cmiles Documentation

cmiles

Mar 27, 2019

Contents:

CMILES generates molecular identifiers for the QCArchive and OpenForceField projects

CMILES provides a way to access data in QCArchive for the purposes of the OpenFF project. Its main function is to provide canonical identifiers for the molecules in the database. These identifiers include:

1. Unique cheminformatics representations
2. A way to map the nodes in the chemical graph to the atom order or coordinates to the QC results

CHAPTER 1

Canonical Identifiers

In the QCArchive database, molecules are uniquely identified by their elements and coordinates. These molecules need to be accessible via an identifier that does not rely on coordinates so that calculations with different geometries of the same molecule, such as torsion scans, can be grouped together.

Several flavors of cheminformatics identifiers exists such as SMILES, InChI, InChIKey etc. `cmiles` provides several flavors of SMILES, InChI, InChIKey, and Hill molecular formula.

`cmiles` addresses several issues related to canonical identifiers:

1. **Canonical SMILES**

   Identifiers must be canonical to reduce redundancy and avoid search failures in the future. However, many cannibalization algorithms exist so SMILES are only canonical within a certain cheminformatics toolkit and many times, only within a specific version of that toolkit.

   `cmiles` is distributed as a Docker image with pinned dependencies to ensure that the SMILES are truly canonical

2. **Standardizing compunds**

   SMILES are unique to each protomeric state. However, sometimes we might want to search all protomers and / or tautomers of a molecule. While InChI provides standardization for charge, it does not standardize all tautomers. Some known tautomers it does not capture are keto-enol and enamine-imine.

   `cmiles` does not yet offer a full solution. However, it provides InChI, the unique protomer from openeye and `MolStandardize` from rdkit. RDkit's solution only addresses tautomers of the same charge states and openeye's solution does not capture indoles, isoindoles and some mesomers. However, the union of these identifiers capture more than each individual solution.

CHAPTER 2

Keeping track of coordinates order

The order of nodes in molecular graphs are arbitrary. While individual cheminformatics toolkits may order the atoms canonically, that order might be different than the order of the elements in the database QC molecule. To ensure that there is no loss of information, `cmiles` generates isomeric, explicit hydrogen mapped SMILES. The map indices correspond to the order of the molecule's coordinates in the database. This SMILES can be used as a SMARTS pattern to retrieve the mapping of the atoms in the cheminformatics molecule to the xyz coordinates and symbols in the QC molecule. `cmiles.utils` provides functions that will do this.

CHAPTER 3

Using CMILES

The main function of cmiles is to generate the identifiers and it is just one function call:

```
from cmiles import generator
identifiers = generator.get_molecule_ids(molecule)
```

identifers is a dictionary with several flavors of SMILES and some other identifiers. It includes:

1. Canonical SMILES
2. Isomeric SMILES
3. Canonical, explicit hydrogen SMILES
4. Isomeric, explicit hydrogen SMILES
5. Isomeric, explicit hydrogen, mapped SMILES
6. Molecular formula in Hill notation
7. InChI
8. InChIKey
9. Unique protomer SMILES
10. Standardized tautomer SMILES

CHAPTER 4

Indices and tables

- genindex
- modindex
- search