Создание уникального репертуара малых молекул на основе образца

Бондарев Тимофей Сергеевич

научный руководитель: П.А. Яковлев

СПб НИАУ РАН

16 июня 2015 г.

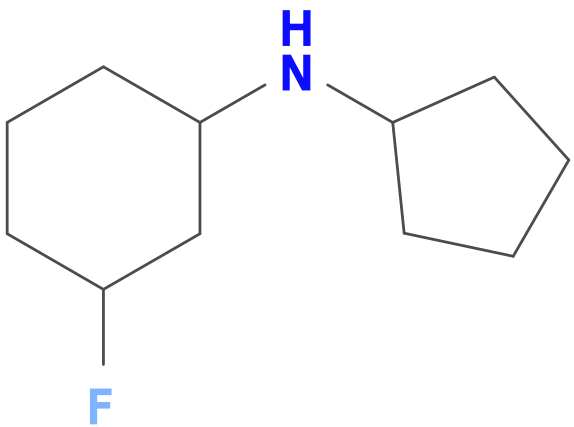
Введение

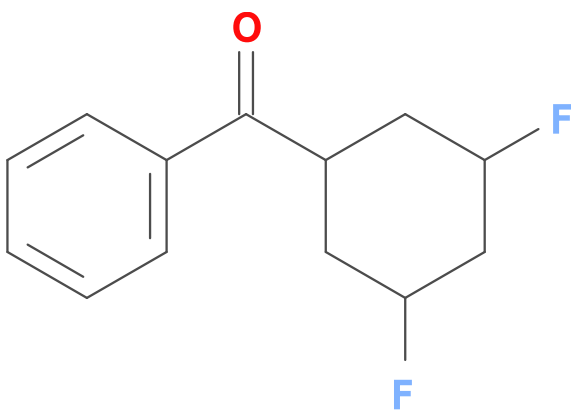
Малые молекулы — молекулы, массой не более 900 а.е.м.

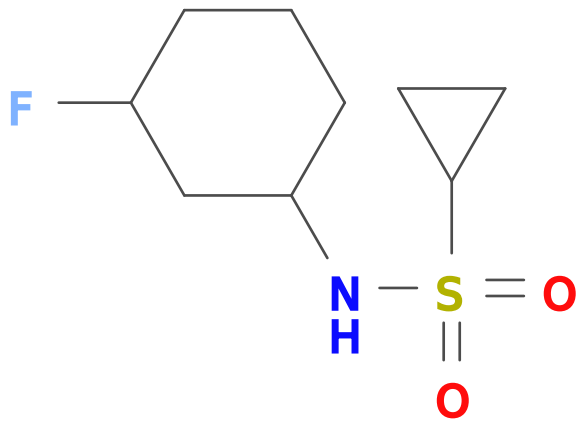
Большинство лекарств являются малыми молекулами.

Один из способов получения лекарственных средств: генерация и выбор подходящего соединения.

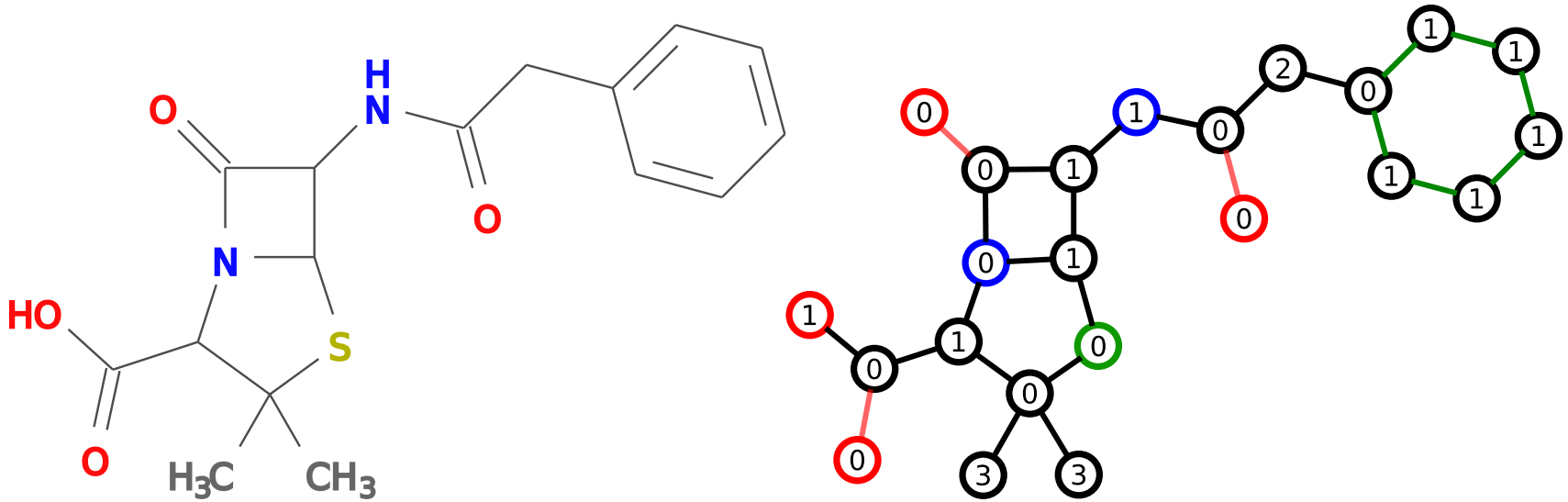
Примеры малых молекул:







Представление молекулы в виде графа



- убраны атомы водорода;
- чёрным вершинам соответствуют атомы углерода (C), красным — кислорода (O), синим — азота (N), зелёным — серы (S);
- внутри вершины указано количество связей с водородом;
- чёрные рёбра соответствуют одинарным связям, красные — двойным, зелёные — ароматическим;

Существующие решения

Решения разбиваются на три вида:

- Полное перечисление всех графов.
  - \+ Нет пропущенных структур.
  - — Очень долгая генерация.
  - — Много бесполезных молекул.
- Составление молекул из фрагментов **[Firth, et al., 2015]**;
  - \+ Меньше бесполезных молекул.
  - \+ Выше вероятность, что полученная молекула синтезируется.
  - — Нет гарантии, что найдётся подходящая молекула.
- Генерация на основе химических реакций **[Hartenfeller, 2012]**.
  - \+ Быстрое перечисление результата.
  - \+ Известен способ синтеза полученных молекул.
  - — Получение только молекул с известными реакциями.
  - — Нет гарантии, что найдётся подходящая молекула.

Цель и задачи работы

Цель

Реализовать генерацию **уникальных** молекул, использующую базовую молекулу и набор радикалов для изменения этой молекулы. Предоставить возможность указывать атомы базовой молекулы, в которых разрешено производить изменения.

Задачи

1. Описать формат для представления молекулы с помеченными атомами.
2. Составить алгоритм для генерации уникальных молекул по базовой молекуле и одному радикалу.
3. Расширить алгоритм на произвольное количество радикалов.

Формат представления молекул

SMILES

— спецификация однозначного описания структуры и состава молекулы в виде последовательности символов.

Атомы записываются символами их химических элементов;

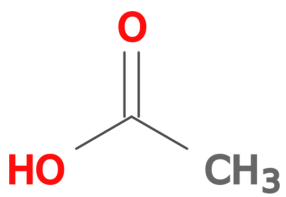
атомы водорода не указываются;

для указания ветвления молекулы используются круглые скобки;

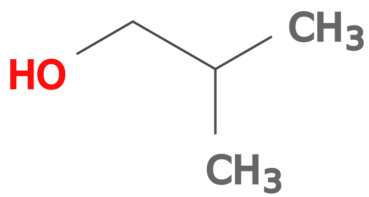
одинарные связи обозначаются символом -, но обычно не указываются, двойные обозначаются символом =, тройные — символом #.

Примеры:

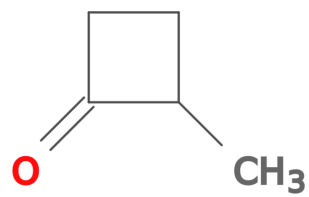
`CC(=O)O`



`C(C)(C)CO`



`CC1C(=O)CC1`



Расширение формата SMILES

STAR-SMILES

— расширение спецификации SMILES.

Для пометки атома, после его появления в строке SMILES, вставляется специальная последовательность, несущая дополнительную информацию.

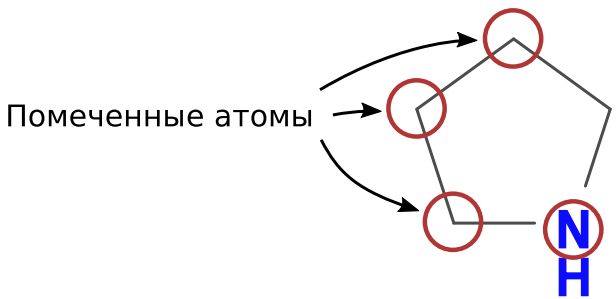
Сейчас используются символы *, **, *** для обозначения возможных изменений вершины,

<"список типов рёбер">, например, <=>, <-=#>, для обозначения свободного ребра фрагмента.

Генерация, использующая один радикал

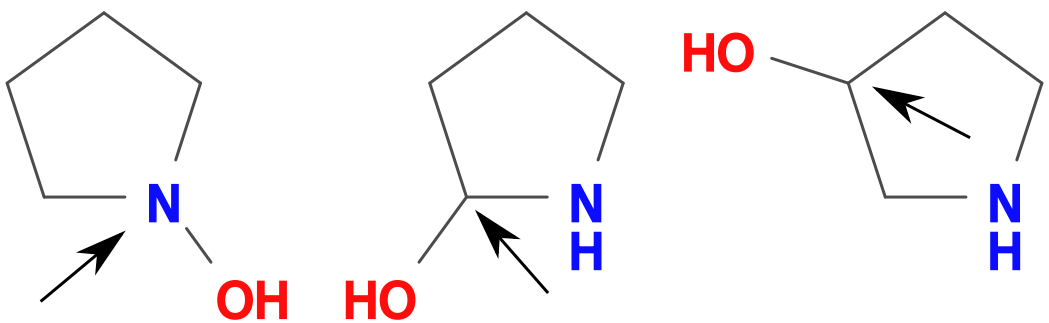
Имеется базовая молекула с помеченными атомами.
Можно изменять структуру этой молекулы, используя радикал.
Необходимо получить минимальный набор уникальных молекул, получаемых изменением базовой молекулы радикалом.

Базовая молекула



Радикал

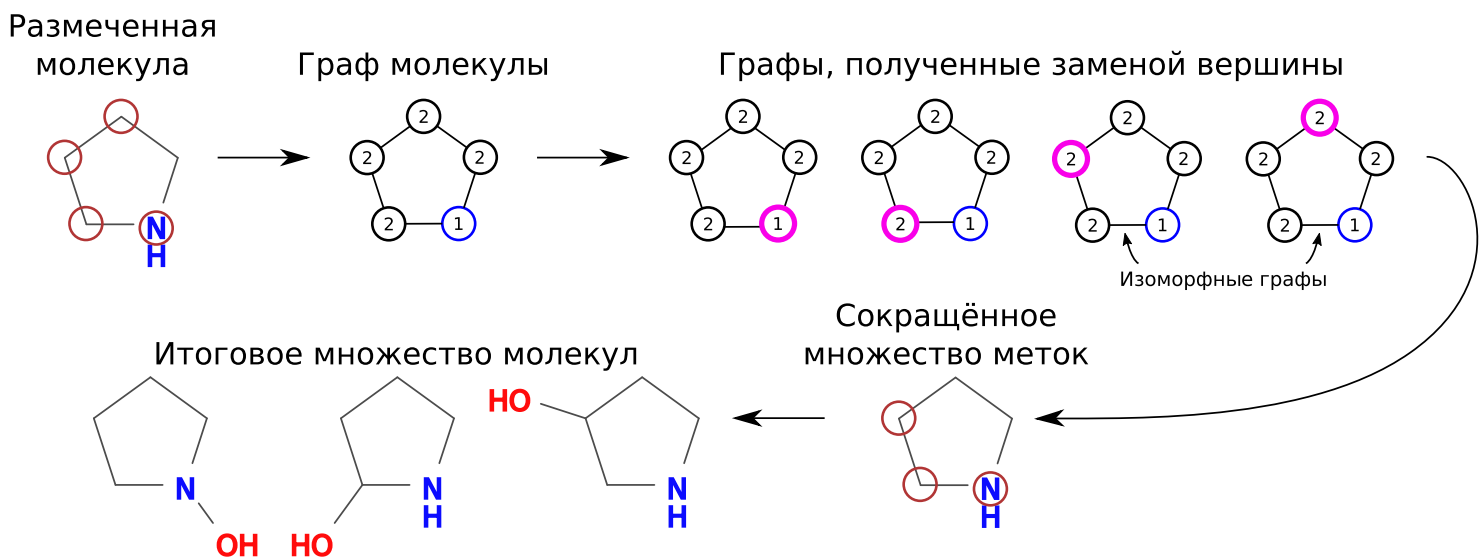
Получаемое с помощью радикала множество молекул



Сокращение множества помеченных атомов

Молекула преобразуется в граф по описанной ранее схеме. Молекулы могут иметь геометрически симметричные фрагменты ⇒ необходимо найти множество вершин графа, изменение в которых достаточно для генерации результата.

Каждый атом молекулы помечается уникальной меткой, после чего запускается проверка изоморфизма всех пар полученных графов. Неизоморфные графы соответствуют меткам, которые требуется оставить.

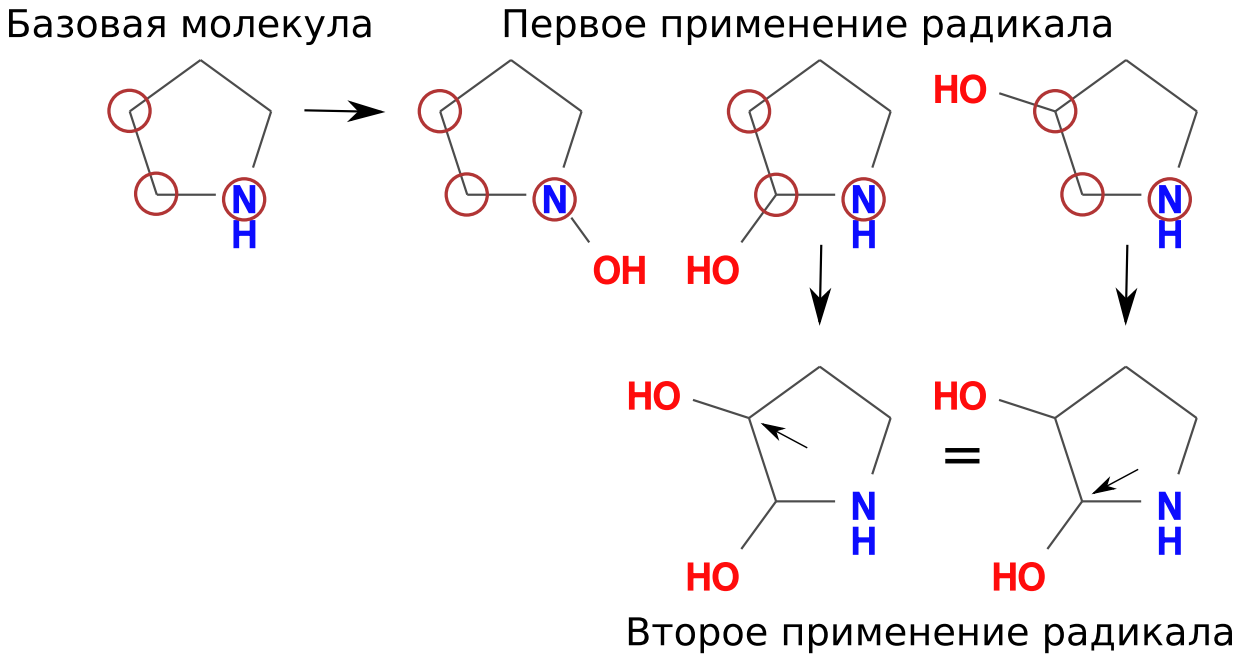


Генерация с произвольным количеством радикалов

Задача состоит в следующем:

1. Применение каждого радикала к помеченным атомам базовой молекулы.
2. Применение каждого радикала к молекулам, полученным на предыдущем этапе.

При выполнении пункта 2 могут появиться совпадающие молекулы:



Удаление повторяющихся молекул

Для удаления повторяющихся молекул используется следующий подход:

- Для каждого графа молекулы вычисляется ядро (Graph Kernel) по алгоритму NSPDK (Neighborhood Subgraph Pairwise Distance Kernel), используя реализацию авторов статьи **[Costa,Grave(2010)]**. Ядро графа представляет собой вектор.
- Для каждой пары векторов находится расстояние друг от друга.
- Отбрасываются графы, ядра которых отличаются на меньшую величину, чем заданная отсечка.

Итоги

- Разработан практичный формат для представления молекулы с помеченными атомами.
- Реализован алгоритм получения уникальных молекул на базе одной молекулы и радикала.
- Реализовано расширение алгоритма на произвольное количество радикалов.
- Реализовано приложение на языке Python 2, с использованием библиотек rdKit для работы с форматом SMILES, networkx для хранения графа молекулы, EDeN для поиска ядер графов. Приложение находится на этапе внедрения в процессы разработки лекарств в компании BIOCAD.

Спасибо за внимание!