**Введение**

Использование опиоидных анальгетиков стало неотъемлемой частью лечения хронической и острой боли. Препараты, такие как морфин и оксикодон, эффективно уменьшают боль, что особенно важно в случаях, когда любые другие обезболивающие не справляются. Однако прием опиодов также сопровождаются серьезными побочными эффектами, включая зависимость и угнетение дыхания. Проблема зависимости от опиоидов стала настоящим общественным кризисом, затрагивающим миллионы людей по всему миру. В условиях растущей обеспокоенности по поводу безопасности традиционных опиоидов возникает необходимость в разработке новых обезболивающих средств, которые могли бы обеспечить эффективное управление болью с минимальным риском для пациентов. Современные технологии, такие как машинное обучение, открывают новые горизонты в области поиска и разработки лекарств. Использование алгоритмов машинного обучения позволяет анализировать огромные объемы данных о химических соединениях и их взаимодействиях с биологическими мишенями. Это не только ускоряет процесс открытия новых анальгетиков, но и помогает предсказывать их эффективность и безопасность еще до начала клинических испытаний. Кроме того, машинное обучение может быть использовано для оптимизации существующих молекул опиоидов, улучшая их терапевтические свойства и снижая риск побочных эффектов. Таким образом, поиск новых опиоидных обезболивающих с использованием моделей машинного обучения не только актуален, но и необходим для обеспечения безопасного и эффективного лечения боли в условиях растущей зависимости от традиционных опиоидов. Это направление исследований может привести к созданию инновационных терапевтических решений, способствующих улучшению качества жизни пациентов и снижению негативных последствий от использования опиоидов. Именно поэтому актуально разрабатывать базы данных эффекторов опиоидных рецепторов.

**Цель:** собрать данные о лигандах опиоидных рецепторов для использования их в качестве данных для обучения регрессионных и/или классификационных моделей.

**Задачи:**

1. Сбор данных

2. Предобработка данных

3. Исследовательский анализ данных

4. Определение и обоснование метрик качества данных

5. Разработка базы данных для хранения данных

6. Оформление пунктов 1, 2 и 5 в отдельный пайплайн для автоматизации

7. Оформление Дашборда

**1. Сбор, предобработка данных и создание бд**

Эти этапы работы представлены в файле OR\_парсинг\_предобработка\_разработка\_бд.py

Использованные версии библиотек и пакетов расположены в файле installed\_packages.txt

Сбор данных о лигандах опиоидных рецепторов представляет собой важную задачу в области биомедицинских исследований и разработки новых лекарственных средств. В этом контексте использование специализированных баз данных, таких как ChEMBL, является более эффективным подходом по сравнению с самостоятельным сбором данных из различных источников. ChEMBL — это тщательно курируемая база данных биоактивных молекул с лекарственными свойствами, которая объединяет химические, биологические и геномные данные.

Преимущества использования ChEMBL:

1. Централизованный доступ к данным: ChEMBL предлагает единый интерфейс для доступа к данным о лигандах, их активности и свойствах. Это значительно упрощает процесс сбора данных, позволяя исследователям сосредоточиться на анализе, а не на поиске информации в разрозненных источниках.

2. Курируемые данные: Все данные в ChEMBL проходят тщательную проверку и стандартизацию, что обеспечивает высокое качество информации. Это особенно важно для специфичных задач, где точность данных играет критическую роль.

3. Анализ данных: На платформе ChEMBL доступны инструменты для анализа данных, позволяющие исследователям проводить статистические и визуализационные исследования непосредственно на собранных данных. Это дает возможность быстро получать результаты и делать выводы на основе достоверной информации.

Опиоидных рецепторов пять, но четыре для которых проведено достаточное количество экспериментов и собрано данных, поэтому загрузим все датасеты, сольем их в один и добавим столбец, в котором будет содержаться информация из какого датасета иначально взяты данные. Для данной задачи эффективнее использовать наработки из готовых БД, чем использовать парсинг данных. Используем бд ChEMBL для поиска данных.

2. Предобработка данных

Удаляем колонки, не несущие информации о молекуле, которая нужна для прогноза структуры и свойств этой молекулы: 'Molecule Name', 'Molecule Max Phase', 'Molecule ChEMBL ID', 'Compound Key', 'Data Validity Comment', 'Comment', 'Uo Units', 'Ligand Efficiency BEI', 'Ligand Efficiency LE', 'Ligand Efficiency LLE','Ligand Efficiency SEI', 'Potential Duplicate', 'Assay ChEMBL ID', 'Assay Description', 'Assay Type', 'BAO Format ID', 'BAO Label', 'Assay Organism', 'Assay Tissue ChEMBL ID', 'Assay Tissue Name', 'Assay Cell Type', 'Assay Subcellular Fraction', 'Assay Parameters', 'Assay Variant Accession', 'Assay Variant Mutation', 'Target ChEMBL ID', 'Target Name', 'Target Organism', 'Target Type', 'Document ChEMBL ID', 'Source ID', 'Source Description', 'Document Journal', 'Document Year', 'Cell ChEMBL ID', 'Properties', 'Action Type', 'Standard Text Value'.

Иногда встречаются поврежденные Smiles представления молекул, которые необходимо удалять, так как для них все равно не получится правильно посчитать дескрипторы. Удаление ошибочных Smiles можно провести одновременно с получением Mol из Smiles, так как для сломанных Smiles или ячеек с Nan Mol не будет получен. Далее можно удалить строки с Nan в колонке с Mol. Таким образом функция отработает быстрее, чем если бы мы проверяли Smiles на валидность. Да и к тому же мы сразу получим еще один необходимый для дальнейших расчетов столбец Mol. Mol – еще один вариант представления химических молекул, реализованный библиотекой rdkit. Данный формат позволяет рассчитать физико-химические дескрипторы для молекулы.

Таргетная величина - pChEMBL Value

Standard Units {'nM', nan} – единицы измерения таргетной величины

Standard Type {'IC50', 'Ki'} – тип таргетной величины, метод измерения активности эффектора в отношении белка

Standard Relation {"'<'", "'='", "'>'", "'>='", nan} – точность оценки таргетной величины

Унифицируем значения таргетной величины:

Удаляем все единицы измерения кроме nM, если бы были значения с другими вариантами М (моль/л), которые было бы возможно перевести в nM – переводим, но это не наш случай, поэтому мы просто удаляем nan.

Оставляем только значения, для которых активность измерена точно (‘=’).

Перевести из IC50 в Ki невозможно, поэтому разделяем на два датасета, один только с Ki, другой с IC50. Далее мы будем работать с бд с Ki, так как в этой бд больше данных.

Так как в бд ChEMBL данные парсились автоматически, то присутствую строки, для которых значения всех столбцов кроме pChEMBL Value являются дубликатами, скорее всего это из-за того, что действительнок значение представлено диапазоном (например: pChEMBL Value 8 – 50 нм), для таких случаев проводим усреднение значения и оставляем одну строку с (например: pChEMBL Value = 29 нм).

Так как для 345 молекул нет экспериментально полученных значений для #RO5 Violations и AlogP, и на данный момент у нас недостаточно данных, описывающих молекулу, то заполнять KNN не столь рационально. Мы можем при помощи библиотеки RDKit получить информацию об AlogP и правилах Липински, поэтому сейчас мы можем удалить эти столбцы из датафрейма df\_ki и заново провести работу с дубликатами.

Затем проверяли на наличие пустых значений. Есть много причин появления пропусков, например, технические - ошибки системы, при передаче, выгрузке данных В данном случае скорее всего не был проведен перерасчет pValue для Value, Но так как все значения Value известны, то мы можем привести перерасчет для pValue.

Итоги подготовки данных:   
----------------------------------------   
Всего строк изначально: 28173   
Всего строк после подготовки: 12727   
Процент строк, которые были удалены: 54.825542185780705 %

Рассчитываем физико-химические дескрипторы RDKit для оставшихся молекул.

Создаем бд при помощи sqlite3.

**2. Анализ данных и Дашборд**

Данный этап работы представлен в файле OR\_Анализ\_данных\_дашборд.ipynb

Был проведен статистический и исследовательский анализ данных.

Используемые статистические тесты: Андерсона-Дарлинга и Крускала-Уоллиса

Выяснили:

- распределение таргетной величины (pChEMBL Value) по тесту Андерсона-Дарлинга не является нормальным

- есть статистическая разница по тесту Крускала-Уоллиса между группами (Target: Mu, Delta, Kappa, Nociceptin)

Вывод: нужно будет учитывать в дальнейшем работе, обучение моделей стоит проводить не на всем датасете, а отдельно для каждой группы по Target.

Исследовательский анализ данных показал какие структурные фрагменты наиболее часто встречаются в базе данных, какие фрагменты наиболее характерны для самых активных и самых неактивных молекул. А также распределение QED характерно для биологически активных молекул. Смещение значений QED смещено в сторону 0.8, что является хорошей тенденцией, так как это показывает смещение распределения молекул в сторону более биодоступных (то есть, способных легче достигнуть необходимого белка в теле человека, за счет растворимости в воде и способности проходить через клеточные мембраны).

Дашборд показывает распределения таргетной величины (pChEMBL Value), QED, наличие характерных фрагментов для молекул.