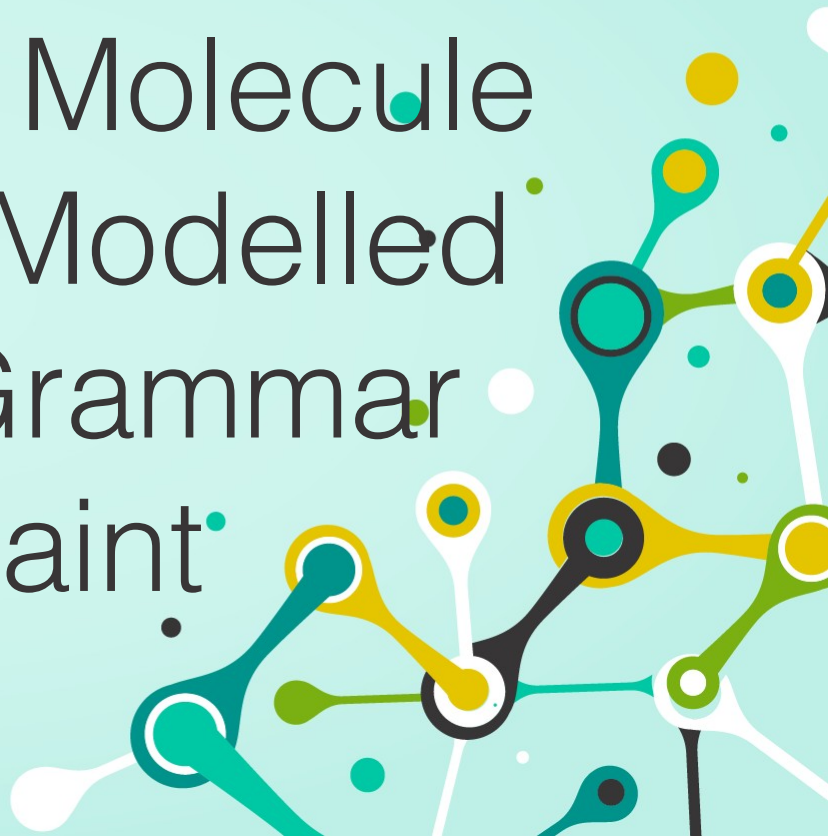
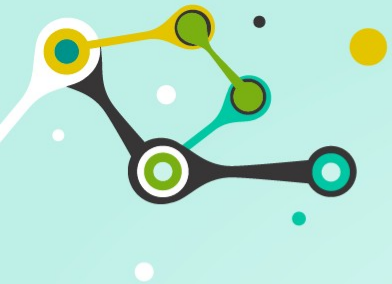
Constrained Molecule Generation Modelled using the Grammar Constraint

David Saikali
Gilles Pesant
Polytechnique Montréal





Contributions

- Application of CP to ***drug discovery***
- Great application for the little-used ***grammar*** constraint
- ***Weighted counting algorithm*** for grammar to guide search

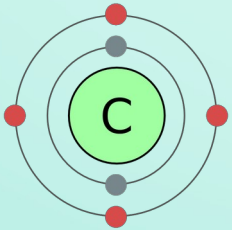
Table of contents

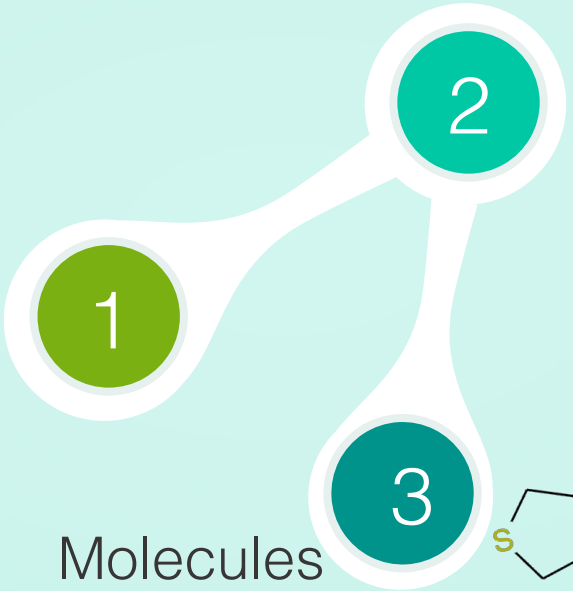
1 Background

A Bit of Chemistry

Valence electrons

The electrons on the outermost layer (valence shell) of an atom



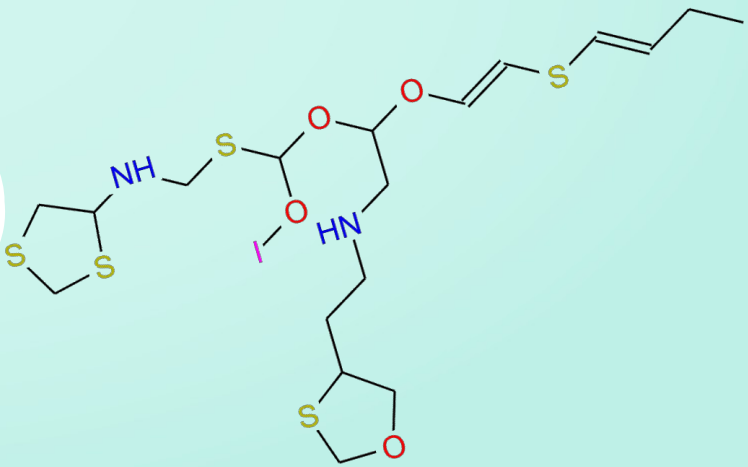


Molecules

Multiple atoms bonded together results in a molecule

Bonds

Valence electrons are used to make bonds with other atoms



A Bit of Chemistry

Lipinski's Rule of 5 (for orally active drugs)

Molecular Weight

Less than 500 Da

Acceptors

Fewer than 10 (2x5) hydrogen-bond acceptors

Donors

Fewer than 5 hydrogen-bond donors

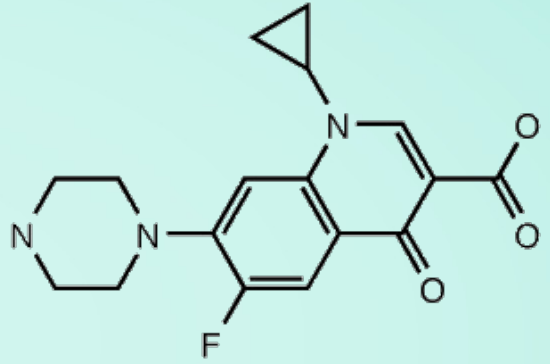
LogP (hydrophobic)

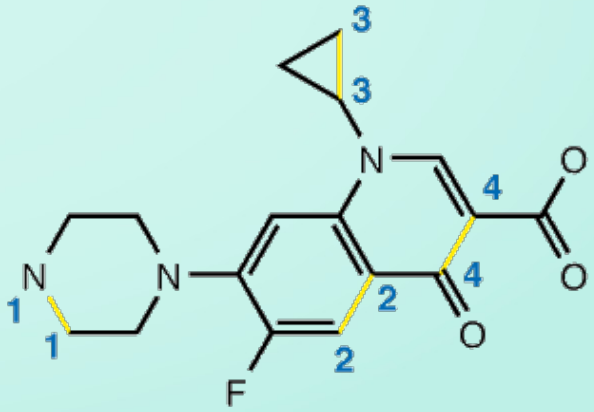
Should not exceed 5

SMILES (standard string representation)

(Carbons and hydrogens are implicit)

1. Number the cycles
2. Break a bond from each cycle
3. Add a token to the previously linked atoms

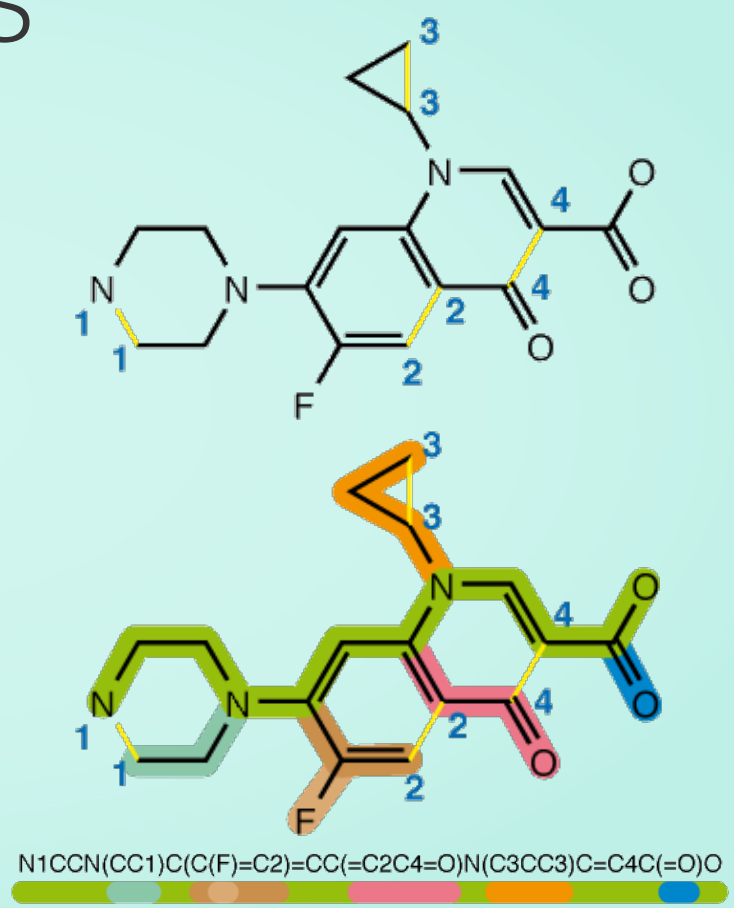




SMILES

4. Choose an arbitrary starting point
5. Explore the molecule using a DFS
6. Write down visited atoms
   (each branch put in parentheses)

There is a context-free grammar for SMILES



Problem

Our Goal

- Explore the design space of molecules efficiently and automatically
  (drug discovery is very time consuming and costly)
- Generate valid molecules using CP
- Actively filter out molecules that don't feature some given target properties

CP Model & Search

SMILES constraints

SMILES syntax

Generated results are a part of the grammar language: SMILES

Grammar

Cycle numbering

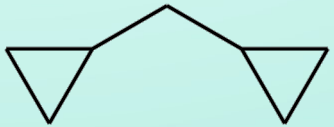
Cycle numbering starts at 1 and doesn't skip a number
C1CC2CC3CC4 still allowed!

Regular

Cycle closing

Cycle numbers come in pairs or don't appear
C1CC1CC2CC2

Among



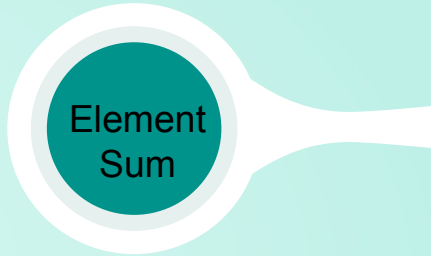
Targeting regions of the design space

Cycle count

Constrain wanted cycle number to appear twice and next cycle to never appear

Molecular weight

Associates an integer weight to each token. Estimation to account for implicit hydrogens.

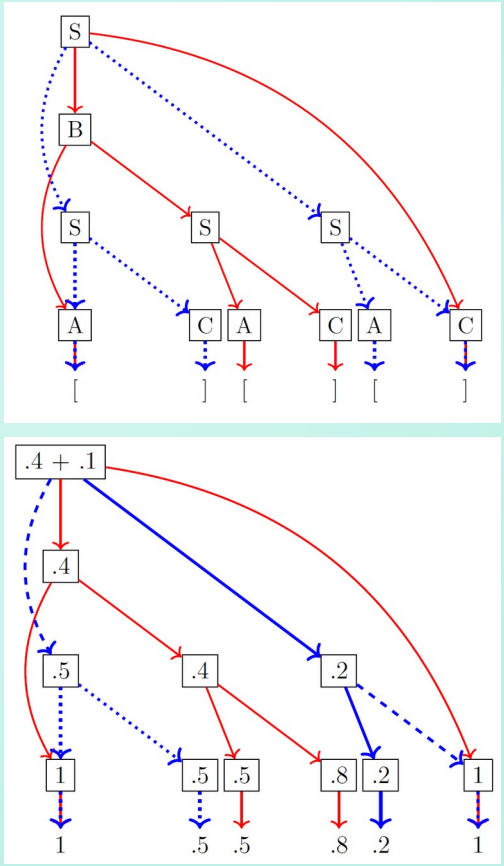


Branch count

Constrain number of occurrences of opening-branch token.

Weighted counting algorithm for ***grammar***

- Belief propagation (*MiniCPBP*): informed messages about likelihood of domain values
- Messages from constraints = weighted counting of their solutions
- ***Guiding search*** using these beliefs (e.g. *maxMarginalStrength*)



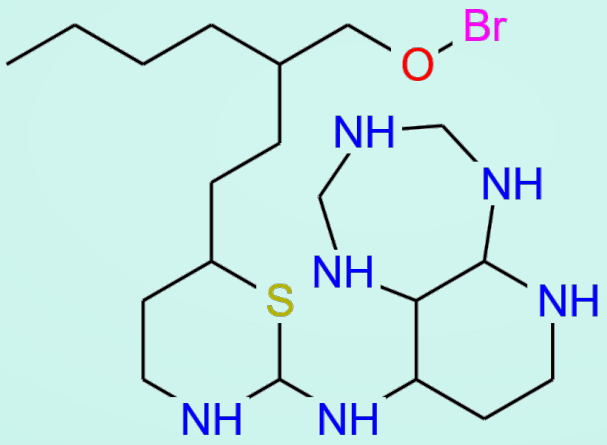
4 Results

De novo 1D molecule optimization

40 characters; 475 Da ≤ molecular weight ≤ 500 Da; one-hour time limit

	domWdeg / random		maxMarginalStrength		maxMarginalStrength LDS	
instance	time (s)	fails	time (s)	fails	time (s)	fails
c1b2	**6.1**	6	6.2	**0**	6.8	**0**
c1b3	5.9	13	**4.2**	**0**	5.2	**0**
c1b4	**5.9**	17	**5.9**	**1**	6.3	**1**
c2b2	23.8	826	**4.9**	**0**	**4.9**	**0**
c2b3	7.7	171	**4.8**	**0**	5.4	**0**
c2b4	10.8	569	**5.9**	**0**	6.0	**0**
c3b2	-	-	**7.3**	**0**	**7.3**	**0**
c3b3	-	-	-	-	**79.6**	**93**
c3b4	-	-	-	-	**12.7**	**17**

Example of generated molecule



c3b3: 6 donors, 8 acceptors, 1.62 logP
BrOCC(CCCC)CCC1CCNC(NC2CCNC(C2NCNC3)N3)S1

Conclusion

Conclusion

Contributions

- Application of CP to ***drug discovery***
- Great application for the little-used ***grammar*** constraint
- ***Weighted counting algorithm*** for grammar to guide search

Current work

- Complete constraints for Lipinski's Rule of 5 (hydrogen-bond acceptors and donors, logP)
- Add « realism » to molecules by combining with NN (ChemBERTa) trained on SMILES molecules dataset