МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
КАФЕДРА БИОМЕДИЦИНСКОЙ ИНФОРМАТИКИ

Исследование архитектур нейронных сетей для предсказания свойств лекарственно-подобных молекул

Курсовая работа

Благодарного Артёма Андреевича
обучающегося 4 курса
специальности «Информатика»

Научный руководитель:
профессор, доктор
физико-математических наук,
Тузиков А.В.

Минск, 2025

Содержание

Введение

Разработка новых лекарственных препаратов (Drug Discovery) представляет собой длительный, дорогостоящий и высокорискованный процесс, в котором основная доля экспериментальных затрат сосредоточена на этапах доклинических и клинических исследований. Одной из ключевых причин неудач на поздних фазах клинических испытаний является несоответствие фармакокинетическим характеристикам и профилю безопасности, объединяемым термином ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity — всасывание, распределение, метаболизм, выведение и токсичность) [1]. Ошибки в оценке этих свойств на ранних этапах приводят к значительным потерям бюджета и времени, что делает задачу точного in silico-прогнозирования ADMET-параметров критически важной для современной хемоинформатики и фармакологии.

За последние годы область компьютерного моделирования химических структур переживает кардинальный сдвиг благодаря стремительному распространению методов глубокого обучения. Архитектуры, изначально разработанные для обработки естественного языка — в частности Transformers — оказались чрезвычайно эффективными при работе с линейными представлениями молекул, такими как строки SMILES [2]. На этой волне сформировался новый класс крупных предварительно обученных моделей — molecular foundation models, заимствующих идеи языковых моделей NLP и адаптирующих их к химическому пространству. Отдельное внимание в современной литературе привлекают модели на основе State Space Models (SSM), к которым относится архитектура Mamba [1]. В отличие от традиционных трансформеров, Mamba обеспечивает линейную вычислительную сложность по длине последовательности, эффективно моделирует длинный контекст и демонстрирует превосходство в ряде бенчмарков, что делает её перспективной для работы с вариативными и часто длинными SMILES-строками, где критично учитывать дальние зависимости.

Тем не менее практическое применение таких архитектур сопряжено с существенными ограничениями. Обучение и инференс современных state-of-the-art моделей требуют высокопроизводительных GPU, поддержки CUDA и значительных объёмов видеопамяти, что ограничивает доступность исследований в условиях скудных вычислительных ресурсов [3]. В связи с этим возникает научно-прикладная задача — понять, насколько оправдан переход к тяжёлым архитектурам в задачах ADMET-прогнозирования и существует ли разумный компромисс между вычислительной сложностью и качеством предсказаний.

Целью настоящей работы является проведение сравнительного анализа эффективности различных нейронных архитектур в задачах прогнозирования

ADMET-свойств молекул. Исследование направлено на сопоставление CPU-ориентированных моделей с результатами, представленными в статье «SMILES-Mamba: A Mamba-based Molecular Foundation Model», а также на оценку применимости архитектуры Mamba в условиях ограниченных вычислительных ресурсов [1]. Важной составляющей является анализ trade-off между вычислительными затратами и качеством предсказания, что позволяет сформулировать практические рекомендации по выбору архитектуры для прикладных задач.

Для достижения поставленной цели в работе решаются следующие взаимосвязанные задачи. Во-первых, проводится теоретический анализ принципов работы моделей пространства состояний (SSM), в частности архитектуры Mamba, и выявляются их ключевые отличия от графовых нейронных сетей (GNN) и трансформеров; особое внимание уделяется методологии и архитектурным решениям, описанным в исходной статье. Во-вторых, реализуется программная часть: разрабатывается и обучается набор нейросетевых моделей, адаптированных для работы на CPU, с применением датасетов из бенчмарка Therapeutics Data Commons (TDC) для задач как классификации, так и регрессии [2]. В-третьих, проводится экспериментальная оценка качества обученных моделей с использованием стандартных метрик (ROCAUC, PR-AUC, RMSE и др.) и анализируется влияние выбора архитектуры на точность предсказания различных классов ADMET-свойств. В-четвёртых, выполняется сравнительный анализ полученных результатов с опубликованными метриками SMILES-Mamba с целью оценки величины отставания и определения классов задач, в которых сложные архитектуры дают значимый прирост качества или, напротив, оказываются избыточными. Наконец, осуществляется оценка эффективности в терминах соотношения вычислительных затрат и качества прогноза, что позволяет сформулировать рекомендации для практического применения при ограниченных вычислительных возможностях.

Объектом исследования являются методы машинного обучения для анализа химических структур, а предметом исследования — влияние выбора архитектуры нейронной сети и доступных вычислительных ресурсов на точность прогнозирования ADMET-свойств молекул. Работа сочетает теоретический разбор современных архитектур и практическую реализацию моделей с целью получения воспроизводимых выводов, полезных как для академических исследований, так и для практических задач доклинической оценки лекарственных кандидатов.

Глава 1. Обзор методов глубокого обучения в хемоинформатике

Прогнозирование физико-химических и биологических свойств молекул (QSAR/QSPR — Quantitative Structure–Activity/Property Relationship) является фундаментальной задачей современной хемоинформатики и играет ключевую роль в ранних этапах разработки лекарственных препаратов. На протяжении десятилетий для решения этих задач применялись ручные молекулярные дескрипторы — атомные параметры, топологические индексы, фингерпринты и другие фиксированные представления структуры. Эти признаки передавались на вход классическим алгоритмам машинного обучения, таким как SVM, Random Forest, kNN [3]. Однако такие методы ограничены качеством вручную сконструированных дескрипторов, что препятствует их масштабируемости и универсальности. Появление глубокого обучения привело к переходу от ручного извлечения признаков к автоматическому, что радикально повысило возможности моделирования молекулярных свойств.

Одним из ключевых аспектов в современной хемоинформатике является выбор представления молекулы, поскольку именно оно определяет, какую информацию получит модель на вход и насколько эффективно сможет её использовать. В зависимости от задач и доступных ресурсов молекулы описываются в виде строк, графов или трёхмерных координат, и каждая из этих парадигм имеет свои преимущества, ограничения и специфические архитектуры. В данной главе приводится систематизированный обзор наиболее значимых подходов к представлению молекул, а также описываются соответствующие архитектуры глубокого обучения — от ранних рекуррентных моделей до трансформеров и современных моделей пространства состояний (State Space Models, SSM), к которым относится архитектура Mamba [5].

1.1. Специфика данных в хемоинформатике

Несмотря на то, что молекула является сложным квантово-механическим объектом с пространственной геометрией, зарядовым распределением и динамическими характеристиками, в задачах QSAR/QSPR требуется привести её к формату, удобному для алгоритмов машинного обучения [7]. Такой формат должен быть одновременно достаточно информативным, чтобы сохранять химическую структуру, и достаточно компактным, чтобы позволять обработку больших датасетов и обучение нейронных сетей в приемлемых вычислительных условиях. В современной хемоинформатике доминируют три класса репрезентаций: строковые (1D), графовые (2D) и трёхмерные (3D). Каждый

подход формирует собственные требования к архитектуре модели и определяет спектр решаемых задач.

Строковые представления (1D)

Наиболее распространённой строковой формой является формат SMILES (Simplified Molecular Input Line Entry System) [4]. Он задаёт молекулу в виде линейной последовательности символов, описывающих атомы, связи, кольца и ветвления. Такая интерпретация структуры позволяет рассматривать молекулу как последовательность токенов и применять методы, унаследованные из обработки естественного языка. Благодаря этому SMILES оказались особенно удобны для использования в рекуррентных сетях, трансформерах и моделях пространства состояний.

Преимуществами строковых представлений являются их близость к парадигме NLP, доступность огромных корпусов неразмеченных данных (таких как ZINC и PubChem), что делает возможным обучение моделей в self-supervised режимах, а также высокая вычислительная эффективность, позволяющая работать с ними даже на CPU [8]. Однако SMILES обладают важными ограничениями: отсутствие единственности представления, зависимость от выбранной канонизации, нарушение топологической близости атомов в линейной записи и высокая хрупкость к синтаксическим ошибкам. Эти недостатки осложняют работу генеративных моделей и могут снижать интерпретируемость. Тем не менее строковый формат остаётся одним из наиболее практичных компромиссов между информативностью, доступностью данных и вычислительной эффективностью.

Графовые представления (2D)

Графовая модель молекулы соответствует её естественной химической структуре: атомы формируют множество вершин, а связи — множество рёбер. Каждая вершина и каждое ребро имеют свои параметры, отражающие их химические свойства. Эта парадигма привела к широкому распространению графовых нейронных сетей, включая GCN, GAT, MPNN, GINE, DMPNN и другие, созданные специально для моделирования молекулярных структур [8].

Ключевым преимуществом графового представления является сохранение топологии молекулы и инвариантность к перестановкам атомов. Благодаря этому сообщение между узлами в графовых нейронных сетях соответствует реальным химическим взаимодействиям, а интерпретируемость таких моделей значительно выше по сравнению со строковыми. Тем не менее графовые модели сталкиваются с рядом фундаментальных ограничений: эффект пересглаживания при увеличении числа слоёв, локальность message passing, затрудняющая моделирование дальнодействующих эффектов, а также высокие требования к памяти, что делает обучение моделей сложным при ограниченных ресурсах [8].

Трёхмерные представления (3D)

Трёхмерные подходы опираются на реальные пространственные координаты атомов и позволяют учитывать конформации молекул, что критически важно в задачах, связанных с докингом, оценкой аффинности или взаимодействием «лиганд–белок». Такие модели наиболее информативны, поскольку включают геометрию, стереохимию, расстояния и углы. Однако их использование связано с серьёзными вычислительными затратами. Получение конформаций часто требует выполнения методов, основанных на DFT или силовых полях, что ограничивает масштабирование и делает такие подходы непрактичными для первичных скринингов, особенно в задачах ADMET, где требуется обработка миллионов структур. Кроме того, выбор конформации может влиять на результат, что добавляет неопределённость при моделировании гибких молекул [6].

Несмотря на значительные успехи графовых и трёхмерных моделей, строковые представления SMILES в сочетании с современными sequence-архитектурами остаются одним из наиболее эффективных и универсальных подходов в хемоинформатике. Они позволяют использовать мощную инфраструктуру NLP, обеспечивают доступ к огромным корпусам данных для предварительного обучения и допускают обучение моделей даже в условиях ограниченных вычислительных ресурсов. Именно поэтому большинство современных foundation-моделей, включая SMILES-Mamba, опираются на строковые представления. В рамках дальнейших разделов рассматриваются архитектуры глубокого обучения, применяемые для обработки таких последовательностей, а также их преимущества и ограничения в задачах прогнозирования свойств молекул.

1.2. Эволюция архитектур для работы с последовательностями

История применения глубокого обучения к строковым представлениям молекул практически полностью повторяет эволюцию NLP-моделей. В разные периоды доминировали различные архитектурные парадигмы: рекуррентные сети, трансформеры, а сегодня — модели пространства состояний (State Space Models, SSM), к которым относится архитектура Mamba [1]. Каждый этап развития отражает попытку найти оптимальный баланс между точностью, вычислительной эффективностью и способностью моделировать длинные зависимости в последовательности.

Рекуррентные нейронные сети (RNN, LSTM, GRU)

Первые работы по анализу и генерации SMILES использовали рекуррентные нейронные сети, которые обрабатывают последовательность последовательно: на шаге t модель принимает токен t и обновляет скрытое состояние h, выступающее в

роли памяти о предыдущем контексте. Модификации RNN — LSTM и GRU — частично устраняют проблему забывания благодаря механизмам ворот, регулирующих поток информации. Применение RNN к SMILES было естественным, поскольку: SMILES представляют собой линейную строку, аналогичного природы предложения в естественном языке; многие задачи (генерация молекул, предсказание следующего токена) формализуются как sequence modeling; RNN устойчивы к переменной длине последовательностей [7].

Ограничения рекуррентных моделей

Несмотря на очевидную применимость RNN, их фундаментальные ограничения стали критичными в хемоинформатике: отсутствие параллелизма; последовательная обработка делает обучение медленным и плохо масштабируемым; для self-supervised обучения на десятках миллионов SMILES такая архитектура становится узким местом; затухание и взрыв градиента. Даже LSTM плохо удерживают информацию на дистанции >100 токенов. Для длинных SMILES (макроциклы, полимеры, разветвленные ароматические системы) это приводит к потере важного контекста и снижению качества предсказаний. Локальность информации: хотя теоретически RNN способны моделировать глобальный контекст, на практике они склонны «залипать» на ближайших токенах, плохо улавливая удалённые химические зависимости. Эти ограничения привели к постепенному уходу от RNN в пользу архитектуры Transformer.

Трансформеры и механизм Self-Attention

Появление Transformer (Vaswani et al., 2017) стало революционным прорывом, радикально изменившим подходы к моделированию последовательностей [3]. Основой архитектуры является механизм самовнимания (Self-Attention), который позволяет учитывать взаимодействие между любыми токенами вне зависимости от их позиции.

В контексте SMILES это означает: модель видит всю молекулу целиком; атомы, находящиеся в начале и конце строки, могут эффективно взаимодействовать; замыкание колец, ветвления и стереохимические элементы моделируются значительно лучше, чем в RNN. На основе Transformer появилось множество специализированных моделей для химии: ChemBERTa, MoLFormer, MegaMolBART, SMILES-BERT, ChemGPT, и другие модели, основанные на self-supervised обучении на миллионах молекул.

Эти модели существенно улучшили качество: предсказания ADMET-свойств, токсичности, растворимости, биодоступности; генерации новых молекул; оптимизации структуры. Преимущества трансформеров для SMILES: глобальный контекст; в self-attention каждый токен видит всю последовательность; полный

параллелизм; все токены обрабатываются одновременно — это резко ускоряет обучение; идеальная совместимость с огромными корпусами. Transformer масштабируются вместе с данными и широко используются для pretraining на десятках миллионов SMILES.

Фундаментальные ограничения трансформеров. Несмотря на успех, Transformer имеет ключевой недостаток: квадратичную вычислительную сложность внимания: $O(N^2)$, где N — длина SMILES. Последствия: высокая стоимость обучения на длинных молекулах; большие требования к GPU-памяти и пропускной способности; ухудшение масштабируемости в задачах массового виртуального скрининга, где необходимо обработать миллионы молекул.

Поскольку индустрия движется к foundation-моделям и обучению на сотнях миллионов примеров, трансформеры становятся узким местом. Это стало основной мотивацией для поиска архитектур, способных: моделировать длинные зависимости, работать параллельно, иметь линейную или около-линейную сложность. Таким классом моделей стали State Space Models (SSM).

1.3. Архитектура Mamba и модели пространства состояний

Появление моделей пространства состояний стало следующим этапом развития архитектур для анализа длинных последовательностей. В отличие от трансформеров, в которых ключевую роль играет механизм самовнимания с квадратичной вычислительной сложностью, SSM рассматривают последовательность как динамическую систему, эволюирующую во времени. Такое представление позволяет эффективно моделировать зависимости между элементами последовательности и выполнять вычисления с линейной сложностью по длине входа. Наиболее значимым достижением в этой области является архитектура Mamba, которая объединяет достоинства рекуррентных сетей и SSM, но по качеству приближается к трансформерам [2].

Классические модели пространства состояний описывают скрытое состояние последовательности с помощью линейной системы, которую затем дискретизируют и используют как рекуррентное обновление. Преимущество таких моделей заключается в том, что вычисления можно выполнять как свёртку, а значит, они эффективны и легко распараллеливаются. Однако первые поколения SSM имели жёсткие ограничения: их параметры динамики оставались фиксированными и плохо адаптировались к разнообразным структурам последовательностей. Это снижало качество моделирования сложных данных, таких как представления молекул в формате SMILES.

Архитектура Mamba решает эти проблемы и вводит два ключевых новшества. Первое — линейная вычислительная сложность. В отличие от трансформеров, где

стоимость обработки растёт квадратично с длиной входа, Mamba использует оптимизированные операции, масштабирующиеся линейно. Это позволяет модели эффективно работать с последовательностями длиной в десятки тысяч токенов. Для химических данных это особенно важно, поскольку SMILES могут быть как короткими, так и очень длинными, например при описании макроциклов или полимеров [1].

Второе новшество — механизм селективного сканирования. В Mamba параметры внутренней динамики зависят от входных данных, то есть модель «решает», какую информацию сохранить, а какую отбросить. Такой механизм делает модель нелинейной и гибкой. Он играет ключевую роль в задачах хемоинформатики, поскольку разные фрагменты молекулы имеют разную значимость. Например, функциональные группы, гетероатомы, ароматические кольца или заряды могут располагаться далеко друг от друга в строковом представлении, но быть важными для определения свойств молекулы. Селективная динамика позволяет Mamba учитывать эти зависимости без затрат, характерных для внимания [2].

Адаптация архитектуры Mamba к SMILES представлена в работе SMILES-Mamba. Модель обучают в два этапа: предварительное обучение на больших корпусах химических структур и последующее дообучение на специализированных задачах. На первом этапе модель учится химической грамматике через предсказание следующего токена. Такой подход оказывается эффективным, поскольку требует от модели понимания валентностей, структуры кольцевых систем, ароматичности и других химических правил, заложенных в формате SMILES. На втором этапе модель дообучают на задачах из набора Therapeutics Data Commons, включающего в себя прогнозирование токсичности, биодоступности, растворимости и других свойств. В этих задачах Mamba демонстрирует результаты, сравнимые или превосходящие трансформеры и графовые нейронные сети.

Несмотря на вычислительную эффективность, Mamba сильно опирается на оптимизации, реализованные на GPU. На графических ускорителях модель показывает высокую скорость благодаря специализированным CUDA-ядрам и оптимизации через Triton. Однако на CPU производительность резко снижается, что усложняет повторение экспериментов в условиях ограниченных вычислительных ресурсов. Этот фактор является важным ограничением и мотивирует исследование применимости Mamba в реальных условиях, где доступность мощных GPU может быть ограничена.

Таким образом, архитектура Mamba представляет собой гибридный подход, сочетающий преимущества рекуррентных моделей и трансформеров. Она

эффективна на длинных последовательностях, способна учитывать сложные нелинейные химические зависимости и экономит память по сравнению с моделями внимания. Всё это делает Mamba одним из наиболее перспективных направлений в анализе SMILES и предсказании свойств молекул.

Глава 2. Методология и постановка эксперимента

Данная глава описывает экспериментальный дизайн, наборы данных, используемые методы предварительной обработки, а также архитектуры моделей, реализованных и обученных в рамках исследования. Особое внимание уделено ограничению вычислительных ресурсов, поскольку все эксперименты проводились на центральном процессоре без использования графических ускорителей. Это определило как выбор моделей, так и характер постановки экспериментов.

2.1. Описание используемых данных

Основой экспериментальной части исследования стали данные из открытой библиотеки Therapeutics Data Commons. Этот ресурс представляет собой обширный набор биомедицинских и фармакологических задач, широко используемый для оценки алгоритмов машинного обучения в области разработки лекарственных препаратов. В рамках работы были выбраны несколько задач по предсказанию молекулярных свойств, включая показатели токсичности, проницаемости, растворимости и взаимодействий с биологическими мишенями. Конкретные свойства, использованные в исследовании, относятся к числу стандартных ADMET-характеристик, таких как проницаемость Caco-2, вероятность блокирования ионного канала hERG, водная растворимость, а также параметры метаболизма. Эти задачи являются репрезентативными и позволяют сравнивать модели по широкому спектру химически значимых предсказаний.

Перед обучением модели данные прошли несколько этапов предварительной обработки. Каждая молекула была представлена в виде строки SMILES, которая затем токенизировалась с использованием простой символьной токенизации. Такой подход сохраняет структуру строки и не требует внешних словарей, что делает его устойчивым к новым химическим фрагментам. Для оценки моделей данные были разделены на обучающую, валидационную и тестовую выборки. Использовались два различных режима разбиения. Первый — случайное разбиение, при котором молекулы распределяются равномерно и случайным образом. Второй — более строгий scaffold split, разделяющий молекулы по их химическому «каркасу». Второй метод позволяет проверить, насколько хорошо модель способна обобщать знания на ранее невидимые химические структуры, что важно для задач поиска новых лекарственных молекул. Такой подход даёт более реалистичную оценку качества модели в условиях открытого химического пространства.

2.2. Бенчмарки для оценки моделей

Оценка качества моделей для предсказания ADMET-свойств требует использования репрезентативных, стандартизированных и широко признаваемых бенчмарков. ADMET-задачи охватывают широкий спектр биофармацевтических свойств — от абсорбции и распределения до метаболизма, токсичности и выведения, — поэтому выбор корректных тестовых наборов данных является критически важным для достоверной валидации моделей. В рамках данного проекта применяются открытые бенчмарки, сформированные из общедоступных биохимических и фармакокинетических источников, которые традиционно используются в исследованиях по компьютерному дизайну лекарств.

Используемые датасеты охватывают как регрессионные задачи (например, прогноз растворимости, липофильности или клиренса), так и задачи бинарной классификации (например, предсказание ингибирования фермента CYP или токсичности соединений). Такой широкий набор позволяет обеспечить комплексную оценку моделей и проверить их способность обобщать знания на различные ADMET-концепции. Данные наборы отличаются по масштабу: от относительно небольших датасетов на несколько сотен соединений до крупных коллекций, включающих десятки тысяч молекул. Это позволяет анализировать устойчивость моделей к разным условиям, включая малые выборки и высокоразмерные пространства [5].

Кроме того, рассматриваемые бенчмарки включают как экспериментально подтверждённые значения, полученные из биологических и фармакологических исследований, так и интегрированные данные, обработанные и унифицированные для моделирования. Использование таких датасетов гарантирует, что сравнение моделей будет объективным, воспроизводимым и сопоставимым с результатами существующих исследований, включая современные архитектуры, такие как Mamba-модели для молекулярных представлений.

В последующих подразделах представлено детальное описание каждого используемого бенчмарка, включая размер, тип задачи, биологическую интерпретацию и роль в общей системе оценки качества предсказаний. Такой подход обеспечивает целостное понимание применяемых данных и создаёт основу для корректной интерпретации результатов, полученных при обучении и тестировании моделей, представленных в данной работе.

2.2.1. Бенчмарки для оценки абсорбции

Абсорбционные свойства описывают способность молекулы проникать через биологические мембраны. Эти свойства являются критически важными на ранних

этапах разработки лекарств, поскольку напрямую связаны с биодоступностью препарата.

Caco-2 permeability (Caco2)

Размер: **906 молекул**. Задача: регрессия. Метрика: MAE.

Датасет моделирует проницаемость через клеточную линию Caco-2, которая является стандартом индустрии для оценки оральной абсорбции препаратов. Значения получены экспериментально, что делает задачу трудной и чувствительной к шуму.

Human Intestinal Absorption (HIA)

Размер: **578 молекул**. Задача: бинарная классификация. Метрика: ROC-AUC.

Классифицирует молекулы как хорошо или плохо абсорбирующиеся в тонком кишечнике. Датасет небольшой и шумный, что обычно приводит к сильной вариативности результатов.

P-glycoprotein substrate/inhibitor (Pgp)

Размер: **1,212 молекул**. Задача: классификация. Метрика: ROC-AUC.

P-gp — транспортёрный белок, активно выталкивающий молекулы из клетки наружу. Изучение взаимодействия с P-gp важно для оценки устойчивости лекарств к активному выведению.

Bioavailability (Bioav)

Размер: **640 молекул**. Задача: классификация. Метрика: ROC-AUC.

Оценивает способность препарата достигать системного кровотока при пероральном приёме. Датасет небольшой, поэтому модели склонны к переобучению.

Lipophilicity (Lipo)

Размер: **4,200 молекул**. Задача: регрессия. Метрика: MAE.

Описывает гидрофобность молекул, обычно выражаемую как logD. Один из самых стабильных и хорошо сбалансированных датасетов в блоке Absorption.

Aqueous Solubility (AqSol)

Размер: **9,982 молекул**. Задача: регрессия. Метрика: MAE.

Самый крупный датасет в данном блоке. Отражает растворимость в воде, ключевой параметр для всех дальнейших ADMET-процессов. Благодаря большому размеру позволяет обучать модели с высокой устойчивостью.

2.2.2. Бенчмарки для оценки распределения

Эти свойства описывают движение молекулы внутри организма после попадания в кровь.

Blood-Brain Barrier Penetration (BBB)

Размер: **1,975 молекул**. Задача: классификация. Метрика: ROC-AUC.

Отражает способность молекулы проходить через гематоэнцефалический барьер. Особенно важно для разработки нейротерапевтических препаратов.

Plasma Protein Binding Rate (PPBR)

Размер: **1,797 молекул**. Задача: регрессия. Метрика: MAE.

Показывает долю молекул, связывающихся с белками плазмы крови. Влияет на распределение, метаболизм и активность препарата.

Volume of Distribution (VD)

Размер: **1,130 молекул**. Задача: регрессия. Метрика: MAE.

Параметр, отражающий степень распределения между кровью и тканями. Один из ключевых фармакокинетических показателей.

2.2.3. Метаболические бенчмарки: взаимодействие с ферментами CYP450

Эта группа датасетов оценивает способность молекул взаимодействовать с ферментами цитохрома P450 — основными элементами системы метаболизма лекарств. В TDC представлены задачи двух типов: ингибирование фермента и субстратность.

Ингибирование

- **CYP2D6-I — 13,130 молекул**
- **CYP3A4-I — 12,328 молекул**
- **CYP2C9-I — 12,092 молекул**

Метрика: PR-AUC. Это крупные датасеты, характеризующиеся сильным дисбалансом классов. Используются для оценки риска лекарственных взаимодействий.

Субстраты

- **CYP2D6-S — 664 молекул**
- **CYP3A4-S — 667 молекул**
- **CYP2C9-S — 666 молекул**

Метрика: PR-AUC. Эти датасеты значительно меньше, что делает задачу особенно сложной и чувствительной к архитектуре модели.

2.2.4. Бенчмарки для оценки выведения

Свойства, связанные со скоростью выведения препарата и его стабильностью в организме.

Half-Life

Размер: **667 молекул**. Метрика: ранговая корреляция Spearman.

Позволяет оценить, как долго препарат сохраняется в организме. Важно для расчёта дозировок.

Clearance-Microsomal (CL-Micro)

Размер: **1,102 молекул**. Метрика: ранговая корреляция Spearman.

Описывает скорость метаболизма в микросомах печени.

Clearance-Hepatocyte (CL-Hepa)

Размер: **1,020 молекул**. Метрика: ранговая корреляция Spearman.

Оценивает скорость метаболизма в гепатоцитах (клетках печени). Сложный датасет с высоким уровнем биологического шума.

2.2.5. Токсикологические бенчмарки

Эта группа задач является наиболее критичной с точки зрения безопасности лекарств.

hERG inhibition

Размер: **648 молекул**. Метрика: ROC-AUC.

hERG-ингибирование связано с риском кардиотоксичности, поэтому используется как один из индустриальных "must-pass" тестов.

AMES mutagenicity

Размер: **7,255 молекул**. Метрика: ROC-AUC.

Очень крупный и сбалансированный датасет, отражающий способность молекулы вызывать мутации.

Drug-Induced Liver Injury (DILI)

Размер: **475 молекул**. Метрика: ROC-AUC.

Один из самых сложных датасетов из-за малого размера и высокой биологической вариативности.

LD50 (oral toxicity)

Размер: **7,385 молекул**. Задача: регрессия. Метрика: MAE.

Оценивает дозу, вызывающую летальный исход у 50% объектов. Датасет крупный, но шумный, поскольку объединяет данные из разных источников.

2.3. Реализованные архитектуры

В рамках данного проекта были реализованы четыре базовые архитектуры, предназначенные для сравнения различных подходов к молекулярному представлению и последующему предсказанию ADMET-свойств. Выбор моделей охватывает наиболее распространённые типы фиче-инженеринга и нейронных сетей в хемоинформатике: векторные отпечатки, последовательные представления SMILES, графовые модели молекул и архитектуры с использованием замороженных предобученных эмбеддингов. Такой спектр моделей позволяет оценить, насколько способы кодирования молекулы влияют на качество предсказания, и создать репрезентативный набор бейзлайнов для последующего сравнения с планируемой архитектурой Mamba.

Каждая модель реализована в виде отдельного класса на базе PyTorch/PyTorch Geometric, что обеспечивает модульность, повторяемость и удобство расширения. Отдельно предусмотрен базовый класс с методом подсчёта числа обучаемых параметров, что помогает сопоставлять вычислительную сложность различных архитектур. В дальнейшем приведено подробное описание каждой модели, включая мотивацию, ключевые идеи и архитектурные особенности.

2.3.1. Morgan + MLP

Первая архитектура представляет собой классический подход, основанный на использовании молекулярных отпечатков (fingerprints). В качестве входного представления применяются Morgan fingerprints размерностью 1024 — один из наиболее распространённых и хорошо изученных дескрипторов в области QSAR-моделирования. Такие отпечатки кодируют локальное окружение атомов с учётом радиусов и типов связей, что позволяет фиксировать важные структурные фрагменты молекулы [3].

Сверху на эту векторную репрезентацию накладывается многоуровневая полносвязная сеть (MLP) с тремя скрытыми слоями. Архитектура включает ReLU-активации и dropout-слои, что снижает риск переобучения. Данная модель является важным бейзлайном, поскольку представляет собой традиционный «индустриальный стандарт» для задач предсказания физико-химических свойств и токсичности. Благодаря простоте и устойчивости, она задаёт нижнюю границу качества, с которой можно сравнивать более сложные последовательные и графовые модели.

2.3.2. SMILES-CNN

Вторая модель работает непосредственно со строками SMILES, используя их как последовательное представление молекулы. Перед подачей в свёрточные слои SMILES-строка токенизируется и кодируется с помощью trainable-эмбеддингов

размерности 64. Далее полученная матрица признаков обрабатывается тремя одномерными свёрточными слоями с различными размерами ядра. Такой приём позволяет извлекать фрагменты различной длины — от коротких подстрок, соответствующих атомарным паттернам, до более протяжённых функциональных групп.

После свёрток применяется max-pooling, агрегирующий признаки по всей длине последовательности, что делает модель инвариантной к длине SMILES. Далее следует небольшой MLP-блок для финального преобразования признаков в предсказание целевой величины [3].

Этот подход обеспечивает компромисс между простотой и способностью модели учитывать локальный контекст символов. SMILES-CNN служит важной точкой отсчёта для методов, которые работают с последовательностями, но не используют механизмы внимания или рекуррентные слои.

2.3.3. GCN для молекулярных графов

Третья архитектура относится к классу графовых нейронных сетей. Молекула в этом случае представляется в виде графа, где атомы являются узлами, а химические связи — рёбрами. Этот метод считается одним из наиболее естественных способов представления молекул, поскольку он напрямую отражает их структурную природу.

Модель состоит из пяти последовательно применяемых графовых свёрток типа GCNConv, каждая из которых передаёт сообщения между атомами, обновляя представление каждого узла. После обработки графа выполняется глобальный агрегационный пуллинг, который суммирует признаки всех атомов, формируя вектор всего соединения. Далее одиночный линейный слой преобразует полученные эмбеддинги в скалярное значение свойства [1].

GCN-архитектура хорошо подходит для задач, где важны топологические и структурные характеристики молекулы. Она служит сильным бейзлайном для сравнения с более современными графовыми моделями и трансформерами.

2.3.4. NeuralFP с замороженным графовым энкодером

Последняя модель — это модификация подхода Neural Fingerprints, использующая предобученный GCN в роли энкодера. Основная идея заключается в том, чтобы отделить обучение структурных эмбеддингов от обучения предсказательной части: графовый энкодер предварительно обучается на крупном датасете, после чего его параметры фиксируются (замораживаются), а поверх него обучается только небольшой MLP-декодер.

Такой подход позволяет анализировать, насколько полезные и универсальные признаки извлекает графовая модель, и снижает риск переобучения на небольших

выборках, поскольку большая часть параметров не обучается. Декодер состоит из нескольких полносвязных слоёв, последовательно уменьшающих размерность и адаптирующих эмбеддинг к конкретной ADMET-задаче.

NeuralFP представляет собой промежуточный шаг между классическим GCN и крупными предобученными foundation-моделями. Эта архитектура хорошо демонстрирует влияние предобучения и служит подготовительной ступенью перед внедрением Mamba-подхода в рамках дипломной работы.

2.4. Ограничения экспериментальной среды

Все эксперименты проводились в среде, где единственным доступным вычислительным ресурсом являлся центральный процессор без графического ускорителя. Это ограничение существенно повлияло на выбор моделей и на невозможность прямого воспроизведения современных SSM-архитектур, таких как Mamba, которые требуют специализированных CUDA-оптимизаций. Дело в том, что ключевое преимущество Mamba — высокая скорость работы благодаря селективному сканированию, реализованному в виде оптимизированных GPU-ядер. На CPU такие реализации либо недоступны. Поэтому обучение полноценной версии Mamba в рамках данного эксперимента было технически невозможным.

Ограничения вычислительной среды повлияли также на время обучения моделей. Тем не менее выбранные архитектуры были тщательно адаптированы, чтобы обеспечить возможность их запуска и достижения сходимости в разумные сроки. Это позволило провести систематическое сравнение моделей и получить объективную оценку их возможностей в условиях ограниченных вычислительных ресурсов.

Глава 3. Обучение моделей и анализ полученных результатов

В данной главе подробно описываются используемые экспериментальные протоколы, стратегии разбиения данных, выбор гиперпараметров, методы регуляризации и критерии оценки качества, применяемые для разных типов ADMET-задач. Особое внимание уделено сопоставлению поведения архитектур при обучении на малых и средних по размеру выборках, а также сравнению устойчивости моделей к дисбалансу классов и вариативности молекулярных структур. Во второй части главы проводится систематический анализ метрик, позволяющий выявить закономерности в качестве предсказаний и определить архитектурные особенности, влияющие на эффективность решения задач ADMET-прогнозирования. Такой подход обеспечивает целостное понимание того, почему каждая модель демонстрирует тот или иной уровень качества, и формирует основу для формулирования направлений дальнейших исследований.

3.1. Процесс обучения моделей

В экспериментальной части работы для унификации процесса обучения всех моделей был разработан универсальный тренер, реализованный в виде класса UniversalTrainer. Основная ответственность этого компонента заключалась в стандартизации этапов подготовки данных, запуска цикла оптимизации, валидации промежуточных результатов и финальной оценки на тестовом наборе. Модель переводилась на целевое вычислительное устройство и инициализировалась оптимизатор AdamW с параметрами, задаваемыми извне. В зависимости от типа задачи тренер автоматически выбирал соответствующую функцию потерь: для задач классификации использовалась бинарная кросс-энтропия с логитами, а для регрессионных задач — L1-функция потерь, эквивалентная MAE. Такая унификация позволила прогонять как классификационные, так и регрессионные задачи в едином пайплайне с минимальными различиями в кодовой базе, что повышало воспроизводимость экспериментов.

Цикл обучения был реализован по классической схеме «эпохи — батчи — обновление параметров», при этом на каждом шаге выполнялись стандартные операции: обнуление градиентов оптимизатора, прямой проход, вычисление лосса, обратное распространение ошибки и шаг оптимизации. Для повышения стабильности применялась пакетная обработка с фиксированным размером батча, а для валидации и теста использовались детерминированные даталоадеры без перемешивания данных. После каждой эпохи проводилась оценка модели на валидационной выборке с вычислением метрики, релевантной рассматриваемой задаче (ROC-AUC, PR-AUC, MAE или Spearman). Механизм сохранения

наилучшего состояния модели был организован на основании изменения валидационной метрики: сохранялся чекпоинт с наилучшей валидационной метрикой, который затем разворачивался для окончательной оценки на тестовой выборке. Такой подход минимизировал риск переобучения и обеспечивает более честную оценку обобщающей способности моделей.

Оценочная логика внутри тренера была реализована единообразно для всех типов моделей и учитывала особенности входных батчей. Пайплайн поддерживал два формата батчей: тензорные пары «признаки — метки» и объекты графовой библиотеки, содержащие признаки узлов, индексы рёбер и информацию о батче. Метод evaluate аккумулировал предсказания и цели по всем батчам, приводил их к одномерным массивам и обрабатывал числовые артефакты с использованием безопасных преобразований. Для классификации логиты переводились в вероятности посредством сигмоидной функции, после чего вычислялись ROC-AUC и PR-AUC с защитой от исключений и падением к дефолтным значениям в случае некорректной статистики (например, единичного класса). Для регрессии вычислялись MAE или ранговая корреляция Спирмена с обработкой NaN-результатов. Такая централизованная обработка метрик упрощала сбор и агрегацию результатов по множеству задач.

Реализация тренера и конфигурация эксперимента были адаптированы под ограниченную вычислительную среду: все тренировки выполнялись на центральном процессоре. Данная техническая предпосылка оказала существенное влияние на архитектурные решения и на организацию эксперимента в целом. Во-первых, выбор модельных конфигураций и их размеров был консервативным: глубина и ширина слоёв подобраны так, чтобы обеспечить сходимость в разумное время на CPU. Во-вторых, масштаб гиперпараметрического поиска был ограничен: вместо широкого grid- или random-search применялись практические эвристики, опирающиеся на литературные рекомендации и предшествующий опыт. Эти инженерные компромиссы позволили обеспечить воспроизводимость прогонов, но одновременно накладывают ограничение на интерпретацию результатов в контексте максимально возможного качества моделей.

Анализ логов обучения показал репродуцируемые закономерности в поведении различных архитектур. Последовательная стратегия «по задачам — по моделям» дала возможность прямо сравнивать финальные тестовые метрики четырёх реализованных архитектур для каждого датасета TDC.Во время проведения эксперимента было замечено, что свёрточная модель, работающая со SMILES, стабильно превосходила конкурентов в задачах классификации токсичности и барьерных свойств; MLP на основе отпечатков Morgan показывал высокую стабильность и лидировал в ряде регрессионных задач, связанных с клиренсом и

физико-химическими параметрами; графовая GCN в условиях CPU-обучения чаще демонстрировала более слабые результаты, чем ожидалось, а подход NeuralFP с замороженным энкодером показал неоднородную эффективность, превосходя конкурентов лишь в отдельных задачах. Эти эмпирические наблюдения подробно обсуждаются в разделе анализа результатов.

Несмотря на общую устойчивость пайплайна, в процессе тренировок выявились практические ограничения и возможные источники систематической ошибки. Во-первых, обучение на CPU ограничивало объём повторных прогонов, следовательно, многие результаты получены на одном или небольшом числе повторов, что повышает вероятность статистической флуктуации. Во-вторых, отсутствие в текущей реализации явных механизмов борьбы с дисбалансом классов (взвешивание потерь, oversampling, focal loss) могло негативно влиять на качество моделей в задачах с редкими положительными примерами, особенно в группе CYP-инхибиторов. В-третьих, были зарегистрированы отдельные случаи числовой нестабильности предсказаний (NaN), указывающие на необходимость дополнительной нормализации входных признаков и усиленной регуляризации при обучении на шумных биологических данных [5].

С позиции методологии, предложенный универсальный тренер обеспечивает прозрачную и воспроизводимую структуру эксперимента, что является важным преимуществом для курсовой работы. Тем не менее для повышения надежности выводов и уменьшения дисперсии результатов необходимо выполнить дополнительные шаги: расширить число повторов с различными seed-ами, организовать более тщательный гиперпараметрический поиск для ключевых архитектур, ввести стратегии балансировки классов и исследовать влияние нормализации и регуляризации на устойчивость обучения. Также целесообразно рассмотреть варианты частичного переноса предобученных эмбеддингов (frozen vs. fine-tuned) и применение техник оптимизации, пригодных для CPU (квантизация, праунинг, distillation), чтобы объективно оценить компромисс между вычислительной стоимостью и качеством предсказаний.

В заключение следует подчеркнуть, что выбранный подход к обучению и валидации представляет собой сбалансированную архитектуру эксперимента, адекватную условиям ограничения аппаратных ресурсов. Унификация логики обучения, аккуратная обработка метрик и централизованная агрегация результатов обеспечили надёжную основу для сравнительного анализа архитектур и позволили выявить ключевые направления дальнейшей работы, включая реализацию и исследование SSM/Mamba-подобных моделей.

3.2. Анализ полученных результатов

Полученные экспериментальные результаты позволяют выявить чёткие поведенческие различия между четырьмя исследуемыми архитектурами — Morgan+MLP, SMILES+CNN, GCN и NeuralFP — на разнообразных наборах задач ADMET, включающих классификацию барьерных свойств, предсказание ингибирования ферментов, токсикологические характеристики и регрессионные биофармацевтические параметры. Несмотря на единую методологию обучения, модели демонстрируют принципиально разное качество в зависимости от используемого представления молекулы и сложности структуры зависимости внутри данных.

Task	Metric	Morgan+MLP	SMILES+CNN	GCN	NeuralFP
BBB	ROC-AUC	0.8429	**0.9058**	0.4588	0.7388
PPBR	MAE	11.9905	**10.4078**	15.9136	16.9043
VD	MAE	2.4211	**2.3123**	2.5624	2.4784
CYP2D6-I	PR-AUC	0.5425	**0.6008**	0.1764	0.2612
CYP3A4-I	PR-AUC	0.7866	**0.7958**	0.5398	0.5812
CYP2C9-I	PR-AUC	**0.7242**	0.6923	0.5775	0.5417
CYP2D6-S	PR-AUC	**0.7373**	0.5490	0.3902	0.4028
CYP3A4-S	ROC-AUC	**0.7239**	0.6355	0.5769	0.5545
CYP2C9-S	PR-AUC	**0.4223**	0.3883	0.3713	0.3680
Half-Life	SPEARMAN	**0.6276**	0.0472	0.1115	-0.1117
CL-Micro	SPEARMAN	**0.4243**	0.1770	0.0974	0.0004
CL-Hepa	SPEARMAN	**0.2577**	0.2452	-0.0030	-0.0504
hERG	ROC-AUC	0.7350	**0.8506**	0.6931	0.6456
AMES	ROC-AUC	0.7880	**0.8103**	0.6800	0.6212
DILI	ROC-AUC	0.8229	**0.8837**	0.7483	0.8264
LD50	MAE	**0.5645**	0.5680	0.7237	0.7192
Caco2	MAE	**0.4822**	0.7722	1.0998	1.0300
HIA	ROC-AUC	**0.8045**	0.7985	0.7652	0.6182
Pgp	ROC-AUC	**0.8897**	0.8817	0.7401	0.8108
Bioav	ROC-AUC	0.5469	0.5612	**0.5833**	0.5182
Lipo	MAE	**0.6952**	0.7462	1.0099	0.9703
AqSol	MAE	1.1097	**0.9871**	1.3727	1.4733

Рис. 1. Полученные метрики после окончания эксперимента

3.2.1. Общие тенденции по архитектурам

Анализ всех задач показывает несколько ключевых закономерностей. Архитектура SMILES+CNN стабильно занимает лидирующие позиции по большинству классификационных задач. Это подтверждает тезис о том, что локально-зависимые текстовые паттерны SMILES хорошо подходят для свёрточных моделей и позволяют эффективно извлекать структурные фрагменты, релевантные токсичности, ингибированию и мембранной проницаемости.

Архитектура Morgan+MLP демонстрирует высокую устойчивость и показывает лучшие или сопоставимые результаты в ряде регрессионных задач (например, Caco-2, Lipo, VD). Такое поведение объясняется природой молекулярных

отпечатков Morgan, которые хорошо кодируют локальные окружения атомов, важные для молекулярных свойств, но менее полезны для реактивных свойств, зависимых от глобальной конформации.

Архитектура GCN, несмотря на свою теоретическую выразительность, существенно уступает остальным моделям практически по всем задачам. Основная причина — ограничение вычислительного ресурса (CPU) и невозможность подобрать оптимальные гиперпараметры и процедуры регуляризации. Как следствие, графовая модель не достигла своего потенциального качества.

Архитектура NeuralFP, основанный на замороженном GCN-энкодере, показывает смешанные результаты: в задачах, где графовая структура важна минимально, модель иногда сопоставима с лидерами, однако в большинстве задач уступает как CNN, так и Morgan+MLP. Это ожидаемо, поскольку фиксированный энкодер не адаптируется к конкретным задачам, и информация, выделяемая GCN, иногда оказывается недостаточной для эффективной декодирующей части.

3.2.2. Барьерные свойства

В задачах, связанных с проницаемостью через биологические барьеры, лидером почти всегда выступает SMILES+CNN, что подтверждается высокими значениями ROC-AUC для BBB (0.9058), hERG (0.8506) и AMES (0.8103). CNN уверенно извлекает локальные химические мотивы, отвечающие за способность молекул проникать через гематоэнцефалический барьер или абсорбироваться в кишечнике.

При этом Morgan+MLP демонстрирует выдающиеся результаты в задаче Caco-2 (MAE = 0.4822), что делает его лучшей моделью среди всех участников для этого регрессионного набора. Вероятно, свойства, определяющие пермеабельность через Caco-2, тесно коррелируют с локальными описателями, отражёнными в отпечатках Morgan, что делает MLP эффективным инструментом в данной задаче.

Интересно отметить, что Bioavailability оказалась сложной задачей для всех моделей: SMILES+CNN и Morgan+MLP показывают средние значения ROC-AUC (0.5612 и 0.5469 соответственно), что подтверждает известную в фармакокинетике трудность предсказания биодоступности из одних только структурных данных.

4.2.3. Ингибирование и субстратность ферментов семейства CYP

Эта группа задач представляет особый интерес, поскольку результат определяется тонкими различиями в локальной электронике молекулы и её пространственном строении. Конкретные результаты показывают, что SMILES+CNN уверенно превосходит остальные архитектуры во всех четырёх задачах ингибирования CYP (например, CYP2D6-I = 0.6008 PR-AUC, CYP3A4-I = 0.7958). Задачи

субстратности оказались сложнее, и разрыв между моделями здесь меньше, но CNN по-прежнему лидирует.

При этом GCN в задачах CYP проявляет себя особенно слабо (например, CYP2D6-I лишь 0.1764), что отражает неспособность графовой модели эффективно обучиться без тщательной настройки гиперпараметров и мощного GPU-обучения.

Модель NeuralFP показывает промежуточные результаты, что ожидаемо: фиксированный графовый энкодер не способен адаптироваться к высокочувствительным ферментативным данным.

4.2.4. Токсикология и безопасность

В токсикологических задачах наблюдается чёткая закономерность: SMILES+CNN вновь демонстрирует лучшие или сопоставимые результаты. Например, в задаче hERG модель достигает ROC-AUC = 0.8506, что заметно выше, чем у остальных моделей.

Задача DILI оказалась одной из немногих, где NeuralFP показала конкурентоспособный результат, достигая ROC-AUC = 0.8264 — сопоставимого с Morgan+MLP (0.8229). Возможно, замороженный графовый энкодер выделяет устойчивые глобальные признаки, актуальные именно для DILI.

В задаче LD50 все модели демонстрируют близкие значения MAE, где Morgan+MLP (0.5645) и SMILES+CNN (0.5680) практически равны. Это может свидетельствовать о том, что предсказание LD50 обладает высоким уровнем стохастичности по самой природе экспериментальных данных.

4.2.5. Фармакокинетические регрессионные задачи

Эта группа задач является наиболее чувствительной к качеству представления молекулы, поскольку временные параметры распада и скорости клиренса определяются широким спектром слабых взаимодействий, конформационных эффектов и метаболических реакций.

Результаты подтверждают высокую сложность: ни одна модель не достигает высоких значений Spearman. Тем не менее Morgan+MLP стабильно занимает первую позицию среди всех архитектур (например, Half-Life = 0.6276, CL-Micro = 0.4243), что делает его предпочтительным выбором для фармакокинетических регрессионных задач. Это отличает его от CNN, которая здесь оказывается едва ли не худшей (Spearman всего 0.0472 для Half-Life).

GCN и NeuralFP также показывают слабые результаты, что указывает на недостаточное количество данных и невозможность графовых моделей полноценно обучиться в условиях CPU.

4.2.6. Переход к архитектурам нового типа

В совокупности наблюдаемый ландшафт результатов подчёркивает важный вывод: ни одна из классических архитектур не обеспечивает уверенное превосходство по всем задачам ADMET. Каждая архитектура проявляет локальные преимущества, но также демонстрирует ограничения, связанные как с типом представления молекулы, так и с вычислительными ограничениями.

Эти результаты подтверждают необходимость перехода к более современным моделям последовательностей — таким как SSM-архитектуры (Structured State Space Models) и, в частности, Mamba, которые, по данным современных исследований, способны объединять преимущества CNN, трансформеров и графовых моделей, эффективно обрабатывая длинные входные последовательности, включая SMILES.

Именно поэтому следующая часть работы посвящена планированию реализации архитектуры Mamba и её последующей проверке на тех же наборах задач.

Группа	Задача	Метрика	Morgan+MLP	SMILES+CNN	GCN	NeuralFP
PhysChem	AqSol	MAE ↓	1.1097	**0.9871**	1.3727	1.4733
	Lipo	MAE ↓	**0.6952**	0.7462	1.0099	0.9703
Absorption	Caco2	MAE ↓	**0.4822**	0.7722	1.0998	1.0300
	HIA	ROC-AUC ↑	**0.8045**	0.7985	0.7652	0.6182
	Pgp	ROC-AUC ↑	**0.8897**	0.8817	0.7401	0.8108
	Bioav	ROC-AUC ↑	0.5469	0.5612	**0.5833**	0.5182
Distribution	BBB	ROC-AUC ↑	0.8429	**0.9058**	0.4588	0.7388
	PPBR	MAE ↓	11.9905	**10.4078**	15.9136	16.9043
	VD	MAE ↓	2.4211	**2.3123**	2.5624	2.4784
Metabolism	CYP2D6-I	PR-AUC ↑	0.5425	**0.6008**	0.1764	0.2612
	CYP3A4-I	PR-AUC ↑	0.7866	**0.7958**	0.5398	0.5812
	CYP2C9-I	PR-AUC ↑	**0.7242**	0.6923	0.5775	0.5417
	CYP2D6-S	PR-AUC ↑	**0.7373**	0.5490	0.3902	0.4028
	CYP3A4-S	ROC-AUC ↑	**0.7239**	0.6355	0.5769	0.5545
	CYP2C9-S	PR-AUC ↑	**0.4223**	0.3883	0.3713	0.3680
Excretion	Half-Life	Spearman ↑	**0.6276**	0.0472	0.1115	-0.1117
	CL-Micro	Spearman ↑	**0.4243**	0.1770	0.0974	0.0004
	CL-Hepa	Spearman ↑	**0.2577**	0.2452	-0.0030	-0.0504
Toxicity	hERG	ROC-AUC ↑	0.7350	**0.8506**	0.6931	0.6456
	AMES	ROC-AUC ↑	0.7880	**0.8103**	0.6800	0.6212
	DILI	ROC-AUC ↑	0.8229	**0.8837**	0.7483	0.8264
	LD50	MAE ↓	**0.5645**	0.5680	0.7237	0.7192

Рис. 2. Результаты моделей в разрезе свойств ADMET

Заключение

В ходе данной курсовой работы была проведена комплексная исследовательская программа, направленная на анализ современных нейросетевых архитектур для предсказания свойств лекарственно-подобных молекул (ADMET). Особое внимание уделялось сравнению моделей, основанных на различных способах молекулярного кодирования: традиционных молекулярных отпечатков (Morgan), последовательных представлений на основе SMILES, графовых нейронных сетей (GCN) и нейронных фингерпринтов (NeuralFP).

Полученные результаты позволили выявить ряд устойчивых закономерностей, важных как для понимания природы ADMET-задач, так и для проектирования более эффективных моделей будущего.

Во-первых, экспериментально подтверждено, что последовательностные модели, использующие SMILES-кодирование (SMILES+CNN), демонстрируют наиболее стабильные и высокие результаты на большинстве классификационных ADMET-бенчмарков, включая задачи токсичности (hERG, AMES, DILI), ингибирования ферментов CYP450 и проникновения через гематоэнцефалический барьер. Это свидетельствует о том, что линейное представление SMILES, при корректной архитектуре, эффективно фиксирует локальные и среднесрочные структурные зависимости, значимые для биофизико-химических свойств молекул.

Во-вторых, модель Morgan+MLP проявила неожиданно высокую конкурентоспособность, особенно на регрессионных задачах (VD, PPBR, Lipo, AqSol) и отдельных классификационных задачах. Это подтверждает, что традиционные молекулярные дескрипторы, несмотря на появление более сложных нейронных архитектур, остаются сильной базовой линией и обеспечивают оптимальный баланс между простотой, скоростью обучения и качеством.

В-третьих, графовые модели (GCN) и NeuralFP, наоборот, продемонстрировали наименее стабильные метрики среди всех архитектур. Это может объясняться как недостаточной емкостью выбранных реализаций GNN, так и особенностями самих ADMET-датасетов, которые часто содержат малые выборки и требуют моделей с высокой устойчивостью к шуму. Анализ ошибок показывает, что графовые подходы сильнее других страдают от переобучения и чувствительны к структуре данных, их размеру и вариативности. Это указывает на необходимость дальнейших исследований в области регуляризации GNN, разработки более глубоких MPNN-архитектур и внедрения современных последовательностно-графовых гибридов.

С точки зрения общей картины можно сделать два ключевых вывода. С одной стороны, архитектуры, основанные на SMILES-последовательностях, на текущем

этапе обеспечивают наиболее высокий уровень качества при умеренной вычислительной стоимости. С другой стороны, дальнейший рост качества предсказаний ADMET-свойств, вероятно, связан с внедрением более современных моделей класса state-space architectures, включая Mamba-подобные модели, которые сочетают преимущества CNN/Transformer-подходов, но при этом обладают улучшенной эффективностью и способностью обрабатывать длинные последовательности.

Таким образом, проведённое исследование не только позволяет системно оценить сильные и слабые стороны различных подходов к молекулярному моделированию, но и формирует прочную теоретическую и методологическую основу для развития более совершенных архитектур, способных повысить точность и надёжность вычислительного прогнозирования свойств лекарственных соединений.

Перечень использованных источников

1. Wang Y., Ouyang R., Zhan W., …, Zhuang Z. SMILES-Mamba: A Mamba-based Molecular Foundation Model. arXiv preprint arXiv:2408.05696, 2024. Available at: https://arxiv.org/abs/2408.05696 (accessed 10.12.2025).

2. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 1988, vol. 28, no. 1, pp. 31–36. https://doi.org/10.1021/ci00057a005

3. Kotsias P. C., Arús-Pous J., Chen H., Papadopoulos K., …, Tyrchan C. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nature Machine Intelligence*, 2020, vol. 2, no. 5, pp. 254–265. https://doi.org/10.1038/s42256-020-0174-5

4. Li Q., Han Z., Wu X. M. Deeper insights into graph convolutional networks for semi-supervised learning. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. Available at: https://ojs.aaai.org/index.php/AAAI/article/view/11604

5. Hochreiter S., Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, vol. 9, no. 8, pp. 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

6. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., …, Polosukhin I. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017, no. 30. (See also: Chithrananda S., Grand G., Ramsundar B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction, 2020.) https://arxiv.org/abs/1706.03762

7. Gu A., Dao T. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752, 2023. Available at: https://arxiv.org/abs/2312.00752

8. Huang K., Xiao C., Glass L. M., Zitnik M., …, Sun J. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. arXiv preprint arXiv:2102.09548, 2021. Available at: https://arxiv.org/abs/2102.09548