ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР «ИНФОРМАТИКА И УПРАВЛЕНИЕ» РОССИЙСКОЙ АКАДЕМИИ НАУК

На правах рукописи

Ерещенко Алексей Владимирович

# ПРИМЕНЕНИЕ ГРАФОВЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ АНАЛИЗА МОЛЕКУЛЯРНЫХ СТРУКТУР

Специальность 1.2.1

«Искусственный интеллект и машинное обучение»

Диссертация на соискание ученой степени

кандидата технических наук

Научный руководитель -

доктор физико-математических наук,

профессор Ревизников Дмитрий Леонидович

Москва 2025

**Оглавление**

# Введение

**Актуальность темы исследования.** Разработка новых лекарственных молекул является высокотехнологичным, многоэтапным и дорогостоящим процессом. Чтобы сократить время и финансовые расходы на ранних этапах разработки, обычно используются различные методы компьютерного моделирования. Это позволяет проанализировать, например, вероятную зону связывания белковой молекулы с лигандом, оценить свойства и возможные механизмы связывания подобной области, выбрать наиболее перспективные молекулы для виртуального скрининга, предсказать их фармакокинетические параметры и оценить другие важные свойства. Ключевым элементом работы современных вычислительных методов является анализ молекулярных структур, как белковых молекул, так и малых молекул. Методы машинного обучения становятся все более популярным инструментом для решения задач разработки лекарственных средств и виртуального анализа биологических соединений, однако их качество прямо зависит от качества и количества доступных данных.

Исторически сложились следующие ключевые подходы к виртуальному скринингу: лиганд-ориентированный дизайн лекарств (LBDD) и структурно-ориентированный дизайн лекарств (SBDD). Оба подхода имеют свои сильные и слабые стороны; однако второй метод считается более точным и информативным. Более того, для выявления новых лигандов ранее неизвестной мишени часто возникает ситуация, когда трехмерная (3D) модель белка-мишени доступна, а информации о потенциальных лигандах недостаточно или она полностью отсутствует [1]. В этом случае возможно применение методов, опирающихся на 3D структуру белка. Благодаря растущему объему структурных данных, депонированных в общедоступном хранилище данных о белковых молекулярных структурах Protein Data Bank [2], применимость методов SBDD стабильно возрастает. В данной диссертации рассмотрены актуальные вычислительные методы, применяемые в данной сфере.

Помимо поиска сайтов связывания белков, значимым является процесс оценки данного пространства с точки зрения возможной биохимической активности. Данная разметка может служить вспомогательным звеном для направленной разработки малых молекул, обладающих высокой вероятностью ингибирования, позволить характеризовать сайты связывания, а также помочь предсказывать вероятности ингибирования данной области уже известных малых молекул.

Несмотря на расширяющийся объем доступных 3D данных о молекулах, многие малые молекулы с экспериментально выявленными свойствами не обладают известными стабильными 3D конформациями, и могут быть рассмотрены только с точки зрения химического состава и связности. Существуют информационные базы с десятками миллионов записей подобных данных, открывая возможность для разработки моделей машинного обучения для решения задач медицинской химии, в частности предсказания фармацевтических свойств малых молекул и первичного анализа их возможных мишеней.

Графовые нейронные сети (GNN) представляют собой перспективную нейросетевую архитектуру для анализа молекулярных структур. GNN модели представляют данные в виде набора вершин и ребер, обладая большой гибкостью возможного для применения математического аппарата за счет реализации различных способов агрегации данных на вершинах. Подобное представление данных является наиболее близким к молекулярной структуре, которые наиболее часто описывают именно графом, позволяя хранить признаковое описание атомов на вершинах, а информацию о связности или расстояниях (в случае работы с 3D структурой) на ребрах графа. Графовые нейронные сети могут работать как с молекулярными структурами с известной 3D структурой, так и с более упрощенным представлением в виде последовательности атомов и матрицы связности, обеспечивая применимость данной нейросетевой архитектуры ко всему спектру доступных для анализа молекулярных структур.

**Степень разработанности.** Для анализа текущего состояния области был проведен обзор существующих решений для анализа молекулярных структур, как

использующих, так и не использующих машинное обучение, созданных за рубежом и в России. В монографии [3] показана применимость различных базовых физико-математических моделей и алгоритмов, а также многомасштабного моделирования, для решения прикладных задач материаловедения и предсказания свойств кристаллических материалов. В работах [4-7] представлены различные вычислительные методы идентификации сайтов связывания лигандов, в том числе основанные на геометрии, на основе сходства, на основе компьютерного зрения с применением сверточных нейросетей. В публикациях [8-10] описываются разработанные передовые нейросетевые методы, анализирующие молекулярные структуры, для решения различных задач данной области, таких как генерация структуры белка [8], молекулярное докирование [9] и де-ново проектирование молекул [10]. Авторами [11-13] представлены алгоритмы классификации сайтов связывания белковых молекул с применением машинного обучения.

С растущей популярностью машинного обучения и искусственного интеллекта в рассматриваемой области, в настоящее время разрабатываются многочисленные решения на основе машинного обучения для более быстрого и точного предсказания аффинности малых молекул. Недавние примеры таких алгоритмов включают InteractionGraphNet [14] и graphLambda [15].

Ввиду своей востребованности, задача поиска возможных мишеней малой молекулы активно изучается, в том числе применение методов машинного обучения для оценки крупных наборов молекул, основываясь на их структурной информации. Были реализованы методы с применением градиентного бустинга [16], опорных векторов [17], случайного леса [18]. В недавнем исследовании были разработаны 32 модели классификации с использованием XGBoost, RF, SVM и деревьев решений [19]. Были разработаны также и нейросетевые модели: модели глубоких нейронных сетей [20,21], модели свёрточных нейронных сетей [22,23], модели графовых нейронных сетей [24,25] и модели на основе трансформеров [26-28]. Известен метод AikPro [29], использующий 3D-конформации для

извлечения дополнительных дескрипторов, а также метод с применением фармакофорного описания малых молекул [30].

Также было разработано множество методов для оценки растворимости малых молекул: полуэмпирические методы [31], методы, которые не используют подогнанные параметры и основываются на квазихимической теории жидких смесей [32], методы, основанные на энергии [33], методы молекулярной динамики [34,35], модели, основанные на количественных данных взаимосвязи между структурой и активностью [36,37]. Решения, использующие машинное обучение также использовались для решения подобного рода задач [38-41].

Опираясь на последние академические исследования и принимая во внимание большой интерес в сфере разработки лекарственных средств, можно сформулировать множество актуальных задач, решение которых приведет к созданию более эффективных решений для применения на различных этапах разработки лекарств, подразумевающих компьютеризованные решения.

**Цель настоящей работы -** разработка и практическое применение алгоритмов машинного обучения на основе графовых нейронных сетей для анализа молекулярных структур. Для достижения данной цели были решены следующие задачи:

1. Разработан алгоритм с применением графовых нейронных сетей для поиска сайтов связывания белковых молекул, создающий объемное представление зоны связывания на поверхности молекулы. Показана эффективность алгоритма по сравнению с существующими актуальными решениями.

2. Предложен алгоритм с применением графовых нейронных сетей для оценки свойств пространства сайтов связывания белковых молекул. Показана эффективность разработанного алгоритма по сравнению с существующими актуальными решениями. Показана применимость оценок, которые производит алгоритм, для разработки и обучения других нейросетевых решений, решающих задачи предсказания дополнительных свойств сайта связывания, а также оценки аффинности малых молекул относительно данной белковой молекулы.

3. Разработан алгоритм с применением графовых нейронных сетей для предсказания возможных мишеней для малой молекулы, основываясь на теоретических 3D конформациях молекулы, без знания истинной 3D структуры. Показана эффективность алгоритма по сравнению с другими архитектурами моделей машинного обучения.

4. Исследовано применение алгоритма, основанного на графовых нейронных сетях, для оценки одного из фармакокинетических свойств малых молекул, а именно растворимости, без использования информации о 3D структуре. Оценена его эффективность по сравнению с другими архитектурами моделей машинного обучения.

В исследовании использовались методы: машинное обучение, градиентная оптимизация, кластеризация, графовые нейронные сети, многослойный перцептрон, одномерные сверточные нейронные сети, градиентный бустинг.

**Основные положения, выносимые на защиту:**

1. Предложен алгоритм поиска сайтов связывания белковых молекул с применением графовых нейронных сетей и графового представления трехмерной структуры белка, продемонстрирована более высокая точность поиска сайтов связывания по сравнению с известными алгоритмами, применяемыми в данной области.

2. Предложен алгоритм аннотации сайтов связывания белковых молекул с применением графовых нейронных сетей и графового представления трехмерной структуры белка, продемонстрирована более высокая точность аннотации по сравнению с распространенными алгоритмами, применяемыми в данной области. Продемонстрирована применимость оценок, генерируемых предложенным алгоритмом, в качестве признакового описания пространства. Показано, что модели, обученные на подобном описании, демонстрируют сравнимую или превосходящую точность по сравнению с существующими аналогами.

3. Предложены алгоритмы предсказания свойств малых молекул, а именно ингибирующую активность против тех или иных мишеней, а также растворимость, с применением графовых нейронных сетей и графового

представления молекулы. Продемонстрировано преимущество графовой нейронной сети по сравнению с известными алгоритмами машинного обучения в задаче предсказания наличия ингибирующей активности против заданной мишени.

**Научная новизна.** Предложен подход применения графовых нейронных сетей и графового представления данных для работы с трехмерной структурой белка и пространством на его поверхности, с созданием по результату работы алгоритма объемного сайта связывания: разработана соответствующая архитектура нейронной сети, проведено обучение моделей и их тестирование, показано преимущество обученных моделей по сравнению с иными моделями, использованными для сравнения. Предложен подход для аннотации сайта связывания белка с применением графовых нейронных сетей и графового представления трехмерной структуры белка: разработана соответствующая архитектура, проведено обучение моделей и их тестирование, продемонстрировано их преимущество по сравнению с аналогами. Изучено применение оценок, сгенерированных обученными аннотационными нейросетями, для обучения других графовых нейронных сетей, решающих иные задачи. Продемонстрированы модели, классифицирующие сайты связывания, и модель, предсказывающая аффинность малых молекул. Экспериментально показано, что созданные подобным образом модели сравнимы или превосходят существующие аналоги. Предложен подход по предсказанию наличия ингибирующей активности против заданной мишени (на примере 75 белковых мишеней) с применением графовой нейронной сети, обученной на ансамблях трехмерных фармакофорных представлений малых молекул, показано преимущество разработанного алгоритма по сравнению с иными подходами машинного обучения.

**Теоретическая значимость.** Теоретическая значимость работы заключается в исследовании графового описания трехмерной структуры белка и преимущества графовых нейронных сетей. Изучена возможность применения оценок обученных графовых сетей аннотации пространства белка для обучения

графовых нейросетевых моделей, в дополнение либо вместо других способов описания макромолекулярной среды точки пространства на поверхности белка. Исследован способ описания малой молекулы в виде ансамбля трехмерных конформаций, оценена результативность графовой нейронной сети, обученной на подобном пространственном описании.

**Практическая значимость.** Разработанные модели поиска сайтов связывания, аннотации пространства сайта связывания, а также созданные на основе моделей аннотации модели классификации сайтов связывания и предсказания аффинности малых молекул были внедрены в разрабатываемой платформе для генерации лекарственных молекул во ФГУП "ВНИИА". Разработанная модель предсказания ингибирующей активности против заданных мишеней была применена для разметки открытой библиотеки из более чем 80 000 соединений, не имеющих известного профиля активности.

**Достоверность и обоснованность результатов** подтверждена их непротиворечивостью и согласованностью с известными фактами и исследованиями в рассматриваемой области, экспертной валидацией результатов работы алгоритмов на представительных наборах данных, апробацией на научных конференциях, а также экспертным тестированием платформы для генерации лекарственных молекул, использующей разработанные алгоритмы.

**Апробация работы.** Основные результаты, освященные в диссертации, были представлены на следующих конференциях: 11-я Московская конференция по вычислительной молекулярной биологии (MCCMB) 3-6 августа 2023 года, XIII International Conference on Chemistry for Young Scientists "MENDELEEV 2024" 2-6 сентября 2024 года, XXX Symposium on Bioinformatics and Computer-Aided Drug Discovery (BCADD-2024) 16-18 сентября 2024 года, Всероссийский форум молодых исследователей ХимБиоSeasons 14-18 апреля 2025 года, с публикацией тезисов.

**Публикации.** По тематике диссертации были опубликованы 4 статьи, в том числе 3 статьи индексируемые в Q1 Scopus/WOS ([42],[43],[44]), и статья в журнале из списка ВАК ([45]). Также было получено авторское право на код:

свидетельство о государственной регистрации программы для ЭВМ № 2023684775, дата регистрации 20.11.2023 «Программа для виртуального поиска сайтов связывания малых молекул на поверхности трехмерных моделей белков “SiteRadar”» // Государственная регистрация программы для ЭВМ, бюллетень №11. Также был зарегистрирован патент № 2 838 984 «Способ разметки лиганд-белковых сайтов связывания».

**Личный вклад.** Работы диссертанта, выполненные с соавторами.

В [42] соискателем была проведена следующая работа: анализ и обработка подготовленных экспертами для обучения данных, формирование обучающей и валидационной выборок, подготовка архитектуры нейросети, тестирование различных архитектур нейросетей, обучение моделей, подбор алгоритмов кластеризации, перебор гиперпараметров кластеризации, разметка комплексов в ходе независимого тестирования (включая применение алгоритмов сравнения). В [43] соискателем была проведена следующая работа: анализ и обработка подготовленных экспертами для обучения данных для моделей аннотации сайтов связывания, формирование обучающей и валидационной выборок, подготовка архитектуры нейросети, тестирование различных архитектур нейросетей, обучение моделей. Также была проведена работа по доработке нейросети по предсказанию аффинности малой молекулы, дополнительных экспериментов по ее обучению, были проведены эксперименты по применению различного признакового описания и их влияния на итоговую точность модели. В [44] соискателем была проведена работа по исследованию применяемых вычислительных решений и анализу алгоритма AlphaFold.

В главе 3, в рамках работ по разработке алгоритма предсказания потенциальных ингибиторов малых молекул, соискателем была проведена следующая работа: анализ и обработка подготовленных экспертами для обучения данных, формирование обучающих, валидационных и тестовых выборок, подготовка архитектуры нейросети, тестирование различных архитектур нейросетей, обучение моделей, настройка модели, разработка архитектур и обучение моделей машинного обучения, использованных для сравнения.

Разработка алгоритма предсказания растворимости и связанные эксперименты были выполнены полностью самостоятельно.

Содержание диссертации и основные положения, выносимые на защиту, отражают персональный вклад автора в опубликованных работах. В диссертацию вошли результаты, которые получены лично автором. Результаты других авторов, упомянутых в диссертации, носят справочный характер и имеют сопутствующие обозначения.

**Структура и объем диссертации.** Диссертация состоит из введения, трех глав и заключения. В работе используется сквозная нумерация формул. В каждой главе используется своя автономная нумерация таблиц и иллюстраций. Полный объём текста диссертации составляет 122 страницы с 19 рисунками и 12 таблицами. Список литературы содержит 97 наименований источников.

# Глава 1. Разработка модели поиска сайта связывания белковой молекулы

В данной главе рассматривается применение GNN для поиска сайтов связывания белков на основании их трехмерной структуры. Рассматриваются два варианта признакового описания белковой структуры – с использованием информации об аминокислотном составе, и с использованием только позиционной информации атомов белка. Описывается алгоритм поиска сайтов связывания с применением обученных моделей. Проводиться сравнение алгоритма с рядом существующих моделей поиска сайтов связывания.

В 1.1 представлено введение в предметную область и ее значимость, проводится обзор существующих подходов к решению данной задачи. В 1.2 приведена постановка задачи, описание формирования входных данных, подхода к описанию белковой структуры и пространства вокруг нее, а также использованный GNN подход для ее решения. В 1.3 описывается архитектура разработанной GNN. В 1.4 представлен подход к обучению GNN моделей, формированию обучающей и валидационной выборок, представлены показатели точности обученных моделей на валидационной выборке. В 1.5 описан алгоритм генерации поиска сайтов связывания, использующий разработанные GNN модели. В 1.6 приведено тестирование алгоритма поиска сайтов связывания с использованием разработанных GNN на независимой выборке данных, подобранных как отличные от обучающей и валидационной выборок, описываются использованные оценочные метрики. Приводится сравнение алгоритма с рядом существующих решений на данной выборке, также приводятся результаты экспертного анализа отдельных примеров аннотации белков разработанным алгоритмом. В 1.7 приведены выводы по исследованиям, представленным в данной главе.

## 1.1 Введение

Современный дизайн лекарств основан на двух основных подходах: дизайне на основе лиганда (LBDD) и дизайне на основе структуры (SBDD). Для

идентификации новых лигандов ранее неизученной мишени характерна ситуация, когда имеется 3D модель целевого белка, а информации о его потенциальных лигандах недостаточно или она полностью отсутствует [1]. Учитывая это, SBDD является наиболее подходящим вычислительным подходом для решения этой проблемы. В связи с растущим объемом структурных данных, хранящихся в Protein Data Bank [2,46], важность методов SBDD неуклонно растет.

Первый этап SBDD включает в себя анализ трехмерной структуры мишени и идентификацию сайта связывания, который может быть потенциально лекарственным. В данной работе сайт связывания лекарственного средства рассматривается как карман, который связывает низкомолекулярный лиганд (включая молекулы лекарственного средства), что приводит к модуляции активности белка-мишени. Идентификация карманов, пригодных для связывания лекарств, расширяет количество белковых мишеней и создает новые возможности для разработки лекарств. Применение вычислительных методов, в частности методов машинного обучения, является менее дорогостоящим и трудоемким по сравнению с биофизическими методами идентификации сайта связывания (например, скрининг фрагментов, сайт-направленный мутагенез и т. д.). В настоящее время существуют различные вычислительные методы идентификации сайтов связывания лигандов, в том числе основанные на геометрии (Fpocket [4]), на основе сходства (ProBiS [5]), на основе компьютерного зрения (DeepSite [6], BiteNet [7]) и т. д. Несмотря на разнообразие существующих решений, их прогностическая способность невелика. Как правило, способность обнаруживать карманы для связывания лекарств не превышает 40% для большинства методов [47]. Современные подходы не повышают точность предсказания значительно по сравнению с алгоритмами немашинного обучения, особенно для идентификации ранее не охарактеризованных, неглубоких экспонированных растворителем [48] и аллостерических сайтов связывания [49], а также сайтов связывания ковалентных ингибиторов [50] и полостей на границе регионов белок-белок (PPI) [51].

Принимая во внимание тот факт, что данная область обладает большим простором для улучшения качества предсказания, был разработан метод на основе графовых нейронных сетей для определения сайтов связывания лекарств.

## 1.2 Задача поиска сайтов связывания белка

Общая задача может быть описана следующим образом: дан набор данных:

$$D = \{(x_k, y_k)\}_{k=1\ldots n} \quad (1)$$

где $x_k$ это некое признаковое представление объекта, а $y_k \in R$ это целевой ответ по данному объекту, при этом примеры $(x_k, y_k)$ независимы друг от друга. В данной работе подходом к решению задачи прогнозирования $y$ является разработка и обучение модели $F: R^m \to R$ которая бы минимизировала выбранную функцию потерь:

$$L(F) \coloneqq EL\big(y, F(x)\big) \quad (2)$$

где $(x, y)$ — это пары независимо выбранных из обучающего набора примеров.

В начале рассмотрим формирование набора данных $D$. Источником информации для $D$ являются трехмерные структурные данные, которые были получены из базы данных sc-PDB [52]. Были отобраны только те белковые субъединицы, тяжелые атомы которых расположены в пределах 7 Å от тяжелых атомов лиганда. Комплексы белок-лиганд, содержащие более 10 000 тяжелых атомов, были исключены. Вокруг каждой белковой структуры была сформирована кубическая решетка с интервалом 2 Å. Из сгенерированных решеток были удалены точки, находящиеся ближе, чем 2 Å от тяжелых атомов белка с целью исключить перекрытие решетки со структурой белка, а также были удалены точки, охарактеризованные как открытые для растворителя (подробнее данный критерий описан в [42]). Таким образом, объектом набора данных $D$ является точка созданной вокруг трехмерной структуры белка кубической решетки.

Каждая точка решетки была классифицирована как принадлежащая или не принадлежащая сайту связывания на основании близости к любому из тяжелых атомов лиганда, расположенного в сайте связывания данного белка, на

расстоянии до 2 Å. Таким образом, $y_k$ для данной задачи является бинарный класс «сайт связывания» и «не сайт связывания». Признаковое описание $x_k$ точки решетки было сформировано в виде графа, созданного на основе окружающей ее макромолекулярной среды. Для этого были собраны 3D-координаты и характеристики аминокислот для тяжелых атомов белка в пределах 7 Å от каждой точки решетки (рисунок 1.1). Координаты использовались для расчета расстояний между тяжелыми атомами белка и точкой решетки, однако не были включены явно как признаки. Каждая точка решетки представлялась в виде графа, состоящего из тяжелых атомов белка и самой точки решетки в виде узлов, а расстояния от тяжелых атомов белка до точки решетки — в виде ребер. Кроме того, также были собраны все расстояния между тяжелыми атомами белка в пределах 7 Å друг от друга. Были собраны специфические для аминокислот данные, закодированы one-hot encoding методом (преобразованы в бинарные векторы) и использованы в качестве узловых признаков. Эти данные представлены названием остатка (Asp, Glu, Phe и т. д.), названием атома (N, CA, C, O, CB и т. д.), названием элемента (C, N, O, S), меткой «основная цепь» или «боковая цепь» и классом аминокислот. Белки с нестандартными атомами (например, C4, O1P, CH3 и т. д.) были пропущены. Подробнее специфические признаки и их отбор описаны в [42]. Также, часть точек решетки класса «не сайт связывания» была исключена из обучения, подробнее описано в разделе 1.4. Т.к. точка решетки не может обладать химическими признаками, для ее описания были созданы особые признаки, отличающие ее от атома белка.

Моделью $F: R^m \to R$ в данной задаче является GNN. GNN обрабатывает молекулярный граф как сочетание узлов и рёбер, где каждый узел описывается набором признаков, относящихся к соответствующему атому, а рёбра задаются списками пар индексов признаков узлов. Существует множество вариантов архитектур операторов графовых нейронных сетей, которые отличаются используемыми математическими методами передачи сообщений и обновления информации на узлах. В выбранном графовом нейронном операторе [53]

применяется механизм многоголового внимания, широко используемый в трансформерных нейронных сетях, и он может быть описан следующим образом:

$$x_i = W_1 x_i + \sum_{j \in N(i)} a_{i,j}\left(W_2 x_j + W_5 e_{ij}\right) \quad (3)$$

где $x_i$ это корень графа, $x_j$ — это узел, от которого передаётся сообщение, $e_{ij}$ — это признаки на ребре, соединяющем узлы $x_i$ и $x_j$, $W_1$ соответствует матрице весов корня графа, $W_2$ — это матрица весов значений, $W_5$ — матрица весов признаков ребра, $N(i)$ — множество индексов всех узлов, связанных с данным узлом.

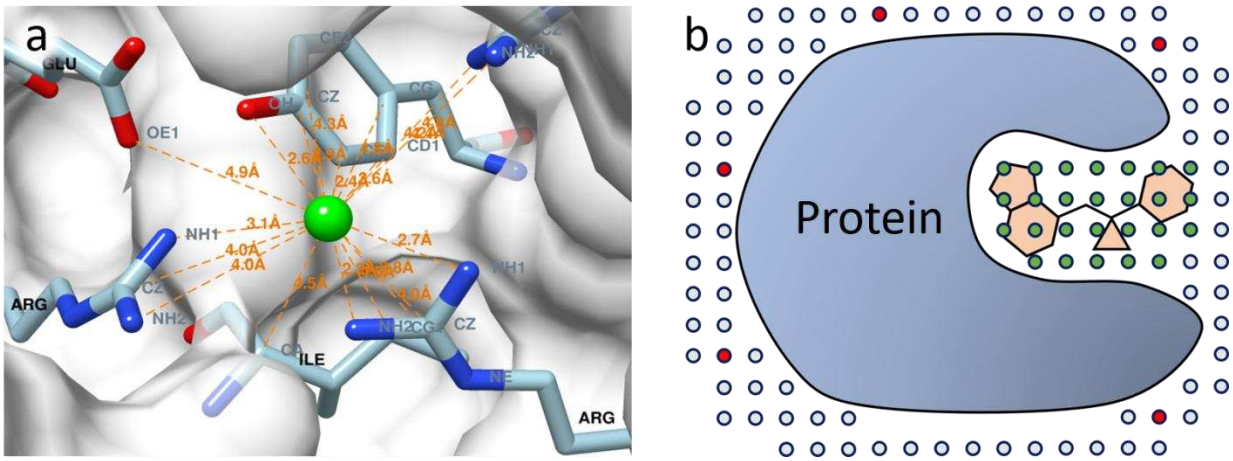


Рисунок 1.1. Визуальное представление процедуры сбора данных (а, зелёный шар обозначает точку сетки) и выбора положительных и негативных примеров (б, лиганд показан оранжевым цветом, точки решетки класса «сайт связывания» показаны зелёным цветом, точки решетки, принадлежащие классу «не сайт связывания» показаны белым цветом, а выбранные отрицательные точки сетки показаны красным цветом). Источник [42]

Коэффициент внимания $a_{i,j}$ вычисляется следующим образом:

$$a_{i,j} = soft\max\frac{(W_3 x_i)^T\left(W_4 x_j + W_5 e_{ij}\right)}{\sqrt{d}} \quad (4)$$

где $d$ это скрытая размерность каждой «головы» внимания, $W_3$ — это матрица весов запроса, $W_4$ — матрица весов ключа. $W_1$, $W_2$, $W_3$ и $W_4$ включают в себя обучаемый параметр суммируемого смещения.

В рамках данной задачи были обучены нейронные сети двух разных видов – с использованием информации о химическом окружении (специфическая к аминокислотам (АК) модель), и модель, учитывающая только расстояния между точкой решетки и окружающими атомами белка (геометрическая модель). Это было сделано с целью выявить вклад признаков, описывающих химическое окружение, в предсказательную силу итоговой модели, а также оценить возможность поиска сайтов связывания используя только пространственную информацию.

### 1.3 Архитектура разработанных нейронных сетей

Как было сказано ранее, были обучены две модели, использующие различное признаковое описание. Обе модели были построены с использованием библиотек Python Pytorch [54] и Pytorch Geometric [55]. Модели используют схожую архитектуру с основным отличием, заключающимся в размерности первого слоя из-за разницы в количестве используемых признаков (72 бинаризованных признака для АК-специфичной модели по сравнению с бинарным признаком (точка решетки или атом белка) для геометрической модели).

В рамках разработанной архитектуры, векторные представления признаков на узлах и ребрах были сгенерированы отдельно. Элементы узла были пропущены через GNN слой с графовым нейронным оператором, описанным выше, который преобразует входящие бинаризованные признаки в 300-мерное представление. Признаками на ребрах при этом были расстояния между вершинами. В данном слое GNN оператор использовался с 3 головами внимания (с усреднением выходного сигнала каждой головы), а к выводу слоя был применен слой случайного обнуления (dropout) с вероятностью 20%. Значения на ребрах, полученных из входящего графа, обрабатывались с помощью слоя размытия гаусса с 50 гауссианами и последовательности двух линейных слоев с функцией активации усеченного линейного преобразования с «утечкой» (leaky ReLU). Данная операция была проведена с целью представления информации о

расстоянии в виде распределения, а не скаляра. Данный подход был продемонстрирован в [56], где показал свою эффективность. Результатом работы данного блока преобразования информации о расстояниях является 32-мерный вектор. Полученные на первом GNN слое представления признаков и 32-мерный вектор поступали в три последовательных GNN слоя, аналогичных вышеописанным, но с 6 головами внимания (выходные сигналы голов объединяются). В качестве активации также использовалась функция активации leaky ReLU. Также между слоями была применена пакетная нормализация и dropout с вероятностью 20%. Выходные сигналы последнего GNN слоя усреднялись по среднему значению для каждого графа. Полученный вектор проходил через слой dropout с вероятностью 30%, затем поступал в выходной линейный слой. Как для АК-специфической, так и геометрической модели, выходным сигналом модели является оценка точки решетки (а именно графа, ее представляющего) в виде действительного числа в диапазоне [0,1]. Полученные оценки использовались для классификации точек на классы «сайт связывания» (1) либо «не сайт связывания» (0), с порогом отсечения 0.5. Итоговая архитектура моделей представлена на рисунке 1.2.

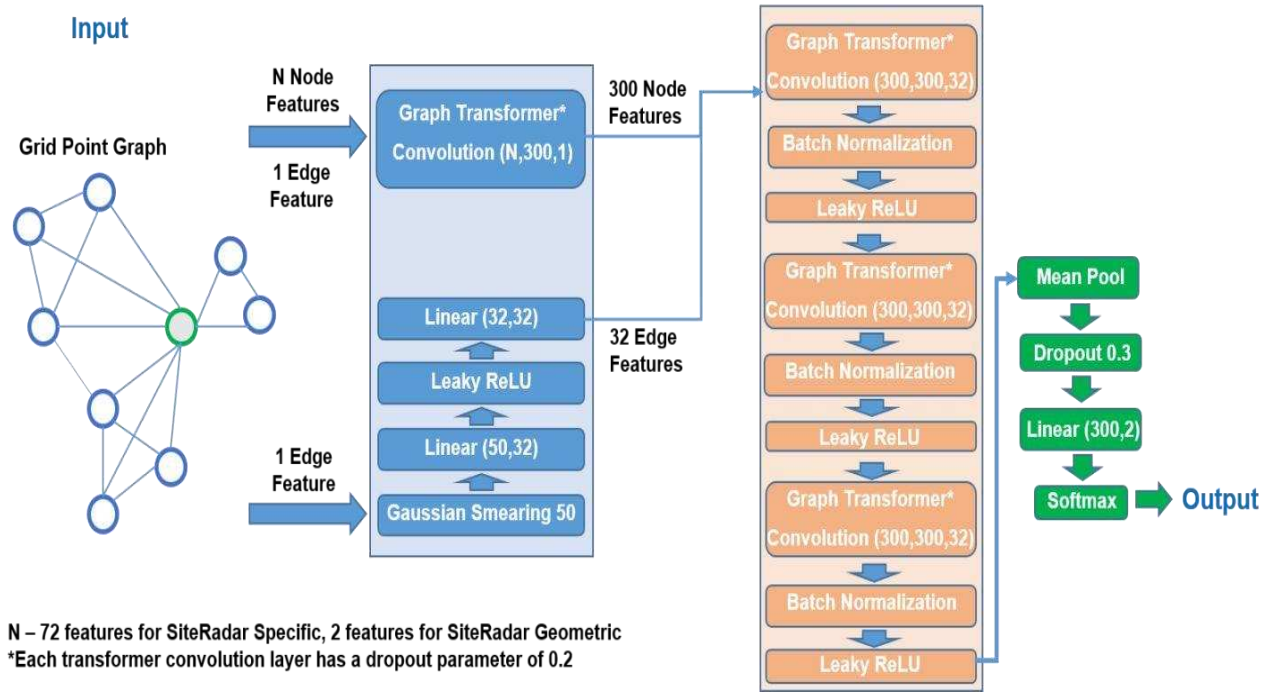


Рисунок 1.2. Архитектура разработанных нейронных сетей. Источник [42] (дополнительный материал)

## 1.4. Обучение моделей

Собранные данные были случайно распределены в наборы для обучения (95%) и валидации (5%) для контроля переобучения. Разбиение было произведено по идентификаторам белковых комплексов, а не индивидуальных точек решетки. Это было сделано с целью исключить возможность попадания точек из пространства одного и того же сайта связывания (которые могут обладать схожей локальной макромолекулярный средой) в обучение и валидацию. Следует отметить, что сформированная выборка данных, с учетом созданного графового описания, является весьма ресурсоёмкой с точки зрения занимаемого пространства при хранении и обработке в оперативной памяти, что представляло сложности для проведения экспериментов на имеющихся вычислительных мощностях. Кроме того, количество точек, описывающих пространство, классифицируемое как «сайт связывания», значительно меньше, чем представителей класса «не сайт связывания» (примерно 1/110). В связи с этим, перед проведением обучения на всей собранной выборке данных, на случайно собранной подвыборке данных, составляющей порядка 1\18 от всего массива данных (331 125 точек из 5 859 318), были проведены эксперименты по обучению на данных, из которых были исключены 90% и 95% точек класса «не сайт связывания». По результатам эксперимента, обучение с использованием полного набора точек класса «не сайт связывания», а также 10% точек данного класса, не давало преимущества по сравнению с вариантом, где были сохранены только 5% точек класса «не сайт связывания». В связи с этим, итоговое обучение было проведено именно в такой конфигурации.

Суммарно итоговая выборка, использованная для обучения и валидациии, включала в себя 5 859 318 графов, где каждый граф представлял собой точку решетки и ее макромолекулярное окружение на расстоянии 7 Å.

Обучающие данные были переданы моделям в батчах, где каждый элемент батча был представлен графом узла решетки и его макромолекулярном окружением, как описано выше. В одном батче разрешалось содержать

информацию о точках решетки из разных белков. Использовался размер батча 1536, и перемешивание данных выполнялась каждую эпоху обучения.

Для обучения моделей использовалась функция потерь кросс-энтропии, которая может быть представлена следующей формулой:

$$L = -\frac{1}{N}\sum_{I=1}^{N} w_{y_i} \log\left(\frac{e^{z_{i,y_i}}}{\sum_{j=1}^{C} e^{z_{i,j}}}\right) \quad (5)$$

где $N$ это размер батча, $C$ = количество классов, $z_{i,j}$ = логит для класса j примера i, $y_i$ это истинный индекс класса для примера i, $w_{y_i}$ = вес для истинного класса примера i (из тензора весов). Чтобы компенсировать разницу в количестве примеров в целевых классах (сайт связывания, не сайт связывания), были применены веса для балансировки классов. Все модели были построены и обучены с использованием библиотек Pytorch и Pytorch Geometric. Видеокарта NVIDIA A100-SXM4 с 80 гигабайтами памяти использовалась для обучения моделей примерно по 1 часу на эпоху. Производительность моделей была проверена с использованием средней невзвешенной точности, средней невзвешенной полноты и оценки F-1.

АК-специфичная модель была обучена с использованием оптимизатора Adam [57] с динамической скоростью обучения и применением уменьшения веса: первая эпоха имела скорость обучения $1*10^{-3}$, для следующих 10 эпох применялось уменьшение веса $5*10^{-5}$, а в течение обучения последних 10 эпох скорость была снижена до $1*10^{-4}$, а снижение веса изменено до $5*10^{-6}$. Геометрическая модель обучалась в течение 10 эпох с использованием оптимизатора Adam со скоростью обучения $1*10^{-3}$ и без снижения веса. Для обучения обеих моделей использовалось отсечение градиента по норме с параметром максимальной нормы установленным как 1.

По результатам тестирования на валидационной выборке, АК-специфичная модель показала более высокую точность с точки зрения макросредней точности, а именно 0,88, по сравнению с 0,82 у геометрической модели. Однако почти не наблюдалось различий в показателях полноты макросреднего значения (0,91 и 0,9 для АК-специфичной и геометрической моделей соответственно).

С целью проверки воспроизводимости результатов и влияния случайного фактора при разделении данных на валидацию и обучение, а также при исключении точек с классом «не сайт связывания», каждая модель была обучена повторно с применение другого значения рандомизации. Результаты продемонстрированы в таблице 1.1. Как было сказано ранее, каждый объект обучающей выборки является графом, включающим в себя точку решетки и тяжелые атомы белка, окружающие ее, на расстоянии 7 Å. Данное расстояние было подобрано на основании экспертной оценки специалистов предметной области, а также из соображений вычислительных затрат – чем больше радиус покрываемого окружения, тем более крупными будут сформированные графы с точки зрения занимаемого пространства в оперативной памяти и при хранении.

| Модель | Обучающий сет | Макро средняя точность | Макро средняя полнота | F-1 мера |
|---|---|---|---|---|
| АК-специфичная | 1 | 0.88 | 0.91 | 0.89 |
| | 2 | 0.88 | 0.92 | 0.90 |
| Геометрическая | 1 | 0.82 | 0.90 | 0.85 |
| | 2 | 0.84 | 0.90 | 0.87 |

Таблица 1.1. Метрики точности на валидационной выборке обученных моделей на разных рандомизациях обучающих данных

Дополнительно был проведен эксперимент с обучением модели на графах, сформированных по окружению с радиусом 5 Å. Данные модели показали более низкую точность на валидационной выборке: АК-специфичная модель показала макро среднюю точность 0.76, макро среднюю полноту 0.84, и F1-меру 0.79. Показатели геометрической модели, обученной по графам, сформированным по с радиусом 5 Å, также были ниже: макро средняя точность 0.69, макро средняя полнота 0.76, и F1-мера 0.72. Эксперименты с расстоянием выше 7 Å не проводились т.к. данный порог был выбран предельным с учетом вычислительной нагрузки при обработке более крупных графов.

## 1.5. Алгоритм поиска сайтов связывания с использованием разработанных и обученных GNN

С применением обученных GNN моделей был разработан подход для поиска сайтов связывания SiteRadar, подробно описанный в [42]. Он состоит из следующих последовательных шагов: 1) подготовка решетки; 2) аннотация решетки; 3) кластеризация; 4) выполнение оценки. GNN, обученные на данных, сформированных тем же способом, что и в ходе работы алгоритма на шаге 1, применяются для аннотации всех точек сформированной решетки (шаг 2), возвращая как оценку в виде действительного числа, так и предсказанный класс.

Результатом работы SiteRadar является объемный сайт связывания. С целью его формирования из аннотированных GNN точек решетки, применяется двухэтапная кластеризация. На первом этапе точки решетки, классифицированные как «сайт связывания», группируются на основе их трехмерных координат с использованием метода агломеративной кластеризации [58] со следующими параметрами: кластеризация точек на основе одиночной связи между точками, с использованием нормы расстояния L2 и порогом расстояния для кластеризации 2,1 Å. Первая процедура кластеризации предназначена для формирования отдельных областей точек, которые могут формировать объемный сайт связывания. На этом шаге по умолчанию сохраняются только карманы с не менее чем 20 точками решетки. Второй этап кластеризации используется для разделения крупных полостей, которые могут включать в себя несколько проксимальных сайтов связывания. Здесь к уже сформированным на первом этапе кластерам применяется метод кластеризации MeanShift [59] с параметром bandwidth 5,8 Å. Точки, которые не были отнесены ни к одному кластеру, присоединяются к ближайшему. Данные параметры были подобраны экспериментально, в том числе с учетом экспертной оценки получаемых по итогу работы SiteRadar объемных сайтов связывания, а также основываясь на параметрах известных мест связывания лекарственных средств, подробнее данные критерии описаны в [42]. Было проведено тестирование

алгоритма при использовании других гиперпараметров кластеризации, эта информация приведена в разделе 1.6.4.

На последнем этапе проводится оценка и ранжирование предсказанных сайтов связывания с точки зрения лекарственной способности. Чтобы ранжировать идентифицированные сайты связывания, оценка для каждого предсказанного кармана рассчитывается путем усреднения оценок всех точек решетки в кармане.

Алгоритм SiteRadar представлен в графическом виде на рисунке 1.3.

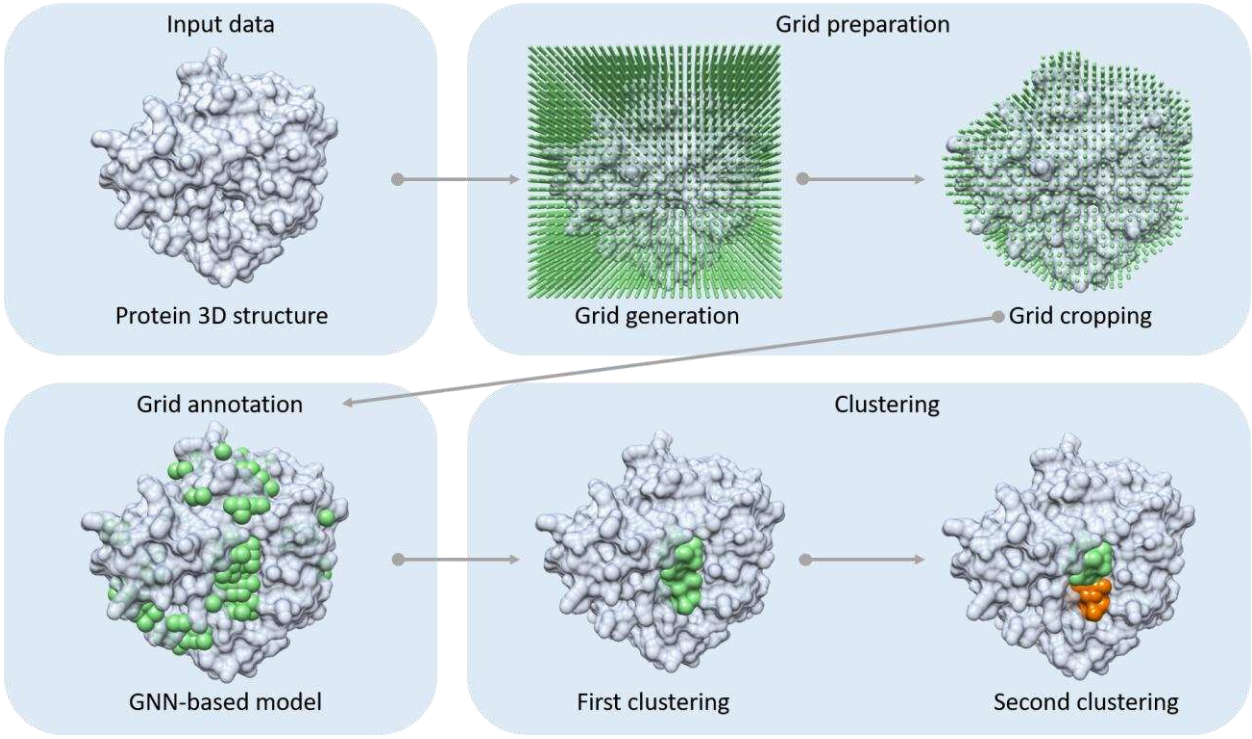


Рисунок 1.3. Архитектура алгоритма SiteRadar. Источник [42]

## 1.6 Тестирование на независимой выборке и сравнение с существующими алгоритмами

### 1.6.1 Подготовка данных для независимого тестирования

Для проведения сравнения алгоритма SiteRadar была выбрана модель, основанная на геометрическом анализе и не использующая машинное обучение Fpocket, а также модель, основанная на сверточных нейронных сетях PUResNet [60]. Чтобы сравнить SiteRadar с этими алгоритмами был подготовлен набор данных кристаллических структур, содержащих лиганды, подобные

лекарственным средствам. В этих целях из базы данных RCSB PDB [2] (дата доступа 28.08.2021) была сформирована выборка из 232 белков и 244 связанных лигандов. Данные были подобраны как наиболее непохожие на обучающие данные, с применением известных метрик оценки схожести белковых структур и экспертного анализа, подробно процесс описан в [42]. Дополнительно методом, описанным в [42], была проведена проверка на то, отличаются ли сайты связывания в валидационной выборке от сайтов в обучающей выборке.

### 1.6.2 Метрики сравнения

Каждая запись из эталонного набора данных была обработана с использованием стандартной процедуры SiteRadar с применением обеих обученных GNN моделей (АК-специфичная и геометрическая). Для сравнения предсказаний использовались программное обеспечение Fpocket [61] и PUResNet со стандартными параметрами и настройками. Чтобы проанализировать полученные результаты в одинаковых условиях, генерируемые Fpocket сферы alpha были преобразованы в решётки с шагом 2 Å.

В этом исследовании применялись два класса метрик. Первая группа предназначена для оценки способности изучаемых методов правильно определять местоположение сайта связывания лекарственного средства: расстояние «центр-центр» (DCC), оценка top N, top N+2, а также среднее количество предсказанных карманов на белок. Вторая группа показателей использовались для оценки способности моделей воспроизводить трехмерную форму указанных лигандов в пределах правильно обнаруженных сайтов связывания: покрытие лиганда (LC), покрытие кармана (PC) и перекрытие дискретного объема (DVO). Метод расчета данных метрик описан в [42]. Кроме того, были рассчитаны объемные параметры размеченных моделями сайтов связывания, они были сопоставлены с пространственными характеристиками известных сайтов связывания лекарственных средств, подробно данное сравнение описано в [42].

### 1.6.3. Статистика и воспроизводимость

Для LC, PC, DVO и количества предсказанных карманов был выполнен статистический анализ с использованием библиотеки SciPy Python [62]. Для проверки распределения данных был применен критерий Шапиро-Уилка. В случае нормального распределения использовался критерий Манна-Уитни; в противном случае использовался t-критерий Стьюдента для независимых выборок. Разница считалась значимой, если p-значение было меньше 0,05.

### 1.6.4 Результаты сравнения

Дискриминационная способность SiteRadar была сравнена с Fpocket и PUResNet с использованием 232 предварительно обработанных кокристаллов (см. подготовку набора для тестирования выше). Входные данные тестирования показали низкое сходство последовательности аминокислот с обучающим набором (максимальное сходство последовательности аминокислот с обучающим набором не превышало 52%). Кроме того, согласно распределению минимальных различий между участками связывания, было обнаружено высокое несоответствие между тестовым и обучающим наборами (подробнее в [42]).

В результате SiteRadar правильно определил 76% (АК-специфичная GNN) и 82% (геометрическая GNN) истинных участков связывания по метрике DCC (Рисунок 1.4, левая часть, Таблица 1.2), что выше, чем у Fpocket и PUResNet (71% и 46% соответственно) (Рисунок 1.5). Среднее количество предсказанных карманов на белок составило 4,2 для SiteRadar с использованием АК-специфичной GNN и 7,6 для SiteRadar с использованием геометрической GNN, в то время как Fpocket и PUResNet генерировали в среднем 19,2 и 1 карман на белок соответственно (Рисунок 1.4, Рисунок 1.5, правая часть). SiteRadar с использованием АК-специфичной GNN правильно идентифицировал 49% и 72% истинных участков связывания по метрикам top N и top N+2 соответственно.

| | DCC | Top N | Top N+2 | LC | PC | DVO | Среднее количество карманов |
|---|---|---|---|---|---|---|---|
| SiteRadar АК-специфичный | 0.76 | 0.49 | 0.72 | 0.82 | 0.47 | 0.40 | 4.2 |
| SiteRadar геометрический | 0.82 | 0.47 | 0.74 | 0.83 | 0.43 | 0.39 | 7.6 |
| Fpocket | 0.71 | 0.31 | 0.46 | 0.90 | 0.34 | 0.35 | 19.2 |
| PUResNet | 0.46 | ND | 0.47 | 0.95 | 0.27 | 0.45 | 1 |

Таблица 1.2. Результаты SiteRadar алгоритма с применением АК-специфичной и геометрической GNN, а также Fpocket и PUResNet на независимой выборке данных

SiteRadar с использованием геометрической GNN показал сопоставимые значения (47% для top N и 74% для метрики top N+2). Успешность Fpocket была намного ниже (31% для top N и 46% для метрики N+2). Поскольку PUResNet не может предоставлять оценку для ранжирования сгенерированных карманов, метрика top N неприменима к этому алгоритму. Однако PUResNet не генерировал более трех карманов для каждого белка из тестового набора, что позволило применить метрику top N+2; модель показала точность 47%, что сопоставимо с Fpocket. Что касается способности соответствовать форме кармана, Fpocket и PUResNet показали лучшие результаты по метрике LC и достигли точности 90% и 95% соответственно. Менее точные, но сопоставимые друг с другом результаты были получены для SiteRadar с использованием АК-специфичной модели (82%) и геометрической модели (83%). SiteRadar превзошел Fpocket и PUResNet по метрике PC (47% и 43% для АК-специфичной GNN и геометрической GNN соответственно). Fpocket продемонстрировал значение PC, равное 34%, а PUResNet показал точность 27%. В рамках метрики DVO, точность SiteRadar (40% для АК-специфичной и 39% для геометрической) была выше, чем у Fpocket (35%), но ниже, чем у PUResNet (45%).

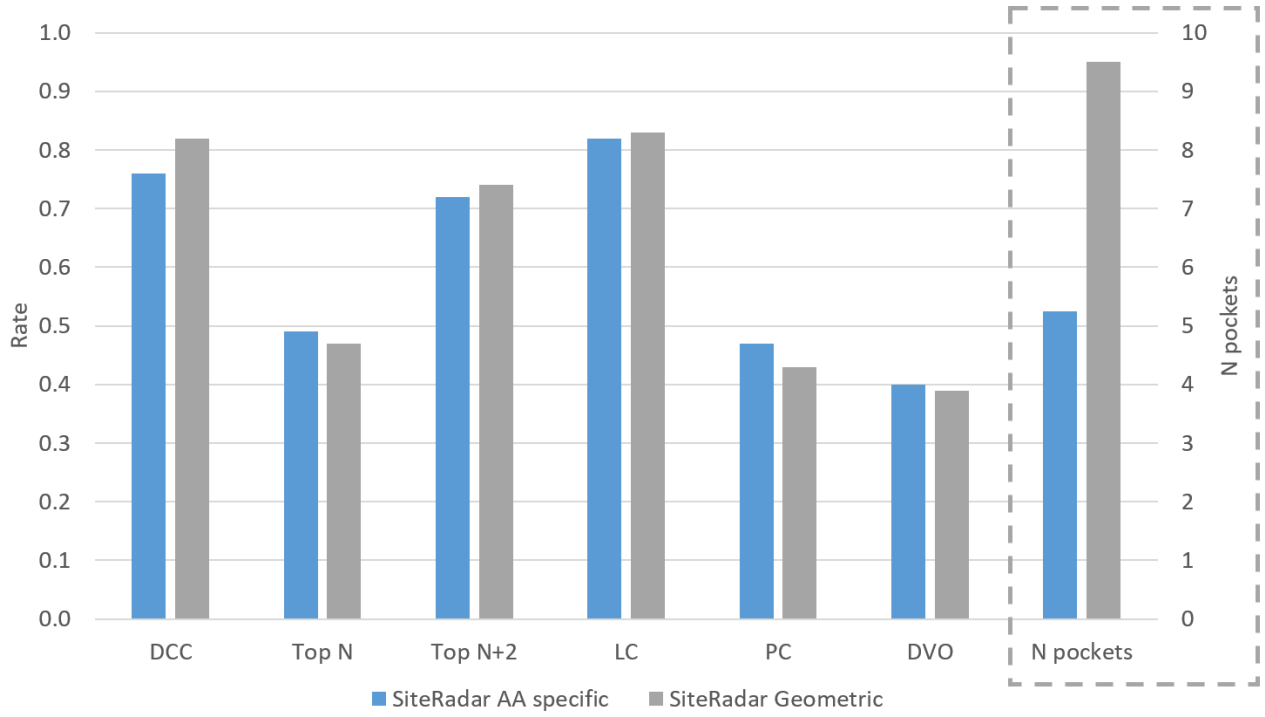


Рисунок 1.4. Результаты, полученные на независимой выборке моделями SiteRadar. Источник [42]

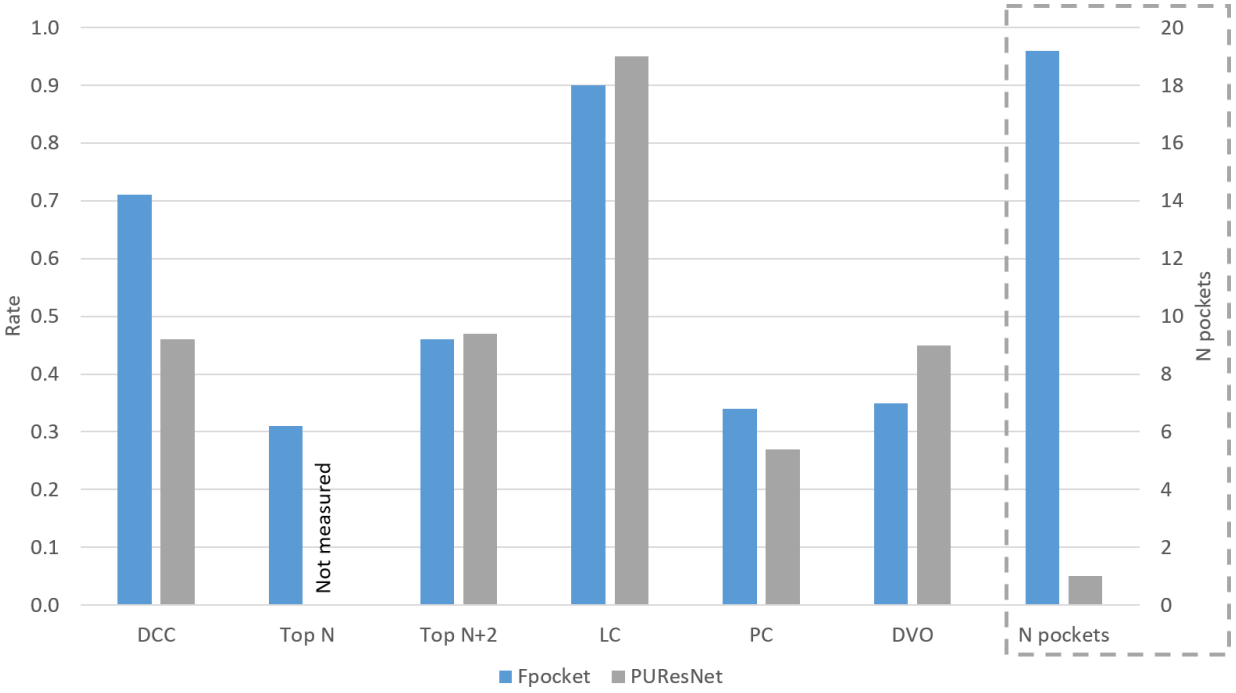


Рисунок 1.5. Результаты, полученные на независимой выборке методами Fpocket и PUResNet. Источник [42]

Далее сгенерированные карманы сравнивались с известными фармакологически доступными участками связывания по объёмным параметрам. Результаты показали, что FPocket и PUResNet более склонны генерировать чрезвычайно большие карманы, которые не соответствуют пространственным

характеристикам известных фармакологически доступных полостей. Только 57% и 19% карманов, сгенерированных FPocket и PUResNet соответственно, обладают теми же свойствами, что и фармакологически доступные карманы, тогда как SiteRadar генерирует карманы, соответствующие параметрам известных участков связывания лигандов в 72% случаев при применении АК-специфичной модели и в 77% при применении геометрической модели.

Кроме того, все упомянутые метрики были рассчитаны для 17 дополнительных настроек с разными комбинациями пороговых значений агломеративной кластеризации (первый этап кластеризации) и параметра bandwidth MeanShift (второй этап кластеризации). Результаты представлены в таблице 1.3.

| Модель | Предельное значение агломеративной кластеризации | Параметр bandwidth MeanShift | DCC | Среднее количество карманов | LC | PC | DVO | Top N | Top N+2 | Доля сгенерированных сайтов с объемными параметрам, характерными для ингибируемых сайтов |
|---|---|---|---|---|---|---|---|---|---|---|
| Геометрическая | 2.1 | 5 | 0.86 | 9.5 | 0.75 | 0.49 | 0.39 | 0.47 | 0.77 | 0.79 |
| | | 5.8 | 0.82 | 7.6 | 0.83 | 0.43 | 0.39 | 0.47 | 0.74 | 0.77 |
| | | 7 | 0.79 | 6.1 | 0.89 | 0.40 | 0.38 | 0.53 | 0.73 | 0.67 |
| | 2.9 | 5 | 0.87 | 11.7 | 0.75 | 0.49 | 0.39 | 0.45 | 0.78 | 0.68 |
| | | 5.8 | 0.83 | 9.2 | 0.82 | 0.43 | 0.38 | 0.45 | 0.72 | 0.7 |
| | | 7 | 0.76 | 7.0 | 0.89 | 0.39 | 0.37 | 0.48 | 0.69 | 0.63 |
| | 3.5 | 5 | 0.89 | 13 | 0.74 | 0.49 | 0.4 | 0.46 | 0.78 | 0.64 |
| | | 5.8 | 0.81 | 9.7 | 0.82 | 0.43 | 0.39 | 0.44 | 0.69 | 0.67 |
| | | 7 | 0.76 | 7.3 | 0.89 | 0.39 | 0.37 | 0.48 | 0.69 | 0.61 |
| АК-специфичная | 2.1 | 5 | 0.79 | 5.2 | 0.77 | 0.50 | 0.40 | 0.48 | 0.75 | 0.80 |
| | | 5.8 | 0.76 | 4.2 | 0.82 | 0.47 | 0.40 | 0.49 | 0.72 | 0.72 |
| | | 7 | 0.72 | 3.5 | 0.87 | 0.43 | 0.39 | 0.50 | 0.69 | 0.65 |
| | 2.9 | 5 | 0.81 | 6.2 | 0.77 | 0.50 | 0.39 | 0.48 | 0.74 | 0.71 |
| | | 5.8 | 0.79 | 5.0 | 0.82 | 0.46 | 0.39 | 0.51 | 0.73 | 0.67 |
| | | 7 | 0.74 | 3.94 | 0.87 | 0.42 | 0.38 | 0.51 | 0.71 | 0.60 |
| | 3.5 | 5 | 0.59 | 5.7 | 0.76 | 0.50 | 0.39 | 0.31 | 0.51 | 0.67 |
| | | 5.8 | 0.80 | 5.2 | 0.82 | 0.46 | 0.39 | 0.49 | 0.73 | 0.64 |
| | | 7 | 0.30 | 2.8 | 0.89 | 0.39 | 0.37 | 0.15 | 0.25 | 0.58 |

Таблица 1.3. Результаты алгоритма SiteRadar с применением АК-специфичной и геометрической моделей на независимой выборке при различных гиперпараметрах кластеризации

### 1.6.5 Экспертная валидация на отдельных примерах

Для демонстрации области применимости разработанного алгоритма SiteRadar был применен к 3D-структурам различных фармакологических мишеней, не имевших аналогов в обучающем наборе данных с точки зрения оценки сходства последовательностей АК (максимальное сходство последовательностей с обучающим набором не превышало 43%), был проведен экспертный анализ полученных результатов. Подробно проведённые эксперименты описаны в [42]. Обобщая полученные результаты, было показано, что SiteRadar способен идентифицировать различные сайты связывания лигандов, в том числе открытые для растворителя полости, аллостерические сайты, карманы связывания ковалентных лигандов, а также сайты PPI. Поведение АК-специфичной и геометрической моделей различно с точки зрения достоинств и ограничений. В целом, использование химических данных обеспечивает более высокую точность в тех случаях, когда форма кармана отличается от обычных сайтов связывания, например, неглубокие полости, открытые для растворителя. Напротив, геометрическая модель демонстрирует более высокую достоверность при использовании необычных комбинаций аминокислот.

## 1.7 Выводы

Были продемонстрированы возможности GNN для анализа молекулярных структур в рамках решения задачи поиска сайтов связывания. С применением обученных на обширной выборке данных трехмерных структур GNN был разработан алгоритм формирования объемных сайтов связывания SiteRadar. Было показано, что данный алгоритм обеспечивает относительно высокую точность прогнозирования как истинного местоположения сайта связывания, так и его объемных параметров. Используя относительные расстояния и графовую топологию, SiteRadar представляет собой эквивариантный подход, который не чувствителен к начальным положениям анализируемого белка, состоянию вращения или смещения в отличие от других подходов, рассматривающих белок как трехмерный объект. В ходе экспертного анализа было показано, что

разработанный алгоритм может быть успешно применен к широкому спектру лекарственных карманов, включая нетрадиционные открытые для растворителя, аллостерические сайты, сайты PPI и полости для ковалентного связывания лигандов.

В целях оценки вклада химических дескрипторов, а также оценки способности GNN корректно определять пространство, относящееся к сайту связывания, используя только информацию о структуре белка, были обучены две GNN модели, АК-специфичная и геометрическая. На внутренней валидации было показано, что обе GNN модели способны корректно определять точки, относящиеся к сайту связывания, при этом АК-специфичная модель показала более высокую макросреднюю точность и F1-меру, со сравнимым показателем макросредней полноты.

Была сформирована отдельная выборка данных, отличная от обучающей выборки, для проведения независимого тестирования алгоритма SiteRadar, с применением разработанных экспертами метриками оценки качества формирования объемных сайтов связывания. Было проведено сравнение SiteRadar с существующими методами поиска сайтов связывания Fpocket и PUResNet. Было показано, что разработанный алгоритм определяет сайты связывания белок-лиганд с большей точностью, чем Fpocket и PUResNet, согласно метрикам DCC, top N и top N+2. Этот результат достигается благодаря более сбалансированному количеству генерируемых карманов и более точной оценки размеченных точек пространства разработанными GNN. PUResNet и Fpocket демонстрируют более высокий охват лигандов, однако чаще создают необоснованно большие сайты связывания, тогда как SiteRadar формирует карманы более избирательно. В рамках метрики LC было показано, что SiteRadar склонен разделять крупные предсказанные карманы на отдельные сайты связывания разумного размера. Это может влиять на значение LC в случае крупных лигандов, например бифункциональных ингибиторов, которые связываются с несколькими близко расположенными субкарманами одновременно. Для конкретной задачи генерации эти субкарманы могут быть объединены в единый объем связывания. Значения

DVO показывают, что SiteRadar заполняет карманы немного хуже, чем PUResNet, но точнее, чем Fpocket. SiteRadar с применением АК-специфичной GNN и геометрической GNN показал в целом сопоставимые результаты, однако геометрическая модель в среднем размечала большее количество карманов как сайты связывания белок-лиганд. С другой стороны, АК-специфичный SiteRadar обеспечивает меньшее количество карманов, не влияя на способность правильно идентифицировать сайты связывания белок-лиганд согласно метрикам top N и top N+2. Таким образом можно сказать, что АК-специфичная GNN более избирательна при поиске сайта связывания. В отношении объемных параметров данные варианты также не показали значительной разницы для правильно обнаруженных сайтов связывания. Однако метод, основанный на химических дескрипторах, более подходит для применения к карманам с неклассической формой (например, мелкие сайты связывания на поверхности растворителя), тогда как геометрический подход демонстрирует хорошие результаты при работе с белковыми структурами, обладающими необычным сочетанием аминокислот.

Основным ограничением подхода является время расчетов, которое во многом зависит от скорости построения и подготовки решетки, кроме того, графовое представление пространства в заявленном виде является ресурсоемким. Это делает это решение не самым лучшим для скрининга большого количества белков. Однако, в первую очередь, алгоритм был разработан для применения в автоматизированном компьютерном дизайне лекарств, где этап идентификации кармана выполняется один раз для целевого белка, что снижает значимость скорости расчета.

# Глава 2. Разработка модели анализа свойств сайта связывания белковой молекулы

В данной главе рассматривается применение GNN для аннотации сайта связывания белковой молекулы на основании макромолекулярного окружения данного пространства. Описывается возможное применение данной разметки, в том числе и для обучения иных ML моделей. Приводится сравнение обученных GNN с существующими моделями разметки пространства. Рассматриваются модели, обученные с применением данных, сгенерированных моделями, в рамках задач классификации сайтов связывания и предсказания аффинности малых молекул.

В 2.1 представлено введение в предметную область и ее значимость, проводится верхнеуровневый обзор возможного применения аннотации пространства. В 2.2 приведена постановка задачи, описание входных данных, а также использованный GNN подход для ее решения. В 2.3 описывается архитектура разработанной GNN. В 2.4 представлен подход к обучению GNN моделей, формированию обучающей и валидационной выборок. В 2.5 описан алгоритм генерации псевдолигандов, использующий разработанные GNN модели. В 2.6 приведено тестирование разработанных GNN, а также сравнение точности аннотации пространства с их применением с существующими алгоритмами. 2.7 посвящен применению аннотации пространства разработанными GNN для формирования признакового пространства для обучения GNN моделей классификации сайтов связывания, а также анализу результатов тестирования и сравнения полученных моделей с опубликованными в научной литературе аналогами. 2.8 описывает применение аннотации пространства разработанными GNN для формирования дополнительных признаков для обучения GNN модели предсказания аффинности лиганд-белок: описывается формирование обучающей выборки и входных данных, архитектура модели, ее обучение, а также результаты тестирования модели на наборе данных, отличном от обучающих, результаты сравнения с существующими ML и не ML решениями; показаны эксперименты,

направленные на выявление вклада признаков, сгенерированных аннотационными GNN моделями, в производительность обученной модели. В 2.9 приведены выводы по исследованиям, раскрытым в данной главе.

## 2.1. Введение

В последние годы глубокое обучение (DL) существенно изменило область структурной биологии и разработки лекарств. Были созданы передовые методы DL для различных целей, таких как генерация структуры белка [8,44], молекулярное докирование [9] и де-ново проектирование молекул [10]. Другой важной задачей, которая еще не была решена с помощью методов машинного обучения, является генерация псевдолигандов. Генерация псевдолигандов относится к процессу выбора подходящих типов атомов для каждой точки внутри участка связывания лиганд-белок. Этот метод имеет широкий спектр применения в области молекулярной биологии и структурного проектирования лекарств. В частности, генерация псевдолигандов может решить такие задачи как классификация участков связывания лиганд-белок и предсказание аффинности лиганд-белок. Традиционные методы генерации псевдолигандов основаны на силовых полях и не учитывают сложные взаимодействия между лигандом и белком.

Принимая во внимание тот факт, что данная область обладает большим простором для методов машинного обучения, был разработан метод на основе GNN для оценки свойств пространства сайта связывания лиганд-белок. Также данный метод был применен для обучения моделей классификации сайтов связывания и оценки аффинности лиганд-белок; продемонстрировано, что ответы обученных GNN, предназначенные для аннотации пространства, могут использоваться как признаки для обучения других GNN моделей.

## 2.2 Задача разметки сайта связывания белка

Задача разметки сайта связывания белка была сформулирована схожим образом с задачей поиска сайта связывания белка, описанной в 1.1, но с

некоторыми отличиями. В данном случае в качестве набора данных $D$ используются лиганд-белковые комплексы, полученные из RCSB PDB [2], где каждым объектом данных является точка в пространстве сайта связывания белка. Однако, в отличии от подхода, описанного в главе 1, в рамках данной задачи не обязательно создается решетка – в процессе обучения, анализируемыми точками были атомы лиганда из лиганд-белкового комплекса, а при применении модели целевым объектом может быть как произвольная точка пространства, так и атом молекулярной структуры.

Каждая точка была классифицирована 13 классами, которые соответствуют тем или иным функциональным свойствам, которыми может обладать данное пространство, они представлены в таблице 2.1. Данный список атомов был экспертно разработан на основе таких свойств атомов, как химический элемент, гибридизация, участие в образовании колец и способность обладать положительным или отрицательным зарядом (подробнее в [43]). Признаковое описание точки было сформировано по аналогии с тем, как это выполнено для задачи поиска сайта связывания, а именно в виде графа, включающего саму точку и ее белковое окружение. Были использованы те же описательные признаки, что и для АК-специфичной модели из главы 1.

| Тип атома | Описание |
|---|---|
| Car | Атом углерода в гибридизации sp2 внутри ароматического кольца |
| Cs2 | Атом углерода в гибридизации sp2 вне ароматического кольца |
| Cs3 | Атом углерода в гибридизации sp3 |
| Csp | Атом углерода в гибридизации sp |
| Nd+ | Ионизируемый азотный атом, связанный с водородным атомом; может служить донором водородной связи |
| Nd0 | Электронейтральный азотный атом, связанный с водородным атомом; может служить донором водородной связи |
| Nac | Электронейтральный азотный атом без водородного атома; может служить акцептором водородной связи |
| .=O | Атом кислорода в гибридизации sp2; может служить акцептором водородной |

| | |
|---|---|
| | связи |
| O_a | Атом кислорода в гибридизации sp3 без водородного атома; может служить акцептором водородной связи |
| O_d | Атом кислорода в гибридизации sp3, связанный с водородным атомом; может служить акцептором водородной связи |
| Sul | Бивалентный атом серы |
| SO2 | Шестивалентный атом серы |
| Hal | Любой галоген, кроме фтора |

Таблица 2.1 Предсказываемые типы атомов и их описание

Для решения данной задачи, для каждого класса была обучена своя бинарная модель – это было сделано для более точечного контроля обучения модели под каждый конкретный класс, а также лучшего понимания ограничений при их обучении. Для каждой из 13 моделей были собраны положительные и отрицательные примеры (например, Car-атом и не-Car-атом). Отрицательные примеры состояли из всех типов атомов, исключая положительные примеры.

Применяемой моделью является GNN, функционирующая схожим образом – используется тот же графовый оператор, а данные подаются в виде графа. Однако, модель обладает иной реализацией в рамках обучения и общей архитектуры (подробнее в 2.4).

Дополнительно была поставлена задача предсказания вышеупомянутых типов атомов, но только в тех случаях, где атом действительно формирует ключевые взаимодействия с белком, например водородные связи, ароматическое стекирование (подробнее данные взаимодействия и их отбор описаны в [43]). Для этого были обучены дополнительные 13 моделей (далее именуемые фармакофорными), которые использовали ту же выборку данных, но с более строго отобранными положительными примерами классов.

## 2.3 Архитектура разработанных GNN для разметки сайта связывания

И фармакофорные, и базовые бинарные модели имели одну и ту же архитектуру и были реализованы с помощью библиотек Pytorch [54] и Pytorch Geometric [55] для Python. Архитектура GNN моделей схожа с архитектурой,

представленной в разделе 1.3, с определёнными отличиями. Также как и для GNN по разметке сайтов связывания, векторные представления признаков на узлах и ребрах были сгенерированы отдельно – признаки узлов были пропущены через GNN слой с графовым нейронным оператором, а признаки на ребрах (расстояния между атомами, посчитанные по 3-мерным координатам) через слой гауссового размытия с двумя последовательным линейными слоями. Однако, в данном случае признаки были преобразованы в более крупные 320 мерные вектора. Полученные векторные представления признаков на ребрах и узлах поступали в GNN блок, состоящий из 8 последовательных GNN слоев, аналогичных вышеописанным, но с применением 16 голов внимания (выходные сигналы голов объединяются), с применением батч нормализации и функции активации leaky ReLU. По сравнению с архитектурой, представленной в разделе 1.3, dropout слой не применялся – вместо этого использовался dropout на уровне вершин графа – у каждой вершины была 20% вероятность передать нулевой сигнал. Вывод GNN блока объединялся с выводом первого GNN слоя, и подавался в отдельный GNN слой (аналогичный оператор, но без применения dropout на уровне вершин). Затем вывод данного GNN слоя проходил слой активации leaky ReLU, усреднялся по среднему (global mean pooling). Усредненный вектор проходил через dropout слой, а затем поступал в линейный слой, выдающий две итоговые оценки – первая соответствовала целевому классу (например, Car), а вторая соответствовала отрицательному классу (не-Car).

### 2.4 Обучение разработанных GNN для разметки сайта связывания

Набор данных из 8150 кристаллов комплексов лиганд-белок был получен из RCSB PDB [2] для подготовки обучающих данных как для базовых, так и для фармакофорных моделей, с основным различием в аннотации и фильтрации данных. На этапе сбора данных для базовых моделей каждый атом лиганда был классифицирован в соответствии с 13 типами атомов (таблица 2.1), и было собрано его графовое представление, состоящее из атома и окружающих его тяжелых атомов белка в радиусе 5 Å. Кроме того, все расстояния между

тяжелыми атомами белка также включались в граф, если они не превышали 5 Å. Была проведена дополнительная фильтрация структур с целью исключения ковалентных ингибиторов, молекул, неглубоко внедренных в свои сайты связывания (72 лиганда), а также белковых структур с нестандартными названиями остатков или атомов (подробнее в [43]).

Обработка данных фармакофорных моделей была подобна, но включала дополнительные шаги, направленные на сбор дополнительных дескрипторов взаимодействий для каждого атома (подробнее в [43]). Сбор этой информации был необходим для аннотации атомов: хотя они все еще придерживаются тех же 13 типов атомов, процесс разметки обучающих данных учитывал наличие соответствующих взаимодействий. Все атомы, которые не могли быть классифицированы таким образом, были помечены как неопределенные и всегда были частью негативных примеров для любой из 13 бинарных фармакофорных моделей.

Для фармакофорных и базовых моделей 10% данных были случайным образом отделены в набор валидации на основе белка, чтобы избежать присутствия графов атомов лиганда из одной структуры в обучающем и валидационном наборах. Для каждой бинарной модели все атомы были либо помечены как 1 (относящиеся к классу), либо 0 (являющиеся либо другим классом, либо неопределенными).

Для каждой модели обучающие данные передавались в модель батчами размером 2048, где каждый элемент батча был графовым представлением атома лиганда и его белкового окружения, сконструированным как описано выше. Каждый батч случайным образом перемешивался во время обучения, что позволяло точкам из разных молекул присутствовать в одном батче данных. Была использована функция потерь бинарная кросс-энтропия с корректировкой весов для учета несбалансированности классов (веса рассчитывались для каждого типа индивидуально). Было использовано усечение градиентов с диапазоном [-1, 1] для стабилизации процесса обучения. Все модели были обучены со скоростью обучения $10^{-5}$ с помощью оптимизатора Adam. Для каждой модели было

использовано 100 эпох обучения. Фиксировались те веса модели, при которых сумма функции потери на обучающей и валидационной выборках была минимальной. Подобный подход к обучению был применен в связи с тем, что в ходе обучения модель выходила на «плато» по значению валидации, показывая небольшие колебания (пример представлен на рисунке 2.1 для обученного классификатора класса Nd0), при этом все еще стабильно снижая показатель функции потери на обучающих данных. Проходя плато, значения валидации резко ухудшались, демонстрируя переобучение модели. Соответственно, для получения лучшего решения (как на данных, схожих с обучением, так и отличающихся), критерии остановки определялся по сумме функций потерь, предотвращая преждевременную остановку обучения модели.

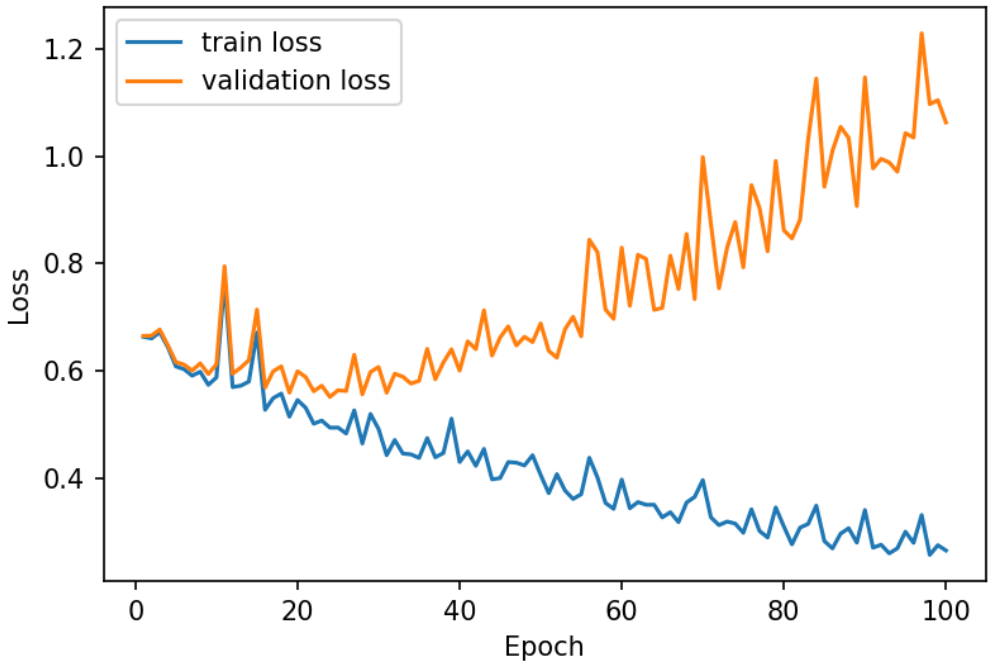


Рисунок 2.1 График потерь на обучающей и валидационной выборках классификатора Nd0

Все модели были построены и обучены с помощью библиотек Pytorch и Pytorch Geometric. Для обучения моделей была использована видеокарта NVIDIA A100-SXM4 с 80 Гб памяти, что занимало примерно 1 час 20 минут для 100 эпох.

## 2.5. Алгоритм аннотации сайта связывания белка и генерации псевдолигандов

С применением обученных GNN был разработан алгоритм разметки псевдолигандов в заданном белковом окружении Skittles, описанный в [43]. Алгоритм состоит из двух последовательных шагов: (1) генерация формы псевдолиганда и (2) предсказание типа атома псевдолиганда (Рисунок 2.2). Для генерации формы необходим алгоритм, способный создать объемное пространство, по окружающим его атомам белка. В рамках работы, описанной в [43], для этого шага использовался вышеупомянутый алгоритм поиска сайтов связывания SiteRadar. Результаты этого шага представлены облаками точек на поверхности белка (силуэт псевдолиганда), которые соответствуют гипотетическим участкам связывания лиганд-белок. В ходе второго шага используются обученные GNN для разметки сайтов связывания, позволяющие получить оценку степени принадлежности каждой точки силуэта псевдолиганда к используемым 13 классам. Для обеспечения более высокой точности классификации точек при совместной работе обученных бинарных моделей были подобраны балансирующие степенные коэффициенты. Таким образом, предсказание одной модели с применением коэффициента может быть описано как $f(x) = x^a$, где $x$ это предсказание модели, а $a$ – подобранный коэффициент. Коэффициенты были подобраны для обеспечения наивысшего среднего показателя F1 на обучающем наборе данных, рассматривая аннотацию набора данных как многоклассовую задачу, с использованием библиотеки Sklearn [63] реализации алгоритма минимизации Пауэлла [64].

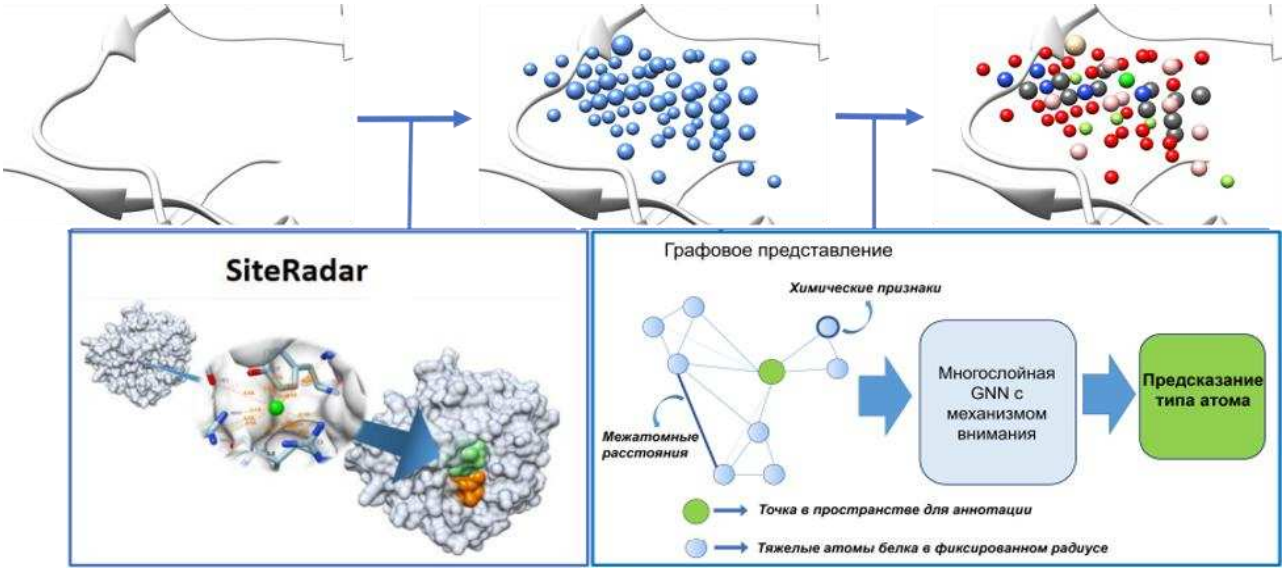


Рисунок 2.2. Схема генерации псевдолиганда с помощью обученных GNN (источник [43])

## 2.6 Тестирование моделей аннотации сайта и сравнение с существующими аналогами

Разработанные базовые модели были протестированы в нескольких экспериментах в двух конфигурациях: индивидуально (без применения балансирующих коэффициентов) и совместно (с применением коэффициентов). Индивидуальные тесты включали расчет различных метрик, таких как полнота, точность, F1 мера; результаты представлены на рисунках 2.3 и 2.4. Данные метрики использовались для оценки моделей в ходе процесса обучения и подбора итоговой архитектуры и подхода к обучению.

Однако, для оценки качества моделей в рамках решаемой задачи аннотации сайтов, данные результаты не являются репрезентативными, поскольку атом в позиции теоретически может выполнять несколько ролей, и модели могут испытывать трудности в предоставлении точной классификации конкретного атома. Поэтому, хотя способность таких бинарных моделей к классификации может быть низкой, в сочетании они способны обеспечивать необходимую информацию для принятия решений о вероятности размещения разных типов атомов в данной макромолекулярной среде. Подробнее данный аспект продемонстрирован в [43].

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Car | 0.75 | 0.39 | 0.51 | 8566 |
| O_a | 0.15 | 0.26 | 0.19 | 667 |
| Cs3 | 0.59 | 0.44 | 0.50 | 6092 |
| Nac | 0.30 | 0.32 | 0.31 | 1196 |
| Nd+ | 0.16 | 0.44 | 0.23 | 342 |
| Nd0 | 0.45 | 0.37 | 0.41 | 806 |
| Cs2 | 0.33 | 0.34 | 0.33 | 1610 |
| .=O | 0.52 | 0.50 | 0.51 | 1575 |
| Hal | 0.08 | 0.46 | 0.13 | 230 |
| O_d | 0.33 | 0.54 | 0.41 | 715 |
| Csp | 0.05 | 0.49 | 0.09 | 102 |
| Sul | 0.03 | 0.35 | 0.06 | 147 |
| SO2 | 0.09 | 0.56 | 0.15 | 153 |
| accuracy | | | 0.40 | 22201 |
| macro avg | 0.29 | 0.42 | 0.30 | 22201 |
| weighted avg | 0.57 | 0.40 | 0.45 | 22201 |

Рисунок 2.3. Показатели полноты, точности и F1 меры для обученных базовых бинарных моделей (источник [43])

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Car | 0.88 | 0.42 | 0.57 | 5661 |
| O_a | 0.16 | 0.44 | 0.23 | 84 |
| Cs3 | 0.58 | 0.43 | 0.49 | 1934 |
| Nac | 0.19 | 0.78 | 0.30 | 152 |
| Nd+ | 0.36 | 0.65 | 0.46 | 184 |
| Nd0 | 0.62 | 0.70 | 0.66 | 433 |
| Cs2 | 0.16 | 0.42 | 0.23 | 326 |
| .=O | 0.72 | 0.77 | 0.74 | 572 |
| Hal | 0.05 | 0.49 | 0.09 | 98 |
| O_d | 0.36 | 0.67 | 0.47 | 267 |
| Csp | 0.13 | 0.64 | 0.22 | 91 |
| Sul | 0.02 | 0.17 | 0.04 | 41 |
| SO2 | 0.24 | 0.64 | 0.35 | 160 |
| accuracy | | | 0.48 | 10003 |
| macro avg | 0.34 | 0.56 | 0.37 | 10003 |
| weighted avg | 0.71 | 0.48 | 0.53 | 10003 |

Рисунок 2.4. Показатели полноты, точности и F1 меры для обученных фармакофорных бинарных моделей (источник [43])

Для проверки точности классификации моделей при работе совместно, была оценена точность top-n (вероятность того, что при ранжировании полученных оценок каждой модели от большего к меньшему, оценка истинного класса окажется на n позиции или выше). Тестирование на валидационной выборке показало, что точность большинства базовых моделей выше 60% по top-n при n равному 4, однако определенные классы были более трудными для точного прогнозирования (Рисунок 2.5 А). Данные результаты могут быть объяснены природой самих классов и неоднозначности их возможного определения.

Например, положение и геометрия Nd0 (азот донора водородной связи), Nac (азот акцептора водородной связи) и .=O (кислород с гибридизацией sp2) могут быть строго определены образованием водородных связей, в то время такие классы как Sul (двухвалентная сера) или Csp (углерод с гибридизацией sd), не имеют такой геометрии, обусловленной взаимодействием. Также стоит отметить, что наиболее слабо предсказываемые классы, а именно Csp, Sul, SO2 и Hal, являются мало представленными в обучающей выборке. Эти группы составляли только 0,47% (9046 атомов), 0,66% (12 595 атомов), 0,76% (14 472 атома) и 1,39% (26 616 атомов) от набора данных соответственно, по сравнению с более часто встречаемыми типами атомов, такими как Car, который представляет 39,41% от набора данных (752 605 атомов), или Cs3 (25,54% от набора данных, 487 749 атомов). Фармакофорные модели показали более точные результаты по метрике top-n, как показано на Рисунке 2.5 Б. Данные результаты можно объяснить более строгими критериями отбора, примененными при сборе данных, которые позволили различить классы более точно благодаря исключению точек, расположенных в зонах высокой неопределенности из как обучающих, так и тестовых наборов данных.

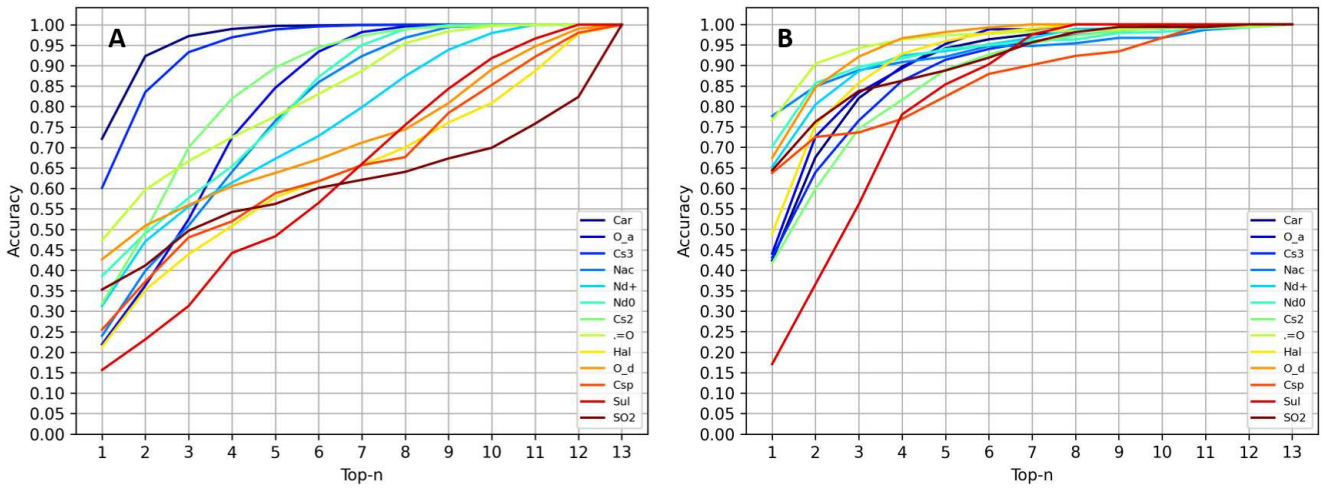


Рисунок 2.5. Top-n точность базовых (A) и фармакофорных (B) моделей в объединенном режиме

Также было проведено сравнение метода с существующими решениями. Для этого был использован независимый набор данных Astex Diverse Set [65], который не использовался для формирования обучающих и валидационных

данных моделей. Были выбраны две модели для сравнения, основанные на расчётах энергий, AutoSite [66] и AutoLigand [67]. Сравнение было проведено следующим образом: пространство было аннотировано с помощью алгоритма Skittles, затем предсказанные классы были сгруппированы в три более широкие категории: доноры водородных связей, акцепторы водородных связей и гидрофобные атомы (таблица 2.2). Группировка была проведена в связи с тем, что AutoSite и AutoLigand были разработаны для прогнозирования именно этих трех более широких классов. Следует отметить, что данные методы и их данные для обучения не находятся в открытом доступе, в связи с чем для оценки алгоритмов AutoSite и AutoLigand использовались опубликованные ими результаты на вышеупомянутом наборе данных. Разработанный алгоритм показал точность, близкую к точности AutoSite, и оба метода превзошли AutoLigand при прогнозировании атомов-доноров водородных связей, акцепторов водородных связей и гидрофобных атомов (Таблица 2.2). Доля корректно предсказанных атомов рассчитывалась как доля атомов лиганда в радиусе 2 Å, для которых существует хотя бы одна точка правильного класса (этот метрика была первоначально использована в статье об AutoSite).

Сравниваясь с AutoSite, созданный на основе обученных GNN алгоритм аннотации проявил более низкую точность в прогнозировании доноров водородных связей, но более высокую точность при предсказании акцепторов водородных связей и гидрофобных атомов. При этом и разработанная GNN и AutoSite превзошли по точности AutoLigand в рамках всех трех классов. В целом, проведенный анализ указывает на возможность применения алгоритма Skittles для определения ключевых свойств места связывания лиганд-белок.

| Типы атомов в методах сравнения | Типы атомов в алгоритме |
|---|---|
| HA | .=O, O_d, Nac, O_a |
| HD | Nd+, Nd0, Hal, O_d |
| HC | Car, Sul, Hal, Cs3, Csp |

Таблица 2.2. Соответствие типов атомов моделей сравнения и разработанной модели

| Модель | Доля корректно предсказанных атомов | | |
|---|---|---|---|
| | HD | HA | HC |
| Разработанная модель | 0.644 | 0.792 | 0.925 |
| AutoSite | 0.657 | 0.686 | 0.908 |
| AutoLigand | 0.201 | 0.447 | 0.717 |

Таблица 2.3. Сравнение точности предсказания типов атомов

В [43] также была продемонстрирована способность различать типы атомов с одинаковыми фармакофорными свойствами. В результатах проведенного эксперимента было показано, что модели Nac и O_a успешно идентифицировали 88% и 90% атомов из тестовой выборки соответственно.

Также была протестирована применимость оценок, получаемых разработанными GNN, для обучения иных моделей машинного обучения, решающих задачи классификации сайтов связывания и оценки аффинности малых молекул.

## 2.7 Применение разработанных моделей разметки сайта связывания для создания алгоритма классификации мест связывания лиганд-белковых комплексов

Чтобы продемонстрировать практическое применение разработанного алгоритма, он был применен к задаче классификации мест связывания лиганд-белковых комплексов с использованием моделей GNN.

Классификация мест связывания лиганд-белковых комплексов является важной задачей в структурной биологии и разработке лекарств, поскольку она может помочь понять функции белков или разработать синтетические модуляторы, имитирующие режим связывания субстратов. Эта классификация может быть основана на различных критериях, таких как химическая структура природного лиганда места связывания, способность к медиации аллостерической регуляции или возможность размещения ковалентных связывающих агентов.

Упомянутые свойства важны, поскольку они могут указать на новые пути фармакологического вмешательства.

Разработанная модель классификации сайтов связывания с лиганд-белковых комплексов была создана соавторами [43], и кратко представлена в данной работе для демонстрации применения моделей аннотации пространства для обучения иных моделей. Она использует обученные бинарные GNN классификаторы, использующие сгенерированные алгоритмом Skittles псевдолиганды как признаковое описание. Конкретно, объектом данных был полносвязный ненаправленный граф точек решетки, которая была сгенерирована по сайту связывания с применением вышеописанного алгоритма SiteRadar, и размечена вышеописанными GNN моделями. Признаками вершин в данном графе были значения предсказаний для всех типов атомов для данной точки, а признаки ребер были расстояниями между точками решетки. Визуальное представление использования признаков, сгенерированных Skittles, для классификации мест связывания лиганд-белок предоставлено на рисунке 2.6.

Разработанные модели были обучены для предсказания пяти химических свойств потенциальных субстратов: нуклеотиды/АТФ, углеводы, гем, пептиды и нуклеиновые кислоты, а также двух дополнительных свойств, таких как способность к аллостерическому модулированию и способность принимать ковалентные лиганды. Подробно обучение данных моделей, а также их архитектура описаны в [43].

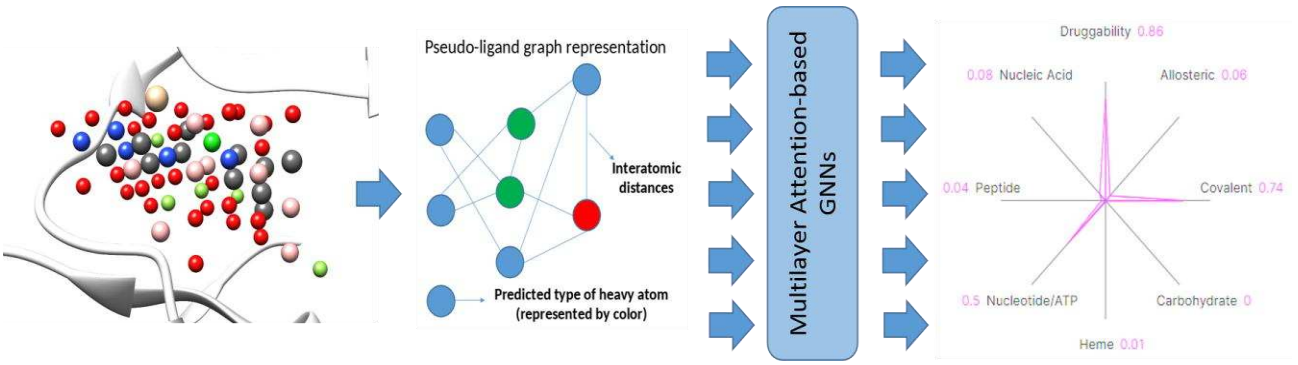


Рисунок 2.6. Упрощенная схема использования представления, сгенерированного Skittles, для классификации мест связывания лиганд-белок (источник [43])

### 2.7.1 Оценка модели классификации

Для оценки моделей классификации, обученных на основании сгенерированных Skittles псевдолигандов, были проведены несколько проверок, включая независимую оценку каждой модели и анализ совместного использования моделей для прогнозирования химической природы субстратов.

Точность классификации моделей была оценена на независимой выборке, подобранной таким образом, чтобы отличаться от обучающих данных по подобию секвенции белков (подробнее в [43]), в качестве метрики был использована площадь под кривой рабочей характеристики приемника (ROC AUC). Результаты оценки представлены в таблице 2.4

Сравнение с существующими методами является одним из ключевых способов оценки разработанного алгоритма, однако, в данной области прямое сравнение созданных классификаторов было затруднено в связи с отсутствием доступа к исходному коду и\или тестовым данным уже опубликованных ML моделей для классификации мест связывания лиганд-белок. В связи с этим сравнение было проведено с результатами, опубликованными авторами аналогичных алгоритмов на разработанных ими тестовых наборах по сравнимым классам – в таблице 2.4, помимо результатов разработанных классификаторов, также представлены опубликованные результаты алгоритмов BindSiteS-CNN [11], DeepDrug3D [12] и BionoiNet [13]. Данное сравнение было проведено соавторами в [43], и представлено в данной работе для демонстрации применимости моделей аннотации, упомянутых ранее. Как можно видеть, алгоритм BionoiNet продемонстрировал более высокую точность по двум из рассматриваемых классов, кроме того, разработанная модель показала самый низкий ROC AUC по классу нуклеотиды/АТФ. В связи с этим было проведено дополнительно тестирование - разработанные модели классификации были применены к тестовому набору данных BionoiNet, так как он был в открытом доступе.

| | Разработанный алгоритм (собственный тест) | BindSiteS-CNN (результаты из статьи) | DeepDrug3D (результаты из статьи) | BionoiNet (результаты из статьи) |
|---|---|---|---|---|
| Нуклеотиды/АТФ | 0,799 | 0,89 | 0,95 | 0,96 |
| Углеводы | 0,915 | - | - | - |
| Гем | 0,921 | 0,74 | 0,82 | 0,935 |
| Пептиды | 0,723 | - | - | - |
| Нуклеиновые кислоты | 0,693 | - | - | - |
| Аллостерический сайт | 0,897 | - | - | - |
| Ковалентный сайт | 0,606 | - | - | - |

Таблица 2.4. ROC AUC предсказания свойств сайта связывания на независимом тесте, а также ROC AUC, опубликованные для аналогичных методов по определенным классам (источник [43])

На тестовом наборе BionoiNet разработанные модели показали более высокие значения ROC AUC, чем на собственном тесте - метрика ROC AUC в случае классификации белков, содержащих Гем, составила 0,957, а в случае белков, взаимодействующих с нуклеотидами, - 0,966, по сравнению со значениями 0,935 и 0,960 соответственно для BionoiNet. Для класса Гем, на тестовом наборе BionoiNet, была зафиксирована схожая по результатам на собственном тесте метрика ROC AUC; он была сопоставима со значением ROC AUC BionoiNet (подробнее в [43]). Как можно видеть, на тестовом наборе BionoiNet, разработанный классификатор показал более высокое значение ROC AUC, чем на собственном тесте, и превзошел ROC AUC, опубликованный для методов BindSiteS-CNN и DeepDrug3D. Это может быть объяснено более высокой степенью сложности собственного сформированного тестового набора по сравнению с датасетом BionoiNet.

Для анализа эффекта совместного использования моделей для предсказания химической природы предполагаемых субстратов пять моделей для определения природы субстрата были применены к тестовому набору данных, и класс с

наивысшей вероятностью был выбран в качестве прогнозируемого класса для данного сайта связывания. Как показано в матрице ошибок (Рисунок 2.7), большинство классов сайтов связывания были правильно предсказаны. Более высокая степень ошибок в классификации сайтов связывания нуклеиновых кислот (были предсказаны как сайты связывания нуклеотидов в 22% случаев), и сайта связывания пептидов (были классифицированы как сайты связывания нуклеиновых кислот в 24% случаев), могут быть объяснены структурными особенностями данных классов (подробнее в [43]).

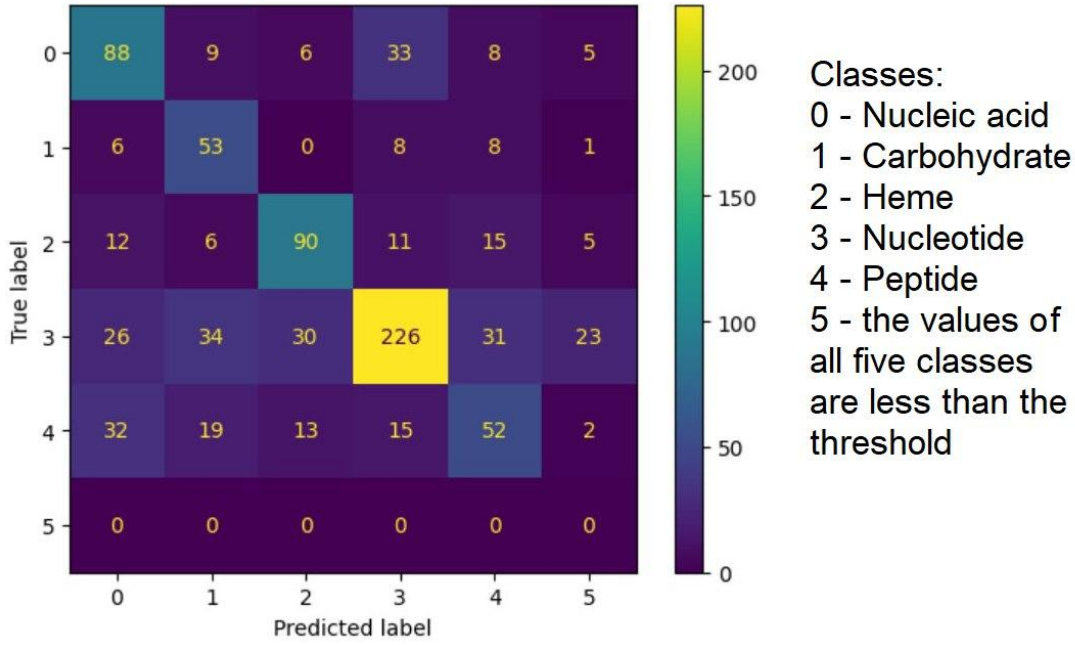


Рисунок 2.7. Матрица ошибок объединенного использования моделей для предсказания химической природы предполагаемого вещества. Источник [43]

## 2.8. Применение разработанных моделей разметки сайта связывания для создания алгоритма предсказания аффинности лиганд-белок

Важной задачей в сфере разработки лекарств является оценка аффинности конкретной малой молекулы к белку-мишени. Сложность межмолекулярных взаимодействий, множество факторов, которые необходимо учитывать, а также наличие несоответствий зафиксированных значений аффинности в доступных данных [68] затрудняют обучения точных моделей, основанных на машинном обучении. Для решения данной задачи существует ряд алгоритмов, не

использующих ML, и реализованных в программном обеспечении, в частности MOE (https://www.chemcomp.com/en/Products.htm), Glide Schrodinger [69], Autodock и его модификаций [70], LeadFinger [71]. С растущей популярностью ML в различных сферах, в настоящее время разрабатываются многочисленные решения на основе машинного обучения для более быстрого и точного предсказания аффинности малых молекул. Недавние примеры таких алгоритмов включают InteractionGraphNet (IGN) [14] и graphLambda [15].

В рамках проведенного научного исследования, продемонстрировано применение обученных GNN, размечающих пространство сайта связывания, для создания и обучения GNN метода предсказания аффинности малой молекулы по отношению к заданному белковому окружению. В разработанном методе, входным объектом является граф малой молекулы, построенный по ее трехмерной структуре, где признаками на вершинах являются предсказания GNN моделей разметки пространства в рамках каждого класса и производные от них признаки (подробнее в [43], а также в разделе 2.8.1), дополненные дескрипторами, характеризующими взаимодействий лиганд-белок, такие как заполнение кармана и количество межмолекулярных контактов (например, водородных связей, солевых мостков, стэкинга, пи-катионных и т. д., подробнее в [43]). Визуальное представление алгоритма показано на рисунке 2.8.

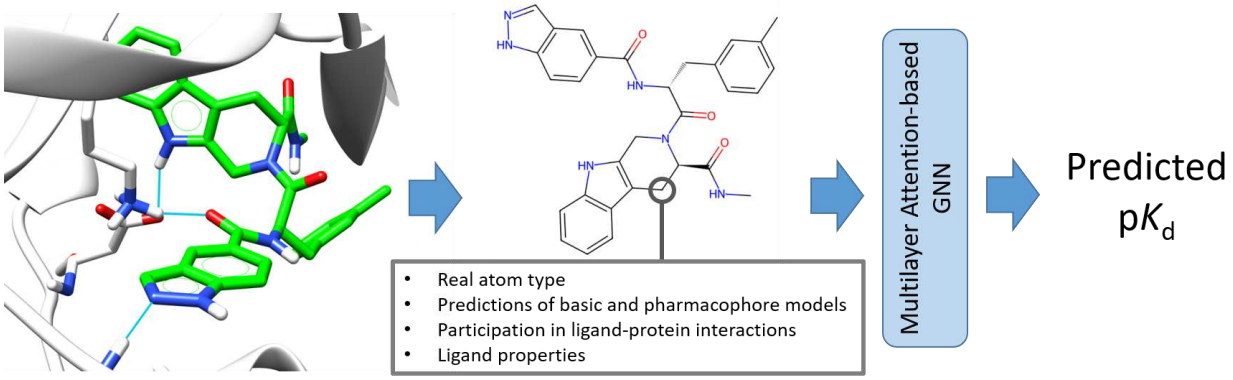


Рисунок 2.8. Упрощенная схема использования представления, сгенерированного Skittles, для предсказания аффинности лиганд-белок. Источник [43]

### 2.8.1 Подготовка обучающих данных GNN для предсказания аффинности лиганд-белок

Выборка данных, использованная для обучения и тестирования модели, была сформирована на основании базы PDBBind v.2016. 19120 комплексов были отфильтрованы на основании экспертно выбранных критериев (подробнее в [43]), в результате чего были сохранены 8093 комплекса. Данные комплексы были дополнительно предобработаны, в частности, было проведено протонирование комплекса с помощью AutoDock Tools, и сохранялась только одна молекула лиганда, подробнее в [43].

С целью формирования валидационной выборки (для контроля переобучения) и выборки независимого тестирования, для комплексов были рассчитаны парные метрики сходства белков (с применением матрицы BLOSUM62) и метрики сходства лигандов (подобие Morgan-фингерпринтов). На основе этих метрик 10% самых непохожих комплексов были случайно разделены на тестовый и валидационный наборы, содержащие 404 и 405 комплексов соответственно, а обучающий набор, после удаления этих 10%, состоял из 7284 комплексов.

Как было описано ранее, входным объектом для разработанной модели предсказания аффинности является графовое представление лиганда, где признаками узлов являются различные взаимодействия между лигандом и его белковым окружением, а также атрибуты самой малой молекулы. Для каждого атома лиганда, кроме фторов и водородов, были собраны следующие признаки (разработанные экспертами соавторами статьи [43]):

1. Тип атома (согласно Таблице 2.1), one-hot кодирование (метод представления категориальных данных в бинарном формате) — 13 столбцов, соответствующих 13 типам атомов.

2. Оценивается ли атом GNN моделями разметки (атомы, удаленные более чем на 5 Å от белка, не оцениваются — 1 столбец).

3. Оценки базовой и фармакофорной моделей, соответствующих типу атома, определенного в 1 (2 столбца).

4. Оценки базовой и фармокофорной моделей для каждого типа атома в данной точке пространства (26 столбцов: по 2 для каждого из 13 типов атомов).

5. Взаимодействия лиганд-белок, в которых участвует атом (водородные связи, гидрофобные, пи-катионные и т.д. — 11 столбцов).

6. Является ли атом частью цикла (1 столбец).

7. Соответствуют ли типы атомов, предсказанные базовыми и фармакофорными моделям как наилучшие, истинному типу атома в данной точке пространства (2 столбца).

8. One-hot закодированные типы атомов, предсказанные базовыми и фармакофорными моделями как наилучшие — 26 столбцов).

9. Численные оценки базовой и фармакофорной моделей, соответствующие предсказанному типу (2 столбца).

Дополнительно, каждый атом также получал следующие признаки, рассчитанные для всей молекулы: 1) количество тяжелых атомов в молекуле; 2) метрика занятости (процент объема места связывания, занимаемого лигандом); 3) метрика интеркаляции (процент атомов лиганда внутри места связывания белка). Подробнее данные метрики и их алгоритм расчета описаны в [43]. Все числовые признаки были масштабированы. Суммарно каждая вершина (атом лиганда) была представлена 87 признаками.

Положение атомов относительно друг друга в 3D-пространстве представляется парными расстояниями, которые записываются как признаки ребер. Данные ребра создавались только между атомами лиганда на расстоянии не более 7 Å. Дополнительно на ребрах хранится признак наличия ковалентной связи между атомами (0 для отсутствия связи, 1 для присутствия связи без учета кратности связи). Целевой предсказываемой величиной являлась активность лиганда против белка (отрицательный десятичный логарифм Kd, Ki или IC50).

### 2.8.2 Архитектура обученных GNN для предсказания аффинности лиганд-белок

Разработанная модель построена на GNN, ее архитектура представлена на рисунке 2.9. Как и в случае архитектур GNN по поиску сайтов связывания (глава 1) и разметке сайтов связывания, признаки на ребрах входного графа преобразовывались отдельным блоком в векторные представления, используемые последующими GNN слоями. Этот блок состоит из слоя гауссового размытия в диапазоне от 0 до 7,01 (в соответствии с максимально возможным расстоянием между двумя вершинами в графе) с использованием 50 гауссиан линейного слоя с 32 нейронами, и функцией активации leaky ReLU. Результирующий вектор размерностью 32, объединяется с признаком присутствия/отсутствия связи, в результате чего получается вектор размерностью 33. 87 признаков на узлах графа и вышеупомянутое векторное представление расстояний, являющееся признаком ребер, обрабатываются блоком из 6 GNN слоев с механизмом многоголового внимания с 512 нейронами и 16 головами внимания (использовался тот же графовый оператор, что и в других обученных GNN), с использованием батч нормализации. Выходы средних четырех слоев передаются через дополнительные слои dropout со значением 0,2 для уменьшения переобучения. Последний слой блока имеет 256 нейронов, на основании его выхода создается два вектора объединения сигналов по вершинам – объединения сигналов методом усреднения и методом взятия максимального значения. Эти векторы объединяются с 9 признаками взаимодействия (исключены стэкование и пи-катионные взаимодействия), усредненными по графу. Результирующий вектор передается в блок из 4 линейных слоев (с пакетной нормализацией и dropout со значением 0,2), который выдает окончательное скалярное значение предсказания активности. Leaky ReLU-активация с параметром отрицательного наклона 0,5 используется в блоках графовых и линейных слоев.

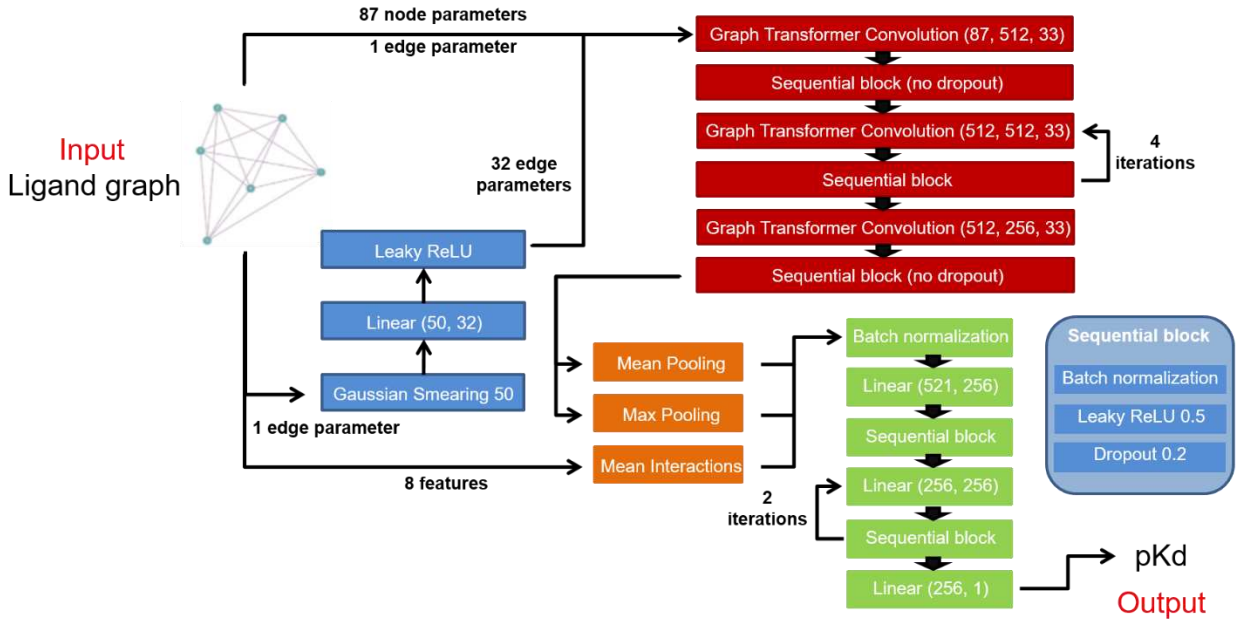


Рисунок 2.9 Схема GNN предсказания аффинности малой молекулы

### 2.8.3 Обучение GNN для предсказания аффинности лиганд-белок

Модель обучалась батчами по 512 элементов, где каждый элемент представлял один молекулярный граф, соответствующий одному комплексу белок-лиганд. Элементы в батчах случайно перемешивались на каждой эпохе обучения. В качестве функции потерь использовалась функция MSELoss (среднеквадратичная ошибка), а для оптимизации применялся алгоритм Adam со скоростью обучения 0,001. Обучение проводилось в течении 400 эпох, и модель сохранялась при минимальном значении потерь на валидационном наборе данных; если значение потерь на валидационном наборе не улучшалось в течение 50 эпох, обучение останавливалось преждевременно. На GPU NVIDIA A100-SXM4 с 80 ГБ видеопамяти обучение занимало примерно 10 минут при ранней остановке; для полного цикла требовалось 30 минут. Процесс обучения повторялся три раза, чтобы обеспечить его стабильность и разумную воспроизводимость полученной нейронной сети.

### 2.8.4 Тестирование модели для предсказания аффинности лиганд-белок

Было проведено сравнение разработанной модели с двумя другими алгоритмами, осуществляющими предсказание аффинности малой молекулы к

белку: IGN, передовым алгоритмом, основанным на ML и GNN, и традиционным не ML методом AutoDock Vina [42]. Оценка моделей проводилась по двум метрикам - коэффициент корреляции Пирсона предсказаний с истинными значениями и их средняя абсолютная ошибка (MAE). Датасетом для сравнения служил созданный независимый датасет (пункт 2.8.1).

Экспериментальные результаты показали превосходство ML методов над AutoDock Vina как в коэффициенте корреляции Пирсона, так и в средней абсолютной ошибке (p-значение $5{,}861 \times 10^{-25}$ и $4{,}048 \times 10^{-22}$ для разработанного метода и IGN соответственно), как показано на рисунке 2.10. Сравнивая разработанный метод и IGN, было обнаружено значимое различие по коэффициенту корреляции Пирсона (0,7408 для разработанной модели и 0,7486 для IGN), но не было обнаружено статистической значимости по метрике MAE (0,948 для разработанной модели и 1,0 для IGN, p-значение 0,529). Следует отметить, что авторы модели IGN не предоставили идентификаторы комплексов, которые были использованы в ходе обучения (была предоставлена информация только об используемой в качестве основы базе данных и ее предобработке), в связи с чем не было возможности оценить степень схожести комплексов использованного тестового набора с обучающим набором IGN. Иными словами, модель IGN может показывать переоцененные результаты на независимом тесте из-за потенциального наличия схожих с обучением комплексов. Резюмируя, IGN и разработанный метод показали значительное улучшение в рамках использованных метрик над AutoDock Vina, в то время как превосходство одного из проанализированных GNN-основанных подходов остается под вопросом.

Как было сказано ранее, хотя разработанная модель в значительной степени использует в качестве признаков оценки 26 обученных моделей (13 базовых и 13 фармакофорных) разметки пространства, также используется и признаки, основанные на описании взаимодействий малой молекул с окружающими атомами белка. Для верификации вклада признаков, основанных на предсказаниях обученных моделей (Skittles-дескрипторы), были проведены два дополнительных эксперимента, в рамках которых были обучены дополнительные GNN модели в

следующих условиях: (1) предсказание сродства лиганд-белок на основании всех признаков, кроме Skittles-дескрипторов; и (2) предсказание сродства лиганд-белок только с помощью Skittles-дескрипторов.

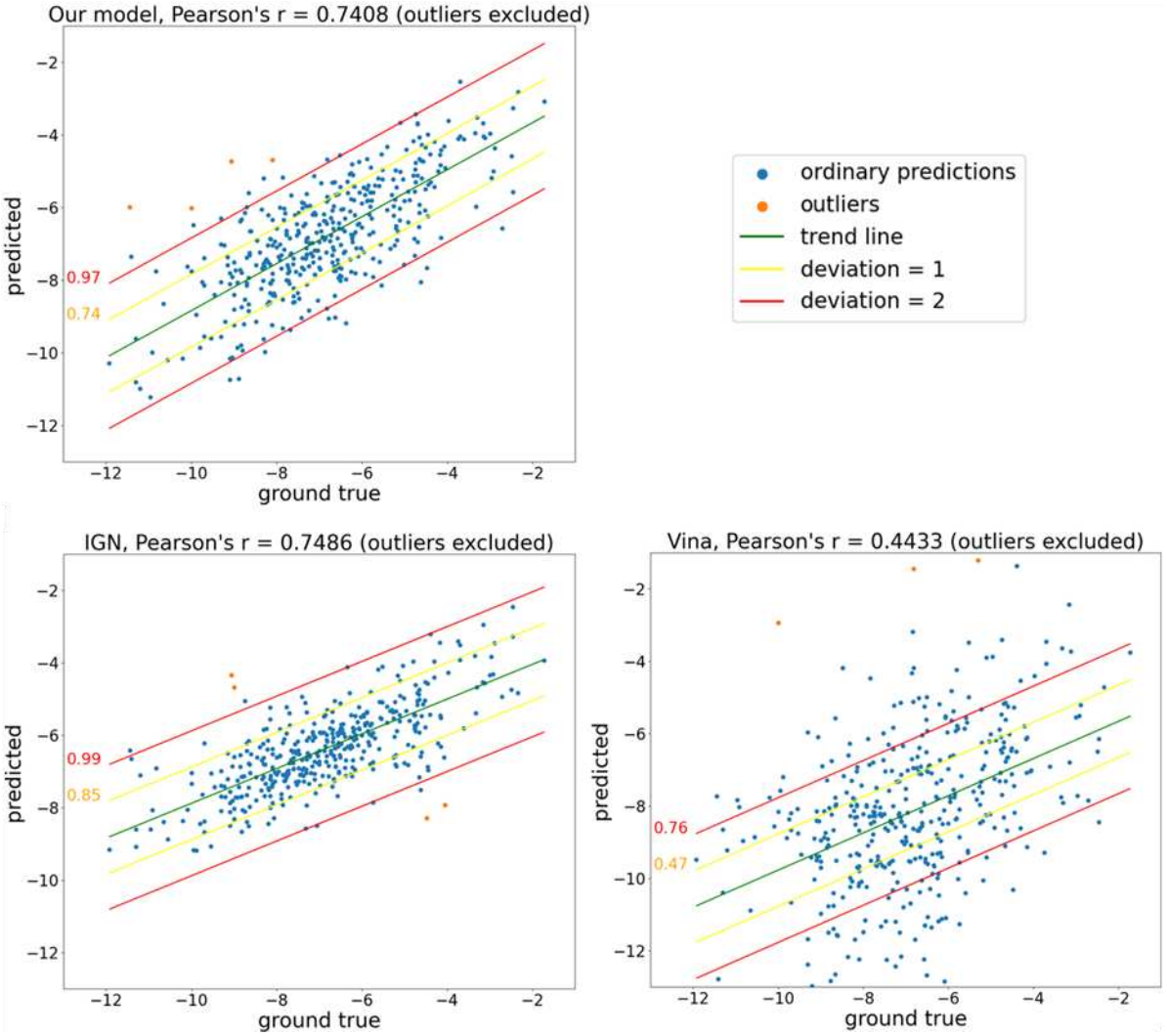


Рисунок 2.10. Сравнение различных методов для предсказания аффинности лиганда к белку (источник [43])

Под Skittles-дескрипторами здесь подразумеваются следующие признаки:

1. Тип атома (согласно Таблице 2.1), one-hot кодирование (метод представления категориальных данных в бинарном формате) — 13 столбцов, соответствующих 13 типам атомов.

2. Оценивается ли атом GNN моделями разметки (атомы, удаленные более чем на 5 Å от белка, не оцениваются — 1 столбец).

3. Оценки базовой и фармакофорной моделей, соответствующих типу атома, определенного в 1 (2 столбца).

4. Оценки базовой и фармокофорной моделей для каждого типа атома в данной точке пространства (26 столбцов: по 2 для каждого из 13 типов атомов).

5. Соответствуют ли типы атомов, предсказанные базовыми и фармакофорными моделям как наилучшие, истинному типу атома в данной точке пространства (2 столбца).

6. Является ли атом частью цикла (1 столбец).

7. One-hot закодированные типы атомов, предсказанные базовыми и фармакофорными моделями как наилучшие — 26 столбцов).

8. Численные оценки базовой и фармакофорной моделей, соответствующие предсказанному типу (2 столбца).

За исключением изменения размерности входного слоя в соответствии с количеством признаков на вершинах, модели обладали той же архитектурой, и были обучены таким же образом, что и вышеописанная модель предсказания аффинности.

По результатам эксперимента обе модели продемонстрировали меньшую точность в рамках коэффициента корреляции Пирсона, но сопоставимую точность в рамках метрики MAE по сравнению с моделью, использующей все дескрипторы в комбинации. В частности, модель без Skittles-дескрипторов дала коэффициент корреляции Пирсона 0,7188 и MAE 0,9571 (p-значение 0,994). Аналогично, модель, обученная исключительно на Skittles-основанных дескрипторах, показала коэффициент корреляции Пирсона 0,7325 и MAE 0,9575 (p-значение 0,725). Следует отметить, что модель, основанная исключительно на Skittles-дескрипторах, превзошла модель без них в терминах коэффициента корреляции Пирсона, однако не было обнаружено статистически значимой разницы в MAE (p-значение 0,7383). Этот эксперимент продемонстрировал, что Skittles-дескрипторы могут служить ценным альтернативным вариантом в предсказании аффинности лиганд-белок, и сочетание Skittles и не Skittles признаков имеет потенциал повысить корреляцию между предсказанным и фактическим сходством, хотя оно не существенно влияет на абсолютную ошибку точных значений предсказания.

В рамках проведенного тестирования, статистическая значимость рассчитывалась с помощью теста Манна-Уитни, примененного к значениям MAE и коэффициентам корреляции Пирсона. Нормальность распределения данных анализировалась с помощью теста Шапиро-Уилка.

## 2.9 Выводы

В данной главе было рассмотрено применение GNN для разметки пространства связывания белка. Для 13 разработанных экспертами классов, описывающих те или иные свойства пространства, были обучены 13 моделей на двух вариантах разметки обучающей выборки (суммарно 26 моделей). Была рассмотрена точность получаемых предсказаний, как отдельно взятых классификаторов, так и при совместной работе в случае необходимости выбора одного, наиболее подходящего класса для конкретной точки в пространстве. Также было проведено сравнение обученных моделей с существующими аналогами.

Были продемонстрированы возможные сферы применения данных моделей, а именно: генерация псевдолигандов в рамках разработанного и описанного в [43] алгоритма, дополнительно использующего модель поиска и создания объемных сайтов связывания, рассмотренного в главе 1; классификации сайтов связывания; предсказания аффинности малой молекулы к заданному белку-мишени.

В рамках задачи классификации, было показано, что с применением предсказаний разработанных GNN моделей аннотации пространства как признаков и алгоритма генерации псевдолигандов, возможно обучение нейросетевых моделей, основанных на GNN, способных предсказывать различные свойства сайта связывания белка. Проведенные в [43] эксперименты показали, что разработанные модели обладают схожей или превосходящей точностью в сравнении с рассмотренными современными аналогами.

В рамках задачи предсказания аффинности малой молекулы была продемонстрирована возможность применения обученных GNN моделей аннотации пространства для генерации признакового описания, применимого для

обучения другой GNN модели предсказания аффинности малой молекулы к заданному белковому окружению. Проведенные эксперименты показали, что разработанная модель способна предсказывать аффинность на уровне точности, сравнимой с существующим машинно обученным аналогом IGN, и существенно превосходящей популярно используемый не-ML подход AutoDock Vina.

Эти результаты подчеркивают значимость проведенного исследования и его практическую актуальность в вышеупомянутых научных областях.

# Глава 3. Разработка моделей оценки свойств малой молекулы

В данной главе рассматривается применение GNN для работы с молекулярной структурой, без использования информации о ее макромолекулярном белковом окружении, для анализа ее свойств, на примере задач предсказания потенциальных белковых мишеней молекулы и анализа растворимости. Проводится сравнение GNN и графового представления данных с другими ML архитектурами и представлениями молекулярных структур.

В 3.1 представляется введение в предметную область и ее значимость, проводится верхнеуровневый обзор формата хранения данных, которые могут быть использованы для обучения ML моделей, рассмотрены подходы к сбору их признакового описания. В 3.2 приводится обзор предметной области в рамках задачи предсказания мишеней и существующих решений. В 3.3 приведена постановка задачи, описание входных данных, применяемых методов ML. В 3.4 детально описывается источник и предобработка данных, формирование обучающих, валидационных и тестовых выборок, проводится анализ распределения предсказываемых меток объектов. В 3.5 приводится информация о формировании представлений данных, необходимых для обучения рассматриваемых методов. В 3.6 описываются архитектуры подготовленных в рамках исследования моделей. В 3.7 представлен метод обучения рассматриваемых моделей. В 3.8 представлены результаты тестирования моделей, их сравнение. В 3.9 представлен эксперимент по применению GNN модели для аннотации крупной библиотеки структур, чьи истинные метки не были известны. В 3.10 приводится описание предметной области задачи по предсказанию растворимости малых молекул. В 3.11 представлена постановка задачи, а также верхнеуровневое описание применяемых для ее решения моделей. В 3.12 описываются использованные для обучения и тестирования данные, их предобработка и анализ. В 3.13 представлена архитектура и подход к обучению рассматриваемых моделей. В 3.14 описаны результаты тестирования на сформированном тестовом наборе данных. В 3.15 описаны результаты

тестирования, проведенные на внешнем наборе данных. В 3.16 представлено заключение по данной главе.

## 3.1. Введение

Важным элементом процесса разработки лекарств является оценка свойств виртуально созданной малой молекулы. В частности, исследователей области медицинской химии зачастую интересует, может ли малая молекула ингибировать не только целевую мишень, но и иные белки, что может влиять на эффективность потенциального лекарства, созданного с применением данной молекулы, ее побочные эффекты. Также анализ покрываемых малой молекулой мишеней осуществляется с целью предварительного скрининга крупных библиотек структур для поиска потенциальных веществ для дальнейшей разработки и анализа под интересуемую мишень.

Помимо этого, анализируются различные физико-химические свойства молекулы.

Одним из важных физико-химических параметров, которые учитываются при разработки лекарственных средств, является водная растворимость. Этот фактор влияет на вариабельность пероральной биодоступности, может приводить к плохому всасыванию и, как следствие, снижению эффективности препарата. Кроме того, недостаточная растворимость усложняет оценку активности соединения и нередко вызывает нежелательные побочные реакции [72]. В связи с этим фармацевтические компании уделяют особое внимание растворимости на этапах оптимизации лекарственных препаратов [73].

Существуют обширные базы данных, содержащие сведения о физико-химических характеристиках молекул, однако большинство таких ресурсов представляют молекулы в виде упрощенной молекулярной входной строковой записи (SMILES), которая является распространенным форматом для описания молекулярных структур. Данная форма записи дает возможность для извлечения различных описательных признаков, а также представления молекулы в виде 2D графа, однако не содержит информации о ее истинной трехмерной конформации.

Использование методов, работающих с трехмерными структурами, остается возможным, однако требует применения генерации 3D конформации по SMILES записи, не гарантируя при этом воспроизведение ее истинной 3D конформации.

В данной главе будет рассмотрено применение GNN к задаче предсказания потенциальных киназных мишеней, с применением ансамблей сгенерированных 3D конформаций на основе SMILES малой молекулы, а также применение GNN к задаче предсказания растворимости, но используя 2D графовое представление. Для оценки результативности GNN методов, были обучены дополнительные ML модели, использующие иные дескрипторы, извлеченные из SMILES записи молекулы.

## 3.2. Предсказание потенциальных мишеней малой молекулы: обзор предметной области

Киназы играют важную роль во многих клеточных процессах, а их дисрегуляция связана с определенными заболеваниями, такими как рак. Хотя этот класс белков был предложен для терапевтического вмешательства несколько десятилетий назад, сохраняется сильный интерес к разработке новых ингибиторов киназ [74-76]. Несколько препаратов, нацеленных на киназы, уже были одобрены FDA, однако управление селективностью ингибиторов киназ представляет собой сложную задачу из-за высокой схожести в структуре их каталитических участков [77]. Недавние исследования обнаружили ранее неизвестные мишени даже для уже установленных ингибиторов киназ, демонстрируя, что истинный ландшафт потенциальных фармакологических мишеней еще не полностью охарактеризован [78]. С другой стороны, способность воздействовать на несколько белков также может быть желательной, поскольку препараты с широким профилем ингибирования могут иметь потенциал достижения более сильного фармакологического эффекта. Хотя пациенты с раком могут развивать устойчивость к селективным препаратам из-за активации альтернативных или компенсаторных путей, пан-ингибиторы киназ менее подвержены этому явлению [79].

Традиционный подход к профилированию селективности включает в себя тестирование in vitro, которое является как дорогостоящим, так и длительным. Поэтому активно разрабатываются вычислительные методы для выявления хитов и оценки селективности, особенно те, которые используют искусственный интеллект (ИИ), чтобы оптимизировать процесс путем сужения разумного химического пространства. Как было упомянуто ранее, большинство публично доступных наборов данных предоставляют информацию о соединениях в формате SMILES, но не содержат деталей об их активных 3D-конформациях. В результате многие существующие модели полагаются на дескрипторы, полученные из SMILES, иногда включая данные о целевых белках. Были реализованы различные методы, не основанные на нейронных сетях, включая метод Kronecker RLS [80], метод на основе градиентного бустинга под названием SimBoost [16], подход на основе метода опорных векторов (SVM) [17] и метод, использующий фармакофорные модели случайного леса (RF) [18]. В недавнем исследовании были разработаны 32 модели классификации с использованием XGBoost, RF, SVM и деревьев решений для прогнозирования ингибиторов ALK [19]. Кроме того, были созданы модели нейронных сетей с различными архитектурами для задачи профилирования молекул: модели глубоких нейронных сетей [20,21], модели свёрточных нейронных сетей (CNN) [22,23], модели графовых нейронных сетей (GNN) [24,25] и модели на основе трансформеров [26- 28]. Хотя большинство существующих решений используют 2D-данные соединений, определенные методы, такие как AikPro [29], строят ансамбли 3D-конформаций для извлечения дополнительных дескрипторов, основанных на расположении атомов молекулы в трехмерном пространстве. Фармакофорное описание малых молекул также было использовано для виртуального скрининга с помощью Psearch [30].

В рамках задачи оценки возможности связывания лигандов с белками доступны два варианта: предсказание точной аффинности препарата к мишени (DTA) или оценка взаимодействия препарата с мишенью (DTI), которая классифицирует соединения как активные или неактивные по отношению к

данной мишени. Известны определенные недостатки методов DTI, включая невозможность ранжирования молекул и необходимость выбора порога активности, который может варьироваться для разных мишеней [28]. Хотя эти проблемы могут быть смягчены путем ранжирования молекул на основе уверенности классификации модели и использования разных порогов для разных мишеней, методы DTA характеризуются несоответствиями в доступных данных относительно значений аффинности соединений [68]. Это несоответствие представляет собой существенный недостаток для любого обучаемого метода, направленного на предсказание точных значений DTA. В рамках данной работы, был применен DTI подход с целью снижения возможного шума в изначальных данных при использовании конкретных значений аффинности.

## 3.3. Предсказание потенциальных мишеней малой молекулы: постановка задачи

Задача может быть описана образом, схожим с описанием в главах 1 и 2: дан набор данных, как представлено в формуле (1), где $x_k$ — это вектор признакового описания объекта, а $y_k \in R$ — соотвествующий ответ по данному объекту. При этом примеры $(x_k, y_k)$ считаются независимыми друг от друга. Для прогнозирования $y$ строится модель $F: R^m \to R$ которая обучается минимизировать заданную функцию потерь, как это показано в формуле (2), где пары *(x,y)* являются случайно и независимо выбранными примерами из обучающего множества.

В рамках данной задачи, набор данных содержит профили активности малых молекул, представленных в SMILES записи, для 75 киназных мишеней, где у каждой молекулы может быть больше одной мишени. Таким образом решается задача классификация по многим меткам, для чего была выбрана функция потери бинарной кросс-энтропии. Данная функция может быть представлена следующим образом:

$$l(x, y) = -w_n\left[y_n \times \log\left(\sigma(x_n)\right) + (1 - y_n) \times \log\left(1 - \sigma(x_n)\right)\right] \quad (6)$$

где $x_n$ это логит для n-го примера по определённому классу, $\sigma(x_n)$ это сигмоидная функция вида:

$$\sigma(x_n) = \frac{1}{1 + e^{(-x_n)}} \quad (7)$$

где $y_n$ это целевая метка для n-го примера по этому классу (0 или 1), а $w_n$ это необязательный вес для n-го примера (в данной задаче не используется). Для общего случая батча размером $N$ с $C$ классами функция потери может быть представлена следующим образом:

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} w_{n,c}\left[y_{n,c} \times \log\left(\sigma\left(x_{n,c}\right)\right) + \left(1 - y_{n,c}\right) \times \log\left(1 - \sigma\left(x_{n,c}\right)\right)\right] \quad (8)$$

где $x_{n,c}$ это логит для n-го примера класса $c$, $y_{n,c}$ это целевая метка для примера $n$ класса $c$, а $w_{n,c}$ это необязательный вес для n-го примера класса $c$.

Признаковое описание объекта $x_k$ зависит от конкретной архитектуры, и в рамках решения задачи были спроектированы и протестированы четыре архитектуры - GNN модель, использующая комбинацию графового представления и таблицы глобальных дескрипторов молекулы, две модели, использующие только представление молекулы в виде таблицы извлеченных дескрипторов, и модели, рассматривающей SMILES запись как символьную последовательность. Три не GNN модели (далее также именуемые как модели сравнения) были обучены с целью выявления наличия либо отсутствия качественного прироста точности предсказаний при использовании GNN и графового представления, по сравнению с иными подходами, при обучении на и применении к идентичным исходным данным (молекулярным структурам).

Исследование было сфокусировано на обучении модели GNN, дополненной данными 3D-фармакофорного представления молекулы. Недавние исследования продемонстрировали значительные улучшения качества моделей при применении GNN к 3D-объектам в области компьютерного проектирования лекарств и структурной биологии [14,42,15], в связи с чем, ввиду отсутствия информации об истинной 3D структуре, было применено 3D моделирование возможных конформаций. 3D-фармакофорное представление является ансамблем 3D-

фармакофоров, которые включают следующие типы: доноры водородных связей (HBD), акцепторы водородных связей (HBA), сопряженные системы (ароматические кольца и атомы углерода в гибридизации $sp^2$ вне любого кольца), гидрофобные группы, группы с положительным зарядом, группы с отрицательным зарядом. Фармакофорные типы были отобраны экспертно. Для генерации конформаций, использовался разработанный экспертами алгоритм.

Одной из моделей сравнения является модель градиентного бустинга с применением программной библиотеки CatBoost [81]. Алгоритм градиентного бустинга является эффективным способом обработки признаков, представленных в табличном виде. Процесс градиентного бустинга представляет собой итеративное построение серии приближений, где каждое новое приближение формируется путём аддитивного добавления к предыдущему:

$$F^t: R^m \to R \tag{9}$$

где $F^t$ получается из предыдущего следующим способом:

$$F^t = F^{t-1} + ah^t \tag{10}$$

где $a$ — это шаг, а функция $h^t$ — это базовая модель - предиктор. В случае CatBoost в качестве базовых моделей — предикторов используются бинарные решающие деревья. Выбор библиотеки CatBoost обусловлен её высокой точностью (данная реализация градиентного бустинга входит в число передовых алгоритмов), богатой документацией, а также удобством и гибкостью в применении. Как было сказано ранее, данный алгоритм наиболее подходит для обработки данных в табличном представлении. Для обучения CatBoost модели табличные представления молекул были сформированы на основании отпечатков Моргана [82] и набора физико-химических дескрипторов.

Отпечаток Моргана — это распространенное в молекулярной информатике описание, извлекаемое из SMILES, которое кодирует структурные особенности молекулы в виде бинарного вектора. Такие молекулярные отпечатки широко используются для анализа структурно-активных связей, однако имеют ограничения: они могут не улавливать тонкие структурные отличия в крупных молекулах и плохо отражают глобальные характеристики, например размер и

форму [83]. В связи с этим, дополнительно были включены часто используемые в молекулярной информатике физико-химические параметры, призванные охарактеризовать общие свойства молекулярной структуры. Их извлечение также осуществлялось на основе SMILES с помощью специализированного программного обеспечения RDKit (https://www.rdkit.org/). Следует отметить, что данные табличные дескрипторы также были использованы и GNN моделью (подробнее в разделе 3.5).

Помимо вышеупомянутого программного пакет RDKit, существуют и иные способы извлечения числовых дескрипторов, в частности, библиотека Mordred [84]. Mordred это программное обеспечение для расчёта дескрипторов, способное вычислять более 1800 двумерных и трёхмерных дескрипторов. Оно свободно доступно через GitHub. Mordred легко устанавливается и может использоваться через интерфейс командной строки, в виде веб-приложения или как гибкий пакет Python на всех основных платформах (Windows, Linux и macOS). Mordred работает быстрее различных аналогов, а также способен рассчитывать дескрипторы для крупных молекул, что недоступно многим другим программам. Благодаря высокой производительности, удобству, большому количеству дескрипторов и отсутствию строгих лицензионных ограничений, Mordred является популярным современным программным пакетом, и был успешно применен для задачи предсказания мишеней малых молекул [85]. Решение о разработке модели, работающей именно с данным признаковым описанием, было принято с целью исследования его применимости по сравнению с графовым подходом и более классической комбинацией из небольшого количества ключевых дескрипторов и отпечатков Моргана. Для работы с данным видом дескрипторов была разработана глубокая нейросеть (далее DNN), основанная на линейных слоях. Данная архитектура была выбрана вместо градиентного бустинга в связи с тем, что бустинговые алгоритмы (в особенности построенные на решающих деревьях, как например Catboost) менее приспособлены для работы с большим количество числовых признаков.

Существует более новый подход к работе с молекулярными данными, основанный на преобразовании символов SMILES в токены и создании для них обучаемых векторных представлений. Молекулярная структура при этом рассматривается как последовательность таких векторов. Одним из методов построения нейросетей для обработки подобных данных является архитектура одномерной сверточной нейронной сети (1D CNN). Этот способ становится всё более популярным для анализа молекулярных структур и показал высокую эффективность на недавнем открытом конкурсе по прогнозированию свойств молекул [85]. В 1D CNN свёртка производится по последовательности векторных представлений преобразованных в токены элементов SMILES записи, что позволяет выявлять структурные закономерности. В проведенном исследовании для реализации 1D CNN была использована версия из библиотеки PyTorch [54], осуществляющая взаимную корреляцию, которую в общем виде можно описать следующим образом (при условии, что фильтр обрабатывает элементы последовательно) [86]:

$$(I * K)(i) = \sum_{u=1}^{s} I(i + u - 1)\, K(u) \tag{11}$$

где *K* это фильтр с размером *s*, *I* это входной сигнал длинны *n*, индекс *i* идет от 1 до *n*, а индекс *u* идет от 1 to *s*. При выполнении свёртки с использованием множества фильтров над входными данными, представленными в виде массива с несколькими столбцами, свёртка может проводиться отдельно для каждого столбца. Набор фильтров при этом описывается тензором третьего порядка размерности *s x p x q*, где *p* это число сверток а *q* это количество колонок. В результате формируется двумерная матрица размером $(n - s + 1)xq$, которую можно рассматривать как массив из q столбцов, каждый из которых представляет собой сумму p свёрток [86]:

$$C_I(i) = \sum_{j=1}^{p} I_j * K_{j,l}(u) \tag{12}$$

где $l - 1,2, \dots q$.

В свёртку также включается обучаемый параметр суммируемого смещения.

## 3.4. Предсказание потенциальных мишеней малой молекулы: подготовка обучающих данных

Основным аспектом исследования было обучение моделей ML на молекулах, для которых была известна активность по всем предсказываемым мишеням. Набор данных был подготовлен экспертами в данной области, с применение открытых баз данных, в частности LINCS (https://lincs.hms.harvard.edu/kinomescan/), PKIS [87], PKIS2 [88], DCPKIS (https://www.sgc-ffm.uni-frankfurt.de/), KCGS [89] и EMD [90]. Соединения были помечены 1 или 0 в зависимости от того, демонстрировали ли они ингибирование не менее 50% при концентрации 1 мкмоль/л. Были сохранены только те соединения, для которых были доступны данные об ингибирующей активности против всех выбранных 75 киназ. Соединения, демонстрирующие противоречивые данные (более 60% ингибирования в некоторых экспериментах и менее 40% ингибирования в других для одной и той же мишени), были исключены. После удаления дубликатов набор данных содержал 1232 записи. В итоговом наборе данных были представлены следующие 75 киназных мишеней: ALK, AXL, BLK, BMX, BRSK1, BRSK2, BTK, CAMK1D, CHEK1, CLK2, CLK3, CSK, DAPK1, DDR2, DYRK2, EGFR, EPHA2, EPHA3, EPHA4, EPHB2, EPHB3, EPHB4, ERBB4, FER, FES, FGFR1, FGFR2, FGFR3, FGFR4, FGR, FLT3, FLT4, FYN, GRK7, GSK3A, GSK3B, HCK, HIPK1, IGF1R, INSR, IRAK4, ITK, KIT, LCK, MAPKAPK2, MARK1, MELK, MET, MUSK, NEK2, NEK6, NEK7, PAK2, PAK3, PAK6, PHKG2, PIM1, PIM2, PIM3, PLK1, PRKX, RET, ROCK1, ROCK2, SGK2, SGK3, SRC, SRPK1, SYK, TBK1, TEC, TNK2, TXK, TYRO3, ZAP70.

Как было сказано ранее, GNN модель использует ансамбль из 3D конформаций (до 20 конформаций на молекулу) для обучения, и для их генерации был применен экспертно разработанный метод, подготовленный для данного научного исследования; помимо него, было протестировано применение современного метода генерации конформаций малых молекул CONFORGE [91].

CONFORGE (при использовании параметров по умолчанию в соответствии с предоставленной документацией по применению) генерировал в среднем 8,6 конформаций на структуру, но не смог сгенерировать конформации для 96 структур. Экспертно разработанный алгоритм генерировал в среднем 16,8 конформаций на структуру и не смог сгенерировать конформации для 56 структур. В связи с полученными результатами было принято решение применить разработанный в целях исследования алгоритм генерации конформаций для создания данных для обучения GNN. После исключения молекул, для которых не удалось сгенерировать ни одной конформации, было оставлено 1176 записей для использования в дальнейших исследованиях.

Для проведения независимого тестирования от подготовленного набора данных была отделена одна тестовая выборка. Ее формирование представляло сложность в связи с неравномерным распределением и частотой встречаемости меток различных мишеней, от наиболее часто встречаемой метки для киназы KIT (294 малых молекул) до наименее встречаемой метки для киназы NEK6 (13 малых молекул). Полный список представлен в таблице 3.1.

В связи с этим, для формирования независимого теста, был применен следующий подход: случайным образом было проведено отделение 15% данных. С целью обеспечения представленности, разделение было проведено со стратификацией, но только по меткам, чьих примеров было менее 40 (т.к. стратификация по всем меткам не была возможной при данном размере тестового набора). Для обеспечения представленности наиболее редких классов, был сохранен первый случайный сплит, который обеспечивал присутствие как минимум 4 представителей всех классов в тестовом датасете. Всего в тестовый набор данных вошло 177 структур.

После отделения независимого датасета, оставшиеся 999 структур были случайно разделены на 5 обучающих и валидационных сплитов с целью контроля переобучения, а также снижения вероятности неоптимального обучения модели из-за неудачно созданного сплита. Валидационные выборки были отделены по тому же принципу, что и тестовая, за исключением снижения минимального

порога до 3 примеров для каждого класса. Данные выборки были использованы для всех обученных в ходе данного исследования моделей.

| Мишень | Ко-во меток | Мишень | Ко-во меток | Мишень | Ко-во меток |
|---|---|---|---|---|---|
| KIT | 294 | TXK | 83 | FER | 52 |
| FLT3 | 251 | MET | 78 | TYRO3 | 51 |
| LCK | 194 | FGFR1 | 77 | CHEK1 | 49 |
| RET | 185 | BMX | 74 | CSK | 48 |
| FLT4 | 177 | TBK1 | 73 | PAK6 | 48 |
| BLK | 168 | MUSK | 72 | TEC | 48 |
| EGFR | 162 | PLK1 | 72 | PIM2 | 48 |
| CLK2 | 160 | ALK | 72 | SGK3 | 46 |
| DDR2 | 146 | BRSK1 | 71 | EPHA3 | 45 |
| FGR | 143 | BTK | 71 | PIM3 | 45 |
| GSK3A | 141 | SYK | 68 | IGF1R | 44 |
| HCK | 140 | PRKX | 67 | PHKG2 | 41 |
| SRC | 131 | BRSK2 | 67 | NEK2 | 40 |
| AXL | 130 | EPHB2 | 66 | FES | 40 |
| FYN | 128 | PAK3 | 64 | GRK7 | 40 |
| ERBB4 | 126 | EPHA2 | 61 | PAK2 | 39 |
| ROCK2 | 118 | EPHB4 | 60 | ITK | 35 |
| ROCK1 | 114 | CLK3 | 60 | CAMK1D | 33 |
| SRPK1 | 101 | EPHA4 | 60 | SGK2 | 31 |
| GSK3B | 100 | MARK1 | 59 | FGFR4 | 30 |
| HIPK1 | 98 | IRAK4 | 58 | ZAP70 | 24 |
| MELK | 96 | PIM1 | 58 | NEK7 | 20 |
| TNK2 | 84 | INSR | 58 | EPHB3 | 18 |
| DYRK2 | 84 | FGFR3 | 55 | MAPKAPK2 | 14 |
| FGFR2 | 83 | DAPK1 | 55 | NEK6 | 13 |

Таблица 3.1 Представленность 75 рассматриваемых меток киназных мишеней

## 3.5. Предсказание потенциальных мишеней малой молекулы: подготовка входных данных и алгоритм работы методов

Для подготовки входных данных для обучения GNN модели, для каждой малой молекулы генерировалось до 20 трехмерных конформаций. На основании данных конформаций было построено их представление в виде фармакофорных точек, на основании которых строились полносвязные графы, где вершинами являются фармакофорные точки, признаком которых является их фармакофорный тип (указанные выше 6 возможных типов), закодированные one-hot encoding способом, в то время как признаками на ребрах были расстояния в Å между фармакофорами. Дополнительно, на основании нотации SMILES, собиралось признаковое описание молекулы в виде 512-битового вектора отпечатка Моргана, а также следующие 11 физико-химических дескрипторов с использованием RDKit: LogP (липофильность), TPSA (топологическая площадь полярной поверхности молекулы), NHOHCount (количество OH и NH групп), NOCount (количество атомов кислорода и азота), NumHAcceptors (количество акцепторов водорода), NumHDonors (количество доноров водорода), NumRotatableBonds (количество вращаемых связей), NumHeteroatoms (количество гетероатомов), FractionCSP3 (количество углеродов в SP3 гибридизации), ExactMolWt (молекулярный вес), NumAromaticRing (количество ароматических колец). К данным дескрипторам был применен метод стандартного масштабирования.

Ансамбль фармакофорных точек в трехмерном пространстве обрабатывался GNN слоями, а их ответ объединялся с табличными признаками для дальнейшей обработки линейным слоями, которые выдавали финальные прогнозы каждого класса для данной молекулы в диапазоне [0,1]. Процесс представлен на рисунке 3.1.

Для подготовки данных DNN модели были собраны все доступные дескрипторы Mordred, исключая те, у которых все значения были нулями или NaN для подобранной выборки данных. Для дальнейшего уменьшения пространства признаков и решения проблемы наличия потенциально избыточных или сильно коррелированных признаков применялось преобразование PCA

(предварительно к признакам было применено стандартное масштабирование). Итеративно проверив различные значения количества компонент, признаковое описание было приведено к 190 компонентам, которые объясняли 99,4% дисперсии, демонстрируя минимальную потерю информации.

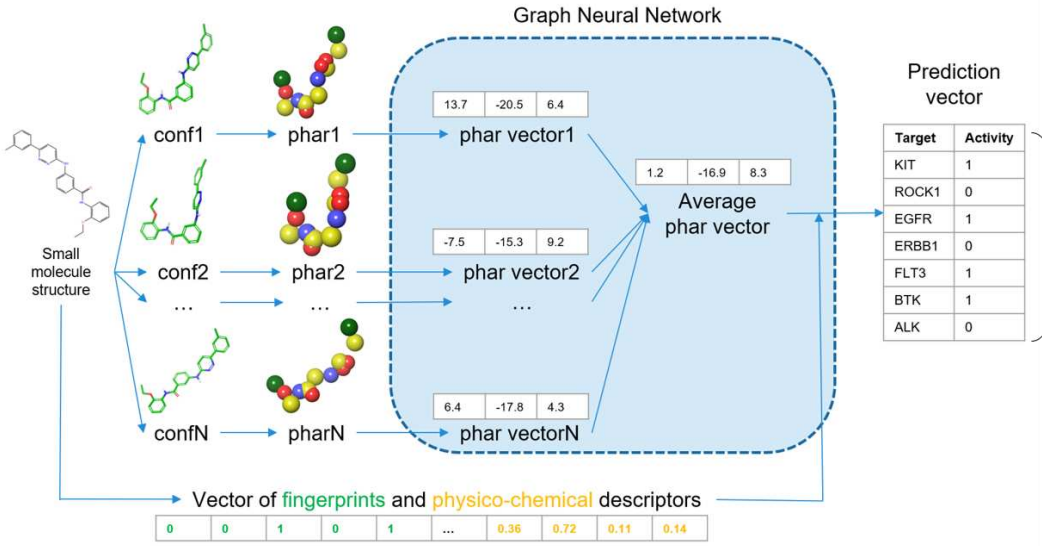


Рисунок 3.1. Схема обработки молекулы и прогнозирования активности

Модель 1D CNN обрабатывала представление молекулы в нотации SMILES как закодированную последовательность. Нотация SMILES токенизировалась, присваивая числовые значения часто встречающимся символам (сочетания, обозначающие единственное свойство или имя атома, например, «Br», кодировались как один токен). Дополнительные токены резервировались для заполнения и других символов, в результате чего получалось в общей сложности 38 уникальных токенов.

Модель Catboost принимала на вход вышеупомянутые отпечатки Моргана размерностью 512 и 11 физико-химических дескрипторов. Принимая во внимание принцип работы Catboost и рекомендации разработчиков данного метода градиентного бустинга, признаки не масштабировались. Следует отметить, что были протестированы и другие варианты отпечатков Моргана, а именно в виде 1024 и 2048 бит соответственно, однако, более крупные векторные представления не давали прироста точности в рамках проведенных тестов, повышая при этом вычислительные затраты.

## 3.6. Предсказание потенциальных мишеней малой молекулы: архитектура GNN и моделей сравнения

Разработанная GNN основана на комбинации GNN слоев (использовался тот же оператор что и в предыдущих главах), обрабатывающих графовые представления каждого из 3D-конформеров в ансамбле, слоев агрегации выходов по всем графам и линейных слоев для интеграции объединенного выходного сигнала GNN с отпечатками Моргана и физико-химическими дескрипторами; финальный линейный слой обеспечивает окончательное предсказание всех 75 классов как множественных бинарных меток, где истинными может быть произвольное количество меток, включая все (случай так называемых пан-киназных ингибиторов). Для GNN блока, признак расстояния на ребрах был предварительно обработан блоком, состоящим из слоя размытия Гаусса с 256 гауссианами, за которым следовала последовательность из двух линейных слоев, использующих функцию активации leaky ReLU, чтобы сгенерировать 64-мерное представление данных. Это представление было использовано для репрезентации признаков на ребрах графа фарамкофорных точек, где признаком на вершинах был тип фармакофорной точки, закодированный one-hot методом. Данные графы поступали в GNN блок, состоящий из 8 GNN слоев с 300 нейронами и 10 головами внимания (кроме первого слоя, у которого было 4 головы внимания), функциями активации утечки ReLU, слоями пакетной нормализации и слоем dropout с вероятностью обнуления 20% (после последнего слоя dropout не применялся). Выходы блока GNN для каждого графа агрегировались путем взятия среднего значения, чтобы получить единственный вектор-ответ для каждого графа. Данные вектора затем усреднялись по всем графам каждого ансамбля, представляющего целевую молекулу, чтобы получить один ответ для всего ансамбля. Следует отметить, что были протестированы и иные методы агрегации информации по ансамблю графов, в частности использование вектора агрегации по среднему и максимуму, применение отдельного GNN слоя для агрегации, однако данные вариации либо снижали точность модели, либо не влияли на нее, требуя при этом больше вычислительных ресурсов. После прохождения еще

одного слоя dropout с параметром вероятности 20%, полученный вектор объединялся с вектором отпечатков Моргана и физико-химическими признаками, а затем подавался в серию 4 линейных слоев, первый слой которого спроектирован по архитектуре, аналогичной ячейке LSTM [92]. Финальный слой генерирует оценку от 0 до 1 для каждого из 75 обученных меток, соответствующих данной молекуле.

DNN модель состояла из шести линейных слоев: входного слоя с архитектурой, аналогичной ячейке LSTM, с размерностью выхода 1024; четырех скрытых линейных слоев, каждый из которых имел размерность выхода, уменьшенную вдвое по сравнению с предыдущим слоем; и выходного линейного слоя. Нейронная сеть включала нормализацию пакетов для ускорения и стабилизации обучения, а также 50% dropout для снижения переобучения. На протяжении всей сети использовалась функция активации ReLU.

1D CNN модель принимала токенизированную последовательность строковой записи SMILES молекулы в слой вложений, который генерировал обучаемое векторное представление размерностью 256 для каждого возможного токена из словаря (38). Этот вектор обрабатывался блоком CNN, состоящим из пяти слоев 1D CNN, каждый из которых имел увеличивающееся количество фильтров, начиная с 32 и удваиваясь в каждом последующем слое, используя размер ядра 5. Выход блока CNN затем обрабатывался многослойным линейным блоком, который использовал функцию активации ReLU и включал слои нормализации пакетов и dropout. Данная архитектура была создана по аналогии с архитектурами 1D CNN, продемонстрированными в [85].

Модель CatBoost использовала отпечатки Моргана и физико-химические дескрипторы в качестве входных данных для предсказания потенциальных мишеней для данной молекулы. Для настройки гиперпараметров модели был выполнен поиск по сетке с трехкратной валидацией на обучающем наборе данных. Финальные гиперпараметры для модели градиентного бустинга включали скорость обучения 0,03, глубину дерева 6 и параметром L2-регуляризации листьев 7.

## 3.7. Предсказание потенциальных мишеней малой молекулы: обучение GNN и моделей сравнения

Для обучения GNN, обучающие данные подавались батчами размером 10, где каждый объект представлял собой ансамбль графов 3D-фармакофоров для целевой молекулы. Батчи перемешивались на каждой эпохе обучения. Модель обучалась с скоростью обучения $10^{-4}$ с использованием оптимизатора Adam и функции бинарной кросс энтропии в течение 60 эпох. Процесс обучения повторялся три раза для каждого сплита обучения и валидации. Итоговые веса фиксировались на эпохе, на которой модель показывала наименьшее значение функции потерь на обучающей выборке, при этом находясь в пределах 1% от самой низкой зафиксированной потери на валидации.

Для каждой из пяти лучших моделей пороги бинаризации для каждой метки корректировались для максимизации макросредней F1-оценки для данного целевого предсказания на соответствующем наборе обучающих данных. Эти пороги рассчитывались с помощью стохастического метода оптимизации Basin-Hopping [93], при этом для локальной минимизации на каждом шаге использовался метод L-BFGS-B [94]. Итоговая GNN модель выбиралась для внешнего тестирования и сравнения с другими разработанными моделями на основе средней F1-оценки на независимом тестовом наборе с применением подобранных порогов бинаризации.

Обучение моделей сравнения проходило по аналогичному принципу – для каждого сплита обучение\валидация обучалась собственная модель, а для итогового сравнения выбиралась модель, показавшая наилучший результат на тестовой выборке с применением порогов бинаризации. Чтобы обеспечить справедливое сравнение с GNN, все модели сравнения обучались и оценивались на тех же сплитах обучения и валидации, и пороги бинаризации для каждой метки рассчитывались аналогичным образом. Для модели Catboost, обучение проводилось на 5000 итераций, с возможностью ранней остановки обучения в случае, если значение функции потери на валидационной выборке не улучшалось в течение 200 итераций.

Модель 1D CNN обучалась батчами размером 128 со скоростью обучения $10^{-4}$, используя оптимизатор Adam и функцию потерь бинарная кросс-энтропия. Обучение проходило в течении 200 эпох, с возможностью ранней остановки, если потеря на валидации не улучшалась в течение 20 эпох. Выбирались веса модели с самым низким значением функции потерь на валидации.

DNN обучалась с размером батча 256, скоростью обучения $10^{-3}$, используя оптимизатор Adam и функцию потерь бинарная кросс-энтропия. Обучение проходило в течении 200 эпох, с возможностью ранней остановки, если потеря на валидации не улучшалась в течение 20 эпох. Выбирались веса модели с самым низким значением функции потерь на валидации.

Все модели создавались и обучались с использованием библиотек PyTorch [54] и PyTorch Geometric [55]. Для обучения моделей использовалась видеокарта NVIDIA A100-SXM4 с 80 ГБ памяти, что занимало примерно 5 минут на 60 эпох.

## 3.8. Предсказание потенциальных мишеней малой молекулы: тестирование и сравнение разработанных моделей

Для оценки производительности различных алгоритмов машинного обучения, модели были сравнены с использованием независимого тестового набора. На основе полученных результатов модель GNN, которая использовала как ансамбль 3D-фармакофоров молекулы, так и признаки, полученные из SMILES, превзошла модели сравнения. Рассматривая лучшие результаты, достигнутые для каждой предложенной архитектуры, модель GNN продемонстрировала макросредний показатель F1 в 0,47 по сравнению с показателем F1 в 0,39 для модели на основе CatBoost, 0,40 для DNN на основе дескрипторов Mordred и 0,23 для 1D CNN. Эффективность лучших моделей также была оценена с использованием коэффициента корреляции Мэттьюса (MCC) и сбалансированной точности (BA), усреднённых по всем предсказываемым меткам. GNN модель показала MCC равный 0,446 и BA равный 0,714, по сравнению с MCC равным 0,392 и BA равным 0,648 для модели CatBoost, MCC

равным 0,363 и BA равным 0,685 для DNN на основе дескрипторов Mordred, и MCC равным 0,224 и BA равным 0,577 для 1D CNN.

При расчете средних показателей по всем пяти использованным сплитам обучение/валидация для каждого метода были получены показатели F1 в 0,426, 0,38, 0,358 и 0,194 для модели GNN, модели CatBoost, DNN на основе Mordred и 1D CNN соответственно (Рисунок 3.2). Модель GNN показала статистически значимую разницу по сравнению с CatBoost (p-значение = 0,03) и другими базовыми методами. Несмотря на то, что архитектура 1D CNN, обученная на последовательностях SMILES, широко использовалась в недавних соревнованиях по хеминформатике и показала один из лучших результатов на соревновании Kaggle по предсказанию мишеней для молекул [85], в данном эксперименте проявила себя слабее всего в рамках используемых метрик. Напротив, два других подхода — DNN, обученная на дескрипторах Mordred и градиентный бустинг с применением отпечатков Моргана и выбранных физико-химических дескрипторов — показали лучшие результаты, демонстрируя более высокую способность к обобщению этих молекулярных представлений при использовании с подходящими архитектурами.

На обучающем наборе данных были получены следующие макро-средние оценки F1: 0,91, 0,54, 0,91 и 0,72 для моделей 1D CNN, DNN, CatBoost и GNN соответственно. Архитектуры CatBoost и 1D CNN показали более высокие метрики, чем GNN, при применении к обучающим данным, но продемонстрировали более низкую точность на независимом тесте. Среди всех обученных моделей модель DNN на основе дескрипторов Mordred показала минимальную разницу в качестве предсказаний на обучающих и тестовых наборах данных. Поскольку все модели были обучены на одних и тех же сплитах валидации и обучения, сохраняя состояние с наименьшим значением потерь на валидации, справедливо предположить, что наблюдаемая способность к обобщению не могла быть вызвана каким-либо врожденным смещением в распределении данных, но вероятно зависит от выбранного представления данных и архитектуры модели. Добавление информации о фармакофорах в 3D

пространстве позволило добиться наиболее точных предсказаний, позволяя судить о важности использования молекулярной 3D структуры. Хотя, исходя из проведенного исследования литературы в данной области, в настоящее время нет подходов к генерации конформаций, которые могли бы гарантировать генерацию активной конформации, данный эксперимент показывает, что обобщенное представление молекулы в виде примеров возможных 3D конформаций предоставляет дополнительную информацию, позволяющую ML модели лучше оценивать способность молекул модулировать активность данного белка.

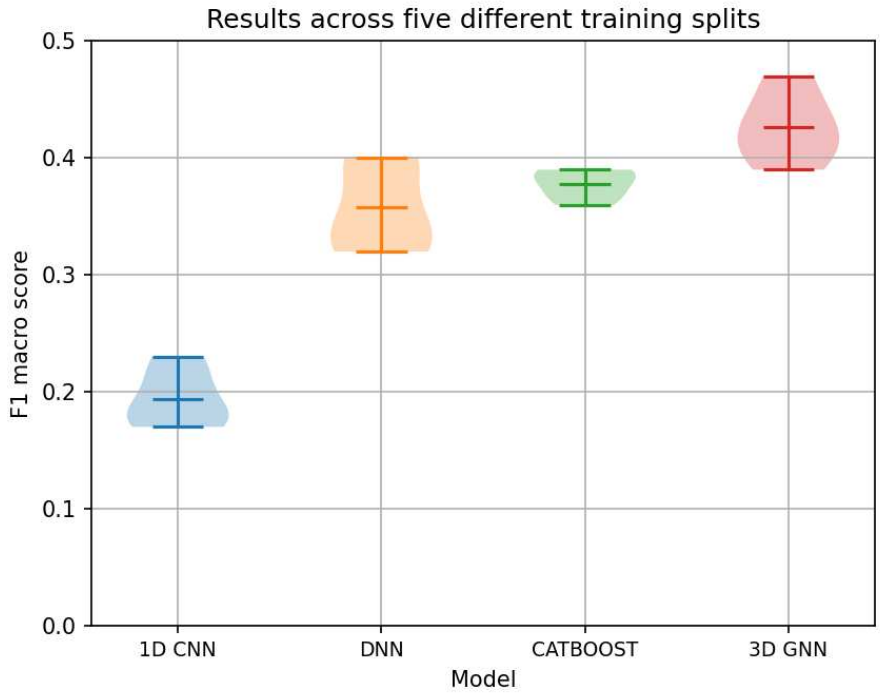


Рисунок 3.2. Сравнение подготовленных моделей на независимом наборе данных

В таблицах 3.2, 3.3, 3.4, 3.5 представлены оценки точности всех разработанных в рамках данного исследования моделей. Как можно видеть, модели показали разные уровни точности в отношении различных мишеней, работая лучше на одних, чем на других. Эта вариативность может быть частично объяснена значительным дисбалансом меток разных мишеней в используемом наборе данных.

| **Мишень** | **Точность** | **Полнота** | **F1** | **Кол-во примеров** |
|---|---|---|---|---|
| ALK | 0.35 | 0.43 | 0.39 | 14 |
| AXL | 0.81 | 0.50 | 0.62 | 26 |
| BLK | 0.57 | 0.68 | 0.62 | 31 |
| BMX | 0.47 | 0.50 | 0.49 | 18 |
| BRSK1 | 0.83 | 0.45 | 0.59 | 11 |
| BRSK2 | 0.29 | 0.33 | 0.31 | 12 |
| BTK | 0.43 | 0.46 | 0.44 | 13 |
| CAMK1D | 0.50 | 0.43 | 0.46 | 7 |
| CHEK1 | 0.20 | 0.50 | 0.29 | 4 |
| CLK2 | 0.68 | 0.46 | 0.55 | 28 |
| CLK3 | 0.08 | 0.10 | 0.09 | 10 |
| CSK | 0.30 | 0.67 | 0.41 | 9 |
| DAPK1 | 0.57 | 0.44 | 0.50 | 9 |
| DDR2 | 0.56 | 0.67 | 0.61 | 21 |
| DYRK2 | 0.50 | 0.36 | 0.42 | 14 |
| EGFR | 0.77 | 0.74 | 0.75 | 31 |
| EPHA2 | 0.71 | 0.42 | 0.53 | 12 |
| EPHA3 | 0.67 | 0.44 | 0.53 | 9 |
| EPHA4 | 0.43 | 0.27 | 0.33 | 11 |
| EPHB2 | 0.45 | 0.62 | 0.53 | 8 |
| EPHB3 | 0.00 | 0.00 | 0.00 | 4 |
| EPHB4 | 0.60 | 0.30 | 0.40 | 10 |
| ERBB4 | 0.81 | 0.81 | 0.81 | 26 |
| FER | 0.57 | 0.33 | 0.42 | 12 |
| FES | 0.44 | 0.57 | 0.50 | 7 |
| FGFR1 | 0.39 | 0.54 | 0.45 | 13 |
| FGFR2 | 0.31 | 0.36 | 0.33 | 11 |
| FGFR3 | 0.45 | 0.45 | 0.45 | 11 |
| FGFR4 | 0.27 | 0.75 | 0.40 | 4 |

| | | | | |
|---|---|---|---|---|
| FGR | 0.50 | 0.67 | 0.57 | 24 |
| FLT3 | 0.71 | 0.71 | 0.71 | 38 |
| FLT4 | 0.42 | 0.67 | 0.52 | 21 |
| FYN | 0.63 | 0.46 | 0.53 | 26 |
| GRK7 | 0.50 | 0.40 | 0.44 | 5 |
| GSK3A | 0.65 | 0.58 | 0.61 | 19 |
| GSK3B | 0.47 | 0.54 | 0.50 | 13 |
| HCK | 0.51 | 0.69 | 0.59 | 26 |
| HIPK1 | 0.28 | 0.29 | 0.29 | 17 |
| IGF1R | 0.50 | 0.56 | 0.53 | 9 |
| INSR | 0.56 | 0.42 | 0.48 | 12 |
| IRAK4 | 0.83 | 0.56 | 0.67 | 9 |
| ITK | 0.67 | 0.25 | 0.36 | 8 |
| KIT | 0.72 | 0.82 | 0.77 | 44 |
| LCK | 0.61 | 0.67 | 0.64 | 33 |
| MAPKAPK2 | 0.00 | 0.00 | 0.00 | 4 |
| MARK1 | 0.40 | 0.57 | 0.47 | 7 |
| MELK | 0.46 | 0.55 | 0.50 | 11 |
| MET | 0.67 | 0.33 | 0.44 | 18 |
| MUSK | 0.43 | 0.43 | 0.43 | 7 |
| NEK2 | 0.71 | 0.50 | 0.59 | 10 |
| NEK6 | 0.00 | 0.00 | 0.00 | 4 |
| NEK7 | 0.40 | 0.50 | 0.44 | 4 |
| PAK2 | 0.50 | 0.50 | 0.50 | 6 |
| PAK3 | 0.43 | 0.38 | 0.40 | 8 |
| PAK6 | 0.33 | 0.14 | 0.20 | 7 |
| PHKG2 | 0.57 | 0.67 | 0.62 | 6 |
| PIM1 | 0.75 | 0.40 | 0.52 | 15 |
| PIM2 | 0.67 | 0.60 | 0.63 | 10 |
| PIM3 | 0.62 | 0.56 | 0.59 | 9 |
| PLK1 | 0.89 | 0.73 | 0.80 | 11 |

| | | | | |
|---|---|---|---|---|
| PRKX | 1.00 | 0.36 | 0.53 | 11 |
| RET | 0.47 | 0.57 | 0.52 | 28 |
| ROCK1 | 0.78 | 0.41 | 0.54 | 17 |
| ROCK2 | 0.69 | 0.43 | 0.53 | 21 |
| SGK2 | 0.57 | 0.57 | 0.57 | 7 |
| SGK3 | 1.00 | 0.30 | 0.46 | 10 |
| SRC | 0.41 | 0.39 | 0.40 | 23 |
| SRPK1 | 0.44 | 0.27 | 0.33 | 15 |
| SYK | 0.32 | 0.64 | 0.42 | 11 |
| TBK1 | 0.60 | 0.40 | 0.48 | 15 |
| TEC | 0.57 | 0.31 | 0.40 | 13 |
| TNK2 | 0.33 | 0.38 | 0.36 | 13 |
| TXK | 0.40 | 0.40 | 0.40 | 15 |
| TYRO3 | 0.45 | 0.62 | 0.53 | 8 |
| ZAP70 | 1.00 | 0.29 | 0.44 | 7 |
| | | | | |
| **Макро среднее** | 0.53 | 0.47 | 0.47 | 1051 |
| **Микро среднее** | 0.54 | 0.52 | 0.53 | 1051 |
| **Взвешенное среднее** | 0.57 | 0.52 | 0.52 | 1051 |
| **Среднее по объекту** | 0.32 | 0.29 | 0.27 | 1051 |

Таблица 3.2 Точность классификации GNN модели на тестовой выборке с применением подобранных порогов бинаризации

| **Мишень** | **Точность** | **Полнота** | **F1** | **Кол-во примеров** |
|---|---|---|---|---|
| ALK | 0.86 | 0.43 | 0.57 | 14 |
| AXL | 0.69 | 0.42 | 0.52 | 26 |
| BLK | 0.74 | 0.65 | 0.69 | 31 |
| BMX | 0.50 | 0.11 | 0.18 | 18 |
| BRSK1 | 0.67 | 0.18 | 0.29 | 11 |

| | | | | |
|---|---|---|---|---|
| BRSK2 | 0.50 | 0.08 | 0.14 | 12 |
| BTK | 0.44 | 0.31 | 0.36 | 13 |
| CAMK1D | 0.50 | 0.14 | 0.22 | 7 |
| CHEK1 | 0.33 | 0.25 | 0.29 | 4 |
| CLK2 | 0.61 | 0.50 | 0.55 | 28 |
| CLK3 | 0.20 | 0.10 | 0.13 | 10 |
| CSK | 0.36 | 0.44 | 0.40 | 9 |
| DAPK1 | 0.60 | 0.33 | 0.43 | 9 |
| DDR2 | 0.58 | 0.52 | 0.55 | 21 |
| DYRK2 | 0.25 | 0.07 | 0.11 | 14 |
| EGFR | 0.85 | 0.71 | 0.77 | 31 |
| EPHA2 | 0.00 | 0.00 | 0.00 | 12 |
| EPHA3 | 1.00 | 0.11 | 0.20 | 9 |
| EPHA4 | 0.00 | 0.00 | 0.00 | 11 |
| EPHB2 | 0.50 | 0.13 | 0.20 | 8 |
| EPHB3 | 0.00 | 0.00 | 0.00 | 4 |
| EPHB4 | 1.00 | 0.20 | 0.33 | 10 |
| ERBB4 | 0.89 | 0.62 | 0.73 | 26 |
| FER | 1.00 | 0.33 | 0.50 | 12 |
| FES | 1.00 | 0.14 | 0.25 | 7 |
| FGFR1 | 1.00 | 0.23 | 0.38 | 13 |
| FGFR2 | 0.43 | 0.27 | 0.33 | 11 |
| FGFR3 | 0.75 | 0.27 | 0.40 | 11 |
| FGFR4 | 1.00 | 0.25 | 0.40 | 4 |
| FGR | 0.70 | 0.58 | 0.64 | 24 |
| FLT3 | 0.76 | 0.66 | 0.70 | 38 |
| FLT4 | 0.67 | 0.38 | 0.48 | 21 |
| FYN | 0.83 | 0.58 | 0.68 | 26 |
| GRK7 | 0.75 | 0.60 | 0.67 | 5 |
| GSK3A | 0.71 | 0.26 | 0.38 | 19 |
| GSK3B | 0.67 | 0.31 | 0.42 | 13 |

| | | | | |
|---|---|---|---|---|
| HCK | 0.81 | 0.50 | 0.62 | 26 |
| HIPK1 | 0.42 | 0.29 | 0.34 | 17 |
| IGF1R | 0.75 | 0.33 | 0.46 | 9 |
| INSR | 0.67 | 0.33 | 0.44 | 12 |
| IRAK4 | 0.33 | 0.11 | 0.17 | 9 |
| ITK | 0.00 | 0.00 | 0.00 | 8 |
| KIT | 0.67 | 0.68 | 0.67 | 44 |
| LCK | 0.84 | 0.48 | 0.62 | 33 |
| MAPKAPK2 | 0.50 | 0.25 | 0.33 | 4 |
| MARK1 | 0.67 | 0.29 | 0.40 | 7 |
| MELK | 0.40 | 0.18 | 0.25 | 11 |
| MET | 0.75 | 0.33 | 0.46 | 18 |
| MUSK | 0.29 | 0.29 | 0.29 | 7 |
| NEK2 | 1.00 | 0.30 | 0.46 | 10 |
| NEK6 | 0.00 | 0.00 | 0.00 | 4 |
| NEK7 | 0.00 | 0.00 | 0.00 | 4 |
| PAK2 | 1.00 | 0.17 | 0.29 | 6 |
| PAK3 | 0.43 | 0.38 | 0.40 | 8 |
| PAK6 | 0.43 | 0.43 | 0.43 | 7 |
| PHKG2 | 1.00 | 0.33 | 0.50 | 6 |
| PIM1 | 1.00 | 0.07 | 0.13 | 15 |
| PIM2 | 0.71 | 0.50 | 0.59 | 10 |
| PIM3 | 0.50 | 0.44 | 0.47 | 9 |
| PLK1 | 0.89 | 0.73 | 0.80 | 11 |
| PRKX | 1.00 | 0.36 | 0.53 | 11 |
| RET | 0.75 | 0.64 | 0.69 | 28 |
| ROCK1 | 0.56 | 0.53 | 0.55 | 17 |
| ROCK2 | 0.57 | 0.38 | 0.46 | 21 |
| SGK2 | 1.00 | 0.43 | 0.60 | 7 |
| SGK3 | 1.00 | 0.40 | 0.57 | 10 |
| SRC | 0.50 | 0.30 | 0.38 | 23 |

| | | | | |
|---|---|---|---|---|
| SRPK1 | 0.38 | 0.20 | 0.26 | 15 |
| SYK | 0.75 | 0.27 | 0.40 | 11 |
| TBK1 | 0.50 | 0.20 | 0.29 | 15 |
| TEC | 0.38 | 0.23 | 0.29 | 13 |
| TNK2 | 0.67 | 0.15 | 0.25 | 13 |
| TXK | 0.50 | 0.33 | 0.40 | 15 |
| TYRO3 | 0.57 | 0.50 | 0.53 | 8 |
| ZAP70 | 0.00 | 0.00 | 0.00 | 7 |
| | | | | |
| **Макро среднее** | 0.61 | 0.31 | 0.39 | 1051 |
| **Микро среднее** | 0.66 | 0.38 | 0.49 | 1051 |
| **Взвешенное среднее** | 0.65 | 0.38 | 0.46 | 1051 |
| **Среднее по объекту** | 0.33 | 0.26 | 0.27 | 1051 |

Таблица 3.3 Точность классификации Catboost модели на тестовой выборке с применением подобранных порогов бинаризации

| **Мишень** | **Точность** | **Полнота** | **F1** | **Кол-во примеров** |
|---|---|---|---|---|
| ALK | 0.38 | 0.36 | 0.37 | 14 |
| AXL | 0.56 | 0.77 | 0.65 | 26 |
| BLK | 0.65 | 0.65 | 0.65 | 31 |
| BMX | 0.42 | 0.28 | 0.33 | 18 |
| BRSK1 | 0.36 | 0.45 | 0.40 | 11 |
| BRSK2 | 0.27 | 0.33 | 0.30 | 12 |
| BTK | 0.33 | 0.23 | 0.27 | 13 |
| CAMK1D | 0.25 | 0.29 | 0.27 | 7 |
| CHEK1 | 0.10 | 0.25 | 0.14 | 4 |
| CLK2 | 0.48 | 0.39 | 0.43 | 28 |
| CLK3 | 0.19 | 0.30 | 0.23 | 10 |
| CSK | 0.29 | 0.67 | 0.40 | 9 |

| | | | | |
|---|---|---|---|---|
| DAPK1 | 0.38 | 0.56 | 0.45 | 9 |
| DDR2 | 0.56 | 0.48 | 0.51 | 21 |
| DYRK2 | 0.24 | 0.29 | 0.26 | 14 |
| EGFR | 0.82 | 0.74 | 0.78 | 31 |
| EPHA2 | 0.43 | 0.25 | 0.32 | 12 |
| EPHA3 | 0.36 | 0.44 | 0.40 | 9 |
| EPHA4 | 0.32 | 0.55 | 0.40 | 11 |
| EPHB2 | 0.33 | 0.38 | 0.35 | 8 |
| EPHB3 | 0.00 | 0.00 | 0.00 | 4 |
| EPHB4 | 0.31 | 0.50 | 0.38 | 10 |
| ERBB4 | 0.83 | 0.77 | 0.80 | 26 |
| FER | 0.56 | 0.42 | 0.48 | 12 |
| FES | 0.33 | 0.29 | 0.31 | 7 |
| FGFR1 | 0.30 | 0.54 | 0.39 | 13 |
| FGFR2 | 0.27 | 0.27 | 0.27 | 11 |
| FGFR3 | 0.33 | 0.18 | 0.24 | 11 |
| FGFR4 | 0.20 | 0.50 | 0.29 | 4 |
| FGR | 0.54 | 0.58 | 0.56 | 24 |
| FLT3 | 0.49 | 0.74 | 0.59 | 38 |
| FLT4 | 0.37 | 0.52 | 0.43 | 21 |
| FYN | 0.67 | 0.46 | 0.55 | 26 |
| GRK7 | 0.14 | 0.20 | 0.17 | 5 |
| GSK3A | 0.53 | 0.53 | 0.53 | 19 |
| GSK3B | 0.47 | 0.62 | 0.53 | 13 |
| HCK | 0.56 | 0.54 | 0.55 | 26 |
| HIPK1 | 0.37 | 0.65 | 0.47 | 17 |
| IGF1R | 0.80 | 0.44 | 0.57 | 9 |
| INSR | 0.83 | 0.42 | 0.56 | 12 |
| IRAK4 | 0.57 | 0.44 | 0.50 | 9 |
| ITK | 0.33 | 0.12 | 0.18 | 8 |
| KIT | 0.65 | 0.70 | 0.67 | 44 |

| | | | | |
|---|---|---|---|---|
| LCK | 0.67 | 0.73 | 0.70 | 33 |
| MAPKAPK2 | 0.00 | 0.00 | 0.00 | 4 |
| MARK1 | 0.17 | 0.14 | 0.15 | 7 |
| MELK | 0.35 | 0.55 | 0.43 | 11 |
| MET | 0.50 | 0.33 | 0.40 | 18 |
| MUSK | 0.38 | 0.43 | 0.40 | 7 |
| NEK2 | 0.38 | 0.60 | 0.46 | 10 |
| NEK6 | 0.00 | 0.00 | 0.00 | 4 |
| NEK7 | 1.00 | 0.50 | 0.67 | 4 |
| PAK2 | 0.33 | 0.17 | 0.22 | 6 |
| PAK3 | 0.16 | 0.38 | 0.22 | 8 |
| PAK6 | 0.14 | 0.14 | 0.14 | 7 |
| PHKG2 | 0.44 | 0.67 | 0.53 | 6 |
| PIM1 | 0.33 | 0.33 | 0.33 | 15 |
| PIM2 | 0.33 | 0.20 | 0.25 | 10 |
| PIM3 | 0.29 | 0.22 | 0.25 | 9 |
| PLK1 | 0.73 | 0.73 | 0.73 | 11 |
| PRKX | 0.57 | 0.36 | 0.44 | 11 |
| RET | 0.47 | 0.68 | 0.56 | 28 |
| ROCK1 | 0.55 | 0.35 | 0.43 | 17 |
| ROCK2 | 0.45 | 0.43 | 0.44 | 21 |
| SGK2 | 0.30 | 0.43 | 0.35 | 7 |
| SGK3 | 0.50 | 0.50 | 0.50 | 10 |
| SRC | 0.50 | 0.52 | 0.51 | 23 |
| SRPK1 | 0.28 | 0.33 | 0.30 | 15 |
| SYK | 0.25 | 0.27 | 0.26 | 11 |
| TBK1 | 0.67 | 0.40 | 0.50 | 15 |
| TEC | 0.40 | 0.31 | 0.35 | 13 |
| TNK2 | 0.23 | 0.23 | 0.23 | 13 |
| TXK | 0.42 | 0.53 | 0.47 | 15 |
| TYRO3 | 0.40 | 0.75 | 0.52 | 8 |

| | | | | |
|---|---|---|---|---|
| ZAP70 | 0.50 | 0.29 | 0.36 | 7 |
| | | | | |
| **Макро среднее** | 0.41 | 0.42 | 0.40 | 1051 |
| **Микро среднее** | 0.45 | 0.49 | 0.47 | 1051 |
| **Взвешенное среднее** | 0.47 | 0.49 | 0.47 | 1051 |
| **Среднее по объекту** | 0.30 | 0.29 | 0.26 | 1051 |

Таблица 3.4 Точность классификации DNN модели на тестовой выборке с применением подобранных порогов бинаризации

| **Мишень** | **Точность** | **Полнота** | **F1** | **Кол-во примеров** |
|---|---|---|---|---|
| ALK | 0.67 | 0.14 | 0.24 | 14 |
| AXL | 0.67 | 0.15 | 0.25 | 26 |
| BLK | 0.55 | 0.19 | 0.29 | 31 |
| BMX | 0.43 | 0.17 | 0.24 | 18 |
| BRSK1 | 0.33 | 0.18 | 0.24 | 11 |
| BRSK2 | 0.29 | 0.17 | 0.21 | 12 |
| BTK | 0.29 | 0.15 | 0.20 | 13 |
| CAMK1D | 1.00 | 0.14 | 0.25 | 7 |
| CHEK1 | 0.17 | 0.25 | 0.20 | 4 |
| CLK2 | 0.80 | 0.14 | 0.24 | 28 |
| CLK3 | 0.00 | 0.00 | 0.00 | 10 |
| CSK | 0.33 | 0.22 | 0.27 | 9 |
| DAPK1 | 0.20 | 0.11 | 0.14 | 9 |
| DDR2 | 0.55 | 0.29 | 0.37 | 21 |
| DYRK2 | 0.50 | 0.14 | 0.22 | 14 |
| EGFR | 0.52 | 0.42 | 0.46 | 31 |
| EPHA2 | 0.25 | 0.08 | 0.13 | 12 |
| EPHA3 | 0.50 | 0.22 | 0.31 | 9 |

| | | | | |
|---|---|---|---|---|
| EPHA4 | 0.67 | 0.18 | 0.29 | 11 |
| EPHB2 | 0.25 | 0.13 | 0.17 | 8 |
| EPHB3 | 0.00 | 0.00 | 0.00 | 4 |
| EPHB4 | 0.50 | 0.10 | 0.17 | 10 |
| ERBB4 | 0.58 | 0.42 | 0.49 | 26 |
| FER | 0.50 | 0.25 | 0.33 | 12 |
| FES | 0.33 | 0.14 | 0.20 | 7 |
| FGFR1 | 0.25 | 0.08 | 0.12 | 13 |
| FGFR2 | 0.27 | 0.27 | 0.27 | 11 |
| FGFR3 | 0.60 | 0.27 | 0.37 | 11 |
| FGFR4 | 0.00 | 0.00 | 0.00 | 4 |
| FGR | 0.67 | 0.17 | 0.27 | 24 |
| FLT3 | 0.76 | 0.34 | 0.47 | 38 |
| FLT4 | 0.40 | 0.19 | 0.26 | 21 |
| FYN | 0.47 | 0.31 | 0.37 | 26 |
| GRK7 | 0.50 | 0.40 | 0.44 | 5 |
| GSK3A | 0.31 | 0.21 | 0.25 | 19 |
| GSK3B | 0.17 | 0.08 | 0.11 | 13 |
| HCK | 0.56 | 0.19 | 0.29 | 26 |
| HIPK1 | 0.33 | 0.12 | 0.17 | 17 |
| IGF1R | 0.33 | 0.22 | 0.27 | 9 |
| INSR | 0.75 | 0.25 | 0.38 | 12 |
| IRAK4 | 0.50 | 0.22 | 0.31 | 9 |
| ITK | 0.20 | 0.13 | 0.15 | 8 |
| KIT | 0.68 | 0.30 | 0.41 | 44 |
| LCK | 0.36 | 0.24 | 0.29 | 33 |
| MAPKAPK2 | 0.00 | 0.00 | 0.00 | 4 |
| MARK1 | 0.25 | 0.14 | 0.18 | 7 |
| MELK | 0.17 | 0.09 | 0.12 | 11 |
| MET | 0.38 | 0.17 | 0.23 | 18 |
| MUSK | 0.20 | 0.14 | 0.17 | 7 |

| | | | | |
|---|---|---|---|---|
| NEK2 | 0.38 | 0.30 | 0.33 | 10 |
| NEK6 | 0.00 | 0.00 | 0.00 | 4 |
| NEK7 | 0.00 | 0.00 | 0.00 | 4 |
| PAK2 | 0.20 | 0.33 | 0.25 | 6 |
| PAK3 | 0.50 | 0.13 | 0.20 | 8 |
| PAK6 | 1.00 | 0.14 | 0.25 | 7 |
| PHKG2 | 0.50 | 0.50 | 0.50 | 6 |
| PIM1 | 0.50 | 0.13 | 0.21 | 15 |
| PIM2 | 0.60 | 0.30 | 0.40 | 10 |
| PIM3 | 0.33 | 0.11 | 0.17 | 9 |
| PLK1 | 0.67 | 0.55 | 0.60 | 11 |
| PRKX | 0.67 | 0.18 | 0.29 | 11 |
| RET | 0.50 | 0.18 | 0.26 | 28 |
| ROCK1 | 0.60 | 0.18 | 0.27 | 17 |
| ROCK2 | 0.50 | 0.24 | 0.32 | 21 |
| SGK2 | 0.33 | 0.14 | 0.20 | 7 |
| SGK3 | 1.00 | 0.10 | 0.18 | 10 |
| SRC | 0.50 | 0.17 | 0.26 | 23 |
| SRPK1 | 0.20 | 0.07 | 0.10 | 15 |
| SYK | 0.13 | 0.09 | 0.11 | 11 |
| TBK1 | 0.29 | 0.13 | 0.18 | 15 |
| TEC | 0.67 | 0.15 | 0.25 | 13 |
| TNK2 | 0.00 | 0.00 | 0.00 | 13 |
| TXK | 0.14 | 0.07 | 0.09 | 15 |
| TYRO3 | 0.13 | 0.13 | 0.13 | 8 |
| ZAP70 | 1.00 | 0.14 | 0.25 | 7 |
| | | | | |
| **Макро среднее** | 0.42 | 0.18 | 0.23 | 1051 |
| **Микро среднее** | 0.42 | 0.20 | 0.27 | 1051 |
| **Взвешенное среднее** | 0.47 | 0.20 | 0.27 | 1051 |

| | | | | |
|---|---|---|---|---|
| **Среднее по объекту** | 0.19 | 0.13 | 0.14 | 1051 |

Таблица 3.5 Точность классификации 1D CNN модели на тестовой выборке с применением подобранных порогов бинаризации

Статистический анализ значимости полученных результатов проводился с помощью библиотеки SciPy для Python. Распределение данных оценок F1 было проверено с помощью теста Шапиро-Уилка. Был выполнен тест Манна-Уитни, при котором различия считались значимыми при значениях p ниже 0,05.

## 3.9. Предсказание потенциальных мишеней малой молекулы: применение GNN модели для аннотации крупной библиотеки соединений

Для проведения дополнительного тестирования, а также предоставления ценных данных для открытия киназоориентированных лекарств, GNN модель (как лучшая из разработанных) была применена к 81 515 структурам из виртуальной библиотеки соединений ChemDiv. Данные структуры были предварительно отобраны и отфильтрованы для анализа экспертами на основе нескольких критериев, в основном направленных на исключение очевидных некиназных ингибиторов и неподобных лекарств через предварительную фильтрацию по подструктуре. Принимая во внимание неравномерную точность предсказания тех или иных мишеней, для аннотации данного набора структур были отобраны 50 лучших мишеней на основании макро средней оценки F1 на тестовой выборке (минимальное макросреднее значение F1 меры на данных 50 мишенях составило 0.44). После применения GNN модели к этим структурам 20 259 молекул были помечены как потенциально активные против хотя бы одной из оцениваемых 50 мишеней (Рисунок 3.3). Распределение количества потенциальных мишеней кажется разумным, причем большинство аннотированных молекул имеют от 1 до 5 потенциальных ингибиторов, и очень немногие демонстрируют более 20 потенциальных мишеней.

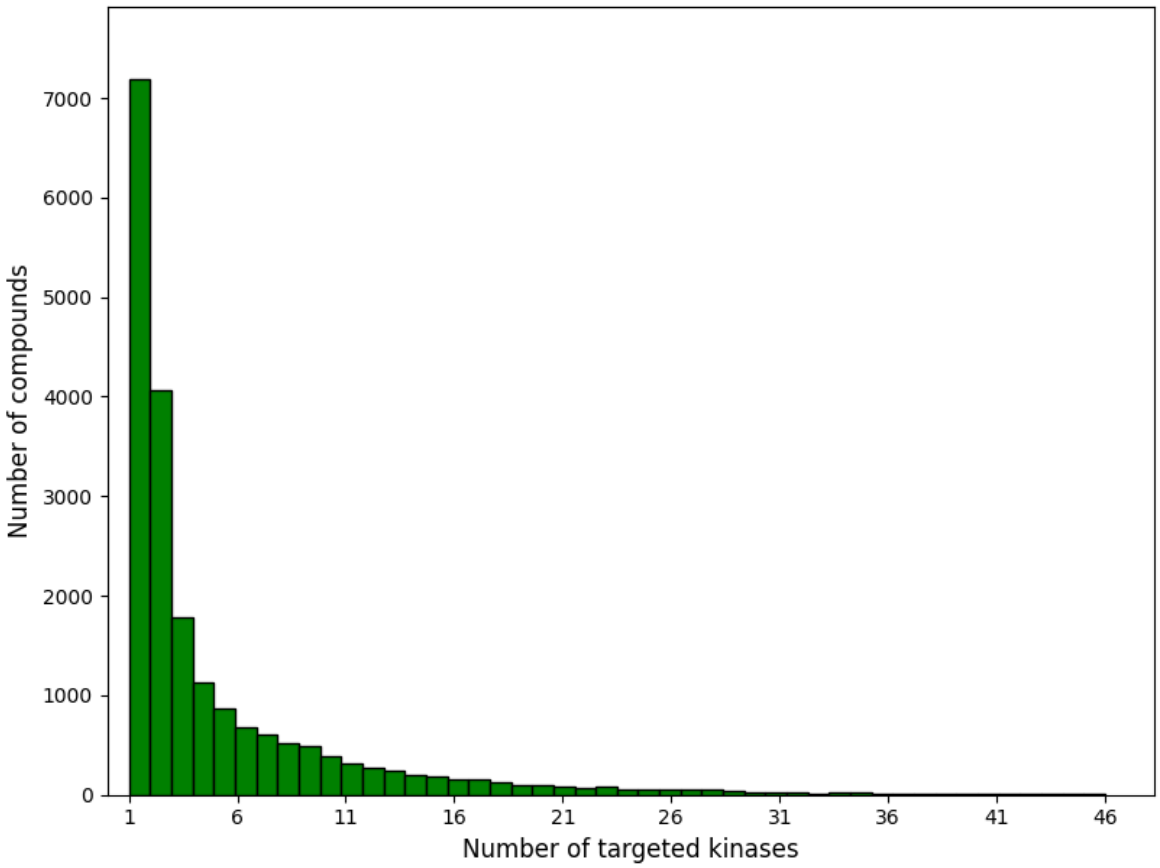


Рисунок 3.3. Статистика по 20 259 аннотированным соединениям из внешней базы данных

Был проведен экспертный анализ аннотированных соединений с целью выявления структур, которые могли бы теоретически связывать активный сайт белка (симуляция была проведена с помощью программ виртуального докирования), под который они были аннотированы. Процедура включала исследование расположения и ключевых связей, формируемых известным ингибитором с сайтом связывания данной молекулы, и выявления возможности подобного размещения и формировании данных связей аннотированной молекулой.

Были идентифицированы потенциальные ингибиторы следующих мишеней: EGFR и ERBB4 (двойной ингибитор), KIT, KIT и FLT3 (двойной ингибитор), LCK. Хотя без физического экспериментального подтверждения активности данные результаты являются только предположительными, проведенные экспертами исследования показали возможность разработанной модели идентифицировать малые молекулы, близкие к реальным ингибиторам.

## 3.10. Предсказание растворимости малой молекулы: обзор предметной области

Несмотря на достаточно обширные наборы доступных данных и значительного объема проведенных в данной области исследований, точное предсказание растворимости малых молекул с применением вычислительных методов (как использующих, так и не использующих ML) является сложной задачей. Появляется все больше информации, подтверждающей, что растворимость лекарств in vitro может недооценивать настоящую растворимость in vivo [72]. Исследование, проведенное в 2010 году, оценило, что 40% доступных лекарств обладают низкой растворимостью [95], а проведенное уже в 2014 году исследование оценило, что около 70% молекул, находящихся в разработке, имеют низкую растворимость [96].

Для решения задачи предсказания растворимости были разработаны различные алгоритмы. В [45] представлен обзор многих существующих алгоритмов для предсказания растворимости: полуэмпирические методы, такие как пересмотренное уравнение растворимости [31], не использующее подогнанные параметры, UNIFAC, метод, объединяющий концепцию функциональных групп с коэффициентами активности, основанными на квазихимической теории жидких смесей [32]. Также применяются методы, основанные на энергии, использующие симуляцию кристаллической решетки и расчет свободной энергии [33], методы молекулярной динамики [34,35], модели, основанные на количественных данных взаимосвязи между структурой и активностью [36,37]. Технологии с использованием ИИ активно внедряются в данную область, с их применением разрабатываются различные вычислительные решения. Основной целью использования ИИ для оценки растворимости является обеспечение аналогичной или более высокой точности по сравнению с полуэмпирическими методами, при этом избегая длительных и сложных расчетов, основанных на энергии и динамическом моделировании. В обзоре, приведенным в работе [45], представлены самые разные ИИ архитектуры: [38] исследовал и сравнил применение случайного леса (RF), метода частичных наименьших

квадратов, метода опорных векторов и искусственных нейронных сетей, [39] разработал модели с применением CNN, рекуррентной нейронной сети, DNN, а также спайковых нейронных сетей. В [40] был реализован ансамбль рекурсивных нейронных сетей. Были проведены эксперименты по сочетанию различных архитектур ИИ с методами вычислительной химии в [41].

Как и в случае ранее рассмотренной задачи предсказания потенциальных мишеней малой молекулы, доступные базы данных о малых молекулах с известными свойствами (обычно данные свойства рассчитываются экспериментально в лабораториях) не хранят информацию о 3D конформациях молекул: в них доступна только структурная информация в SMILES формате. В связи с этим, как было в том числе показано в ходе обзора существующих решений, большинство современных методов машинного обучения в том или ином виде используют дескрипторы, основанные на SMILES записи, из которой можно извлечь как физико-химические признаки, так и структурную информацию. Важным отличительным аспектом подобных решений является подход к признаковому представлению молекулы. В рамках исследования решения задачи предсказания мишеней, был показан результат применения различных признаковых описаний, основанных на SMILES записи; эти же дескрипторы (за исключением Mordred дескрипторов) были применены и для решения задачи предсказания растворимости. Данное решение было принято, во-первых, с целью дополнительно исследовать влияние различных представлений данных на точность предсказаний подготовленных на их основании моделей, но в другом приложении; во-вторых, сбор различных физико-химических свойств, например, число тяжелых атомов или молекулярный вес, является общим шагом для почти всех существующих решений, использующих машинное обучение, для предсказания растворимости.

Таким образом, в данном разделе рассмотрены следующие из ранее примененных подходов к декодированию SMILES молекул: представление в виде отпечатков Моргана, токенизация символов SMILES и их рассмотрение как последовательности закодированных токенов, а также представление молекулы в

виде графа. Важным дополнительным отличием является то, что для построения GNN модели оценки растворимости, не использовалась генерация трехмерных конформаций – как описано в [45], молекулярный граф был построен на основании доступной информации об атомах и их связях, без использования трехмерного моделирования и соответствующих признаков (межатомных расстояний), а также без использования фармокофорного моделирования. Данное изменение было введено по следующей причине: хотя GNN на основании ансамбля 3D фармакофрного представления молекулы показала себя как наиболее точный подход из рассмотренных на задаче предсказания мишеней, подобное 3D моделирование является вычислительно затратным: в частности, разметка выше описанной библиотеки из 81 515 структур заняла порядка 3 часов на имеющимся вычислительном оборудовании. В связи с этим, представляло интерес проверить, сохранится ли преимущество GNN подхода при работе с более простым графовым представлением.

С использованием набора данных из 70 710 SMILES записей молекул с известной растворимостью, были обучены и сравнены ИИ-модели на основе трех различных архитектур и подходов к представлению данных: GNN с графовым оператором (3), модель градиентного бустинга с использованием библиотеки CatBoost, а также 1D CNN.

### 3.11. Предсказание растворимости малой молекулы: постановка задачи

Так как решается схожая по своей основе работа (предсказание свойства молекулы по ее SMILES представлению), постановка задачи схожа с представленной в 3.3. Основным отличием является источник набора данных $D$, а также целевая величина $y_k$.

Использованный в данном эксперименте набор данных был взят из недавно опубликованного соревнования Kaggle «1st EUOS/SLAS Joint Challenge: Compound Solubility», для которого данные подготовила организация EU-OPENSCREEN ERIC [97]. Набор данных содержит 101 017 соединений, из которых 70 710 доступны для обучения моделей, а 30 307 были отделены

организаторами для тестирования моделей – для этих данных не были предоставлены метки классов.

Данные были предоставлены в табличном виде, где каждая строка содержала запись молекулы в SMILES формате, идентификатор соединения, а также ее класс растворимости (для тестовых выборок данная колонка отсутствовала). Целью конкурса было правильно предсказать метку растворимости соединения, которая могла быть высокой, средней или низкой (в предоставленном наборе данных обозначены как 2, 1 и 0 соответственно), в соответствии с классификацией, проведенной организаторами соревнования. Таким образом, $y_k$ в данной задаче это метка одного из 3 возможных классов.

Подход к разработке используемых для решения данной задачи моделей схож с описанным в 3.3, и был построен по следующему принципу – каждая архитектура использует свой подход к описанию структуры молекулы (отпечатки Моргана для модели градиентного бустинга, токенизированная последовательность для 1D CNN модели, графовое представление для GNN модели), с добавлением вышеописанных 11 физико-химических дескрипторов для описания общих свойств молекулы. Таким образом, модель градиентного бустинга Catboost была применена без изменений типа входящих дескрипторов. 1D CNN была применена с дополнительным применением физико-химических дескрипторов. GNN модель перетерпела наибольшие изменения: как было сказано ранее, GNN использует графы, построенные по доступной информации об атомах и их связях из SMILES молекулы, также с добавлением физико-химических дескрипторов.

## 3.12. Предсказание растворимости малой молекулы: предобработка обучающих данных

Как описано в [45], в обучающей выборке из 70 710 соединений содержалось 65 834 соединения с высокой растворимостью, 2 835 соединений со средней растворимостью и 2 041 соединение с низкой растворимостью. Для подготовки обучающих, валидационных и независимых тестовых наборов

данных, была использована та часть опубликованного набора данных, которая была отмечена как обучающая, без дополнительной аугментации. Валидационные выборки были использованы для контроля переобучения, тестовая выборка была использована для сравнения моделей. 30 307 соединений из исходного набора, которые были отмечены организаторами как тестовые (и не имели значений растворимости в открытом доступе) были использованы для дополнительного независимого тестирования моделей.

В рамках предобработки данных, описанной в [45], была проанализирована длина строковых описаний соединений в формате SMILES. Хотя 99,8% всех соединений в формате SMILES содержали от 18 до 95 символов, были выявлены некоторые выбросы, достигающие 226 символов и минимально 7 символов. Поскольку все 3 разрабатываемые модели чувствительны к размеру входной структуры, выбросы были удалены. Независимый тестовый набор был сформирован путем отделения 10% данных случайным образом. Оставшиеся данные были разделены на 5 случайных обучающих и валидационных разбиений. Как и в ходе проведения исследования по предсказанию наличия ингибирующей активности молекул против белковых мишеней, одни и те же разбиения были использованы для обучения моделей всех трех архитектур, для обеспечения обучения моделей на одинаковых исходных данных. Как и в ранее рассмотренной задаче, это было сделано для обеспечения более объективного сравнения моделей.

Как было сказано ранее, классы растворимости в представленных данных не были сбалансированы (класс высокой растворимости был намного более представлен). В связи с этим к предсказаниям обученных моделей были применены экспоненциальные балансировочные коэффициенты. Коэффициенты были настроены для получения максимального квадратичного коэффициента Каппа (коэффициента Каппа) на обучающем наборе данных, используя реализацию алгоритма минимизации Пауэлла [64] в библиотеке Python Sklearn [63].

## 3.13. Предсказание растворимости малой молекулы: подготовка и обучение моделей

Входные признаки для модели Catboost состояли из 11 физико-химических дескрипторов и вектора отпечатка Моргана размером 1024 бита с радиусом 2. Данные параметры отличаются от параметров, использованных в рамках задачи предсказания наличия ингибирующей активности, но они был получены с применением такой же методологии, а именно с применением поиска по сетке с пятикратной кросс-валидацией. Кросс валидация была выполнена на объединении валидационной и обучающей выборок. Итоговые параметры для модели градиентного бустинга были следующими: скорость обучения 0,1, глубина дерева 10 и регуляризация листьев L2 1. Для каждого из пяти тренировочных и валидационных разделений данных была обучена отдельная модель. Обучение каждой модели проводилось с возможностью ранней остановки в случае, если значение функции потерь на валидации не улучшалось в течение 400 итераций, с фиксацией итерации с наименьшим значением потери на валидации. Обучение одной модели требовало около 2,5 секунд на 100 итераций.

Для модели 1D CNN входная последовательность состояла из токенизированной последовательности символов. Токенизация строкового представления SMILES была проведена аналогично с тем, как это было сделано в ходе эксперимента по предсказанию ингибирующей активности малых молекул: числовые значения были присвоены часто встречающимся символам (с комбинациями, обозначающими единственное свойство или имя атома, например "Cl" или "Na", закодированными как один токен). Общее количество уникальных токенов (включая токены, зарезервированные для заполнения или неизвестных случаев) составило 40. Все последовательности SMILES были токенизированы данным образом. Максимальный размер последовательности был равен 95, соответствуя максимальному размеру строки SMILES (в символах) в обучающих данных, после исключения выбросов. Архитектура 1D CNN модели, представленной в [45], включала слой с размерностью 256 для обучения векторных представлений токенов. Векторное представление передавалось в блок

CNN, состоящий из пяти слоев 1D CNN, каждый из которых имел увеличивающееся количество фильтров, начиная с 64 и удваивающееся в каждом последующем слое, с фильтром размерностью 5, использующим функцию активации ReLU. Выходной ответ CNN блока проходил через слои максимальной и средней субдискретизации, выходные вектора которых были объединены. К объединенному вектору дополнительно добавлялись 11 физико-химических дескрипторов, описанных ранее. Итоговый вектор обрабатывался многослойным линейным блоком с использованием функции активации ReLU с «утечкой», батч нормализации и 20% dropout слоев между линейным слоями. Модель обучалась батчами размером 2048, со скоростью обучения $10^{-4}$, используя алгоритм оптимизации Adam и функцию потерь кросс-энтропия. Для каждого из пяти тренировочных и валидационных разделений обучение проводилось в течение 60 эпох, с возможностью ранней остановки, если значение функции потерь на валидации не улучшалось в течение 10 эпох. Фиксировались веса моделей, полученных на эпохе обучения на которой было достигнуто наименьшее значение функции потерь на валидационной выборке. Обучение одной модели требовало около двух секунд на одну обучающую эпоху. В целом данная архитектура и подход к обучению были аналогичным выбранным в ходе эксперимента по предсказанию мишеней малой молекулы, но с введением физико-химических дескрипторов.

Для подготовки данных для GNN модели, как описано в [45], было сформировано 2D графовое представление. Оно было создано по информации о связях и свойствах атомов, извлеченных из SMILES с помощью библиотеки RDKit. Узлы представляли атомы молекулы и хранили в себе соответствующее признаковое описание. Ребра графа создавались только между атомами, имеющими связи. Так как связи между атомами в данном случае являются двунаправленными, графы были сделаны неориентированными.

Всего для обучения модели были использованы следующие признаки: имя атома (12 категорий), тип гибридизации (5 категорий), степень атома, определенная количеством связанных с ним соседей (в диапазоне от 0 до 5),

общее количество неявных и явных водородов (в диапазоне от 0 до 5), неявная валентность (в диапазоне от 0 до 5), является ли атом частью кольца и является ли он ароматическим (бинарные значения). Были собраны следующие признаки связей, соединяющих атомы: тип связи (4 категории), является ли связь частью кольца и является ли она сопряженной (бинарные признаки). Данное признаковое описание было извлечено с применением библиотеки RDKit, используя SMILES запись молекулы.

Все числовые признаки были нормализованы. Часть признаков были категориальными, и для них были созданы обучаемые векторные представления: 16-мерные векторы генерировались для признаков имени атома и гибридизации, и 8-мерный вектор генерировался для признака типа связи. Результирующие векторы затем объединялись с оставшимися числовыми признаками ребер и узлов. Признаки связей дополнительно передавались через многослойный линейный блок с выходным размером 16.

В соответствии с архитектурой, представленной в [45], основным элементом разработанной нейросети является GNN блок состоящий из 7 GNN слоев. Данные слои использовали графовый оператор, аналогичный GNN сетям, описанным в предыдущих разделах. Количество нейронов и голов внимания были подобраны под данную задачу: было использовано 8 голов внимания и 64 нейрона в каждом GNN слое. Для ускорения обучения и снижения переобучения в GNN блоке использовался 20% dropout на уровне узлов (позволяя случайным образом обнулять сигнал от тех или иных связанных ребрами вершин) и батч нормализация. В данный блок поступал входящий 37-мерный вектор, описывающий молекулярную структуру. Вывод данного блока проходил через максимальную, среднюю и суммирующую субдискретизацию. Получаемые в результате три вектора объединялись, затем проходили батч нормализацию. После нормализации данный вектор объединялся с 11 физико-химическими дескрипторами. Итоговый вектор проходил через многослойный линейный блок с использованием функции активации ReLU с «утечкой» (параметр отрицательного наклона был равен 0.2), батч нормализации и 50% dropout слоев между линейным

слоями. Данный линейный блок выдавал финальный вектор с предсказаниями по каждому из рассматриваемых классов растворимости.

Обучение модели было проведено батчами размером 2048, со скоростью обучения $10^{-3}$, используя алгоритм оптимизации Adam и функцию потерь кросс-энтропия. Для каждого из пяти тренировочных и валидационных разделений обучение проводилось в течение 100 эпох, с возможностью ранней остановки, если значение функции потерь на валидации не улучшалось в течение 10 эпох. Фиксировались веса моделей, полученных на эпохе обучения, на которой было достигнуто наименьшее значение функции потерь на валидационной выборке. Обучение одной модели требовало около 7 секунд на одну эпоху.

Все модели были обучены на графической карте NVIDIA RTX 4090 с 24 Гб видеопамяти.

### 3.14. Предсказание растворимости малой молекулы: результаты на независимом тестовом наборе

В соответствии с методологией, описанной в [45], сравнение моделей было произведено по коэффициенту Каппа Коэна. Данная метрика была использована организаторами конкурса в качестве основного оценочного показателя, и сравнение моделей на представленных организаторами тестовых данных было возможно только по ней. Дополнительно, на созданной внутренней тестовой выборке, модели были оценены по средней точности (AP). Для обеспечения более надежного сравнения, предсказания моделей, обученных на каждом из пяти разбиений обучающих данных, были усреднены. В ходе тестирования, для оценки коэффициента Каппа Коэна, использовались предсказания моделей с применением соответствующих балансировочных коэффициентов, т.к. данная метрика использует конечные метки классов. Для метрики AP использовались предсказания без коэффициентов.

Как показало исследование, проведенное в [45], наилучшие результаты по выбранным метрикам качества показала модель CatBoost. Рассматривая усредненные предсказания пяти моделей каждой архитектуры на тестовой

выборке, 1D CNN, GNN и CatBoost показали коэффициенты Каппа 0.0356, 0.0597 и 0.0708 соответственно, продемонстрировав значительное различие (p-значение < 0.05) между собой по данной метрике. Коэффициенты Каппа каждой из 5 обученных моделей для каждой архитектуры представлены на рисунке 3.4.

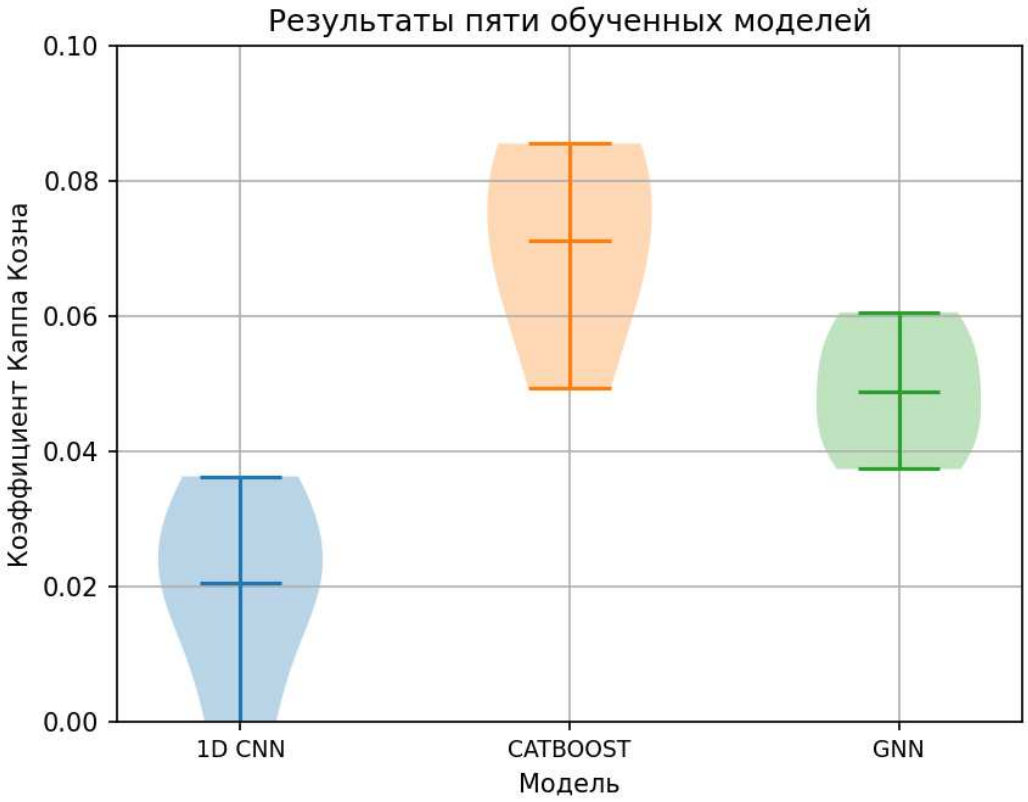


Рисунок 3.4. Сравнение обученных моделей на независимой выборке данных. Источник [45]

Как можно видеть, разные тренировочные разбиения показали схожие результаты: модели CatBoost дали лучшие результаты, а 1D CNN – худшие, с статистически значимой разницей (p-значения < 0,05). Модель на основе CatBoost также показала преимущество в рамках метрики AP, но с меньшей разницей. Усредненные модели CatBoost, GNN и 1D CNN показали AP 0,347, 0,343 и 0,337 соответственно на внутренней тестовой выборке данных, с статистически значимой разницей (p-значения < 0,05). Как можно видеть, в сравнении с результатами, полученными в ходе решения задачи предсказания мишеней малой молекулы, 1D CNN вновь показала себя самой слабой (в рамках оцениваемых критериев), несмотря на дополнительное использование физико-химических

дескрипторов. Catboost модель, которая использовала то же признаковое пространство что и в задаче предсказания мишеней, показала наивысшую точность по используемым метрикам. GNN модель, которая более не использует трехмерное представление молекул, оказалась менее точной, чем Catboost, но более точной, чем 1D CNN.

В ходе проведенного в [45] эксперимента было обнаружено, что все модели показали значительно более слабые результаты на классах с низкой и средней растворимостью. В частности, при показателе AP 0,94 для класса с высокой растворимостью, модель CatBoost показала только 0,05 AP для классов с низкой и средней растворимостью. Такие результаты могут быть связаны с природой данных, а не с выбранными архитектурами. Данное предположение исходит из результатов, которых разработанные модели достигли на тестовых выборках, сформированных организаторами соревнования «1st EUOS/SLAS Joint Challenge: Compound Solubility», в сравнении с лучшими представленными на конкурсе вычислительными решениями.

## 3.15. Предсказание растворимости малой молекулы: результаты на тестовых выборках соревнования

Как было упомянуто ранее, организаторами соревнования были опубликованы два тестовых набора – закрытый и открытый. Открытый набор использовался в ходе проведения соревнования, для предоставления участниками возможности примерно оценить результативности разработанных ими методов. Закрытый набор использовался для выбора победителей конкурса: рассчитанные по нему оценки не были видны в ходе соревнования, но доступны в настоящий момент. Для оценки разработанных в ходе проведенного исследования моделей на этих данных, размеченные таблицы по данным выборками загружались на веб-сайт Kaggle для обработки скриптом, разработанным организаторами. Данный скрипт был единственным способом рассчитать коэффициенты Каппа Коэна, т.к. истинные метки по этим данным не были предоставлены.

Разметка закрытого и открытого тестов проводилась так же, как и в ходе тестирования на внутренней тестовой выборке: усредненные прогнозы 5 обученных моделей каждой рассмотренной архитектуры, с применением балансирующих коэффициентов, использовались для классификации всех соединений из предоставленных тестовых выборок данных. Как показано в [45], в данном тесте разработанные модели продемонстрировали такую же сравнительную точность, что и в ходе локального тестирования. CatBoost показал наивысшие оценки, с 0,09676 на приватном наборе и 0,11903 на публичном наборе, 1D CNN показала наименьшие оценки, с 0,06678 на закрытом наборе и 0,07219 на открытом наборе, а GNN показала результаты лучше чем у 1D CNN, но хуже чем у CatBoost: 0,07047 на приватном наборе и 0,11456 на публичном наборе. Как можно видеть, GNN модель показала наибольшую разницу в своей точности между открытым и закрытым наборами данных – если на открытом наборе данных ее показатели близки к показателям CatBoost, то на закрытом тесте она проигрывает значительно больше.

Представленный организаторами набор данных был использован в первую очередь для сравнения разработанных моделей, а не для сравнения с другими представленными в ходе соревнования решениями. Основной причиной для этого является тот факт, что для достижения наиболее высоких метрик на соревновании, лучшие решения были адаптированы непосредственно под данный набор данных, в то время как целью проведенного исследования было сравнение подходов для решения данной задаче в более общем применении, с возможность использования полученных в ходе работы выводов для других выборок данных. Кроме того, лучшие решения соревнования дополнительно использовали (исходя из опубликованной информации) идентификационный номер соединения как часть признакового описания структур, поскольку во время соревнования было обнаружено, что он коррелирует с предсказываемыми классами растворимости. Очевидно, что идентификатор соединения не влияет на растворимость и не должен оказывать никакого влияния на точность предсказаний, однако при его

применении обученные модели показывали более высокие значения целевой метрики.

В работе [45] был проведен дополнительный эксперимент для демонстрации этого эффекта. Были дополнительно обучены пять моделей CatBoost, которые учитывали идентификаторы соединений помимо других используемых дескрипторов, при этом процедура обучения проводилась так же, как и в рамках основного эксперимента, описанного выше. Усреднение полученных в результате пяти моделей показало оценку 0,10903 на закрытой выборке и оценку 0,13323 на открытой выборке, что позволило бы этому решению занять третье место в соревновании, где лучшее решение показало коэффициент Каппа 0,11562 (исключая четыре решения, которые были дисквалифицированы организаторами соревнования). Одна из пяти обученных моделей показала еще более высокую оценку 0,11021 на закрытой выборке [45]. Таким образом явно продемонстрировано, что использование признака, не имеющего практического или физического смысла, позволяет повысить точность моделей на данных выборках ввиду того, как именно были сформированы идентификаторы соединений организаторами соревнования.

Статистический анализ проводился с помощью библиотеки SciPy Python [62]. Распределение данных коэффициентов Каппа проверялось с помощью теста Шапиро-Уилка. Если распределения были нормальными, выполнялся т-тест Стьюдента, в противном случае выполнялся тест Манна-Уитни. Различия считались значимыми для p-значений ниже 0,05.

## 3.16. Выводы

В данной главе было продемонстрировано применение GNN для анализа свойств молекулярных структур, на примере задач предсказания потенциальных мишеней малой молекулы (на примере выбранных 75 киназ) и предсказания растворимости. Для оценки относительной результативности GNN, были дополнительно обучены модели сравнения, основанные на других способах описания малых молекул и иных ML архитектурах, в частности DNN, 1D CNN,

градиентный бустинг с применением CatBoost. Таким образом, данное исследование позволило как сравнить GNN подход с другим методами для обработки молекулярных данных, так и изучить различные подходы к признаковому представлению молекулярных данных.

В рамках задачи по предсказанию мишеней, обучение моделей было проведено на относительно небольшой (1176), экспертно подобранной выборке данных. Данный масштаб позволил применить более вычислительно сложную реализацию GNN, а именно с применением 3D симуляции возможных конформаций молекулы (их истинная активная конформация не была известна) и фармакофорного моделирования. По результатам проведенных на пяти разных разделениях выборки на обучение и валидацию экспериментов, построенная подобным образом GNN превзошла модель градиентного бустинга, которая была основана на представлении данных в виде молекулярных отпечатков (для описания структуры) и физико-химических признаков (для описания общих свойств молекулы), 1D CNN модель, основанной на токенизации SMILES записей молекул и обучения на последовательности их векторных представлениях, и DNN модель, основанной на крупном наборе физико-химических дескрипторов, полученных с применением библиотеки Mordred, и обработанных PCA методом с целью создания более компактного представления.

В рамках экспертной валидации, была аннотирована крупная база из 81 515 структур, в рамках которой 20 259 из них был размечены как потенциальной активные против одной либо более из рассматриваемых киназ. Экспертный анализ выявил несколько примеров структур, схожих с реальными ингибиторами киназ, по которым они были размечены как активные.

При решении задачи предсказания растворимости молекул, была задействована более крупная выборка (70 000 структур) для обучения моделей, разработанная EU-OPENSCREEN ERIC и опубликованная в рамках открытого соревнования на Kaggle. В связи с этим было принято решение протестировать GNN, принимающую на вход упрощенное графовое представление молекул, без генерации 3D конформаций и фармакофорного представления. По результатам

проведенных на пяти разных разделениях выборки на обучение и валидацию экспериментов, модель CatBoost, а не GNN, проявила себя как наиболее точную, как на локально подготовленном независимом тестовом наборе, так и на тестовых наборах, опубликованных в вышеупомянутом соревновании. Это может быть объяснено тем, что табличное представление данных в виде молекулярных отпечатков и физико-химических дескрипторов лучше подходит для данной задачи: также возможно, что 3D представление молекулы позволяло GNN архитектуре наиболее полно задействовать свои преимущества по сравнению с другим архитектурами, чем в случае «плоского» графа. Стоит отметить, что задача предсказания растворимости, в рамках проведенного исследования, показала себя как весьма сложную для основанных на ИИ подходах. Все обученные модели продемонстрировали значительно более сильную предсказательную способность для класса высокой растворимости, чем для классов средней и низкой растворимости. Кроме того, модель градиентного бустинга на основе CatBoost, показавшая лучшие из обученных моделей результаты, смогла достичь коэффициента Каппа около 0,10, что указывает на достаточно слабую предсказательную способность. Лучшие из отобранных организаторами решений так же показали близкие к 0.1 значения. Подобные результаты могу быть вызваны неточностями разметки и обработки данных, которую провели организаторы соревнования, однако данный аспект не был исследован.

Вместе с тем, проведенные эксперименты показывают, что представления молекул, извлеченных из SMILES в виде графов или векторных представлений символов, не обязательно превосходят более традиционные табличные представления, состоящие из отпечатков Моргана и химических дескрипторов.

## Заключение

Основные результаты диссертационной работы.

1. Разработан алгоритм поиска сайтов связывания белковых молекул с применением графовых нейронных сетей и представления трехмерной структуры белка и окружающего его пространства в виде графа. Продемонстрировано, что разработанный алгоритм способен более точно определять сайты связывания белка, чем известные решения, применяемые в данной области.

2. Предложен алгоритм аннотации сайтов связывания белковых молекул с применением графовых нейронных сетей. Показано, что графовые нейронные сети и графовое представление трехмерной структуры белка могут быть применены для аннотации сайтов связывания белковых молекул. Продемонстрирована более высокая точность аннотации по сравнению с распространенными алгоритмами, применяемыми в данной области, построенными на иных принципах. Продемонстрирована применимость оценок, генерируемых разработанным алгоритмом, в качестве признакового описания пространства поверхности белка. Показано, что модели, обученные с использованием данного признакового описания, демонстрируют сравнимую или превосходящую точность по сравнению с существующими аналогами.

3. Исследовано применение графовых нейронных сетей для предсказания свойств малых молекул, а именно ингибирующей активности против выбранных 75 белков, и растворимости. Продемонстрирована более высокая точность разработанного алгоритма по сравнению с известными алгоритмами машинного обучения в задаче предсказания ингибирующей активности.

## Список литературы


1. Leelananda, S. P., Lindert, S. Computational methods in drug discovery // Beilstein J Org Chem. – 2016. – 12. – Pp. 2694-2718. DOI: 10.3762/bjoc.12.267
2. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E. The protein data bank. Nucleic Acids Res. – 2000. – Vol. 28. – Pp. 235−242. DOI: 10.1093/nar/28.1.235
3. Абгарян К.К., Журавлев А.А. Методы многомасштабного моделирования в задачах цифрового материаловедения. – Москва: МАКС ПРЕСС, 2022. DOI: 10.29003/m3138.978-5-317-06870-7
4. Le Guilloux, V., Schmidtke, P., Tuffery, P. Fpocket: an open source platform for ligand pocket detection // BMC Bioinformatics. – 2009. – Vol. 10. – Pp. 168-179. DOI: 10.1186/1471-2105-10-168
5. Konc, J., Janezic, D. ProBiS: a web server for detection of structurally similar protein binding sites // Nucleic Acids Res. – 2010. – Vol. 38. – Pp. 436-440. DOI: 10.1093/nar/gkq479
6. Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A.S., De Fabritiis G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks // Bioinformatics. – 2017. – Vol. 33. – Pp. 3036-3042. DOI: 10.1093/bioinformatics/btx350
7. Kozlovskii, I., Popov, P. Spatiotemporal identification of druggable binding sites using deep learning // Commun Biol. – 2020. – Vol. 3. – Pp. 618-630. DOI: 10.1038/s42003-020-01350-0
8. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., … Hassabis, D. Highly accurate protein structure prediction with AlphaFold // Nature. – 2021. – Vol. 596(7873). – Pp. 583–589. DOI: 10.1038/s41586-021-03819-2

9. Corso G., Jing B., Stark H., Barzilay R., Jaakkola T. Blind Protein-Ligand Docking with Diffusion-Based Deep Generative Models // Biophys. J. – 2023. – Vol. 122 (3). – 143a. DOI: 10.1016/j.bpj.2022.11.937
10. Ivanenkov, Y., Zagribelnyy, B., Malyshev, A., Evteev, S., Terentiev, V., Kamya, P., Bezrukov, D., Aliper, A., Ren, F., Zhavoronkov, A. The Hitchhiker's Guide to Deep Learning Driven Generative Chemistry // ACS medicinal chemistry letters. – 2023. – Vol. 14(7). – Pp. 901–915. DOI: 10.1021/acsmedchemlett.3c00041
11. Scott, O. B., Gu, J., Chan, A. E. Classification of protein-binding sites using a spherical convolutional neural network // Journal of Chemical Information and Modeling. – 2022. – Vol. 62(22). – Pp. 5383-5396. DOI: 10.1021/acs.jcim.2c00832
12. Pu, L., Govindaraj, R. G., Lemoine, J. M., Wu, H. C., Brylinski, M. DeepDrug3D: Classification of ligand-binding pockets in proteins with a convolutional neural network // PLoS computational biology. – 2019. – Vol. 15(2). – P. e1006718. DOI: 10.1371/journal.pcbi.1006718
13. Shi, W., Lemoine, J. M., Shawky, A. A., Singha, M., Pu, L., Yang, S., Ramanujam, J., Brylinski, M. BionoiNet: ligand-binding site classification with off-the-shelf deep neural network // Bioinformatics (Oxford, England). – 2020. – Vol. 36(10). – Pp. 3077–3083. DOI: 10.1093/bioinformatics/btaa094
14. Jiang, D., Hsieh, C.Y., Wu, Z., Kang, Y., Wang, J., Wang, E., Liao, B., Shen, C., Xu, L., Wu, J., Cao, D., & Hou, T. InteractionGraphNet: A Novel and Efficient Deep Graph Representation Learning Framework for Accurate Protein-Ligand Interaction Predictions. // Journal of medicinal chemistry. – 2021. – Vol. 64(24). – Pp. 18209–18232. DOI: 10.1021/acs.jmedchem.1c01830
15. Mqawass, G., Popov, P. graphLambda: Fusion Graph Neural Networks for Binding Affinity Prediction // Journal of chemical information and modeling. – 2024. – Vol. 64(7). – Pp. 2323–2330. DOI: 10.1021/acs.jcim.3c00771
16. He, T., Heidemeyer, M., Ban, F., et al. SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines //

Journal of Cheminformatics. – 2017. – Vol. 9:24. DOI:10.1186/s13321-017-0209-z

17. Ma, X. H. et al. Virtual screening of selective multitarget kinase inhibitors by combinatorial support vector machines // Molecular Pharmaceutics. – 2010. – Vol. 7. – Pp. 1545–156. DOI: 10.1021/mp100179t
18. Bora, A., Avram, S., Ciucanu, I., Raica, M., Avram, S. Predictive Models for Fast and Effective Profiling of Kinase Inhibitors // Journal of Chemical Information and Modeling. – 2016. – Vol. 56 (5). – Pp. 895-905. DOI: 10.1021/acs.jcim.5b00646
19. Qu, D., Yan, A. Classification models and SAR analysis of anaplastic lymphoma kinase (ALK) inhibitors using machine learning algorithms with two data division methods // Molecular Diversity. – 2024. DOI: 10.1007/s11030-024-10990-x
20. Rohani, N., Eslahchi, C. Drug–drug interaction predicting by neural network using integrated similarity // Scientific Reports. – 2019. – Vol. 9. – Pp. 1–11. DOI: 10.1038/s41598-019-50121-3
21. Vijay, S., Gujral, T. S. Non-linear deep neural network for rapid and accurate prediction of phenotypic responses to kinase inhibitors // iScience – 2020. – Vol. 23. – 101129. DOI: 10.1016/j.isci.2020.101129
22. Öztürk, H., Özgür, A., Ozkirimli, E. DeepDTA: deep drug-target binding affinity prediction // Bioinformatics. – 2018. – Vol. 34. – Pp. i821–i829. DOI: 1093/bioinformatics/bty593
23. Thafar, M.A., Alshahrani, M., Albaradei, S., et al. Affinity2Vec: drug-target binding affinity prediction through representation learning, graph mining, and machine learning // Scientific Reports. – 2022. – Vol. 12. – Pp. 1–18. DOI: 10.1038/s41598-022-08787-9
24. Nguyen, T., Le, H., Quinn, T.P., et al. GraphDTA: predicting drug target binding affinity with graph neural networks // Bioinformatics. – 2021. – Vol. 37. – Pp. 1140–7. DOI: 10.1093/bioinformatics/btaa921

25. Yang, Z., Zhong, W., Zhao, L., et al. MGraphDTA: deep multiscale graph neural network for explainable drug-target binding affinity prediction // Chemical science. – 2022. – Vol. 13. – Pp. 816–33. DOI: 10.1039/d1sc05180f
26. Shin, B., Park, S., Kang, K., et al. Self-attention based molecule representation for predicting drug-target interaction // Proceedings of Machine Learning Research. – 2019. – Vol. 106. – Pp. 230–48. DOI: 10.48550/arXiv.1908.06760
27. Zhao, Q., Duan, G., Yang, M., et al. AttentionDTA: drug-target binding affinity prediction by sequence-based deep learning with attention mechanism // IEEE/ACM Transactions on Computational Biology and Bioinformatics. – 2023. – Vol. 20. – Pp. 852–63. DOI: 10.1109/TCBB.2022.3170365
28. Brahma, R., Shin, J.M., Cho, K.H. KinScan: AI-based rapid profiling of activity across the kinome // Briefings in Bioinformatics. – 2023. – Vol. 24(6). DOI: 10.1093/bib/bbad396
29. Park, H., Hong, S., Lee, M., et al. AiKPro: deep learning model for kinome-wide bioactivity profiling using structure-based sequence alignments and molecular 3D conformer ensemble descriptors // Scientific Reports. – 2023. – Vol. 13. – Pp. 1–12. DOI: 10.1038/s41598-023-37456-8
30. Kutlushina, A., Khakimova, A., Madzhidov, T., Polishchuk, P. Ligand-Based Pharmacophore Modeling Using Novel 3D Pharmacophore Signatures // Molecules. – 2018. – Vol. 23(12). – P. 3094. DOI: 10.3390/molecules23123094
31. Ran, Y., Samuel H. Yalkowsky, S.H. Prediction of Drug Solubility by the General Solubility Equation (GSE) // Journal of Chemical Information and Computer Sciences. – 2001. – Vol. 41 (2). – Pp. 354-357. DOI: 10.1021/ci000338c
32. Fredenslund, A., Jones, R. L. Prausnitz, J. M. Group-contribution estimation of activity coefficients in nonideal liquid mixtures // AIChE J. – 1975. – Vol. 21. – Pp. 1086-1099. DOI: 10.1002/aic.690210607
33. Palmer, D.S., McDonagh, J.L., Mitchell, J.B.O., Mourik, T., Fedorov, M.V. First-Principles Calculation of the Intrinsic Aqueous Solubility of Crystalline Druglike

Molecules // Journal of Chemical Theory and Computation. – 2012. – Vol. 8. (9). – Pp. 3322-3337. DOI: 10.1021/ct300345m

34. Li, L., Totton, T., Frenkel, D. Computational methodology for solubility prediction: Application to the sparingly soluble solutes // J. Chem. Phys. – 2017. – Vol. 146 (21). – P. 214110. DOI: 10.1063/1.4983754
35. Boothroyd, S., Anwar, J. Solubility prediction for a soluble organic molecule via chemical potentials from density of states // The Journal of Chemical Physics. – 2019. – Vol. 151. – Pp. 184113. DOI: 10.1063/1.5117281
36. Duchowicz, P.R., Castro, E.A. QSPR Studies on Aqueous Solubilities of Drug-Like Compounds // Int. J. Mol. Sci. – 2009. – Vol. 10. – Pp. 2558-2577. DOI: 10.3390/ijms10062558
37. Yu, X., Wang, X., Wang, H., Li, X.,Gao, J. Prediction of Solubility Parameters for Polymers by a QSPR Model // QSAR Comb. Sci. – 2006. – Vol. 25. – Pp. 156-161. DOI: 10.1002/qsar.200530138
38. Palmer, D.S., O'Boyle, N.M., Glen, R.C., Mitchell, J.B.O. Random Forest Models To Predict Aqueous Solubility // Journal of Chemical Information and Modeling. – 2007. – Vol. 47 (1). – Pp. 150-158. DOI: 10.1021/ci060164k
39. Deng, T., Jia, G. Prediction of aqueous solubility of compounds based on neural network // Molecular Physics. – 2019. – Vol. 118(2). DOI: 10.1080/00268976.2019.1600754.
40. Lusci, A., Pollastri, G., Baldi, P. Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules // Journal of Chemical Information and Modeling. – 2013. – Vol. 53 (7). – Pp. 1563-1575. DOI: 10.1021/ci400187y
41. Boobier, S., Hose, D.R.J., Blacker, A.J. et al. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water // Nature Communciations. – 2020. – Vol. 11. – P. 5753. DOI: 10.1038/s41467-020-19594-z
42. Evteev, S.A., Ereshchenko, A.V., Ivanenkov, Y.A. SiteRadar: Utilizing Graph Machine Learning for Precise Mapping of Protein-Ligand-Binding Sites //

Journal of chemical information and modeling. – 2023. – Vol. 63(4). – Pp. 1124–1132. DOI: 10.1021/acs.jcim.2c01413

43. Evteev, S., Ereshchenko, A., Adjugim, D., Vyacheslavov, A., Pastukhova, A., Malyshev, A., Terentiev, V. and Ivanenkov, Y. Skittles: GNN-Assisted Pseudo-Ligands Generation and Its Application for Binding Sites Classification and Affinity Prediction // Proteins. – 2025. DOI: 10.1002/prot.26816
44. Ivanenkov, Y., Evteev, S., Malyshev, A., Terentiev, V., Bezrukov, D., Ereshchenko, A., Korzhenevskaya, A., Zagribelnyy, B., Shegai, P., Kaprin, A. AlphaFold for a medicinal chemist: tool or toy? // Russ. Chem. Rev. – 2024. – Vol. 93 (3). – P. RCR5107. DOI: 10.59761/RCR5107
45. Ereshchenko A.V. Applying machine learning for solubility prediction: comparing different representations of molecular data // Modelling and Data Analysis. – 2025. – Vol. 15, no. 1. – Pp. 35–50. DOI: 10.17759/mda.2025150103
46. van Montfort, R.L.M., Workman, P. Structure-based drug design: aiming for a perfect fit // Essays Biochem. – 2017. – Vol. 61. – Pp. 431-437. DOI: 10.1042/EBC20170052
47. Stepniewska-Dziubinska, M. M., Zielenkiewicz, P., Siedlecki, P. Improving detection of protein-ligand binding sites with 3D segmentation // Sci Rep. – 2020. – Vol. 10. – Pp. 5035-5044. DOI: 10.1038/s41598-020-61860-z
48. Gagliardi, L., Raffo, A., Fugacci, U., Biasotti, S., Rocchia, W., Huang, H., Amor, B. B., Fang, Y., Zhang, Y., Wang, X., Christoffer, C., Kihara, D., Axenopoulos, A., Mylonas, S., Daras, P. SHREC 2022: Protein–ligand binding site recognition // Computers & Graphics. – 2022. – Vol. 107 – Pp. 20-31. DOI: 10.1016/j.cag.2022.07.005
49. Zhang, W., Li, R., Shin, R., Wang, Y., Padmalayam, I., Zhai, L., Krishna, N. R. Identification of the binding site of an allosteric ligand using STD-NMR, docking, and CORCEMA-ST calculations // ChemMedChem. – 2013. – Vol. 8. – Pp. 1629-1633. DOI: 10.1002/cmdc.201300267
50. Zhang, Y., Zhang, D., Tian, H., Jiao, Y., Shi, Z., Ran, T., Liu, H., Lu, S., Xu, A., Qiao, X., Pan, J., Yin, L., Zhou, W., Lu, T., Chen, Y. Identification of Covalent

Binding Sites Targeting Cysteines Based on Computational Approaches // Molecular Pharmaceutics. – 2016. – Vol. 13. – Pp. 3106-3118. DOI: 10.1021/acs.molpharmaceut.6b00302

51. Li, H.; Kasam, V., Tautermann, C. S., Seeliger, D., Vaidehi, N. Computational method to identify druggable binding sites that target protein-protein interactions // Journal of Chemical Information and Modeling. – 2014. – Vol. 54. – Pp. 1391-1400. DOI: 10.1021/ci400750x
52. Desaphy, J., Bret, G., Rognan, D., Kellenberger, E. sc-PDB: a 3D-database of ligandable binding sites--10 years on // Nucleic Acids Research. – 2015. – Vol. 43. – Pp. 399-404. DOI: 10.1093/nar/gku928
53. Shi, Y., Huang, Z., Wang, W., Zhong, H., Feng, S., Sun, Y. Masked Label Prediction: Unified Massage Passing Model for Semi-Supervised Classification // Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. – 2021. – Pp. 1548−1554. DOI: 10.24963/ijcai.2021/214
54. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library // NeurIPS. – 2019. – Pp. 8024-8035.
55. Fey, M., Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric // ArXiv. – 2019.
56. Schütt, K., Kindermans, P. J., Sauceda Felix, H. E., Chmiela, S., Tkatchenko, A., & Müller, K. R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions // Advances in neural information processing systems. – 2017. – P. 30.
57. Kingma, D. P., Ba, J. Adam: A Method for Stochastic Optimization // CoRR. – 2015.
58. Ackermann, M. R., Blömer, J., Kuntze, D., Sohler, C. Analysis of Agglomerative Clustering // Algorithmica. – 2014. – Vol. 69. – Pp. 184-215. DOI: 10.1007/s00453-012-9717-4

59. Comaniciu, D., Meer, P. Mean Shift: A Robust Approach Toward Feature Space Analysis // IEEE Trans Pattern Anal Mach Intell. – 2002. – Vol. 24. – Pp. 603-619.
60. Kandel, J., Tayara, H., Chong, K.T. PUResNet: prediction of protein-ligand binding sites using deep residual neural network // Journal of Cheminformatics. – 2021. – Vol. 8. – Pp. 65-79. DOI: 10.1186/s13321-021-00547-7
61. Fpokcet: сайт. – URL: https://github.com/Discngine/fpocket
62. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. SciPy 1.0: fundamental algorithms for scientific computing in Python // Nature Methods. – 2020. – Vol. 17. – Pp. 261-272. DOI: 10.1038/s41592-019-0686-2
63. Pedregosa et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. – 2011. – Vol. 12(85). – Pp. 2825−2830.
64. Powell, M. J. D. An efficient method for finding the minimum of a function of several variables without calculating derivatives // The Computer Journal. – 1964. – Vol. 7, Issue 2. – Pp. 155–162. DOI: 10.1093/comjnl/7.2.155
65. Hartshorn M. J. et al. Diverse, high-quality test set for the validation of protein-ligand docking performance // J Med Chem. – 2007. – Vol. 50. – Pp. 726–741.
66. Ravindranath, P. A., Sanner, M. F. AutoSite: an automated approach for pseudo-ligands prediction-from ligand-binding sites identification to predicting key ligand atoms // Bioinformatics (Oxford, England). – 2016. – Vol. 32(20). – Pp. 3142–3149. DOI: 10.1093/bioinformatics/btw367
67. Harris, R., Olson, A. J., Goodsell, D. S. Automated prediction of ligand-binding sites in proteins // Proteins. – 2008. – Vol. 70(4). – Pp. 1506–1517. DOI: 10.1002/prot.21645

68. Landrum, G. A., Riniker, S. Combining IC50 or Ki Values from Different Sources Is a Source of Significant Noise // Journal of chemical information and modeling. – 2024. – Vol. 64(5). – Pp. 1560–1567. DOI: 10.1021/acs.jcim.4c00049
69. Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., & Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening // Journal of medicinal chemistry. –. 2004. –. Vol. 47(7). – Pp. 1750–1759. DOI: 10.1021/jm030644s
70. Eberhardt, J., Santos-Martins, D., Tillack, A. F., Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings // Journal of chemical information and modeling. – 2021. – Vol. 61(8). – Pp. 3891–3898. DOI: 10.1021/acs.jcim.1c00203
71. Novikov, F. N., Stroylov, V. S., Zeifman, A. A., Stroganov, O. V., Kulkov, V., Chilov, G. G. Lead Finder docking and virtual screening evaluation with Astex and DUD test sets // Journal of computer-aided molecular design. – 2012. – Vol. 26(6). – Pp. 725–735. DOI: 10.1007/s10822-012-9549-y
72. Fink, C., Sun, D., Wagner, K., Schneider, M., Bauer, H., Dolgos, H., Mäder, K., Peters, S.-A. Evaluating the Role of Solubility in Oral Absorption of Poorly Water-Soluble Drugs Using Physiologically-Based Pharmacokinetic Modeling // Clin. Pharmacol. Ther. – 2020. – Vol. 107. – Pp. 650-661. DOI: 10.1002/cpt.1672
73. Ameta, R.K., Soni, K., Bhattarai, A. Recent Advances in Improving the Bioavailability of Hydrophobic/Lipophilic Drugs and Their Delivery via Self Emulsifying Formulations // Colloids Interfaces. – 2023. – Vol. 7. – P. 16. DOI: 10.3390/colloids7010016
74. Huang D., Yang J., Zhang Q., Zhou X., Wang Y., Shang Z., Li J., Zhang B. Design, synthesis, and biological evaluation of 2,4-dimorpholinopyrimidine-5-carbonitrile derivatives as orally bioavailable PI3K inhibitors // Front Pharmacol. – 2024. – Vol. 15. – P. 1467028. DOI: 10.3389/fphar.2024.1467028
75. Evteev S., Ivanenkov Y., Semenov I., Malkov M., Mazaleva O., Bodunov A., Bezrukov D., Sidorenko D., Terentiev V., Malyshev A., Zagribelnyy B.,

Korzhenevskaya A., Aliper A., Zhavoronkov A. Quantum-assisted fragment-based automated structure generator (QFASG) for small molecule design: an in vitro study // Front Chem. – 2024. – Vol. 12. – P. 1382512. DOI: 10.3389/fchem.2024.1382512

76. Ryad N., Elmaaty A.A., Selim S., Almuhayawi M.S., Al Jaouni S.K., Abdel-Aziz M.S., Alqahtani A.S., Zaki I., Abdel Ghany L.M.A. Design and synthesis of novel 2-(2-(4-bromophenyl)quinolin-4-yl)-1,3,4-oxadiazole derivatives as anticancer and antimicrobial candidates: in vitro and in silico studies // RSC Adv. – 2024. – Vol. 14(46). – Pp. 34005-34026. DOI: 10.1039/d4ra06712f

77. Smyth, L. A., Collins, I. Measuring and interpreting the selectivity of protein kinase inhibitors // Journal of Chemical Biology. – 2009. – Vol. 2. – Pp. 131–151. DOI: 10.1007/s12154-009-0023-9

78. Klaeger, S., Heinzlmeir, S., Wilhelm M., et al. The target landscape of clinical kinase drugs // Science. – 2017. – Vol. 358. – P. eaan4368. DOI:10.1126/science.aan4368

79. Petrelli, A, Giordano, S. From single- to multi-target drugs in cancer therapy: when aspecificity becomes an advantage // Current Medicinal Chemistry. – 2008. – Vol. 15(5). – Pp. 422-32. DOI: 10.2174/092986708783503212

80. Pahikkala, T., Airola, A., Pietilä, S., et al. Toward more realistic drug–target interaction predictions // Briefings in Bioinformatics. – 2015. – Vol. 16, Issue 2. – Pp. 325–337. DOI: 10.1093/bib/bbu010

81. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A. CatBoost: unbiased boosting with categorical features // NeurIPS. – 2018. DOI: 10.48550/arXiv.1706.09516

82. Morgan, H.L. The generation of a unique machine description for chemical structures — a technique developed at chemical abstracts service // Journal of Chemical Documentation. – 1965. – Pp. 107–113. DOI: 10.1021/c160017a018

83. Capecchi, A., Probst, D., Reymond, J.L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome // Journal of Cheminformatics. – 2020. – Vol. 12. – P. 43. DOI: 10.1186/s13321-020-00445-4

84. Moriwaki, H., Tian, YS., Kawashita, N. et al. Mordred: a molecular descriptor calculator // Journal of Cheminformatics. – 2018. – Vol. 10, (4). DOI: 10.1186/s13321-018-0258-y

85. Andrew Blevins, Ian K Quigley, Brayden J Halverson, Nate Wilkinson, Rebecca S Levin, Agastya Pulapaka, Walter Reade, and Addison Howard. NeurIPS 2024 - Predict New Medicines with BELKA [Электронный ресурс] // Kaggle. – 2024. – Режим доступа: https://kaggle.com/competitions/leash-BELKA

86. Cacciari, I., Ranfagni, A. Hands-On Fundamentals of 1D Convolutional Neural Networks — A Tutorial for Beginner Users // Applied Sciences. – 2024. – Vol. 14(18). – P. 8500. DOI: 10.3390/app14188500

87. Elkins J.M., Fedele V., Szklarz M., Abdul Azeez K.R., Salah E., Mikolajczyk J., Romanov S., Sepetov N., Huang X.P., Roth B.L., Al Haj Zen A., Fourches D., Muratov E., Tropsha A., Morris J., Teicher B.A., Kunkel M., Polley E., Lackey K.E., Atkinson F.L., Overington J.P., Bamborough P., Müller S., Price D.J., Willson T.M., Drewry D.H., Knapp S., Zuercher W.J. Comprehensive characterization of the Published Kinase Inhibitor Set // Nat Biotechnol. – 2016. – Vol. 34(1). – Pp. 95-103. DOI: 10.1038/nbt.3374

88. Drewry D.H., Wells C.I., Andrews D.M., Angell R., Al-Ali H., Axtman A.D., Capuzzi S.J., Elkins J.M., Ettmayer P., Frederiksen M., Gileadi O., Gray N., Hooper A., Knapp S., Laufer S., Luecking U., Michaelides M., Müller S., Muratov E., Denny R.A., Saikatendu K.S., Treiber D.K., Zuercher W.J., Willson T.M.. Progress towards a public chemogenomic set for protein kinases and a call for contributions // PLoS One. – 2017. – Vol. 2;12(8). – P. e0181585. DOI: 10.1371/journal.pone.0181585

89. Wells C.I., Al-Ali H, Andrews D.M., Asquith C.R.M., Axtman A.D., Dikic I., Ebner D., Ettmayer P., Fischer C., Frederiksen M., Futrell R.E., Gray N.S., Hatch S.B., Knapp S., Lücking U., Michaelides M., Mills C.E., Müller S., Owen D., Picado A., Saikatendu K.S., Schröder M., Stolz A., Tellechea M., Turunen B.J., Vilar S., Wang J., Zuercher W.J., Willson T.M., Drewry D.H.. The Kinase Chemogenomic Set (KCGS): An Open Science Resource for Kinase

Vulnerability Identification // Int J Mol Sci. – 2021 – Vol. 8;22(2). – P. 566. DOI: 10.3390/ijms22020566

90. Gao Y., Davies S.P., Augustin M., Woodward A., Patel U.A., Kovelman R., Harvey K.J. A broad activity screen in support of a chemogenomic map for kinase signalling research and drug discovery // Biochem J. – 2013 – Vol. 15;451(2). – Pp. 313-28. DOI: 10.1042/BJ20121418
91. Seidel, T., Permann, C., Wieder, O., Kohlbacher, S., Langer, T. High-Quality Conformer Generation with CONFORGE: Algorithm and Performance Assessment // Journal of Chemical Information and Modeling. – 2023. – Vol. 63 (17). – Pp. 5549-5570. DOI: 10.1021/acs.jcim.3c00563
92. Hochreiter, S., Schmidhuber, J. Long Short-Term Memory // Neural Comput. – 1997. – Vol. 9 (8).–.Pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735
93. Wales, D.J, Doye, J. P. K. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms // Journal of Physical Chemistry A. – 1997. – Vol. 101. – P. 5111. DOI: 10.1021/jp970984n
94. Zhu, C., Byrd, R. H., Lu, P., Nocedal, J. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization // Association for Computing Machinery. – 1997. – Vol. 23 (4) – Pp. 550–560. DOI: 10.1145/279232.279236
95. Loftsson, T., Brewster, M.E. Pharmaceutical applications of cyclodextrins: basic science and product development // Journal of Pharmacy and Pharmacology. – 2010. – Vol. 62, Issue 11. – Pp. 1607–1621. DOI: 10.1111/j.2042-7158.2010.01030.x
96. Basavaraj, S., Guru V. Betageri, G.V. Can formulation and drug delivery reduce attrition during drug discovery and development—review of feasibility, benefits and challenges // Acta Pharmaceutica Sinica B. – 2014. – Vol. 4, Issue 1. – Pp. 3-17, ISSN 2211-3835. DOI: 10.1016/j.apsb.2013.12.003

97. Zaliani, A., Tang, J., Martin, J., Harmel, R., Wang, W. 1st EUOS/SLAS Joint Challenge: Compound Solubility [Электронный ресурс] // Kaggle. – 2022. – Режим доступа: https://kaggle.com/competitions/euos-slas