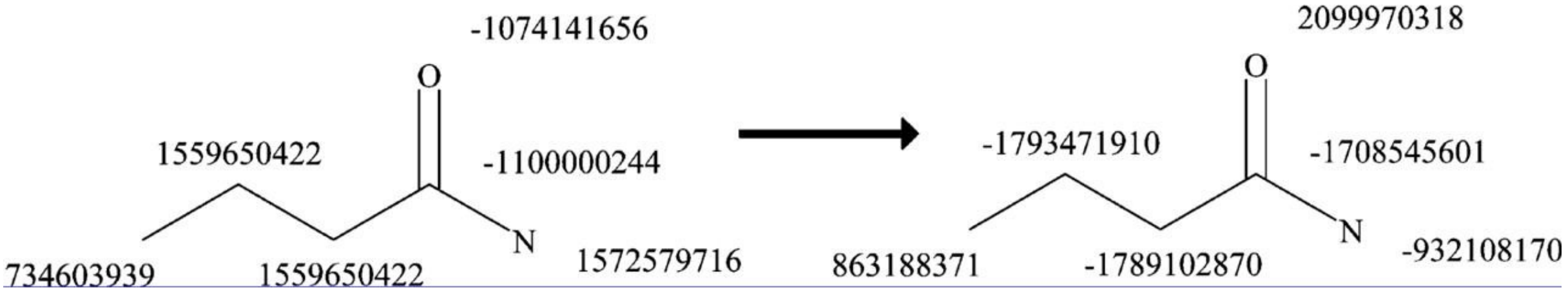
Расширение открытий с помощью структурных моделей: Мультимодальная комбинация локальных и структурных свойств для предсказания химических особенностей

Автор: *Рекут Н.А.*

Научный руководитель: *к.ф.м. Безносиков А.Н.*

Химическая часть проекта: SMILES и ECFP

- Пример SMILES-кодировки: COc(c1)cccc1C#N (молекулярная формула C8H7NO)
- Построение ECFP (пример ECFP для этанола: [['2246728737', '2245384272', '864662311'], ['3542456614', '4018048386', '1535166686']])



Распределения по количеству уникальных токенов

- ECFP0: 50k
- ECFP1: 100k
- ECFP2: 600k
- ECFP3: 1.5kk

Таким образом, работа с ECFP в чистом виде невозможна, нужна некоторая токенизация

Идеи токенизации

- BPE
- Кластеризация (data-driven)
- Сделать собственные фингерпринты и использовать bpe на них

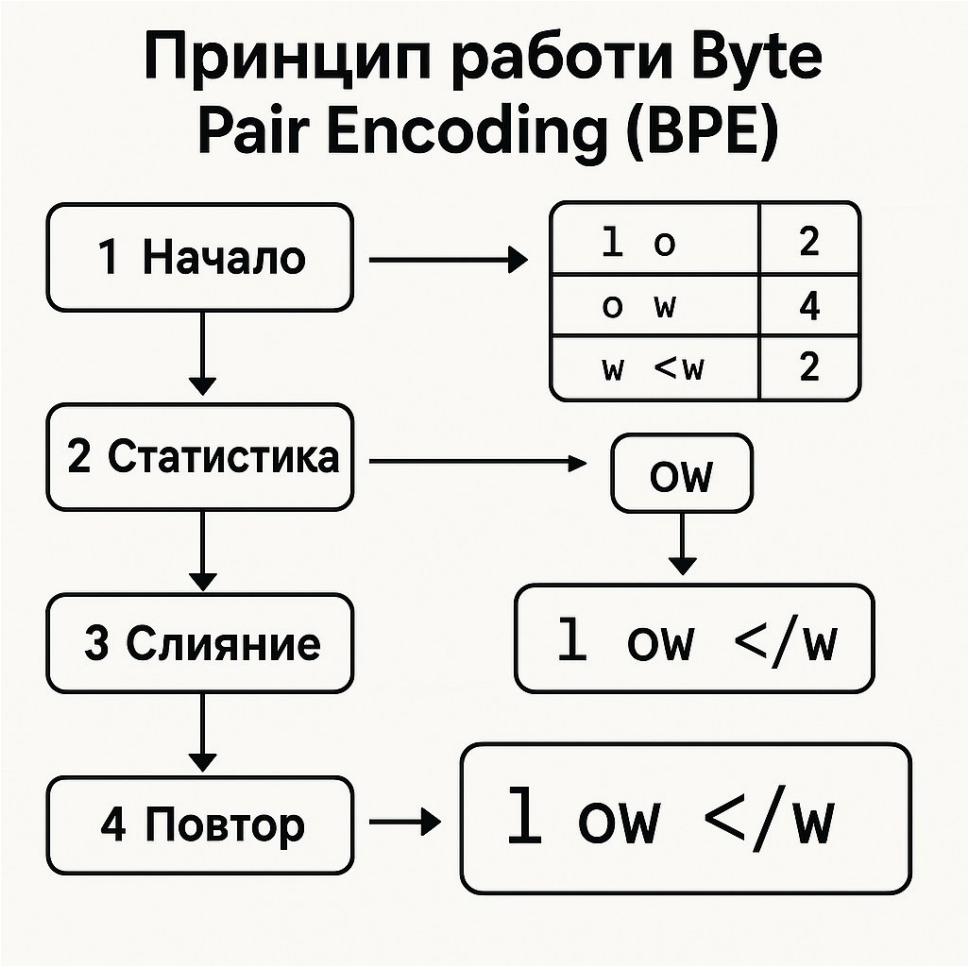
BPE

Byte Pair Encoding (BPE) — это алгоритм субсловного разбиения, позволяющий эффективно кодировать слова, встречающиеся редко, путём объединения часто встречающихся пар символов.

Этапы работы:

- **Начало**: каждое слово разбивается на отдельные символы + специальный символ конца слова (например, "low" → l o w </w>).
- **Статистика**: считаем частоту пар символов во всем корпусе текста.
- **Слияние**: находим самую частую пару и объединяем её в новый токен.
- **Повтор**: повторяем шаги 2–3 до достижения заданного числа слияний (мерж-операций).

Единственная проблема — мы раздуваем исходный словарь, вместо того чтобы сжимать его.



Data-driven подход

Этапы:

- Обучаем на всём корпусе контекстуальные эмбеддинги (word2vec)
- На их основе делаем кластеризацию при помощи KNN, задавая незначительное количество элементов в каждом кластере (около 10)
- Объединяем токены из одного кластера в один новый токен

Существенный недостаток заключается в нехимичности данного подхода.

Создание новых отпечатков

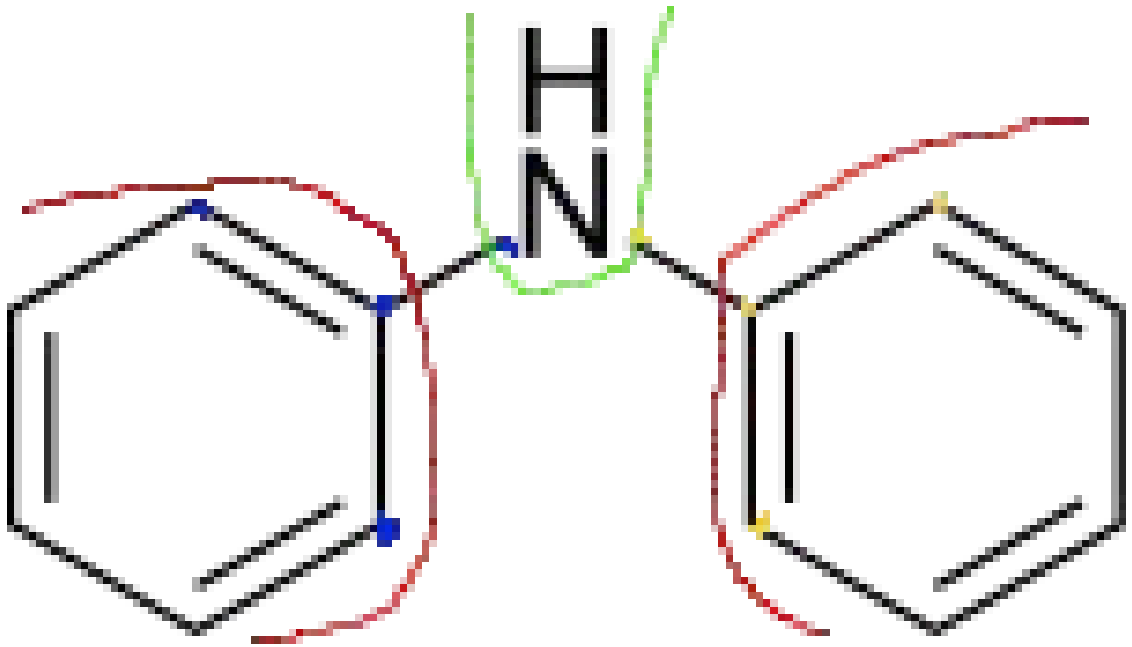
Идеи:

- Будем строить отпечатки для подструктур, а не для атомов по-отдельности.
- Постараемся задать свойства так, чтобы размерность отпечатков для подструктур любых размеров будет одинакова.
- В рамках каждой подструктуры применим BPE
- Далее для получения межподструктурного взаимодействия можем воспользоваться предыдущим подходом.

Создание новых отпечатков

Идеи:

- Нарезать молекулу на подструктуры будем на основе уже имеющейся концепции IUPAC
- Для каждой подструктуры вычисляем 12 свойств, и превращаем каждую подструктуру в массив из 12 элементов.



Обучение GCN/GIN

- Маскируем несколько атомов (10%-15%)
- Учим модель сближать эмбеддинги аугментаций одной и той же молекулы и отдалять аугментации различныхх