На правах рукописи

Ерещенко Алексей Владимирович

ПРИМЕНЕНИЕ ГРАФОВЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ АНАЛИЗА МОЛЕКУЛЯРНЫХ СТРУКТУР

Специальность 1.2.1

«Искусственный интеллект и машинное обучение»

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Москва 2025

Работа выполнена в Федеральном исследовательском центре «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН).

Научный руководитель: **Ревизников Дмитрий Леонидович**
доктор физико-математических наук, профессор,
профессор кафедры вычислительной математики и программирования Московского авиационного института

Официальные оппоненты: **Куравский Лев Семёнович**
доктор технических наук, профессор,
декан факультета информационных технологий Московского государственного психолого-педагогического университета

Иванков Дмитрий Николаевич
кандидат физико-математических наук,
старший преподаватель Центра молекулярной и клеточной биологии Сколковского института науки и технологий

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт системного программирования им. В.П. Иванникова Российской академии наук

Защита состоится «___»_____20 г. в____:____на заседании диссертационного совета 24.1.224.03 при Федеральном исследовательском центре «Информатика и управление» Российской академии наук по адресу: 119333, Россия, Москва, ул. Вавилова, д. 42.

С диссертацией можно ознакомиться в библиотеке Федерального исследовательского центра «Информатика и управление» Российской академии наук и на сайте http://www.frccsc.ru.

Автореферат разослан «____»_____ года.

Ученый секретарь
диссертационного совета
24.1.224.03,
кандидат технических наук Рейер Иван Александрович

Общая характеристика работы

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. 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. В рамках задачи по выявлению новых химических соединений, которые могли бы формировать взаимодействия с ранее неизвестной мишенью, часто возникает ситуация, когда трехмерная (3D) модель белка-мишени доступна, а информации о химических соединениях, которые формируют с ней взаимодействия, недостаточно или она полностью отсутствует. 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, применимость методов SBDD стабильно возрастает. В данной диссертации рассмотрены актуальные методы машинного обучения и интеллектуального анализа данных, 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 вероятность 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 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. Опираясь на последние академические исследования и принимая во внимание большой интерес в сфере разработки лекарственных средств, можно сформулировать множество актуальных задач, решение которых с применением алгоритмов машинного обучения приведет к созданию более эффективных компьютеризованных подходов для использования на различных этапах разработки лекарств.

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.

1. Разработана архитектура графовой нейронной сети для классификации пространства на поверхности трехмерной белковой структуры. Предложен комплексный алгоритм, который состоит из алгоритмического формирования трехмерного пространства вокруг белковой структуры, классификации данного пространства разработанными и обученными графовыми нейронными сетями, кластеризации классифицированного пространства для формирования

объемных объектов, а также ранжирования данных объектов с помощью оценок реализованных классификаторов. Показано преимущество алгоритма в задаче поиска и формирования объемных сайтов связывания по сравнению с иными методами, использованными для сравнения.

2. Реализована модифицированная архитектура графовой нейронной сети классификации пространства поверхности трехмерной белковой структуры для предсказания экспертно разработанных функциональных групп. Предложен новый подход для аннотации сайта связывания белка с применением разработанной архитектуры и графового представления трехмерной структуры белка, использующий экспертно сформированную и размеченную выборку данных для обучения моделей. Продемонстрировано преимущество разработанного подхода по сравнению с аналогами.

3. Изучено применение оценок, сгенерированных разработанными графовыми нейросетями аннотации пространства, для обучения других графовых нейронных сетей, решающих иные задачи. Представлены разработанные нейросетевые модели, классифицирующие сайты связывания, и разработанная модель, предсказывающая аффинность малых молекул. Экспериментально показано, что созданные и обученные подобным образом модели сравнимы или превосходят существующие аналоги.

4. Разработана архитектура графовой нейронной сети, способной принимать в качестве входных данных ансамбль графов трехмерных фармакофорных представлений рассматриваемой малой молекулы, а также ее физико-химические дескрипторы. Предложен новый подход по предсказанию наличия ингибирующей активности против заданной мишени с применением разработанной и обученной графовой нейронной сети и экспертного метода трехмерного моделирования возможных конформаций малых молекул. Показано преимущество реализованного алгоритма по сравнению с иными подходами машинного обучения.

Теоретическая значимость. Было проведено исследование графового описания трехмерной структуры белка и преимуществ применения графовых нейронных сетей в рассматриваемой предметной области. Изучена возможность применения оценок разработанных и обученных графовых нейронных 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 разработанной графовой нейронной сети, обученной на подобном пространственном описании.

Практическая значимость. Разработанные модели поиска сайтов связывания, аннотации пространства сайтов 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.

Разработанные модели применимы для различных задач медицинской химий и компьютеризованной разработки лекарственных средств, таких как поиск сайтов связывания в белках без известных ингибиторов, поиск аллостерических сайтов связывания, классификация сайтов связывания, предсказание аффинности химического соединения к заданному белковому окружению, оценка селективности лекарственных молекул.

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 апрель 2025 года, с публикацией тезисов.

Публикации. По тематике диссертации были опубликованы 4 статьи, в том числе 3 статьи индексируемые в Q1 Scopus/WOS ([1],[2],[3]), и статья в журнале из списка ВАК ([4]). 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 [1] соискателем была проведена следующая работа: анализ и обработка подготовленных экспертами для обучения данных, формирование обучающей и валидационной выборок, разработка 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 [2] соискателем была проведена следующая работа: анализ и обработка подготовленных экспертами для обучения данных для моделей аннотации сайтов связывания, формирование обучающей и валидационной выборок, разработка 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 [3] 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, разработка 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.

Краткое содержание диссертационной работы

Во введении обоснованы актуальность темы исследования, научная новизна и практическая значимость полученных результатов.

В первой главе рассматривается разработка и применение GNN для классификации пространства на поверхности трехмерных белковых структур с целью поиска специфических областей, а именно сайтов связывания. В данной работе сайт связывания рассматривается как карман, который связывает низкомолекулярное химическое соединение посредством

формирования межатомных взаимодействий (далее именуемое лиганд), что приводит к модуляции активности белка-мишени. Рассматриваются два варианта признакового описания белковой структуры – с использованием экспертно разработанных дескрипторов, и с использованием только позиционной информации атомов белка. Описывается реализованный в ходе проведенной исследовательской работы алгоритм поиска и формирования объемных сайтов связывания с применением разработанных и обученных моделей. Проводится сравнение алгоритма с рядом существующих моделей, выполняющих подобные задачи.

Был разработан новый метод на основе графовых нейронных сетей для определения сайтов связывания малых молекул. 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 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 решение задачи реализовано путем прогнозирования y путем разработки и обучения 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 примеры, независимо выбранные из обучающего набора.

Источником информации для D являются трехмерные структурные данные, которые были получены из открытой базы данных sc-PDB. Вокруг трехмерных структур были сформированы кубические решетки, созданные с помощью экспертно разработанного алгоритма на основании знаний предметной области. Объектом набора данных D является точка кубической решетки.

Признаковое описание x_k было сформировано в виде графа, созданного на основе окружающих рассматриваемый объект атомов белка, кроме водородов (далее именуемые тяжелые атомы). Были собраны 3D-координаты и характеристики аминокислот для всех тяжелых атомов белка в пределах 7 Å от каждой точки решетки. Координаты использовались для расчета расстояний между тяжелыми атомами белка и рассматриваемым объектом, однако не были включены явно как признаки. Каждый рассматриваемый объект представлялся в виде графа, где узлами являлись тяжелые атомы белка на заданном радиусе и сама точка решетки, а ребрами являлись рассчитанные расстояния. Использовались не только расстояния «атом – точка решетки», но и расстояния «атом – атом». Были собраны экспертно подобранные признаковые описания, закодированы one-hot encoding методом (преобразованы в бинарные векторы) и использованы в качестве узловых признаков. Т.к. точка решетки не может обладать химическими признаками, для ее описания были созданы особые признаки, отличающие ее от атомов белка.

y_k для данной задачи является бинарный класс «сайт связывания» и «не сайт связывания». Классификация объектов набора данных была проведена на основании предметной экспертизы.

Моделью $F: R^m \to R$ в данной задаче является разработанная нейросетевая архитектура на основе GNN. GNN обрабатывает молекулярный граф как комбинацию узлов и рёбер, где узлы представлены набором признаков, соответствующих данному атому, а рёбра представлены списками пар индексов узловых признаков. Применяемый графовый нейронный оператор использует механизм многоголового внимания, широко используемый в трансформерных нейронных сетях, и 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 — узел, от которого передаётся сообщение, e_{ij} — 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, $N(i)$ — индексы всех узлов, связанных с данным узлом.

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. Следует отметить, что W_1, W_2, W_3 и W_4 включают обучаемый параметр суммируемого смещения.

Таким образом, для решения поставленной задачи была разработана архитектура нейронной сети с применением описанного выше графового нейронного оператора, принимающая на вход описание пространства в виде графа, и возвращающая числовую оценку возможности рассматриваемого пространства принадлежать или не принадлежать сайту связывания. 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). Обе 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). Архитектура моделей представлена на рисунке 1.

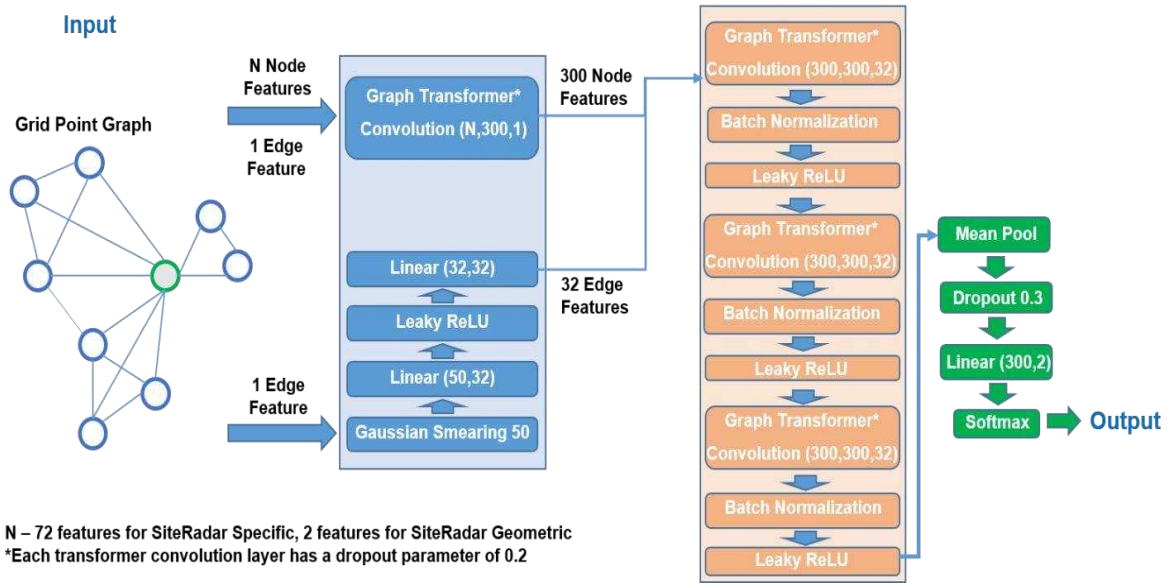

Рисунок 1. Архитектура разработанных нейронных сетей

Одной из важных особенностей рассматриваемой задачи является необходимость корректно учесть расстояния «атом – атом» и «атом – точка решетки» в рамках архитектуры нейросети. Истинные позиции атомов не полностью совпадают с позициями, зафиксированными в ходе сбора кристаллографической информации, и обладают подвижностью. Более того, модели необходимо обладать возможностью по-разному оценивать различные диапазоны расстояний, а не воспринимать их как скалярную величину. В связи с этим, по рассчитанным величинам расстояний было сформировано обучаемое векторное представление посредством применения размытия Гаусса к изначальной величине, результат которого затем проходил через последовательность линейных слоев с функцией активации усеченного линейного преобразования с «утечкой» (leaky ReLU). Данное векторное представление использовалось как признаковое описание ребер в основном многослойном GNN блоке. Так как GNN слой возвращает ответы на каждую вершину графа, в то время как в рамках данной задачи необходим один ответ на весь объект, после GNN блока применяется слой агрегирования графа, который затем проходит через слой выбивания (dropout) и поступает в линейный слой, формирующий ответ модели. Нейросеть была разработана экспериментально в ходе тестирования на обучающей и валидационной выборках. Также проводилась валидация предметными экспертами на специально отобранных примерах.

Собранные данные были случайно распределены в наборы для обучения (95%) и валидации (5%) для контроля переобучения. Разбиение было произведено по идентификаторам белковых комплексов. Итоговая выборка, использованная для обучения и валидациии, включала в себя 5 859 318 графов. Обучение проводилось батчами. Для обучения моделей использовалась функция потерь кросс-энтропии:

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

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 матрицы весов). Чтобы компенсировать разницу в количестве примеров в целевых классах (сайт связывания, не сайт связывания), были применены веса для балансировки классов.

С целью проверки воспроизводимости результатов и влияния случайного фактора при разделении данных на валидацию и обучение, каждая модель была обучена повторно с применением другого значения рандомизации.

С применением разработанных и обученных GNN моделей был реализован подход для поиска сайтов связывания SiteRadar. Он состоит из следующих последовательных шагов: 1) подготовка решетки экспертно разработанным алгоритмом; 2) классификация точек решетки; 3) кластеризация; 4) выполнение оценки. Разработанные GNN, обученные на данных, сформированных тем же способом, что и в ходе работы алгоритма на шаге 1, применяются для классификации всех точек сформированной решетки (шаг 2), возвращая как оценку в виде действительного числа, так и предсказанный класс. Результатом работы SiteRadar является объемный сайт связывания. Алгоритм SiteRadar представлен в графическом виде на рисунке 2.

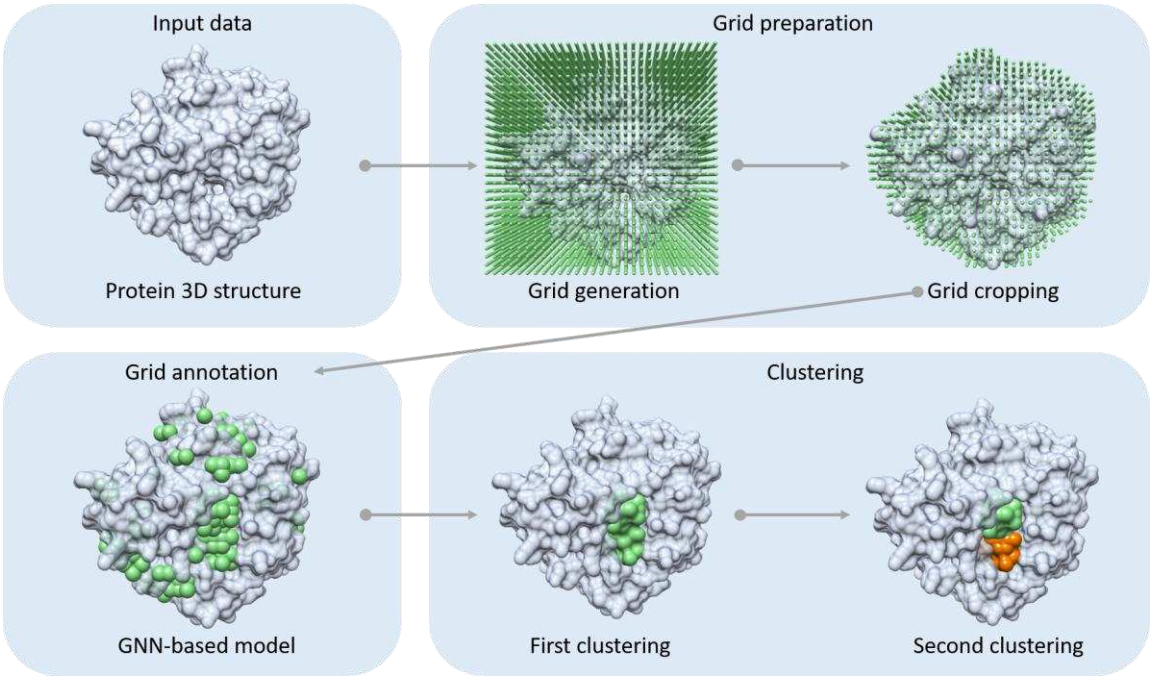

Рисунок 2. Архитектура алгоритма SiteRadar

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. Чтобы сравнить SiteRadar с этими алгоритмами был подготовлен экспертно сформированный набор данных кристаллических

структур из 232 белков и 244 связанных лигандов. Данные были подобраны как наиболее непохожие на обучающие данные и показали низкое сходство последовательности аминокислот с обучающим набором (максимальное сходство последовательности аминокислот с обучающим набором не превышало 52%).

В этом исследовании применялись два класса метрик. Первая группа предназначена для оценки способности изучаемых методов правильно определять местоположение сайта связывания: расстояние «центр-центр» (DCC), оценка top N, top N+2, а также среднее количество предсказанных сайтов на белок. Вторая группа показателей использовалась для оценки способности моделей воспроизводить трехмерную форму лигандов в пределах правильно обнаруженных сайтов связывания: покрытие лиганда (LC), покрытие кармана (PC) и перекрытие дискретного объема (DVO). Кроме того, были рассчитаны объемные параметры размеченных моделями сайтов связывания, они были сопоставлены с пространственными характеристиками известных сайтов связывания лекарственных средств. Данные метрики были экспертно подобраны и применяются для оценки подобных моделей в современных научных исследованиях. Основные результаты представлены в таблице 1.

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

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 сайтов 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 разработанный алгоритм формирует их более избирательно.

Во второй главе рассматривается разработка и применение GNN для разметки сайта связывания трехмерной белковой структуры в соответствии с экспертно разработанными

функциональными группами, на основании белкового окружения данного пространства. Описывается возможное применение данной разметки, в том числе и для обучения иных ML моделей. Приводится сравнение разработанных и обученных GNN с существующими моделями разметки пространства. Рассматриваются модели, обученные с применением данных, сгенерированных реализованными GNN, в рамках задач классификации сайтов связывания и предсказания аффинности малых молекул.

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, набором данных D являются лиганд-белковые комплексы, полученные из открытой базы трехмерных белковых структур RCSB PDB, а каждым объектом данных является точка в пространстве сайта связывания белка. Однако, в отличие от подхода, использованного в главе 1, рассматриваемыми объектами были не точки алгоритмически созданных решеток, а атомы молекул, которые могли находиться в произвольных позициях, расширяя сферу применимости разработанных GNN.

Признаковое описание точки было сформировано по аналогии с тем, как это выполнено для задачи поиска сайта связывания, но с радиусом сбора данных 5 Å, были использованы те же экспертно подобранные описательные признаки, что и для АК-специфичной модели из главы 1. Каждая точка была классифицирована 13 экспертно разработанными классами, которые соответствуют тем или иным функциональным свойствам, которыми может обладать данное пространство.

Для решения поставленной 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. Отрицательные примеры состояли из всех функциональных групп, исключая положительные. Таким образом, целевым ответом для каждой из 13 обученных моделей был бинарный ответ по принадлежности целевому классу.

Применяемой моделью является GNN, использующая архитектуру, разработанную на основе архитектуры GNN, представленной в главе 1. Основные изменения включали в себя добавление дропаута на уровне графа, увеличения количества слоев и нейронов, добавления перед слоем агрегирования графа GNN слоя, принимающего на вход конкатенированные ответы предшествующего многослойного GNN блока и первого GNN слоя. Нейросеть была разработана экспериментально в ходе тестирования на обучающей и валидационной выборках, а также по результатам экспертной валидации на специально отобранных примерах. Был применен схожий подход к обучению.

Дополнительно была поставлена задача предсказания вышеупомянутых функциональных классов, но только в тех случаях, где атом действительно формирует ключевые взаимодействия с белком. Для этого были обучены дополнительные 13 моделей (далее именуемые фармакофорными), которые использовали ту же выборку данных, но с более строго отобранными положительными примерами классов.

Для обучения использовался экспертно обработанный набор данных, сформированный из 8150 комплексов лиганд-белок, полученных из открытой базы RCSB PDB. Обработка данных фармакофорных моделей была подобна, но включала дополнительные шаги, направленные на сбор дескрипторов взаимодействий для каждого атома.

10% данных были случайным образом отделены в набор валидации на основе идентификаторов белков. Для каждой бинарной модели все атомы были либо помечены как 1 (относящиеся к классу), либо 0 (являющиеся либо другим классом, либо неопределенными). Обучение моделей проводилось батчами. Для обучения моделей использовалась функция потерь кросс-энтропии (5).

С применением разработанных и обученных GNN был реализован алгоритм разметки пространства в соответствии с его функциональными свойствами в заданном белковом окружении (также именуемый алгоритм генерации псевдолигандов). Алгоритм состоит из двух последовательных шагов: (1) генерация объемного пространства для разметки и (2) предсказание функциональных классов точек пространства (Рисунок 3). В рамках данной работы для шага (1) использовался алгоритм, описанный в главе 1. На шаге (2) используются разработанные GNN для разметки сайтов связывания, позволяющие получить оценку степени принадлежности каждой точки к используемым 13 классам.

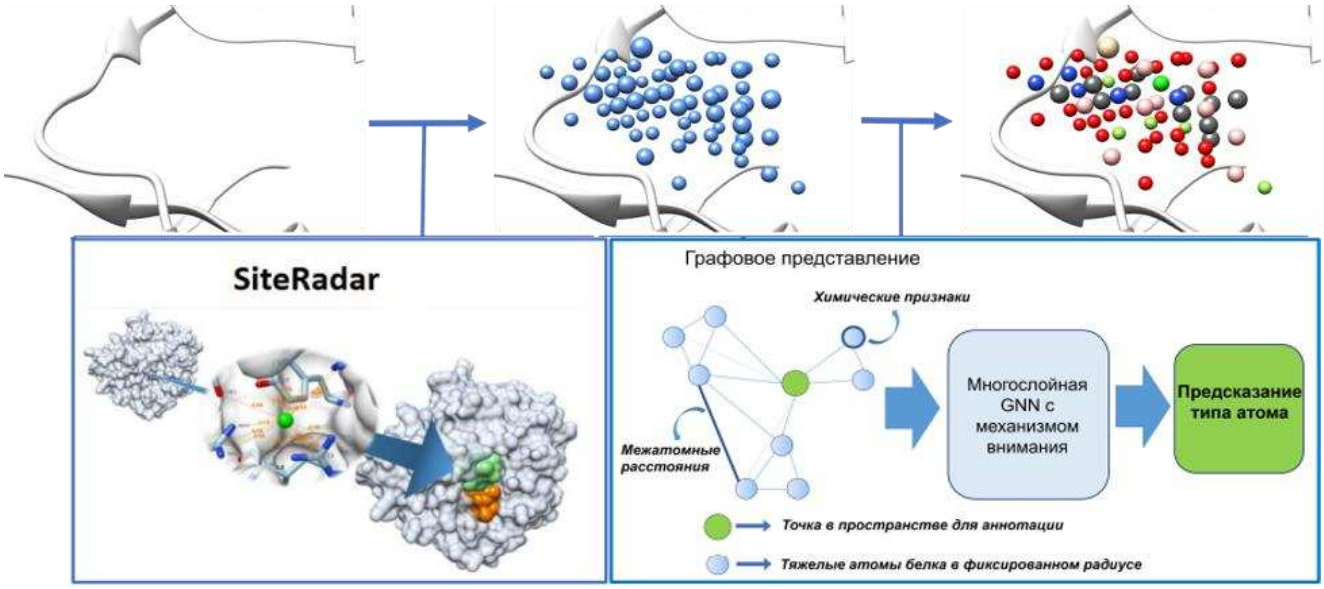

Рисунок 3. Схема генерации псевдолиганда с помощью разработанных GNN

Для обеспечения более высокой точности классификации точек при совместной работе обученных бинарных моделей были подобраны балансирующие коэффициенты.

Разработанные модели были протестированы в нескольких экспериментах. Основные результаты представлены на рисунке 4, где продемонстрирована точность классификации моделей при работе совместно по метрике top-n (вероятность того, что при ранжировании полученных оценок каждой модели от большего к меньшему, оценка истинного класса окажется на n позиции или выше) на валидационной выборке. Точность большинства базовых моделей выше 60% по top-n при n равному 4, однако определенные классы были более трудными для точного прогнозирования (Рисунок 4 А). Данные результаты могут быть объяснены природой самих классов и неоднозначности их возможного определения. Также стоит отметить, что наиболее слабо предсказываемые классы, а именно Csp, Sul, SO2 и Hal, являются мало представленными в обучающей выборке. Фармакофорные модели показали более точные результаты по метрике top-n, как показано на Рисунке 4 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.

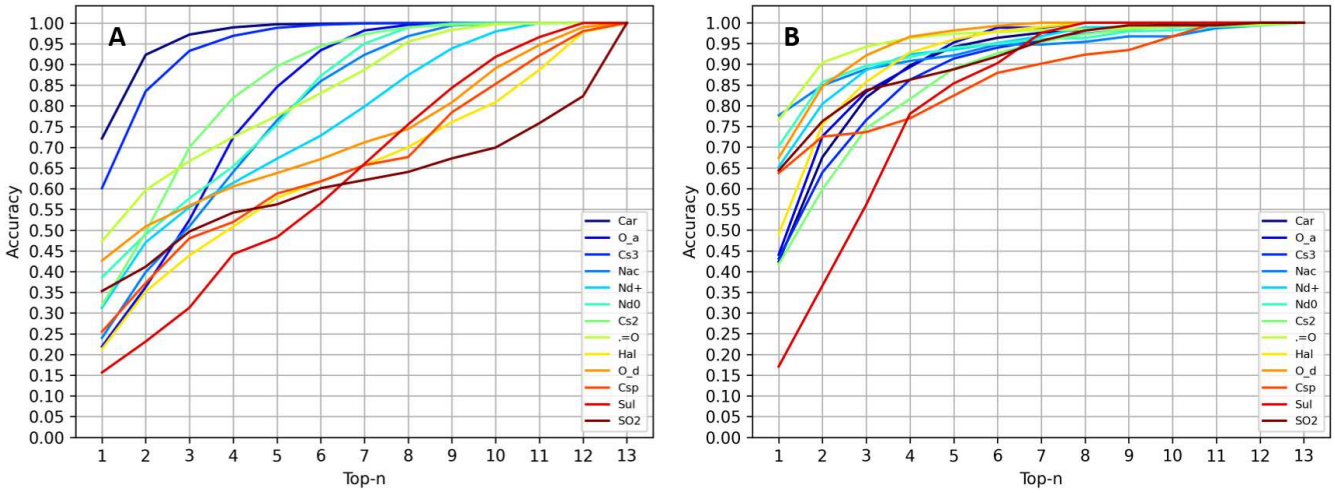

Рисунок 4. Top-n точность базовых (А) и фармакофорных (В) моделей в объединенном режиме

Также было проведено сравнение метода с существующими решениями. Для этого был использован независимый набор данных Astex Diverse Set, который не использовался для формирования обучающих и валидационных выборок. Были выбраны две модели для сравнения, основанные на расчётах энергий, AutoSite и AutoLigand. Сравнение было проведено следующим образом: пространство было аннотировано с помощью алгоритма генерации псевдолигандов, затем предсказанные классы были сгруппированы в три более широкие категории: доноры водородных связей (HD), акцепторы водородных связей (HA) и

гидрофобные атомы (HC). Группировка была проведена в связи с тем, что AutoSite и AutoLigand были разработаны для прогнозирования именно этих трех более широких классов. Разработанный алгоритм превзошел модель AutoSite по точности предсказания HA и HC, и оба метода превзошли AutoLigand при прогнозировании всех рассматриваемых классов (Таблица 2). Доля корректно предсказанных атомов была рассчитана в соответствии с методологией, описанной в статье об AutoSite.

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. Сравнение точности предсказания типов атомов

Также была протестирована применимость оценок, получаемых разработанными GNN, для обучения иных моделей машинного обучения, решающих задачи классификации сайтов связывания и оценки аффинности малых молекул. Разработаны GNN, использующие сгенерированные представленным алгоритмом псевдолиганды как признаковое описание для классификации сайтов связывания (Рисунок 5).

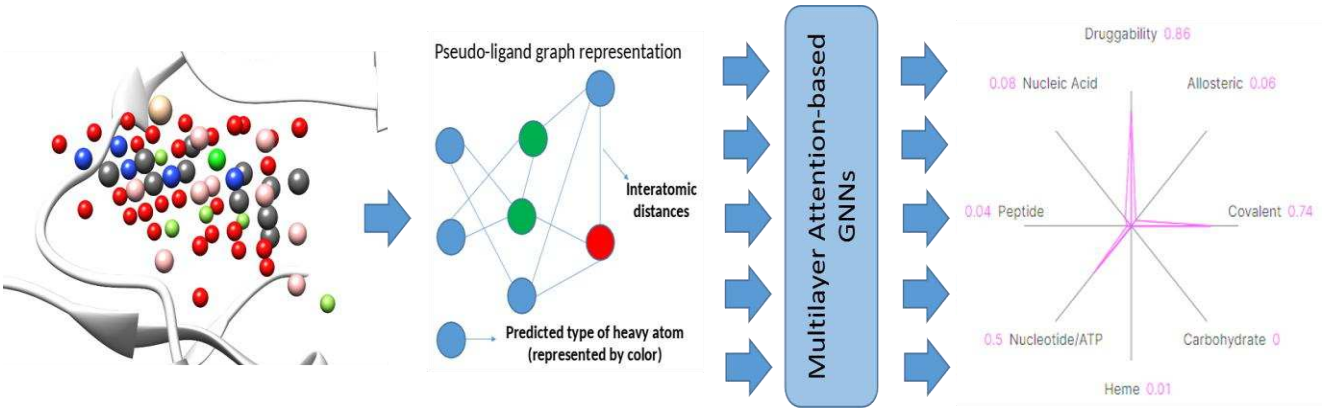

Рисунок 5. Упрощенная схема использования сгенерированного алгоритмом псевдолиганда для классификации мест связывания лиганд-белок

Разработанные модели были обучены для предсказания пяти химических свойств: нуклеотиды/АТФ, углеводы, гем, пептиды и нуклеиновые кислоты, а также двух дополнительных свойств, таких как способность к аллостерическому модулированию и способность принимать ковалентные лиганды. Было проведено сравнение модели классификации с существующими аналогами, в частности с моделью BionoiNet по классам Нуклеотиды/АТФ и Гем с применением метрики площадь под кривой рабочей характеристики

приемника (ROC AUC). При использовании открытой тестовой выборки данных, созданной авторами BionoiNet, разработанный алгоритм показал значения ROC AUC 0,966 по классу Нуклеотиды/АТФ и 0,957 по классу Гем, превосходя результаты BionoiNet (0,96 по классу Нуклеотиды/АТФ и 0,935 по классу Гем).

В рамках проведенного научного исследования, продемонстрировано применение разработанных GNN, размечающих пространство сайта связывания, для создания и обучения GNN, способной предсказывать аффинность малой молекулы по отношению к заданному белковому окружению. В реализованном 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, дополненные экспертно подобранными дескрипторами, характеризующими физико-химические взаимодействия. Визуальное представление алгоритма показано на рисунке 6. В рамках данного метода была разработана соответствующая архитектура нейронной сети.

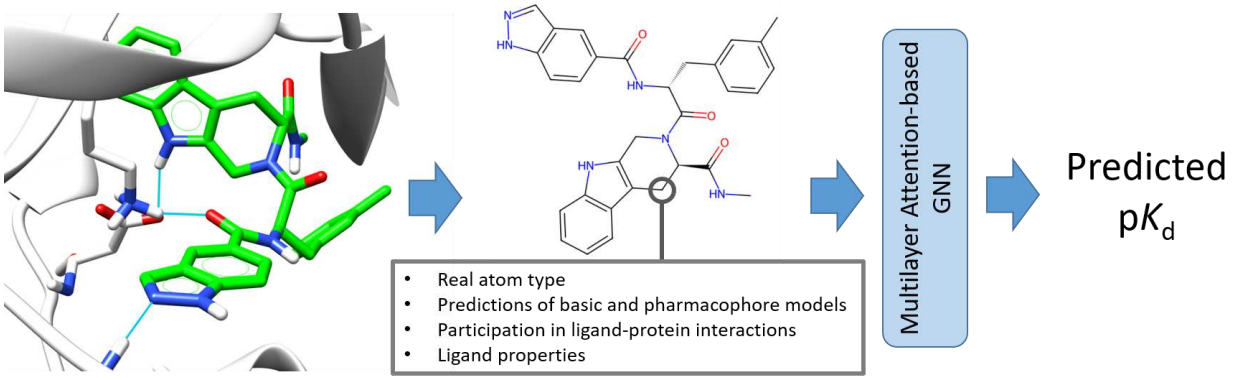

Рисунок 6. Упрощенная схема использования представления, сгенерированного GNN моделями разметки, для предсказания аффинности лиганд-белок

Выборка данных, использованная для обучения и тестирования модели, была экспертно сформирована на основании открытой базы PDBBind v.2016. 5% самых непохожих комплексов (404) были случайно отделены для независимого тестирования. На данной выборке было проведено сравнение 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. Оценка моделей проводилась по двум метрикам - коэффициент корреляции Пирсона предсказаний с истинными значениями (Рисунок 7) и их средняя абсолютная ошибка (MAE).

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. ML методы превзошли AutoDock Vina как в коэффициенте корреляции Пирсона, так и в MAE. 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, в связи с чем не было возможности оценить степень схожести комплексов использованного тестового набора с обучающим набором IGN.

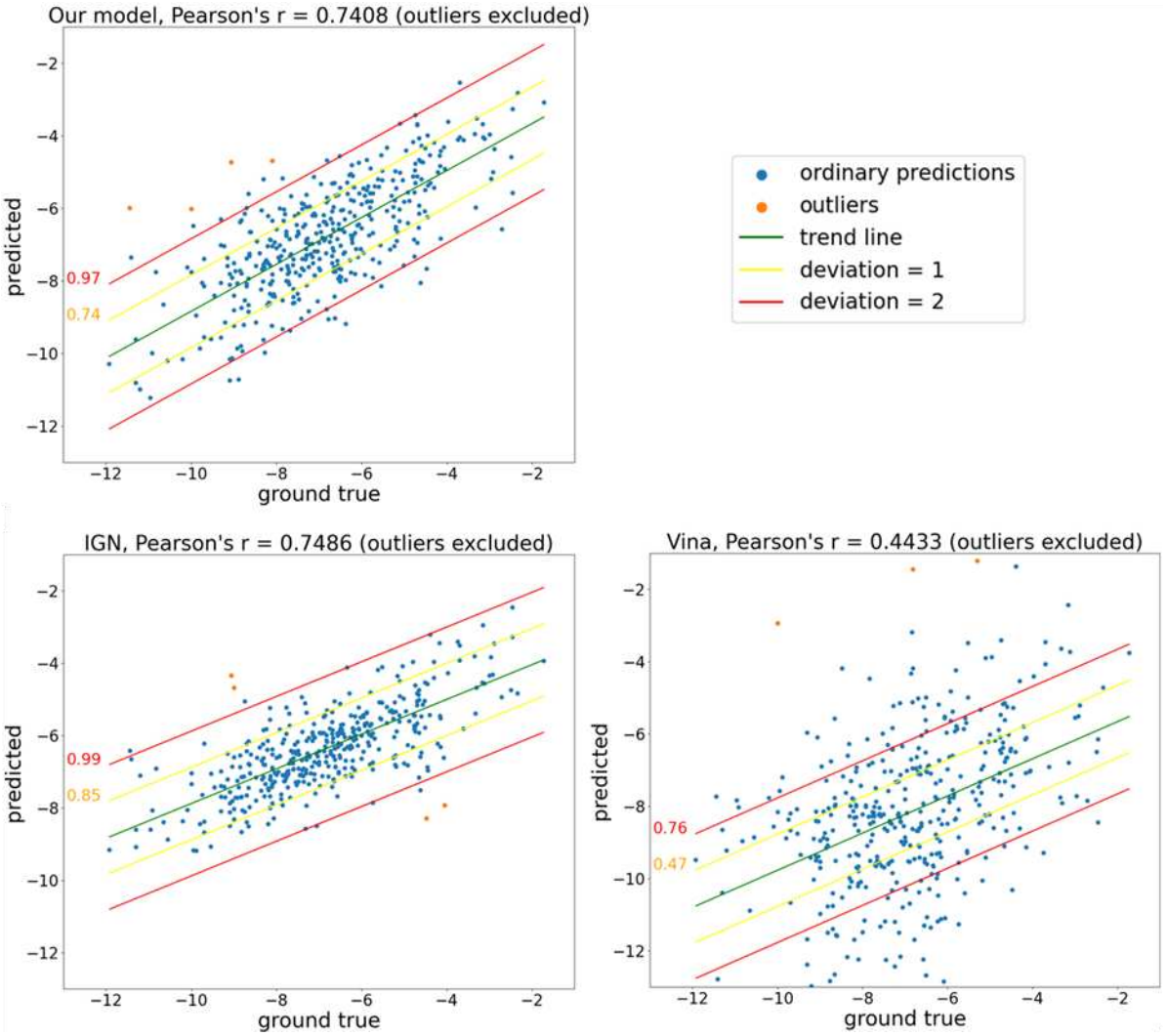

Рисунок 7. Сравнение различных методов для предсказания аффинности лиганда к белку

В третьей главе рассматривается применение GNN для работы с молекулярной структурой без использования информации о ее белковом окружении, для анализа ее свойств, на примере задач предсказания потенциальных белковых мишеней молекулы и анализа растворимости. Изучаются подходы для работы с различными признаковыми описаниями, которые можно извлечь из упрощенной молекулярной входной строковой записи (SMILES). Проводится сравнение разработанных GNN и графового представления данных с другими ML архитектурами и представлениями молекулярных структур.

Постановка задачи схожа с представленными в главах 1 и 2, с основным различием в входных данных и предсказываемых величинах. В рамках задачи предсказания наличия

ингибирующей активности, набор данных содержит профили активности малых молекул, представленных в SMILES, для 75 белковых 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:

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

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

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

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 разработаны 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 модель, рассматривающая SMILES как символьную последовательность. С целью выявления наличия либо отсутствия качественного прироста точности предсказаний при использовании разработанной GNN и графового представления по сравнению с иными подходами, были обучены три модели (далее именуемые модели сравнения): модель на основе градиентного бустинга с использованием библиотеки CatBoost, одномерная сверточная нейросеть (1D CNN), и глубокая нейросеть (DNN). CatBoost модель была обучена на сочетании физико-химических дескрипторов и представления данных в виде отпечатков Моргана, являющихся распространенным способом кодирования структурных характеристик молекулы в вектор, 1D CNN модель была обучена на токенизированной последовательности SMILES молекул, а DNN модель была обучена на крупном наборе дескрипторов, собранных с помощью библиотеки Mordred.

Обучение моделей было проведено на молекулах, для которых была известна активность по всем предсказываемым мишеням. Набор данных содержал 1176 записей и был подготовлен экспертами в данной области, с применением открытых баз данных. Соединения были

помечены 1 или 0 в зависимости от того, демонстрировали ли они ингибирование не менее 50% при концентрации 1 мкмоль/л.

Для формирования независимого теста, было отделено 15% данных (177 молекул). Оставшиеся 999 структур были случайно разделены на 5 обучающих и валидационных сплитов.

Для подготовки входных данных для обучения GNN модели, для каждой малой молекулы генерировалось до 20 трехмерных конформаций. На основании данных конформаций было экспертно построено их представление в виде функциональных зон (фармакофоров), на основании которых строились полносвязные графы, где вершинами являются фармакофорные точки, признаком которых является их фармакофорный тип, закодированный 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 11 физико-химических дескрипторов, подобранных экспертами. К данным дескрипторам был применен метод стандартного масштабирования.

В рамках разработанной архитектуры нейронной сети, 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 8.

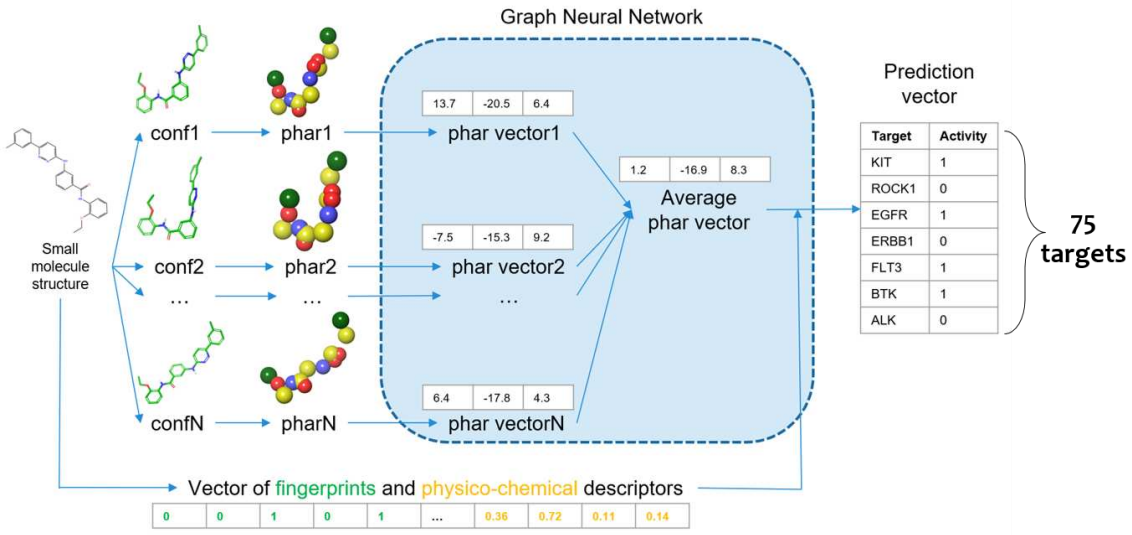

Рисунок 8. Схема обработки молекулы и прогнозирования активности

Обучение всех моделей проводилось батчами. После обучения для всех моделей были подобраны пороги бинаризации на обучающей выборке. Сравнение всех обученных методов было проведено на тестовой выборке и представлено на рисунке 9. 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. Модель GNN показала статистически значимую разницу по сравнению с CatBoost (p-значение = 0,03) и другими базовыми методами.

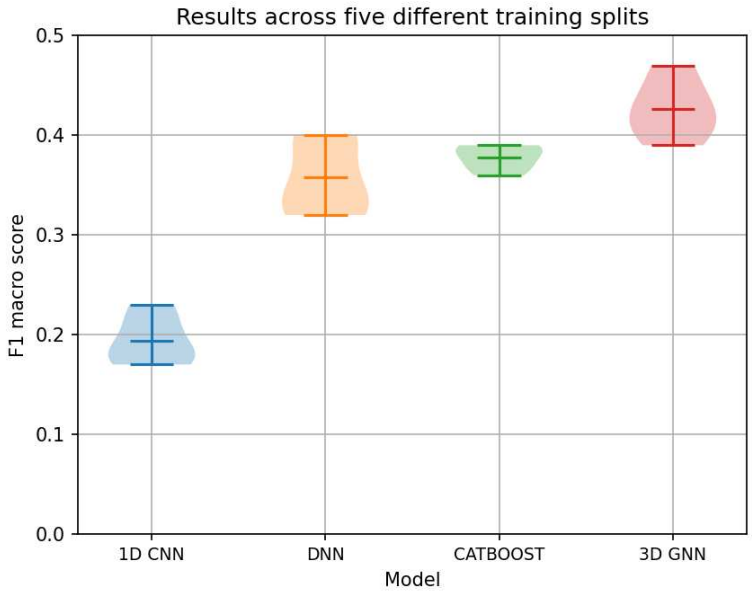

Рисунок 9. Сравнение подготовленных моделей на независимом наборе данных

В рамках решения задачи предсказания растворимости, 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 графовое представление.

В рамках данной задачи не использовалась генерация трехмерных конформаций – молекулярный граф был построен на основании доступной информации об атомах и их связях. Данное изменение было введено по следующей причине: хотя GNN на основании ансамбля 3D фармакофорного представления молекулы показала себя как наиболее точный подход из рассмотренных на задаче предсказания мишеней, подобное 3D моделирование является вычислительно затратным. В связи с этим, представляло интерес проверить, сохранится ли преимущество GNN подхода при работе с более простым графовым представлением. Для работы с подобным представлением данных была разработана соответствующая архитектура графовой нейронной сети.

Используя публично доступный набор данных, содержащий более 70 000 молекул с известной растворимостью, размеченный по трем классам растворимости (высокая, низкая, средняя), были обучены и сравнены ИИ-модели на основе трех различных архитектур и подходов к представлению данных: модель градиентного бустинга в реализации CatBoost, 1D CNN и графовая нейронная сеть на основе механизма внимания. Данные были разделены на тестовую выборку, а также 5 сплитов обучение\валидация. Основное сравнение моделей было произведено по коэффициенту Каппа Коэна.

По результатам проведенных на пяти разных разделениях выборки на обучение и валидацию экспериментов, модель CatBoost, а не GNN, проявила себя как наиболее точную. 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.

В заключении приведены основные результаты работы:

1. Разработан алгоритм поиска сайтов связывания белковых молекул с применением созданных 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 разработанных 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 разработанные 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.

Основные публикации по теме диссертации

1. 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). – P. 1124–1132. DOI: 10.1021/acs.jcim.2c01413
2. 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. – Vol. 93 (7). – P. 1269–1280. DOI: 10.1002/prot.26816
3. 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
4. Ereshchenko A.V. Applying machine learning for solubility prediction: comparing different representations of molecular data // Modelling and Data Analysis. – 2025. – Vol. 15 (1). – P. 35–50. DOI: 10.17759/mda.202515010