[слайд 2]

Цель работы – провести сравнительный анализ эффективности различных методов машинного обучения в решении задачи предсказания свойств малых молекул и объяснить полученные результаты.

Задачи работы:

- Обоснование актуальности задачи предсказания свойств малых молекул
- Поиск и выборка различных методов МО для сравнительного анализа
- Проверка эффективности моделей для предсказания свойств малых молекул (например, растворимости) на едином датасете.
- Сравнительный анализ результатов различных моделей.
- Трактовка результатов и подведение итогов.

Суть проделанной работы заключается в

- анализе проблем создания лекарственных препаратов,
- установлении важности решения задачи предсказания свойств малых молекул,
- выборке моделей МО различных архитектур и сравнении их эффективности в решении задач предсказания свойств малых молекул,
- сравнительном анализе и подведении итогов, трактовке и объяснению полученного результата.

1. Постановка задачи

Основные свойства малой молекулы, влияющие на биодоступность и эффективность применения лекарственного препарата. [слайд 3]

[слайд 4]

В рамках практической части настоящей работы были выполнены следующие задачи:

1. Составлена выборка из моделей - представителей популярных архитектур машинного обучения. Главными критериями отбора были:
    a. разнообразность архитектур
    b. актуальность конкретной архитектуры
    c. открытость
2. Выбран датасет, основные критерии:
    a. открытость
    b. универсальность
    c. соответствие тематике
3. Выбраны метрики

4. Составлен механизм дообучения, оценки эффективности модели на выбранном датасете и добавления результатов этой оценки в общую статистику.
5. Применение механизма и анализ результатов

Модели машинного обучения: описание и принцип действия (BERT, GCN, RF)

BERT [слайд 6]

ChemBERTa - это модель, действующая по принципу архитектуры Transformer и адаптированная для обработки химических молекул, представленных в виде строк SMILES (Simplified Molecular Input Line Entry System). Модель основана на архитектуре RoBERTa (Robustly Optimized BERT Pretraining Approach).
После предобучения модель может быть дообучена (fine-tuned) на конкретных задачах предсказания молекулярных свойств — как регрессионных, так и классификационных.

Эта модель была рассмотрена в рамках сравнительного анализа по следующим причинам:

- Модель активно применяется в научных целях
- Способствует архитектурной полноте выборки
- Модель и обученные веса находятся в свободном доступе
- Применяется к молекулам в формате SMILES,
- Может быть дообучена и применена на стандартной CPU-машине

Модель ChemBERTa представляет собой современное, доступное и архитектурно уникальное решение для предсказания свойств малых молекул.

GCN [слайд 7]

GCNPredictor — это модель графовой нейронной сети (GNN), реализованная в библиотеке GDL-LifeSci, специально разработанной для биоинформатики и хемоинформатики. Она основана на **Graph Convolutional Network (GCN)** [18].
GCNPredictor моделирует молекулы как **неориентированные графы**, где узлы (nodes) соответствуют атомам, а рёбра (edges) — химическим связям. Модель обучается предсказывать молекулярные свойства, обобщая информацию, проходящую через структуру молекулярного графа.
Код и примеры обучения доступны на GitHub. Не требует особых вычислительных ресурсов - обучение возможно на CPU при работе с небольшими датасетами.

Обучение и воспроизведение результатов может быть проведено с использованием общедоступных молекулярных баз.

RF [слайд 8]

Random Forest не требует обучения на GPU или большого объема оперативной памяти, имеет небольшое число гиперпараметров и предоставляет удобные средства анализа важности признаков. Это полезно при анализе влияния различных молекулярных дескрипторов на результат. Для оценки эффективности и проведения сравнительного анализа будет использоваться реализация Random Forest из библиотеки sklearn.

Датасет [слайд 8]

Датасет ESOL (Estimated Solubility) представляет собой широко используемую открыто доступную коллекцию экспериментальных данных, отражающих водную растворимость малых органических молекул. Впервые упомянут в работе «ESOL: Estimating Aqueous Solubility Directly from Molecular Structure» J.S.Delaney в 2004 г.

Основной предсказываемой характеристикой в данном датасете является логарифм растворимости молекул в воде (logS), выраженный в логарифмической шкале концентрации (молярность). Значения logS были получены на основе экспериментальных измерений и охватывают широкий диапазон растворимостей, что обеспечивает репрезентативность выборки.

В датасет входит 1128 малых молекул, каждая из которых представлена в формате SMILES. Такой формат делает датасет универсальным: он может быть использован с моделями любых архитектур — от случайных лесов до BERT-подобных моделей и графовых нейросетей.

Дополнительно к строкам SMILES предоставляются вычисленные молекулярные дескрипторы, включая молекулярную массу, число водородных доноров и акцепторов, площадь полярной поверхности (TPSA) и другие параметры.

Таким образом, ESOL отвечает ряду ключевых требований:

- открытый,
- актуальный.
- универсальный
- не требует объемных ресурсов

Предобработка датасета

[слайд 9]

Перед применением датасет был разбит на обучающую и тестовую выборки в отношении 4/1. Каждая модель была дообучена на обучающей выборке и

оценена на тестовой. Затем оценки всех моделей были объединены в общую статистику [слайд 10]

Результаты и их трактовка, сравнительный анализ

Модель ChemBERTa, основанная на архитектуре трансформера RoBERTa, продемонстрировала высокое качество предсказаний с RMSE = 0.63 и R^2 = 0.78. Эти показатели отражают её способность точно количественно оценивать логарифм растворимости и объяснять значительную часть вариабельности целевого признака. Архитектурные преимущества ChemBERTa связаны с механизмом self-attention, который позволяет эффективно моделировать сложные и дальнодействующие взаимосвязи между атомами и функциональными группами, представленными в виде последовательностей SMILES. Предобучение на больших химических корпусах способствует устойчивости модели к различиям в структуре молекул.

Графовая сверточная сеть (GCN), с результатами RMSE ≈ 1.04 и R^2 ≈ 0.77, демонстрирует хорошие показатели, однако с определёнными ограничениями. Модель эффективно кодирует локальную топологию молекулы за счёт агрегации признаков соседних вершин графа, что позволяет захватывать локальные химические взаимодействия. Тем не менее, ограниченная глубина сети и эффект переусреднения ограничивают способность учитывать глобальные и более отдалённые зависимости, которые важны для точного описания молекулярных свойств.

Модель Random Forest достигла RMSE около 0.59 и R^2 около 0.62, показывая удовлетворительную точность предсказаний, однако с более низкой объяснительной способностью по сравнению с глубокими архитектурами. Этот ансамблевый метод, оперирующий табличными признаками молекул, успешно выявляет нелинейные зависимости и устойчив к переобучению, что делает его простым и надёжным инструментом для предварительного анализа данных. Тем не менее, отсутствие встроенного механизма обработки молекулярной топологии и пространственных отношений атомов ограничивает глубину химического анализа, что сказывается на способности модели к комплексному учёту химических особенностей.