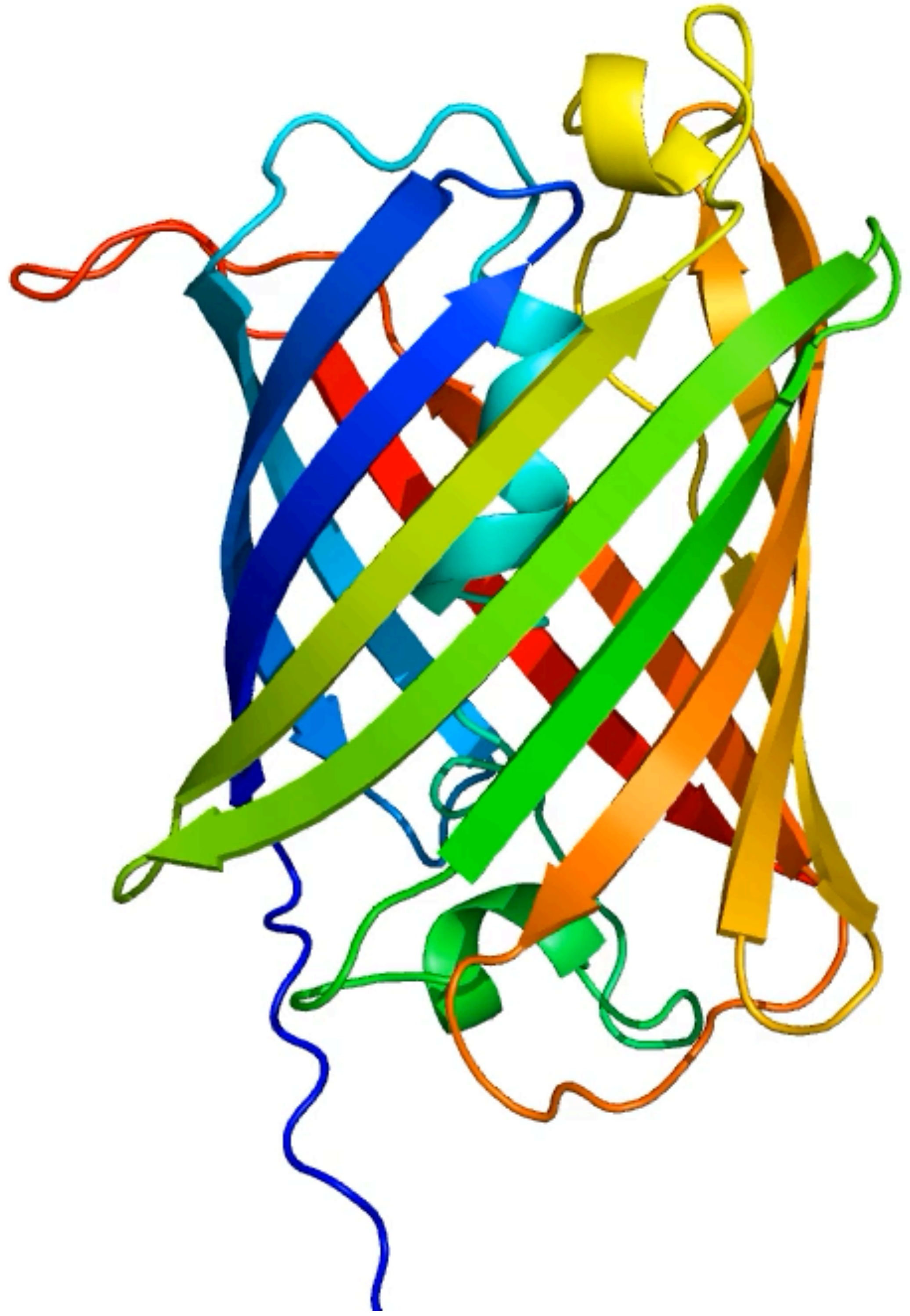
Table of Contents

- Why ML for molecules?
- Molecule dataset
- Term Project Description
- Conclusion

Why ML for molecules?

- Lately, research on predicting molecular properties using ML has been actively conducted.

- It reduces the computational cost required for predicting molecular properties



Green fluorescent protein 1
modelled by Alphafold

SARS-Cov-2 ORF8

Why ML for molecules?

$$i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2} + V\Psi$$

↓

Training & Inference

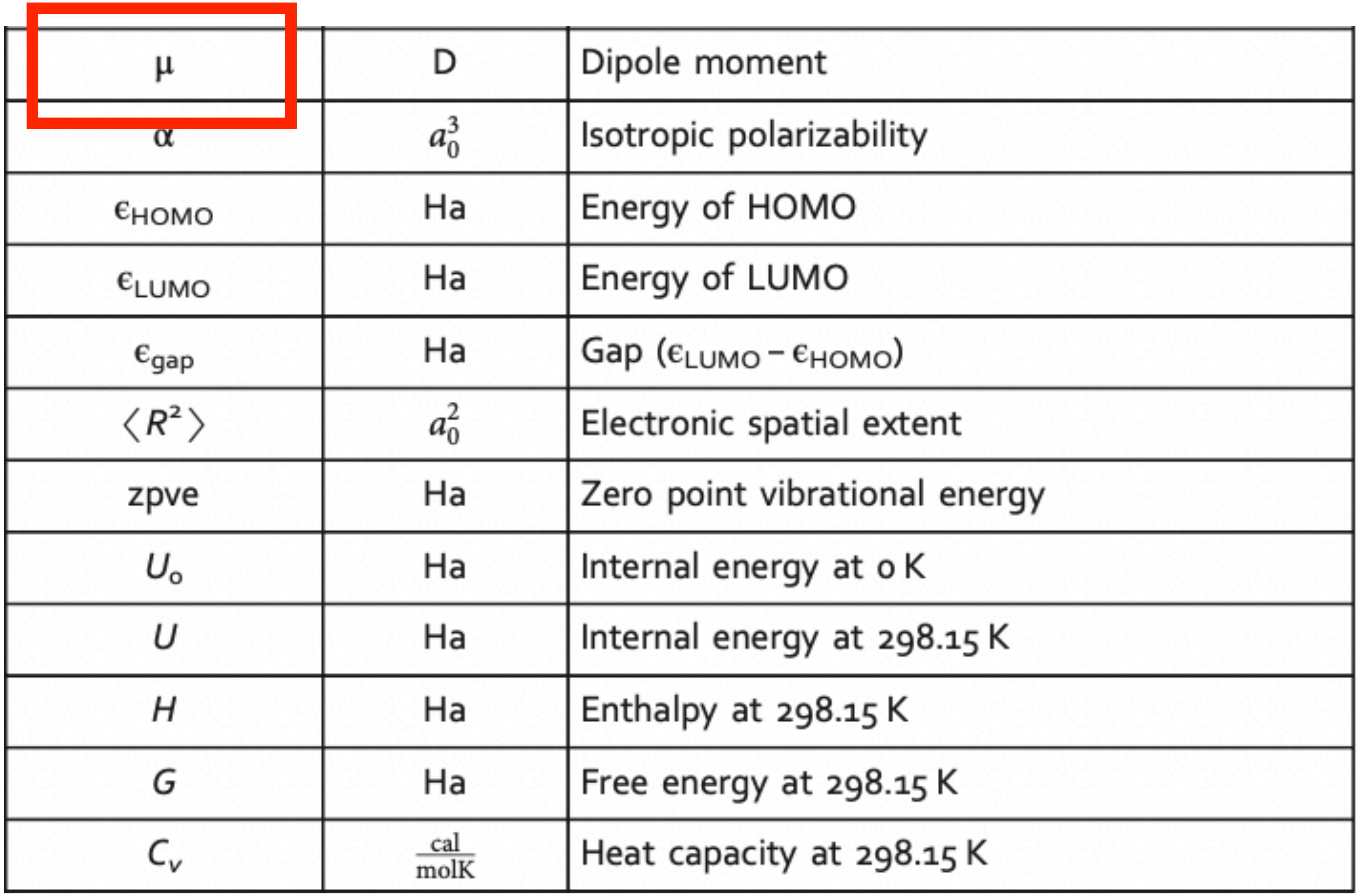
Molecule dataset

- 12k stable small organic molecules made up of **CHONF**.
- This dataset contains some small amino acids

  e.g) glycine, alanine, as well as nucleobases cytosine, uracil, and thymine

- This dataset provides dipole moment μ, which is one of quantum chemical properties calculated at the DFT/B3LYP/6-31G(2df, p) level of theory.

Molecule dataset



μ	D	Dipole moment
α	a_0^3	Isotropic polarizability
ϵ_{HOMO}	Ha	Energy of HOMO
ϵ_{LUMO}	Ha	Energy of LUMO
ϵ_{gap}	Ha	Gap ($\epsilon_{\text{LUMO}} - \epsilon_{\text{HOMO}}$)
$\langle R^2 \rangle$	a_0^2	Electronic spatial extent
zpve	Ha	Zero point vibrational energy
U_0	Ha	Internal energy at 0 K
U	Ha	Internal energy at 298.15 K
H	Ha	Enthalpy at 298.15 K
G	Ha	Free energy at 298.15 K
C_v	$\frac{\text{cal}}{\text{molK}}$	Heat capacity at 298.15 K

Table 1

Molecule dataset

- **Training Data Attributes:**

Atom type (CHNOF)

Scalar properties (target values)

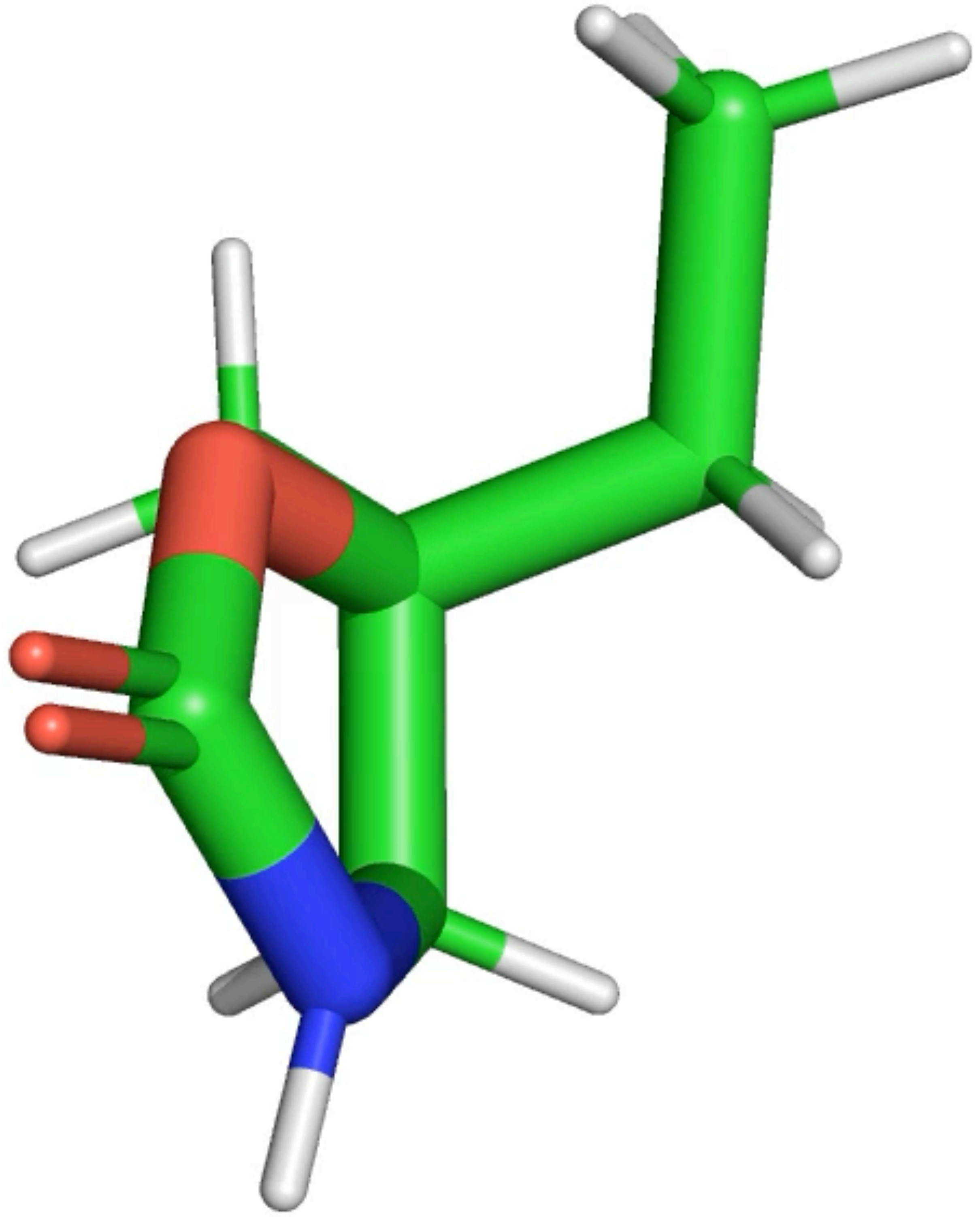
XYZ coordinates of atoms

Bond(Edge) index

Bond(Edge) type

SMILES string

+ α (Hand-crafted features)



Molecule dataset

- **Test Data Attributes:**

  Atom type (CHNOF)

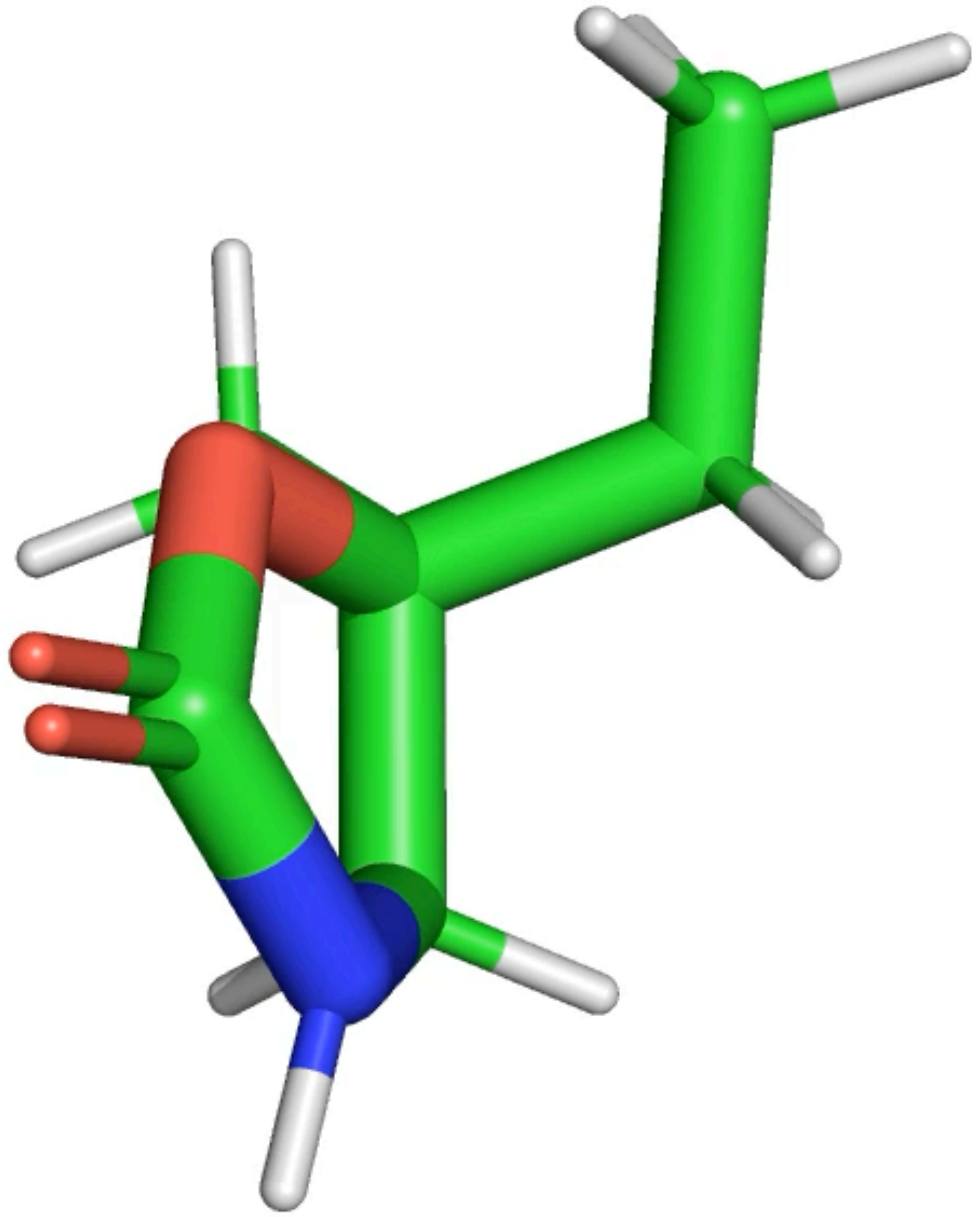
  XYZ coordinates of atoms

  Bond(Edge) index

  Bond(Edge) type

  SMILES string

  + α (Hand-crafted features)



Molecule dataset

- **Atom type:**

$\mathbb{N}$ = {C, H, N, O, F} → mapping into Atomic number {6, 1, 7, 8, 9}

- **Scalar properties (Target values):**

(see Table 1)

- **XYZ coordinates of atoms:**

$X \in \mathbb{R}^{|\mathscr{V}| \times 3}$ matrix, where $|\mathscr{V}|$ is the number of atoms

Molecule dataset

- **Bond(Edge) index**

  The bond index is defined by [2, 2x|$\mathscr{E}$|], which represents all edges from $\mathscr{V}_i$ to $\mathscr{V}_j$.

- **Bond(Edge) type**

  The bond type is defined as a string.

  e.g) "SINGLE", "DOUBLE", "TRIPLE", "AROMATIC"

- **SMILES string**

  e.g) "CCc1c(non1)CO"

Term Project Description

- 과제 제출 파일 형식
  - Kaggle: .csv with correct form
  - 보고서: .pdf
- 과제 제출 기한:
  - Kaggle: ~2022.12.14 (수) 23:59
  - 보고서: ~2022.12.18(일) 23:59
- 제출 플랫폼 (링크):
  - Kaggle: https://www.kaggle.com/t/4791bb471a804f4098ccf38036b4de44
  - 보고서: LMS
- Kaggle 리더보드 평가 Metric

Mean absolute error(MAE)

$$MAE = \frac{1}{N}\sum_{j=1}^{N} |y_j - \hat{y}_j|$$

Conclusion

- We have introduced molecule dataset.

- This dataset enable us to train with molecule data as graph and inference quantum molecular properties within merely milliseconds.

  e.g) dipole moment (μ)

- To make your model perform well, you can use hand-crafted features, such as SBF, RBF. (optional)

Q&A