Molecular Graph Generation

Corporate partner mentors from Merck: Ti-chiun Chang, Xiang Yu – Data Mine student: Maria Berardi

PURDUE UNIVERSITY | The Data Mine

Project Overview

Motivation for AI based Molecular Generation

- Creating new chemical compounds satisfying desired properties is a fundamental goal in drug discovery, hence Merck's interest in this project.
- Creating such novel molecules is expensive in terms of time and cost, so researchers have found ways to automate part of the search process using various machine learning algorithms.

Proposed Strategy

- Combine features of reinforcement learning and molecular graph encodings.

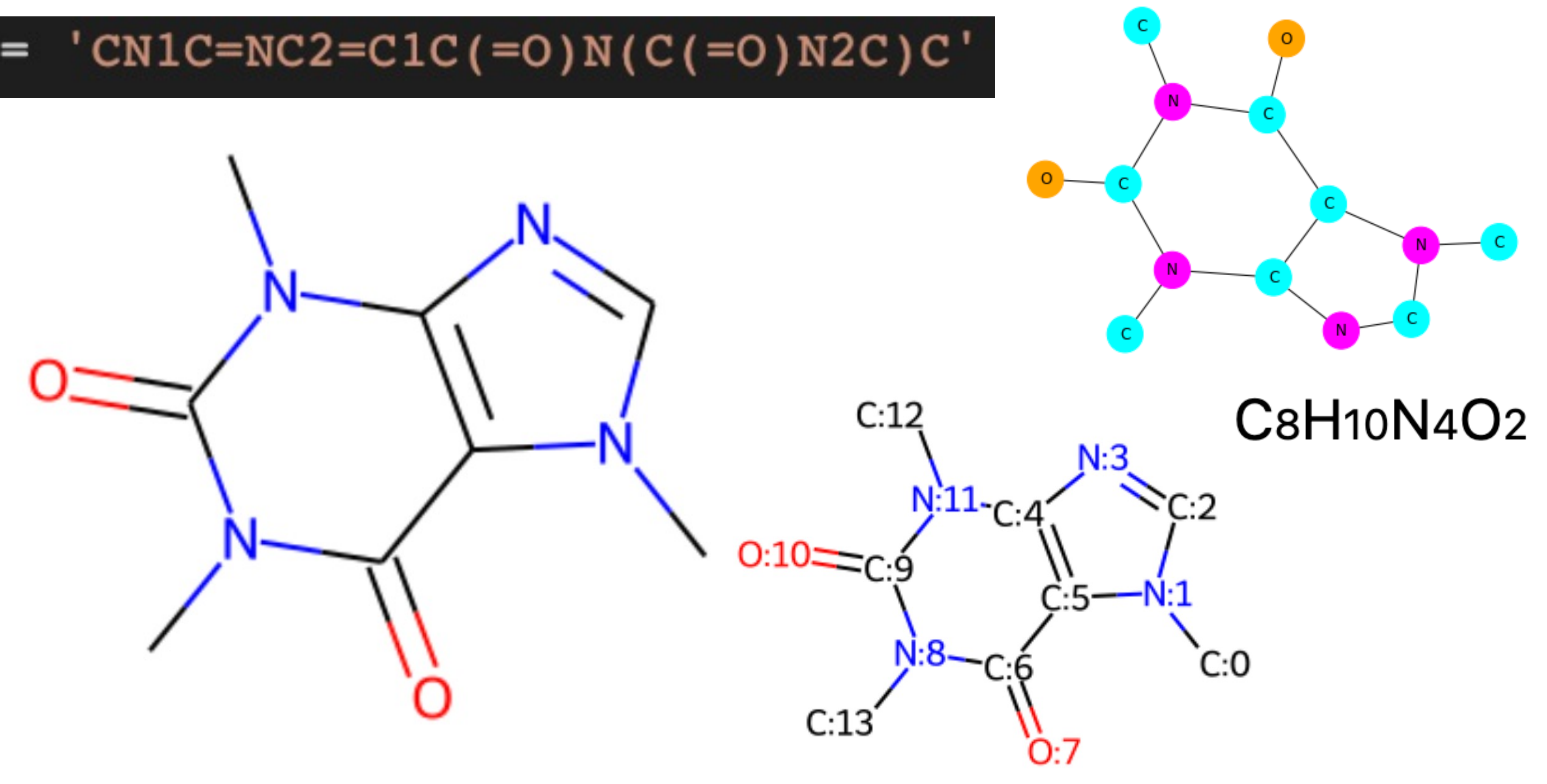
Molecular Representations in Python

Possible Approaches

- RDKit is a Python library designed to deal with chemical compounds.
- Using RDKit, we can represent a molecule in different ways, depending on the context.
- Common ways to represent a molecule are:
  - A SMILES string,
  - A graph,
  - A graph connectivity matrix.

```
caffeine_smiles = 'CN1C=NC2=C1C(=O)N(C(=O)N2C)C'
```

```
[[0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
 [1. 0. 1. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0.]
 [0. 1. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
 [0. 0. 1. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
 [0. 0. 0. 1. 0. 1. 0. 0. 0. 0. 0. 1. 0. 0.]
 [0. 1. 0. 0. 1. 0. 1. 0. 0. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 1. 0. 1. 1. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0. 1. 0. 0. 1. 0. 0. 0. 1.]
 [0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 1. 1. 0. 0.]
 [0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0.]
 [0. 0. 0. 0. 1. 0. 0. 0. 0. 1. 0. 0. 1. 0.]
 [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0.]
 [0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0.]]
```



$C_8H_{10}N_4O_2$

Figure 1: Different representations of a caffeine molecule.

Traditional use of Reinforcement Learning

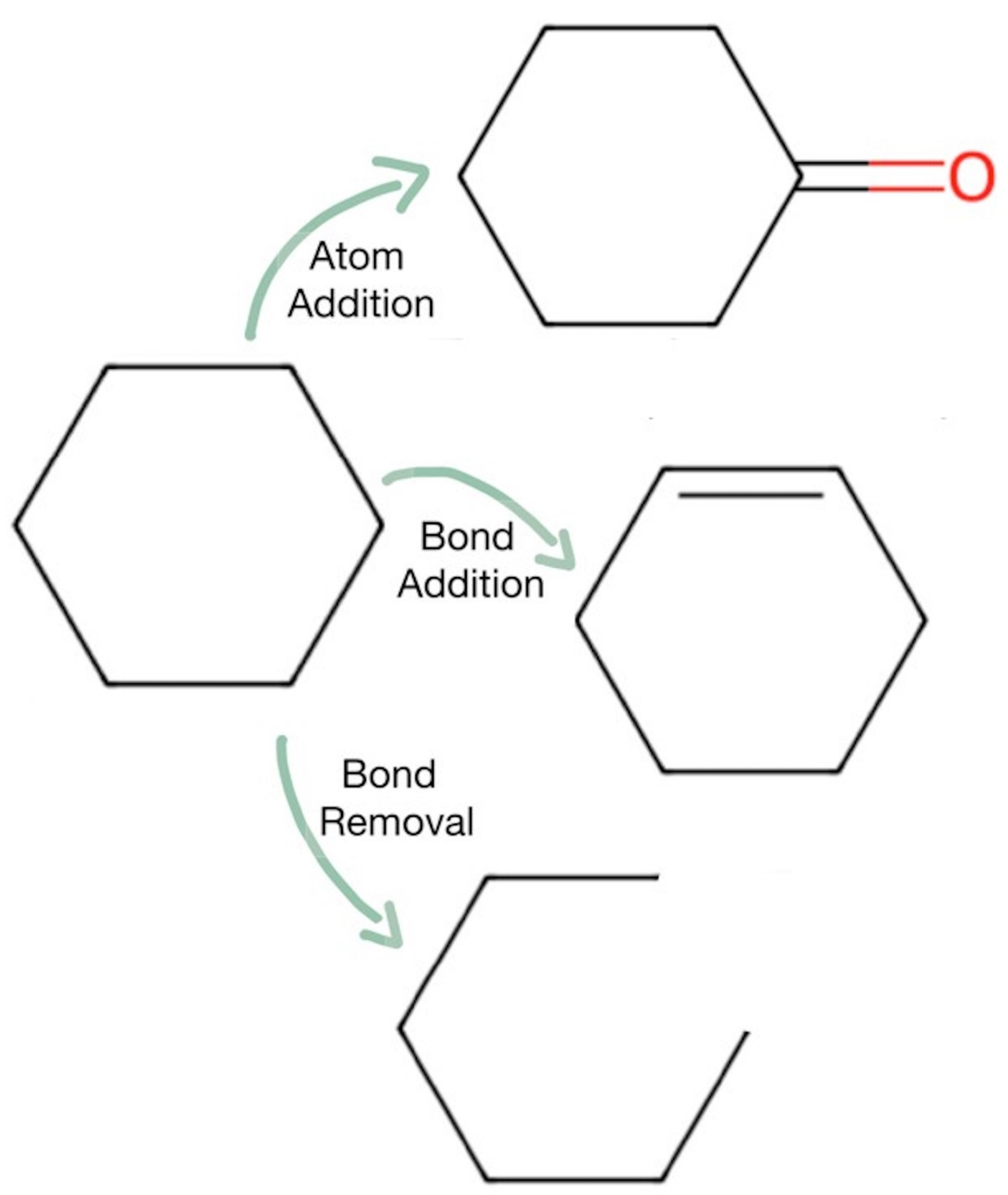


Figure 2: A molecular generation model can be trained to explore the chemical space via a sequence of **actions**, iteratively adding and removing bonds and atoms.

At each step, a **reward** is given, weighted based on when in the process the action is taken: **exploration** is encouraged in the earlier stages of training, **exploitation** is encouraged after some knowledge has been acquired.

Learning happens through the **optimization** of the **cumulative reward**.

It is possible to design the reward to obtain **optimization** of **predetermined molecular properties**.

Such a model can be implemented relying on a **SMILES string** representation.

Replacing Actions on Atoms with Actions on Larger Molecular Substructures

Advantages of Graph Representation

Adopting a molecular graph representation offers the possibility to:

- Include 3D features, such as distance between atoms;
- Attach a feature vector to each node and edge, characterizing node type (ex: C, O, N...) and edge type (ex: single, doble, aromatic...);
- Rely on GNNs (graph neural networks) for model training;
- Measure molecular similarity more accurately than what can be computed from strings;
- Make use of graph substructures to encode reinforcement learning actions, resulting in more accurate molecular reaction encoding and in the ability to deal with larger chemical compounds than those that can be handled with string encodings.

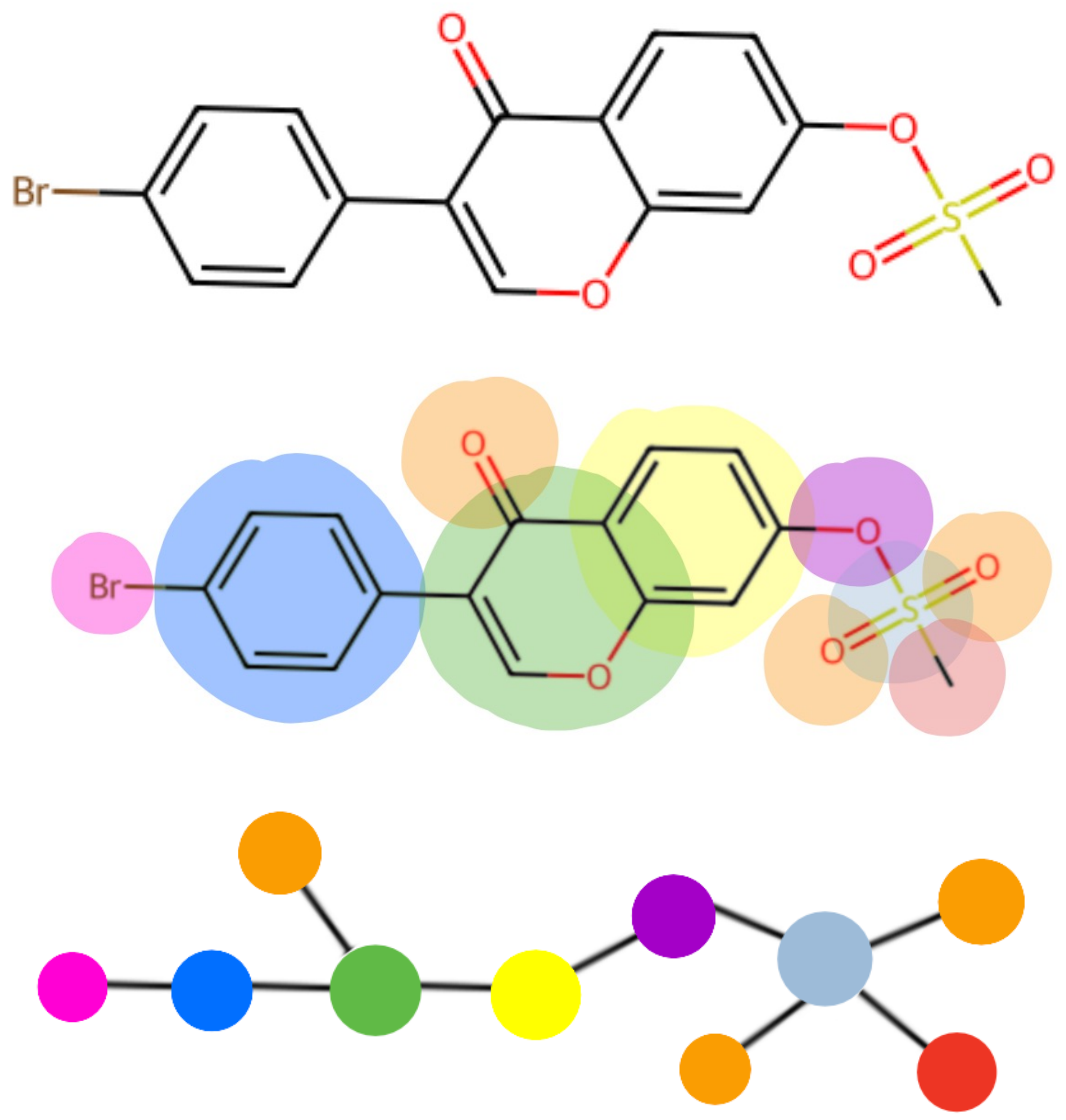


Figure 3: A molecular graph and its junction tree representation

Subgraphs: Trees and Motifs

- When several data points in a training set of molecules exhibit frequently occurring substructures, we may have a reinforcement learning algorithm rely on larger molecular substructures consisting of groups of atoms, rather than single atoms or bonds.
- Such molecular substructures are called junction trees or motifs.
- Main advantage: better reconstructions of larger compounds (ex: polymers).

Possible Future Developments

- Combine current work with multi-objective optimization to achieve simultaneous control of competing molecular properties.

The Data Mine Corporate Partners Symposium 2023

References

1. Jin, Wengong, Regina Barzilay, and Tommi Jaakkola. "Hierarchical generation of molecular graphs using structural motifs." In *International conference on machine learning*, pp. 4839-4848. PMLR, 2020.
2. Jin, Wengong, Regina Barzilay, and Tommi Jaakkola. "Junction tree variational autoencoder for molecular graph generation." In *International conference on machine learning*, pp. 2323-2332. PMLR, 2018.

Acknowledgements

I would like to thank my mentors, Ti-chiun Chang and Xiang Yu, for all their guidance and advice, and the Data Mine staff for their support.