Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Уральский федеральный университет имени первого Президента России Б. Н. Ельцина»
Химико-технологический институт
Кафедра технологии органического синтеза

ДОПУСТИТЬ К ЗАЩИТЕ В ГЭК
Зав.кафедрой_____В.А.Бакулев
«___» _____2025 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

РАЗРАБОТКА И ИССЛЕДОВАНИЕ ГРАФОВОЙ НЕЙРОСЕТЕВОЙ МОДЕЛИ ДЛЯ ПРЕДСКАЗАНИЯ ТОКСИЧНОСТИ АЗОЛОПИРИМИДИНОВЫХ СОЕДИНЕНИЙ

Руководитель, доц., канд. мед. наук В. В. Мелехин
Нормоконтролер, доц., канд. хим. наук М. А. Безматерных
Студент группы ХВМ-330024 Е. В. Самойлова

Екатеринбург
2025 год

Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет имени первого Президента России Б.Н.Ельцина»

Институт___*Химико-технологический*_____
Кафедра/Департамент____*Технологии органического синтеза*_____
Направление (специальность) _*19.04.01 Биотехнология*_____
Образовательная программа __Молекулярная биотехнология и биоинженерия_____

УТВЕРЖДАЮ
Зав.кафедрой_____В.А.Бакулев
«___» _____2025 г.

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

студента _*Самойловой Елены Вячеславовны*_____ группы _*ХВМ-330024*_
(фамилия, имя, отчество)

1 Тема ВКР ___Разработка и исследование графовой нейросетевой модели для предсказания токсичности азолопиримидиновых соединений_____
Утверждена распоряжением по институту от «24» февраля 2025 г. № 33-09-05/26

2 Руководитель _*Мелехин В.В., доц., канд. мед. наук*_____
(Ф.И.О., должность, ученое звание, ученая степень)

3 Исходные данные к работе_*данные предоставлены Лабораторией первичного биоскрининга, клеточных и генных технологий НОиИЦ ХФТ ХТИ УрФУ*_____

4 Перечень демонстрационных материалов _*презентация в PowerPoint*_____

5 Календарный план

Наименование этапов выполнения работы	Срок выполнения этапов работы	Отметка о выполнении
Аналитический обзор литературы	10.08.2025	
Экспериментальная часть	15.09.2025	
Анализ полученных данных	02.10.2025	
Оформление пояснительной записки	15.10.2025	
Магистерская дисертация в целом	07.11.2025	

Руководитель _____ Мелехин В.В.
(подпись) Ф.И.О.

Задание принял к исполнению _____
(подпись)

8 Выпускная квалификационная работа закончена «07» ноября 2025 г. Считаю возможным допустить *Самойлову Елену Вячеславовну* к защите его выпускной квалификационной работы в экзаменационной комиссии.

Руководитель_____В.В. Мелехин

9 Допустить *Самойлову Елену Вячеславовну* к защите выпускной квалификационной работы в экзаменационной комиссии (протокол заседания кафедры №15 от «_20_» ноября_ 2025 г.)

Зав. кафедрой _____В.А.Бакулев

ОГЛАВЛЕНИЕ

РЕФЕРАТ

Выпускная квалификационная работа на соискание степени магистра – 69 стр., 13 рис., 19 табл., 49 источников, 6 приложений.

Работа состоит из введения, обзора литературы, методологии, экспериментальной части, результатов, списка использованных источников и приложения.

Ключевые слова: графовые нейронные сети, токсичность, аффинность, азолопиримидины, QSAR-моделирование, хемоинформатика.

В работе предложен подход к прогнозированию токсичности малых органических молекул на основе их аффинности к токсикологически значимым белковым мишеням. Построена графовая нейронная модель, использующая данные аффиности к белкам, вовлечённых в механизмы метаболизма, клеточного стресса и апоптоза, как прокси-показатели токсического потенциала.

В качестве примера применения методики рассмотрен класс азолопиримидинов: перспективных фармакофоров с разнообразной биологической активностью и обладающих структурой, позволяющей проводить рациональный дизайн и оптимизацию свойств.

Модель продемонстрировала способность воспроизводить молекулярные закономерности токсичности. Разработанный подход может служить основой для генеративного дизайна молекул по целевым признакам.

АННОТАЦИЯ

Представлена графовая нейронная модель прогнозирования токсичности малых органических молекул, основанная на идее связи между аффинностью соединений к белковым мишеням и их токсическим потенциалом. Для обучения использованы данные BindingDB по значениям IC_{50} в отношении белков, участвующих в метаболизме ксенобиотиков, апоптозе и клеточном стрессе.

Модель показала высокую воспроизводимость и способность выявлять структурные детерминанты, ассоциированные с токсичностью. В качестве демонстрационного примера рассмотрен класс азолопиримидинов.

Разработанный подход может быть использован в задачах генеративного моделирования соединений с заданными свойствами и оптимизации структуры лекарственных кандидатов.

ABSTRACT

This work presents a graph neural network model for predicting the toxicity of small organic molecules, grounded in the hypothesis that toxicological outcomes can be approximated through compound affinity toward toxicity-relevant protein targets. The model is trained on BindingDB IC_{50} measurements for proteins implicated in xenobiotic metabolism, apoptotic signaling, and cellular stress responses.

The resulting framework exhibits high reproducibility and effectively captures structural determinants associated with increased toxic potential. Azolopyrimidines are examined as a representative chemical class to demonstrate the model's applicability to closely related scaffolds.

The proposed methodology provides a foundation for downstream tasks such as generative design of molecules with controlled toxicological profiles and structure-based optimization of drug candidates.

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

Граф	Математическая структура, описывающая множество вершин (атомов) и соединяющих их рёбер (химических связей), применяемая для представления молекул.
Клиффы (activity cliffs)	Пары молекул с высокой структурной схожестью, но существенно различающейся биологической активностью.
Скаффолд (scaffold)	Структурный каркас молекулы, формирующий её основную химическую основу без учёта периферических заместителей.
Murcko scaffold	Тип представления химического каркаса, включающий кольцевые системы и соединяющие их звенья без заместителей.
Transfer learning (перенос обучения)	Метод, при котором предварительно обученная модель дообучается на новой задаче с ограниченным набором данных.
Multitask-обучение	Архитектурный подход, при котором одна модель решает несколько взаимосвязанных задач (например, прогноз различных параметров ADMET).
Embedding (встраивание)	Векторное представление объекта (молекулы, атома и т. д.), формируемое на скрытых слоях нейронной сети.
Коэффициент Танимото (Tanimoto coefficient)	Мера структурного сходства между двумя молекулами, вычисляемая по их бинарным отпечаткам.
Morgan-отпечатки (Extended Connectivity Fingerprints, ECFP)	Векторное кодирование молекулы на основе подструктур атомных окружений заданного радиуса.
Random split	Равномерное случайное распределение соединений между обучающей, валидационной и тестовой выборками.
Scaffold split	Разбиение данных по уникальным химическим каркасам, исключающее пересечение структур между выборками.
Neighbor split	Разбиение данных с учётом молекулярного сходства, при котором структурно близкие соединения направляются в разные выборки для оценки обобщающей способности модели.

УСЛОВНЫЕ ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

БД	база данных
ЛС	лекарственное средство
QSAR	Quantitative Structure–Activity Relationship, количественная зависимость структура–активность
ADMET	Absorption, Distribution, Metabolism, Excretion, Toxicity
HTS	High-Throughput Screening, высокопроизводительный скрининг
SMILES	Simplified Molecular Input Line Entry System, упрощённая линейная запись структуры молекулы
SMARTS	язык шаблонов для поиска химических подструктур
GNN	Graph Neural Network, графовая нейронная сеть
MPNN, D-MPNN, GCN, GAT	типы архитектур графовых нейросетей
RDKit	библиотека Python для хемоинформатики и расчёта дескрипторов
ML	Machine Learning, машинное обучение
MAE	Mean Absolute Error, средняя абсолютная ошибка
MSE	Mean Squared Error, среднеквадратичная ошибка
RMSE	Root Mean Square Error, корень из среднеквадратичной ошибки
R^2	коэффициент детерминации
AUC	Area Under the Curve, площадь под ROC-кривой
pIC_{50}, IC_{50}, K_i, K_d	биохимические константы взаимодействия лиганда с мишенью
ECFP4	Extended Connectivity Fingerprint (radius 4), расширенные молекулярные отпечатки
cross-validation	кросс-валидация, метод оценки качества модели
Random Forest	ансамблевый метод на множестве деревьев решений
XGBoost, LightGBM	реализации алгоритма градиентного бустинга
RMSE, MAE, R^2	основные метрики регрессии
GNN Chemprop	реализованная архитектура графовой нейросети (D-MPNN)
MLP	Multi-Layer Perceptron, многослойный персептрон

ВВЕДЕНИЕ

Современные исследования в области молекулярного дизайна направлены на ускорение и удешевление этапов поиска и оптимизации лекарственных соединений. Одним из ключевых критериев при оценке перспективности молекул является токсичность, определяющая как их потенциальную безопасность, так и возможность последующего подбора селективных соединений, активных в отношении патологических клеток при минимальном воздействии на нормальные ткани. Оценка токсичности соединений является необходимым этапом в доклинических исследованиях и во многом определяет успех дальнейших стадий разработки лекарственных средств.

По данным отраслевых аналитических отчётов, до 59 % кандидатов на этапе доклинических и ранних клинических исследований исключаются из разработки именно по причине проявления токсических эффектов. Экспериментальные методы токсикологической оценки требуют значительных материальных и временных ресурсов, что ограничивает объём доступных данных и существенно замедляет поиск безопасных соединений. В этой связи особое значение приобретают вычислительные методы (*in silico*), позволяющие прогнозировать токсичность и активность молекул по их структуре ещё до проведения биологических испытаний.

Токсичность не является самостоятельным свойством вещества, а представляет собой функциональное следствие его взаимодействия с биомишенями. Избыточное ингибирование или аномальная активация белков, участвующих в метаболизме, клеточном стрессе или сигнальной регуляции, способно запускать каскады реакций, приводящие к апоптозу, воспалению и другим нежелательным эффектам. Поэтому для прогнозирования токсичности целесообразно учитывать не только физико-химические свойства молекулы, но и её аффинность к токсикологически значимым белкам.

Наиболее распространённым направлением таких исследований остаются QSAR-модели, основанные на дескрипторах и статистических зависимостях между структурой и активностью. Эти подходы доказали эффективность при изучении отдельных классов соединений, но требуют ручного подбора признаков и ограничены линейными зависимостями. Развитие глубокого обучения позволило перейти к моделям, напрямую работающим со структурой молекулы.

Графовые нейронные сети (GNN) рассматривают соединение как граф, где вершины соответствуют атомам, а рёбра химическим связям. Такой подход позволяет извлекать признаки, отражающие как локальные, так и глобальные топологические свойства молекулы. В отличие от классических QSAR-моделей, графовые архитектуры не требуют

ручного описания структуры, но нуждаются в больших и биологически релевантных выборках для обучения, что ограничивает их применение в специализированных задачах.

В настоящей работе предложен подход, связывающий молекулярную аффинность и токсичность. Предполагается, что значения IC_{50} по белкам, вовлечённым в механизмы детоксикации, клеточного стресса и сигнальной регуляции, могут рассматриваться как прокси-показатели молекулярного токсического потенциала. Такая постановка задачи позволяет использовать доступные данные о взаимодействии «лиганд–белок» для построения графовой модели.

Предмет исследования: графовая модель для предсказания свойств химических соединений, связывающих аффинность и токсичность.

Объект исследования: азолопиримидиновые производные.

Цель работы: разработка и валидация модели прогнозирования молекулярной токсичности на основе графовых нейронных сетей и данных о взаимодействии соединений с токсикологически значимыми белками.

Для достижения поставленной цели решались следующие **задачи**:

1. Провести анализ существующих методов прогнозирования токсичности и определить их ограничения;
2. Сформировать и стандартизировать набор данных с фильтрацией по белковым мишеням, связанным с токсичностью;
3. Построить и обучить графовую модель аффинности с таргетом pIC_{50};
4. Проверить воспроизводимость и устойчивость модели на независимых данных.

Научная новизна заключается в использовании данных об аффинности в качестве количественной основы для моделирования токсичности. В отличие от существующих моделей

Практическая значимость работы заключается в том, что разработанный подход позволяет сократить объём дорогостоящих экспериментов, ускорить отбор перспективных соединений и может быть адаптирован к другим классам молекул. Модель служит фундаментальной основой для прогнозирования токсичности и оценки структуры–активности при ограниченных данных.

1 ЛИТЕРАТУРНЫЙ ОБЗОР

1.1 Текущее состояние разработки лекарственных средств

Разработка лекарственных средств представляет собой многоэтапный процесс поиска и всестороннего изучения молекул, обладающих потенциальной терапевтической активностью и подходящим профилем безопасности [1]. Этот путь включает в себя скрининг, доклинические и клинические испытания и отличается высокой степенью научной и финансовой сложности, значительными затратами ресурсов и высоким риском неудач.

Классический процесс поиска лекарств обычно начинается с идентификации биологической мишени, связанной с заболеванием, эмпирический скрининг тысяч соединений in vitro и in vivo, целевую оптимизацию химической структуры, многократные итерации синтеза, проверки их активности в биологических тестах [2].

Исследования показывают, что средняя стоимость разработки нового препарата в период с 2008 по 2019 год в США составляет приблизительно 172,7 млн долларов (в долларах США 2018 года), включая пострегистрационные исследования. С учетом стоимости неудачных попыток стоимость увеличивается до 515,8 млн долларов, а с учетом каптиализации стоимость разработки препарата составиляет около 879,3 млн долларов [3]. К 2020 году семь из 16 обследованных компаний «Большой фармы» имели отрицательную производительность НИОКР [4].

Согласно экономической оценке JAMA Network Open [3] с момента старта исследований до одобрения клинический путь лекарств составляет 10 лет. В таблице 1 представлены проценты затрат по блокам и сроки по фазам из исследования, охватывающего 13 терапевтических областей.

Таблица 1 – Структура финансовых и временных затрат на разработку ЛС, согласно экономической оценке JAMA Network Open [3]

Блок разработки	**Доля затрат , %**	**Средние сроки (месяцы)**
Поиск и доклинические испытания. Открытие и оптимизацию кандидатов, а также доклинические исследования in vitro и на животных	6.8% (95% ДИ 3.7–9.1)	31.2
Клинические испытания. Все стадии клинических испытаний на людях, направленные на оценку безопасности, дозировок и терапевтической эффективности	68.0% (95% ДИ 45.8–73.3)	I фаза: 27.8 II фаза: 34.0 III фаза: 38.0

Регуляторная подача и регистрация. Подготовка досье, подача заявки и рассмотрение материала регулирующими органами	1.5% (95% ДИ 1.3–2.0)	16.2
Пострегистрационные исследования. Исследования после выхода препарата на рынок, включая мониторинг долгосрочной безопасности и дополнительную оценку эффективности	23.7% (95% ДИ 17.7–47.7)	36.6

Если учесть не только прямые денежные расходы, но и риски неудач, и, так как разработка длится годы, учитывать стоимость дисконтирования вложений. Таким образом, реальные доли затрат смещаются (см. рисунок 1).

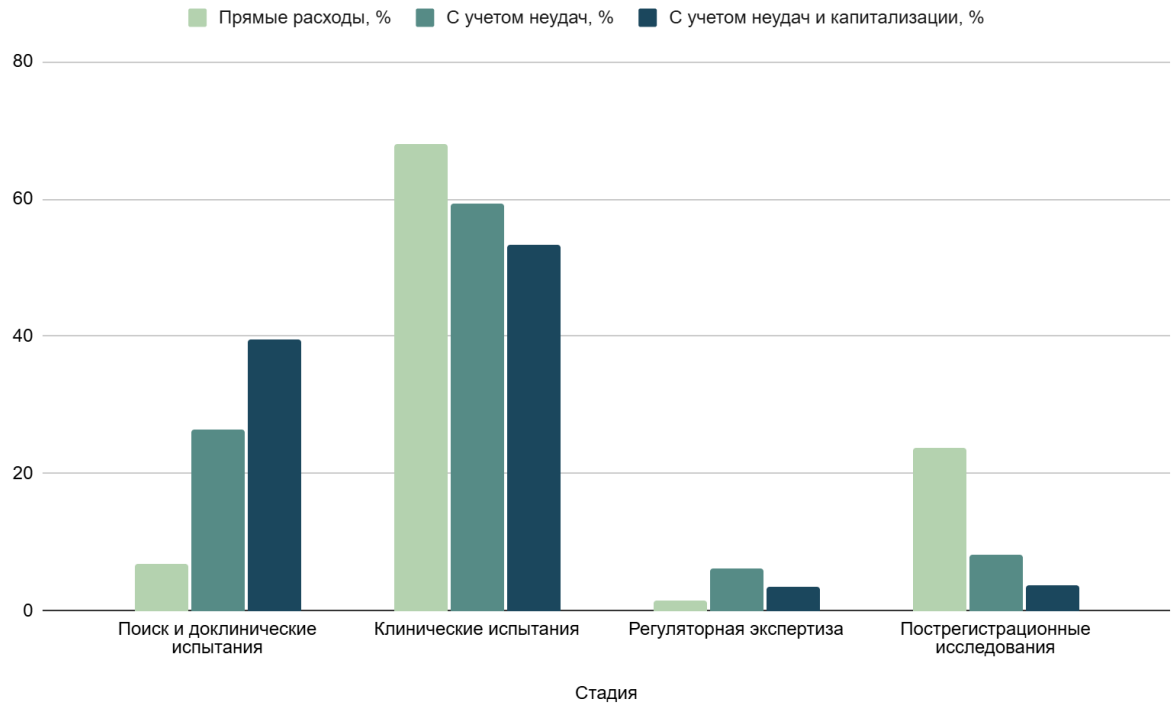


Рисунок 1 – Расходы на разработку ЛС с учетом неудач и дисконтирования за период разработки [3]

Анализ затрат показывает: если учитывать риск неудач и стоимость капитала, на поиск и доклинический этап приходится всё больше расходов. Это происходит потому, что на этих стадиях отсеивается большинство кандидатов, и каждое неудачное соединение добавляет скрытые издержки.

Исследование Waring, основанное на данных четырёх международных фармкомпаний в период за 2000–2010 г., показывает, что на доклиническом этапе первичной причиной прекращения разработки выступает неклиническая токсикология

(59.3 % случаев), фармакокинетика и биодоступность - 5.3 % случаев. Суммарно токсичность и ADME составляют около 65 % доклинического отсева кандидатов [5].

Таким образом, особое значение сегодня имеет доклинический этап разработки лекарств. Именно здесь решается, какие молекулы дойдут до клиники, а какие отпадут на раннем этапе. Чтобы этот процесс стал быстрее и точнее, нужны новые подходы к оценке активности и токсичности соединений.

Современные методы вычислительной химии и автоматизированный поиск соединений позволяют раньше увидеть потенциал молекулы и избежать затрат на бесперспективные варианты. Это снижает стоимость, ускоряет разработку и повышает шансы на успех.

1.2 Современные подходы к созданию лекарственных средств

Процесс поиска новых лекарственных средств прошёл путь от эмпирического использования природных экстрактов к рациональному молекулярному дизайну [6]. Развитие химии и биохимии позволило выделять активные компоненты, синтезировать соединения с заданными свойствами и изучать взаимодействие лигандов с мишенями. В дальнейшем появились комбинаторные библиотеки и технологии высокопроизводительного скрининга (HTS), позволившие тестировать десятки тысяч соединений и резко увеличить объём данных [7, 8].

Рост объёма экспериментальных данных и усложнение задач поиска активных соединений потребовали новых инструментов, способных обрабатывать большие массивы информации и прогнозировать свойства ещё до лабораторных испытаний. Так сформировалось направление *in silico*-исследований, основанное на компьютерном моделировании и анализе молекулярных структур.

1.2.1 Компьютерные методы (*in silico*-подходы)

Выбор метода поиска биологически активных молекул во многом определяется объемом доступной информации [1, 8]:

Известна трёхмерная структура мишени. В этом случае применяют структурно-ориентированные методы: молекулярный докинг (моделирование позы и энергии связывания лиганда в активном центре белка), структурно-обусловленный виртуальный скрининг крупных химических библиотек, построение фармакофорных моделей на основе структуры белка, а также молекулярную динамику для уточнения конформаций и оценки стабильности комплекса.

Известно активное соединение, но отсутствует информация о мишени. Используют лиганд-ориентированные стратегии: QSAR-моделирование, фармакофорный

анализ по известным лигандам, методы поиска «обратного докинга» и химико-геномные подходы для идентификации потенциальных белков-мишеней.

Нет ни структуры мишени, ни подтверждённых активных соединений. Применяют поисковые и фенотипические подходы, включая виртуальный скрининг больших химических коллекций по критериям пригодности соединений в качестве лекарств и прогнозируемым ADMET-параметрам, а также фенотипический скрининг с последующей биоинформатической обработкой данных для выделения первичных «хитов».

Частично известны и структура мишени, и набор активных соединений. Эффективной оказывается гибридная стратегия, комбинирующая лиганд- и структурно-ориентированные методы: предварительная фильтрация библиотек QSAR- или фармакофорными моделями с последующим докингом и молекулярной динамикой для приоритизации кандидатов.

Таблица 2 – Сравнительная таблица методов QSAR [2]

Информация о структуре	**Структура лиганда известна**	**Структура лиганда неизвестна**
Структура мишени известна (прямой дизайн)	Докинг. Моделирование связывания лиганда с активным центром белка и оценка энергии взаимодействия	De novo дизайн. Генерация новых молекул непосредственно в активном центре мишени
Структура мишени неизвестна (непрямой дизайн)	Аналоговый дизайн / QSAR. Поиск и оптимизация соединений на основе известных активных лигандов и количественных зависимостей «структура–активность»	Скрининг и поиск структурного сходства. Фильтрация больших библиотек по признакам, фармакофорам и прогнозируемым свойствам

У методов разная эффективность и есть ограничения. При докинге и дизайне de novo многое зависит от качества структуры мишени, гибкости, проработки учета воды, протонировки и других параметров. В QSAR-методах успех зависит от качества обучающей выборки, ее объема и сбалансированности данных [10].

Классические методы *in silico* остаются важным инструментом рационального дизайна, однако их эффективность ограничена зависимостью от качества исходных структур, точности параметризации и невозможностью учесть всю сложность биологических систем. Рост объёмов данных и разнообразие источников информации потребовали подходов, способных выявлять скрытые зависимости и интегрировать разнородные данные.

На этом фоне всё большую роль начинают играть методы искусственного интеллекта и машинного обучения, которые позволяют осуществлять поиск закономерностей в огромных объемах данных.

1.2.2 Задачи машинного обучения в поиске лекарственных средств

Машинное обучение позволяет формализовать эмпирическую зависимость между структурными характеристиками молекулы и её биологическими свойствами. Исходные данные обычно представляют в виде вектора дескрипторов или признаков (топологические, физико-химические, электронные параметры, отпечатки подструктур и др.).

Машинное обучение решает в фармацевтических исследованиях несколько ключевых задач:

Классификация помогает различать активные и неактивные соединения (например, ингибиторы киназ против неэффективных аналогов) или прогнозировать токсичность и проницаемость через мембраны.

Регрессия используется для расчёта количественных параметров, таких как pIC_{50}, logP или LD_{50}. Например, для оценки силы связывания лиганда с белком.

Кластеризация и поиск аналогов позволяют группировать соединения по структурному или фармакофорному сходству и находить потенциальные аналоги уже известных лекарств.

Генеративное моделирование применяют для создания новых структур с заданными свойствами. Например, молекул с улучшенной селективностью и сниженной токсичностью.

Эти методы превращают накопленные экспериментальные данные в источник предсказаний, помогая выявлять закономерности между структурой и активностью и формируя основу современного QSAR-моделирования.

1.2.3 Методы машинного обучения

Методы машинного обучения позволяют создать зависимость между структурными характеристиками молекулы и её биологическими свойствами. Исходные данные представляют в виде признаков (топологические, физико-химические, электронные параметры, отпечатки подструктур и др.). Алгоритм обучается на наборе молекул с известной активностью или токсичностью, выявляя статистические закономерности и формируя предсказательную модель.

Выбор метода машинного обучения зависит от типа данных и цели исследования. Если признаки молекулы хорошо описаны числовыми параметрами (масса, заряд, липофильность), применяют классические методы - логистическую регрессию, SVM, деревья решений. В работе “Предсказание лекарственно-индуцированной токсичности печени с использованием метода опорных векторов (SVM)” построили модели QSAR с SVM для прогнозирования гепатотоксичности на выборке 1 253 соединения, используя отбор признаков и SVM с кросс-валидацией [11].

Когда нужно повысить точность, используют ансамблевые методы, например, Random Forest или XGBoost. Они объединяют множество классических моделей и улучшают стабильность предсказаний. Random Forest широко используется в задачах QSAR и виртуального скрининга. Например, в исследовании “Random-forest model for drug–target interaction prediction via Kullback–Leibler divergence” модель на основе признаков показала точность ~0,882 и AUC ~0,990 при предсказании взаимодействий лекарств и белков [12]. Алгоритмы градиентного бустинга показали высокую точность в задачах ADMET. Например, веб-сервис ADMETBoost использует XGBoost для прогнозирования множества параметров ADMET; в исследованиях градиентный бустинг и LightGBM превосходили другие методы при прогнозировании биологической активности и свойств ADMET [13].

Большинство таких моделей используют только молекулярные признаки в качестве входных данных, что упрощает реальную молекулу. В обзоре [14] подчёркивается, что признаки могут не захватывать всю сложность процессов ADMET.

Поэтому при работе с большими и разнородными данными, где трудно заранее выделить признаки, применяют нейросетевые подходы. Глубокие нейронные сети способны сами извлекать информативные представления из «сырых» структур, например, из SMILES или молекулярных графов.

1.2.4 Глубокие нейросети и графовые модели

Классические алгоритмы могут оперировать только данными в эвклидовом пространстве (вектора фиксированной длины), тогда как в молекулах есть другие важные структурные данные: сети взаимодействий, белковые структуры, молекулярные графы. Недавно внедренные архитектуры глубокого обучения позволяют работать и с такими неевклидовыми данными (например, последовательности или графы), что открывает путь для более полной цифровой репрезентации химии в моделях [15].

Глубокие нейронные сети способны автоматически извлекать информативные представления из сырых данных - будь то графовое описание молекулы, последовательность SMILES или трёхмерная структура. В отличие от классических методов, где обучение идёт на статичном наборе признаков, нейросети строят свои представления признаков и улавливают сложные нелинейные зависимости, что делает их особенно перспективными для прогнозирования биологической активности и ADMET-характеристик [16].

1.2.5 Сравнение методов

Для более наглядного понимания различий в подходах было целесообразно сопоставить классические методы машинного обучения и нейросетевые модели, применяемые в поиске лекарственных средств. Такое сравнение позволяет подчеркнуть сильные и слабые стороны каждого класса алгоритмов и обосновать выбор направления дальнейшего анализа.

Таблица 3 – Сравнение классических методов машинного обучения и нейронных сетей в поиске лекарственных средств

Характеристика	Классические методы ML	Глубокие нейронные сети
Представление молекулы	Зависит от заранее рассчитанных дескрипторов (отпечатки подструктур, физико-химические параметры, топологические индексы)	Автоматическое извлечение признаков из представлений (SMILES, молекулярные графы, 3D-координаты)
Работа с нелинейностью	Улавливают нелинейные зависимости через ансамбли или ядровые функции, но в ограниченном объёме	Многоуровневые представления позволяют описывать сложные нелинейные зависимости и взаимодействия
Интерпретируемость	Хорошая (важности признаков, анализ правил)	Более ограниченная, требует специальных методов пояснения
Требования к данным	Эффективны при умеренных наборах данных	Требуют больших массивов данных для надёжного обучения
Производительность	Высокая при небольших данных и простых зависимостях; быстрая настройка	Превосходят при больших и разнородных датасетах; лучше обобщают на новые химические классы
Применение в фармацевтике	QSAR-моделирование, ADMET-прогнозы, фильтрация библиотек	Виртуальный скрининг, генеративный дизайн молекул, предсказание активности, селективности и токсичности

Сравнение классических алгоритмов машинного обучения и нейросетевых подходов показывает, что первые остаются востребованными благодаря простоте реализации, интерпретируемости и высокой предсказательной способности на ограниченных выборках. Однако по мере роста доступных данных и усложнения задач именно нейросетевые модели демонстрируют более высокую эффективность и универсальность.

1.3 Графовые нейросети как инструмент предсказания свойств

1.3.1 Принцип работы графовых моделей

Современные методы представления молекул в задачах машинного обучения делятся на линейные и графовые. В линейных подходах молекула кодируется строкой SMILES компактной, но чувствительной к перестановкам: небольшое изменение символов может соответствовать другой структуре или даже невалидной молекуле.

Графовые модели опираются на естественное представление химического соединения как графа: атомы выступают вершинами, а химические связи рёбрами. Такой формат сохраняет топологию и позволяет моделировать взаимодействия между атомами напрямую, без преобразования в заранее рассчитанные признаки.

На рисунке 2 показан общий принцип работы графовых нейросетей:

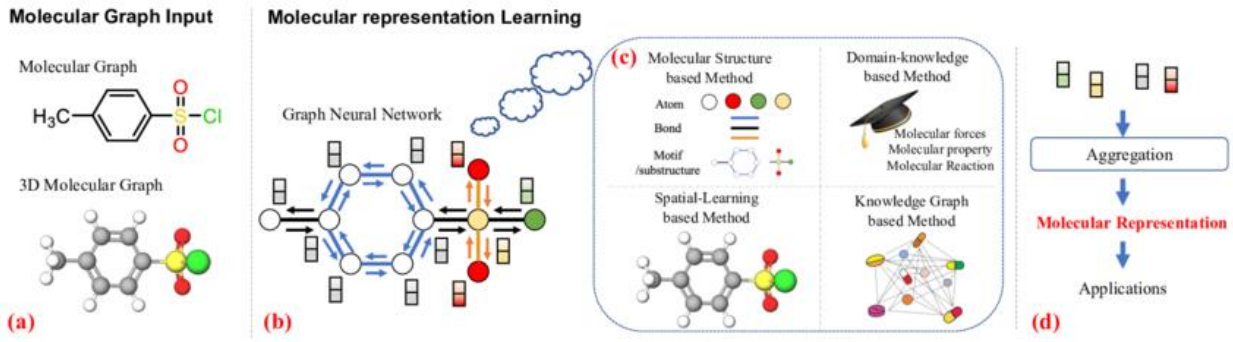


Рисунок 2 – Принцип работы графовой нейросети [17]

1. На вход подаётся молекулярный граф, где каждому атому приписан вектор признаков (атомный номер, валентность, ароматичность и др.), а для каждой связи характеристики типа связи и стереохимии.
2. Далее сеть выполняет итеративное распространение сообщений: информация от соседних атомов и связей агрегируется и обновляет внутреннее представление узла. На каждом слое модель «видит» всё более широкий фрагмент молекулы: от ближайшего окружения атома до всей структуры.
3. После нескольких итераций скрытые представления атомов объединяются (aggregation) в общий вектор молекулы: молекулярное вложение (molecular representation), которое используется для предсказания активности, токсичности или других свойств.

Таким образом, ключевое отличие графовых моделей от методов на основе признаков заключается в том, что признаки не задаются вручную, а извлекаются автоматически из топологии молекулы и распределяются по уровням представления, отражая как локальные подструктуры, так и глобальную архитектуру соединения [17].

1.3.2 Применение графовых моделей в прогнозировании

Исследования показывают, что применение GNN для предсказания свойств молекул даёт более высокую точность по сравнению с традиционными алгоритмами, основанными на фиксированных признаках [18]. В частности, GNN доказали свою эффективность в решении задач прогнозирования физических и ADMET-свойств [19].

В 2018 г. MoleculeNet представляет собой набор эталонных датасетов для задач хемоинформатики (активность, растворимость, ADMET и др.), использованный для систематического сравнения ML- и GNN-моделей. В ряде задач графовые модели показали устойчивое преимущество перед классическими подходами. Это исследование стало ключевым для стандартизации оценки GNN в области drug discovery [20].

В исследовании Stokes et al., 2020 авторы применили глубокую нейросетевую архитектуру с графовым представлением молекул для поиска новых антибиотиков. Модель предсказала соединение халицин, обладающее принципиально новым механизмом действия и активностью против широкого спектра патогенов, включая резистентные штаммы. Это стало одним из первых примеров, когда GNN фактически привели к открытию нового класса лекарств [21].

В реестре последних достижений в области глубокого графового обучения для рассматриваются архитектуры (MPNN, GCN, GAT), примеры применения в прогнозировании активности, ADMET и молекулярного дизайна. GNN становятся одним из основных инструментов современной фармацевтической разработки [22].

Согласно последним данным GNN рассматриваются как один из наиболее перспективных и быстро развивающихся инструментов, способных существенно изменить практику поиска и оптимизации лекарственных средств, а количество исследований и приложений в этой области неуклонно растёт: авторы отмечают экспоненциальное увеличение числа работ по применению GNN в фармацевтической разработке за последние 5 лет, особенно в США и Китае [23] (см. рис. 3).

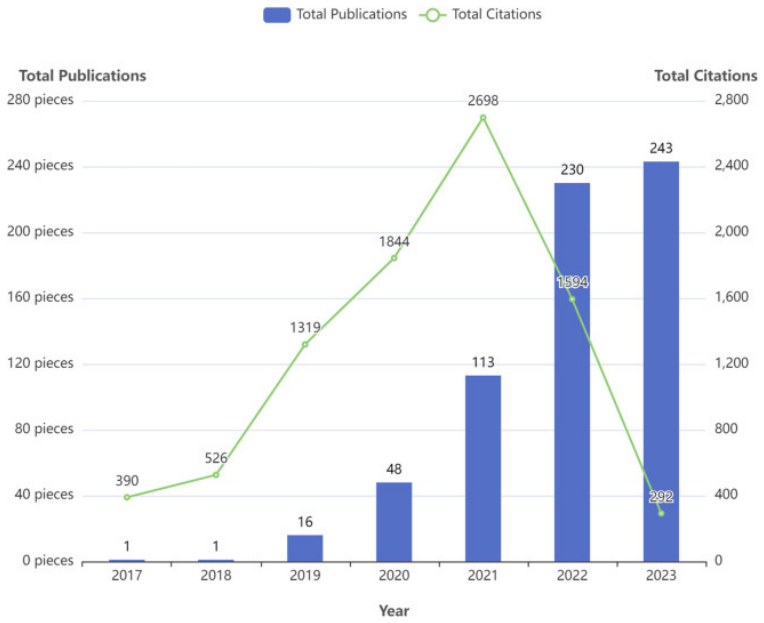


Рисунок 3 – Частота публикаций о GNN с 2017 по 2023 г. [23]

1.3.3 Ограничения графовых моделей

Хотя GNN имеет очевидные преимущества в поиске лекарств, они также создают проблемы, такие как недостаточная интерпретируемость модели, потребность в больших объемах маркированных данных для обучения и высокое потребление вычислительных ресурсов [24].

Сложность интерпретации. Одно из традиционных преимуществ классических моделей - это относительная простота и интерпретируемость (например, можно увидеть вклад каждого признака в активность). Глубокие же модели трудно поддаются объяснению. Это касается и GNN: хотя они отражают структуру молекулы, понять, какие именно фрагменты или взаимодействия привели к высокому предсказанию, не всегда легко [25].

Зависимость от больших данных. GNN требуют значительных объёмов качественных примеров для обучения, особенно если структура химических пространств сложна и разнообразна. Во многих задачах фармацевтики таких наборов просто нет [26].

Высокие вычислительные расходы. Обучение GNN на больших молекулярных библиотеках и с многочисленными слоями сообщений потребляет значительные ресурсы GPU/CPU и время, особенно при оптимизации гиперпараметров или масштабировании на сотни тысяч соединений.

1.3.4 Пример реализации: Chemprop

Chemprop представляет собой открытую программную платформу, разработанную в Массачусетском технологическом институте (MIT), основанную на архитектуре Directed Message Passing Neural Network (D-MPNN). В отличие от классических MPNN, в D-MPNN сообщения распространяются вдоль ориентированных рёбер, что позволяет избежать дублирования информации и улучшает стабильность обучения [27]. На этапе агрегации формируется молекулярное векторное представление, которое далее используется для решения задач регрессии (например, предсказание pIC_{50}), классификации или в многозадачных постановках.

Одним из преимуществ Chemprop является возможность добавления внешних признаков, которые включаются как дополнительный канал в модель. Это делает архитектуру «гибридной»: она объединяет силу автоматического извлечения признаков графовой сетью и дополнительную информацию из физико-химических характеристик, что особенно полезно при небольших объёмах данных или задачах с выраженной зависимостью от глобальных молекулярных свойств.

Эффективность Chemprop подтверждена сравнительными исследованиями: в классической работе Yang et al. D-MPNN часто превосходит лучшие модели на базе MoleculeNet [27]. В описании самого пакета Chemprop [28] приводятся результаты, где D-

MPNN-модели достигают передовой производительности на наборах MoleculeNet и SAMPL по свойствам вроде logP, реакционных барьеров и атомных зарядов. Обзорные статьи также указывают на то, что D-MPNN остаётся одним из лидеров среди графовых архитектур для молекулярных задач [29].

Chemprop – это одна из самых сильных архитектур для молекулярных данных на сегодняшний день. Учитывая универсальность, репутацию и валидированные результаты, выбор Chemprop для моделирования свойств является обоснованным и отвечает современным стандартам анализа молекулярных данных.

1.4 Современные подходы к прогнозированию токсичности

1.4.1 Традиционные подходы к прогнозированию ADMET-свойств

Оценка характеристик ADMET является ключевым этапом при отборе кандидатов. Как правило, прогноз токсичности строится по данным экспериментальных тестов, представленных в специализированных базах Tox21, TOXRIC, ADMETlab и других, где собраны значения цитотоксичности, гепатотоксичности и других конечных эффектов для множества соединений.

Традиционно прогноз таких свойств осуществляется с помощью QSAR-моделей, в которых используются заранее рассчитанные свойства: молекулярная масса, доноры и акцепторы водородных связей, функциональные группы и др.) [30, 31]. Подобные модели хорошо работают для ограниченных химических семейств, однако с трудом обобщаются на новые классы соединений и зависят от качества экспериментальных данных.

Переход к графовому представлению молекул позволил повысить обобщающую способность моделей и точность прогнозов [32]. Модели учитывающие пространственную конфигурацию молекул, показали точность, сопоставимую с лучшими 2D-моделями, особенно на крупных и разнородных датасетах [35].

Появляются и специализированные архитектуры, направленные на отдельные типы токсичности: например, GraphADT учитывает связи между атомами как отдельные узлы и выделяет функциональные группы, повышая интерпретируемость прогноза [36].

1.4.2 Альтернативные методы: аффинность к белкам

Когда данных о токсичности соединений недостаточно, перспективным решением становится использование косвенных признаков, в частности, аффинности молекул к белкам. Такой подход основан на том, что токсическое действие часто связано с взаимодействием соединений с белками, участвующими в жизненно важных процессах клетки.

Известно, что связывание с ферментами метаболизма (например, цитохромами P450), белками окислительного стресса или компонентами сигнальных путей апоптоза может приводить к развитию нежелательных эффектов. Поэтому высокая аффинность к подобным мишеням рассматривается как возможный признак токсического потенциала соединения [37].

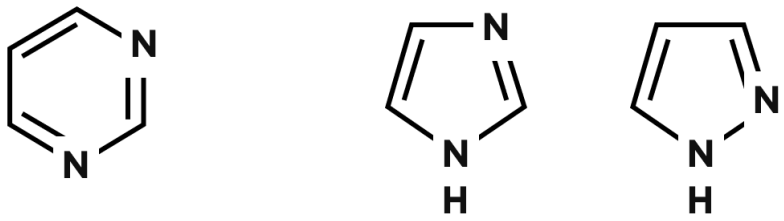
Подобный перенос информации позволяет обходить дефицит прямых данных и обобщать закономерности на более широкое химическое пространство.

1.5 Азолопиримидины как объект исследования

1.5.1 Биологическая активность и значимость

В данной работе объектом исследования выбраны азолопиримидины - класс соединений, обладающий высоким потенциалом как противоопухолевые агенты. Химически они содержат один азольный цикл, примеры на рис. 4 (б), и пиримидиновое кольцо (рис. 4, а).

Азолы - это пятичленные гетероциклы, имеющие в цикле не менее двух гетероатомов, один из которых атом азота, а также би- и полициклические соединения, включающие азольный цикл. Пиримидин (C4N2H4, 1,3- или м-диазин, миазин) - гетероциклическое соединение, простейший представитель 1,3-диазинов.



а) б)

Рисунок 4 – Структурные части азолопиримидинов

Такая комбинация обеспечивает богатое химическое разнообразие и возможность варьирования заместителей, что делает его удобным для направленного структурного дизайна.

Гетероциклические соединения имеют решающее значение в разработке новых противораковых средств. Более 85% молекул лекарственных препаратов, одобренных FDA, содержат гетероциклы, и, что наиболее важно, многочисленные гетероциклические лекарственные молекулы демонстрируют потенциальную эффективность в лечении ряда злокачественных новообразований [38]. В статье FDA-approved heterocyclic molecules for cancer treatment прямо подчёркивается, что целый ряд родственных пиримидин-фьюжн

соединений входят в число одобренных FDA противоопухолевых препаратов. Пиразолопиримидиновый каркас встречается в одобренных препаратах и обладает подтверждённой активностью против широкого спектра терапевтических мишеней, включая CDK и B-Raf киназы, что демонстрирует его значимость для противоопухолевой терапии [39]. Производные этого ряда способны ингибировать широкий спектр биомишеней, однако их селективность и профиль токсичности остаются недостаточно изученными.

Таким образом, выбор азолопиримидинов обоснован как с точки зрения химической структуры, так и фармакологического потенциала.

1.5.2 Особенности для графовой модели

Биологическая активность азолопиримидинов в значительной степени определяется настройкой заместителей, которые могут изменять электронные, стерические и фармакокинетические характеристики соединений. Это требует от модели высокой чувствительности к малым структурным различиям.

Исключительно графовые модели иногда теряют глобальную статистическую информацию, которая полезна при малых сдвигах активности. В литературе отмечается, что комбинация графовых моделей с классическими молекулярными признаками обеспечивает выигрыш в точности предсказаний при работе с узкими сериями структурных аналогов и при ограниченных объёмах данных [40].

Поэтому добавление признаков и отпечатков даёт модели дополнительный канал информации. Признаки фиксируют свойства, напрямую влияющие на ADMET-профиль (logP, TPSA, доноры/акцепторы водородных связей и др.), а Morgan-отпечатки позволяют явно закодировать повторяющиеся фрагменты, характерные для исследуемого ряда соединений. Таким образом, гибридный подход сочетает автоматическое извлечение признаков через GNN и использование заранее известных молекулярных характеристик.

1.6 Источники и подготовка данных для моделирования

Применение методов машинного обучения в химии и фармакологии начинается с формирования качественного набора данных. Наличие репрезентативных и корректно размеченных данных определяет воспроизводимость модели и её прогностическую силу. Существующие базы данных можно условно разделить на два класса: крупные токсикологические коллекции (ориентированные на молекулярные и биохимические показатели); наборы, основанные на данных по клеточным линиям, позволяющие оценивать цитотоксичность и селективность.

1.6.1 Базы данных

Большинство открытых токсикологических коллекций представляют данные в бинарном формате: соединения классифицируются как «токсичные» или «нетоксичные» на основе пороговых значений активности. Для построения количественных моделей требуются источники, содержащие показатели IC_{50} или родственные величины, позволяющие оценивать степень взаимодействия соединений с биологическими мишенями.

В ряде исследований показано, что аффинность к определённым белкам может рассматриваться как косвенный индикатор токсичности. Так, Raunio et al. (2015) [40] продемонстрировали, что связывание ксенобиотиков с ферментами цитохрома P450 связано с риском метаболической токсичности; Cohen (1997) показал, что селективное связывание реактивных метаболитов с белками-мишенями коррелирует с поражением органов; Rashid et al. (2025) установили связь между аффинностью микропластиков к CYP1A1 и изменением активности фермента, приводящим к токсическим эффектам. Ниже приведены ключевые источники, подходящие для подобных исследований:

TOXRIC: база, агрегирующая более 1400 токсикологических эндпоинтов из разнородных источников (PubChem, ToxRefDB, eChemPortal) для обучения мультизадачных моделей [41].

TDC ADMET: Модуль платформы TDC, включающий наборы по абсорбции, распределению, метаболизму, экскреции и токсичности. Часть коллекций содержит количественные данные на определенные типы токсичности [34].

BindingDB: База данных, включающая миллионы записей о взаимодействиях лиганд–белок с количественными показателями (IC_{50}, K_i, K_d) и унифицированными структурами в формате SMILES. Используется для построения QSAR-моделей и оценки аффинности соединений. В отличие от клеточных баз, отражает молекулярные взаимодействия [42].

1.6.2 Практики предобработки и стандартизации данных

Большинство публичных и лабораторных баз содержат ошибки, дубликаты, неоднородные форматы записи или несогласованные измерения активности, что требует систематической предобработки. Корректная предобработка структур и биологических измерений является ключевым условием воспроизводимости результатов в машинном обучении для хемоинформатики. В литературе описаны следующие наиболее распространённые практики:

Единообразие биологических значений. Параметры активности рекомендуется приводить к логарифмическому масштабу, что улучшает распределение данных и делает возможным сопоставление результатов из разных экспериментов [31].

Стандартизация структур. Для устранения неоднозначностей применяют канонизацию SMILES, удаление солей и протонованных форм, нейтрализацию зарядов, выбор крупнейшего фрагмента по числу тяжёлых атомов, стандартизация таутомеров [43].

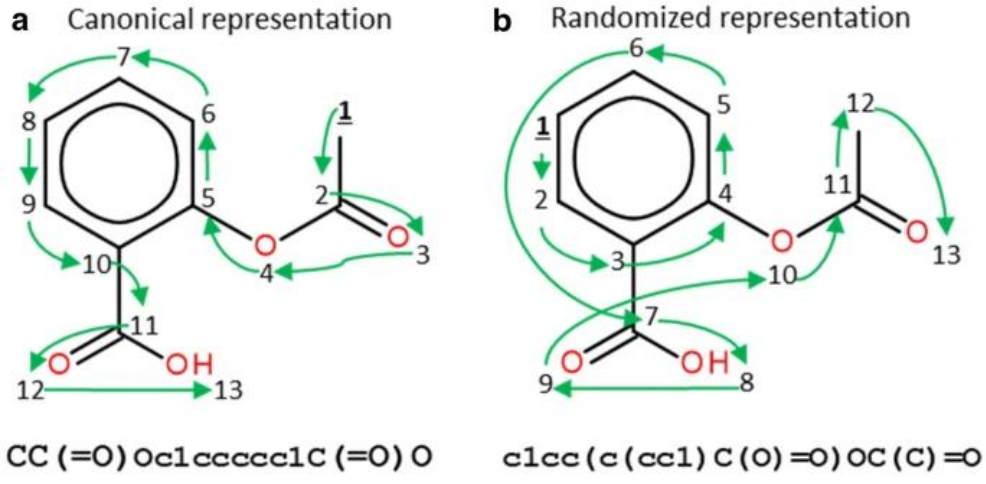


Рисунок 5 – Разные типы прочтения SMILES [44]

Эти шаги позволяют минимизировать дублирование и повысить согласованность наборов .

Удаление неорганических и металлических соединений. Многие ML-модели рассчитаны на органические молекулы, поэтому такие записи исключают или выделяют в отдельные задачи.

Агрегирование повторных измерений и удаление дубликатов. Стандартизованные протоколы предусматривают удаление полных дубликатов, а в случаях, когда для одного соединения доступны несколько значений активности, рекомендуется использовать медиану или среднее значение, чтобы снизить влияние выбросов и межэкспериментальной вариабельности [45].

Ограничение диапазонов. Часто обрезают значения выше определённого предела, приравнивая их к пороговым. Данные вне 1-го и 99-го перцентиля могут удаляться, чтобы убрать экспериментальный шум.

Таким образом, лучшие практики подготовки данных предполагают поэтапную стандартизацию структур и значений активности, а также их агрегирование и фильтрацию. Эти действия снижают риск появления ошибок при моделировании.

1.6.3 Молекулярные дескрипторы

Важнейшей частью подготовки данных является расчёт молекулярных дескрипторов. Дескрипторы представляют собой числовые характеристики, описывающие химическую структуру в форме, пригодной для алгоритмов машинного обучения. Они позволяют связать молекулярное строение с биологической активностью и фармакологическими свойствами.

Физико-химические параметры.

Базовые показатели, такие как молекулярная масса, коэффициент липофильности (LogP), площадь полярной поверхности (TPSA), число доноров и акцепторов водородных связей, отражают свойства, связанные с растворимостью, проницаемостью и связыванием с мишенями. Подобные дескрипторы широко применяются, например, для оценки «правила пяти Липински» при отборе кандидатов в лекарственные средства [46].

Топологические и электронные дескрипторы.

Наборы, реализованные в RDKit (SlogP_VSA, EState_VSA, BCUT2D, Kappa и др.), позволяют описывать распределение электронной плотности, топологическую сложность и вариации поверхности молекулы. Эти параметры успешно применялись в задачах QSAR для предсказания растворимости, токсичности и сродства к белковым мишеням [].

Каркасная структура.

Определение каркасов по Мёрко [47] используется для описания и сравнения химических структур на уровне их базовой архитектуры. Каркас по Мёрко представляет собой «скелет» молекулы, совокупность кольцевых систем и соединяющих их звеньев, без учёта заместителей. Такой подход позволяет выделить ядро соединения, общее для группы аналогов, и анализировать, как различные функциональные группы влияют на активность.

Функциональные фрагменты.

Использование SMARTS-шаблонов позволяет количественно оценивать присутствие отдельных подструктур: амидов, сульфонамидов, нитрилов, простых эфиров и др. Такие фрагменты нередко ассоциированы с характерными фармакофорами или токсофорами, что даёт возможность выявлять закономерности «структура–активность».

Таблица 4 – Основные группы молекулярных дескрипторов и их применение

Группа дескрипторов	Примеры показателей	Что характеризуют	Использование в исследованиях
Физико-химические	Молекулярная масса, коэффициент липофильности (LogP), площадь полярной поверхности (TPSA), число доноров и акцепторов водородных связей, число вращательных связей	Растворимость, проницаемость через мембраны, способность к образованию межмолекулярных взаимодействий	Отбор соединений с приемлемыми свойствами для биодоступности и начального скрининга
Топологические	Количество колец, число ароматических колец, индексы формы (Kappa)	Геометрия молекулы, структурная сложность, наличие циклических систем	Сравнение структур, оценка разнообразия химических библиотек

Продолжение таблицы 4

Электронные	Зарядовые индексы (EState), распределение поверхности (VSA-параметры), дескрипторы BCUT	Электронная структура, распределение заряда и полярности	Прогноз растворимости, сродства к белковым мишеням, метаболической стабильности
Фрагментные	Подсчёт функциональных групп (амиды, эфиры, сульфонамиды, нитрилы и др.)	Наличие и частота встречаемости характерных химических мотивов	Связь между отдельными функциональными группами и активностью или токсичностью
Каркасные	Каркасы по Мёрко (Murcko scaffold)	Базовая архитектура молекулы (скелетная структура)	Анализ структурного разнообразия, поиск новых серий соединений

Таким образом, молекулярные дескрипторы охватывают разные уровни представления молекулы. Их использование является стандартной практикой и обеспечивает основу для построения прогностических алгоритмов.

1.6.4 Разбиение данных

Корректное разбиение данных на обучающую, валидационную и тестовую выборки является базовом условием честной оценки моделей: обучение ведётся на train (обучающая), параметры/ранняя остановка подбираются по validation (валидационная), а единственная итоговая оценка даётся на отложенном test-наборе (тестовая), не использованном ни на одном этапе настройки. Такой протокол и его вариации (включая кросс-валидацию) подробно обсуждаются в учебной литературе по статистическому обучению [48]. От выбора метода зависит, насколько объективно будет оценена способность модели воспроизводить свойства новых молекул.

В работе были протестированы три подхода:

Случайное разбиение (random split). Наиболее простой способ, при котором молекулы случайным образом распределяются между обучающей и тестовой выборками. Данный метод обеспечивает равномерное распределение по активности, но имеет существенный недостаток: в тестовую выборку могут попасть молекулы, структурно почти идентичные тем, что использовались для обучения. В таком случае модель демонстрирует завышенную точность, не отражающую её способность к обобщению. В химии простое случайное разделение часто переоценивает качество, потому что в тест попадают

молекулы, очень похожие на обучающие. Поэтому широко применяют химически осмысленные схемы [20]

Разбиение по скэффолдам (scaffold split). Здесь для каждой молекулы выделяется так называемый «каркас». Разбиение проводится так, чтобы молекулы с одинаковым каркасом полностью относились либо к обучающей, либо к тестовой выборке. Такой подход позволяет проверить, насколько модель способна предсказывать активность на принципиально новых химических каркасах. Однако у метода есть ограничения: распределение активностей между выборками может оказаться неравномерным, а очень крупные скэффолды могут смещать баланс данных.

Разбиение по соседям (neighbor split). Суть метода заключается в том, что «соседями» считаются молекулы, имеющие высокое сходство. Для каждой пары соединений рассчитывается коэффициент Танимото (Tanimoto); значения выше порогового интерпретируются как высокая структурная близость. Он позволяет исключить из теста молекулы, структурно слишком похожие на обучающие, то есть проверить модель на «новых» химических пространствах, сохранив разнообразие. Но в то же время выявить слабые места модели именно на структурно близких, но резко различающихся по активности молекулах, что невозможно при случайном разбиении и более точно, чем при скэффолд-разбиении.

Таблица 5 –Сравнение методов разбиения данных

Метод	Принцип разбиения	Сильные стороны	Слабые стороны
Random split	Молекулы распределяются случайно по выборкам	Простая реализация; равномерное распределение по активности	В тесте могут оказаться почти идентичные молекулы → завышение точности
Scaffold split	Молекулы с одинаковым Bemis–Murcko каркасом полностью относятся к одной из выборок	Проверка способности модели предсказывать новые каркасы; исключает «утечку скэффолда»	Дисбаланс: крупные каркасы могут смещать данные; распределение активностей между выборками искажено
Neighbor split	В тест попадают только молекулы, у которых нет ближайшего соседа в train (ECFP4, Tanimoto)	Строгая проверка переносимости; исключает «близнецов»; сохраняет разнообразие химических классов	Более сложная реализация; часть молекул исключается при жёстком пороге

1.7 Итоги анализа литературы

Современная система разработки лекарственных средств остается затратной и длительной, а основной причиной неудач на поздних этапах служат неблагоприятные ADMET-характеристики кандидатов. Это подчёркивает необходимость точного *in silico*-прогнозирования токсичности уже на ранних стадиях исследования соединений.

Анализ литературы показывает, что традиционные методы машинного обучения ограничены заранее выбранными дескрипторами, тогда как глубокие нейронные сети, особенно графовые архитектуры, способны напрямую использовать топологию молекулы, автоматически извлекать признаки и демонстрируют более высокую точность при прогнозировании активности и токсичности. С учётом сравнительного анализа методов и источников данных, наиболее целесообразным является применение графовой нейросетевой модели Chemprop для предсказания токсичности.

В рамках данной работы выбран класс азолопиримидинов: перспективных противоопухолевых соединений с выраженной структурной вариабельностью. Использование профилей аффинности к белкам как косвенного показателя токсичности позволяет компенсировать дефицит прямых экспериментальных данных и выявлять закономерности, характерные для редких химических семейств. Прямых исследований, где такой подход применялся, немного, что придаёт нашей работе элемент научной новизны.

2 ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

2.1 Методы и инструменты

Настоящий раздел содержит описание методологической основы исследования: используемых источников данных, вычислительных средств, принципов построения, обучения и оценки моделей.

2.1.1 Источники данных

В работе используется база BindingDB [42], содержащая данные о биологической активности более 1 миллиона соединений, включая SMILES-представления молекул, UniProt-ID целевых белков и значения аффиности IC_{50}.

2.1.2 Инструменты и программная среда

Обучение моделей проводилось на облачных вычислительных ресурсах Kaggle и Google Colab, с использованием графических ускорителей NVIDIA Tesla T4 / P100.

Анализ данных и моделирование выполнялись в среде Python 3.x с использованием библиотек pandas, numpy, scikit-learn и RDKit. Для построения графовых нейросетевых моделей применялся пакет Chemprop. В качестве средства визуализации были использованы библиотеки matplotlib, seaborn.

2.1.3 Построение моделей

Для построения моделей использовалась графовая нейросетевая архитектура Chemprop. Модель обучалась в режиме регрессии для предсказания значений pIC_{50}. В качестве входных данных использовались молекулярные графы, при необходимости дополненные вычисленными признаками из RDKit (молекулярные дескрипторы и отпечатки Morgan).

2.1.4 Разделение данных

Использовались стратегии random split, scaffold split в Модели №1 и neighbor split в Модели №2.

2.1.5 Метрики

Для оценки качества предсказаний применялись метрики:

MAE (Mean Absolute Error) – среднее отклонение предсказанных значений от экспериментальных.

$$MAE = \frac{1}{n} \sum_{i} \quad | \hat{y}_i - y_i |$$

RMSE (Root Mean Square Error) – корень из среднеквадратичной ошибки. Чувствителен к крупным отклонениям и отражает разброс ошибок.

$$RMSE = \sqrt{\left(MAE = \frac{1}{n} \sum_{i} \quad | \hat{y}_i - y_i | \right)^2}$$

R^2 (коэффициент детерминации) – показывает долю объяснённой моделью вариации признака. Следует учитывать, что величина R^2 зависит от разброса истинных значений и не всегда сопоставима между различными сплитами.

$$R^2 = 1 - \frac{\sum_i \quad | \hat{y}_i - y_i |^2}{\sum_i \quad | \hat{y}_i - y_i |^2}$$

Дополнительные метрики:

bias (средняя знаковая ошибка) – показатель систематического смещения (завышение или занижение предсказанных значений).

$$bias = \frac{1}{n} \sum_{i} \quad | \hat{y}_i - y_i |$$

slope и intercept – параметры линейной регрессии «предсказание против эксперимента», отражающие калибровку модели:

$$\hat{y} \approx slope * y + intercept$$

slope свидетельствует о сжатии или расширении диапазона (экстремальные

intercept указывает на глобальный сдвиг предсказаний.

2.1.6 Визуализации и представления результатов

Для углублённой диагностики качества использовались графические методы:

Parity plot (истина против предсказания) – точки, расположенные вдоль диагонали y = x, соответствуют идеальным предсказаниям. Смещение облака относительно диагонали указывает на bias, а наклон облака – на масштабное искажение (slope).

Гистограмма остатков (predicted/true) – позволяет оценить распределение ошибок: центр около нуля указывает на отсутствие систематического смещения, ширина распределения – на дисперсию ошибки. Для удобства интерпретации использовался перевод в кратности IC_{50}: ошибка 0,3 pIC_{50} соответствует приблизительно двукратному изменению IC_{50}, а 0,6 pIC_{50} – четырёхкратному.

Диаграмма Бланда–Альтмана (ошибка против среднего значения) – отображает систематические смещения и гетероскедастичность.

Горизонтальное облако вокруг линии bias ≈ 0 без тренда свидетельствует о корректной калибровке модели по всему диапазону значений. Узкие границы согласия (bias ± 1,96·SD) интерпретировались как признак стабильности предсказаний.

2.1.7 Ограничения и допущения

Наличие ограничений по вычислительным ресурсам определяло выбор архитектур и глубину моделей. Они требовали уменьшения числа эпох обучения и глубины архитектуры нейронных сетей.

Кроме того, при построении модели использовались медианные значения IC_{50} по ограниченному числу белковых мишеней, что также следует учитывать при интерпретации результатов.

2.2 Результаты обучения Модели №1.

2.2.1 Данные для Модели № 1

Исходная выборка для обучения модели была сформирована на основе данных BindingDB и содержала 990 630 записей. После этапа предварительной обработки, включавшего удаление дубликатов, очистку и стандартизацию структур, в рабочей выборке осталось 65964 уникальных молекулярных описаний. На рисунке 6 представлено итоговое распределение количества молекул по значениям pIC_{50} после всех этапов очистки и фильтрации.

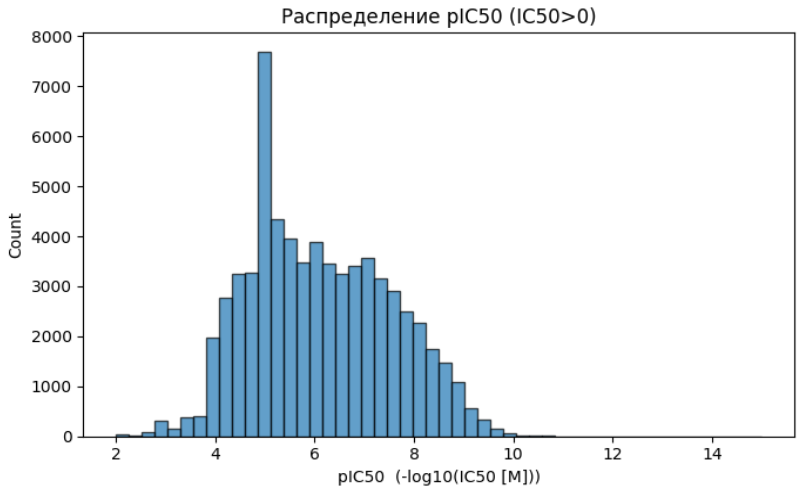


Рисунок 6 – Распределение количества молекул в рабочей выборке относительно pIC_{50} после предобработки

Статистические характеристики распределения значений pIC_{50} после предобработки выглядели следующим образом:

Таблица 9 – Статистические значения pIC_{50} в рабочей выборке

Количество	**Mean**	**Std**	**Min**	**Median**	**Max**
65 964	6.111	1.416	2.000	5.979	15.000

Для обучения модели использовались пары вида «структура –активность», где структура задавалась в формате SMILES, а активность в виде экспериментального значения pIC_{50}.

Таблица 10 – Примеры записей обучающей выборки

smiles	**pIC_{50}**
Bc1ccc(F)c2c1CC[C@H]2Nc1ccc([C@H]2C[C@@H]2C(=O...	4.349485
Bc1ccc(S(=O)(=O)N2CCN(C(=O)N3CCN(c4cc(C)ncn4)C...	5.000000
CN1[C@@H]2C[C@H](OC(=O)[C@H](CO)c3ccccc3)C[C@H...	3.876148
Cc1ccc(C(=O)Cn2c3c(sc2=N)CCCC3)cc1	6.750845
BrC1=CC[C@@H]2C(=C1)[C@@H]1c3cc(Br)ccc3C[C@@H]...	5.299296

2.2.2 Сравнение разбиений

Результаты проверки разбиений на random и scaffold представлены в таблице:

Таблица 11 – Сравнение результатов разбиений Модели №1

	Кол-во	**RMSE**	**MAE**	**R^2**	**Spearman ρ**	**Improvement**	**Bias**	**Slope**	**Intercept**
Наивный предиктор	6 595	1.4156	1.1913	-	-	-	-	-	-
Random split	6 595	0.7827	0.5585	0.6943	0.8323	44.7 %	0.029	0.745	1.591
Scaffold split	6 593	0.8914	0.6365	0.6027	0.7716	37.0 %	−0.055	0.693	1.804

Независимо от способа разбиения данных, все протестированные модели демонстрировали значимое улучшение качества прогнозирования по сравнению с наивным предиктором, снижая RMSE на 37–45 %. При этом параметры линейной аппроксимации выявили характерные особенности работы моделей: значение наклона (slope) ниже 1 и смещение intercept указывали на тенденцию к сжатию диапазона предсказаний и систематическую ошибку, связанную с неполной калибровкой. Такая ситуация может привести к недооценке или переоценке крайних значений активности: модель хуже различает наиболее активные и неактивные соединения. Подобные эффекты часто наблюдаются при обучении на данных с широкой вариативностью химических структур и небольшим числом информативных дескрипторов.

Рисунок 7 –Результаты Модели №1

С учетом технических ограничений, умеренное различие между результатами random и scaffold split позволило нам остановится на scaffold split при выборе фундаментальной архитектуры, поскольку это признанный стандарт оценки для молекулярных задач. Показатель R^2 выше 0,6 был признан удовлетворительным для фундаментальной модели, в связи с широтой диапазона изучаемых молекул и ограниченного числа признаков.

2.3 Результаты обучения Модели №2

2.3.1 Данные для Модели 2

После применения структурного фильтра (SMARTS-паттерн), для выделения азолопиримидинов из исходной выборки и фильтрации по токсикологически релевантным мишеням из базы было получено 732 уникальные молекулы, соответствующие 42 белкам из панели.

Наиболее представленные мишени: MTOR, EGFR, ERBB2, AKT1, KCNH2, VEGFR2, HTR2C, GRM5, PDGFRA и CYP2D6. Эти белки относятся к ключевым узлам сигнальных и метаболических путей, задействованных в регуляции роста, дифференцировки и жизнеспособности клеток. Их аффинность рассматривается как прокси-показатель потенциальной цитотоксичности, поскольку ингибирование данных мишеней может приводить к нарушениям клеточного метаболизма, митохондриальной функции и мембранного потенциала.

На рисунке 8 представлено распределение количества молекул по значениям pIC_{50} после этапов очистки и стандартизации.

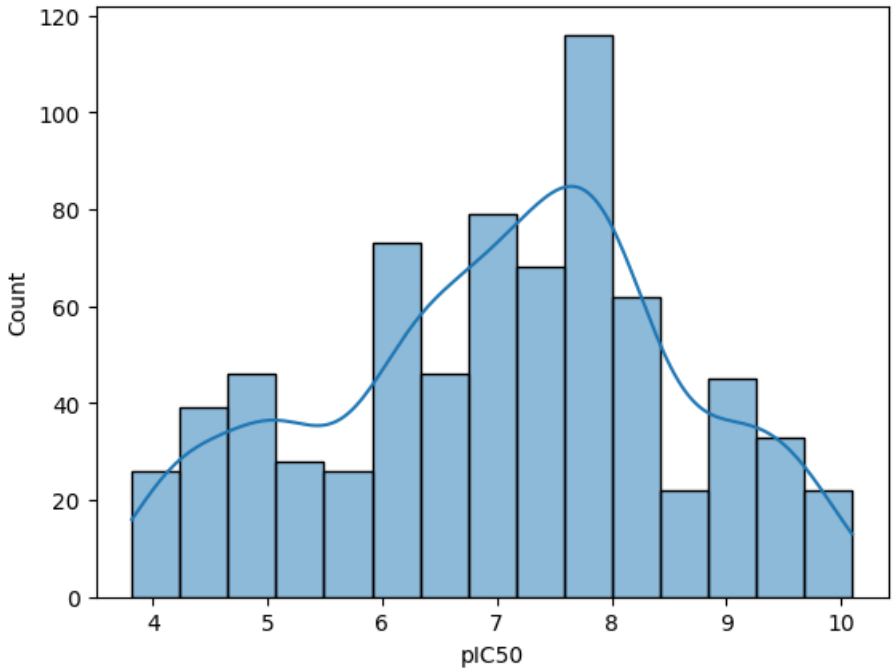


Рисунок 8 –Распределение количества молекул азолопиримидинов в выборке относительно pIC_{50} после предобработки

Дескрипторы. Для повышения информативности и устойчивости моделей в качестве входных признаков были добавлены 86 дополнительных дескрипторов. В итоге входные данные представляли собой таблицу, где каждая строка соответствовала уникальной молекуле, а столбцы набору структурных и физико-химических характеристик.

В Приложении № 5 представлена тепловая карта z-нормализованных дескрипторов. Большинство признаков демонстрируют узкий диапазон значений, что отражает структурное сходство азолопиримидинов, однако отдельные параметры (в том числе полярность, степень ароматичности, липофильность) проявляют заметную вариативность и позволяют дифференцировать соединения по физико-химическим свойствам.

Клиффы. Для анализа способности модели различать структурно близкие соединения с разной активностью были сформированы пары. Одна и та же молекула могла частвовать в разных парах одновременно. Пример клиффовой пары молекул на рисунке 9.

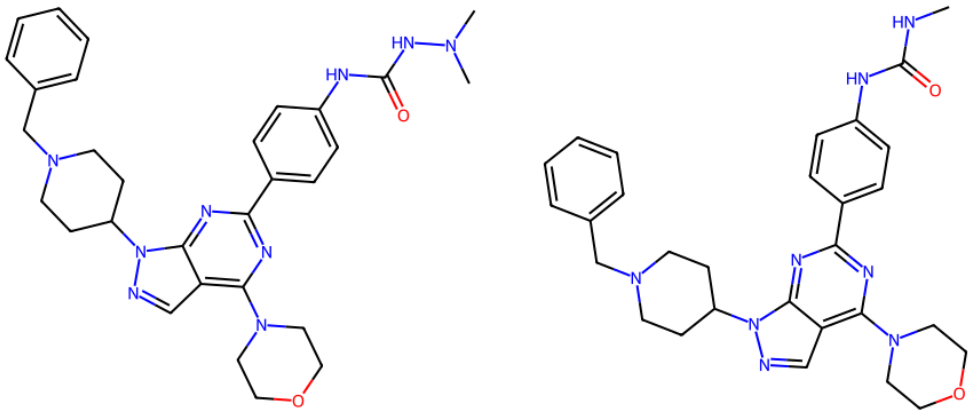


Рисунок 9 – Пример клиффовой пары молекул из выборки азолопиримидинов

В варианте, где все клифф-молекулы были направлены в тестовую выборку их кол-во составило 143 пары.

В варианте, где молекулы были и в тестовой, и в обучающей выборке, для тестирования остались только 37 пар..

Все варианты Модели № 2 обучались на основе предобученной Модели № 1. Для объективного сравнения приведены также показатели наивного предиктора, вычисленные по средним значениям обучающей выборки (табл. 14).

Таблица 12 –Средние значения обучающей выборки

Кол-во	**RMSE**	**MAE**	**R^2**	**bias**
108	1.485	1.238	-0.013	-0.171

2.3.2 Оценка различных вариантов выборок.

Базовая модель. В начальной конфигурации модель обучалась и тестировалась на случайных разбиениях, без специального выделения клифф-пар. Были получены следующие метрики (табл. 2):

Таблица 13 – Показатели Базовой Модели №2 на neighbor split

Кол-во	**RMSE**	**MAE**	**R^2**	**bias**	**slope**	**intercept**
108	0.787	0.568	0.740	0.022	0.722	1.952

Несмотря на высокие значения метрик на тестовой выборке, такая оценка могла не отражать реальную способность модели различать структурно близкие соединения с существенно различающейся активностью, поэтому в следующей итерации были построены модели, в которых клифф-пары были целенаправленно включены в состав выборок в различных форматах.

Для более строгой проверки устойчивости модели клифф-пары были целенаправленно включены в тестовую выборку. Результаты представлены в таблице 16.

Таблица 14 – Показатели Базовой Модели № 2 при клифф-парах в тесте

Кол-во	RMSE	MAE	R^2	bias	slope	intercept
108	1.063	0.808	0.480	0.114	0.518	3.574

Ухудшение метрик отражает ожидаемое снижение предсказательной точности на случаях с резкими различиями в активности.

Дообученная модель: клифф-пары в обучении и тесте. Во второй конфигурации клифф-молекулы присутствовали и в тренировочной, и в тестовой выборках. Метрики улучшились как на всем тесте, так и на клифф-подмножестве (табл. 17).

Таблица 15 – Показатели Модели № 2 при клифф-парах в обучении и тесте

Split	Кол-во	RMSE	MAE	R^2	Bias	Slope	Intercept
test	108	0.818	0.588	0.710	0.007	0.721	1.982
test_cliffs	37	0.702	0.575	0.705	0.055	0.829	1.341

На рисунке 10 показаны результаты оценки Модели № 2 при включении клифф-молекул как в обучающую, так и в тестовую выборку. Верхний ряд иллюстрирует качество предсказаний на всем тестовом наборе, нижний –на подмножестве клифф-пар.

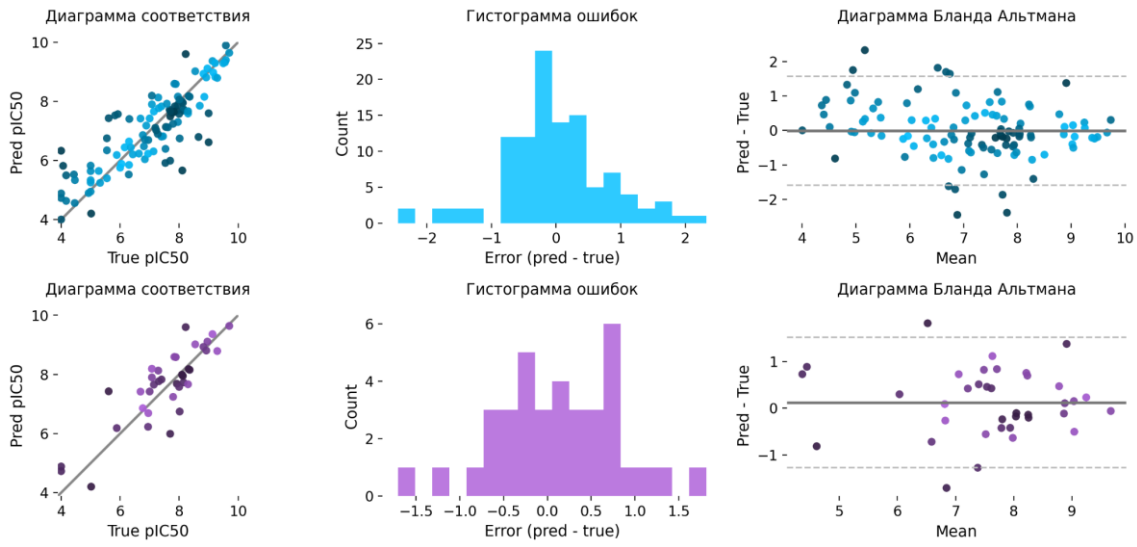


Рисунок 10 –Сравнение предсказанных и экспериментальных значений активности для Модели № 2

На общей выборке (верхний ряд) наблюдается плотное расположение точек вдоль диагонали на диаграмме соответствия и симметричное распределение ошибок около нуля, что свидетельствует о хорошей согласованности предсказаний с экспериментом. На

подмножестве клиффов (нижний ряд) разброс ошибок несколько выше, а на диаграмме Бланда–Альтмана отмечается большая вариативность при крайних значениях активности, что ожидаемо для структурно близких, но фармакологически контрастных соединений. Тем не менее модель сохраняет адекватное направление зависимости и отсутствие систематического смещения.

На рисунке 11 представлена диаграмма соответствия с калибровочными прямыми. Светло-голубыми точками показаны все молекулы тестовой выборки, фиолетовыми – подмножество клиффов. Линии регрессии рассчитаны отдельно для полного тестового набора и для подмножества клиффов, что позволяет оценить различия в калибровке модели на обычных и сложных для предсказания случаях.

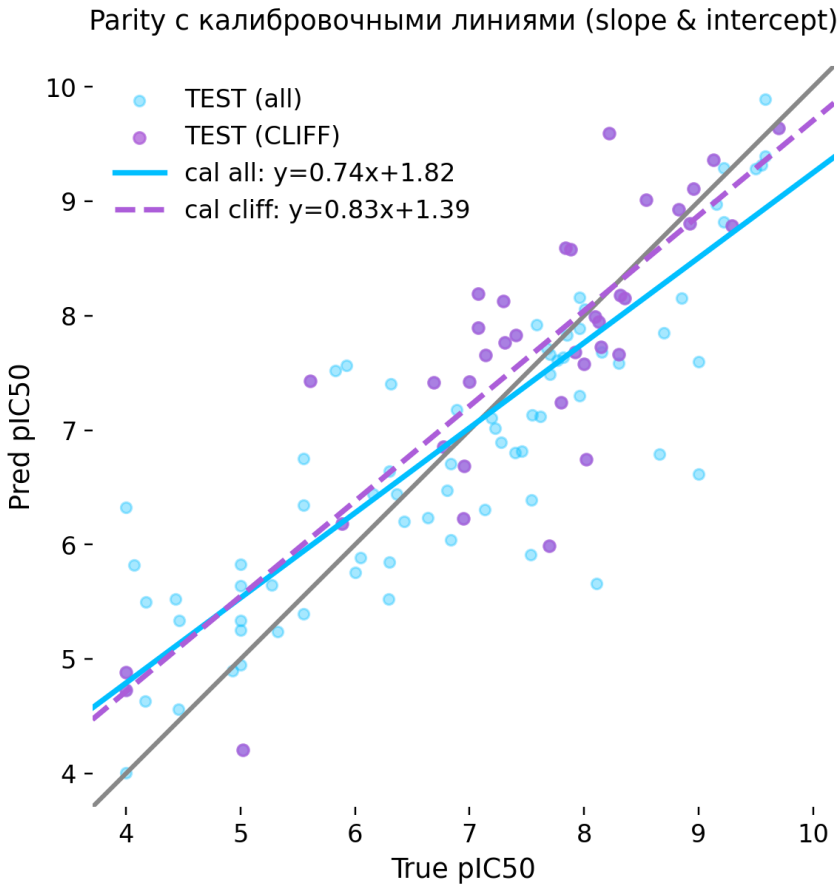


Рисунок 11 – Сравнение предсказанных и экспериментальных значений активности для Модели № 2

Наклоны и смещения калибровочных линий (соответственно 0.74 и 1.82 для всей выборки; 0.83 и 1.39 для клиффов) показывают, что модель сохраняет адекватную линейную зависимость между предсказанными и экспериментальными значениями pIC_{50}. При этом линия для клиффов располагается ближе к идеальной диагонали, что указывает на более точное согласование предсказаний в диапазоне средних и высоких активностей.

Незначительное отклонение общей линии вниз по оси y отражает тенденцию к занижению предсказаний для наиболее активных соединений, характерную для регрессионных моделей при обучении на структурно близких молекулах.

2.3.3 Сравнение всех моделей

Краткое сравнение на всех тестовых данных и на клифф-парах (табл.):

Таблица 16 – Сводное сравнение вариантов Модели № 2

Модель	Данные	Кол-во	RMSE	MAE	R^2	bias	slope	intercept
Наивный предиктор	все соединения	108	1.485	1.238	-0.013	-0.171	-	-
Базовая модель	все соединения	108	0.787	0.568	0.740	0.022	0.722	1.952
	только клиффы	37	1.063	0.808	0.480	0.114	0.518	3.574
Дообученая модель	все соединения	108	0.818	0.588	0.710	0.007	0.721	1.982
	только клиффы	37	0.702	0.575	0.705	0.055	0.829	1.341

В результате сравнительного анализа наиболее сбалансированной оказалась конфигурация, где клифф-молекулы присутствуют в обучении и тесте: она сохраняет высокую точность на всей выборке (R^2 = 0.71, MAE = 0.588) и при этом показывает наилучшее качество на клифф-подмножестве (R^2 = 0.70, MAE = 0.575).

Эта версия Модели № 2 принята в качестве основной для дальнейшего анализа.

2.4 Оценка переносимости

2.4.1 Проверка корреляции на лабораторном наборе данных

Экспериментальные данные включали разные клеточные линии и индивидуальные наборы соединений. Диапазон pIC_{50} варьирует всего от 3 до 5, что ограничивает статистическую мощность.

К структурам SMILES были рассчитаны молекулярные десрпиторы и применена Модель №2. Распределение аффинности также показало небольшую вариативность от 5.2 до 6.6, по сравнению с диапазоном модели, что может указывать на корректное отображение признаков (см рис. 12).

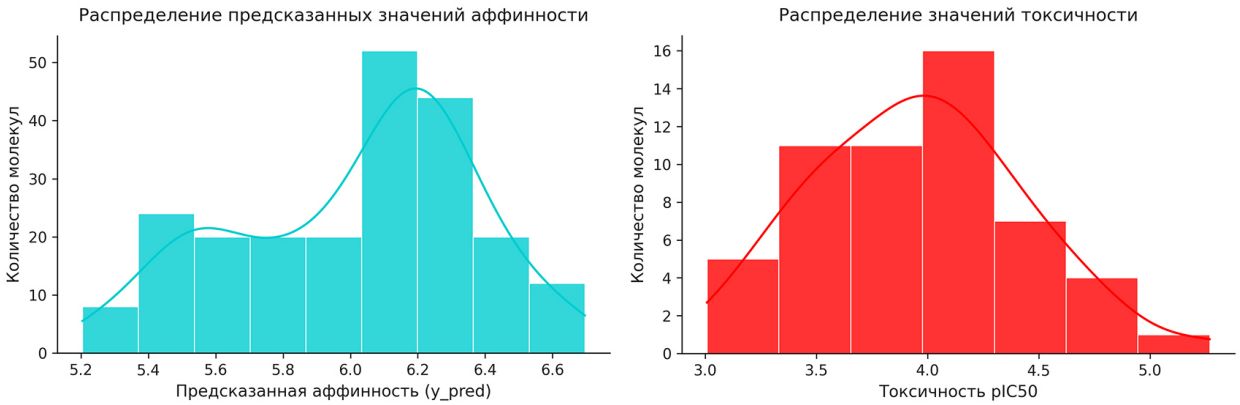


Рисунок 12 – Распределение значений аффиности и токсичности в лабораторном наборе данных

Полученые значения предполагаеммой аффиности были включены в общую таблицу с данными (см. табл. 17)

Таблица 17 – Аггрегированая таблица значений аффиности и токсичности в лабораторном наборе данных

smiles	pIC_{50}	cell_line	y_pred
CN(C)c1ccc(C2C(C#N)=CNc3nnnn32)cc1	3.23825	HepG2	6.048781
CSc1nc2n(n1)C(c1ccc(N(C)C)cc1)C(C#N)=CN2	3.89279	HepG2	6.463467
[CH2]c1nc2n(n1)C(c1ccc(N(C)C)cc1)C(C#N)=CN2	3.68447	HepG2	6.299982

Для количественной оценки согласованности между предсказанными моделью значениями аффинности (y_pred) и экспериментальной токсичностью (pIC_{50}) был выполнен анализ остатков относительно линии регрессии. По результатам распределения соединений (см. табл. 18) получено:

Таблица 18 – Распределение корреляции значений аффиности/токсичности

Категория	Количество (n)
Средняя	154
Высокая корреляция	36
Низкая корреляция	30

Поскольку предсказанные значения аффинности касаются исключительно химической структуры молекулы, а экспериментальные значения pIC_{50} отражают реакцию в различных клеточных линиях, прямая групповая корреляция между y_pred и pIC_{50} была признана методически ограниченной. Поэтому нами была составлена тепловая карта диапазонов с корреляцией токсисичности/аффинности (рисунок 13)

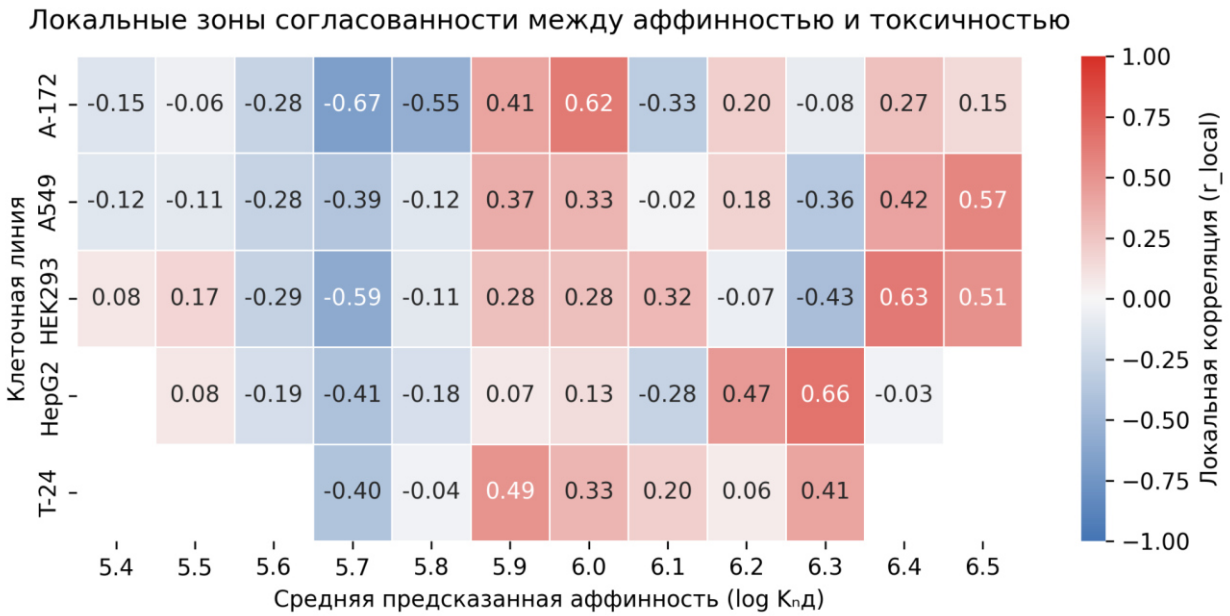


Рисунок 13 – Тепловая карта диапазонов с корреляцией токсисичности/аффинности

На тепловой карте представлены значения локальной корреляции между предсказанной аффинностью и экспериментальной токсичностью для различных клеточных линий в диапазоне средней аффинности 5.4–6.5. Видно, что выраженные положительные зависимости наблюдаются лишь в отдельных диапазонах и преимущественно при средних значениях аффинности около 5.9–6.4 и для линий A-172, HepG2 и T-24. На остальных участках корреляция колеблется вокруг нуля или отрицательна. Это указывает на то, что взаимосвязь между связыванием и токсичностью носит локальный характер, она выражена только для отдельных групп соединений. Можно предположить, что исследуемые лабораторные молекулы воздействуют не только через связывание с токсикологическими белками, но и через дополнительные клеточные механизмы.

Таким образом, высокая локальная корреляция в отдельных диапазонах подтверждает наличие областей, где предсказанная аффинность может быть связана с цитотоксичностью, однако разнородность клеточных ответов указывает на необходимость включения дополнительных признаков в обучение для более полной валидации модели.

2.4.2 Интерпретация признаков

Поскольку графовые нейросетевые модели не поддаются прямой интерпретации, для анализа использовались молекулярные дескрипторы, рассчитанные для соединений с различными уровнями предсказанной аффинности. Это позволило предположить, какие типы молекул модель считает более аффинными, и сопоставить эти выводы с известными

химическими закономерностями, а также оценить возможную связь таких свойств с токсичностью.

Для выявления структурных факторов, определяющих аффинность по мнению модели, были рассчитаны средние значения дескрипторов в трёх подгруппах молекул: с низкой, средней и высокой предсказанной аффинностью. Разница между крайними группами (высокой и низкой) позволила количественно оценить вклад каждого признака в предсказания модели.

Наибольшие положительные различия ($\Delta > 10$) наблюдались для дескрипторов, связанных с молекулярной массой, поверхностными и электронными характеристиками (таблица 19).

Таблица 19 – Дельта различия характеристик высокоаффинных соединений

Дескриптор	**Δ**	**Интерпретация**
desc_MolWt (молекулярная масса)	+98.6	более тяжёлые соединения, увеличенная объёмность и липофильная поверхность
desc_TPSA (полярная площадь поверхности)	+30.6	наличие полярных гетероатомов и донорно-акцепторных групп
desc_SlogP_VSA2, desc_SlogP_VSA5–10	+24.5 … +5.8	расширение гидрофобных площадей, более выраженная липофильность
desc_SMR_VSA1,6,10	+7–17	возрастание поляризуемости и распределённых π-систем
desc_EState_VSA2–4	+6–13	рост локальных зарядовых различий, характерных для акцепторных центров
desc_PEOE_VSA1,7–9	+9–10	наличие выраженных частично отрицательных областей – участков потенциального связывания
desc_HeavyAtoms, scaf_len	+7–8	усложнение скелета и увеличение длины каркаса молекулы

Таким образом, высокоаффинные структуры по мнению модели характеризуются повышенной массой, липофильностью и поляризуемостью, что соответствует комплексным ароматическим или гетероциклическим системам, способным к множественным неспецифическим взаимодействиям с белками.

Для дополнительной проверки химической согласованности Модели № 2 молекулы тестовой выборки были разделены на три подгруппы –высокоаффинные (верхний 10-й перцентиль), низкоаффинные (нижний 10-й перцентиль) и промежуточные.

Анализ показал, что модель чаще относит к аффинным структуры, содержащие амидные, анилиновые фрагменты и третичные амины, что совпадает с её предпочтением более функционально насыщенных и полярных соединений.

Выявленные характеристики показывают, что модель последовательно ассоциирует высокие значения предсказанной аффинности, которые в данном контексте интерпретируются как потенциально повышенная токсичность, с более тяжёлыми, поляризуемыми и липофильными структурами. Напротив, низкоаффинные по мнению модели соединения отличаются компактностью, умеренной полярностью и меньшей функциональной насыщенностью, что согласуется с представлением о более безопасных и метаболически стабильных молекулах.

Эти связи являются модельными ассоциациями и их следует рассматривать как гипотезы, требующие дополнительной статистической проверки и экспериментальной валидации.

2.5 Обобщение и ограничения

Разработанная Модель № 2 продемонстрировала способность воспроизводить закономерности связывания в классе азолопиримидинов, однако её применимость ограничена рядом факторов, связанных как с исходными данными, так и с архитектурой модели.

1. Ограниченность исходных данных.

Обучающая выборка была составлена на основе экспериментальных данных из базы BindingDB и включала преимущественно соединения, исследованные в узком диапазоне мишеней. Это определяет смещение распределения по химическим пространствам и может ограничивать способность модели к обобщению за пределами этого класса структур.

2. Связь аффинности и токсичности носит косвенный характер.

Поскольку модель обучалась на данных по аффинности, её предсказания отражают не прямую цитотоксичность, а скорее склонность соединений к сильному (в том числе неспецифическому) связыванию с белками. Соответственно, высокая предсказанная аффинность может служить только косвенным признаком потенциальной токсичности и требует экспериментального подтверждения.

3. Графовая архитектура ограничивает интерпретацию.

Модель Chemprop формирует скрытые представления на основе топологии молекулы, но не учитывает динамику взаимодействия или контекст клеточной среды. Поэтому результаты следует рассматривать как статистическую аппроксимацию химического пространства, а не как прямое моделирование механизмов токсичности.

В совокупности эти факторы определяют границы применимости модели. Она способна выявлять внутренние закономерности в рамках исследованного класса соединений, но её предсказания следует рассматривать как гипотетические оценки

вероятности токсического взаимодействия, требующие дополнительной экспериментальной проверки.

3 ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

В данной главе представлена последовательность этапов построения моделей, предназначенных для прогнозирования токсичности соединений по их молекулярной структуре. Целью эксперимента являлось создание графовой модели, способной предсказывать аффиность азолопиримидинов на основе накопленных данных о молекулах с известной цитотоксичностью.

Исследование проводилось поэтапно - от построения фундаментальной модели токсичности на большом корпусе данных до её дообучения на наборе азолопиримидинов. Часть исследования была отведена для проверки способности модели определять тонкие различия в структуре.

В рамках экспериментальной части исследования были реализованы несколько вариантов постановки задачи, различающиеся схемами разделения данных на обучающую и тестовую выборки. Такой дизайн эксперимента позволяет сравнить предсказательную способность моделей при различных стратегиях разделения, отражающих как случайное распределение соединений, так и более жёсткие условия проверки генерализации на новые химические структуры.

3.1 Наборы данных и целевые показатели

В ходе анализа было протестировано несколько открытых баз данных, включая TOXRIC, Tox21, ChEMBL и TDC ADMET. Однако для класса азолопиримидинов количество доступных записей оказалось крайне ограниченным. В связи с этим прямое построение модели токсичности оказалось методологически затруднительным, и была выбрана альтернативная стратегия: моделирование молекулярной аффинности к токсикологически значимым белковым мишеням.

3.1.1 Обучающий набор данных

В качестве обучающей базы данных была выбрана BindingDB, так как содержала наибольший набор качественных данных. Ключевым преимуществом BindingDB является высокий уровень стандартизации. База данных как источник количественных данных об аффинности к токсико-релевантным белкам, что позволило построить модель токсического потенциала азолопиримидинов через фундаментальные закономерности структуры и активности.

С помощью нее мы сформировали выборку для обучения фундаментальной регрессионной модели общей химии, а также выделили из этой базы подкорпус азолопиримидинов для профильного анализа.

3.1.2 Валидационный набор данных

В качестве валидационного набора использовались данные из Лаборатории первичного биоскрининга, клеточных и генных технологий о 55 протестированных соединений азолопиримидинового класса со значениями токсичности IC_{50}.

3.2 Предобработка данных

Предварительная очистка включала удаление записей с нулевыми либо отрицательными значениями активности. Далее значения IC_{50} перевели в pIC_{50} согласно общепринятой практике. Для молекул с несколькими измерениями pIC_{50} использовалась медиана, что снижает влияние выбросов и вариабельности экспериментов.

3.2.1 Стандартизация и очистка молекул

Перед расчётом признаков и обучением моделей исходные структуры SMILES приводились к единообразному виду. Процедура реализована на базе RDKit и включала следующие шаги:

Удаление солей и малых нецелевых компонентов. Для каждой записи выполнялось автоматическое удаление нековалентных ионных компонентов с помощью RDKit.Chem.SaltRemover.

Выбор основного ковалентного фрагмента. В случае многокомпонентных записей выбирался крупнейший фрагмент по числу тяжёлых атомов. Этот критерий позволяет выделять фармакофорный основной компонент и отбрасывать малозначимые примеси.

Удаление явных атомов водорода. После очистки структуры удалялись явные H атомы, что приводило SMILES к компактной записи и предотвращало дублирование за счёт разных вариантов гидрогенизации.

Канонизация SMILES. Итоговая структура переводилась в канонический SMILES для устранения разных записей одной и той же молекулы. Если молекулы не могли пройти этот фильтр, их приходилось отбрасывать для чистоты базы обучения.

3.2.2 Фильтрация данных по токсикологически значимым белковым мишеням

Нами была выполнена фильтрация по UniProt-белкам, вовлечённым в механизмы метаболизма ксенобиотиков, апоптоза и клеточного стресса, что позволило ограничить анализ взаимодействиями, потенциально влияющими на токсичность соединений. Из исходных данных BindingDB были отобраны взаимодействия с токсикологически значимыми белками, идентифицированными по UniProt. Панель включала ферменты

метаболизма, транспортеры, ядерные рецепторы и мишени, связанные с прочей токсичностью.

Таблица 6 – Классы белков с потенциальной токсичностью взаимодействия

Класс белков	Примеры	Основной токсикологический риск
Цитохромы P450	CYP1A1–CYP3A7	Реактивные метаболиты, DDI, гепатотоксичность
Фаза II ферменты	UGT, SULT, GST, NAT	Нарушение детоксикации
Транспортёры	ABC, SLC-семейства	Холестаз, нефротоксичность
Ионные каналы	KCN, SCN, CACNA	Кардиотоксичность
Нейро- и рецепторные мишени	HTR, ADRA, DRD, CHRM	ЦНС-эффекты, серотониновый синдром
Митохондриальные ферменты	CPT1A, ACADVL, SLC25	Энергетический стресс

3.3 Модель №1. Фундаментальная модель

Фундаментальная модель подразумевала в себе обучение на графовых структурах без дополнительных признаков на большом объеме данных.

3.3.1 Разбиение данных

Для проверки способности модели обобщать закономерности на новые химические структуры были реализованы две стратегии разбиения данных.

Первая, random split, основана на случайном равномерном распределении соединений между выборками и используется как стандартный подход для оценки качества модели. Однако при таком разбиении структурно близкие молекулы могут одновременно присутствовать и в обучающем, и в тестовом наборах, что приводит к завышению метрик.

Вторая стратегия, scaffold split, исключает этот эффект, распределяя соединения по уникальным химическим каркасам. Такой подход формирует более строгие условия для проверки способности модели обобщать знания на новые химические скелеты. При планировании эксперимента рассматривалось, что для модели, обученной на широком химическом пространстве, scaffold-разделение может оказаться чрезмерно жёстким.

3.3.2 Параметры обучения Модели №1

Обучение графовой нейронной сети Chemprop для задачи регрессии по pIC_{50} проводилось на очищенном датасете, с использованием следующих параметров:

Таблица 7 – Параметры обучения графовой Модели №1

Категория	**Характеристика**
Тип модели	Графовая нейросеть Chemprop
Структура сети	3 слоя по 400 нейронов
Тип задачи	Регрессия по целевой переменной pIC_{50}
Ансамбль	3 модели, dropout = 0.2
Размер мини-батча	32 (для моделей с весами –16)
Оптимизатор	Adam
Функция потерь	Среднеквадратичная ошибка (MSE)
Тип вычислений	Обучение на GPU
Молекулярное представление	Графовая структура SMILES
Дополнительные признаки	-

Размеры разбиения составили 80 % train и по 10% для test и val.

3.4 Модель №2. Азолопиримидины

Для построения специализированной модели на классе азолопиримидинов из очищенного датасета был выделен соответствующий набор соединений с помощью специально реализованной функции структурного фильтра. После отбора проведена дополнительная обработка данных, включающая расчет расширенного набора молекулярных дескрипторов.

3.4.1 Генерация SMARTS-паттерна

Для выделения из общей базы данных соединений, относящихся к классу азолопиримидинов, была реализован автоматизированный структурный фильтр на основе SMARTS-шаблонов.

Структура должна было удовлетворять двум необходимым условиям:

1. Наличие пиримидинового (1,3-диазинового) шестичленного ядра (ровно два атома азота в кольце, расположенные в позициях 1 и 3 относительно циклического обхода)
2. Конденсация с пятичленным азольным кольцом по общему ребру,представляет собой пятичленный гетероцикл, включающий не менее двух гетероатомов, среди которых обязательно присутствует атом азота (имидазол, пиразол, оксазол, тиазол и др.).

Допускались последующие замещения и модификации вне каркасного скелета. Для этого использовалась 2 ступенчатая генерация:

1. Каркасная генерация: конструирование минимального скелета «пиримидин + азол» посредством формальных правил конденсации колец без заранее заданных структурных шаблонов.
2. Замещённая генерация: добавление заместителей (R-групп) к допустимым вершинам каркаса с соблюдением валентностей, правил ароматичности и фармакофорных ограничений.

Каждый сгенерированный кандидат прошел валидацию ароматичности, структурную фильтрацию, дедупликацию и приоритизацию по набору медико-химических критериев (синтетическая доступность и т. д.).

Реализация выполнена в RDKit [Приложение №1]:

— Проверка топологии пиримидина выполняется на уровне локального кольцевого обхода (две позиции N разделены одним атомом по циклу). Эта проверка независима от SMILES-нумерации и устойчиво переносится между изомерами

— Контроль фьюжна по ребру. В момент слияния колец проверяется, что два общих узла смежны в обоих циклах (общая грань, а не два случайных общих атома), и что результирующая степень узлов не превышает валентность.

— Репродуцируемость. Все шаги (позиции гетероатомов в азоле, выбор точек замещения, выбор фрагментов) выполняются в случайном порядке; промежуточные наборы сохраняются в список.

Генерированые изображения SMARTS-паттернов в Приложении №2

3.4.2 Функция фильтрации

Все молекулы, удовлетворяющие SMARTS-критерию, автоматически отбирались в новый поднабор. Для этого была составлена функция используемая для фильтрации баз данных и генеративной модели [Приложение №3]. В результате отбора по SMARTS-фильтру была получена выборка молекул, содержащих конденсированную систему «азол + пиримидин».

3.4.3 Вычисление молекулярных дескрипторов

В рамках подготовки данных для модели 2 были рассчитаны молекулярные дескрипторы, включающие стандартный набор двухмерных физико-химических параметров RDKit, а также структурные фрагменты и характеристики каркасной структуры молекулы.

Базовые физико-химические дескрипторы. Для каждой молекулы вычислялись следующие показатели: молекулярная масса, коэффициент липофильности, полярная поверхность, число доноров и акцепторов водородных связей, число вращательных связей, доля sp^3-гибридизованных атомов углерода, общее число колец и число ароматических колец, количество стереоцентров, число тяжёлых атомов, формальный заряд, количество амидных связей, а также содержание атомов N, O, S и галогенов (F, Cl, Br, I).

Расширенные RDKit-дескрипторы. Дополнительно применялись предопределённые наборы безопасных дескрипторов. Данные показатели основаны на распределении поверхностей Ван-дер-Ваальса, электростатическом заряде и топологических характеристиках, что позволяет более полно описывать электронную структуру и трёхмерную организацию молекулы в двумерном приближении.

Структурные фрагменты. Для выделения функциональных групп был использован набор SMARTS-шаблонов, позволивший количественно оценить содержание типичных фрагментов: третичных аминов, четвертичных аммониевых солей, амидов, мочевин, сульфонамидов, простых эфиров, трифторметильных групп, нитрилов, алкинов и нитрогрупп. Для каждого соединения подсчитывалось количество вхождений соответствующего подструктурного мотива.

Каркасная структура. Дополнительно для каждой молекулы определялся каркас по Мёрко, и в качестве дескриптора использовалась длина SMILES-представления данного каркаса, что позволяет косвенно учитывать архитектурную сложность скелета.

3.4.4 Разбиение данных

Для анализа активности азолопиримидинов и последующей диагностики резких различий активности у структурно близких молекул был выбран метод разбиения по соседям (neighbor split), для более строгой оценки обобщающей способности.

Степень химического сходства оценивалась по ECFP4 с радиусом 2 и с использованием коэффициента Танимото. То есть, определяемый фрагмент это атом и все, что его окружает на расстоянии двух химических связей. Пары молекул с Tanimoto ≥ 0,90 считались соседями и не допускались к попаданию в разные подвыборки.

Особый интерес представляла проверка модели на наиболее сложных для прогнозирования случаях –парах структурно сходных молекул с существенно различающейся активностью. Для

выявления и анализа была реализована отдельная обработка клифф-пар, в которой мы учитывали не только структурное сходство, но и значение pIC_{50}. В качестве клиффа отбиралась только та пара, которая одновременно удовлетворяла двум критериям: высокая структурная схожесть (Tanimoto ≥ 0,7) и разница биологической активности ($\Delta pIC_{50} \geq 0,7$). Примеры выделенных клиффов в Приложении №4.

Все эксперименты реализованы с едиными настройками разбиения neighbor split, 15% тест, 10% валидация и идентичной архитектурой Chemprop. Отличие моделей заключалось лишь в стратегии обращения с клиффами, а также в порядке использования предобученной фундаментальной модели.

3.4.5 Параметры обучения

Модель №2 обучалась на фундаменте в виде Модели №1. Важным отличием было добавление дескрипторов с описательными характеристиками. Комбинации графовых признаков и дополнительных дескрипторов обеспечивала более полное представление молекулы и повышала чувствительность модели к мелким вариациям заместителей в пределах фиксированного ядра.

Общая архитектура вариантов Модели №2 представлена в таблице:

Таблица 8 – Параметры обучения графовой Модели №2

Категория	**Характеристика**
Тип модели	Графовая нейросеть Chemprop
Структура сети	4 скрытых слоя по 800 нейронов
Тип задачи	Регрессия по целевой переменной pIC_{50}
Ансамбль	3 модели, dropout = 0.3
Размер мини-батча	32 (для моделей с весами –16)
Оптимизатор	Adam
Параметр early stopping	Patience = 15
Функция потерь	Среднеквадратичная ошибка (MSE)
Тип вычислений	Обучение на GPU
Молекулярное представление	Графовая структура SMILES
Дополнительные признаки	2D-дескрипторы и физико-химические характеристики, рассчитанные с помощью RDKit
Базовая модель	Предобученная на данных BindingDB

3.5 Валидация модели

Мы сравнивали предсказанные моделью значения с помощью графического анализа (диаграмма соответствия, диаграмма Бланда-Альтмана и др.) и расчета стандартных метрик качества регрессии (MAE, RMSE, R^2, bias, slope, intercept).

Также проводился анализ остатков и выявлялись случаи систематического смещения предсказаний, что позволило более детально охарактеризовать области успеха и ограничения модели.

3.4.1 Оценка переносимости

Для независимой проверки переносимости модели использовались лабораторные данные по 55 соединениям, для которых была экспериментально определена цитотоксичность на ряде клеточных линий. Для каждой молекулы вычислялся интегральный показатель токсической аффинности (агрегированный pIC_{50} по отобранным UniProt-мишеням). Полученные значения сопоставлялись с экспериментальными показателями жизнеспособности клеток.

Для количественной оценки согласованности между предсказанными моделью значениями аффинности (y_pred) и экспериментальной токсичностью (pIC_{50}) был выполнен анализ остатков относительно линии регрессии. Линия регрессии описывалась выражением

$$\hat{y} = \underline{y} + r * \frac{s_y}{s_x} (x - \underline{x})$$

где x –предсказанная аффинность (y_pred),

y –экспериментальные значения pIC_{50},

r –коэффициент корреляции Пирсона,

s_x и s_y –стандартные отклонения.

Для каждой точки вычислялось отклонение от этой линии (residual = y − ŷ). Отклонения с модулем меньше одного стандартного отклонения рассматривались как согласованные с моделью, в то время как большие отрицательные и положительные значения относили к зонам «высокой» и «низкой» корреляции соответственно.

3.4.2 Косвенное определение признаков в модели

Для получения представления о химической интерпретируемости модели был проведён анализ вкладов молекулярных признаков. Мы сопоставили структурные мотивы и дескрипторы с тем, как их оценивает модель, а не с истинными значениями (для интерпретации именно модели).

1. Определили группы пороги высокой и низкой токсичности

2. С помощью SMARTS-паттерна функциональных групп, определили какие наиболее часто встречаются в обеих группах
3. Определили величину вклада молекулярных дескрипторов
4. Проверили изменения заместителей. Отобрали молекулы с одинаковым каркасов и вычислили пары с наибольшей разницей в токсичности, чтобы определить, какие замены повышают/занижают предсказанную активность

Это позволило идентифицировать, какие структурные элементы и дескрипторы модель считает ключевыми для высокой и низкой токсичности. Анализ этих случаев позволил критически оценить, насколько предсказания модели согласуются с химико-биологической логикой и реальными экспериментальными наблюдениями.

4 БЛОК-СХЕМА ЭКСПЕРИМЕНТА

Логика реализации проекта представлена в нотации BPMN [49], поскольку она позволяет формализовать последовательность этапов обработки данных, обучения модели и получения предсказаний. BPMN используется здесь как универсальная технологическая схема, описывающая поток вычислений, а не архитектуру программного кода.

Первая схема (см. Приложение 6) описывает полный процесс предобработки данных: выбор и очистку исходной базы, стандартизацию SMILES, генерацию и выделение азолопиримидинового каркаса, а также фильтрацию соединений по токсико-релевантным белкам на основе UniProt ID.

Вторая схема отражает этап моделирования, включающий обучение фундаментальной и специализированной моделей, оценку чувствительности к активити-клиффам и последующую валидацию предсказаний на независимом лабораторном наборе.

ЗАКЛЮЧЕНИЕ

Разработан и обоснован графовый метод количественного прогнозирования токсичности малых молекул, в котором pIC_{50} по токсико-релевантным UniProt-мишеням из BindingDB используется как прокси-показатель токсического потенциала. Использование данных об аффинности в качестве основы для моделирования токсичности позволяет учитывать биологические механизмы действия и компенсировать дефицит прямых экспериментальных меток.

Построенная модель показала характеристики, сравнимые с публикациями по графовым моделям: $R^2 = 0,71$; MAE = 0,59. Модель также стабильна на соединениях с тонкими структурными различиями, что указывает на устойчивость к локальным вариациям внутри фиксированного каркаса.

Интерпретационный анализ дескрипторов показал, что модель ассоциирует повышенную аффинность с более тяжёлыми, поляризуемыми и липофильными структурами и с присутствием основных функциональных групп, тогда как низкие значения связаны с компактностью и умеренной полярностью. Эти выводы согласуются с известными тенденциями неспецифического связывания.

Проверка переносимости на независимом лабораторном наборе показала локальные области положительной связи между предсказанной токсикоаффинностью и экспериментальной цитотоксичностью, при общей высокой вариативности ответов. Это подтверждает частичную валидируемость гипотезы и указывает на необходимость расширения признаков модели.

Практическая значимость:

1. Разработанный подход может быть использован в задачах генеративного моделирования соединений с заданными свойствами и оптимизации структуры лекарственных кандидатов.
2. Методика основана на модульной структуре, где функции выбора белков-мишеней и фильтра химических каркасов реализованы отдельно. Это позволяет применять модель к узким панелям мишеней одного типа, а также адаптировать её под другие классы соединений.

СПИСОК ИСТОЧНИКОВ

1. Аткинсон А. Дж. [и др.] Принципы клинической фармакологии [Электронный ресурс] / науч. ред. и пер. с англ. Г. Т. Сухих и др // М.: Практическая медицина, 2022. URL: https://www.trauma-books.ru/principi_klin_farm_sod_i_prim_str.pdf (дата обращения: 29.10.2025).
2. Деривативное электронное издание на основе печатного аналога: Фармацевтическая химия : учебник / под ред. Г. В. Раменской. // М. : БИНОМ. Лаборатория знаний, 2015. URL: http://lib.ysu.am/disciplines_bk/83a96e12f01b59c50ef6f56527e8df72.pdf (дата обращения: 29.10.2025).
3. Sertkaya A. и др. Costs of Drug Development and Research and Development Intensity in the US, 2000-2018 // JAMA Netw Open. American Medical Association, 2024. Т. 7, № 6. doi: 10.1001/jamanetworkopen.2024.15445
4. Schuhmacher A. и др. Analysis of pharma R&D productivity – a new perspective needed // Drug Discovery Today. 2023. Т. 28. № 10. doi: 10.1016/j.drudis.2023.103726
5. Waring M. J. и др. An analysis of the attrition of drug candidates from four major pharmaceutical companies // Nat Rev Drug Discov. 2015. Т. 14. № 7. doi: 10.1038/nrd4609
6. Schedler D. J. A. Drug Discovery: A History (Sneader, Walter) // J. Chem. Educ. 2006. Т. 83. № 2. С. 215. doi: 10.1021/ed083p215.1
7. Головко Ю.С., Ивашкевич О.А., Головко А.С. Современные методы поиска новых лекарственных средств [Электронный ресурс] // Вестник БГУ. Серия 2, Химия. Биология. География. 2012. № 1. С. 7-15. URL: https://elib.bsu.by/handle/123456789/36095 (дата обращения: 29.10.2025)
8. Хушпульян Д.М., Гайсина И.Н., Никулин С.В., Чубарь Т.А., Савин С.С., Газарян И.Г., Тишков В.И. Высокопроизводительный скрининг при поиске лекарств: проблемы и решения [Электронный ресурс] // Вестник Моск. ун-та. Сер. 2. Химия. 2024. Т. 65. № 2. С. 96-112. URL: https://www.chem.msu.ru/rus/vmgu/242/96.pdf (дата обращения: 29.10.2025).
9. Садыбеков А.В., Катритч В. Computational approaches streamlining drug discovery [Электронный ресурс] // Nature. 2023. Vol. 616, No. 7958. DOI: 10.1038/s41586-023-05905-z. URL: https://www.nature.com/articles/s41586-023-05905-z (дата обращения: 29.10.2025).
10. Bender B. J. и др. A practical guide to large-scale docking // Nat Protoc. 2021. Т. 16. № 10. doi: 10.1038/s41596-021-00597-z.

11. Jaganathan K., Tayara H., Chong K. T. Prediction of Drug-Induced Liver Toxicity Using SVM and Optimal Descriptor Sets // IJMS. 2021. Т. 22. № 15. doi: 10.3390/ijms22158073
12. Ahn S., Lee S. E., Kim M.-H. Random-forest model for drug–target interaction prediction via Kullback–Leibler divergence [Электронный ресурс] // Journal of Cheminformatics. 2022. Vol. 14, Article 67. URL: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-022-00644-1 (дата обращения: 29.10.2025).
13. Venkataraman M. и др. Leveraging machine learning models in evaluating ADMET properties for drug discovery and development // ADMET DMPK. 2026. Т. 13. № 3. doi: 10.5599/admet.2772
14. De Carlo A. и др. Predicting ADMET Properties from Molecule SMILE: A Bottom-Up Approach Using Attention-Based Graph Neural Networks // Pharmaceutics. 2024. Т. 16. № 6. doi: 10.3390/pharmaceutics16060776
15. Yao R. и др. Knowledge mapping of graph neural networks for drug discovery: a bibliometric and visualized analysis // Front. Pharmacol. 2024. Т. 15. doi: 10.3389/fphar.2024.1393415
16. Cherkasov A. и др. QSAR Modeling: Where Have You Been? Where Are You Going To? // J. Med. Chem. 2014. Т. 57. № 12. doi: 10.1021/jm4004285
17. Guo Z. и др. Graph-based Molecular Representation Learning // arXiv 2022. doi: 10.48550/ARXIV.2207.04869
18. Shahin R., Jaafreh S., Azzam Y. Tracking protein kinase targeting advances: integrating QSAR into machine learning for kinase-targeted drug discovery // Future Science OA. 2025. Т. 11. № 1. doi: 10.1080/20565623.2025.2483631
19. Berry K., Cheng L. A Survey of Graph Neural Networks for Drug Discovery: Recent Developments and Challenges [Электронный ресурс] // arXiv. 2025. URL: https://arxiv.org/html/2509.07887v1 (дата обращения: 29.10.2025).
20. Wu Z. и др. MoleculeNet: a benchmark for molecular machine learning // Chem. Sci. 2018. Т. 9. № 2. doi: 10.1039/C7SC02664A
21. Stokes J. M. и др. A Deep Learning Approach to Antibiotic Discovery // Cell. 2020. Т. 180. № 4. doi: 10.1016/j.cell.2020.01.021
22. Drug Discovery Today [Электронный ресурс] // Elsevier BV. –URL: https://scispace.com/journals/drug-discovery-today-3g2e0xk4/ (дата обращения: 29.10.2025).

23. Yao R. и др. Knowledge mapping of graph neural networks for drug discovery: a bibliometric and visualized analysis // Front. Pharmacol. 2024. Т. 15. doi: 10.3389/fphar.2024.1393415
24. Zhou J. и др. Graph neural networks: A review of methods and applications // AI Open. 2020. Т. 1. doi: 10.1016/j.aiopen.2021.01.001
25. Wu Z., Wang J., Du H., Jiang D., Kang Y., Li D., Pan P., Deng Y., Cao D., Hsieh C-Y., Hou T. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking [Электронный ресурс] // Nature Communications. 2023. Vol. 14. URL: https://www.nature.com/articles/s41467-023-38192-3 (дата обращения: 29.10.2025).
26. Saxena R. R., Saxena R. Applying Graph Neural Networks in Pharmacology // Techrxiv. 2024. doi: 10.36227/techrxiv.170906927.71541956/v1
27. Yang K. и др. Analyzing Learned Molecular Representations for Property Prediction // J. Chem. Inf. Model. 2019. Т. 59. № 8. doi: 10.1021/acs.jcim.9b00237
28. Heid E. и др. Chemprop: Machine Learning Package for Chemical Property Prediction // 2023. doi: 10.26434/chemrxiv-2023-3zcfl
29. Wang Z. и др. Advanced graph and sequence neural networks for molecular property prediction and drug discovery // Bioinformatics. 2022. Т. 38. № 9. doi: 10.1093/bioinformatics/btac112
30. Venkataraman M. и др. Leveraging machine learning models in evaluating ADMET properties for drug discovery and development // ADMET DMPK. 2026. Т. 13. № 3. doi: 10.5599/admet.2772
31. Wu F. и др. Computational Approaches in Preclinical Studies on Drug Discovery and Development // Front. Chem. 2020. Т. 8. doi: 10.3389/fchem.2020.00726
32. Feinberg E. N. и др. Step Change Improvement in ADMET Prediction with PotentialNet Deep Featurization // 2019. doi: 10.48550/arXiv.1903.11789
33. Swanson K. и др. ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries // Bioinformatics. 2024. Т. 40. № 7. doi: 10.1093/bioinformatics/btae416
34. TDC ADMET Benchmark Group [Электронный ресурс] // TDC: ADMET Group. URL: https://tdcommons.ai/benchmark/admet_group/overview/ (дата обращения: 29.10.2025).
35. Cremer J. и др. Equivariant Graph Neural Networks for Toxicity Prediction // 2023. doi: 10.26434/chemrxiv-2023-9kb55-v2

36. Ma X. и др. GraphADT: empowering interpretable predictions of acute dermal toxicity with multi-view graph pooling and structure remapping // Bioinformatics. 2024. Т. 40. № 7. doi: 10.1093/bioinformatics/btae438
37. Moukheiber L. и др. Identifying Protein Features and Pathways Responsible for Toxicity Using Machine Learning and Tox21: Implications for Predictive Toxicology // Molecules. 2022. Т. 27. № 9. doi: https://doi.org/10.3390/molecules27093021
38. Hossain M. и др. FDA-approved heterocyclic molecules for cancer treatment: Synthesis, dosage, mechanism of action and their adverse effect // Heliyon. 2024. Т. 10. № 1. doi: 10.1016/j.heliyon.2023.e23172
39. Elgemeie G. H. и др. Medicinal Chemistry of Pyrazolopyrimidine Scaffolds Substituted with Different Heterocyclic Nuclei // CPD. 2022. Т. 28. № 41. doi: 10.2174/1381612829666221102162000
40. Yang K. и др. Analyzing Learned Molecular Representations for Property Prediction // J. Chem. Inf. Model. 2019. Т. 59. № 8. doi: 10.1021/acs.jcim.9b00237
41. TOXRIC [Электронный ресурс] // Database. Comprehensive toxicological database for intelligent computation. URL: https://toxric.bioinforai.tech/ (дата обращения: 29.10.2025)
42. BindingDB [Электронный ресурс] // https://www.bindingdb.org/ (дата обращения: 29.10.2025).
43. Wu Z. и др. MoleculeNet: a benchmark for molecular machine learning // Chem. Sci. 2018. Т. 9. № 2. doi: doi.org/10.1039/C7SC02664A
44. David L. и др. Molecular representations in AI-driven drug discovery: a review and practical guide // J Cheminform. 2020. Т. 12. № 1. doi: 10.1186/s13321-020-00460-5
45. Muratov E. N. и др. QSAR without borders // Chem. Soc. Rev. 2020. Т. 49. № 11. doi: 10.1039/D0CS00098A
46. Kaneko H. Molecular Descriptors, Structure Generation, and Inverse QSAR/QSPR Based on SELFIES // ACS Omega. 2023. Т. 8. № 24. doi: 10.1021/acsomega.3c01332
47. Bemis G. W., Murcko M. A. The Properties of Known Drugs. 1. Molecular Frameworks // J. Med. Chem. 1996. Т. 39. № 15. doi: 10.1021/jm9602928
48. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning: with Applications in R [Электронный ресурс]. // New York: Springer, 2021. URL: https://www.statlearning.com/ (дата обращения: 29.10.2025).
49. Business Process Model and Notation (BPMN) [Электронный ресурс] // Elsevier BV. – URL: http://bpmn.org/ (дата обращения: 29.09.2025).

ПРИЛОЖЕНИЕ 1. Алгоритм генерации SMARTS-паттернов азолопиримидинов

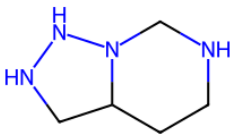
```
from rdkit import Chem
from rdkit.Chem import AllChem, Draw
from rdkit.Chem.rdchem import BondType
from IPython.display import display
import random
random.seed(42)
# базовое ядро пиримидина (N в поз. 1 и 3)
def make_pyrimidine():
    em = Chem.EditableMol(Chem.Mol())
    ring_atoms = [em.AddAtom(Chem.Atom(sym)) for sym in
["N","C","N","C","C","C"]]
    for i in range(6):
        em.AddBond(ring_atoms[i], ring_atoms[(i+1)%6], BondType.SIN-
GLE)
    m = em.GetMol()
    Chem.SanitizeMol(m)
    return m, ring_atoms

# выбор триплета для азола (≥2 гетеро, ≥1 N)
def sample_azole_triplet():
    return random.choice([
        ["N","N","C"], ["N","O","C"], ["N","S","C"],
        ["N","N","O"], ["N","N","S"]
    ])

# слияние 5-членного азола на ребро пиримидина
def fuse_azole_on_edge(pyr, pyr_ring, edge_idx, azole_triplet):
    i = pyr_ring[edge_idx]
    j = pyr_ring[(edge_idx+1) % 6]
    em = Chem.EditableMol(Chem.Mol(pyr))
    A,B,C = [em.AddAtom(Chem.Atom(sym)) for sym in azole_triplet]
    em.AddBond(i,A,BondType.SINGLE)
    em.AddBond(A,B,BondType.SINGLE)
```

```
            em.AddBond(B,C,BondType.SINGLE)
            em.AddBond(C,j,BondType.SINGLE)
            m = em.GetMol()
            try:
                Chem.SanitizeMol(m)
            except:
                return None
            return m

        # проверка: есть 1,3-диазин + общее ребро с 5-членным азолом
        def is_valid_azole_pyrimidine(m):
            ri = m.GetRingInfo().AtomRings()
            five = [r for r in ri if len(r)==5]
            six  = [r for r in ri if len(r)==6]
            if not five or not six:
                return False
            ok6 = False
            for r in six:
                atoms = [m.GetAtomWithIdx(a) for a in r]
                npos = [k for k,a in enumerate(atoms) if a.GetAtomic-
Num()==7]
                # два N в шестичленнике; позиции 1 и 3 по обходу (расстояние
2 по кольцу)
                if len(npos)==2 and ((npos[1]-npos[0])%6 in (2,4)):
                    ok6=True; pyr6=r; break
            if not ok6:
                return False
            def edges(r): return {tuple(sorted((r[k],r[(k+1)%len(r)]))) for
k in range(len(r))}
            e6 = edges(pyr6)
            for r in five:
                atoms=[m.GetAtomWithIdx(a) for a in r]
                het = sum(a.GetAtomicNum() in (7,8,16) for a in atoms)
                nn  = sum(a.GetAtomicNum()==7 for a in atoms)
                if het>=2 and nn>=1 and (e6 & edges(r)):
```

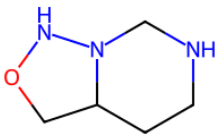
```
                return True
        return False

    # простая декорация галогенами на sp2-C
    def decorate_halogen(m, p_attach=0.3):
        em = Chem.EditableMol(Chem.Mol(m))
        cand = [a.GetIdx() for a in m.GetAtoms()
                if a.GetAtomicNum()==6 and a.GetDegree()<=3 and
a.GetIsAromatic()]
        random.shuffle(cand)
        hal = random.choice(["F","Cl","Br"])
        for idx in cand:
            if random.random()<p_attach:
                h = em.AddAtom(Chem.Atom(hal))
                em.AddBond(idx,h,BondType.SINGLE)
        mm = em.GetMol()
        try:
            Chem.SanitizeMol(mm); return mm
        except: return m

    # генератор библиотеки
    def generate_azolopyrimidines(n=100, decorate=True):
        base, ring = make_pyrimidine()
        smiles=set(); trials=0
        while len(smiles)<n and trials<n*50:
            trials+=1
            edge = random.randrange(6)
            trip = sample_azole_triplet()
            fused = fuse_azole_on_edge(base,ring,edge,trip)
            if fused and is_valid_azole_pyrimidine(fused):
                if decorate: fused = decorate_halogen(fused,0.25)
                smi = Chem.MolToSmiles(fused, canonical=True)
                smiles.add(smi)
        return sorted(smiles)
```

```
    # проверка корректности молекулы
    def check_and_kekulize(smiles_list, quoted=False):
        out = []
        for sm in smiles_list:
            m = Chem.MolFromSmiles(sm)
            if not m:
                continue
            try:
                Chem.SanitizeMol(m)
                Chem.Kekulize(m, clearAromaticFlags=True)
                smi = Chem.MolToSmiles(m, canonical=True)
                out.append(f"'{smi}'" if quoted else smi)
            except Exception:
                continue
        return out

    def show_grid(smiles_list, per_row=5, size=(220,220), outfile="az-
olopyrimidines_grid.png"):
        # если вдруг пришли с кавычками —снимем их
        clean = [s.strip("'\"") for s in smiles_list]
        mols = [Chem.MolFromSmiles(s) for s in clean]
        legends = [f"#{i+1} | {s}" for i,s in enumerate(clean)]
        img = Draw.MolsToGridImage(mols, molsPerRow=per_row, subImg-
Size=size,
                                   legends=legends, useSVG=False)  #
PIL.Image
        if hasattr(img, "save"):
            img.save(outfile)
        else:
            with open(outfile, "wb") as f:
                f.write(img.data)
        return outfile, img
```

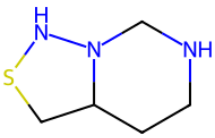
ПРИЛОЖЕНИЕ 2. Генерированные структуры азолпиримидинов



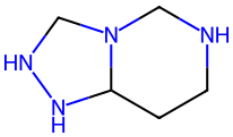
#1 | C1CC2CNNN2CN1



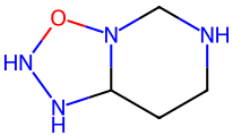
#2 | C1CC2CONN2CN1



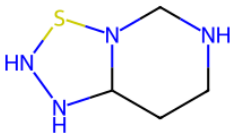
#3 | C1CC2CSNN2CN1



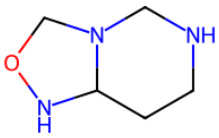
#4 | C1CC2NNCN2CN1



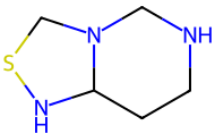
#5 | C1CC2NNON2CN1



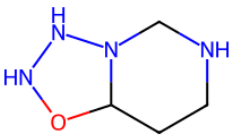
#6 | C1CC2NNSN2CN1



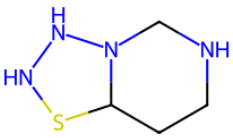
#7 | C1CC2NOCN2CN1



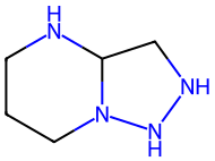
#8 | C1CC2NSCN2CN1



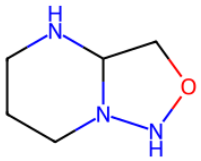
#9 | C1CC2ONNN2CN1



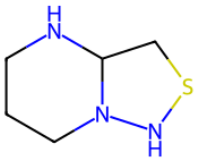
#10 | C1CC2SNNN2CN1



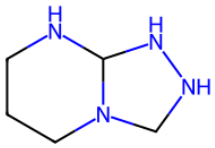
#11 | C1CNC2CNNN2C1



#12 | C1CNC2CONN2C1

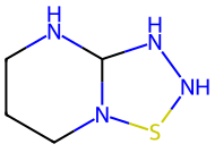


#13 | C1CNC2CSNN2C1



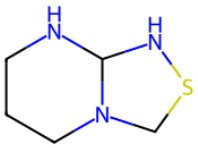
#14 | C1CNC2NNCN2C1

#15 | C1CNC2NNON2C1

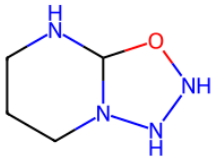


#16 | C1CNC2NNSN2C1

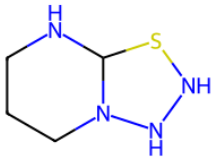
#17 | C1CNC2NOCN2C1



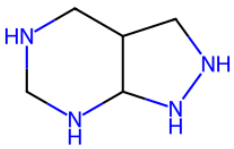
#18 | C1CNC2NSCN2C1



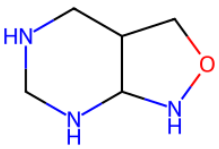
#19 | C1CNC2ONNN2C1



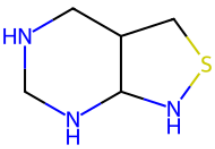
#20 | C1CNC2SNNN2C1



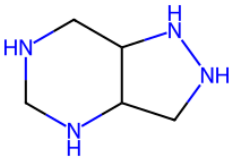
#21 | C1NCC2CNNC2N1



#22 | C1NCC2CONC2N1

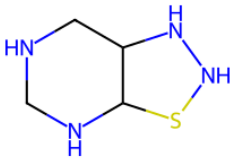


#23 | C1NCC2CSNC2N1



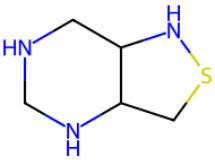
#24 | C1NCC2NNCC2N1

#25 | C1NCC2NNOC2N1

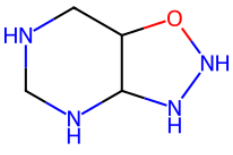


#26 | C1NCC2NNSC2N1

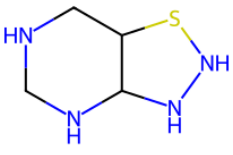
#27 | C1NCC2NOCC2N1



#28 | C1NCC2NSCC2N1



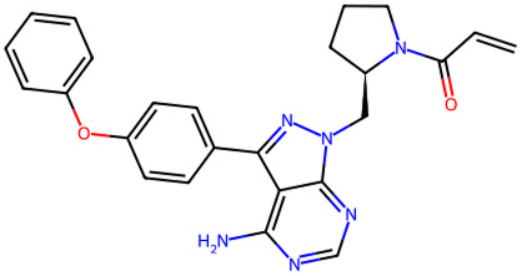
#29 | C1NCC2ONNC2N1



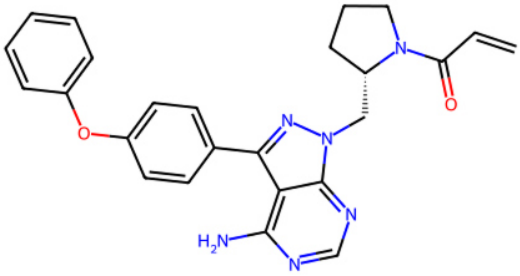
#30 | C1NCC2SNNC2N1

ПРИЛОЖЕНИЕ 3. Функция фильтрации азолопиримидинов

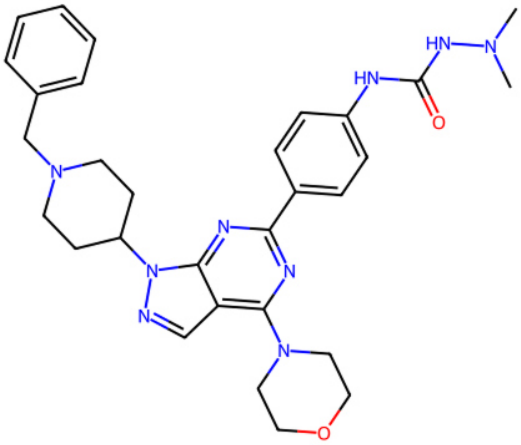
```
    AZOLOPYRIMIDINE_SMARTS =
['c1cc2cnnn2cn1','c1cc2conn2cn1','c1cc2csnn2cn1','c1cc2nncn2cn1','c1cc2nno
n2cn1','c1cc2nnsn2cn1','c1cc2nocn2cn1','c1cc2nscn2cn1','c1cc2onnn2cn1','c1
cc2snnn2cn1','c1cnc2cnnn2c1','c1cnc2conn2c1','c1cnc2csnn2c1','c1cnc2nncn2c
1','c1cnc2nnon2c1','c1cnc2nnsn2c1','c1cnc2nocn2c1','c1cnc2nscn2c1','c1cnc2
onnn2c1','c1cnc2snnn2c1','c1ncc2cnnc2n1','c1ncc2conc2n1','c1ncc2csnc2n1','
c1ncc2nncc2n1','c1ncc2nnoc2n1','c1ncc2nnsc2n1','c1ncc2nocc2n1','c1ncc2nscc
2n1','c1ncc2onnc2n1','c1ncc2snnc2n1']
    _AZOLO_PATTERNS = [Chem.MolFromSmarts(s) for s in AZOLOPYRIMI-
DINE_SMARTS if Chem.MolFromSmarts(s)]

    def has_azolopyrimidine(smiles: str, patterns=_AZOLO_PATTERNS,
min_atoms: int = 7) -> bool:
        mol = Chem.MolFromSmiles(smiles)
        if not mol or mol.GetNumAtoms() < min_atoms:
            return False
        for p in patterns:
            if mol.HasSubstructMatch(p):
                return True
        return False

    def filter_azolopyrimidines(df, smiles_col="SMILES", add_flag=True,
min_atoms=7):
        mask = df[smiles_col].apply(lambda s: has_azolopyrimidine(s,
_AZOLO_PATTERNS, min_atoms))
        out = df[mask].copy()
        if add_flag:
            out['is_azolopyrimidine'] = True
        print(f"Азолопиримидины {len(df)} : {len(out)} строк")
    return out
```



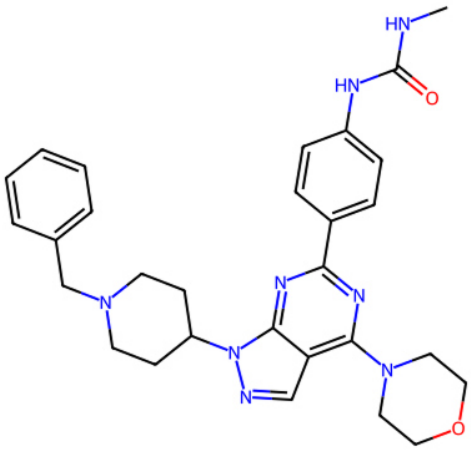
pIC50 = 7.73



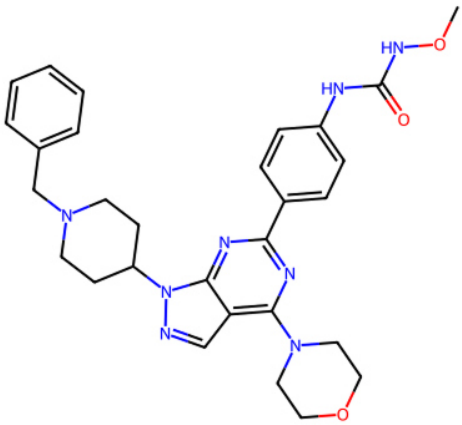
pIC50 = 8.64



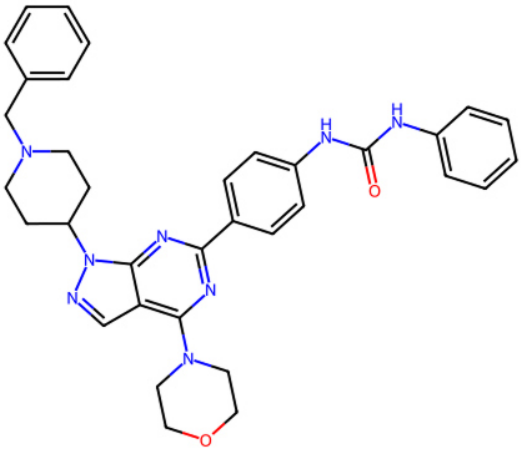
pIC50 = 7.08



pIC50 = 9.29

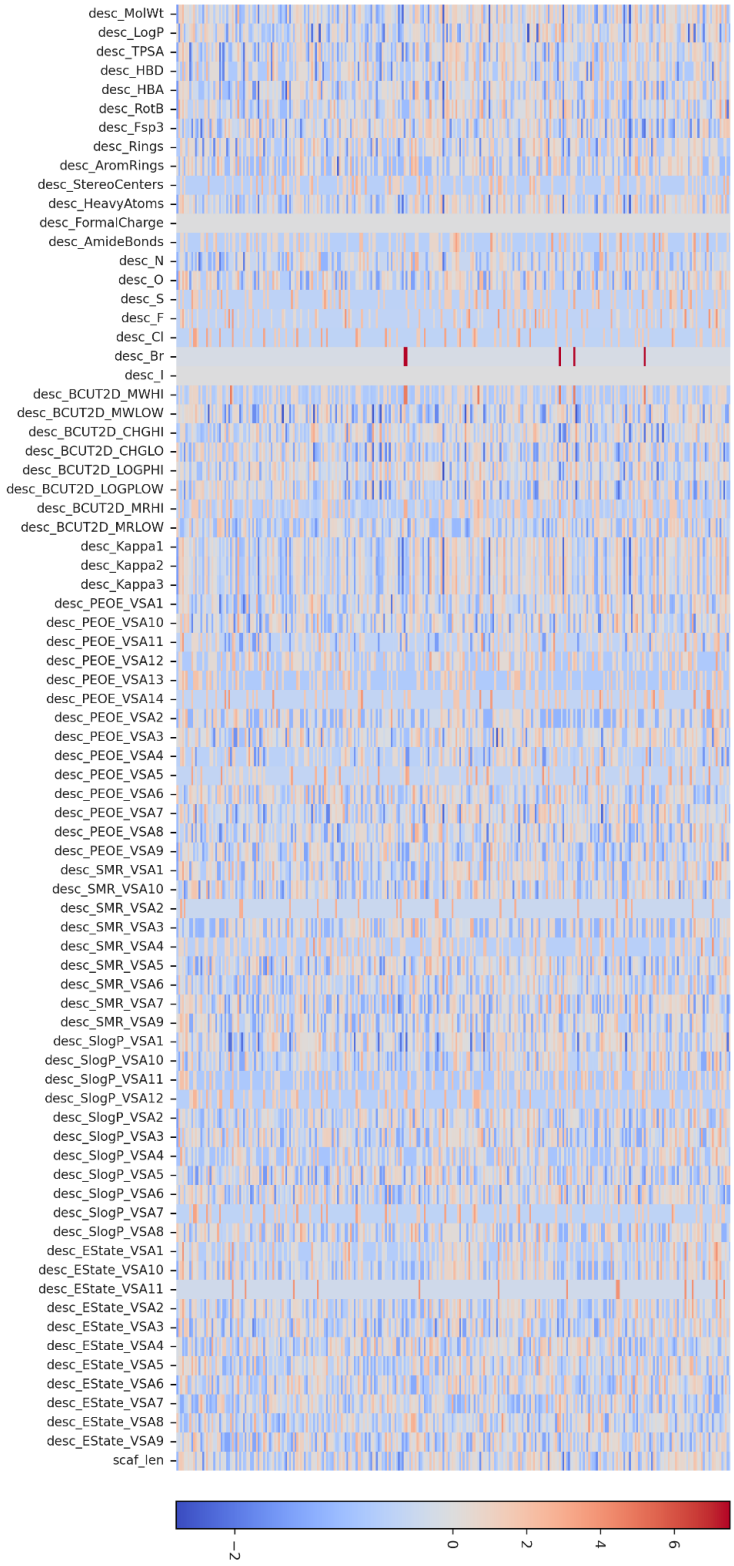


pIC50 = 7.80

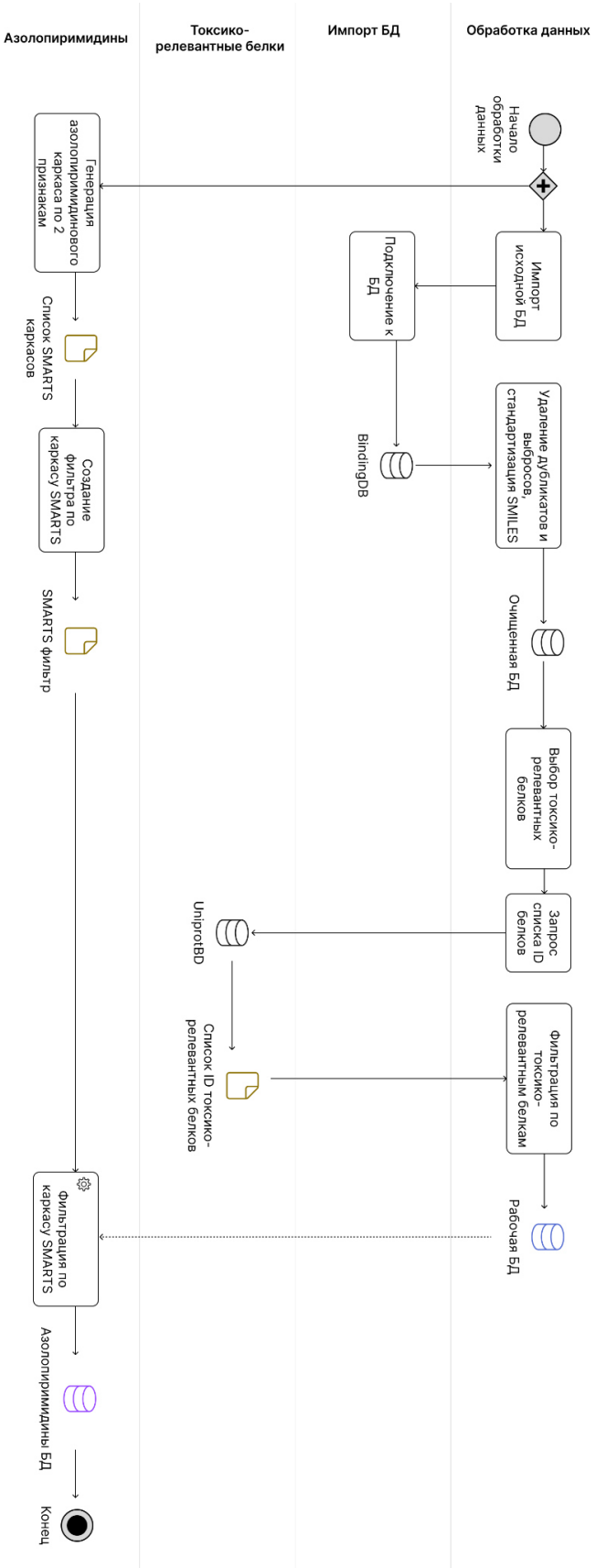


pIC50 = 8.62

ПРИЛОЖЕНИЕ 5. Тепловая карта дескрипторов



ПРИЛОЖЕНИЕ 6. Технологическая схема проекта в нотации BPMN



Моделирование
Проверка
Модель №1
Модель №2
Признаки
Валидация
Начало моделирования
Рабочая БД
Обучение Модели №1
Оценка на тестовом наборе
Модель №1
Результаты обучения
Азолопиримидины БД
Подключение к Модели №1
Обучение Модели №2
Оценка чувствительности модели к тонким изменениям структуры
Изменение параметров обучения
низкая оценка
высокая оценка
Оценка на тестовом наборе
Модель №2
Результаты обучения
Конец моделирования
Начало вычисления признаков
Создание функции вычисления признаков
Применение к предсказанным значениям
Получение признаков высокоаффинных соединений
Список признаков
Конец вычисления признаков
Начало валидации
Лабораторные данные
Удаление дубликатов и выбросов, стандартизация SMILES
Очищенные лабораторные данные
Предсказание аффинности с помощью Модели №2
Сравнение предсказаний с лабораторными pIC50
Результаты сравнения
Конец сравнения