На правах рукописи

БИЗЮКОВА НАДЕЖДА ЮРЬЕВНА

ФОРМИРОВАНИЕ ЗНАНИЙ О БИОЛОГИЧЕСКОЙ АКТИВНОСТИ НИЗКОМОЛЕКУЛЯРНЫХ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ НА ОСНОВЕ АВТОМАТИЗИРОВАННОГО АНАЛИЗА ТЕКСТОВ

1.5.8. - Математическая биология, биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата биологических наук

Москва – 2026

Работа выполнена в Федеральном государственном бюджетном научном учреждении «Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича» (ИБМХ)

Научный руководитель:	Кандидат биологических наук, Тарасова Ольга Александровна
Научный консультант:	Доктор биологических наук, кандидат физико-математических наук, профессор, академик РАН, Поройков Владимир Васильевич
Официальные оппоненты:	Орлов Юрий Львович, доктор биологических наук, ФГАОУ ВО Первый МГМУ имени И.М. Сеченова Минздрава России (Сеченовский Университет), Институт цифровой медицины, профессор кафедры информационных и интернет-технологий Девяткин Дмитрий Алексеевич, кандидат физико-математических наук, ФГБНУ «Институт проблем искусственного интеллекта», руководитель 73 отдела

Ведущая организация: Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук»

Защита состоится «23» апреля 2026 года в 11 часов на заседании диссертационного совета 24.1.172.01 (Д 001.010.01) в Федеральном государственном бюджетном научном учреждении «Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича» по адресу: 119121, Москва, ул. Погодинская, д. 10, стр. 8.

С диссертацией можно ознакомиться в библиотеке ИБМХ и на сайте www.ibmc.msk.ru

Автореферат разослан «____» _____ 2026 г.

Ученый секретарь диссертационного совета,
кандидат химических наук  Карпова Е.А.

1. ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

1.1. Актуальность темы и степень ее разработанности

Стремительный рост объёма научных публикаций в области биологии и медицины значительно усложняет проведение литературного поиска, который является необходимым этапом большинства исследований. Увеличение числа журналов и статей по биомедицинской тематике сопровождается расширением круга изучаемых объектов и методов, что многократно повышает риск неполного учёта уже имеющихся данных.

Особенно остро эта проблема проявляется при изучении биологической активности низкомолекулярных органических соединений - ключевого аспекта медицинской химии, фармакологии и токсикологии при разработке лекарственных препаратов. Существенная доля временных и финансовых издержек в процессе создания лекарств обусловлена необходимостью анализа сведений о механизмах действия, фармакокинетике и токсичности соединений (Poroikov V. et al., 2019; Аладышева Ж. И. и др., 2019), которые рассредоточены по многочисленным публикациям. Возможности ручной обработки таких данных крайне ограничены, что делает применение автоматизированных методов интеллектуального анализа текстов критически необходимым (Peng Y. et al., 2019; Moreau E. et al., 2024; Cohen A.M. et al., 2005).

Существующие технические решения не позволяют решить проблему в полной мере. Подходы, в которых интегрированы большие массивы данных и фактографические базы знаний (Bizon C. et al., 2019; Bachman J. A. et al., 2023), требуют постоянного обновления и фокусируются на ограниченном наборе источников, что снижает полноту получаемых сведений. Напротив, специализированные методы, направленные, например, на поиск репозиционируемых молекул или построение графов заболеваний (Tandra G. et al., 2023; Xing H. et al., 2023), демонстрируют высокую эффективность только в узких областях и с трудом масштабируются на другие задачи.

Отсутствие универсального подхода, обеспечивающего воспроизводимое, интерпретируемое и системное извлечение данных непосредственно из текстов

научных публикаций, определяет потребность в новых подходах, объединяющих преимущества интегральных методов и специализированных алгоритмов.

Таким образом, разработка метода интеллектуального анализа текстов, ориентированного на выявление и структурирование сведений о биологической активности низкомолекулярных органических соединений, является актуальной научной задачей, имеющей важное теоретическое и прикладное значение для биоинформатики, медицинской химии и фармакологии.

1.2. Степень разработанности

Задачи автоматизированного анализа биомедицинских текстов развиваются в течение последних двух десятилетий, и за это время сформировалась группа подходов, направленных на распознавание и интерпретацию сведений, содержащихся в научных публикациях. Созданы специализированные коллекции и наборы данных для извлечения биомедицинских сущностей и отношений (Krallinger M. et al., 2015; Wei C.-H. et al., 2016), разработаны алгоритмы на основе правил (Ben Abacha A. et al., 2011; Ravikumar K.E. et al., 2021), статистические модели и современные нейросетевые архитектуры (Nguyen D.Q. et al., 2018; Wang W. et al., 2017; Lai P.-T. et al., 2021), обеспечивающие высокую точность при решении отдельных видов задач. Тем не менее большинство существующих решений фокусируются на выделении ограниченных типов сущностей или конкретных видов связей и не предназначены для интегрального анализа данных о биологической активности низкомолекулярных органических соединений. Это приводит к фрагментарности существующих подходов и подчёркивает необходимость разработки универсальной методологии, объединяющей распознавание наименований биологических и химических объектов, определение взаимосвязей между ними и дальнейшую систематизацию полученной информации.

1.3. Цель и задачи исследования

Целью данной работы является разработка, реализация и тестирование метода для извлечения взаимосвязей между наименованиями низкомолекулярных органических соединений и их известной биологической активностью.

Для достижения цели были поставлены следующие задачи:

1. Создание коллекции текстов с информацией о наименованиях низкомолекулярных органических соединений и данными об их биологической активности.
2. Разработка, тестирование, и программная реализация метода извлечения ассоциаций между наименованиями низкомолекулярных органических соединений и видами их биологической активности.
3. Создание базы данных о наименованиях низкомолекулярных органических соединений, их синонимах, и ассоциированных с ними видах биологической активности с возможностью автоматического пополнения с применением разработанного метода.
4. Проверка согласованности и полноты информации о биологической активности низкомолекулярных органических соединений.

1.4. Научная новизна работы

Впервые предложен интегральный метод автоматизированного извлечения всесторонних сведений о биологической активности низкомолекулярных органических соединений из биомедицинских текстов. Разработанный метод объединяет этапы отбора релевантных публикаций, извлечения наименований биомедицинских сущностей, определения взаимосвязей между ними и последующую нормализацию данных, обеспечивая их системное и воспроизводимое структурирование.

Создана специализированная база данных, включающая тексты, наименования объектов, ассоциации и сведения из фактографических ресурсов, что позволило объединить разнородные источники информации в единую логическую модель. Предложенный метод обеспечивает интерпретируемость и достаточную точность при сравнительно небольших вычислительных затратах, а также демонстрирует устойчивость при применении к гетерогенным коллекциям биомедицинских публикаций.

Проведённая валидация показала согласованность полученных данных с существующими базами знаний и позволила выявить отсутствующие в них новые сведения, что свидетельствует о применимости метода к расширению имеющихся представлений о биологической активности химических соединений.

1.5. Теоретическая и практическая значимость работы

Теоретическая значимость работы заключается в развитии подходов интеллектуального анализа биомедицинских текстов на основе предложенного нами интегрального метода, который объединяет распознавание наименований объектов, выявление ассоциаций и их нормализацию в единую структуру знаний. Разработанный подход формирует основу для системного представления сведений о биологической активности низкомолекулярных органических соединений и позволяет осуществлять объединение разнородных источников текстовой и фактографической информации в единую логическую модель.

Практическая значимость работы определяется возможностью использования предложенного метода для решения широкого круга задач в биоинформатике, медицинской химии и фармакологии. Метод позволяет автоматизировать анализ больших массивов биомедицинской литературы, выявлять новые сведения о биологической активности соединений. Разработанная база данных и программная реализация метода могут быть использованы для поддержки исследований в области разработки лекарственных средств, анализа механизмов заболеваний и формирования гипотез о потенциальных биологических эффектах лекарственно-подобных соединений.

1.6. Личный вклад автора

Лично автором проведён анализ литературных источников и сформирована коллекция биомедицинских текстов, использованная в исследовании. Разработаны и реализованы алгоритмы интегрального метода автоматизированного извлечения сведений о биологической активности низкомолекулярных органических соединений, а также выполнено их тестирование на репрезентативных коллекциях текстов.

Автор принимала непосредственное участие в создании базы данных, предназначенной для хранения и структурирования полученной информации, разработке инструментов, обеспечивающих практическое применение предложенного подхода. Лично автором выполнена серия вычислительных экспериментов, включающая оценку точности, полноты и воспроизводимости извлечённых сведений, а также интерпретацию результатов с позиции их биологической значимости.

1.7. Основные положения, выносимые на защиту

- Разработан интегральный метод извлечения ассоциаций между наименованиями низкомолекулярных органических соединений и видами их биологической активности, который обеспечивает высокую точность и полноту извлекаемой информации.
- С применением разработанного метода создана база данных о лекарственно-подобных соединениях и их биологической активности, обеспечивающая систематизацию информации и возможность автоматического пополнения новыми сведениями.
- Проведен анализ согласованности и полноты извлечения данных, который показал, что применение методов интеллектуального анализа текстов позволяет значительно обогатить информацию, доступную в существующих базах данных.

1.8. Степень достоверности и апробация результатов работы

Достоверность полученных результатов обеспечена использованием репрезентативных коллекций биомедицинских текстов (PubMed, PMC), применением профильных онтологий и фактографических баз данных, корректностью используемых математических методов и многократной проверкой алгоритмов на независимых выборках. Согласованность извлечённой информации с данными экспериментальных и клинических исследований подтверждает обоснованность научных выводов.

Основные положения работы были представлены и обсуждены на российских и международных научных мероприятиях, включая XXVI–XXVIII Симпозиумы

«Биоинформатика и компьютерное конструирование лекарств» (2020–2022), конференцию «МОБИ-ХимФарма2020 VI» (2020), MedChem-Russia (2021), XIII и XIV Multiconferences BGRS/SB (2022, 2024), ACS Fall (2023), VI Международную конференцию «ПОСТГЕНОМ’2024», II Школу молодых учёных (2025), XXVI Харитоновские тематические научные чтения (2025), VI Международную научную конференцию и X Всероссийский молодёжный научный форум «Наука будущего – наука молодых».

1.9. Публикации

По теме диссертации опубликовано 17 работ, из которых 6 статей в рецензируемых научных журналах и 11 публикаций в материалах научных конференций.

1.10. Объем и структура диссертации

Диссертация изложена на 150 страницах машинописного текста, содержит 12 таблиц и 17 рисунков. Работа включает следующие разделы: введение, обзор литературы, материалы и методы, результаты и их обсуждение, заключение и выводы. Кроме того, диссертация содержит список сокращений, словарь терминов, список литературы, сведения о финансировании работы, а также два приложения, содержащие дополнительные материалы, иллюстрирующие проведённые исследования.

2. МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

2.1. Формирование коллекции биомедицинских текстов

Основным источником данных для исследования являлась библиографическая база данных PubMed для обеспечения доступа к заголовкам и аннотациям биомедицинских публикаций. Эти тексты использовались в качестве материала для анализа, поскольку они доступны для всех записей и содержат основные сведения о целях, объектах и результатах исследований.

Отбор релевантных публикаций осуществлялся двумя методами. В первом - на основе поиска по MeSH-терминам – проводилось итеративное уточнение запросов: перечень терминов последовательно корректировался в зависимости от результатов анализа полученной выборки, что позволило повысить долю релевантных источников. Дополнительно применялись фильтры по типам публикаций для исключения обзоров с целью отбора оригинальных исследований.

Во втором методе использовали классификацию текстов с помощью алгоритмов машинного обучения. Были сформированы коллекции публикаций с ручной разметкой наличия описаний *in vitro* тестирования противовирусной активности. Для обучения применялись различные алгоритмы – метод опорных векторов, k-ближайших соседей, случайный лес и рекуррентные нейронные сети (LSTM). Представление текстов осуществлялось на основе TF-IDF-векторов или ключевых терминов. Дополнительно рассматривалась классификация по типам биологических тестов на основе меток из BioAssay Ontology (раздел “Assays” базы данных ChEMBL).

2.2. Методы распознавания наименований биологических и химических объектов

Для последующего анализа сведений о биологической активности низкомолекулярных органических соединений в текстах необходимо корректно идентифицировать упоминания ключевых биомедицинских объектов. В исследовании учитывались химические соединения, белки и гены, заболевания, биологические виды, клеточные линии, микроРНК, а также однонуклеотидные полиморфизмы и аминокислотные замены.

Извлечение наименований этих объектов осуществлялось на основе комбинации методов машинного обучения и алгоритмов, основанных на правилах. Для распознавания наименований химических соединений, белков, генов, заболеваний, биологических видов использовались модели машинного обучения, включая алгоритмы условных случайных полей (CRF) и HunFlair (Weber L. et al., 2021; Sänger M. Et al., 2024), основанную на архитектуре LSTM-CRF. Мы обучали эти модели на специализированных размеченных выборках, что позволило учитывать широкий спектр вариативных форм записи терминов, включая систематические названия и аббревиатуры.

Для сущностей, отличающихся строгой формальной структурой, применялись алгоритмы, основанные на регулярных выражениях. Такой подход использовался для распознавания микроРНК и описаний генетических вариантов (однонуклеотидных и аминокислотных замен), что обеспечивало высокую точность за счёт использования заранее сформированных шаблонов.

2.3. Методы извлечения и нормализации ассоциаций

После распознавания наименований биологических и химических объектов в текстах осуществлялось извлечение ассоциаций между ними, что необходимо для последующего анализа биологической активности низкомолекулярных соединений. В работе рассматривались связи между химическими соединениями, белками и генами, видами биологической активности, заболеваниями, биологическими видами и другими объектами, упоминаемыми в научных публикациях.

Для извлечения ассоциаций применялся комбинированный подход, включающий правила и шаблоны, отражающие устойчивые лингвистические конструкции, а также методы машинного обучения. Алгоритмы, основанные на правилах, использовали набор типовых синтаксических структур (например, «ингибирует», «связан с», «вызывает», «обусловливает»), позволяющих выделять упоминания взаимодействий и функциональных связей между объектами. Для увеличения точности соотнесения пар объектов с лингвистической конструкцией, характеризующей взаимосвязь между ними, была разработана численная оценка, основанная на расстоянии между релевантными фрагментами текста.

После извлечения проводилась нормализация ассоциаций: каждому объекту присваивались устойчивые идентификаторы из внешних контролируемых словарей и баз данных (химические соединения и белки – ChEMBL, биологические виды – NCBI Taxonomy, заболевания – Human Disease Ontology). Это позволяло объединять различные варианты написания одного и того же объекта и сформировать единое представление связей между сущностями. Нормализация обеспечила согласованность данных и возможность их дальнейшей интеграции с внешними источниками знаний и фактографическими базами данных.

2.4. Создание и наполнение базы данных

Для хранения структурированной информации, извлечённой из биомедицинских текстов, нами разработана реляционная база данных. Её архитектура предусматривала разделение данных по функциональным модулям. В отдельные таблицы мы помещали: тексты публикаций, распознанные наименования различных типов, результаты нормализации, ассоциации между объектами.

Проектирование логической схемы базы данных выполнялось с учётом необходимости фиксировать множественные варианты написания сущностей и их сопоставление с внешними идентификаторами. Поэтому каждая запись включала ссылки на источник текста, тип сущности, позицию в аннотации и нормализованный идентификатор. Для ассоциаций предусматривались поля, в которых для них указаны типы взаимосвязей и контекстов, из которых они были извлечены. Реализация базы данных произведена с использованием системы управления базами данных (СУБД) MySQL.

Наполнение базы данных осуществлялось автоматически, посредством скриптов на языке Python. Эти скрипты последовательно обрабатывали результаты распознавания сущностей и извлечения связей, выполняли нормализацию терминов, проверяли корректность форматов и вносили данные в соответствующие таблицы. При загрузке данных применялись процедуры устранения дубликатов и проверки согласованности, что обеспечивало целостность информации при обработке больших массивов публикаций.

2.5. Методы оценки качества и проверки полноты сведений

Оценка корректности работы разработанных алгоритмов проводилась на основе комплекса процедур, направленных на проверку точности извлечения сущностей и ассоциаций, а также на анализ полноты сведений, получаемых из текстов.

Для оценки качества распознавания сущностей использовались стандартные метрики информационного поиска: точность (precision), полнота (recall) и их гармоническое среднее (F_1-score). Проверка выполнялась на вручную размеченных выборках публикаций, содержащих упоминания химических соединений, биологических объектов и экспериментальных характеристик. Для каждого типа сущностей сравнивался предсказанный набор упоминаний с эталонной разметкой.

Качество извлечения ассоциаций оценивалось аналогичным образом: на размеченных примерах мы проверяли способность алгоритмов корректно выделять типы связей. Для анализа случаев расхождения применялись процедуры ручной валидации, позволяющие уточнять формулировки правил и корректировать используемые шаблоны.

Для оценки полноты сведений, получаемых из коллекции текстов, нами проводилось сопоставление извлечённых данных с информацией, содержащейся во внешних фактографических ресурсах. При этом мы анализировали совпадения и расхождения в перечнях химических соединений, биологических мишеней и описанных видов активности. Дополнительно проверена избыточность извлечённых сведений и наличие уникальных данных, отсутствующих в фактографических базах.

3. ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

3.1.1. Оценка эффективности методов отбора релевантных публикаций

Проверка качества сформированных коллекций текстов показала сопоставимую эффективность обоих упомянутых выше подходов к выявлению релевантных публикаций. Отбор на основе MeSH-терминов обеспечивает высокую точность – более 80 % публикаций содержали сведения о биологической активности

низкомолекулярных соединений при ручной проверке подвыборок. Данный подход позволяет формировать устойчивые к тематическим различиям коллекции и обеспечивает хорошую полноту охвата данных.

При применении методов машинного обучения на размеченных выборках для задач бинарной классификации наиболее высокой точностью характеризуется модель на основе рекуррентных нейронных сетей (LSTM), которая обеспечивала значения F_1-score 0,86 при равных долях классов в обучающей выборке и 0,76 при существенном дисбалансе (25/75 релевантные/нерелевантные). Классические алгоритмы машинного обучения (метод опорных векторов, случайный лес) давали несколько более низкие значения показателей, но сохраняли приемлемый уровень точности и полноты при значительно меньших вычислительных затратах.

Для задач многоклассовой классификации по категориям экспериментов по классификации BioAssay Ontology (BAO) было получено значение F_1-score, равное 0,71, что подтвердило возможность автоматического выделения публикаций, различающихся по типам биологического тестирования.

Проведённое сравнение показало, что использование методов машинного обучения для отбора текстов действительно позволяет формировать коллекции с высокой точностью и узкой тематической направленностью, однако требует значительных трудозатрат на подготовку обучающих наборов для каждой отдельной задачи. В отличие от этого, отбор по MeSH-терминам характеризуется универсальностью, устойчивой применимостью к различным предметным областям и достаточно высоким уровнем релевантности публикаций при минимальных затратах на формирование запроса. Именно эти преимущества определили выбор подхода на основе MeSH-аннотаций в качестве основной стратегии формирования коллекций, использованных на последующих этапах исследования.

3.1.2. Состав и характеристики сформированных коллекций текстов

Для экспериментальной проверки разработанных алгоритмов были сформированы тематические коллекции аннотаций публикаций из базы данных PubMed, различающиеся по объёму и области исследования (таблица 1).

Таблица 1. Коллекции текстов, сформированные в ходе исследования

Коллекция	**Объём**	**Тематика и особенности содержания**
XenoMet	~1 000	Биотрансформация химических соединений, метаболиты и реакции
Противовирусная активность	>400 000	Биологическая активность соединений, ингибирующих вирусные инфекции
Сигнальные пути при заболеваниях	>160 000	Роль межклеточных сигнальных путей в патогенезе разнообразных патологий
Путь Hedgehog	~10 000	Участие сигнального пути Hedgehog в онкогенезе
Большое депрессивное расстройство	~30 000	Патогенез и терапия большого депрессивного расстройства

Первая коллекция включает около 1000 публикаций, посвящённых биотрансформации низкомолекулярных соединений. Она была вручную аннотирована с учетом упоминаний родительских соединений, метаболитов и типов метаболических реакций и использовалась для обучения и валидации алгоритма в задаче извлечения сведений о путях биотрансформации (Biziukova N. et al., 2025).

Коллекция публикаций о противовирусной активности соединений включает более 400 тыс. текстов и охватывает широкий спектр механизмов ингибирования вирусных инфекций. Она применялась для проверки масштабируемости алгоритмов и анализа активности соединений (Тарасова О.А. и др., 2024).

Набор публикаций, отражающих роль сигнальных путей в патогенезе заболеваний, насчитывает свыше 160 тыс. аннотаций и использовался для оценки

корректности выделения каскадов функциональных взаимодействий и возможности их регуляции при помощи химических соединений (Biziukova N. et al., 2023).

Также сформированы две специализированные выборки: около 10 тыс. публикаций по участию сигнального пути Hedgehog при онкогенезе – для валидации на узкой тематике (Biziukova N. et al., 2024), и около 30 тыс. публикаций по большому депрессивному расстройству - для демонстрации применимости метода при анализе патологических процессов (Тарасова О.А. и др., 2024).

Сформированные коллекции обеспечили разнообразие предметных областей и объем данных, необходимый для оценки устойчивости и универсальности предложенного подхода.

3.2. Разработка и апробация интегрального метода

Нами разработан интегральный метод автоматического извлечения сведений о биологической активности низкомолекулярных органических соединений из текстов научных публикаций, включающий этапы отбора релевантных публикаций, распознавания биомедицинских сущностей, выделения взаимосвязей между объектами и их нормализации с использованием внешних онтологий и автоматических запросов к фактографическим базам данных. Такой подход позволяет формировать структурированное представление знаний непосредственно из литературных источников и обеспечивает возможность последующей аналитической обработки данных.

Проверка качества разработанного интегрального метода проводилась на вручную размеченных выборках публикаций, сформированных из различных тематических коллекций. При оценке истинно положительными считались не только корректно выявленные ассоциации между объектами, но и их верное соотнесение с типами биологической активности. По итогам валидации значение F_1-score достигло 0,82, что свидетельствует о высокой точности идентификации как самих связей, так и их функциональной направленности.

Разработанный подход был апробирован при решении различных научных задач. В частности, проведён анализ публикаций, посвящённых поиску низкомолекулярных соединений с противовирусной активностью, что позволило

систематизировать сведения о потенциальных ингибиторах вирусных инфекций и об их молекулярных мишенях. Кроме того, выполнено исследование текстов, описывающих роль сигнального пути Hedgehog в опухолевых заболеваниях, а также процессов, связывающих воспаление и развитие большого депрессивного расстройства. Во всех случаях извлечённые сведения позволили выделить группы генов, белков и химических соединений, функционально вовлечённых в развитие соответствующих патологических состояний, что продемонстрировало возможность применения метода как в исследованиях биологически активных соединений, так и в анализе молекулярных механизмов заболеваний различной природы.

Представление результатов извлечения в виде взаимосвязанных структур (таблиц и графов) позволяет выявлять недостающие звенья в цепочках биологических взаимодействий, формировать гипотезы о потенциальных механизмах действия соединений и, тем самым, служить инструментом для расширения существующих знаний на основе анализа литературных данных.

3.3. База данных сведений о биологической активности низкомолекулярных соединений

На основе результатов автоматизированного анализа текстов сформирована база данных, обеспечивающая хранение структурированной информации о низкомолекулярных органических соединениях, их биологических мишенях и описанных видах активности. В базу включены нормализованные наименования объектов с сопоставлением внешним идентификаторам, связи между ними, а также сведения об источниках, в которых эти данные были обнаружены. Такая организация обеспечивает прослеживаемость происхождения каждой записи и возможность агрегирования информации, полученной из разных публикаций.

Разработанная структура данных позволила интегрировать сведения, извлечённые из различных тематических коллекций, в первую очередь относящихся к исследованиям противовирусной активности низкомолекулярных соединений. Информация представлена в нормализованном виде с указанием источников и возможностью расширения перечня объектов по мере дальнейшего пополнения базы данными из новых публикаций. Созданный ресурс ориентирован на систематизацию данных, получаемых в результате анализа литературы, и может использоваться в

задачах поиска биологически активных соединений и изучения молекулярных аспектов противовирусных механизмов действия.

Для обеспечения практического применения разработанного подхода создан веб-интерфейс (https://way2drug.com/viruses/nlp/), предоставляющий доступ к информации базы данных и позволяющий выполнять навигацию, поиск, фильтрацию и выборку извлечённых сведений в удобном формате.

3.4. Проверка полноты и согласованности полученных сведений

Для оценки согласованности извлечённых сведений проведено сопоставление результатов автоматизированного анализа текстов с данными фактографической базы ChEMBL, содержащей экспериментальные сведения о биологической активности соединений. Установлено, что только 1,86 % публикаций из нашей выборки одновременно присутствуют в ChEMBL, что обусловлено как более широким охватом тематик (включая исследования репозиционирования лекарственных препаратов и клинические данные). В то же время, 42,22 % публикаций из ChEMBL оказались представленными в нашей выборке, тогда как остальные были исключены преимущественно из-за отсутствия MeSH-индексации по противовирусной активности или из-за принадлежности к обзорам.

Сопоставление ассоциаций между химическими соединениями и биологическими объектами показало, что 23,77 % извлечённых связей полностью совпадают с записями в ChEMBL, что подтверждает корректность результатов извлечения. При этом 47,29 % ассоциаций отсутствуют в базе данных, что связано с более широким охватом информации в текстах публикаций и отличиями в степени детализации аннотаций объектов. В частности, нормализованные записи в ChEMBL могут содержать стереоизомеры или информацию о лекарственных формах соединений, тогда как в текстах публикаций, как правило, фиксируются международные непатентованные наименования препаратов или коды кандидатов в них.

Полученные результаты показали, что предложенный подход воспроизводит значительную часть сведений, уже представленных в существующих

фактографических ресурсах, и при этом позволяет выявлять дополнительные связи между объектами, не отражённые в ChEMBL.

4. ЗАКЛЮЧЕНИЕ

Научные публикации являются основным источником сведений о биологической активности низкомолекулярных органических соединений, однако постоянно растущий объём биомедицинской литературы делает ручной анализ таких данных затруднительным. Автоматизация извлечения и систематизации информации из текстов научных статей становится необходимым условием для эффективного использования накопленных знаний при поиске и оптимизации лекарственных соединений.

В ходе исследования разработан и реализован интегральный метод автоматизированного извлечения сведений о биологической активности низкомолекулярных соединений. Предложенный метод объединяет отбор релевантных публикаций, распознавание биомедицинских объектов, выявление функциональных взаимосвязей между ними и их нормализацию с использованием внешних онтологий и фактографических ресурсов, что формирует структурированное представление знаний непосредственно из текстов.

Апробация подхода на коллекциях текстов различного объема и тематики показала высокую точность извлечения ассоциаций, устойчивость к изменению предметной области и воспроизводимость значительной части сведений, содержащихся в существующих базах данных, при одновременном увеличении набора фиксируемых фактов за счёт информации из публикаций.

Структурированные результаты могут представляться в табличном и графовом формате, что облегчает анализ взаимосвязей между объектами, выявление недостающих элементов молекулярных процессов и формирование новых гипотез о механизмах действия соединений. Разработанный метод обеспечивает практическое применение полученных результатов в биоинформатике, медицинской химии и исследовании молекулярных механизмов заболеваний.

5. ВЫВОДЫ

1. Создана коллекция текстов для извлечения релевантной структурированной информации о наименованиях низкомолекулярных органических соединений и данных об их биологической активности.
2. Разработан и протестирован интегральный подход извлечения знаний, с помощью которого было извлечено более трех миллионов уникальных ассоциаций между наименованиями низкомолекулярных органических соединений, белками, генами, миРНК. Разработанный подход позволяет извлекать сведения о биологической активности низкомолекулярных органических соединений на основе интеллектуального анализа текстов.
3. На основе выявленных ассоциаций создана база данных о низкомолекулярных органических соединениях, включая наименования, синонимы, и ассоциированные с ними виды биологической активности.
4. Проведена проверка согласованности и полноты информации о биологической активности низкомолекулярных соединений, извлекаемой из текстов научных публикаций. В результате было выявлено, что разработанный метод позволяет извлекать значительную часть информации, содержащейся в базе данных ChEMBL (v.34), а также выявлять новые взаимосвязи, отсутствующие в этой базе данных.

6. СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Tarasova O.A., **Biziukova N.Y.**, Rudik A.V., Dmitriev A.V., Filimonov D.A., Poroikov V.V. Extraction of Data on Parent Compounds and Their Metabolites from Scientific Abstracts // Journal of Chemical Information and Modeling. – 2021. – Т. 61. – № 4. – С. 1683–1690.
2. **Biziukova N.Yu.**, Tarasova O.A., Rudik A.V., Filimonov D.A., Poroikov V.V. Automatic Recognition of Chemical Entity Mentions in Texts of Scientific Publications // Automatic Documentation and Mathematical Linguistics. – 2020. – Т. 54. – № 6. – С. 306.
3. **Biziukova N.Yu.**, Ivanov S.M., Tarasova O.A. Identification of Proteins and Genes Associated with Hedgehog Signaling Pathway Involved in Neoplasm Formation Using

Text-Mining Approach // Big Data Mining and Analytics. – 2024. – Т. 7. – № 1. – С. 107–130.

4. Тарасова О.А., **Бизюкова Н.Ю.**, Столбова Е.А., Столбов Л.А., Такташов Р.Р., Карасев Д.А., Ионов Н.С., Иванов С.М., Дмитриев А.В., Рудик А.В., Дружиловский Д.С., Соболев Б.Н., Филимонов Д.А., Поройков В.В. Извлечение информации о взаимодействии вирусов с организмом человека и о противовирусных соединениях на основе интеллектуального анализа больших коллекций текстов // Биомедицинская химия. 2024. – Т. 70. – № 6. – С. 469-474.
5. Тарасова О.А., **Бизюкова Н.Ю.**, Иванов С.М., Поройков В.В. Выявление взаимосвязей между воспалительным процессом при инфекционных заболеваниях и развитием депрессии на основе интеллектуального анализа текстов // Иммунопатология, аллергол, инфектол. 2024. – Т.4. – С. 63-69.
6. **Biziukova N.Y.**, Rudik A.V., Dmitriev A.V., Tarasova O.A., Filimonov D.A., Poroikov V.V. XenoMet: A Corpus of Texts to Extract Data on Metabolites of Xenobiotics // ACS Omega. – 2025. – Т. 10. – № 3. – С. 2459–2471.
7. **Бизюкова Н.Ю.**, Тарасова О.А., Филимонов Д.А., Поройков В.В.. Автоматизированное извлечение наименований биологически активных соединений из текстов научных публикаций // XXVI Симпозиум «Биоинформатика и компьютерное конструирование лекарств». Тезисы докладов. – Москва, Россия, 2020. – С. 53.
8. **Бизюкова Н.Ю.**, Иванов С.М., Тарасова О.А. Идентификация генов, функционально значимых для прогрессии ВИЧ-инфекции на основе интеллектуального анализа текстов // XXVII Симпозиум «Биоинформатика и компьютерное конструирование лекарств». Сборник научных трудов. – Москва, Россия, 2021. – С. 53.
9. **Biziukova N.Yu.**, Tarasova O.A., Filimonov D.A., Poroikov V.V. Identification of virus-host interaction mechanisms based on text-mining approach // XXVIII Symposium «Bioinformatics and Computer-Aided Drug Discovery». Proceedings book. – Moscow, Russia, 2022. – P. 66.
10. **Бизюкова Н.Ю.**, Тарасова О.А. Извлечение наименований низкомолекулярных органических соединений и белков из текстов научных публикаций // VI Междисциплинарная конференция «Молекулярные и биологические аспекты

химии, фармацевтики и фармакологии». Тезисы докладов. – Москва, Россия, 2021. – С. 13.

11. **Бизюкова Н.Ю.**, Тарасова О.А., Поройков В.В. Идентификация потенциальных белков-мишеней организма человека для оптимизации терапии ВИЧ-инфекции на основе интеллектуального анализа текстов // 5-я Российская конференция по медицинской химии с международным участием «МедХим-Россия 2021». Материалы конференции. – Москва, Россия, 2021. – Т. 1. – С. 179.
12. **Biziukova N.**, Ivanov S., Filimonov D., Tarasova O., Poroikov V. Identification of virus-host interaction mechanisms based on text-mining approach // BGRS/SB-2022: The Thirteenth International Multiconference. Abstracts. – Novosibirsk, Russia, 2022. – С. 341-342.
13. **Biziukova N.**, Sobolev B., Karasev D., Ionov N., Sukhachev V., Taktashov R., Rudik A., Ivanov S., Tarasova O. Application of text mining methods to extract comprehensive information about the biological activity of drugs: case-study for antiviral compounds // BGRS/SB-2024: The Fourteenth International Multiconference. Abstracts. – Novosibirsk, Russia, 2024. – С. 625–630.
14. **Бизюкова Н.Ю.**, Тарасова О.А. Извлечение всесторонних сведений о спектре биологической активности противовирусных соединений из больших массивов текстов научных публикаций // VI Международная конференция ПОСТГЕНОМ’2024. XI Российский симпозиум «Белки и пептиды». Российско-китайский конгресс по наукам о жизни. Сборник тезисов. – Москва, Россия, 2024. – С. 81.
15. **Бизюкова Н.Ю.**, Тарасова О.А. Формирование знаний о совокупности молекулярных механизмов потенциальных противовирусных соединений // Современные вызовы молекулярной биологии. Материалы II школы молодых учёных. – Шерегеш, Россия, 2025. – С. 22.
16. **Бизюкова Н.Ю.**, Филимонов Д.А., Поройков В.В., Тарасова О.А. Интеллектуальный анализ больших массивов текстов с целью формирования знаний о биологической активности низкомолекулярных химических соединений // XXVI Харитоновские тематические научные чтения «Искусственный интеллект и большие данные в технических, промышленных,

природных и социальных системах». Материалы конференции. – Саров, Россия, 2025. – С. 25–26.

17. **Бизюкова Н.Ю.** Разработка мультилингвальной модели для извлечения структурированных данных для биологии и медицины из больших массивов текстов // VI Международная научная конференция и X Всероссийский молодёжный научный форум «Наука будущего – наука молодых». Сборник тезисов. – Саратов, Россия, 2025. – С. 130.

7. ФИНАНСИРОВАНИЕ

Программа фундаментальных научных исследований в Российской Федерации на долгосрочный период (2021–2030 гг.) (№ 124050800018-9); Программа фундаментальных научных исследований в Российской Федерации на долгосрочный период (2021–2030 гг.) (№ 122030100170-5); проект по созданию и развитию научных центров мирового уровня «Цифровой дизайн и персонализированное здравоохранение» при финансовой поддержке Министерства образования и науки Российской Федерации (соглашение № 075-15-2022-305); грант Российского научного фонда № 24 25 00453.

8. БЛАГОДАРНОСТИ

Автор выражает глубокую признательность научному руководителю, Ольге Александровне Тарасовой, за постоянную поддержку, ценные рекомендации и внимание к работе на всех этапах её выполнения. Особая благодарность научному консультанту, Владимиру Васильевичу Поройкову, за консультации, профессиональные советы и помощь в формировании научных результатов. Автор также благодарит коллег, принимавших участие в обсуждении, получении и анализе данных, за содействие и конструктивное взаимодействие в ходе подготовки диссертационной работы.