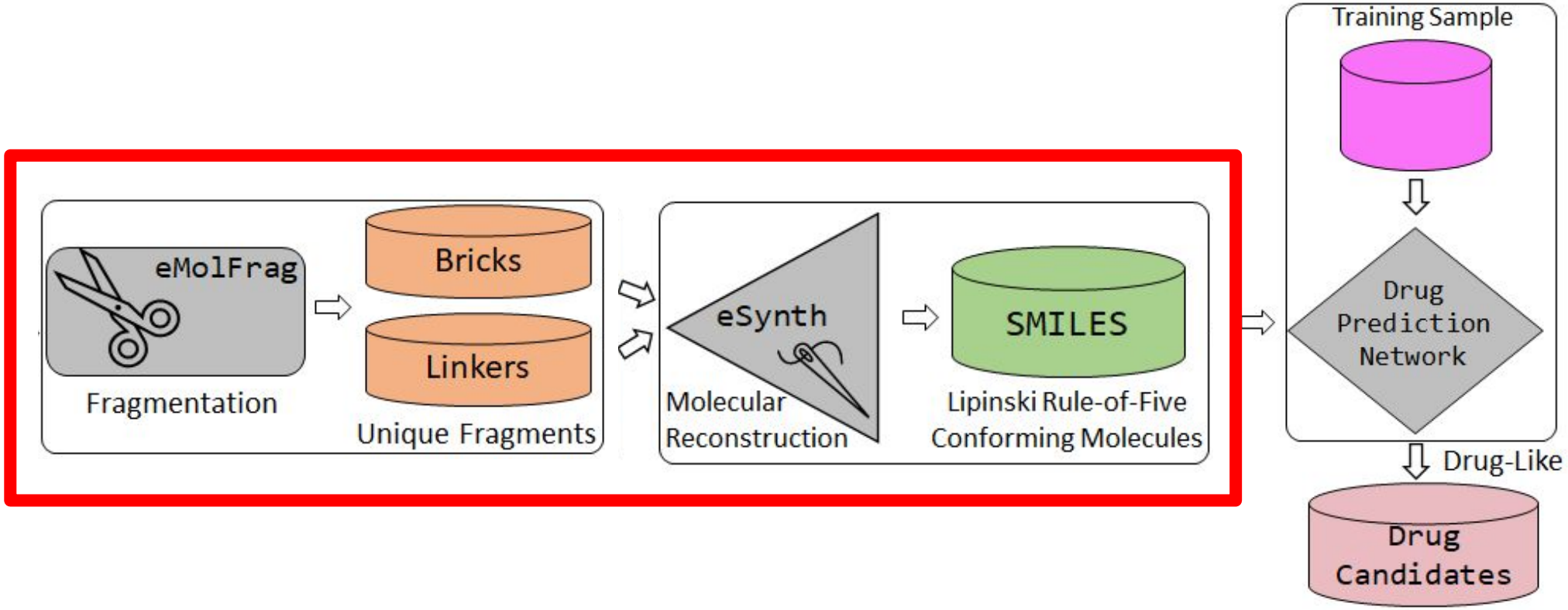
In Silico Synthesis of Molecules

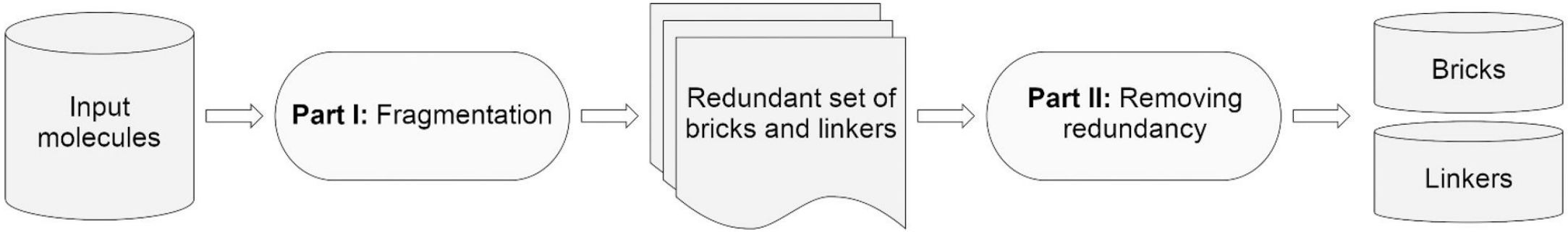
Jonathan Dewey, Ting Chen

Advisor: Chris Alvin

Drug Discovery Process

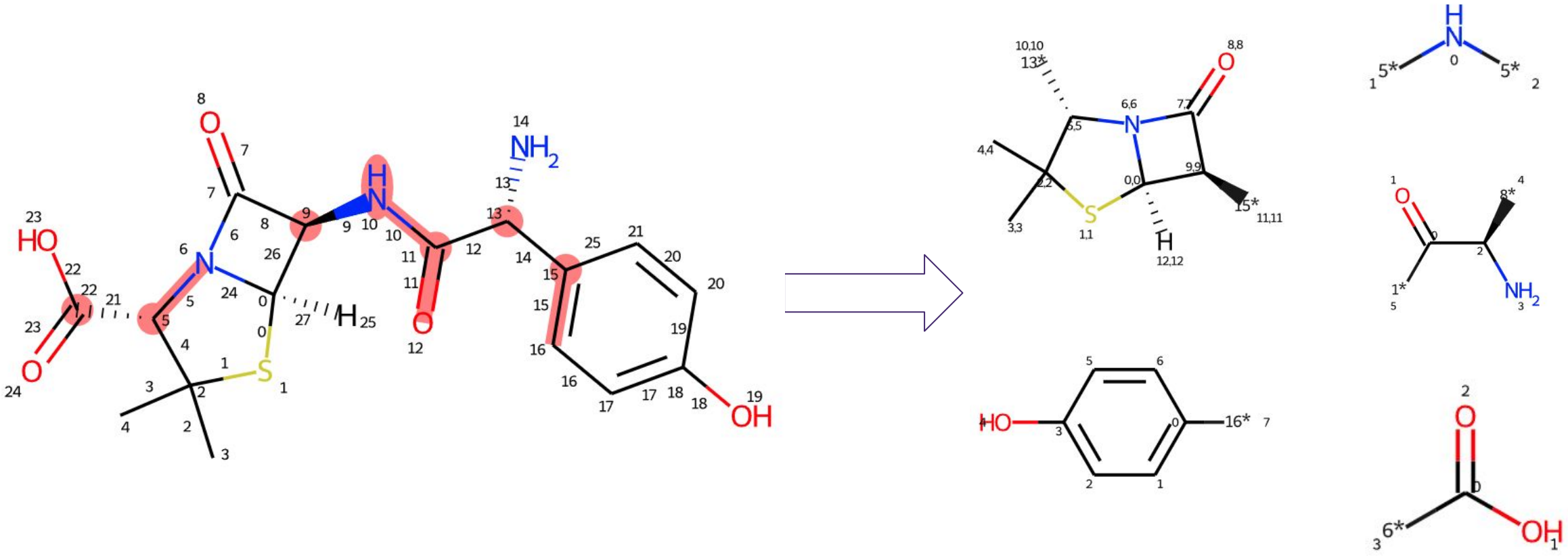


eMolFrag – Molecular Fragmentation



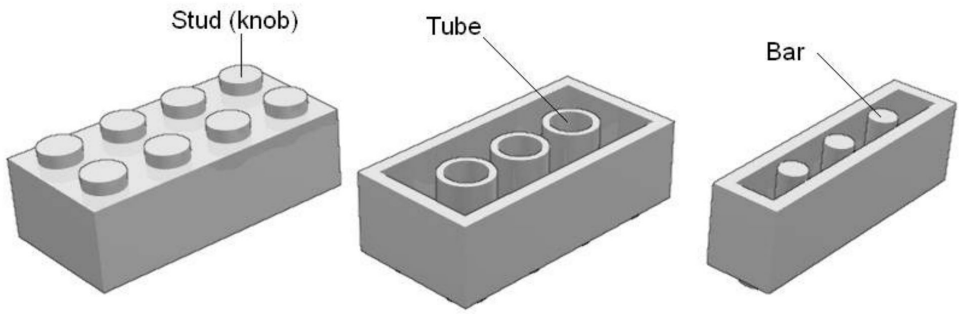
- Fragmentation – making a copy of original molecule and then removes surrounding atoms to retain atom connectivity information.
  - Connectivity is critical for eSynth computation
- Remove Redundancy – compare molecular fingerprint similar by calculating Tanimoto Coefficient(TC) and remove fragments with desire TC threshold
- Results include fragment information in .sdf format and similar fragments

Molecular Fragmentation of Amoxicillin



eMolFrag v1.0 Problems

- Only .mol2 input
- **Major bug found, missing atom connectivity information**
- String manipulation
- Use BRICS algorithm to chop all types of bonds and reconnect double bonds
- No unit tests to verify the result




```
@<TRIPOS>MOLECULE
*****
 11 10 0 0 0
SMALL
GASTEIGER

@<TRIPOS>ATOM
      1 C          -2.1425   -0.2107    0.0000 C.2     1  UNL1        0.1130
      2 N          -1.4286    0.2026    0.0000 N.am    1  UNL1       -0.2796
      3 C          -0.7146   -0.2065    0.0000 C.3     1  UNL1        0.0985
      4 C          -0.0006    0.2068    0.0000 C.3     1  UNL1        0.0297
      5 C           0.7133   -0.2024    0.0000 C.3     1  UNL1        0.0607
      6 O          -2.1415   -1.0383    0.0000 O.2     1  UNL1       -0.5358
      7 O          -1.4304    1.0293    0.0000 O.3     1  UNL1       -0.4195
      8 P           1.4280    0.2132    0.0000 P.3     1  UNL1        0.1172
      9 O           1.4280    1.0383    0.0000 O.2     1  UNL1       -0.4601
     10 O           2.1425   -0.1992    0.0000 O.3     1  UNL1       -0.3571
     11 O           1.6415   -0.5836    0.0000 O.3     1  UNL1       -0.3571
@<TRIPOS>BOND
     1     1     6    2
     2     1     2    am
     3     2     7    1
     4     3     4    1
     5     5     8    1
     6     4     5    1
     7     2     3    1
     8     8     9    2
     9     8    10    1
    10     8    11    1
```

```python
def GenerateMolblock(atomInfo, bondInfo):
    tempMolblockList = []
    tempMolblockList.append('\n')
    tempMolblockList.append('     RDKit          3D\n')
    tempMolblockList.append('\n')

    newAtomNum = len(atomInfo[0])
    newBondNum = len(bondInfo)
    newHead=str(newAtomNum).rjust(3)+str(newBondNum).rjust(3)+'  0  0  0  0  0  0  0  0999 V2000\n'

    tempMolblockList.append(newHead)
    atomIndMapList = [] # [new] <-> [old]
    atomIndMapList.append(list(range(1,newAtomNum+1)))

    dummyIndList = []
    dummyAtomLineList = []
    normalIndList = []
    normalAtomLineList = []

    for i in range(newAtomNum):
        if len(atomInfo[4][i]) > 50:
            if (atomInfo[5][i] == 'R'):
                tempAtomLine = atomInfo[4][i][:31] + 'R ' + atomInfo[4][i][33:]
                dummyIndList.append(atomInfo[0][i])
                dummyAtomLineList.append(tempAtomLine)
            else:
                normalIndList.append(atomInfo[0][i])
                normalAtomLineList.append(atomInfo[4][i])

    atomIndMapList.append(normalIndList+dummyIndList)
    newAtomList = normalAtomLineList + dummyAtomLineList

    newBondList = []
    for bond in bondInfo:
        tempInd1 = int(bond[0:3])
        tempInd2 = int(bond[3:6])
        tempInfo = bond[6:]
        newInd1 = atomIndMapList[0][atomIndMapList[1].index(tempInd1)]
        newInd2 = atomIndMapList[0][atomIndMapList[1].index(tempInd2)]
        newBond = str(newInd1).rjust(3) + str(newInd2).rjust(3) + tempInfo
        newBondList.append(newBond)

    tempMolblockList.append('\n'.join(newAtomList) + '\n')
    tempMolblockList.append('\n'.join(newBondList) + '\n')
    tempMolblockList.append('M  END\n')

    molblockReturn = ''.join(tempMolblockList)
    return molblockReturn
```

eMolFrag v1.0 code example

- Text-based manipulation
- Line Count: 3920
- Source Line Of Code (SLOC): 2817
- ~1000 lines of whitespace and comments
- 1 line of comment for 5 lines of source code
- Poor readability
- Difficult to detect errors in code
- Runtime: **23s** for **100** input molecules

FURMAN UNIVERSITY

eMolFrag v2.0 code example

```
def molBRICSBonds(mol):
    """
    De-Duplicated BRICS Bonds List

        Parameters:
                mol (Rdkit.Mol): Molecule to get BRICS Bonds for

        Returns:
                a (list of tuples): De-Duplicated BRICS Bonds List
    """
    snips = [(a, b) for (a, b), (c, d) in list(BRICS_custom.FindBRICSBonds(mol))]

    # reorder tuples as set
    return {(a, b) if (a < b) else (b, a) for a, b in snips}
```

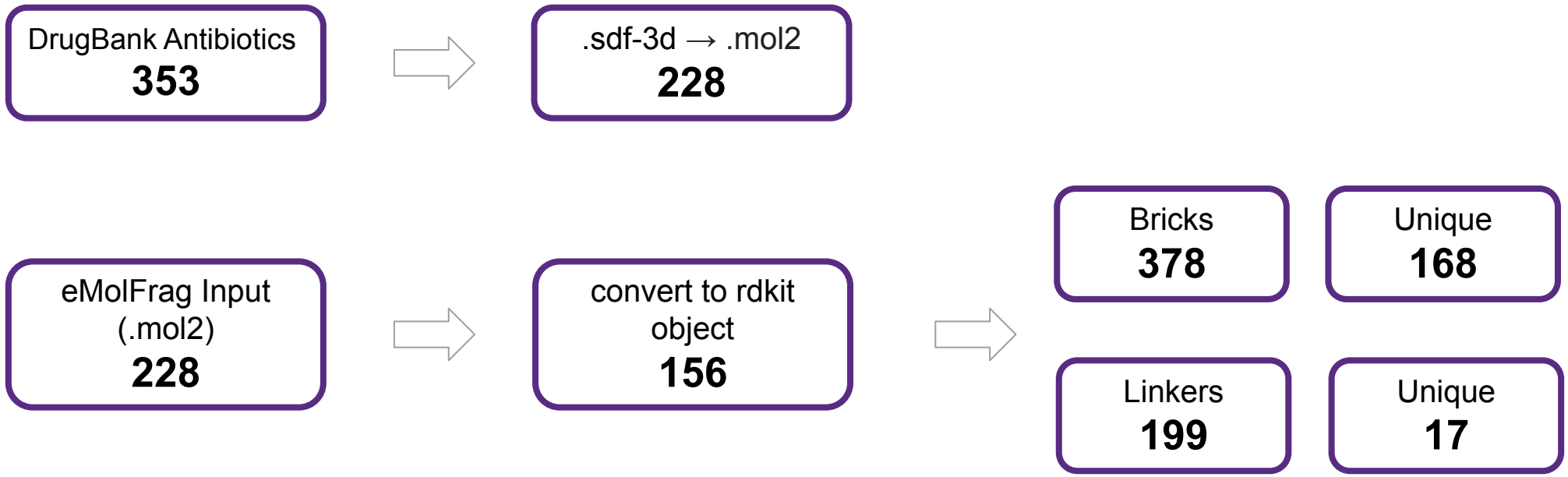
- Line Count: 1818
- Source Line Of Code (SLOC): 546
- ~1300 lines of whitespaces and comments
- 2 lines of comment for 1 line of code
- Runtime: **3s** for **200** input molecules

eMolFrag v2.0 Improvement

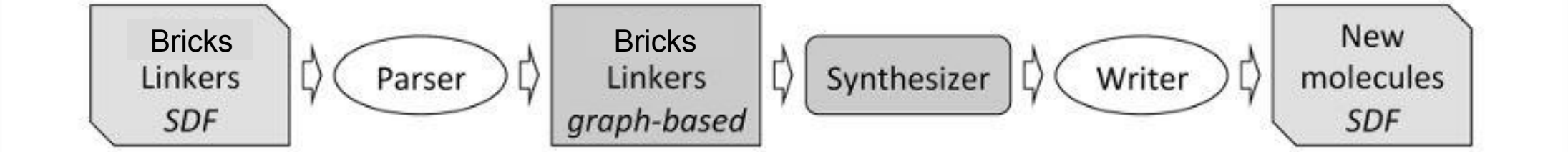
A complete re-implementation that includes:

- **Correctly computing connectivity information from original molecule**
- Graph-based analysis and fragmentation
- All input molecule formats rdkit accepts*
- Improve input system for better efficiency
- Use BRICS custom version (without removing double bonds)
- Implement unit tests for main functionalities

eMolFrag v2.0 Sample Results



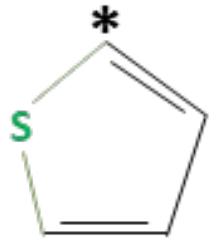
eSynth - Algorithm



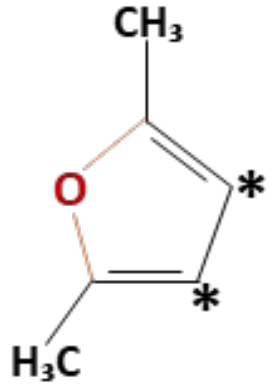
- **Graph of Bricks and Linkers**: The input set of fragments are made into a graph where fragments are consolidated to remove redundancies among the dataset:
  - Bricks consolidate connectivity information so similar bricks can consolidate into one entity that contains all variations in connectivity information
  - Linkers combine linkers that are attached to each other to form longer linkers
- **Synthesizer**: Composes all possible molecules from a set of fragments. The amount of molecules synthesized grows exponentially. eSynth implements a *Bloom Filter* to remove redundant molecules.

Molecular Reconstruction

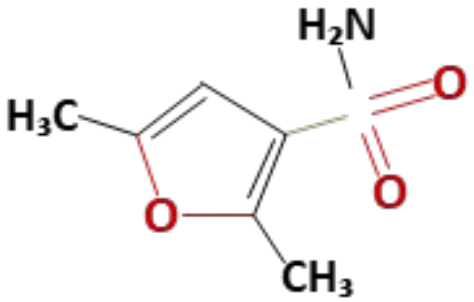
Bricks



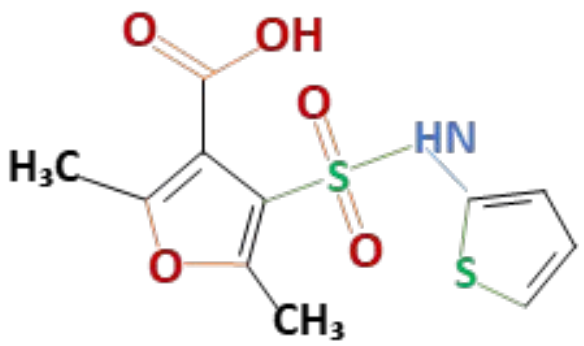
Thiophene

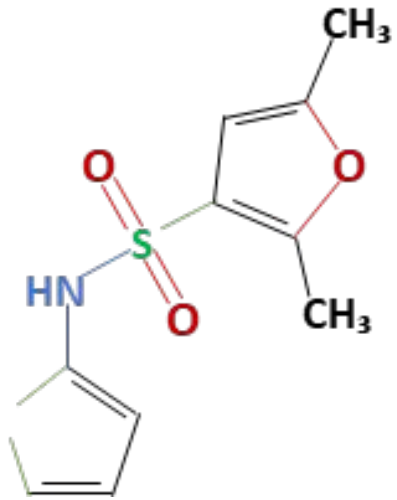


2,5-Dimethylfuran

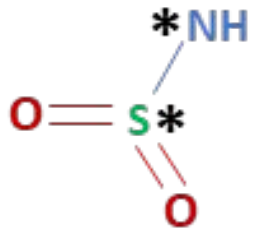


New Molecules

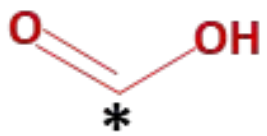




Linkers



Sulfonamide



Carboxylic Acid

FURMAN UNIVERSITY

What is missing from eSynth?

We have a tool for molecular reconstruction, however we need a way to:

- ❏ Analyze input fragments to identify redundancy and similarity among the sets
- ❏ Formally validate molecule reconstruction from its own fragments

How do we “compare” molecules and fragments to find similarity?

Tanimoto Coefficient (TC)

- ❏ TC is a score to measure the similarity of two sets of elements. Our molecules when represented by rdkit molecules can run comparisons like this
- ❏ When TC is close to 1, then the two elements are similar
- ❏ When TC is close to 0, then the two elements have no similarities

Fragment Analyzer

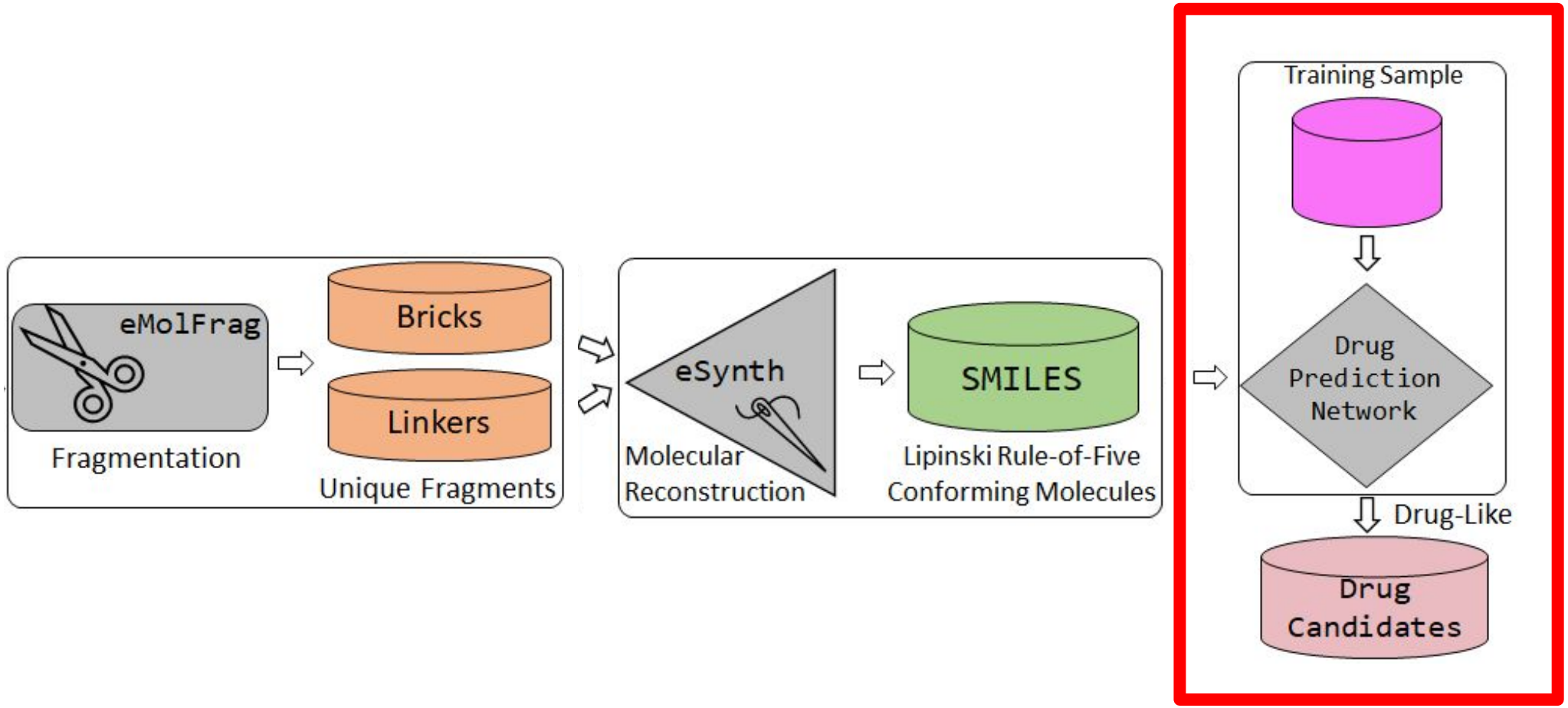
- Frequency Analysis:
  - Takes in a set of fragments
  - Runs a comparison with each fragment in the set with every other fragment in the set.
  - If two fragments are similar the analysis records the relationship between the two

- Distribution Analysis:
  - Takes in a fragment and a set of fragments
  - Compares the fragment to the elements of the set
  - The analysis will tell us how similar our fragment is to the set

On-The-Fly-Validation

- Takes in one file of one or more molecules in mol2 format
- Writes a frequency analysis of how similar the generated molecule is to the input molecule
  - Calculate the Tanimoto Coefficient (TC) of the validation molecule and each generated molecule
  - Keep track of molecules with the same TC value
  - Validates the generated molecule before eSynth writes the result to a file

Next Step


eMolFrag
Fragmentation
Bricks
Linkers
Unique Fragments
eSynth
Molecular
Reconstruction
SMILES
Lipinski Rule-of-Five
Conforming Molecules
Training Sample
Drug
Prediction
Network
Drug-Like
Drug
Candidates

Experimental Results

This is a work in progress, results are pending.

Questions?