Single-Step Retrosynthetic Reaction Prediction with Energy-Weighted Molecular Graphs

David Toomer | CS106X Fall 2019 | Jerry Cain

BACKGROUND

DRUG DISCOVERY

- Drug discovery is a process that is rapidly expanding in biomedicine and pharmacy
- For small drugs (molecules), drug discovery requires knowledge of how to synthesize the molecule
- Would be advantageous for individualized medicine

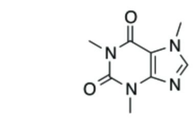
RETROSYNTHESIS

- Retrosynthesis: developing a reaction route to create a target molecule
- Often consists of various **decomposition reactions** navigated by different catalysts and environmental conditions → very time-consuming to predict by hand
- Could be aided using Computer-Aided Synthesis Prediction (CASP)

REPRESENTING MOLECULES

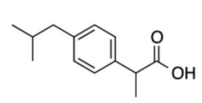
SMILES STRINGS

- Simplified Molecular-Input Line-Entry System (SMILES)
- Can represent any molecule (not unique)



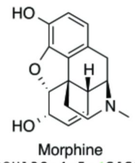
Caffeine

CN1c2ncn(C)c2C(=O)N(C)C1=O



Ibuprofen

CC(C)Cc1ccc(cc1)C(C)C(0)=0

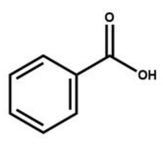


Morphine

[H][C@]12C=C[C@H](O)[C@@H]3Oc4c5c(C[C@H]1N(C)CC[C@@]235)ccc40

MOLFILE

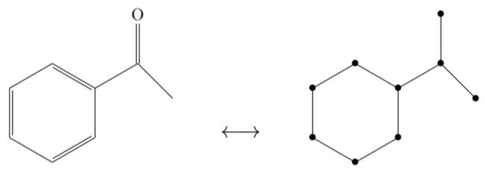
- need to convert SMILES to a more informative data structure → parsing
  - atoms and atom properties (charge, isotopes...)
  - bonds and bond energies, conjugated π-orbitals



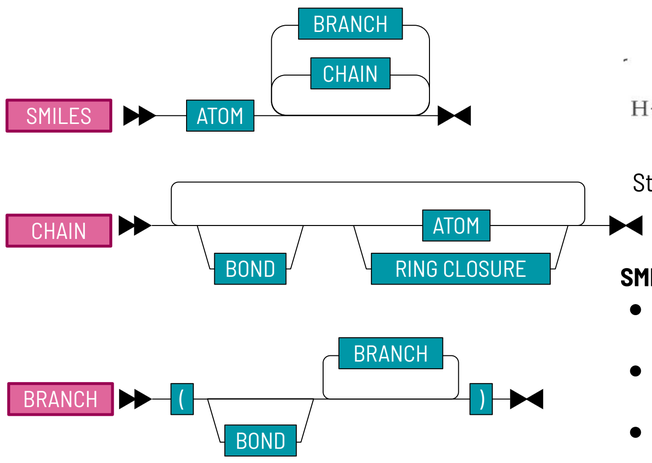
```
  9  9  0  0  0  0  0  0  0  0999 V2000
   -1.4289   -0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.4289   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145   -1.2375    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000   -0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.4289    0.0000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145    1.2375    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0
  2  3  2  0
  3  4  1  0
  4  5  2  0
  5  6  1  0
  6  1  2  0
  5  7  1  0
  7  8  1  0
  7  9  2  0
```

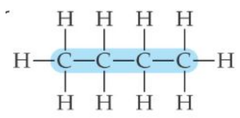
MOLECULAR GRAPHS

- Molecules can be represented as **graphs** where nodes are atoms/functional groups and arc are bonds

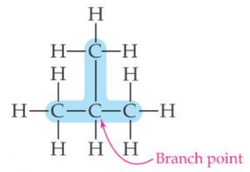


SMILES PARSING: A FORMAL GRAMMAR





Straight-Chain Alkane



Branched Alkane

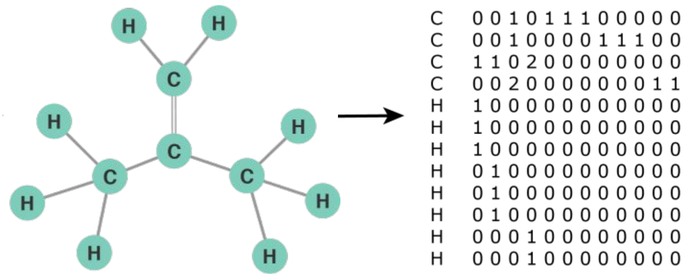
SMILES FORMAL GRAMMAR

- Many different possible chemical configurations to account for
- Researchers have summarized all chemical possibilities as a formal grammar
- Parsing the SMILES with the grammar results in a molecule

REACTION PREDICTION: SPECTRAL CLUSTERING

MOLECULAR GRAPHS

- Molecules can be represented as undirected, weighted graphs
- Edge weights hold approximate bond dissociation energies
- Adjacency matrix holds all the connections (bonds) in the molecule
- Each entry in the adjacency matrix is weighted by its bond dissociation energy → weighted adjacency matrix



```
C  0 0 1 0 1 1 1 0 0 0 0 0
C  0 0 1 0 0 0 0 1 1 1 0 0
C  1 1 0 2 0 0 0 0 0 0 0 0
C  0 0 2 0 0 0 0 0 0 0 1 1
H  1 0 0 0 0 0 0 0 0 0 0 0
H  1 0 0 0 0 0 0 0 0 0 0 0
H  1 0 0 0 0 0 0 0 0 0 0 0
H  0 1 0 0 0 0 0 0 0 0 0 0
H  0 1 0 0 0 0 0 0 0 0 0 0
H  0 1 0 0 0 0 0 0 0 0 0 0
H  0 0 0 1 0 0 0 0 0 0 0 0
H  0 0 0 1 0 0 0 0 0 0 0 0
```

RETROSYNTHESIS PREDICTION PIPELINE:

SMILES STRING → MOLFILE → MOLECULAR GRAPH → RETROSYNTHESIS

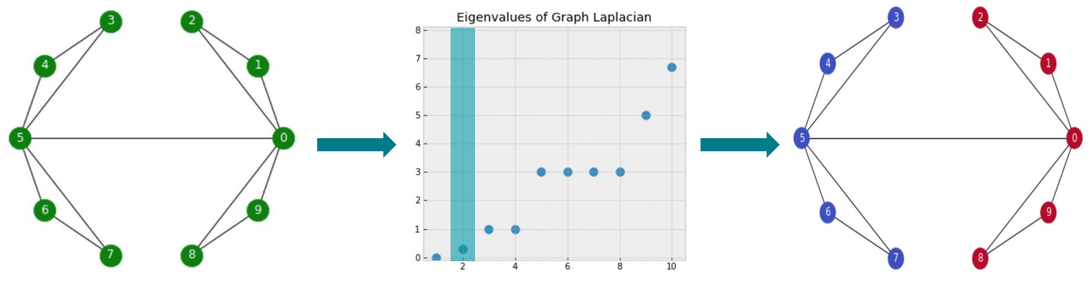
GRAPH LAPLACIAN MATRIX AND SPECTRAL CLUSTERING

- Represent molecules as **undirected, weighted graphs**
- Edge weights hold approximate bond dissociation energies
- Can be represented as a Laplacian matrix → spectral analysis reveals properties of the molecule
  - second-smallest eigenvalue = Fiedler value: shows the **minimum graph cut to make two graphs**
- Can use the Fiedler eigenvector of the Laplacian to make a spectral cut → equivalent to breaking a bond = reaction!

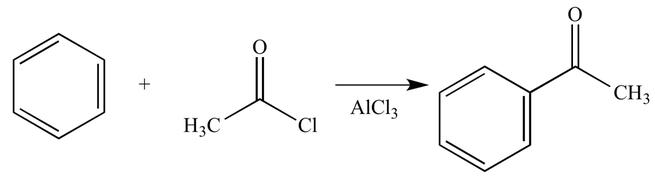
$$D(i,i) = \sum_j W(i,j)$$

$$L = D - W = V \Lambda V^{-1}$$

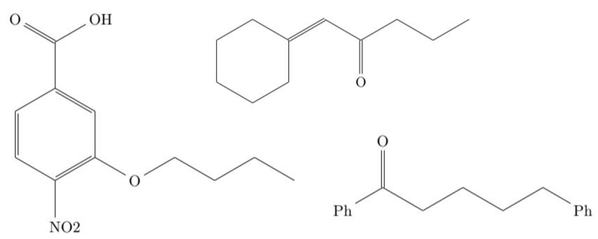
$$\Lambda = diag(\lambda_i) \quad \lambda_i \geq 0$$



EXAMPLES



- Example 1: successful prediction of the formation of acetophenone (right) from benzene (left)
- Catalyst-independent prediction



- Example 2, 3, 4: unsuccessful prediction based on misclustering
- Clustering only off by one atom that is always part of the bond that is broken

SUMMARY

SUCCESSES

- Was able to parse SMILES into a custom data structure
- Used graphs to represent chemicals
- Spectral properties of the graph can represent chemical reactions → accurate for small molecules

IMPROVEMENTS

- Clustering molecules into more than one part by interpreting the next order eigenvector
- More accurate bond energy modeling

FUTURE IDEAS

- Sequentially perform retrosynthesis in order to represent multi-step retrosynthesis
- Predict forward reactions as well
- Validate SMILES for chemical accuracy
- Use GAMESS to create (x,y,z) coordinates for each atom

ACKNOWLEDGEMENTS

CENGAGE Learning™

ChemDoodle by iChemLabs

Armadillo
C++ library for linear algebra & scientific computing