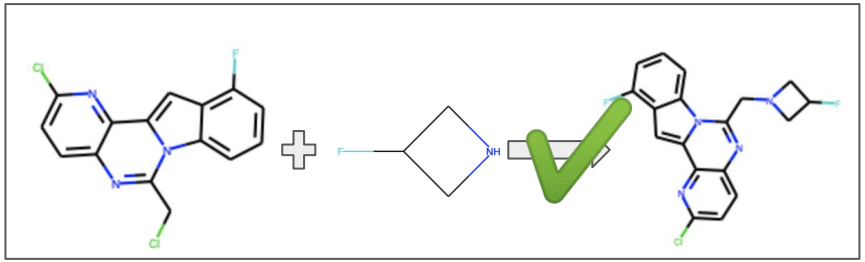
Предсказание синтезируемости молекул

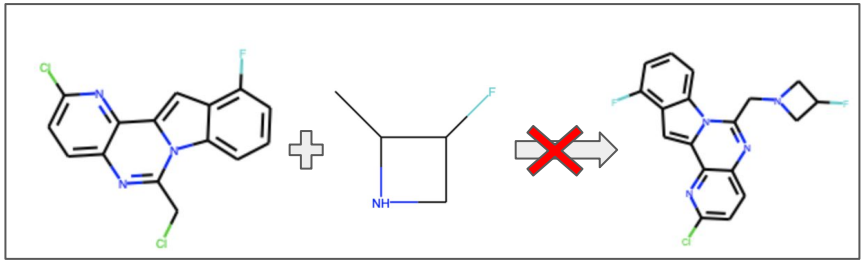
Федотов Александр Сергеевич
Научный руководитель: Шпильман Алексей Александрович
Санкт-Петербургская школа физико-математических
и компьютерных наук

НИУ ВШЭ - Санкт-Петербург, 2020 г.

Введение в область

- Этапы получения лекарства:
  - Создаётся набор молекул кандидатов, обладающих нужными свойствами
  - Для каждого из них создаются пути ретросинтеза - последовательности реакций для синтеза вещества
  - Получившиеся реакции тестируются на реализуемость
- Проблема
  - Существующие методы предлагают огромное множество возможных реакций
  - Среди них большая часть некорректна или не осуществима химически
  - Необходимо быстро отсеивать такие реакции перед тем, как ставить химические эксперименты





Варианты решения

- Квантовая физика [1]
  - Отличное качество, общность
  - Высокая стоимость, возрастающая с размером молекул
- Системы, основанные на химических правилах [2]
  - Хорошее качество, дешевизна применения
  - Отсутствует общность, дорогая расширяемость
- **Машинное обучение [3, 4]**
  - Общность, дешевизна применения
  - Невысокое качество, необходимость большого числа данных
  - Большинство работает только с одним типом реакций и не использует не проходящие реакции для обучения

1. Wang, B., & Cao, Z. (2010). Mechanism of acid-catalyzed hydrolysis of formamide from cluster-continuum model calculations: concerted versus stepwise pathway. The Journal of Physical Chemistry A, 114(49), 12918-12927.
2. Chen, J. H., & Baldi, P. (2009). No electron left behind: a rule-based expert system to predict chemical reactions and reaction mechanisms. Journal of chemical information and modeling, 49(9), 2034-2043.
3. Fooshee, D., Mood, A., Gutman, E., Tavakoli, M., Urban, G., Liu, F., ... & Baldi, P. (2018). Deep learning for chemical reaction prediction.
4. Kayala, M. A., & Baldi, P. F. (2011). A machine learning approach to predict chemical reactions.

Цель и задачи

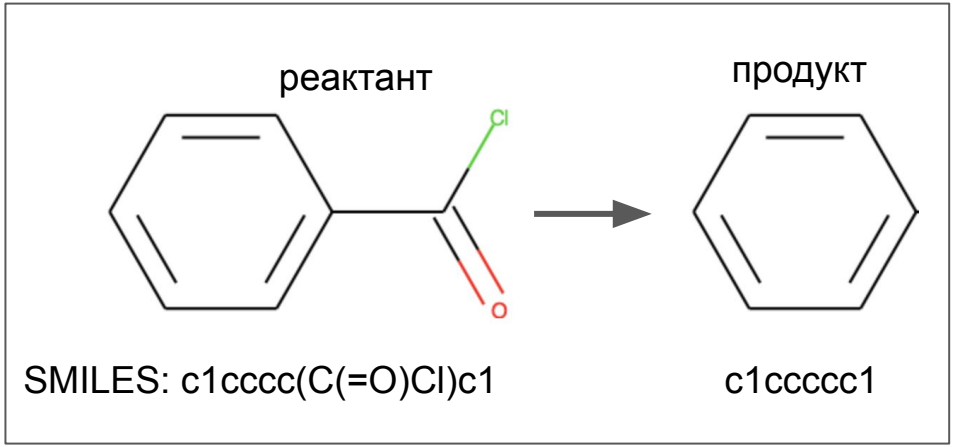
Цель:

Создать модель для предсказания успешности химической реакции, используя методы глубокого обучения

Задачи:

- Разработать генеративную модель для добавления негативных примеров в датасет
- Создать классификатор для обучения на полученном датасете
- Оценить качество полученного решения автоматически и с помощью экспертной оценки

Данные (BIOCAD)



- SMILES [1]
  - Кодировка молекул в виде строки
- Структура данного датасета
  - 400 000 примеров реакций
  - Реакции с 1 или 2 реактантами и одним продуктом
- Приведены только успешные реакции - отсутствуют размеченные негативные примеры для классификации
- В датасете могут быть ошибки
- Нужно дополнить датасет новыми примерами, при этом они должны быть достаточно сложными

1. Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of chemical information and computer sciences, 28(1), 31-36.

Разработка генеративной модели

Необходимо протестировать варианты:

- Модель
  - Используются существующие алгоритмы для генерации молекул, полученные из обработки естественного языка
- Способ кодировки реакций
  - Рассматриваются только строчные форматы
  - Адаптируются различные существующие подходы для кодировки молекул под реакции

Модели:

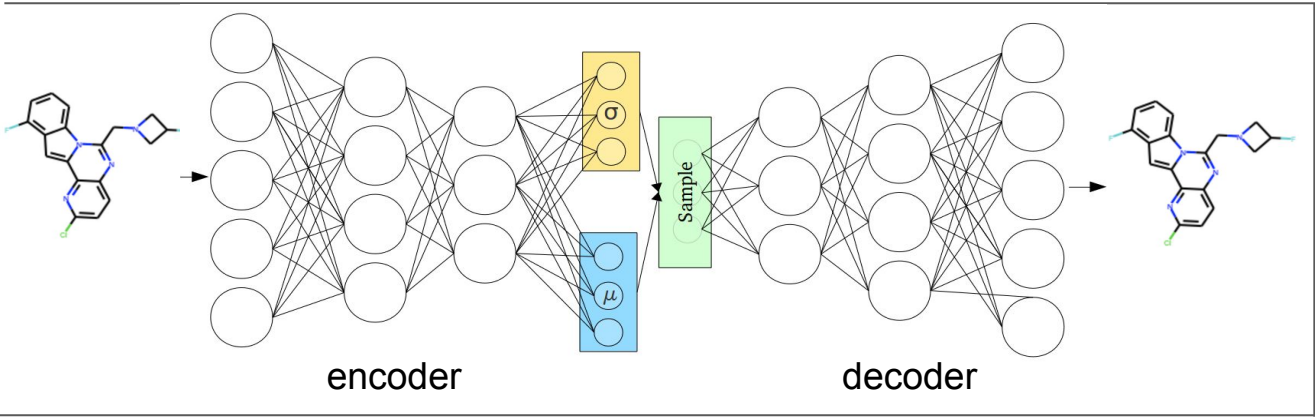
- Variational Autoencoders (VAE)
- Grammar VAE
- Char RNN

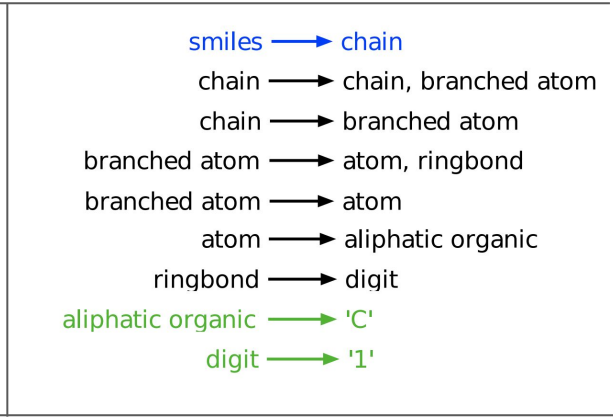
Кодировки реакций:

- SMILES
- DeepSMILES
- SELFIES

Variational Autoencoder (VAE) [1]

- Используется для генерации молекул [2]
- Grammar VAE [3]
  - Молекула представляется набором правил грамматики для строк в формате SMILES
  - Гарантия генерации синтаксически корректных строчек

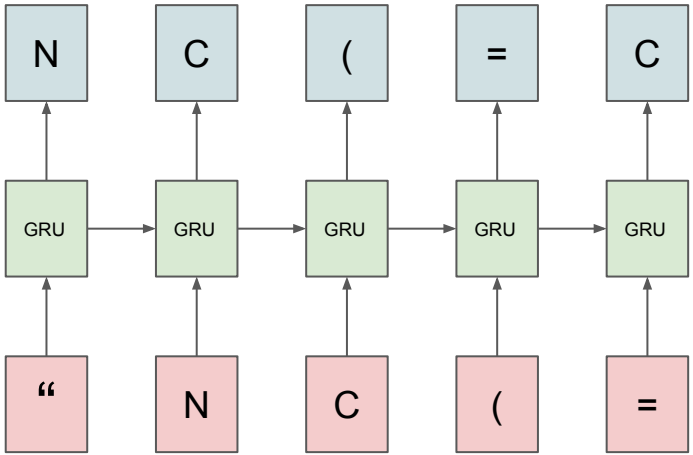




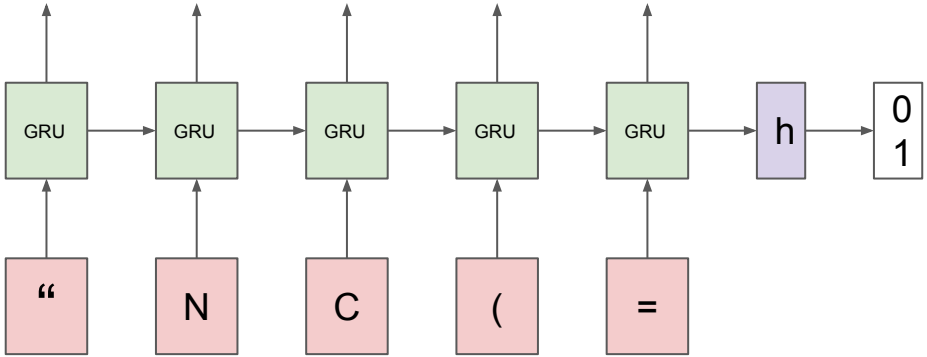
1. Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., & Carin, L. (2016). Variational autoencoder for deep learning of images, labels and captions.
2. Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J., & Chen, H. (2018). Application of generative autoencoder in de novo molecular design
3. Kusner, M. J., Paige, B., & Hernández-Lobato, J. M. (2017, August). Grammar variational autoencoder.

Char RNN [1]

Также используется для генерации молекул [2]



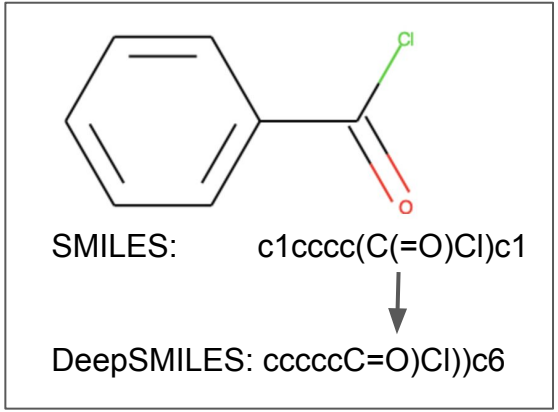
Архитектура была адаптирована для решения задачи классификации



1. Choi, K., Fazekas, G., & Sandler, M. (2016). Text-based LSTM networks for automatic music composition.
2. Segler, M. H., Kogej, T., Tyrchan, C., & Waller, M. P. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks.

Базовые строковые форматы

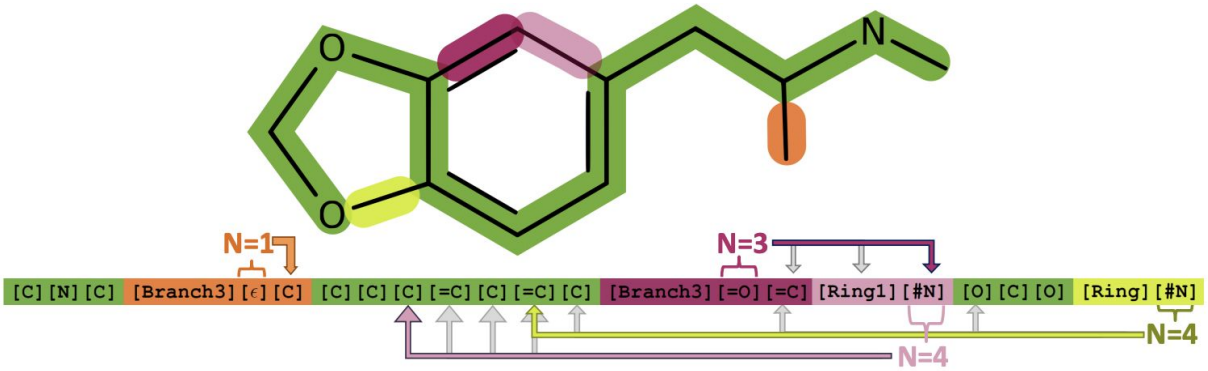
- Адаптация строковых кодировок молекул к реакциям
  - “c1cccc(C(=O)Cl)c1.C=O>>c1cccc(C=O)c1"
- SMILES
  - Плохо подходит для машинного обучения
- DeepSMILES [1]
  - Нет открывающих скобок и только одна цифра для цикла
  - Упрощение генерации



1. O'Boyle, Noel; Dalke, Andrew (2018): DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. ChemRxiv. Preprint.

SELFIES [1]

- Строковый формат
- Грамматика для него контролирует корректность циклов и ветвей
- Таким образом при генерации молекул все из них являются корректными



1. Krenn, M., Häse, F., Nigam, A., Friederich, P., & Aspuru-Guzik, A. (2019). SELFIES: a robust representation of semantically constrained graphs with an example application in chemistry. arXiv preprint arXiv:1905.13741.

Измерение качества генерации

- **Valid%** - процент реакций, все участвующие в которых молекулы корректны
- **Similar%** - процент реакций с разницей по количеству атомов каждого элемента не более 2
- **MCS mean** - средняя общая часть молекул в реакции

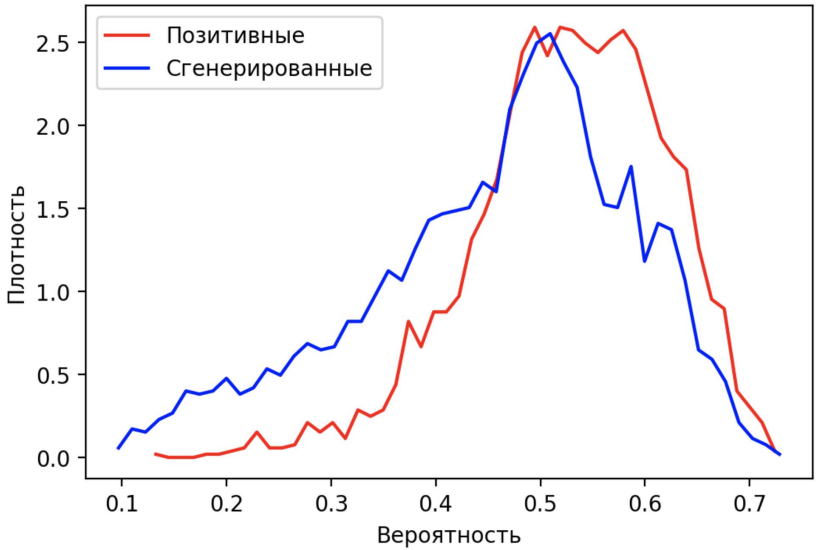
Результаты на датасете с 1 реактантом

Метод	Valid%	Similar%	MCS mean
Исходный датасет	100	35	15.4
Случайный датасет	100	9	9.4
SMILES + VAE	0.8	-	-
DeepSMILES + VAE	1.3	-	-
Grammar VAE	0.1	-	-
SELFIES + VAE	100	20	9.4
SELFIES + Char RNN	**100**	**20**	**12.8**
SELFIES + Char RNN + аугментация	82	24	10.9

Классификация

- Датасет с одинаковым количеством исходных и сгенерированных примеров
- Классы
  - 1 - Примеры из исходного датасета
  - 0 - Сгенерированные примеры
- Вывод - большое число сгенерированных примеров близки к исходным и могут являться позитивными

Плотности предсказанных классов



Экспертная оценка

- Специалисты из BIOCAD разметили по 50 примеров из обоих классов с большими и маленькими предсказанными вероятностями

Количество реакций из предложенных

Настоящий класс	Предсказанный класс	**Проходят**	**Проходят в несколько стадий**	**Не проходят**
Исходные	**Исходные**	5	37	8
Исходные	**Сгенерированные**	21	17	12
Сгенерированные	**Исходные**	1	17	32
Сгенерированные	**Сгенерированные**	3	5	42

Результаты

- С помощью разработанной генеративной модели исходный датасет дополнен потенциально негативными примерами
- Создан и обучен классификатор на полученном датасете
- Экспертная оценка показала, что в сгенерированных примерах содержатся потенциально позитивные, а в исходном датасете - потенциально негативные
- Классификатор показал способность к обучению на таких данных
- В планах - реализация поэтапного обучения моделей классификации и генерации

Результаты на других датасетах

Результаты на датасете с 2 реактантами

Метод	Valid%	MCS mean
Исходный датасет	**83**	**5.7**
Случайный датасет	10	4.9
VAE	28	4.9
Char RNN	**61**	**5.4**

Результаты на полном датасете

Метод	Valid%	MCS mean
Исходный датасет	**53**	**11.9**
Случайный датасет	8	7.8
VAE	22	7.7
Char RNN	**39**	**10.4**