**[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.

В процессе работы возникли трудности, связанные с недостатком ресурсов, поскольку некоторые модели, реально использующиеся для решения этой задачи либо не находятся в открытом доступе, либо требуют внушительных вычислительных ресурсов для функционирования. В целях получения наиболее полной и реалистичной статистики для сравнительного анализа выборка экспериментальных моделей была составлена из открытых аналогов моделей, реально использующихся для решения задачи предсказания свойств малых молекул, и реализаций классических методов МО, не предназначенных *специально* для решения указанной задачи.

[слайд 3]

1. **Постановка задачи**

Малая молекула – химическое соединение со сравнительно малой молекулярной массой (не более 900-1000 дальтон), обладающее той или иной биологической активностью и размером порядка 1 нм. Определяющим свойством малых молекул является её способность проникать в клетку и покидать её без необходимости взаимодействия с другими молекулами. Научный и практический интерес к малым молекулам и их свойствам обусловлен тем фактом, что большинство лекарств являются малыми молекулами.

Основные свойства [слайд 4]

*Молекулярная масса (Molecular Weight, MW)* – меньшая масса способствует увеличению биодоступности.

*Растворимость в воде (aqua solubility)* – важна для абсорбации и доставки в кровоток.

*Липофильность (Lipophilicity, жирорастворимость, гидрофобность)* – влияет на биодоступность и растворимость. Измеряется как логарифм коэффициента распределения между октанолом и водой.

*Мембранная проницаемость (permeability) –* характеризует способность молекулы проходить через биологические барьеры.

*ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity, абсорбация, распределение, метаболизм, выведение и токсичность)* – комплекс критериев, описывающих положение соединения в организме.

*Полярность (PSA)* – число водородных связей. Высокое значение этого параметра может снизить биодоступность.

*Избирательность (selectivity)* – параметр, определяющий активность взаимодействия молекулы с другими соединениями, помимо мишени. Влияет на эффективность и наличие побочных эффектов

[слайд 5]

В рамках практической части настоящей работы были выполнены следующие задачи:

1. Составлена выборка из моделей - представителей популярных архитектур машинного обучения. Главными критериями отбора были:
   1. разнообразность архитектур
   2. актуальность конкретной архитектуры
   3. открытость
2. Выбран датасет, основные критерии:
   1. открытость
   2. универсальность
   3. 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).

Модель ChemBERTa использует энкодерную часть трансформера, подобно оригинальному BERT, и обучается по задаче Masked Language Modeling (MLM). Основные компоненты модели включают:

* Tokenizer: модифицированный byte-level BPE-токенизатор, специально адаптированный для химической нотации SMILES.
* Энкодер: несколько слоёв самовнимания (multi-head self-attention), реализованных в рамках архитектуры RoBERTa.

Целевая функция обучения ChemBERTa формулируется как:

где - множество маскированных токенов

- истинный токен

- последовательность с маскированным токеном

– вероятность восстановить маскированный токен по контексту

Модель обучается на крупной выборке молекул из базы ZINC15, содержащей около 10 миллионов молекул. Во время обучения случайные токены маскируются, и задача модели — предсказать их на основе контекста.

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, продемонстрировала конкурентоспособность на задачах регрессии и классификации свойств малых молекул
* Способствует архитектурной полноте выборки и вносит требуемое разнообразие
* Модель и обученные веса находятся в свободном доступе на Hugging Face.
* Применяется к молекулам в формате SMILES, что исключает необходимость в 3D-геометрии или квантово-химических расчётах.
* Может быть дообучена и применена на стандартной CPU-машине, что критически важно в условиях ограниченных вычислительных ресурсов.

Модель ChemBERTa представляет собой современное, доступное и архитектурно уникальное решение для предсказания свойств малых молекул. Её включение в выборку моделей для сравнительного анализа усиливает научную объективность исследования за счёт архитектурной разнородности, повышает прикладную значимость за счёт актуальности подхода в научной и промышленной практике и обеспечивает воспроизводимость экспериментов в условиях ограниченных ресурсов.

**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.

Основные компоненты GCNPredictor:

1. Graph Convolutional Layers: множественные слои GCN осуществляют агрегацию информации о соседях каждого узла:

(формула на слайде

где - вектор признаков узла i на слое l,

- множество соседей узла i,

– степень узла i,

- обучаемая матрица весов слоя l,

– нелинейная функция активации (например, ReLU

1. Readout-функция. После свёрток производится агрегация узловых векторов в вектор молекулы (глобальное представление графа), с помощью mean/sum/max pooling или learnable attention pooling.
2. MLP head. Глобальный вектор подаётся на полносвязную сеть (MLP), которая и производит финальное предсказание.

GCNPredictor основан на **графовой нейросетевой архитектуре**, принципиально отличной от трансформеров (как ChemBERTa) и моделей на основе SMILES (например, RNN или CNN). Это делает модель ценной для сравнения с иными подходами, так как она опирается на **структурное представление молекулы.**

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.

Таким образом включение GCNPredictor в выборку моделей усиливает архитектурное разнообразие, обеспечивает реалистичное сравнение с GNN-классом моделей, соответствует современным научным и прикладным практикам в предсказании свойств малых молекул и при этом не нарушает критерии доступности и воспроизводимости, что делает эту модель оптимальным кандидатом.

**RF [слайд 8]**

Random Forest (RF) — ансамблевый алгоритм машинного обучения, представляющий собой комбинацию решающих деревьев, обученных на случайных подмножествах признаков и объектов.

Эта модель активно применяется для предсказания ADMET-свойств, токсичности, биоактивности молекул и других свойств малых молекул, применима к дескрипторам молекул, которые можно извлекать без графовой структуры, что даёт возможность проводить аналитически интерпретируемые эксперименты.

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.

RF часто используется как базовая модель, с которой сравниваются более сложные архитектуры. Включение её в исследование обеспечивает точку отсчета и позволяет оценить, насколько архитектуры типа GNN или трансформеров действительно превосходят классические методы.

Также использование RF позволяет выполнять воспроизводимые эксперименты, что особенно важно в рамках практической части настоящей работы.

Для оценки эффективности и проведения сравнительного анализа будет использоваться реализация Random Forest из библиотеки sklearn.

**Датасет [слайд 9]**

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:

- Он входит в состав MoleculeNet и поддерживается библиотеками вроде DeepChem),

- может быть использован с различными типами моделей) и представляет актуальное физико-химическое свойство, напрямую влияющее на биодоступность молекул.

- благодаря умеренному объёму данных датасет также может использоваться в условиях ограниченных вычислительных ресурсов, что делает его особенно удобным для практических экспериментов и сравнительных исследований.

Таким образом выбор датасета ESOL полностью соответствует методологическим требованиям настоящего исследования, гарантирует архитектурную независимость применения, предоставляет реальные и значимые молекулярные свойства, обладает открытым форматом и обеспечивает воспроизводимость экспериментов и сравнимость с результатами предыдущих экспериментов.

Растворимость молекул (в логарифмической шкале — logS) была собрана автором из экспериментальных публикаций, химических справочников и научной литературы и экспериментальные базы данных в фармацевтической и агрохимической промышленности.

Молекулы, содержащие нестабильные фрагменты, сильно ионизированные формы или редкие атомы, были исключены.

Все молекулы были приведены к формату SMILES для универсального описания структуры.

Таким образом в настоящей работе для сравнительного анализа эффективности различных подходов машинного обучения будет использоваться современный датасет, отвечающий тематике исследования и подходящий для использования моделями любой архитектуры в условиях ограниченных вычислительных ресурсов.

**Результаты и их трактовка, сравнительный анализ**

Модель ChemBERTa, основанная на архитектуре трансформера RoBERTa, продемонстрировала высокое качество предсказаний с RMSE = 0.63 и R² = 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 R² ≈ 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.

GCN сохраняет значительную эффективность, учитывая структурный характер входных данных и относительно низкие требования к предварительной обработке.

Модель Random Forest достигла RMSE около 0.59 и R² 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 меньшем значении R² и, соответственно, на способности модели к комплексному учёту химических особенностей.

**Итоги**

Таким образом, ChemBERTa, благодаря своей трансформерной архитектуре и предобучению на масштабных химических корпусах, обеспечивает более глубокое и контекстно обоснованное представление молекулярных данных, что отражается в наилучших метриках качества. GCN эффективно использует структурную информацию молекул, однако его локальные механизмы агрегации и ограничения глубины снижают точность по сравнению с ChemBERTa. Random Forest служит полезным базовым методом с хорошей устойчивостью и простотой применения, но уступает глубоким моделям в способности моделировать сложные химические зависимости.

Сопоставление архитектур показывает, что наиболее перспективным направлением для задач предсказания молекулярных свойств являются глубокие модели, оперирующие с оригинальными химическими представлениями без необходимости ручной инженерии признаков. Среди них ChemBERTa обладает наилучшей способностью к обобщению за счёт self-attention механизма и обширного предобучения. GCN представляет собой мощный инструмент, эффективно использующий графовую природу молекул, но требующий дополнительных улучшений, таких как введение глобального внимания (например, GAT или GINE). Random Forest, несмотря на ограниченность в выразительности, остаётся надёжным и интерпретируемым базовым методом, особенно полезным в условиях ограниченных вычислительных ресурсов или для анализа важности признаков.