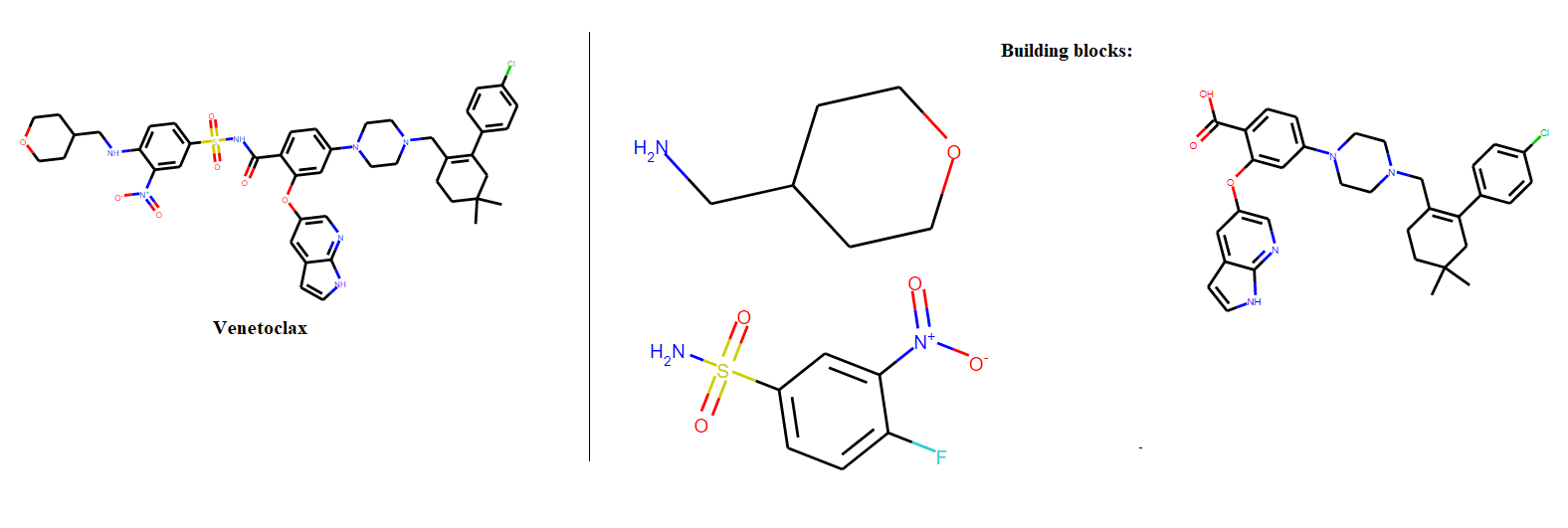
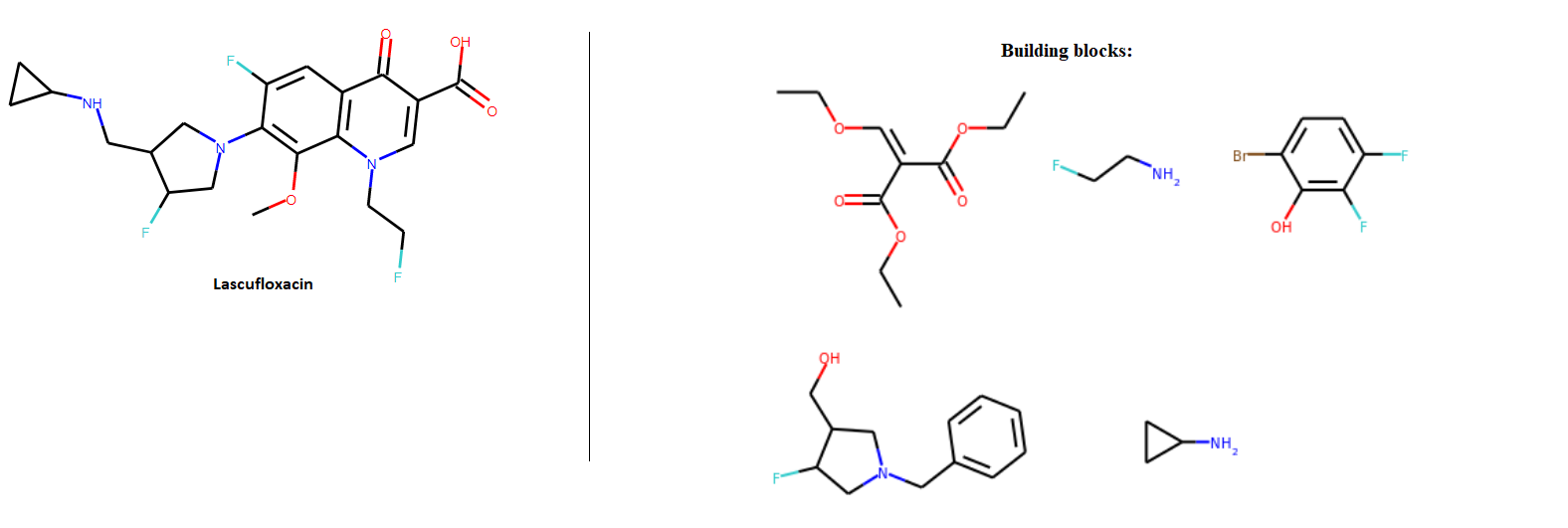
**Цель:** построить регрессионную классификационную модель/алгоритм предсказания конверсии (степень разложения) молекулы в конечные билдинг блоки, полученные в ходе ретросинтеза или представленные в литературном синтезе, на основе сформулированных механистических правил.

Модель/алгоритм получает на вход SMILES исходной молекулы и SMILES билдинг блоков и возвращает один из двух классов: **low** (низкая конверсия), **high** (высокая конверсия). SMILES передаваемых молекул содежат атом маппинги, обозначающие соответствие между атомами целевой молекулой и этими же атомами, обнаруживаемыми в билдинг блоках. Примеры очень низкой и высокой конверсии:



**Рис.1** Venetoclax и билдинг блоки из ретросинтетического пути с очень низкой конверсией. Для облегчения визуализации атом маппинги убраны.



**Рис.2** Lascufloxacin и билдинг блоки из литературного синтеза с высокой конверсией. Для облегчения визуализации атом маппинги убраны.

**Задачи:**

1. Используя критерии, разработанные для expert-based оценивания конверсии датасета с синтетическими путями (см. ниже), алгоритмизировать расчет следующих фичей:

* Проверка наличия в билдинг блоке трициклического конденсированного цикла (*бинарное значение фичи - есть/нет*)
* Есть ли в билдинг блоке бициклический конденсированный гетероцикл с двумя и более гетероатомами хотя бы в одном из билдингов и двумя и более заместителями (*бинарное значение - есть/нет*)
* Подсчет в билдинге числа связей, которые разбиваются робастной реакцией (*целочисленное значение фичи*): достаточно имплементировать **половину из них**

1. Сузуки

2. Михаэль

3. Виттиг/HWE

4. Ацилирование (O, N)

5. Алкилирование (O, N, S, SN), кроме OMe/Et, NMe/Et

6. SNAr (O, N, S), кроме OMe/Et, NMe/Et

7. Бухвальд-Хартвиг/Ульман, кроме NMe/Et

8. Гриньяр

9. Фридель-Крафтс

*Наличие центра реакции можно проверять матчингом подструктуры с использованием SMARTS, описывающего реакционный центр. При подсчете таких связей игнорировать защитные группы (проверять атом маппинг внутри реакционного центра)*

* Если ли билдинг с бициклическим конденсированным циклок с двумя связями, разбиваемыми робастной реакцией (*бинарное значение*)
* Оценка, есть ли среди билдингов как минимум два билдинга с бициклическими конденсированными гетероциклами с двумя гетероатомами и как минимум двумя заместителями (*бинарное значение*)
* В ходе ретросинтеза проведены только минимальные модификации целевой молекулы (замена одной-двух функциональных групп), *бинарное значение*. Можно оценивать соотношение/распределение числа замапленных атомов в билдинг блоках относительно числа атомов в целевой молекуле.
* Подсчет числа циклов в билдинге. (*целочисленное значение*)

Важное замечание - не учитывать циклы, входящие в состав защитных и уходящих групп билдинга. В представленном датасете путей такие группы не содержат атом маппингов в SMILES билдинг блоков.

* Подсчет числа заместителей в карбоциклических и гетероциклических системах: простых или из состоящих из двух циклов (*целочисленное значение*)
* Подсчет числа гетероатомов в гетероциклических системах (*целочисленное значение*)