ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ НАУЧНОЕ УЧРЕЖДЕНИЕ «НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ БИОМЕДИЦИНСКОЙ ХИМИИ ИМЕНИ В.Н. ОРЕХОВИЧА»

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

Диссертация

на соискание ученой степени кандидата биологических наук

Научный руководитель:

кандидат биологических наук
Тарасова Ольга Александровна

Научный консультант:

академик РАН, профессор,
кандидат физико-математических наук,
доктор биологических наук
Поройков Владимир Васильевич

Москва - 2025

СОДЕРЖАНИЕ

СОДЕРЖАНИЕ 2
СПИСОК СОКРАЩЕНИЙ 6
ВВЕДЕНИЕ 9
ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ 17
1.1. Формирование выборки релевантных текстов для анализа 18
1.1.1. Источники биомедицинских текстов 18
1.1.2. Определение релевантности текстов для поставленной цели исследования 20
1.2. Предобработка текстов 26
1.2.1. Подготовка текста 27
1.2.2. Токенизация 28
1.2.3. Нормализация 30
1.3. Обзор инструментов для обработки естественного языка 31
1.3.1 Natural Language Toolkit (NLTK) 33
1.3.2. spaCy 35
1.3.3. Natasha 36
1.3.4. DeepPavlov 37
1.4. Распознавание наименований объектов 38
1.4.1. Метрики для оценки точности алгоритмов интеллектуального анализа текстов 39
1.4.2. Алгоритмы распознавания наименований на основе словарей 40
1.4.3. Алгоритмы распознавания наименований на основе правил 42

1.4.4. Алгоритмы распознавания наименований на основе методов машинного обучения 43

1.4.4.1. Корпусы и наборы данных для обучения 44

1.4.4.2. Методы машинного обучения, используемые для распознавания наименований объектов 46

1.5. Алгоритмы извлечения ассоциаций между сущностями 49

1.5.1. Алгоритмы извлечения ассоциаций, основанные на правилах 49

1.5.2. Алгоритмы извлечения ассоциаций, основанные на методах машинного обучения 51

1.5.2.1. Примеры корпусов, которые используются для построения моделей извлечения ассоциаций из текстов 51

1.5.2.2. Примеры методов машинного обучения, которые используются для построения моделей извлечения ассоциаций из текстов 53

ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ 56

2.1. Создание коллекции текстов с информацией о наименованиях низкомолекулярных органических соединений и данными об их биологической активности 56

2.1.1. Стратегия формирования репрезентативной коллекции текстов 56

2.1.1.1. Метод итеративного поиска с использованием MeSH-терминов 56

2.1.1.2. Метод машинного обучения для категоризации текстов 57

2.1.1.3. Мультиклассовая и многометочная категоризация на основе базы данных ChEMBL 58

2.2. Разработка интегрального подхода для извлечения сведений о биологической активности низкомолекулярных органических соединений из текстов 58

2.2.1. Формирование выборки релевантных текстов 59

2.2.2. Распознавание наименований объектов 61

2.2.3. Извлечение ассоциаций 66

2.3. Создание структуры базы данных для хранения извлеченных сведений о биологической активности низкомолекулярных органических соединений 69

2.4. Валидация разработанных алгоритмов интеллектуального анализа текстов в задаче поиска сведений о химических соединениях, перспективных с точки зрения противовирусной терапии 70

2.5. Проверка полноты сведений о биологической активности химических соединений в фактографических базах данных и базах знаний 71

ГЛАВА 3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ 73

3.1. Коллекции текстов, содержащих знания о биологической активности низкомолекулярных химических соединений 73

3.1.1. Применение методов фильтрации текстов для отбора релевантных публикаций 73

3.1.1.1. Использование автоматизированных запросов к PubMed 74

3.1.1.2. Применение алгоритмов машинного обучения для классификации текстов 79

3.1.2. Специализированные коллекции текстов 85

3.2. Интегральный подход для извлечения сведений о биологической активности низкомолекулярных органических соединений из текстов 88

3.2.1. Распознавание наименований 90

3.2.2. Извлечение ассоциаций 94

3.2.3. Форматы представления результатов извлечения знаний о биологической активности низкомолекулярных органических соединений 101

3.2.4. Применение разработанного интегрального подхода для извлечения информации из биомедицинских текстов 107

3.2.4.1. Исследование роли сигнального пути Hedgehog в патогенезе неопластических заболеваний 107

3.2.4.2. Анализ механизмов большого депрессивного расстройства и возможных методов терапии 108

3.3. База данных о наименованиях низкомолекулярных органических соединений и ассоциированных с ними видах биологической активности 110

3.3.1. Логическая структура базы данных 110

3.3.2. Доступ к базе данных через веб-интерфейс 113

4. Проверка согласованности и полноты информации о биологической активности низкомолекулярных органических соединений. 116

ЗАКЛЮЧЕНИЕ 120

ВЫВОДЫ 122

ФИНАНСИРОВАНИЕ РАБОТЫ 123

СПИСОК ЛИТЕРАТУРЫ 124

Приложение 1 142

Приложение 2 150

СПИСОК СОКРАЩЕНИЙ

Аббревиатура	Расшифровка
ABCA1	ATP-binding cassette sub-family A member 1
ACS	American Chemical Society
ADME	Absorption, Distribution, Metabolism, Excretion
AKT1	AKT serine/threonine kinase 1
API	Application Programming Interface
ATP	Adenosine triphosphate
BAO	BioAssay Ontology
BEL	Biological Expression Language
BERT	Bidirectional Encoder Representations from Transformers
BRCA1	Breast cancer type 1 susceptibility protein
CD14	Cluster of Differentiation 14
CDR	Chemical–Disease Relation corpus
CHEMDNER	Chemical Compound and Drug Name Recognition corpus
CNIPA	China National Intellectual Property Administration
CNN	Convolutional Neural Network
CORD-19	COVID-19 Open Research Dataset
COVID-19	Coronavirus disease 2019
CRF	Conditional Random Field
DDI	Drug–Drug Interaction
EGFR	Epidermal Growth Factor Receptor
EPO	European Patent Office
EU-ADR	European Adverse Drug Reactions project
FOXP3	Forkhead box P3
GENIA	Genomics Informatics corpus
GNN	Graph Neural Network

GPT	Generative Pretrained Transformer
GPU	Graphics Processing Unit
GRB2	Growth factor receptor-bound protein 2
HCK	Hematopoietic cell kinase
HCQ	Hydroxychloroquine
HGNC	HUGO Gene Nomenclature Committee
HIV	Human Immunodeficiency Virus
HLA	Human Leukocyte Antigen
HTML	HyperText Markup Language
ICAM1	Intercellular Adhesion Molecule 1
ID	Identifier
IDF	Inverse Document Frequency
JNLPBA	Joint NLP in Biomedicine and its Applications corpus
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	k-Nearest Neighbors
LDA	Latent Dirichlet Allocation
LSTM	Long Short-Term Memory
LSTM-CRF	Long Short-Term Memory – Conditional Random Field
MAPK3	Mitogen-Activated Protein Kinase 3
MEDLINE	Medical Literature Analysis and Retrieval System Online
MIMIC	Medical Information Mart for Intensive Care
MTOR	Mechanistic Target of Rapamycin
NCBI	National Center for Biotechnology Information
NER	Named Entity Recognition
NLM	National Library of Medicine
NLP	Natural Language Processing
NLTK	Natural Language Toolkit

NS	Not Specified
OCR	Optical Character Recognition
PMC	PubMed Central
PMID	PubMed Identifier
POS	Part-of-Speech
RE	Regular Expression
RF	Random Forest
RNN	Recurrent Neural Network
SNP	Single Nucleotide Polymorphism
SOBIE	Single-Out-Begin-Inside-End
SVM	Support Vector Machine
TF-IDF	Term Frequency – Inverse Document Frequency
UMLS	Unified Medical Language System
USPTO	United States Patent and Trademark Office
WIPO	World Intellectual Property Organization
АПФ	Ангиотензин-Превращающий Фермент
БД	База Данных
ДНК	Дезоксирибонуклеиновая Кислота
НМОС	Низкомолекулярное Органическое Соединение
РНК	Рибонуклеиновая Кислота
ХС	Химическое Соединение
ЭПМЗ	Электронная Персональная Медицинская Запись

ВВЕДЕНИЕ

Актуальность избранной темы. Большая часть исследований в области биологии и медицины сопровождаются предварительным анализом литературных источников. Это необходимо для учета преемственности знаний и дальнейшего развития в конкретной предметной области. Подобный анализ может занимать значительную часть времени всего исследования, особенно в случаях, когда поставлена задача системного анализа отдельных функций биологических организмов.

Развитие науки в области биологии и медицины привело к появлению различных междисциплинарных областей знаний, таких как биоинформатика, биотехнология, а углубление знаний привело к развитию узких специализаций (гистология, эмбриология, геронтология и др.). Закономерным результатом стал и рост количества научных рецензируемых журналов, которые стремятся покрыть все области знания человека о биологических системах. Анализ динамики появления новых научных журналов в области биологии и медицины демонстрирует исключительно высокие темпы развития этих дисциплин. Согласно данным SCImago Journal Rank [1], ежегодно количество журналов в период с 2015 по 2024 год увеличивалось: от 395 (в 2015 году) до 56 (в 2024 году) (рисунок 1). Особенно показателен рост в 2019-2021 годах – период пандемии COVID-19, когда ежегодно появлялось от 459 до 1438 новых журналов. Такие темпы могут указывать на формирование новых специализаций и междисциплинарных направлений исследований, а также высокий интерес к определенной предметной области.

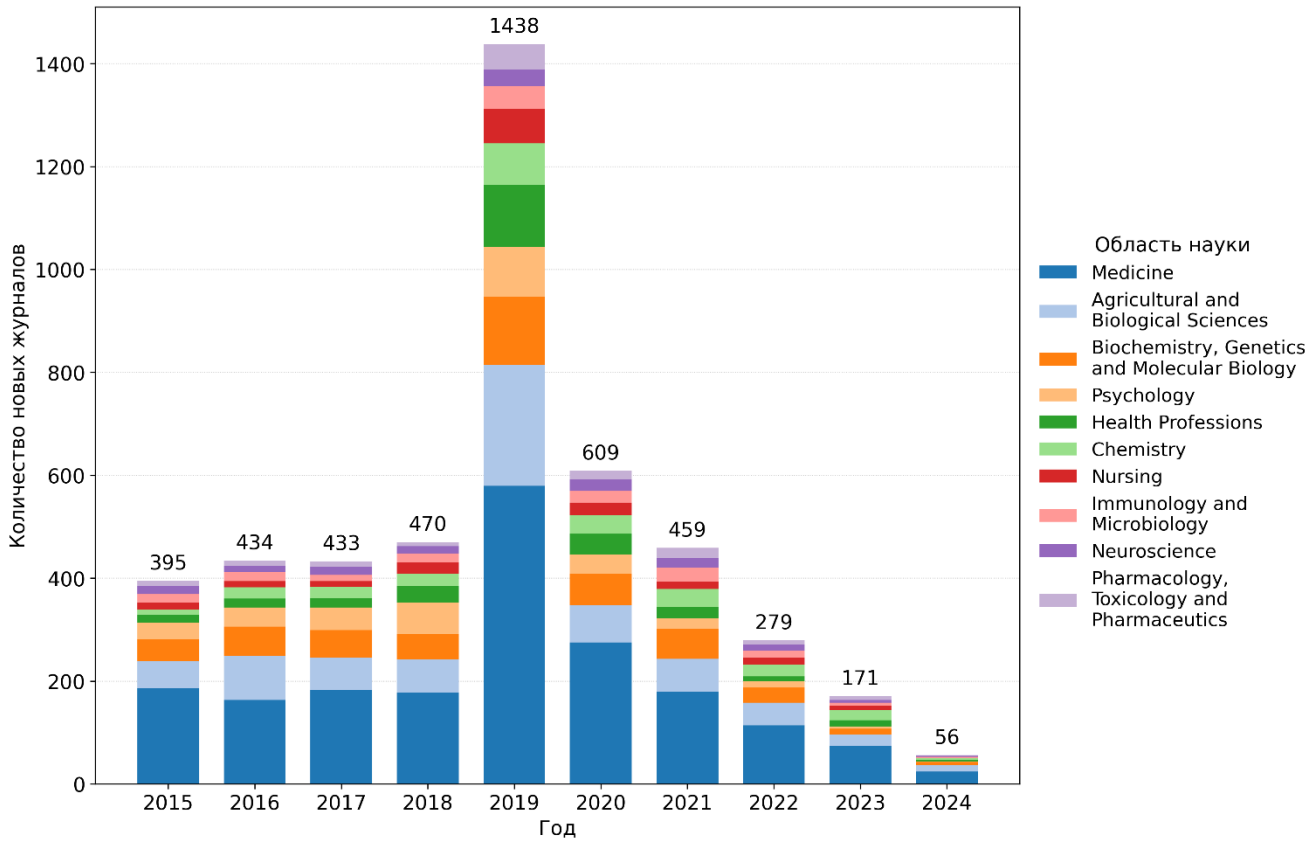

Рисунок 1 – Динамика появления новых рецензируемых журналов в биомедицинских науках в 2015–2024 гг. с разбивкой по ключевым направлениям.

Быстрый рост количества научных изданий формирует принципиально новые вызовы для исследователей. Если для глубокого изучения проблематики ранее было достаточно анализа нескольких десятков профильных журналов, то сегодня даже простое перечисление всех новых изданий, появившихся за последнее десятилетие, является непростой задачей. В таких условиях традиционные методы работы с научной литературой становятся неэффективными, а риск невозможности исследования значимых публикаций или дублирования уже проведенных исследований существенно возрастает.

Особую актуальность в этой связи приобретает разработка алгоритмов автоматизированного анализа научных текстов. Современные методы машинного обучения и обработки естественного языка позволяют не только оперативно выявлять релевантные публикации среди тысяч новых источников, но и

отслеживать эволюцию научных концепций, выявлять междисциплинарные связи, прогнозировать перспективные направления исследований и систематизировать научные знания [2]. Создание таких аналитических инструментов становится важным условием для поддержания высокой эффективности исследовательской работы.

Одной из областей применения таких интеллектуальных систем, имеющей высокую практическую значимость, является извлечение и систематизация знаний о биологической активности низкомолекулярных химических соединений. Эта задача лежит в основе современных исследований в области фармакологии, токсикологии и разработки лекарственных средств. Процесс создания нового препарата характеризуется исключительной длительностью в десятки лет работы и высокой стоимостью в миллиарды долларов [3, 4]. Значительная часть этих издержек связана со всесторонним изучением биологических эффектов химических соединений, включая их механизмы действия, метаболизм, токсичность и потенциальную терапевтическую эффективность.

Методы интеллектуального анализа текстов выступают ключевым инструментом для агрегации и структурирования информации, распределенной по миллионам публикаций. Важным практическим результатом применения этих методов может выступать возможное сокращение временных и финансовых затрат на доклинические и клинические исследования, связанное с отсутствием необходимости проведения дополнительных тестов в условиях наличия полной информации об исследованной биологической активности химического соединения другими авторами [3, 4]. Автоматизированный анализ научной литературы позволяет исследователям в сжатые сроки получить наиболее полную картину об уже изученных свойствах соединения, избегая повторения проведенных исследований и сосредотачивая экспериментальные работы на наиболее перспективных и малоизученных аспектах проблемы.

В условиях стремительного роста объема научных публикаций и необходимости систематизации знаний о биологической активности

низкомолекулярных соединений возникает потребность в разработке специализированных методов интеллектуального анализа текстов. Решение этой задачи позволит автоматизировать процесс выявления связей между химическими структурами и их биологическими эффектами, что особенно важно для целей разработки лекарств.

Степень разработанности темы. Задачи автоматизированного анализа биомедицинских текстов активно развиваются в последние два десятилетия. Созданы корпусы и наборы данных для распознавания биомедицинских сущностей и отношений (CHEMDNER, CDR, DrugProt, BioRED), разработаны методы на основе правил, условных случайных полей (Conditional Random Fields, CRF), нейросетевые архитектуры и современные трансформерные модели. Эти инструменты доказали свою эффективность в отдельных задачах, однако большинство из них ориентированы на ограниченные типы извлекаемых сущностей или отношений и не обеспечивают интегрального подхода к систематизации знаний о биологической активности низкомолекулярных соединений. Таким образом, проблема автоматизации установления ассоциаций между химическими структурами и их биологическими эффектами остаётся актуальной и требует разработки новых методов.

В связи с этим в настоящей диссертационной работе поставлена цель – разработка, реализация и тестирование метода для извлечения взаимосвязей между наименованиями низкомолекулярных органических соединений и их известной биологической активностью.

Для достижения поставленной цели были сформулированы 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.

Научная новизна

В работе предложен интегральный метод автоматизированного анализа биомедицинских текстов, обеспечивающий системное выявление и структурирование сведений о биологической активности низкомолекулярных органических соединений. Разработанный метод отличается достаточной точностью при сравнительно низких требованиях к вычислительным ресурсам, а получаемые результаты являются интерпретируемыми и удобными для последующего анализа специалистами.

Новизна работы заключается также в создании системы, обладающей возможностью масштабирования и адаптации к новым биомедицинским задачам. Такой подход обеспечивает расширяемость базы данных за счёт автоматического пополнения и позволяет применять разработанные методы не только для анализа сведений о низкомолекулярных органических соединениях, но и в смежных областях биомедицинских исследований.

Теоретическая и практическая значимость

Разработанные методы интеллектуального анализа текстов позволяют существенно сократить временные и финансовые затраты на поиск информации и дальшейшее исследование биологической активности соединений.

Особое значение эти методы приобретают для решения задачи репозиционирования лекарственных средств [5]. Алгоритмы анализа текстов способствуют выявлению ранее неизвестных видов биологической активности уже одобренных препаратов и обнаруживанию новых показаний к применению для существующих молекул [6]. Такой подход позволяет существенно сократить сроки и риски вывода нового лекарства на рынок, поскольку его фармакокинетика и профиль безопасности уже хорошо изучены.

Кроме того, интеллектуальный анализ больших массивов публикаций позволяет выявлять неочевидные связи между наименованиями химических соединений, биологическими мишенями и патологическими процессами, что служит основой для генерации новых научных гипотез. Также, алгоритмы интеллектуального анализа текстов могут позволить проводить исследования, схожие по принципу с мета-анализами, только в качестве объекта будут выступать не структурированные данные, а научные публикации [7]. Это позволяет повысить достоверность знания путем агрегации малочисленных результатов.

Практическая значимость работы заключается также в создании базы данных, которая может применяться в фармакологии, молекулярной биологии и смежных областях. Полученные результаты полезны как для фундаментальных исследований, так и для прикладных задач, включая разработку новых терапевтических стратегий.

Кроме того, разработанная система может использоваться при обучении работе с большими массивами данных и в качестве демонстрации возможностей современных методов анализа научных текстов.

Методология и методы диссертационного исследования

Методологическую основу работы составили современные подходы к обработке естественного языка и машинному обучению. В качестве методов исследования использовались: анализ и систематизация литературных источников; формирование корпуса биомедицинских текстов; алгоритмы машинного обучения (CRF, нейронные сети, трансформеры); применение словарей и онтологий (MeSH,

UMLS, ChEBI и др.); вычислительные эксперименты по валидации результатов (оценка точности, полноты, воспроизводимости); сопоставление извлечённых ассоциаций с биологически известными фактами.

Личный вклад автора

Автором самостоятельно проведён анализ литературных источников и сформирована коллекция биомедицинских текстов для последующего исследования. Разработаны алгоритмы интегрального метода автоматизированного извлечения знаний о биологической активности низкомолекулярных органических соединений, реализована и протестирована их работа на репрезентативных корпусах текстов.

Автор участвовал в построении базы данных, обеспечивающей хранение и структурирование полученной информации, а также в разработке инструментов, обеспечивающих практическое применение предложенного подхода. Автором проведена серия вычислительных экспериментов, включая оценку точности, полноты и воспроизводимости полученных результатов, а также их интерпретацию с точки зрения биологической значимости.

Положения, выносимые на защиту:

- Разработан интегральный метод извлечения ассоциаций между наименованиями низкомолекулярных органических соединений и видами их биологической активности, который обеспечивает высокую точности и полноту извлекаемой информации.
- С применением разработанного метода создана база данных о 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.

Степень достоверности результатов

Достоверность научных результатов обеспечивается репрезентативностью использованных корпусов текстов (PubMed, PMC и др.), применением предметно-ориентированных онтологий и профильных баз данных , корректностью математических методов анализа, многократной проверкой алгоритмов на независимых выборках, а также согласованностью полученных результатов с данными экспериментальных и клинических исследований.

Апробация работы. Основные положения диссертации были представлены на российских и международных конференциях и симпозиумах: XXVI, XXVII, XXVIII Symposia «Биоинформатика и компьютерное конструирование лекарств» (онлайн, 2020, 2021, 2022), Междисциплинарная конференция «Молекулярные и биологические аспекты химии, фармацевтики и фармакологии» (МОБИ-ХимФарма2020 VI, Нижний Новгород, 2020), Российская конференция MedChem-Russia (Волгоград, 2021), XIII и XIV International Multiconferences «Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2022, BGRS/SB-2024)» (Новосибирск, 2022, 2024), ACS Fall (Сан-Франциско, виртуально, 2023), VI Международная конференция «ПОСТГЕНОМ’2024» (Московская область, 2024), II Школа молодых учёных (Шерегеш, 2025), XXVI Харитоновские тематические научные чтения (Саров, 2025).

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ

Основная цель разработки алгоритмов в области интеллектуального анализа текстов заключается в извлечении релевантной структурированной информации из неструктурированных данных [8]. Как правило, такие алгоритмы включают в себя несколько этапов.

Первый этап – формирование выборки релевантных текстов для анализа. Он необходим для ограничения предметной области исследования, обеспечения достоверности последующего анализа и формирования репрезентативной базы данных.

Второй этап – распознавание наименований биологических и химических объектов, относящихся к выбранной области исследования. Это могут быть наименования химических соединений и заболеваний, белков/генов, микроРНК (miRNA) или клеточных линий, биологических видов или однонуклеотидных полиморфизмов [9].

Третий этап заключается в поиске ассоциаций между распознанными объектами. Он подразумевает автоматическое установление наличия или отсутствия семантической взаимосвязи между объектами, отражённой в тексте авторами. Эти взаимосвязи могут быть как бинарными (например, «есть/нет связь»), так и многоклассовыми, где каждый класс характеризует определённый тип взаимодействия между объектами [10].

Последний этап связан с обработкой извлечённой информации. В зависимости от поставленных задач он может включать фильтрацию и оценку достоверности найденных ассоциаций, соотнесение объектов с базами данных и онтологиями, а также представление результатов в удобной форме – например, графической или табличной. Процесс интеллектуального анализа текстов представляет собой многоступенчатую систему, где каждый этап оказывает влияние на полноту и точность конечного результата.

1.1. Формирование выборки релевантных текстов для анализа

1.1.1. Источники биомедицинских текстов

Первоисточниками биомедицинских знаний являются тексты научных публикаций, электронные истории болезни, патенты, а также сообщения пользователей и пациентов в социальных сетях [11-15].

Научные публикации представляют собой универсальный источник знаний, поскольку многие из них содержат экспериментально подтверждённые результаты в различных областях науки. Для поиска публикаций наиболее часто применяется библиографическая база данных PubMed [16]. Ограниченный доступ к полнотекстовым публикациям компенсируется использованием аннотаций, где обычно отражены ключевые результаты работы [14]. Для полнотекстового анализа используется библиографическая база данных PubMed Central (PMC) [17], где статьи представлены в гипертекстовом формате, что упрощает автоматическую обработку. Для препринтов применяются bioRxiv [18] и medRxiv [19], что особенно важно при исследовании новых направлений: несмотря на отсутствие рецензирования, такие ресурсы обеспечивают более быстрый обмен информацией между исследователями.

Ценным источником знаний о структуре и биологической активности химических соединений являются патенты. Их автоматизированный анализ осложнён необходимостью обработки изображений химических структур, а также особенностями представления в виде так называемых «формул Маркуша», что позволяет описывать целые классы соединений, но затрудняет идентификацию отдельных структур [15]. В литературе отмечается, что лишь около 6% биологически активных соединений, упомянутых в патентной документации, также встречаются и в научных публикациях. Это указывает на то, что патенты и статьи во многом содержат разнородные и взаимодополняющие сведения, и анализ этих типов источников позволяет формировать более полное представление о биологической активности соединений [15].

Хотя патентные базы данных, такие как WIPO (World Intellectual Property Organization, Всемирная организация интеллектуальной собственности) [20], предоставляют значительный объём патентной информации в глобальном масштабе, специализированные национальные и региональные ресурсы позволяют получить более детализированные данные. Так, в США используется база United States Patent and Trademark Office (USPTO) [21], в БД Европейском Союзе – European Patent Office (EPO) [22], в Китае – БД China National Intellectual Property Administration (CNIPA) [23], в России – БД Роспатента [24], и многие другие. Поиск в патентных офисах отдельных стран может требовать значительное количество времени, поэтому были разработаны агрегаторы, такие как Google Patents [25], объединяющий данные USPTO, EPO и WIPO, или коммерческие решения, например, Derwent Innovation [26], которые, помимо собственно патентной информации, предоставляют также аналитические инструменты для её обработки и интерпретации.

Социальные сети становятся дополнительным источником информации, ценным для здравоохранения. Мониторинг сообщений позволяет выявлять случаи злоупотребления психоактивными веществами [27], фиксировать факторы риска (курение, переедание, злоупотребление алкоголем) [28], а также отслеживать сведения о побочных эффектах лекарств и вакцин, чему уделялось особое внимание в период пандемии COVID-19 [29]. Наиболее часто в исследованиях используется Twitter (X) [29-32].

Электронные персональные медицинские записи (ЭПМЗ, EHRs), являющиеся частью электронных историй болезни, отражают не только назначаемое лечение и побочные эффекты, но и персональные особенности пациентов [33, 34]. Возросший интерес к анализу ЭПМЗ [11, 35] свидетельствует о смещении акцентов с выявления общих закономерностей на персонализированный подход. Для свободного использования доступны обезличенные наборы, такие как MIMIC-IV [36] и eICU Collaborative Research Database [37, 38].

Протоколы и отчёты клинических исследований также представляют важный источник информации о новых терапевтических подходах и тенденциях в медицине. Наиболее известные базы – ClinicalTrials.gov [39], EU Clinical Trials Register [40] и WHO ICTRP [41].

Значимым источником знаний являются клинические руководства и отчёты регуляторов (ВОЗ, FDA, EMA). В частности, FDA Drug Labels (DailyMed, [42]) и European Medicines Agency Assessment Reports [43] содержат сведения о стандартах лечения, безопасности препаратов и побочных эффектах.

1.1.2. Определение релевантности текстов для поставленной цели исследования

В зависимости от поставленной цели исследования различаются и наборы текстов для анализа. Область исследования может быть ограничена как исследуемым объектом или процессом (например, симптомы отдельного заболевания, или функциональная роль и биохимическая значимость отдельного белка), так и областью и методологией исследования (например, биологическая активность химических соединений в отношении вирусов, продемонстрированная в *in vivo* экспериментах; побочные эффекты лекарственных препаратов, описанные в результатах клинических испытаний).

Ограничение цели исследования и четкая формализация критериев релевантности текстов позволяет сосредоточиться на конкретной предметной области и снизить количество ошибок даже при использовании относительно простых методов [44]. Для достижения максимально качественного результата необходима не только кропотливая экспертная работа, но и применение инструментов, позволяющих стандартизировать и частично автоматизировать процесс отбора – в частности, использование MeSH-терминов [45], уточнение по типам публикаций и формализация поисковых запросов. Это делает процедуру более воспроизводимой и снижает вероятность субъективных и случайных погрешностей.

Поскольку тексты научных публикаций являются наиболее универсальным источником знаний, в дальнейшем литературный обзор будет сосредоточен только на особенностях работы именно с ними.

Самым простым решением для формирования выборки текстов для анализа является использование уже готовых наборов (корпусов) текстов, которые отобраны экспертами вручную или полуавтоматически в соответствии с определенной темой. Примером является набор полных текстов (более 400 000) и резюме (более 1 млн) научных публикаций, посвященных исследованию различных аспектов патологии COVID-19 и возбудителя SARS-CoV-2 – CORD-19 (COVID-19 Open Research Dataset) [46, 47]. Однако при работе с готовыми корпусами текстов следует учитывать, что они являются статичными или только периодически обновляемыми, что ограничивает возможности работы с вновь опубликованной информацией. Так, например, последнее обновление корпуса CORD-19 датируется июнем 2022 года, а в период с 2023 по 2025 год только в PubMed были представлены еще более 150 000 тысяч работ по данной тематике.

Ввиду ограничений готовых корпусов, в ряде случаев оптимальным выбором для формирования коллекции релевантных текстов станет прямая работа с библиографическими базами данных.

Наиболее простым способом в данном случае будут являться автоматизированные запросы к библиографическим базам данных на основе набора ключевых слов. Подобный поиск не позволяет получить строго релевантную заданной тематике выборку текстов, однако с помощью него можно произвести разведочный анализ и сформулировать более жесткие критерии релевантности для последующего отбора.

Еще одним распространенным и несложным способом формирования выборки релевантных текстов является использование надстроек к поиску, доступных в самой базе данных библиографической информации. Так, например, для PubMed можно использовать MeSH-термины [45] и типы публикаций. Стоит отметить, что MeSH-термины являются контролируемым словарем

биомедицинской терминологии, а статьи аннотируются с их использованием вручную, что повышает точность отбора релевантных публикаций в сравнении с обычными ключевыми словами. Типы публикаций позволяют сфокусироваться на отдельных форматах (видах) исследования: например, на клинических испытаниях, или только экспериментальных данных – исключая литературные обзоры, главы из книг, и другие. Дополнительные возможности для систематического поиска предоставляет инструмент OpenAlex [48, 49], который интегрирует метаданные из широкого круга источников и позволяет использовать гибкие фильтры для отбора научных публикаций, включая дисциплинарную принадлежность, цитируемость и институциональную принадлежность.

Требующая значительных временных затрат группа методов – это алгоритмы на основе машинного обучения. Как в биологии и медицине, так и во многих других областях, работа с алгоритмами машинного обучения подразумевает реализацию нескольких этапов: формирование обучающей выборки (не обязательно, если речь идет про алгоритмы обучения без учителя), описание объектов в выборке с помощью набора признаков, выбор метода машинного обучения и валидация, формирование окончательной модели для прогноза необходимых признаков свойств объектов, и непосредственно сам прогноз. Рассмотрим последовательно каждый из этапов, применяемых для анализа текстов.

В случае, если используются алгоритмы машинного обучения с учителем, необходимо сформировать обучающую выборку - набор объектов, обладающими известными свойствами. В качестве объекта выступает биомедицинский текст, а в качестве свойства – в простейшем варианте – категории «релевантный» и «не релевантный». Здесь же сразу необходимо обозначить ограничение такого подхода: для формирования качественной обучающей выборки требуется значительная по объему ручная работа, поскольку для точной предсказательной модели необходим большой набор аннотированных данных.

Применение методов машинного обучения без учителя сводится к попытке структурировать весь объем текстов по их признакам в соответствии с

определенным количеством категорий, которые не установлены заранее; более того, может не быть известно и количество этих категорий. Такой подход основан на определении наиболее схожих по признакам текстов (кластеризация), и чаще всего используется в комбинации с другими подходами (например, поиск по ключевым словам/MeSH-терминам для ограничения области поиска – кластеризация с дальнейшей ручной верификацией содержимого кластеров и их описанием – классификация в соответствии с полученными кластерами с применением методов машинного обучения с учителем) [50].

Еще одним способом, схожим с кластеризацией текстов, является метод тематического моделирования. Наиболее часто использующийся для данной задачи подход – это латентное размещение Дирихле (Latent Dirichlet Allocation, LDA) [51], статистический подход, который подразумевает построение распределения вероятностей над словами, встречающимися в коллекции текстов, и выявления наиболее значимых из них – так называемых, скрытых тем. Этот метод не позволяет достичь выявления узких тематик научных текстов, однако дает возможность сузить область поиска, структурируя большие массивы документов по конкретным темам [52].

С развитием нейросетевых технологий появились более продвинутые методы тематического моделирования. Например, метод BERTopic является комбинацией нескольких подходов к работе с текстами [53]. Первым этапом является построение векторных представлений документов коллекции и выделение кластеров, наиболее схожих по этому признаку. Затем рассчитывается метрика TFIDF (term frequency–inverse document frequency) для отдельных слов, однако, в отличие от ее классического применения, каждый из кластеров рассматривается как отдельный документ в результате конкатенации текстов, входящих в этот документ. Далее выделение тем документов в коллекции сводится к стандартной работе со значениями метрики TF-IDF.

Следующий этап в построении модели – это определение группы признаков, которые будут использоваться для описания текстов. Здесь могут быть рассмотрены

всевозможные характеристики текстов, начиная от бинарных векторов, отражающих наличие/отсутствие ключевых слов в документе и заканчивая векторными представлениями всего текста. Рассмотрим более подробно некоторые из них.

Одной из наиболее известных и применяемых оценок слова в тексте является метрика TF-IDF (Term Frequency - Inverse Document Frequency) [54, 55], которая учитывает не только частоту упоминания слова в документе, но и частоту его упоминания в каждом документе коллекции текстов. Наличие поправки IDF (Inverse Document Frequency), являющейся логарифмом обратной частоты упоминания слова во всей коллекции текстов, позволяет сместить акцент с часто употребимых в научной лексике слов на действительно значимые для понимания тематики ключевые термины. Так, например, если слово часто упоминается в одном тексте, но редко – во всей коллекции текстов, то значение меры TF-IDF будет низким, что будет свидетельствовать о том, что данный термин характерен только для одного текста. Если слово, наоборот, редко употребляется в одном тексте, но присутствует практически в каждом тексте коллекции, то значение метрики TF-IDF будет также низким, что можно будет интерпретировать как признак того, что слово не является специфичным для выбранной тематики (в эту категорию, например, будут попадать вводные слова «however», «despite» и другие). Наиболее значимыми для определения тем внутри коллекции текстов будут обладать слова (ключевые термины), обладающие наивысшим значением метрики TF-IDF: то есть те, которые упоминаются в отдельных текстах достаточно часто, но при этом не входят в большинство текстов коллекции документов. Определив список ключевых терминов для коллекций, можно рассчитать для них значения метрик TF-IDF в случае каждого отдельного документа, и, представив полученные значения в качестве вектора, использовать для обучения алгоритма [56].

Еще одним способом описания текстов является применение векторных представлений документов. Эти представления получаются путем усреднения всех векторных представлений слов внутри документа. Векторные представления слов

– это вектор числовых значений, полученных путем многократного подбора весов в искусственных нейронных сетях, таким образом, чтобы выходные векторы схожих по контекстному употреблению слов обладали наибольшим сходством [57]. Если усреднить значение всех векторных представлений слов, то можно получить векторные представления документов, которые можно интерпретировать по тому же признаку: чем более схожи документы по своему смыслу, тем более схожи будут их векторы. Схожесть между многомерными векторами определяют, как правило, после снижения размерности до двух-трех параметров (или – показателей?) и оценкой косинуса угла между ними. Эта метрика является косинусной мерой сходства: чем ближе угол к 0°, тем ближе расположены друг к другу векторы, и тем меньше значения косинуса угла между ними, а значит, тем больше сходство между ними [58]. Использование косинусной меры сходства может быть использовано как самостоятельный метод классификации текстов в случае, если есть эталонная (даже небольшая) выборка релевантных текстов. Чаще всего векторные представления документов используют как отличительный признак текста при обучении алгоритмов.

Следует отметить, что качество векторных представлений документов строго зависит от разнообразия контекста отдельных слов внутри документа. При формировании векторных представлений слов необходимо использовать как можно более разнообразные тексты, причем в большом объеме. Именно поэтому к настоящему моменту готовы уже предобученные на большом числе документов модели искусственных нейронных сетей. Одним из примеров для английского языка является модель bioBERT [59]: адаптированная специально для биомедицинских текстов трансформерная модель BERT (Bidirectional Encoder Representations from Transformers) [60]. В отличие от своего предшественника, обученного на массиве текстов книг и статей в англоязычной Википедии, bioBERT корректирует векторные представления слов с учетом биомедицинского домена, что стало возможным при дообучении BERT на текстах из PubMed и PMC.

Выбор методов машинного обучения для построения классификатора текстов довольно разнообразен: зачастую используются нейросетевые архитектуры [61, 62], метод случайного леса (Random Forest, RF) [63], метод опорных векторов [64], метод k-ближайших соседей [65] и многие другие.

Выбор способа описания и метода классификации текстов во многом зависит от поставленной задачи и наличия/отсутствия временных ограничений. Что касается методов машинного обучения с учителем, авторы литературного обзора [66], анализируя точность различных способов описания текстов и алгоритмов машинного обучения, применительно к ним, на основе корпуса CORD-19 пришли к выводу, что наиболее качественный прогноз можно получить при использовании метрики TF-IDF для описания текстов, а для обучения модели – алгоритмов случайного леса и искусственных нейронных сетей (BERT).

Методы формирования выборки релевантных текстов отличаются по степени сложности и уровню автоматизации, начиная от готовых корпусов и экспертного отбора и заканчивая современными алгоритмами машинного обучения. Их применение позволяет сфокусироваться на узкой предметной области, сократить количество нерелевантной информации и обеспечить высокую точность последующего анализа.

1.2. Предобработка текстов

Автоматизированный анализ биомедицинских текстов – от научных публикаций до электронных медицинских карт – требует тщательной предобработки, которая преобразует неструктурированные данные в формат, пригодный для методов машинного обучения. Биомедицинские тексты обладают рядом особенностей: обилием узкоспециализированных терминов (например, «трансформирующий фактор роста бета» (Epidermal Growth Factor Receptor (EGFR)), сложных аббревиатур, EGFR или COVID-19 (Coronavirus Disease 2019), наличием числовых показателей (дозировки, лабораторные данные) и контекстно-зависимых сокращений. Кроме того, такие тексты часто содержат шум – артефакты

форматирования, опечатки, специальные или нестандартные символы (например, «α, β, ≤, ≥»).

Эффективная предобработка критически влияет на качество результатов анализа на последующих этапах: ошибки могут привести к некорректной интерпретации данных [67, 68]. К числу ключевых шагов предобработки относятся токенизация, нормализация регистра, удаление «стоп-слов», лемматизация или стемминг, обработка числовых данных и единиц измерения, а также удаление дубликатов и шумовых фрагментов.

1.2.1. Подготовка текста

Поскольку речь идёт о текстах, загружаемых из электронных баз данных, они зачастую включают в себя фрагменты гипертекстового документа, или HTML (HyperText Markup Language) - теги, указывающие на наличие разметки внутри статьи. Поскольку теги существенно отличаются от основного текста и в ряде случаев используются в большом количестве (например, теги курсива и полужирного шрифта), они могут вносить значительный шум при анализе. При работе с полными текстами публикаций может потребоваться удаление целых разделов: например, для задачи извлечения сведений о биологической активности химических соединений нет необходимости анализировать список литературы, раздел благодарностей и даже введение, поскольку оно содержит обобщённую информацию, доступную в других источниках. Кроме того, в некоторых случаях исключаются метаданные публикации (название журнала, список авторов), чтобы избежать смещения анализа.

Как и многие другие тексты, научные статьи содержат большое количество так называемых «стоп-слов». В компьютерной лингвистике под ними понимаются элементы текста, не несущие значимой смысловой нагрузки при анализе содержания. К ним относятся предлоги, союзы, вводные слова, а в английском языке – также артикли и формообразующие частицы. Кроме того, к стоп-словам в прикладных задачах часто относят общеупотребительную лексику, встречающуюся в большинстве документов корпуса и не помогающую различать их тематику

(например, слова «study», «result», «analysis» в научных статьях) [69]. При этом готовые списки стоп-слов, доступные в инструментах NLP (например, NLTK, spaCy), требуют адаптации для биомедицинской предметной области, так как иначе возможно удаление терминов, имеющих значение для анализа. При строгом семантическом анализе такие слова могут быть важны для понимания синтаксической структуры, однако при задачах классификации и тематического моделирования они вносят шум и могут снижать информативность метрик TF-IDF и векторных представлений документов [68, 70].

1.2.2. Токенизация

Описывая различные признаки, которые могут использоваться при классификации текстов, мы упоминали, что они, в основном, опираются на отдельные слова. В целом, даже естественное (человеческое), а не компьютерное прочтение текста сводится к пониманию отдельных слов, словосочетаний, предложений и абзацев. В сравнении с алгоритмами человек интуитивно определяет, где начинается и заканчивается слово, предложение или абзац. Для автоматических методов такое разделение не всегда очевидно: например, точка может использоваться не только для обозначения конца предложения, но и для сокращений, а наличие символа (скобка, запятая) не всегда означает конец слова (как в систематических наименованиях химических соединений).

В области Natural Language Processing (NLP, обработка естественного языка) процесс разбиения текста на его элементарные единицы называется токенизацией, а сами элементарные единицы – токенами [71].

Простейший принцип токенизации заключается в разбиении текста по пробелам и знакам препинания (точка, запятая, вопросительный или восклицательный знак и др.). Такой подход не всегда оптимален при обработке биомедицинских текстов. Например, систематические наименования химических соединений содержат больше смысловых частей, чем те, которые выделяются простым разбиением по пробелам [44]. Выбор стратегии токенизации должен

соответствовать целям анализа, так как от него во многом зависит точность последующего извлечения информации [71].

Современные методы токенизации включают не только простые правила разбиения по пробелам и знакам препинания, но и так называемые подсловные методы токенизации. Их основная идея заключается в том, чтобы разбивать редкие или длинные слова на более мелкие единицы – подслова (subwords), которые встречаются чаще и имеют устойчивое значение. Такой подход позволяет существенно сократить размер словаря и одновременно сохранить способность модели обрабатывать ранее не встречавшиеся слова.

Наиболее распространённые подсловные методы – это Byte Pair Encoding (BPE) [72], WordPiece [60] и SentencePiece [73]. Они используют статистику частоты встречаемости: сначала текст разбивается на отдельные символы, затем наиболее часто встречающиеся пары символов объединяются в подслова, и процесс повторяется до тех пор, пока не будет достигнут требуемый размер словаря. В результате редкое или сложное слово, отсутствующее в словаре, может быть представлено как комбинация более простых подслов.

Применительно к биомедицинским текстам подсловная токенизация имеет особое значение. Здесь часто встречаются длинные термины (например, «3-hydroxy-3-methylglutaryl-coenzyme A reductase»), химические соединения с систематическими названиями, а также уникальные генные и белковые обозначения (например, «BRCA1», «IL2Rγ»). Традиционные методы токенизации по пробелам и пунктуации могут некорректно обрабатывать такие слова, дробя их на бессмысленные фрагменты или, наоборот, не разделяя значимые части. Подсловные методы позволяют сохранить информативность этих терминов, поскольку даже если полное название встречается редко, отдельные его части («hydroxy», «methyl», «glutaryl») будут представлены в словаре и корректно интерпретированы моделью [74].

Таким образом, подсловная токенизация обеспечивает баланс между полнотой словаря и его обобщающей способностью. Она позволяет моделям

работать с редко встречающимися и ранее невидимыми терминами, что особенно важно для биомедицинских текстов, отличающихся высоким уровнем терминологического разнообразия и постоянным появлением новых понятий.

1.2.3. Нормализация

В некоторых случаях работы с текстами может потребоваться так называемая нормализация – процесс, необходимый для того, чтобы стандартизовать текст в соответствии с задачей.

Самый яркий пример процесса нормализации текста – это приведение его к единому регистру (верхнему или нижнему). Данный этап может быть необходим, например, при поиске строки в подстроке, или термина из словаря во всем тексте. Необходимость нормализации по регистру связана с тем, что любой компьютерный символ соответствует определенному коду в кодировке, причем одна и та же буква, но в разном регистре, будет иметь различный код, и поэтому различаться с точки зрения компьютерного анализа текста. Слово в начале предложения с заглавной буквы и то же слово внутри предложения с прописной будут считаться разными, хотя идентичны в понимании естественного прочтения текста.

Иногда в предобработку текстов также включают и процесс лемматизации [75]. Лемма – это базовая форма слова, которая имеет схожее семантическое значение со всеми словоформами, образованными от нее. Например, для слова «бегал» в качестве леммы может выступать слово «бегать», для слова «пробежка» - «бег», и так далее. Существует уже немалое количество готовых инструментов для лемматизации текста. Самый распространенный из них – это функция лемматизации текстов, реализованная внутри библиотеки для Python NLTK (Natural Language Toolkit) [76]. Эта библиотека не специализирована для биологических и медицинских текстов, которые изобилуют предметной терминологией, и поэтому получение лемм специализированных терминов может быть затруднено или ошибочно. По этой причине были разработаны специальные инструменты для лемматизации биомедицинских текстов. Одним из наиболее распространенных для английского языка является BioLemmatizer [75], в котором

достигается более 97% точности на биомедицинских корпусах текстов. Для текстов на других языках также разработаны и продолжают разрабатываться инструменты обработки естественного языка: например, в работе [77] представлен инструмент для лемматизации медицинских текстов на немецком языке, а также проекты DeepPavlov [78] и Natasha [79], которые адаптированы для русского языка.

Похожий на лемматизацию способ предобработки текстов – стемминг – позволяет выделить основу слова, то есть морфологический фрагмент без окончания [80]. В английском языке разнообразие окончаний не столь велико: самые разнообразные из них входят в группу окончаний глаголов, которые используются для формирования временных форм (прошедшего, настоящего, будущего).

Большинство способов нормализации текстов используются, в основном, при анализе общего содержания текстов и их классификации [81-83], поскольку в значительной мере изменяют синтаксическую структуру отдельных предложений.

1.3. Обзор инструментов для обработки естественного языка

Прежде чем переходить к обзору задач распознавания наименований объектов и поиска взаимосвязей между ними, следует остановиться на общих или локальных инструментах обработки естественного языка, которые используются или могут быть использованы на всех предыдущих и последующих этапах. Краткое описание рассмотренных инструментов представлено в таблице 1.

Таблица 1 – Примеры инструментов обработки естественного языка.

Инструмент NLP	**Реализованные функции работы с текстами**	**Поддерживаемые языки**	**Ссылки**
NLTK	• Токенизация • Стемминг • Лемматизация • Удаление стоп-слов • Нормализация (регистр, числа)	• В основном - английский • Для многих языков доступно только ограниченное число модулей	[76, 84]

	• Разметка по частям речи (POS-tagging) • Синтаксический разбор • Извлечение наименований объектов • Семантический анализ • Классификация текста (Naive Bayes, Decision Trees) • Анализ тональности • Кластеризация документов • Векторизация текстов (Bag of Words, TF-IDF) • Метрики (напр., числовые характеристики схожести строк)	(токенизация, стемминг) • Возможно использование внешних модулей	
spaCy	• Токенизация • Разметка по частям речи • Лемматизация • Синтаксический анализ • Распознавание наименований объектов (NER) • Сегментация предложений • Векторные представления слов • Поиск по шаблонам • Классификация текста • Морфологический разбор предложений	• Более 75 языков – включая английский, русский, китайский	[85]
Natasha	• Токенизация • Лемматизация • Распознавание наименований объектов (NER) • Синтаксический разбор	Русский	[79]

	• Нормализация текста (исправление опечаток, приведение к стандартному формату) • Поиск по шаблонам • Работа с датами (извлечение, нормализация)		
DeepPavlov	• Токенизация • Разметка по частям речи • Морфологический анализ • Лемматизация • Синтаксический анализ (включая зависимости) • Распознавание наименований объектов (NER) • Классификация текста • Анализ тональности • Диалоговые системы (чат-боты, вопросы-ответы) • Поддержка предобученных языковых моделей (BERT, RoBERTa, GPT)	Русский, английский и другие языки (через предобученные модели)	[78, 86]

1.3.1 Natural Language Toolkit (NLTK)

Ранее, описывая процесс лемматизации текстов, мы уже упоминали библиотеку NLTK [84]. Ее функционал гораздо шире ранее описанных процессов, и позволяет удобно и быстро работать с текстами на естественном языке.

Библиотека NLTK позволяет довольно быстро и точно проводить разметку частей речи, которая в зарубежной литературе называется «PoS-tagging» (PoS – part-of-speech, часть речи). При использовании функции такой разметки следует учитывать, что распознавание частей речи производится на основе контекста употребления слова, поэтому без него получить корректный тег представляется

маловероятным. PoS-теги могут быть использованы в различных задачах, наиболее очевидная из которых – это описание слов при распознавании наименований (например, биологические и химические сущности, как правило, представлены существительным) [44].

Внутри библиотеки NLTK предусмотрен быстрый доступ к списку стоп-слов. Словари стоп-слов доступны не только для английского языка, но и для многих других, включая русский. Удобство работы со списком стоп-слов с использованием NLTK заключается в гибком управлении словарем: новые стоп-слова могут быть добавлены в него, а те, которые в работе не нужны, наоборот, удалены. Стоп-слова являются частью блока *corpus* библиотеки NLTK: помимо них, в этот блок включены множество возможностей, включая работу с аннотированными корпусами текстов, доступ к названиям разнообразных категорий текстов и их фрагментов, которые часто используются в задачах обработки естественного языка (например, «положительный» и «отрицательный» - для анализа тональностей, отражающих настроение автора текста), и многое другое.

Еще одной возможностью библиотеки NLTK является генерация текста по заданным правилам. Речь идет не о содержательной, «смысловой» генерации, а о формальной – на основе заранее определенных шаблонов, например, последовательности частей речи (существительное–глагол–существительное) или фиксированной длины предложения. Также с применением библиотеки NLTK можно производить распознавание наименований объектов: готовый алгоритм уже находится внутри библиотеки, однако он является универсальным для всех областей, поскольку основные классы наименований для распознавания – это имена собственные (географические места, названия организаций, и др.).

Перечисленные функции представляют собой лишь часть доступных в библиотеке NLTK; их гораздо больше, и с полным списком можно ознакомиться в документации проекта [87].

1.3.2. spaCy

Инструмент spaCy, как и NLTK, является набором библиотек для Python, которые позволяют проводить различные операции над текстами на естественном языке. spaCy - исторически более «молодой» инструмент в сравнении с NLTK, но обладает схожим функционалом. Тем не менее, spaCy обладает рядом значительных преимуществ, в числе которых: (1) поддержка графических процессоров (Graphics Processing Unit, GPU), что позволяет ускорять работу алгоритмов, (2) возможность настройки так называемых конвейеров обработки текстов на естественном языке благодаря компактному синтаксису основных функций для обработки текста, (3) мультиязычная поддержка. Последний пункт требует особого пояснения: начало работы с любым текстом с использованием библиотеки spaCy заключается в определении подходящей модели и ее загрузки. При выборе модели основными критериями являются ее язык и размер. Последний параметр влияет и на скорость работы алгоритма, и на ее качество: чем больше текстов использовалось при ее обучении, тем более точные результаты можно получить, однако время работы алгоритмов будет больше. Модель spaCy позволяет проводить анализ более чем 75 языков, включая русский. Модель spaCy для обработки текстов на русском языке (как и любом другом) может быть относительно просто встроена в любой сценарий анализа текстов, например, структурирование информации о состояниях пациентов на основе ЭПМЗ [88].

В spaCy реализован алгоритм распознавания наименований объектов, однако он остается универсальным, то есть рассчитанным на широкий спектр текстов и предметных областей без адаптации к специфике биомедицинского языка. Стандартные модели позволяют распознавать такие сущности, как имена людей, географические локации, организации, даты и другие общие категории, но не специализированные объекты, характерные для научных текстов – химические соединения, гены или заболевания.

Кроме того, библиотека предлагает инструменты для синтаксического разбора предложений [89]. Важным преимуществом является возможность

визуализации зависимости между главными и второстепенными членами предложений, что может использоваться на дальнейших этапах автоматизированного анализа текстов, например, при выявлении взаимосвязей между сущностями.

В рамках проекта spaCy доступно также решение для научных биомедицинских документов – scispaCy [90]. Библиотека scispaCy позволяет использовать все доступные в spaCy функции, но с применением моделей, обученных на биологических и медицинских текстах. Дополнительно, с применением библиотеки scispaCy можно провести процедуру соотнесения сокращений с их полными названиями, а также автоматически сопоставить извлеченные наименования сущностей с объектами в базах данных.

1.3.3. Natasha

Ранее мы рассмотрели универсальные решения, позволяющие работать с текстами на различных языках. Однако для русскоязычных текстов существуют также и локальные, целенаправленные решения. Такая цель стоит перед проектом Natasha. Авторы проекта заявляют, что Natasha – это набор библиотек для Python; и также отмечают, что основная цель разработки – это скорость работы алгоритмов, а не их качество.

На сайте проекта [79] можно ознакомиться с функционалом библиотек в режиме реального времени. Через веб-версию ресурса возможно проведение синтаксического (главные и второстепенные члены предложения и связь между ними), морфологического (часть речи, число, род и др.) разбора текста, а также выявление наименований объектов (люди, локации и др.). Хотя проект ориентирован на разработку алгоритмов для общей обработки естественного языка, некоторые разделы также включают в себя и специализированные для биологии и медицины ресурсы. В частности, в разделе с наборами обучающих данных для алгоритмов интеллектуального анализа текстов [91] можно найти корпусы русскоязычных текстов, в которых аннотированы наименования лекарственных препаратов и связь с их эффективностью/нежелательными реакциями [92].

1.3.4. DeepPavlov

В отличие от проекта Natasha, который в первую очередь ориентирован на скорость работы и решение базовых задач, библиотека DeepPavlov [78, 86] разрабатывалась как более универсальное и масштабируемое решение. DeepPavlov – это открытый программный комплекс, созданный в Московском физико-техническом институте, предназначенный для широкого круга задач обработки текстов на естественном языке.

В состав библиотеки входят готовые модели и инструменты для таких задач, как морфологический и синтаксический разбор, выделение наименований объектов, определение части речи, анализ тональности и классификация текстов. Кроме того, DeepPavlov предоставляет возможность конструировать собственные цепочки обработки текста, объединяя несколько этапов анализа в единую систему.

Особенностью DeepPavlov является поддержка современных методов машинного обучения и нейросетевых архитектур, включая предобученные языковые модели BERT [60], RoBERTa (Robustly Optimized BERT Pretraining Approach) [93] и GPT (Generative Pre-trained Transformer) [94]. Это обеспечивает высокую точность при решении прикладных задач и позволяет учитывать широкий контекст при анализе текстов.

Библиотека распространяется с открытым исходным кодом и имеет подробную документацию, что делает её удобной как для исследователей, так и для разработчиков.

1.4. Распознавание наименований объектов

Распознавание наименований объектов, или иначе распознавание именованных сущностей (Named Entity Recognition, NER) – это процесс выделения из всего массива символов текста только тех последовательностей символов, которые являются частью строк с названиями представляющих интерес объектов. Если речь идет об области биологии и медицины, такими объектами могут быть химические соединения, заболевания, биологические виды, клеточные линии, белки или гены, однонуклеотидные полиморфизмы генов и многое другое.

Выделяют три основных группы методов: основанные на словарях, правилах и алгоритмах машинного обучения [95, 96]. Встречаются также методы, которые являются гибридными, то есть комбинируют несколько подходов.

Методы, основанные на словарях, широко используются в биологии и медицине благодаря наличию специализированных терминологических баз данных, таких как MeSH, UniProt [97, 98], ChEBI (Chemical Entities of Biological Interest) [99, 100] и HGNC (HUGO Gene Nomenclature Committee) [101, 102] и других. Эти подходы предполагают сопоставление текста с заранее составленными перечнями биомедицинских терминов, что обеспечивает высокую точность распознавания стандартизированных названий. Их эффективность снижается при работе с синонимами, аббревиатурами и вариативными формами терминов, характерными для научных публикаций. Постоянное появление новых биомедицинских понятий требует регулярного обновления словарей, что затрудняет поддержание их актуальности.

Методы, основанные на правилах, используют лингвистические шаблоны, учитывающие морфологические, синтаксические и контекстуальные особенности биомедицинских текстов. Например, названия генов часто содержат сочетания букв и цифр (напр., BRCA1 (Breast Cancer gene 1)), наименования лекарственных препаратов могут включать специфические суффиксы/окончания (напр., -mab/-маб для моноклональных антител), названия ферментов имеют окончание -ase/-аза. Такие правила особенно эффективны для распознавания стандартизированных

номенклатур, однако их разработка требует глубоких знаний предметной области и трудоемкой настройки при адаптации к новым типам сущностей или текстовым корпусам.

Машинное обучение в последние годы становится основным инструментом для решения задач распознавания наименований в биомедицине, поскольку позволяет автоматически выявлять сложные закономерности в текстах без явного задания правил. Их применение требует больших объемов размеченных данных, которые в биомедицинской области часто ограничены и требуют привлечения экспертов для аннотирования. Задача алгоритмов машинного обучения, обобщая, заключается в том, чтобы обнаружить закономерности строения и употребления наименований объектов в аннотированных текстах, и при анализе новых, не входящих в обучающую выборку, найти те условия, в которых могут встречаться наименования объектов.

1.4.1. Метрики для оценки точности алгоритмов интеллектуального анализа текстов

Прежде чем перейти к описанию методов распознавания наименований и извлечения ассоциаций между ними, целесообразно охарактеризовать метрики качества алгоритмов, наиболее часто применяемые в области интеллектуального анализа текстов.

Общепринятые в математическом моделировании показатели – чувствительность и специфичность – в данной области практически не используются. Это связано с тем, что аннотированные корпусы, применяемые для обучения и тестирования моделей, как правило, существенно несбалансированы: фрагменты текста, содержащие наименования объектов, составляют лишь незначительную долю от общего объёма текста. В результате метрика специфичности теряет информативность, поскольку её значение в большинстве случаев будет близко к единице и не позволит объективно оценить качество работы алгоритма [103].

Для оценки эффективности алгоритмов интеллектуального анализа текстов наиболее часто используют три метрики – precision, recall и F1-score.

Метрика precision (в русскоязычной литературе часто обозначаемая как точность) отражает долю верных среди всех положительных прогнозов модели:

$$precision = \frac{TP}{TP + FP}$$

Метрика recall (или полнота) характеризует долю корректно выявленных положительных примеров среди всех положительных объектов, присутствующих в аннотированной выборке:

$$recall = \frac{TP}{TP + FN}$$

Наконец, F1-score используется для интегральной оценки качества работы алгоритма и представляет собой гармоническое среднее между precision и recall:

$$F_1 - score = \frac{2 * precision * recall}{precision + recall}$$

Таким образом, применение метрик precision, recall и F1-score позволяет объективно оценивать качество алгоритмов, ориентированных на выявление редких, но значимых фрагментов текста. Эти показатели наиболее адекватно отражают способность модели корректно распознавать целевые сущности и избегать ложных срабатываний.

1.4.2. Алгоритмы распознавания наименований на основе словарей

Одним из наиболее ранних и вместе с тем до сих пор востребованных подходов к распознаванию биомедицинских сущностей являются методы на основе словарей. Их работа заключается в сопоставлении последовательностей символов в тексте с заранее составленными перечнями терминов. Такие словари формируются из авторитетных источников – биомедицинских онтологий и баз знаний, где каждому объекту сопоставлен уникальный идентификатор, а также зафиксированы его каноническое название и синонимы. Наиболее часто применяются MeSH, UMLS [104, 105], ChEBI, PubChem [106, 107], UniProt, а также специализированные ресурсы, такие как Disease Ontology [108, 109] и Gene

Ontology [110, 111]. В клинической практике также широко используется SNOMED CT (Systematized Nomenclature of Medicine–Clinical Terms) [112, 113] как крупнейший справочник медицинской терминологии.

Принцип работы методов распознавания наименований на основе словарей заключается, как правило, в строгом строковом сопоставлении (exact string matching), при котором термин из словаря ищется в тексте как подстрока. При совпадении оно интерпретируется как упоминание сущности. Такой подход обеспечивает высокую точность при распознавании стандартизированных терминов, однако плохо справляется с вариативностью: синонимами, аббревиатурами или длинными составными наименованиями, где характерны изменения порядка слов и пунктуации [114]. Для частичного преодоления этих ограничений применяется нечёткое сопоставление (approximate string matching), например с использованием расстояния Левенштейна [115], позволяющего учитывать опечатки и мелкие различия в написании. Такие методы не решают проблему устаревания словарей: в условиях быстрого роста числа биомедицинских терминов они требуют регулярного обновления [116].

Сильной стороной словарных методов является возможность нормализации: сопоставленный термин немедленно связывается с уникальным идентификатором в базе данных, что облегчает интеграцию результатов анализа с фактографическими ресурсами. Так, система MetaMap, разработанная Национальной медицинской библиотекой США (National Library of Medicine, NLM), обеспечивает сопоставление терминов текста с онтологией UMLS и использует гибридный подход, сочетающий строгий и приближённый поиск [117]. Для извлечения химических сущностей применялся инструмент tmChem, ориентированный на идентификацию названий лекарственных препаратов и других химических соединений [118]. Для заболеваний аналогичный функционал реализован в системе DNorm, которая объединяет словарные методы с моделью нормализации терминов на основе машинного обучения [119].

Словарные методы играют ключевую роль в задачах, требующих сопоставления терминов с фактографическими базами и онтологиями, поскольку обеспечивают интерпретируемость и воспроизводимость результатов. Их эффективность ограничена неполнотой словарей, полисемией и высокой вариативностью биомедицинской терминологии. На практике они часто применяются в составе гибридных систем, сочетающих правила и статистические методы, что позволяет повысить полноту и надёжность извлечения сущностей.

1.4.3. Алгоритмы распознавания наименований на основе правил

Алгоритмы распознавания на основе правил представляют собой классический подход к извлечению наименований объектов, который основывается на созданных вручную правилах, разработанных экспертами-лингвистами или предметными специалистами. Принцип работы этих алгоритмов заключается в сопоставлении текста с заранее определенными шаблонами, кодирующими морфологические, синтаксические, семантические и контекстные характеристики целевых объектов [120-122].

К числу таких правил относится использование регулярных выражений для поиска морфологических признаков, например, идентификации слов, написанных только заглавными буквами, что характерно для аббревиатур (как «EGFR», «ATP»), или наличия специфических суффиксов, таких как «-ase» для ферментов («kinase», «polymerase»). Помимо поиска по форме, правила можно эффективно использовать для поиска объектов по их функции или характеристикам через анализ контекстного окружения. Это особенно актуально для извлечения сущностей, которые сложно идентифицировать только по форме. Так, в биохимических текстах фраза «metabolite of» с высокой вероятностью указывает на родительское соединение, а конструкция «metabolites are» предваряет список самих метаболитов [123, 124]. Аналогичным образом, контекстные маркеры типа «inhibits», «binds to» или «expression of» являются надежными индикаторами для указания на гены, белки или лекарственные средства [125].

Ключевым достоинством методов, основанных на правилах, является их высокая точность и интерпретируемость на тех данных, для которых они были разработаны. Работа алгоритма производится строго в соответствии с заданной логикой, что делает его предсказуемым и не требующим больших вычислительных ресурсов для обучения. Такие системы можно относительно быстро создать и протестировать для узкой предметной области. Кроме того, подходы на основе правил обладают важным преимуществом – они обеспечивает прозрачность и интерпретируемость извлечённых результатов.

Вместе с тем, как и любая методология, он имеет ряд особенностей, которые необходимо учитывать при практическом применении. Создание достаточно полного набора правил требует внимательного анализа текстов и участия экспертов, а сами правила подлежат регулярному обновлению в связи с появлением новых терминов и изменениями в языке. Кроме того, правила, сформированные для одного типа документов, могут нуждаться в адаптации при работе с другими источниками, что связано с разнообразием стилей и форм изложения.

Как и методы на основе словарей, подход, основанный на правилах, демонстрирует высокую точность в узких рамках, но не является универсальным и масштабируемым решением для задачи распознавания наименований объектов в разнородных текстах. Его часто применяют в качестве компонента в гибридных системах или на начальных этапах разработки для разметки данных. К примеру, авторы исследования [126] применили набор регулярных выражений и методы машинного обучения (искусственные нейронные сети) для извлечения информации о генетических вариантах для *C. Elegans*, записанных как в стандартизованной, так и в описательной формах.

1.4.4. Алгоритмы распознавания наименований на основе методов машинного обучения

В отличие методов на основе словарей и правил, использующих строгое сопоставление строк или заранее заданные лингвистические шаблоны, алгоритмы машинного обучения позволяют автоматически выявлять скрытые закономерности

в текстах. Их основная идея заключается в том, что на основе размеченной выборки формируется модель для распознавания последовательности символов, характерных для тех или иных типов сущностей, и затем может применяться к новым, ранее не встречавшимся текстам. Такой подход обеспечивает значительно более высокую гибкость и способность к обобщению: модель позволяет распознать вариативные формы терминов, аббревиатуры и новые комбинации слов, которые не были явно представлены в словарях или правилах.

Ключевым ограничением машинного обучения в области биомедицины остаётся зависимость от наличия больших объёмов качественно размеченных данных [127, 128]. Для создания эффективных моделей требуются корпусы текстов, где биомедицинские сущности аннотированы экспертами, а при необходимости также дополнены связями с идентификаторами в онтологиях и базах данных. Подобная разметка требует значительных временных затрат.

Методы машинного обучения стали основным направлением развития автоматического распознавания наименований химических и биологических объектов, поскольку дают возможность проводить распознавание наименований объектов, которые отсутствуют в словарях или не соответствуют определенному шаблону.

1.4.4.1. Корпусы и наборы данных для обучения

Ключевым условием успешного применения алгоритмов машинного обучения для распознавания биомедицинских сущностей является наличие размеченных корпусов текстов, где наименования объектов выделяются вручную экспертами. Под разметкой в данном контексте понимается процесс, при котором эксперт анализирует текст и отмечает в нём последовательности символов, соответствующие наименованиям объектов.

Результаты разметки сохраняются в структурированном виде, удобном для использования. Как правило, для каждого упоминания фиксируются:

1. тип сущности (например, *Chemical*, *Disease*, *Gene*),
2. начальная и конечная позиция символа в тексте (номер первого и последнего символа строки, соответствующей термину).

В упрощённом виде пример аннотации может выглядеть так:

- Текст
  «Oseltamivir is effective against influenza A virus.»
- Разметка:
  - *Oseltamivir* → тип: *Chemical*, позиции: 0–10
  - *influenza A virus* → тип: *Species*, позиции: 25–42

В некоторых корпусах дополнительно указывается нормализованный идентификатор объекта в базах данных или онтологиях.

Одним из наиболее значимых ресурсов является CHEMDNER (Chemical Compound and Drug Name Recognition), созданный в рамках BioCreative IV [129]. Он включает 10 000 резюме научных публикаций из PubMed с более чем 84 000 аннотаций химических сущностей. В корпусе проводится классификация по типам упоминаний (систематические названия, тривиальные, аббревиатуры, идентификаторы и др.), что позволяет моделям учитывать вариативность форм представления химических соединений.

Для задач, связанных с заболеваниями, широко используется NCBI Disease Corpus [130]. Он содержит 793 резюме научных публикаций из PubMed с примерно 6 900 размеченными упоминаниями заболеваний, каждое из которых сопоставлено с уникальным идентификатором Disease Ontology.

Распознавание наименований генов и белков традиционно тестируется на корпусе JNLPBA (Joint Workshop on Natural Language Processing in Biomedicine and Its Applications) [131]. В него включено около 2 000 рефератов, где аннотированы

пять категорий сущностей: гены/белки, ДНК, РНК, клеточные линии и клеточные типы. Стоит упомянуть здесь, что хотя белки и гены представляют собой различные биологические объекты, в задачах интеллектуального анализа текстов они традиционно объединяются в одну категорию. Это связано с тем, что в научной литературе одни и те же термины могут обозначать как ген, так и кодируемый им белок, а также с тем, что контекст их употребления во многом совпадает.

Для области генетических вариаций используется корпус, сопровождающий систему tmVar [132]. Он состоит из 500 аннотаций научных публикаций, в которых аннотированы упоминания однонуклеотидных полиморфизмов (SNPs) и мутаций в биомедицинских текстах. Этот ресурс играет важную роль, так как такие наименования сложно идентифицировать из-за разнообразия форматов записи.

Помимо англоязычных корпусов, активно используются и специально созданные русскоязычные наборы данных [133]. Так, RuDReC (Russian Drug Reaction Corpus) включает тексты отзывов пациентов о лекарственных препаратах и аннотации, отражающие эффективность, побочные эффекты и эмоциональную окраску высказываний, что позволяет его использовать в задачах автоматизированного извлечения сущностей, связанных с побочными эффектами препаратов, а также анализа тональностей [134]. RuCCoN (Russian Clinical Concept Normalization) содержит фрагменты электронных историй болезни с разметкой клинических сущностей – заболеваний, лекарственных средств, симптомов и процедур. Термины соотнесены с идентификаторами UMLS, обеспечивая основу для автоматической нормализации понятий [135].

1.4.4.2. Методы машинного обучения, используемые для распознавания наименований объектов

Исторически первыми в задаче распознавания биомедицинских сущностей начали использоваться алгоритмы классического машинного обучения, такие как метод опорных векторов (Support Vector Machines, SVM). Их принцип работы

заключался в том, что каждый токен текста представлялся в виде набора признаков, отражающих его свойства. Для этого применялись простые способы векторизации, например «мешок слов» (bag-of-words), где документ описывается вектором частот встречающихся слов без учёта порядка, или метрики TF-IDF, а также морфологические и синтаксические характеристики (POS-теги, окончания слов, принадлежность к словарю терминов). В работе [136] такой подход был протестирован на корпусе GENIA, где аннотированы упоминания генов и белков. Полученные результаты (F_1-score от 0,42 до 0,73 для различных типов объектов) показали ограниченность возможностей применения SVM к биомедицинским текстам: модель не учитывала последовательность слов и зависимость между соседними токенами.

Более точные результаты обеспечил алгоритм условных случайных полей (Conditional Random Fields, CRF) [137]. В отличие от SVM, CRF учитывают линейную структуру текста и взаимосвязь между соседними метками, что особенно важно при распознавании составных терминов (например, «human immunodeficiency virus type 1»). Для обучения CRF, как правило, используются различные морфологические (окончание), синтаксические (наличие символов) признаки, длина токена, принадлежность к словарю заранее определенных терминов, метки части речи. Все эти признаки подавались на вход модели как вектор бинарных или числовых значений, описывающий каждый токен. В работе [138] алгоритм CRF позволил достигнуть значений F_1-score около 0,78 на корпусе GENIA. Позднее Leaman и Gonzalez [139] применили CRF для извлечения химических наименований в корпусе CHEMDNER, где использовали комбинацию словарных индикаторов, морфологических шаблонов и контекстных признаков, достигнув F1-score 0.82.

Активное развитие алгоритмов нейронных сетей привело к их широкому использованию и в области интеллектуального анализа текстов. Как правило, для описания текстов применяют уже не списки различного рода признаков, а

векторные представления слов. В более ранних исследованиях для распознавания наименований объектов используется архитектура рекуррентных искусственных нейронных сетей (Recurrent Neural Networks, RNN) и ее модификации – LSTM (Long Short-Term Memory) – с ячейками памяти, которые способны моделировать длинные зависимости в тексте. Еще одной модификацией рекуррентных нейронных сетей является архитектура BiLSTM (Bidirectional LSTM), которая, в отличие от обычной LSTM, обрабатывает текст сразу в двух направлениях – слева направо и справа налево. Это позволяет учитывать как предшествующие, так и последующие слова при определении сущности. Архитектура BiLSTM зачастую используется в комбинации с CRF: первый алгоритм используется для кодирования контекста, а второй – для обеспечения согласованности меток, учитывая, что, например, внутри одного термина не может встречаться чередование меток «начало» и «вне сущности». Такой подход был реализован в работе [140], где алгоритм BiLSTM-CRF, обученный на векторных представлениях слов Word2Vec (PubMed + PMC), позволил достигнуть F1-score 0,84 при распознавании наименований химических соединений (наибольшее значение) и 0,77 при распознавании наименований клеточных линий (наименьшее значение).

В последнее время наиболее высокая точность выявлена для моделей на основе трансформеров, которые благодаря механизму самовнимания (self-attention) способны учитывать взаимосвязи между всеми словами в предложении, независимо от их положения. Базовая архитектура BERT показала высокие результаты на общем корпусе, а её биомедицинские адаптации – BioBERT, SciBERT [141] – стали использоваться для различных задач интеллектуального анализа биомедицинских текстов, включая распознавание наименований. В этих моделях векторные представления токенов формируются автоматически на основе контекста, а на выходе каждая позиция классифицируется по типу сущности. Для повышения согласованности меток на финальном уровне часто используется CRF-слой, аналогичным образом, как он используется BiLSTM-CRF. Так, BioBERT

достиг F1-score 0,88 на NCBI Disease и 0,89 на CHEMDNER, что существенно превысило показатели качества более ранних моделей.

1.5. Алгоритмы извлечения ассоциаций между сущностями

Распознавание наименований биологических и химических объектов само по себе не обеспечивает возможности систематизации знаний: для построения базы данных и выявления закономерностей необходимо учитывать также ассоциации между объектами. Под извлечением ассоциаций (relation extraction, RE) понимается автоматическое установление семантических связей, которые авторы публикаций явно или неявно фиксируют в тексте. Если задача NER отвечает на вопрос «какие объекты упомянуты?», то задача RE уточняет «каким образом они связаны?». В биомедицине такие связи могут отражать широкий спектр взаимодействий: от ингибирования мишени химическим соединением до генетических ассоциаций с заболеваниями.

Как и в алгоритмах распознавания наименований, в алгоритмах извлечения ассоциаций можно выделить два основных направления: основанные на правилах и машинном обучении.

1.5.1. Алгоритмы извлечения ассоциаций, основанные на правилах

Одним из ключевых направлений в извлечении ассоциаций являются методы на основе правил и словарей, которые широко применяются в биомедицинских задачах благодаря своей интерпретируемости и воспроизводимости. Суть подхода заключается в формализации лингвистических конструкций, указывающих на наличие связи между объектами. Так, шаблоны вида «*X inhibits Y*», «*interaction between X and Y*» или «*X is a biomarker of Y*» могут использоваться для извлечения конкретных типов отношений. Дополнительно применяются словари синонимов ключевых глаголов и существительных, что расширяет охват возможных формулировок.

Наряду с этим особое значение имеет метод, основанный на совместной встречаемости объектов в пределах одного предложения или абзаца. Такой подход позволяет фиксировать потенциально значимые биологические связи даже в тех случаях, когда авторы публикации не используют явные семантические связки. Например, совместное упоминание вируса и мутации может указывать на её роль в резистентности, хотя сама формулировка «*mutation causes resistance*» в тексте отсутствует. Подобные методы особенно востребованы при построении баз данных, где важна полнота покрытия, а уточнение семантики может быть выполнено на последующих этапах анализа.

Практические результаты подтверждают эффективность этих подходов. В алгоритме под названием SemRep [142] используются синтаксические правила и шаблоны для выявления семантических связей между медицинскими сущностями. При извлечении отношений типа *treatment* было продемонстрировано, что метод характеризуется высокой точностью извлечения ассоциаций (precision составил 0,84**)**, но при этом полнота остается довольно низкой (recall составил 0,54). Другой пример – BELMiner [143], ориентированный на извлечение событий в формате Biological Expression Language (BEL) – формализованного языка для представления знаний о биологических процессах и молекулярных взаимодействиях, например, «*p(HGNC:AKT1) increases p(HGNC:MTOR)*». При использовании этого алгоритма авторам удалось достигнуть значений precision в диапазоне 0,65–0,70, а recall – 0,40–0,55, в зависимости от категории отношений.

Методы на основе совместной встречаемости терминов нередко показывают высокую полноту, что делает их удобным инструментом для первичного выявления связей. Авторы исследования [144] построили сеть генов на основе их совместной встречаемости в проанализированных микробных геномах. Такой подход позволил выявить функционально связанные гены без явной аннотации взаимодействий, что позволило провести реконструкцию генетических сетей. В другой работе [145] авторы применили метод анализа совместной встречаемости для построения сети взаимодействий «ген–болезнь». Было продемонстрировано, что использование

совместной встречаемости в текстах может позволить выявить ассоциации, пересекающиеся с известными базами данных, однако точность остаётся ограниченной и требует дополнительной фильтрации. Схожий принцип применялся и в диссертационной работе Пономаренко Е.А. (2009), где ассоциации между белками оценивались через совместное упоминание в близких по смыслу рефератах MEDLINE и PubMed; построенная таким образом сеть согласовывалась с метаболическими путями KEGG и аннотациями Gene Ontology [146].

1.5.2. Алгоритмы извлечения ассоциаций, основанные на методах машинного обучения

Для решения задач извлечения ассоциаций с применением методов машинного обучения, как и в случае задач распознавания наименований объектов, требуются специализированные корпусы, в которых аннотация включает не только сами сущности, но и наличие или тип взаимосвязи между ними. В отличие от корпусов, предназначенных для распознавания наименований, здесь единицей аннотации выступает пара объектов, дополненная указанием класса взаимодействия. Это позволяет использовать корпусы как для бинарной классификации («есть связь/нет связи»), так и для задач многоклассовой категоризации, где каждому взаимодействию соответствует отдельный тип.

1.5.2.1. Примеры корпусов, которые используются для построения моделей извлечения ассоциаций из текстов

Одним из наиболее известных ресурсов является BioCreative V Chemical – Disease Relation (CDR) корпус, включающий 1 500 статей из PubMed (заголовки и аннотации), где вручную были размечены связи типа «химическое соединение–заболевание». Корпус содержит 4 409 уникальных ассоциаций и используется как стандартный набор для оценки алгоритмов по извлечению связей между химическими соединениями и заболеваниями [147].

Для выявления взаимодействий «химическое соединение – белок/ген» был создан корпус ChemProt (2 500 аннотированных рефератов) [148] в рамках

BioCreative VI, и позже опубликованное в рамках BioCreative VII его расширение – DrugProt (5 000 аннотированных рефератов) [149]. Каждая аннотированная пара объектов сопровождается меткой типа взаимодействия, таких как «ингибитор», «агонист», «субстрат» и другие. Задача классификации в ChemProt имеет многоклассовый характер.

Ещё один крупный корпус – Drug-Drug Interaction (DDI) [150], содержащий 792 текста из MEDLINE [151] и DrugBank [152]. В нём размечено около 5 000 взаимодействий между лекарственными препаратами. Для каждой пары определён один из четырёх классов взаимодействий: «совет» (advice), «эффект» (effect), «механизм» (mechanism) или «неуточнённое взаимодействие» (int). Корпус DDI стал одним из первых широко используемых стандартов для тестирования систем по выявлению лекарственных взаимодействий.

Особый интерес представляет EU - ADR corpus, который охватывает 300 аннотированных MEDLINE-аннотаций [153]. Его особенностью является объединение сразу трёх типов связей: «лекарственный препарат – заболевание», «лекарственный препарат – лекарственный препарат» и «ген – заболевание». Благодаря этому EU - ADR стал удобным ресурсом для тестирования универсальных систем извлечения ассоциаций.

Современным ресурсом является BioRED (Biomedical Relation Extraction Dataset), включающий около 600 PubMed-аннотаций. В нём размечены семь различных типов отношений: «ген – заболевание», «химическое соединение – заболевание», «химическое соединение – ген» и другие. Важной особенностью корпуса является наличие как положительных, так и отрицательных примеров ассоциаций, что делает его особенно ценным для обучения алгоритмов машинного обучения с учителем [154].

1.5.2.2. Примеры методов машинного обучения, которые используются для построения моделей извлечения ассоциаций из текстов

Методы машинного обучения, применяемые для извлечения ассоциаций из биомедицинских текстов, можно разделить на две основные группы: классическое машинное обучение и разнообразные архитектуры нейронных сетей.

В первой группе – методов классического машинного обучения - наиболее часто используется метод опорных векторов (SVM), позволяющий классифицировать пары сущностей на основе заранее заданных признаков. Так, в рамках задачи DDIExtraction на соревновании SemEval-2013 Task 9 были представлены системы, основанные на SVM, которые используют для описания текстов признаки bag-of-words, TF-IDF и синтаксические зависимости. Лучшие решения достигали F_1-score в диапазоне 0,65–0,69 на задаче классификации взаимодействий между лекарственными препаратами [155].

Вторая группа включает в себя различные вариации архитектур искусственных нейронных сетей, включая уже упомянутые расширения рекуррентных LSTM и BiLSTM, сверточные (Convolutional Neural Networks, CNN), графовые (Graph Neural Networks, GNN), а также трансформеры. Эти модели позволяют автоматически извлекать контекстные признаки и учитывать порядок слов.

В исследовании [156] была использована архитектура CNN для анализа клинических текстов, направленная на выявление отношений между медицинскими сущностями. В другой работе авторы использовали CNN для задачи извлечения связей «химическое соединение–заболевание» на корпусе BioCreative V CDR [157]. Авторы дополнили архитектуру символьными векторными представлениями, что позволило улучшить обработку редких и ранее не встречавшихся словоформ, характерных для биомедицинской терминологии.

В работе Wang et al. [158] была предложена Bi-LSTM модель для извлечения взаимодействий между лекарственными препаратами из корпуса DDI. Благодаря

включению синтаксической информации (dependency features) модель достигла значений F_1-score, равных 0,72.

Трансформерные модели, как и в случае с распознаванием наименований, занимают лидирующую позицию среди всех алгоритмов машинного обучения для извлечения ассоциаций между объектами. BioBERT, обученный на текстах PubMed и PMC, показал F_1-score 0,78 на задаче извлечения связей «химическое соединение – заболевание» (CDR) и 0,76 на задаче извлечения связей «химическое соединение–белок» (ChemProt). SciBERT, обученный на более широкой коллекции научных текстов, также продемонстрировал хорошие значения точности (F_1-score 0,74–0,77) на задачах извлечения ассоциаций.

В работе [159] предложен подход к извлечению ассоциаций, выходящих за пределы одного предложения. Авторы рассматривают задачу извлечения связей, включающих сразу несколько сущностей, которые могут быть распределены по разным предложениям в тексте. Для её решения используется комбинация предобученной трансформерной модели BERT и графовых трансформеров (Graph Transformer). Дополнительно реализован механизм «внимания к соседям» (neighbor-attention), который ограничивает область внимания токена его ближайшими соседями в графе, что снижает уровень шума при анализе длинных текстов и удалённых зависимостей. Эксперименты показали, что предложенный метод обеспечивает значимый прирост качества. В частности, на задачах извлечения выявления связей «химическое соединение – заболевание» (корпус CDR) значения F_1-score составили 0,66, что примерно на 2,55 лучше, чем в ранее опубликованной наиболее точной модели.

Проведённый анализ литературы показывает, что в области интеллектуального анализа биомедицинских текстов накоплен значительный опыт как в задаче распознавания наименований объектов, так и в задаче извлечения ассоциаций между ними. Использование методов на основе словарей и правил позволяет достигать высокой точности при решении профильных задач, однако

такие подходы требуют значительных усилий при формировании словарей и правил. Алгоритмы машинного обучения, включая методы глубокого обучения, обеспечивают высокую эффективность и возможность анализа сложных лингвистических зависимостей, однако их применение ограничено необходимостью наличия больших размеченных корпусов текстов и значительными вычислительными затратами.

Несмотря на прогресс развития методов интеллектуального анализа текстов, остаются нерешёнными несколько ключевых проблем. Во-первых, большинство существующих решений разрабатывались для ограниченного набора сущностей и типов отношений, что снижает их применимость в более широких биомедицинских исследованиях. Во-вторых, значительная часть моделей остаётся «чёрным ящиком», что затрудняет интерпретацию результатов и интеграцию с существующими базами данных. Наконец, в литературе отсутствует единый подход, который одновременно обеспечивал бы распознавание различных типов сущностей, извлечение их ассоциаций и удобное представление результатов в форме, пригодной для последующего анализа и интеграции с другими источниками данных.

ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ

2.1. Создание коллекции текстов с информацией о наименованиях низкомолекулярных органических соединений и данными об их биологической активности

2.1.1. Стратегия формирования репрезентативной коллекции текстов

Основой для формирования коллекции текстов послужила библиографическая база данных PubMed, поскольку она является ключевым агрегатором научных публикаций в области биологии и медицины, а также содержит инструменты для удобного программируемого доступа. Для обеспечения релевантности отбираемых публикаций и фокуса на первоисточниках, содержащих оригинальные экспериментальные данные, были разработаны и апробированы два независимых подхода: метод итеративного поиска на основе контролируемого словаря MeSH и метод машинного обучения. Выбор подхода определялся спецификой решаемой задачи.

2.1.1.1. Метод итеративного поиска с использованием MeSH-терминов

Данный метод был применен для задач, требующих формирования тематически узкой и релевантной выборки публикаций.

На первом этапе формулировался начальный поисковый запрос с использованием ключевых слов и логических операторов для PubMed. Запрос структурировался таким образом, чтобы каждая его часть отражала ключевой аспект исследуемой проблемы. Окончательный запрос формировался путем объединения концептов оператором AND: (Компонент 1) AND (Компонент 2) AND (Компонент 3) AND (Компонент 4).

На втором этапе осуществлялась экспертная оценка релевантности результатов, полученных по первоначальному запросу. Для релевантных статей анализировался набор присвоенных им MeSH-терминов. Выявлялись пересечения между множествами MeSH-терминов каждой отдельной статьи. Полученный список использовался для формирования уточненного поискового запроса,

частично или полностью заменявшего ключевые слова. Данный процесс повторялся итеративно. Критерием завершения цикла модификации запроса служила субъективная экспертная оценка, направленная на максимизацию доли релевантных публикаций в выдаче при сохранении репрезентативного объема выборки текстов.

На протяжении всех итераций применялись фильтры PubMed для исключения определенных типов публикаций, таких как, например, обзоры литературы, не являющиеся первоисточником знаний (NOT Review[Publication Type]).

2.1.1.2. Метод машинного обучения для категоризации текстов

Этот метод был исследован в рамках работы как потенциальный инструмент автоматизации, однако был признан целесообразным для применения лишь в отдельных сценариях ввиду высокой трудоемкости подготовки обучающих данных.

Формирование выборки релевантных текстов путем ручного отбора и разметки текстов релевантных публикаций, например, описывающих исследования по *in vitro* тестированию соединений. Затем тексты обучающей выборки (и в последствии всех валидационных/тестовых) были подвергнуты предобработке, которая включала в себя удаление стоп-слов с применением пакета NLTK для Python.

Прежде, чем приступить к обучению модели, тексты были описаны тремя разными способами:

- Бинарные векторы. На основе корпуса релевантных текстов вычислялась мера TF-IDF для всех терминов. После ручной верификации и исключения нерелевантных высокочастотных терминов (например, however), отбиралось N терминов с наивысшими значениями TF-IDF (исследовались значения N = 100, 150, 200). Каждый документ затем представлялся в виде вектора фиксированной длины N, где значениями выступали нули и единицы. Первые присваивались позиции вектора в том случае, если ключевое слово встречалось в тексте, а последние – если не встречалось.

- Векторы со значением метрики TF-IDF. При формировании вектора используется принцип, описанный ранее. Однако на позициях вектора вместо нулей и единиц указаны значения метрики TF-IDF для каждого отдельного ключевого слова.
- Векторные представления документов (эмбеддинги). Для получения семантических представлений целых документов использовалась предобученная модель, доступная в библиотеке spaCy.

На векторизованных данных обучались классические алгоритмы машинного обучения (метод опорных векторов, случайный лес, метод k-ближайших соседей) из библиотеки scikit-learn [160]. Отдельно исследовалась модель на основе нейронной сети с архитектурой LSTM, которая принимала на вход только векторные представления документов; данная модель была реализована с использованием библиотеки Keras [161] для Python 3.11.

2.1.1.3. Мультиклассовая и многометочная категоризация на основе базы данных ChEMBL

Для задачи категоризации текстов по типам исследований использовались данные базы ChEMBL, полученные через соответствующий API (Application Programming Interface, программный интерфейс приложений). Поскольку тексты публикаций могли описывать несколько типов исследований, были опробованы два подхода: единая модель мультиклассовой классификации и ансамбль из нескольких бинарных классификаторов (по одному на каждую категорию, например, cell-based, organism-based) с последующим объединением получаемых прогнозов для присвоения документу нескольких меток.

2.2. Разработка интегрального подхода для извлечения сведений о биологической активности низкомолекулярных органических соединений из текстов

При разработке интегрального подхода для извлечения сведений о биологической активности низкомолекулярных органических соединений из

текстов мы придерживались общепринятого порядка действий, описанного в разделе «Обзор литературы».

2.2.1. Формирование выборки релевантных текстов

Для интегрального подхода нами был выбран метод для отбора релевантных публикаций, основанный на MeSH-терминах, ввиду своей универсальности. Дополнительно, с применением программного интерфейса приложения (API) для базы данных MeSH-терминов мы предприняли попытки автоматизации процесса уточнения запроса: ввод ключевого слова в качестве объекта поиска позволяет получить список MeSH-терминов, отсортированных по релевантности в соответствии с заложенным внутри алгоритмом (рисунок 2), что позволяет довольно точно сформировать запрос при итеративном поиске.

```python
from Bio import Entrez
Entrez.email = "nad.smol@gmail.com"

handle = Entrez.esearch(db="mesh", term="antiviral")
record = Entrez.read(handle)

for elem in record["IdList"][:10]:
    handle = Entrez.esummary(db="mesh", id=elem)
    record = Entrez.read(handle)
    print(elem, record[0]['DS_MeshTerms'][0])
```

```
82000998 Antiviral Agents
68064547 Myxovirus Resistance Proteins
68064130 CRISPR-Associated Proteins
68064113 CRISPR-Cas Systems
68064112 Clustered Regularly Interspaced Short Palindromic Repeats
68054788 Ribosome Inactivating Proteins
68051199 Toll-Like Receptor 7
68047208 Eosinophil-Derived Neurotoxin
68039282 Mirabilis
68038301 Virus Inactivation
```

А)

```python
from Bio import Entrez
Entrez.email = "nad.smol@gmail.com"

handle = Entrez.esearch(db="mesh", term="side effect")
record = Entrez.read(handle)

for elem in record["IdList"][:10]:
    handle = Entrez.esummary(db="mesh", id=elem)
    record = Entrez.read(handle)
    print(elem, record[0]['DS_MeshTerms'][0])
```

```
68064420 Drug-Related Side Effects and Adverse Reactions
68016048 Dideoxyadenosine
68016047 Zalcitabine
68014150 Antipsychotic Agents
68011584 Psychology
68010115 Oxyphenonium
68003520 Cyclophosphamide
67559524 UTL-5g compound
67110016 oxoindole lysozyme
67083898 proacame
```

Б)

Рисунок 2 - Фрагмент кода на Python и результаты выдачи при обращении к базе данных MeSH-терминов через программный интерфейс для ключевых слов А) «antiviral» и Б) «side effect».

В данной работе предусмотрена возможность автоматической загрузки текстов в соответствии с запросом пользователя. Более того, поскольку интерес пользователя может быть ограничен определенным, заранее сформированным объемом данных, мы также предоставляем возможность загрузки готовых текстов или списка идентификаторов (PMID) статей.

Обращение к базам данных MeSH и PubMed реализовано с применением библиотеки Python Bio (модуль Entrez) [162].

2.2.2. Распознавание наименований объектов

Для корректного установления биологической активности низкомолекулярных органических соединений в текстах необходимо учитывать не только сами соединения, но и широкий спектр сопутствующих объектов, с которыми они взаимодействуют. В данной работе к числу таких объектов были отнесены химические соединения, белки и гены, заболевания, биологические виды, клеточные линии, генотипы HLA, миРНК, а также однонуклеотидные полиморфизмы и аминокислотные замены. Учет этих сущностей принципиально важен: например, для корректного понимания упоминания «ингибитор» необходимо знать, на какой белок действует соединение, а сообщение о «противовирусной активности» требует указания конкретного вида вируса.

Для распознавания наименований объектов применялись как алгоритмы машинного обучения, так и методы, основанные на правилах. Для распознавания наименований химических соединений, белков, генов, заболеваний, биологических видов и клеточных линий применялись модели машинного обучения, тогда как для HLA-генотипов, миРНК и полиморфизмов – алгоритмы, основанные на регулярных выражениях.

Для распознавания химических соединений, белков и генов, а также заболеваний была реализована модель на основе условных случайных полей. Обучение проводилось на специализированных аннотированных корпусах: для химических соединений использовался корпус CHEMDNER, для белков и генов - DrugProt, для заболеваний - NCBI Diseases и CDR. На предварительном этапе

тексты подвергались препроцессингу, включавшему токенизацию. При этом для каждого класса объектов применялись разные правила: химические соединения сегментировались по пробелам и символам, что позволяло учитывать семантически значимые структурные элементы в систематических названиях (например, указания на заместители), тогда как для заболеваний и белков/генов токенизация проводилась преимущественно по знакам препинания и скобкам, что отражает особенности их записи в биомедицинских текстах (рисунок 3).

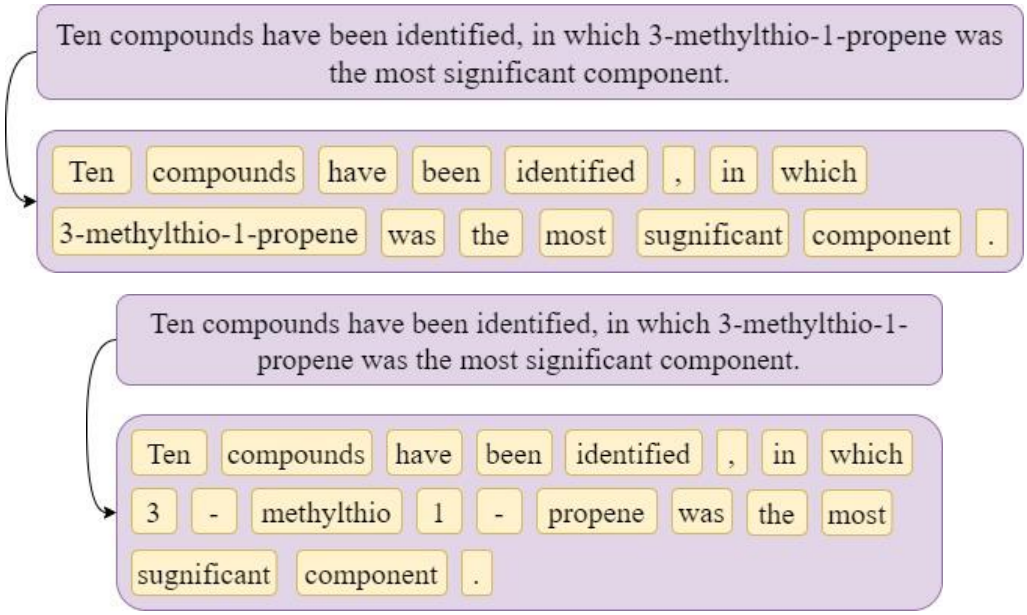

Рисунок 3 - Различные подходы к токенизации текста.

Каждый токен описывался набором признаков, включавших семантические и морфологические дескрипторы. Среди семантических дескрипторов учитывалась принадлежность к стоп-словам и к неспецифическим терминам, которые могут указывать на объект, но не называть его напрямую (например, «chemical» или «inhibitor» для химических соединений, «protein» или «interaction» для белков и генов). Морфологические признаки включали длину токена, наличие цифр и специальных символов, что особенно важно для распознавания номенклатурных форм. Поскольку значение токена определяется контекстом, в описание добавлялись признаки соседних токенов, что позволяло модели учитывать локальные связи в пределах фразы. Пример набора дескрипторов для отдельных

токенов приведён в таблице 2, а схема формирования признаков с учётом контекста представлена на рисунке 4.

Таблица 2 – Набор признаков токенов, который использовался при обучении модели на основе алгоритма условных случайных полей

№	Признак	Тип	Значение
1	**word**	string	Токен
2	**lower**	string	Токен в нижнем регистре
3	**isUpper**	Boolean	Записан ли токен в верхнем регистре
4	**isTitle**	Boolean	Является ли первый символ токена заглавным
5	**isDigit**	Boolean	Является ли токен числом
6	**hasDigits**	Boolean	Содержит ли токен цифры
7	**isNonSpecific**	Boolean	Является ли токен неспецифическим термином
8	**isStopWord**	Boolean	Является ли токен стоп-словом
9	**hasSymbols**	Boolean	Содержит ли токен знаки
10	**word[n-3:n]**	string	Последние три символа токена
11	**word[n-2:n]**	string	Последние два символа токена
12	**firstChar**	string	Первый символ токена
13	**length**	integer	Число символов в токене
14	**posTag**	string	Часть речи

Catalase-like and peroxidase-like catalytic activities ***of silicon nanowire*** arrays.

{***'word'***: 'silicon', ***'word.lower()'***: 'silicon', ***'word.isupper()'***: 0, ***'word.istitle()'***: False,
'word.isdigit()': False, ***'word.havedigits()'***: 0, ***'word.isNonSpecific'***: 0,
'word.isStopWord': 0, ***'word.isSymbol'***: 0, ***'word[-3:]'***: 'con', ***'word[-2:]'***: 'on',
'word.FirstSymbol': 's', ***'word.CharNumber'***: 7, ***'postag'***: 'NN',
'-1:word': 'of', ***'-1:word.lower()'***: 'of', ***'-1:word.isupper()'***: False, ***'-1:word.istitle()'***: False,
'-1:word.isdigit()': False, ***'-1:word.havedigits()'***: 0, ***'-1:word.isNonSpecific'***: 0,
'-1:word.isStopWord': 1, ***'-1:word.isSymbol'***: 0, ***'-1:word[-3:]'***: 'f', ***'-1:word[-2:]'***: 'of',
'-1:word.FirstSymbol': 'o', ***'-1:word.CharNumber'***: 2, ***'-1:postag'***: 'IN',
'+1:word': 'nanowire', ***'+1:word.lower()'***: 'nanowire', ***'+1:word.isupper()'***: False,
'+1:word.istitle()': False, ***'+1:word.isdigit()'***: False, ***'+1:word.havedigits()'***: 0,
'+1:word.isNonSpecific': 0, ***'+1:word.isStopWord'***: 0, ***'+1:word.isSymbol'***: 0,
'+1:word[-3:]': 'ire', ***'+1:word[-2:]'***: 're', ***'+1:word.FirstSymbol'***: 'n',
'+1:word.CharNumber': 8, ***'+1:postag'***: 'NN'}

Рисунок 4 – Иллюстрация учета признаков контекста при описании токенов в формате словаря Python.

Разметка сущностей осуществлялась по схеме SOBIE (Single, Outside, Beginning, Inside, End), которая позволяет одновременно корректно обрабатывать однословные и многословные наименования. После разметки выполнялось обучение модели и оценка её качества методом пятикратной кросс-валидации. Для повышения точности проводилась оптимизация гиперпараметров CRF, включая коэффициенты регуляризации L1- и L2, позволяющие учитывать разреженность признаков и степень переобучения, а также параметры, отвечающие за вес переходов между соседними метками последовательности. Результатом работы модели являлся набор токенов с метками принадлежности, после чего выполнялась процедура конкатенации токенов в цельные наименования, приближенные к их исходному виду в тексте.

Помимо собственных моделей, в работе была использована предобученная система HunFlair, показавшая высокую эффективность на ряде биомедицинских корпусов. Однако при анализе новых текстов, особенно статей по COVID-19, нами было отмечено, что HunFlair хуже справляется с ранее не встречавшимися

наименованиями [163]. В связи с этим применялся консенсусный подход: итоговый результат определялся на основании совпадений предсказаний модели CRF и HunFlair, что позволило повысить точность распознавания. Данные о точности работы HunFlair приведены в оригинальной публикации, а сводные результаты её работы представлены в таблице 3. Хотя в статье и документации по модели HunFlair упомянуто, что она также позволяет проводить распознавание наименований клеточных линий, точность для этой функциональной компоненты нам не удалось найти.

Таблица 3 – Точность распознавания наименований с применением модели HunFlair.

Группа наименований	**LSTM-CRF (F1-score)**
Химические соединения	0,91
Белки/гены	0,89
Заболевания	0,87
Биологические виды	0,81

Для объектов, обладающих устойчивыми форматами записи, применялись методы, основанные на регулярных выражениях в синтаксисе библиотеки *re* языка Python. Шаблоны создавались вручную на основании анализа текстов и позволяли корректно идентифицировать, например, HLA-генотипы (HLA-A*02:01), миРНК (hsa-miR-21) и SNP (rs1234567).

После распознавания все наименования автоматически соотносились с уникальными идентификаторами в соответствующих базах данных: для химических соединений и белков/генов – ChEMBL, для заболеваний – Human Disease Ontology, для биологических видов – NCBI Taxonomy [164]. Соотнесение считалось успешным, если распознанное наименование совпадало с одним из синонимов в записи базы данных. Такой подход позволил устранить дублирующие связи и унифицировать результаты для последующего анализа.

2.2.3. Извлечение ассоциаций

После распознавания объектов следующим этапом являлось извлечение ассоциаций между ними. Для этой задачи мы использовали подход, основанный на правилах, предполагающий поиск в тексте так называемых фраз-шаблонов. Под фразой-шаблоном понимается фрагмент текста (одно или несколько слов), который авторы научных статей используют для обозначения связи между двумя объектами. Например, чтобы показать, что химическое соединение применяется в лечении заболевания, часто употребляются слова *treatment, regimen, drug, intervention*. Если же речь идёт о побочных эффектах, в текстах встречаются выражения *adverse effect, side effect, lead to* и другие.

Список таких фраз-шаблонов был собран нами вручную на основе анализа корпуса статей. Так как одни и те же связи могут описываться синонимичными выражениями, все фразы были объединены в смысловые группы, как было рассмотрено ранее (*treatment, regimen, drug, intervention* – применение в терапии, *adverse effect, side effect, lead to* – побочные эффекты и токсичность)

Для того чтобы учесть все возможные словоформы одного термина, мы использовали полуавтоматический подход: сначала сгенерировали формы слова с помощью библиотеки NLTK, а затем вручную проверили корректность полученного списка. Это позволило повысить полноту извлечения ассоциаций.

При извлечении связей учитывались все возможные комбинации вида «объект 1 – объект 2 – фраза-шаблон» внутри одного предложения. Чтобы исключить ложные результаты, сразу исключались комбинации, которые не имеют биологического смысла. Например, клеточная линия не может выступать в качестве метода терапии заболевания, поэтому даже если в предложении встречаются слова термины, обозначающие клеточную линию и заболевание, а также фраза-шаблон, например, «treatment», они не интерпретируются как связанная этой фразой-шаблоном пара.

В одном предложении может встречаться несколько объектов и несколько фраз-шаблонов. Использование исключительно совместной встречаемости наименований и фраз-шаблонов может привести к генерации значительного числа ложно положительных примеров [165]. Для того, чтобы этого избежать, мы интегрировали в алгоритм численную оценку (score), которая является мерой расстояния в количестве токенов между фразой-шаблоном и парой объектов. Основной принцип ее использования заключается в следующем: если фраза-шаблон действительно связывает именно эту пару, то в предложении она должна находиться ближе к объектам, чем к другим словам.

Формально оценка вычисляется как сумма минимальных расстояний между позициями токенов объектов и токенами фразы-шаблона:

$$score = min(|I_{BO_1} - I_{BO_2}|, |I_{BO_1} - I_{EO_2}|, |I_{EO_1} - I_{BO_2}|, |I_{EO_1} - I_{EO_2}|) + min(|I_{BO_1} - I_{BPH}|, |I_{BO_1} - I_{EPH}|, |I_{EO_1} - - I_{BPH}|, |I_{EO_1} - I_{EPH}|) + min(|I_{BO_2} - I_{BPH}|, |I_{BO_2} - I_{EPH}|, |I_{EO_2} - I_{BPH}|, |I_{EO_2} - I_{EPH}|)$$

где:

- I_{BO1}, I_{EO1} – позиции первого объекта (начало и конец),
- I_{BO2}, I_{EO2} – позиции второго объекта,
- I_{BPH}, I_{EPH} – позиции фразы-шаблона (начало и конец).

Оценка отражает минимальное расстояние между объектами и фразой-шаблоном в тексте. Чем оно меньше, тем выше вероятность, что данная фраза действительно связывает конкретно эту пару объектов. Для каждой комбинации в предложении вычислялась своя оценка, и достоверными считались те ассоциации, у которых значение было минимальным (рисунок 5).

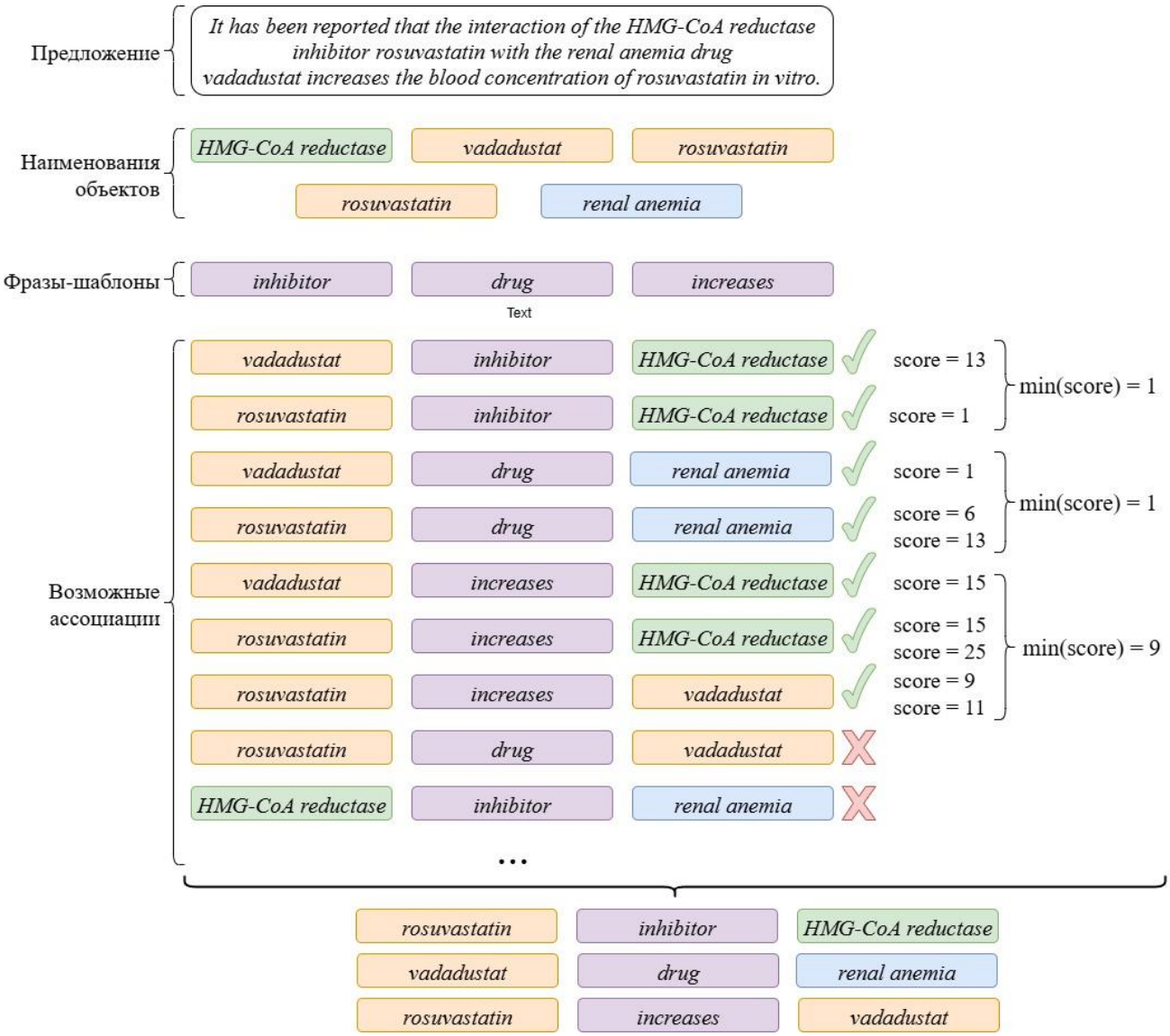

Рисунок 5 – Иллюстрация работы алгоритма извлечения ассоциаций, основанного на использовании фраз-шаблонов.

В результате такой процедуры можно извлечь из текстов набор пар «объект – объект», связанных между собой биологически осмысленными фразами-шаблонами.

2.3. Создание структуры базы данных для хранения извлеченных сведений о биологической активности низкомолекулярных органических соединений

При разработке базы данных для хранения извлеченных сведений мы предусмотрели реализацию трех необходимых составляющих:

1) Тексты научных публикаций. При каждом запуске интегрального подхода инициализируется процедура загрузки и обработки текста. Загруженные и обработанные тексты необходимо хранить для дальнейшего анализа извлеченных ассоциаций с возможностью их сортировки. Во-первых, это помогает интерпретировать полученные результаты без необходимости дополнительного поиска текста первоисточника в специализированных базах данных текстовой информации, а во-вторых, позволяет избежать повторного анализа уже включенных текстов. Включение в базу данных метаданных о публикации (год публикации, авторы, журнал, и др.) позволяют ограничивать область поиска (например, периодом освещения в литературе, спектром специализированных научных журналов и так далее)
2) Наименования распознанных сущностей. За счет наличия этой компоненты, возможно автоматизированное составление словарей наименований, которые упоминаются в текстах, с учетом их регистра и возможных односимвольных модификаций в написании (например, в наименованиях белков зачастую варьируется написание изоформ с греческой буквы: в них может быть включен или не включен дефис)
3) Извлеченные ассоциации. Помимо факта установления ассоциации с помощью разработанного алгоритма необходимо хранить сработавшие правила при извлечении, а также фрагменты текста, в которых была упомянута ассоциация. Хранение подобных сведений позволяет интерпретировать извлеченную ассоциацию: как с точки зрения семантики (сработавшее правило), так и с точки зрения ее силы (часть текста; фиксация известного факта, как во введении, или описание полученных результатов, как в обсуждении)

4) Сведения об объектах в фактографических базах данных. Данный компонент позволяет связать наименование с его реальным объектом. Здесь должны хранится сведения, полученные при нормализации наименований на внешние, уникальные идентификаторы. Как и в случае с компонентом для хранения текстов, это избавляет от необходимости прямого обращения к внешним базам данных для уточнения дополнительной информации, что может значительно сократить время при анализе полученных при автоматизированном анализе текстов результатов.

Нами была разработана логическая структура реляционной базы данных, учитывающая реализацию всех описанных компонентов. Реализация базы данных произведена с использованием системы управления базами данных (СУБД) MySQL.

2.4. Валидация разработанных алгоритмов интеллектуального анализа текстов в задаче поиска сведений о химических соединениях, перспективных с точки зрения противовирусной терапии

Качество работы и полнота извлечения данных с помощью разработанных алгоритмов интеллектуального анализа текстов оценивались на примере группы фармацевтических субстанций с установленной или потенциальной противовирусной активностью.

Для проведения валидации были сформированы две коллекции текстов:

1. Коллекция текстов, посвященных известным противовирусным препаратам. Формирование выборки проводилось на основе предварительно составленного перечня лекарственных средств, имеющих в базе данных ChEMBL показание «viral disease» среди прочих в поле рекомендованных назначений. С использованием API ChEMBL была агрегирована основная информация и синонимичные наименования для каждого препарата из сформированного списка. Полученный список наименований использовался для поиска релевантных публикаций в базе данных PubMed. Дополнительная фильтрация результатов поиска осуществлялась с применением MeSH-

терминов и фильтров по типам публикаций для выделения исследований различных категорий: *in vitro*, *in vivo* и клинические испытания.

2. Коллекция текстов, посвященных соединениям с потенциальной противовирусной активностью. Формирование этой выборки проводилось путем поиска по PubMed с использованием MeSH-термина «Antiviral agents» с последующей аналогичной фильтрацией по типам исследований. Для исключения дублирования тексты, вошедшие в первую коллекцию, были исключены из второй выборки.

К обеим сформированным коллекциям текстов был применен разработанный алгоритм интеллектуального анализа. Дополнительно, на этапе последующей обработки результатов проводилась фильтрация извлеченных ассоциаций с целью сохранения только тех связей, которые включали объекты типа «Species» с указанием на вирусы. Для этого при нормализации наименований биологических видов с использованием базы данных NCBI Taxonomy учитывалась иерархическая классификация, что позволило идентифицировать и сохранить только ассоциации, связанные с вирусами.

Результаты были систематизированы и представлены в табличном (для детального анализа) и графическом (для визуальной интерпретации) форматах.

2.5. Проверка полноты сведений о биологической активности химических соединений в фактографических базах данных и базах знаний

Для оценки полноты и репрезентативности данных, извлеченных из текстовых источников, был проведен сравнительный анализ с информацией, содержащейся в фактографической базе данных ChEMBL. Обращение к базе данных осуществлялось программно через API в 2024 году (актуальная на момент обращения версия 32).

В качестве тестового набора данных были выбраны противовирусные соединения. Для формирования выборки исследований (bioassays) из ChEMBL был применен фильтр по таксономической принадлежности биологической мишени: отбирались только те записи, у которых в поле «Organism Taxonomy L1» было

указано «Viruses». Данный подход позволил сфокусироваться именно на тех экспериментах, которые непосредственно касались тестирования противовирусной активности.

Процедура проверки полноты сведений состояла из двух основных этапов. На первом этапе проводилось сравнение охвата объектов исследований. Для каждой установленной текстовыми методами ассоциации «химическое соединение – биологическая мишень» выполнялся поиск соответствующей записи в ChEMBL. Идентификация химических соединений осуществлялась по их уникальным идентификаторам ChEMBL (ChEMBL ID). Идентификация биологических мишеней проводилась с использованием: ChEMBL ID - для белков и генов, NCBI Taxonomy ID - для вирусов.

На втором этапе анализировалась полнота охвата первичных литературных источников. Сравнивались перечни публикаций, на которые имелись ссылки в отобранных записях об исследованиях в ChEMBL, с перечнем публикаций, обработанных в рамках настоящего исследования.

Для количественной оценки результатов анализа использовался показатель процентного совпадения, рассчитываемый как отношение числа ассоциаций, обнаруженных в обеих системах (текстовой и фактографической), к общему числу ассоциаций, извлеченных из текстов.

ГЛАВА 3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

3.1. Коллекции текстов, содержащих знания о биологической активности низкомолекулярных химических соединений

Нами были разработаны и апробированы различные методики для автоматизированного отбора релевантных публикаций, которые были затем использованы для формирования представительной коллекции текстов, содержащих знания о биологической активности низкомолекулярных органических соединений. Сформированные корпусы охватывают различные аспекты взаимодействия биологически активных соединений с живыми системами, что обеспечивает комплексную проверку разработанных алгоритмов интеллектуального анализа текстов.

3.1.1. Применение методов фильтрации текстов для отбора релевантных публикаций

Как было указано в разделе «Материалы и методы», в работе использовались две основные группы подходов, каждая из которых обладает своими преимуществами и ограничениями.

Алгоритмы, основанные на методах машинного обучения, обеспечивают более высокую точность при формировании выборки релевантных текстов. Однако их применение требует предварительной подготовки обучающих (аннотированных) выборок для каждой конкретной задачи, что делает процесс трудоёмким и снижает универсальность подхода.

В то же время метод, основанный на использовании поисковых запросов с применением MeSH-терминов, отличается простотой реализации и позволяет формировать выборки для широкого круга тематических областей. Его недостатком является меньшая полнота и точность результатов, обусловленные отсутствием аннотаций у части публикаций, а также влиянием человеческого фактора при назначении терминов.

Сравнительный анализ показал, что оба подхода могут быть эффективны в зависимости от целей исследования: методы машинного обучения предпочтительны в задачах, требующих высокой точности и контроля качества выборки, тогда как MeSH-ориентированные запросы более уместны для быстрого получения обобщённых наборов текстов и первичного анализа.

3.1.1.1. Использование автоматизированных запросов к PubMed

Один из подходов подразумевает использование MeSH-терминов и фильтрацию по типам публикации для отбора релевантных источников. Анализ MeSH-терминов статей, релевантных различным биологическим задачам выявил, что типы исследования аннотируются определенным списком MeSH-терминов. Так, например, *in vitro* исследования на клетках в аннотации, как правило, имеют MeSH-термин, характеризующий клеточную линию, а *in vivo* исследования – животных (собаки, мыши и другие).

Одним из преимуществ использования MeSH-терминов является то, что при формировании запроса нет необходимости его детализировать, указывая элементы словаря для каждого конкретного объекта. В запрос могут быть включены родительские, группирующие MeSH-термины; и в случае, если статья аннотирована одним из дочерних, она все равно попадет в результаты выдачи (рисунок 6).

Рисунок 6 – Иллюстрация возможности использования иерархической структуры MeSH-терминов для формирования обобщенного запроса к PubMed.

По этой причине в нашей работе мы использовали следующие обобщенные MeSH-термины для получения научных публикаций, описывающих результаты доклинических исследований:

- Cells, Cultured – *in vitro*;
- Animals – *in vivo*.

В качестве демонстрации мы приводим результаты анализа случайных публикаций, которые содержали обозначенные MeSH-термины для *in vitro* и *in vivo* исследований (см. Приложение 1). Анализ показал, что они в целом отражают

исследования *in vivo* и *in vitro* соответственно. Для Animals большинство статей действительно описывали эксперименты на животных, однако встречались и нерелевантные случаи (обзоры, биобанки, анализы данных). В то же время Cells, Cultured оказался более точным индикатором: почти все публикации содержали эксперименты на клеточных линиях или культурах. Для отбора научных публикаций, описывающих результаты клинических испытаний, применение соответствующего MeSH-термина (Humans) не позволило получить удовлетворительных результатов, поскольку данный термин во многих случаях сопровождает работы, посвященные болезням человека, даже если напрямую пациент не выступает в качестве объекта исследования.

В отличие от доклинических исследований, которые могут быть описаны с использованием соответствующих MeSH-терминов (например, *in vitro* или *in vivo*), публикации по клиническим испытаниям имеют дополнительные особенности. Такие исследования требуют обязательной регистрации и фиксированного протокола, что отражается не только в содержании статей, но и в типах этих публикаций. Во многих научных журналах наряду с обозначением «research article» дополнительно указывается более специфическая категория, например «clinical trial» или её вариации, что позволяет однозначно идентифицировать публикации данного типа. Аналогичная классификация воспроизводится и в базе данных PubMed. Поэтому при формировании выборки статей, посвящённых клиническим испытаниям, мы использовали именно характеристику типа публикаций как основной критерий поиска, что оказалось более надёжным, чем использование MeSH-терминов.

Полный список типов публикаций, который используется для аннотации, доступен по ссылке [165]. Мы проанализировали данный список и выделили только те типы публикаций, которые соответствуют клиническим испытаниям на людях; клинические испытания ветеринарных лекарственных препаратов не рассматривались. Полный список типов публикаций для отбора статей, описывающих результаты клинических испытаний, приведен ниже:

- Adaptive Clinical Trial
- Case Reports
- Clinical Study
- Clinical Trial
- Clinical Trial, Phase I
- Clinical Trial, Phase II
- Clinical Trial, Phase III
- Clinical Trial, Phase IV
- Controlled Clinical Trial
- Meta-Analysis
- Multicenter Study
- Network Meta-Analysis
- Observational Study
- Pragmatic Clinical Trial
- Randomized Controlled Trial
- Twin Study

Хотя описания отдельных клинических случаев (Case Reports) обладают меньшим уровнем доказательности по сравнению, например, с рандомизированными клиническими испытаниями, мы их включили в рассмотрение по причине необходимости анализа всесторонних сведений о биологической активности низкомолекулярных химических соединений.

Использование фильтрации по типам публикаций также позволило нам включать в выборки текстов только первоисточники информации, то есть оригинальные статьи, не дублируя извлекаемую информацию за счет литературных и систематических обзоров:

- Review
- Systematic Review

Таким образом, комбинируя MeSH-термины, описывающие объекты исследования, с фильтрами по типам публикаций и MeSH-терминами,

релевантными различным форматам исследований, возможно составить запрос для отбора коллекции текстов, сфокусированной на поставленной задаче (рисунок 7). Формирование корректного запроса производилось с использованием ряда логических операторов: «OR» для перечисления синонимичных или схожих по смыслу (например, типы клинических испытаний) фрагментов запроса, «AND» для объединения различных по смыслу компонентов запроса (например, типы исследований) и «NOT» для фрагментов запроса, которые должны быть строго исключены из результатов (например, систематические обзоры).

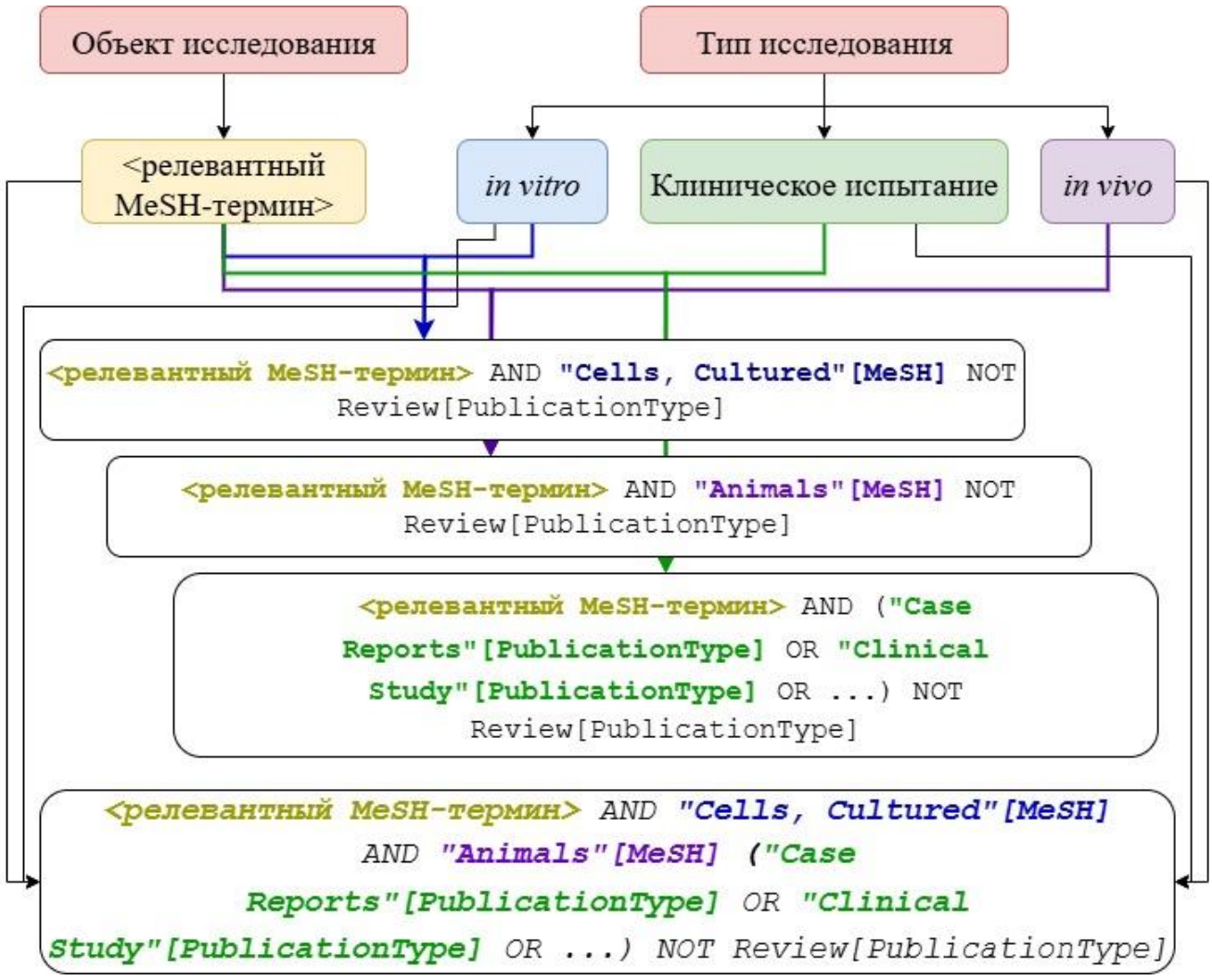

Рисунок 7 – Иллюстрация алгоритма формирования запроса на основе MeSH-терминов и публикаций для отбора релевантных публикаций, описывающих доклинические и клинические исследования объекта.

3.1.1.2. Применение алгоритмов машинного обучения для классификации текстов

Для обучения и валидации моделей классификации текстов были сформированы две выборки публикаций. Первая включала аннотации статей, посвящённых *in vitro* тестированию противовирусной активности в отношении SARS-CoV-2. В неё вошли 300 текстов, половина из которых была отнесена к релевантным, а половина – к нерелевантным. Релевантными считались статьи, в аннотациях которых прямо или косвенно упоминалось проведение тестирования на выделенных и очищенных белках или заражённых клетках вируса. К нерелевантным были отнесены публикации, описывающие *in silico* исследования, популяционные наблюдения и другие подходы, не связанные с непосредственным экспериментальным тестированием противовирусной активности соединений.

Вторая выборка была сформирована из аннотаций публикаций, связанных с биологическими тестами (bioassays), представленных в базе данных ChEMBL и посвящённых исследованию активности химических соединений в отношении вирусов. Общий объём этой выборки составил около 1000 текстов. Для неё использовалась многоклассовая разметка: статьи распределялись по типам экспериментов, а также по категориям, определённым в BioAssay Ontology (BAO) [166, 168]. В ChEBML выделяются следующие типы экспериментов: *binding assays*, характеризующие способность соединений связываться с биомишенями; *functional assays*, описывающие функциональный ответ биологической системы; *ADME assays*, связанные с процессами абсорбции, распределения, метаболизма и выведения; *toxicity assays*, направленные на выявление токсичности соединений; *physicochemical assays*, определяющие физико-химические свойства; и *unclassified assays*, объединяющие тесты, не попадающие в указанные категории.

Форматы BAO позволяют более детально охарактеризовать условия проведения экспериментов. В частности, *organism-based format* описывает тесты, проводимые на уровне целого организма; *tissue-based format* – эксперименты с использованием тканей; *cell-based format* – тесты на клеточных культурах; *cell line или cell-free format* – эксперименты на клеточных линиях или в бесклеточной

системе; *subcellular format* – тесты на выделенных субклеточных структурах (например, митохондриях или ядрах); *cell membrane format* – анализ взаимодействий на уровне мембранных фракций; *single protein format и protein complex format* – тесты, в которых в качестве мишени выступает один белок или комплекс белков; *nucleic acid format* – эксперименты с использованием ДНК или РНК в качестве мишени; категория *other* объединяет все форматы, не подходящие под перечисленные.

Для оценки качества классификации текстов были протестированы различные алгоритмы машинного обучения с применением разных типов текстовых дескрипторов. В таблице 4 представлены средние показатели точности (Precision), полноты (Recall) и F_1-score для выборок с соотношением классов (50/50) и (25/75).

Таблица 4 – Средние значения точностей моделей классификации текстов, полученные при пятикратной кросс-валидации.

	50/50			25/75		
	P	R	F1	P	R	F1
	TF-IDF					
Случайный лес	0,79	0,77	0,73	0,57	0,80	0,67
Метод опорных векторов	0,80	0,67	0,73	0,50	0,60	0,55
Метод k-ближайших соседей	0,67	0,33	0,44	0,56	0,50	0,53
	Векторные представления слов					
Искусственные нейронные сети	0,82	0,90	0,86	0,73	0,80	0,76

Полученные результаты показали, что использование TF-IDF и классических моделей (случайный лес, метод опорных векторов) позволяет достигать приемлемого уровня качества классификации, особенно на сбалансированных выборках, однако эти модели оказались менее устойчивы к дисбалансу классов. Алгоритм k-ближайших соседей продемонстрировал наихудшие результаты в обоих сценариях, что объясняется его низкой устойчивостью к размерности признакового пространства и неравномерному распределению классов.

Наилучшие результаты при прогнозе меток релевантности текстов тестовой выборки были достигнуты при использовании рекуррентной нейронной сети с архитектурой LSTM (рисунок 8). На вход модели подавался вектор длиной 300, содержащий последовательность индексов слов в словаре размером 30 000 терминов, сформированном на этапе токенизации. Слой *Embedding* преобразовывал этот вектор в матрицу размерностью 300 × 600, где число строк соответствовало длине входной последовательности, а число столбцов – размерности векторного представления слова. Каждое слово текста представлялось в виде плотного вектора признаков.

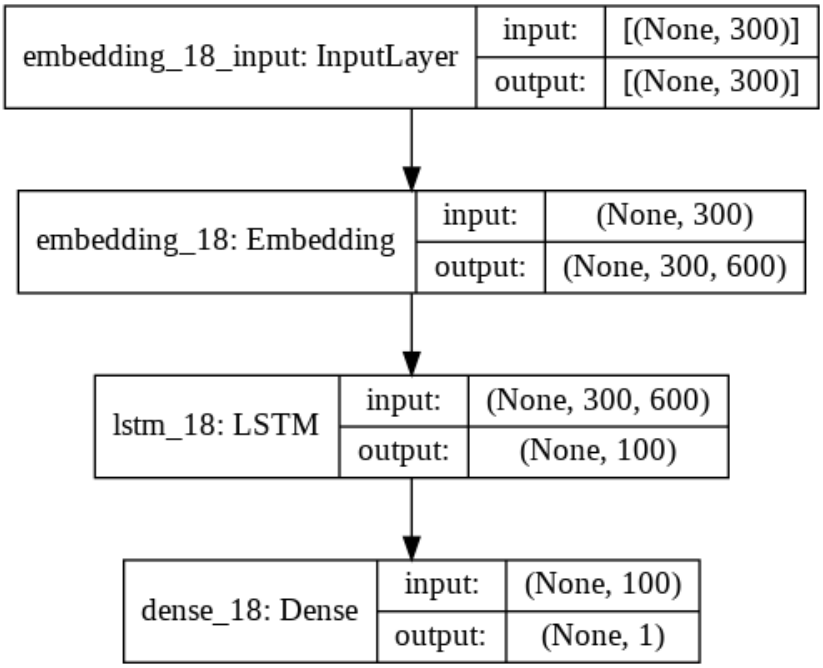

Рисунок 8 – Схема архитектуры искусственных нейронных сетей со слоем LSTM, которая использовалась для классификации текстов.

Далее данные поступали в слой LSTM со 100 нейронами, что позволяло учитывать контекстные зависимости между словами и эффективно работать с длинными предложениями. Для предотвращения переобучения в модели использовался механизм *dropout* (0,1). Выходной слой с функцией активации *sigmoid* выполнял бинарную классификацию текстов на релевантные и

нерелевантные. В качестве функции потерь применялось Mean squared error, обучение проводилось в течение пяти эпох.

Применение LSTM обеспечило значительное повышение полноты при сохранении высокой точности классификации по сравнению с традиционными моделями на основе TF-IDF. Такой результат подтверждает важность учёта последовательного характера текста и использования контекстно-зависимых признаков для анализа научных публикаций.

Для задачи мультиклассовой классификации текстов аннотаций, на которые ссылаются исследования (bioassays) ChEMBL, были использованы два подхода на базе одного и того же инструмента – классификатора spaCy с предобученной моделью SciSpaCy (en_core_sci_lg) [169]. В первом случае модель применялась в режиме мультиклассового категоризатора, что позволило напрямую относить каждый текст к одному из классов BAO. Во втором случае был построен набор бинарных классификаторов, каждый из которых решал задачу вида «относится к классу – не относится к классу».

Применение мультиклассовой модели показало выраженную зависимость качества от выбранного порога оценки, которая является мерой вероятности отнесения к классу. При низких порогах (0,1) полнота достигала очень высоких значений (recall ≈ 0,89), однако точность была низкой (precision ≈ 0,49) (таблица 5). При увеличении порога до диапазона 0,35–0,45 наблюдался оптимальный баланс: precision ≈ 0,79–0,82, recall ≈ 0,78–0,80, а F1-мера достигала 0,80 (таблица 5). При дальнейшем повышении порога точность продолжала расти, но за счёт снижения полноты, что не улучшало итогового качества.

Таблица 5 – Значения точности и полноты для моделей классификации форматов BAO (порог 0,35–0,45)

Подход/Формат	**Precision**	**Recall**	**F1-score**
Multiclass (spaCy)	0,79	0,80	0,71
Single protein format	0,83	0,89	0,86
Cell-based format	0,82	0,63	0,71
Assay format	0,74	0,86	0,80

Бинарные классификаторы обеспечивают более высокие и устойчивые результаты для отдельных форматов. Наилучшие показатели были получены для *single protein format*: при пороге 0,35–0,45 точность составила 0,83, полнота – 0,89, F1 – 0,86 (таблица 5). Для *assay format* оптимальные значения также были высоки (precision 0,74, recall 0,86, F1 = 0,80) (таблица 5). Для *cell-based format* качество оказалось несколько ниже (F1 = 0,71) из-за ограниченной полноты (0,63) при высокой точности (0,82) (таблица 5). Остальные форматы были исключены из анализа в связи с недостаточным числом примеров.

На рисунке 9 представлена динамика значений precision, recall и F_1-score в зависимости от порога классификации для мультиклассовой модели и бинарного классификатора для *single protein format*. Видно, что бинарная модель остаётся более устойчивой, тогда как мультиклассовая демонстрирует характерный компромисс между полнотой и точностью.

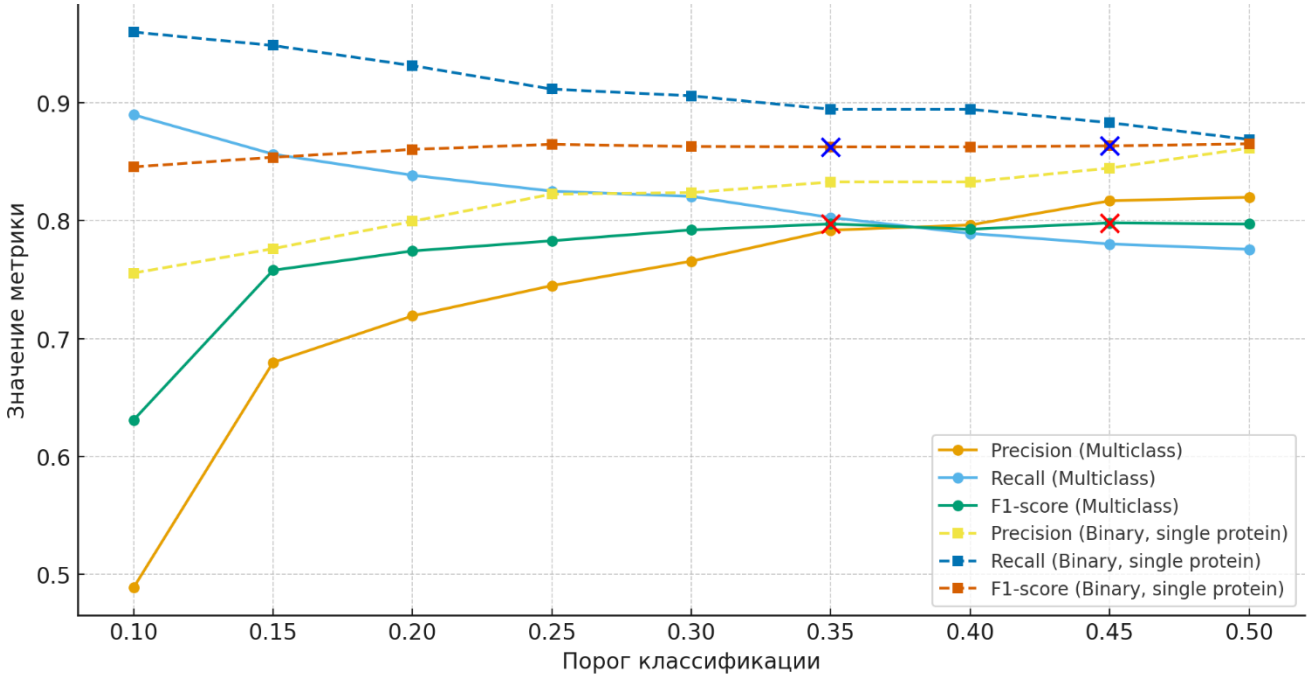

Рисунок 9 – Precision, Recall и F1-score в зависимости от порога для мультиклассовой модели и бинарного классификатора (*single protein format*)

Мультиклассовая модель spaCy подходит для общей категоризации текстов, но бинарные классификаторы обеспечивают более высокое качество для отдельных категорий BAO, особенно для *single protein format*. Это подтверждает целесообразность комбинированного подхода: мультиклассовая классификация – для широкой категоризации, бинарные модели – для уточнения результатов в ключевых категориях.

Несмотря на то, что алгоритмы машинного обучения характеризуются высокой точностью, их разработка является трудоемким процессом ввиду необходимости аннотации обучающих коллекций текстов, что ограничивает их применение для широкого спектра задач, в том числе для тех, для которых пока не получены соответствующие выборки обучающих коллекций текстов. В противопоставление этому, применение сконструированных запросов на основе MeSH-терминов и типов публикаций несколько уступает в точности, но не отнимает много времени у эксперта и может быть автоматизировано. Именно по

этой причине для дальнейшего извлечения информации мы использовали преимущественно второй подход.

3.1.2. Специализированные коллекции текстов

В ходе работы были сформированы и проанализированы несколько коллекций аннотаций публикаций из базы PubMed, полученных с использованием MeSH-запросов и уточняющих ключевых слов (таблица 6). Все коллекции содержат тексты аннотаций, идентификаторы PMID и метаданные (журнал, MeSH-термины, ключевые слова и др.), сохранённые в формате TXT с разделителем табуляций. Указанные коллекции текстов были сформированы в процессе работы над конкретными задачами в рамках отдельных научных проектов.

Таблица 6 – Характеристики специализированных коллекций текстов

Коллекция	**Объём**	**Запрос/критерий отбора**	**Особенности содержания**	**Назначение**
XenoMet	~1000	(«Chemicals and Drugs Category»[Mesh]	Вручную аннотирован: соединения, метаболиты, реакции биотрансформации, ассоциации	Обучение CRF и валидация извлечения знаний о путях биотрансформации
Противовирусная активность	>400 000	Наименования противовирусных препаратов + «Antiviral Agents»[Mesh]	Публикации о разносторонней биологической активности химических соединений, ингибирующих вирусные инфекции	Масштабное тестирование алгоритмов поиска и анализа противовирусных соединений
Сигнальные пути в	>160 000	«Diseases Category"[Mesh]	Роль сигнальных	Валидация извлечения

патогенезе заболеваний		AND "Signal Transduction»[Mesh]	путей при нейродегенеративных, неопластических и инфекционных заболеваниях	информации о механизмах патогенеза
Сигнальный путь Hedgehog	~10 000	«Hedgehog Proteins»[Mesh] AND «Signal Transduction»[Mesh] AND «Neoplasms»[Mesh]	Исследования роли пути Hedgehog в онкогенезе	Проверка точности алгоритмов на узкоспециализированной тематике
Большое депрессивное расстройство	~30 000	«Depressive Disorder, Major»[Mesh]	Патогенез и терапия депрессии	Изучение возможности применения разработанных алгоритмов в случае, если заданный объект исследования отличен от химического соединения или его мишени

Аннотированный корпус XenoMet (~1000 текстов) был получен по запросу *(«Chemicals and Drugs Category»[Mesh] OR «Metabolism»[Mesh] OR «metabolism»[Subheading]) AND (Humans[MeSH])* и вручную аннотирован. В нём выделялись родительские соединения, метаболиты, другие химические соединения и реакции биотрансформации, а также фиксировались ассоциации между соединениями и реакциями (какая реакция приводит к образованию конкретного метаболита). Этот корпус использовался не только для извлечения информации, но

и для обучения модели CRF, применяемой к задаче распознавания наименований и анализа путей биотрансформации.

Корпус по противовирусной активности (>400 000 текстов) был сформирован на основе запроса, который включал как наименования лекарственных препаратов с известной противовирусной активностью, так и MeSH-термин *«Antiviral Agents»[Mesh]*. Коллекция включает публикации как о препаратах с клинически подтверждённой активностью, так и о соединениях, исследуемых на стадии доклинических испытаний.

Корпус по сигнальным путям в патогенезе заболеваний (>160 000 текстов) был сформирован с использованием запроса *«Diseases Category»[Mesh] AND «Signal Transduction»[Mesh]* и включает публикации о роли сигнальных путей при нейродегенеративных, неопластических, инфекционных и других заболеваниях.

Специализированный корпус по сигнальному пути Hedgehog (~10 000 текстов) получен на основе запроса *«Hedgehog Proteins»[Mesh] AND «Signal Transduction»[Mesh] AND «Neoplasms»[Mesh]*. Коллекция сфокусирована на изучении роли сигнального пути Hedgehog в онкогенезе.

Корпус по большому депрессивному расстройству (~30 000 текстов) был собран по запросу *«Depressive Disorder, Major»[Mesh]*. Он включает публикации, посвящённые молекулярным механизмам патогенеза и подходам к терапии депрессии.

При формировании каждого из корпусов текстов к запросу был добавлен фрагмент, который позволяет проводить фильтрацию типов исследования: доклинические и клинические, не включая литературные и систематические обзоры.

Сформированные коллекции различаются как по объёму (от 1 000 до более чем 400 000 текстов), так и по тематике (метаболизм, противовирусная активность, сигнальные пути, психические заболевания). Это позволило протестировать алгоритмы в условиях, требующих как высокой масштабируемости при анализе больших массивов, так и высокой точности при работе с узкоспециализированными

данными. Аннотированный корпус XenoMet, в отличие от остальных коллекций, предоставил возможность для точной количественной оценки качества извлечения информации.

3.2. Интегральный подход для извлечения сведений о биологической активности низкомолекулярных органических соединений из текстов

Разработанный нами интегральный подход к анализу научных текстов представляет собой многоэтапную процедуру, включающую формирование выборки релевантных публикаций, распознавание наименований химических и биологических объектов, извлечение ассоциаций между ними, фильтрацию и нормализацию данных с привязкой к объектам в специализированных базах данных, а также представление результатов в табличном или графическом формате (рисунок 10).

Рисунок 10 – Общая схема разработанного интегрального подхода

Необходимость разработки подобного подхода обусловлена растущим объёмом научной литературы и фрагментарностью представленных в ней данных.

Несмотря на наличие отдельных методов решения задач по выделению сущностей или извлечению связей, их использование не позволяет получить целостного представления о биологической активности соединений. Интегральный подход позволяет объединить отдельные этапы в единую последовательность и обеспечивает системный анализ сведений, охватывающий как молекулярные механизмы действия соединений, так и их возможные терапевтические применения, побочные эффекты и роль в патогенезе заболеваний.

Важное значение разработанного метода обусловлено тем, что его применение впервые позволяет агрегировать широкий спектр информации о низкомолекулярных органических соединениях, создавая основу для построения знаний о взаимосвязях между химическими соединениями и биологическими объектами различного уровня организации.

3.2.1. Распознавание наименований

Разработанный интегральный подход включает в себя как распознавание наименований биомедицинских объектов, так и извлечение ассоциаций между ними. На этапе распознавания сущностей нами были использованы две модели: собственная модель на основе условных случайных полей (CRF) и предобученная модель HunFlair (LSTM-CRF).

Полученные значения точности для разработанной модели CRF (таблица 7) оказались сопоставимыми с результатами, опубликованными ранее в соответствующих исследованиях [118, 140, 170]. Особенно важно отметить, что модель CRF проявила устойчивость к вариативности наименований, включая систематические названия, аббревиатуры и морфологические вариации, что важно для анализа новых объектов.

Таблица 7 – Точность модели распознавания наименований химических и биологических объектов на основе алгоритма условных случайных полей

	Химические соединения (CHEMDNER)			Белки (DrugProt)			Патологии (NCBI Disease + CDR)		
	10 000 аннотаций			15 000 аннотаций			2 293 аннотаций		
	P	R	F_1	P	R	F_1	P	R	F_1
Средние значения	0,91	0,87	0,89	0,87	0,84	0,85	0,79	0,68	0,72

P – precision, R – recall, F_1 – F_1-score

На основе разработанного алгоритма распознавания наименований с применением метода условных случайных полей был разработан свободно-доступный веб-ресурс SigNER [171]. Этот веб-сервис предоставляет возможность распознавания наименований в научных публикациях путем прямой загрузки файлов или текстов в систему, или же с использованием автоматической загрузки текстов по PubMed идентификаторам статей. Фрагмент его веб-интерфейса с результатами распознавания наименований представлен на рисунке 11.

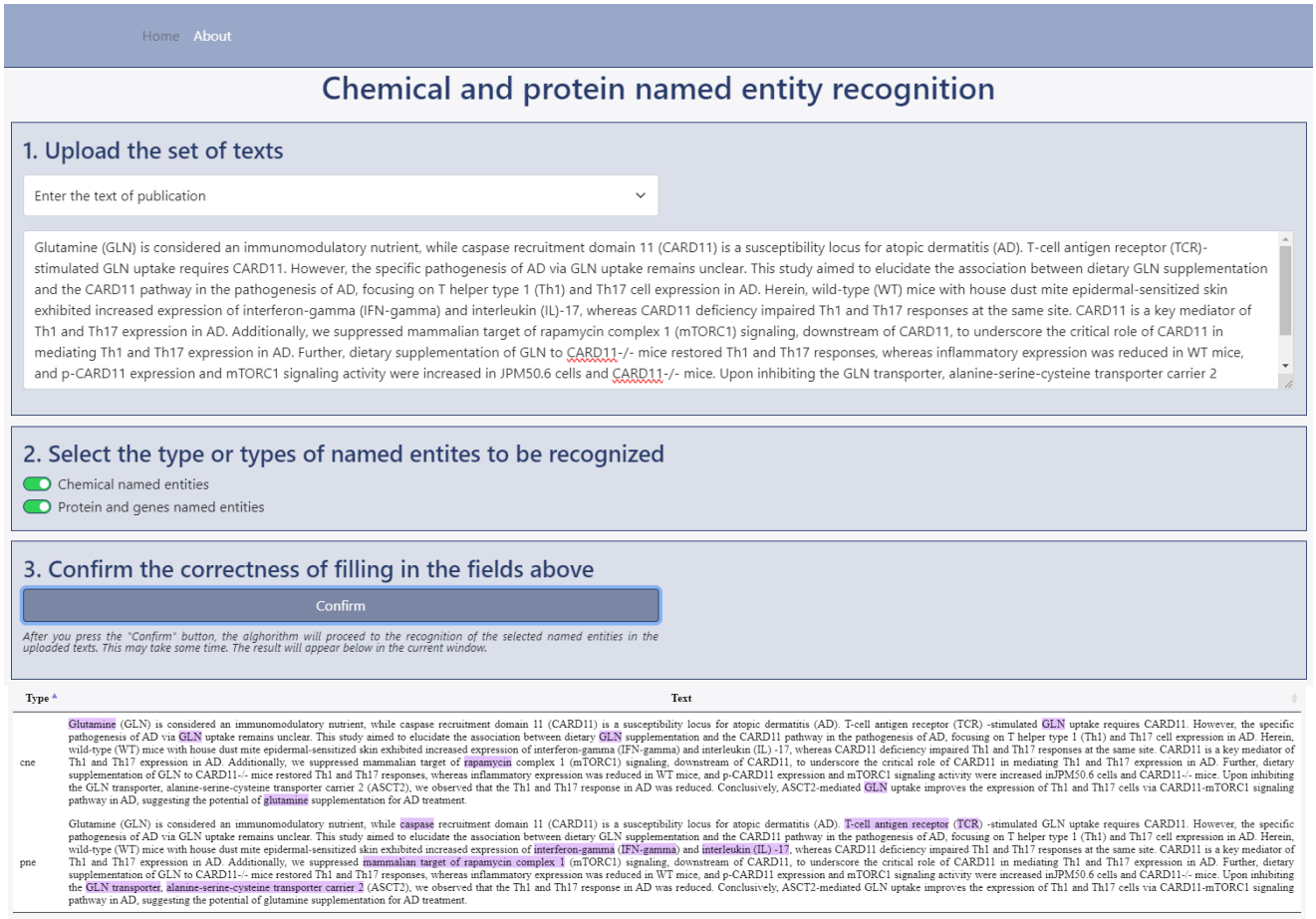

Рисунок 11 – Фрагмент веб-интерфейса ресурса SigNER с выполненным прогнозом

Для распознавания наименований миРНК, генотипов HLA и однонуклеотидных полиморфизмов/аминокислотных замен были разработаны списки шаблонов (регулярных выражений). Их список представлен в таблице 8.

Таблица 8 – Список регулярных выражений, которые использовались для распознавания наименований HLA-генотипов, миРНК и однонуклеотидных полиморфизмов/аминокислотных замен

Тип наименования	**Регулярные выражения**
миРНК	• miR-\w+-\d[a-z], MiR-\w+-\d[a-z] • miR-\w+, MiR-\w+ • microRNA-\w+-\d[a-z], MicroRNA-\w+-\d[a-z] • microRNA-\w+, MicroRNA-\w+
Полиморфизмы	• rs\d+ • [A-Za-z]\d+[A-Za-z] • [A-Za-z]\d+[A-Za-z]/[A-Za-z]
Генотипы HLA	• HLA-[A-Z]*\d+ • HLA[A-Z]\d+ • HLA-[A-Z]*\d+:\d+ • HLA-[A-Z]\d+ • \(HLA\)-[A-Z]\d+ • \(HLA\)-[A-Z]\(*\)\d+ • HLA-[A-Z]\(*\)\d+ • HLA allele [A-Z]\d+ • HLA-[A-Za-z]+*\d+ • HLA-[A-Z]\d+:\d+ • HLA-[A-Z]+\d+*\d+:\d+ • HLA-[A-Z]+\d+*\d+:\d+:\d+[A-Z]

Данный список фраз-шаблонов позволяет извлекать подавляющее большинство наименований указанных объектов, что было выявлено путем экспертной оценки.

Распознанные наименования объектов с помощью предоставленных инструментов программируемого доступа были соотнесены с соответствующими объектами в базах данных, как было описано в разделе Материалы и методы. В

случае, если среди списка синонимов всех объектов в базах данных наименование не было найдено, оно сохранялось в том виде, в котором присутствовало в тексте.

3.2.2. Извлечение ассоциаций

Нами был сформирован список фраз-шаблонов для извлечения ассоциаций между различными объектами. Как уже упоминалось ранее, для некоторых из ассоциаций рассматривалась исключительно совместная встречаемость (таблица 9); это обусловлено отсутствием специфических семантических связок в текстах. Например, в предложении «…native and H275Y mutant (oseltamivir-resistant) neuraminidases (NAs) of influenza A virus…» [172] отсутствуют любые семантические связки между наименованиями «H275Y», «neuraminidases» и «influenza A virus», не считая предлога «of», который является общераспространенным.

Таблица 9 – Список извлекаемых ассоциаций между различными типами объектов с примерами биологических активностей и указанием метода извлечения ассоциаций.

N	**Типы объектов**	**Способ извлечения ассоциаций**	**Примеры биологических активностей**
1	Chemical-Chemical	Фразы-шаблоны	Комбинации препаратов, используемых в терапии; влияние лекарственного препарата на биохимические показатели; межлекарственные взаимодействия; метаболизм и биотрансформация
2	Chemical-Species	Фразы-шаблоны	Применение в терапии (часто наименования заболеваний, вызванных вирусом, распознаются, как вид); применение в профилактике; резистентность/восприимчивость к лекарственным препаратам

3	Chemical-Gene	Фразы-шаблоны	Влияние лекарственного препарата на биохимические показатели; воздействие на белок/ген
4	Chemical-Disease	Фразы-шаблоны	Побочные эффекты и токсичность; применение в терапии; применение в профилактике; возможные биомаркеры заболеваний
	Chemical-miRNA	Совместная встречаемость	Воздействие химических соединений на регуляторные механизмы
5	Chemical-SNP	Фразы-шаблоны; Совместная встречаемость	Взаимосвязь между аминокислотной/нуклеотидной заменой и устойчивостью/восприимчивостью к препарату
6	Chemical-Genotype	Фразы-шаблоны	(Как правило – ассоциации генотипов HLA с гиперчувствительностью к лекарственным препаратам)
7	Species-Species	Фразы-шаблоны	Ко-инфекции
8	Species-Gene	Фразы-шаблоны	Возможные биомаркеры инфекций; белки и гены, вовлеченные в молекулярные механизмы инфекций
9	Species-Disease	Фразы-шаблоны	Причина заболевания
10	Species-miRNA	Совместная встречаемость	миРНК, вовлеченные в молекулярные механизмы инфекций
11	Species-SNP	Совместная встречаемость	Как дополнительный параметр – уточнить, замены белков каких вирусов приводят к устойчивости к лекарственным препаратам
12	Gene-Gene	Фразы-шаблоны	Взаимодействия между белками (в т.ч. сигнальные пути); ассоциации типа «часть-целое» (белок-семейство)
13	Gene-Disease	Фразы-шаблоны	Возможные биомаркеры; молекулярные механизмы патогенеза; способ терапии

14	Gene-miRNA	Совместная встречаемость	Участие в регуляции
15	Gene-SNP	Совместная встречаемость	Как дополнительный параметр – уточнить, замены каких белков приводят к устойчивости к лекарственным препаратам (зачастую эта информация отсутствует)
16	Disease-Disease	Фразы-шаблоны	Основное заболевание-симптом; коморбидности; многокомпонентные заболевания
17	Disease-miRNA	Совместная встречаемость	Как компонент ассоциации Gene-miRNA, для уточнения
18	Disease-SNP	Совместная встречаемость	Ассоциации изменений в белках/генах с манифестацией и патогенезом заболеваний
19	Disease-Genotype	Есть связки; Совместная встречаемость	Как компонент ассоциации Chemical-Genotype; «гиперчувствительность» сама по себе распознается как наименование заболевания

При формировании списка фраз-шаблонов мы использовали иерархическую классификацию с той целью, чтобы в дальнейшем возможно было проводить выявление интересующих исследователя биологических активностей. На верхнем уровне она включает такие категории (биологические активности), как «применение в терапии», «применение в профилактике», «побочные эффекты и токсичность», «биомаркеры», «воздействие на гены и белки» (положительная и отрицательная регуляция, связывание, ингибирование, активация), «метаболизм и биотрансформация», «лекарственные взаимодействия», «коинфекции и коморбидности», «основное заболевание-симптом», «методы терапии», «лекарственная резистентность» и ряд других. Внутри каждой категории были выделены смысловые группы (например, для терапии: *treat, administer, prescribe, regimen, co-administration, pharmacotherapy*), а для каждой группы определены конкретные лексемы и их словоформы, включая различные грамматические варианты.

Всего было сформировано более тысячи фраз-шаблонов, объединённых в 44 категории, соответствующих типам биологической активности (таблица 10). Наибольшее их количество относится к категориям, связанным с терапией, токсичностью и молекулярным воздействием лекарственных соединений, тогда как для областей, связанных с миРНК и однонуклеотидными полиморфизмами, число устойчивых выражений оказалось крайне ограниченным. В этих случаях извлечение ассоциаций осуществлялось на основе совместной встречаемости терминов в пределах одного предложения. Например, для связей *Chemical–miRNA* авторы часто ограничиваются описанием совместного упоминания вещества и миРНК без использования характерных глаголов. Аналогичным образом, в случае *Disease–SNP* информация об ассоциации указывается через упоминание варианта гена у пациентов без выделения явного маркера.

Таблица 10 – Типы ассоциированных объектов и соответствующие им типы биологической активности, выделенные при классификации фраз-шаблонов, с указанием числа смысловых групп и среднего количества фраз-шаблонов для них.

Типы ассоциированных объектов	**Категории (типы биологических активностей)**	**Количество смысловых групп**	**Среднее количество фраз-шаблонов на смысловую группу**
Chemical - Disease	Side effects	6	23,5
Chemical - Microorganism	Side effects	6	22,5
Chemical - Chemical	Drug-drug interactions (unspecified), changes in clinical indicators	5	7,2
Chemical - Gene/Protein	Impact on protein/gene	5	36,4
Gene/Protein - Gene/Protein	Impact on protein/gene	4	34,0
Chemical - Mutation	The study of mutations associated with	4	7,8

	stability/susceptibility to therapy		
Disease - Microorganism	Cause of the disease	3	22,7
Gene/Protein - Gene/Protein	Change in clinical indicators/level of macromolecules OR unspecified effect	3	23,0
Chemical - Gene/Protein	Change in clinical indicators/ level of macromolecules	3	29,0
Microorganism - Microorganism	Co-infections	3	14,0
Chemical - Disease	Has an effect	2	50,0
Chemical - Microorganism	Has an effect	2	50,0
Disease - Gene/Protein	Cause of the disease	2	14,5
Chemical - Chemical	Biotransformation	2	70,5
Chemical - Chemical	Drug-drug interactions	2	10,0
Disease - Disease	Comorbidities	2	52,0
Gene/Protein - Gene/Protein	Regulation indicating impact type	2	9,0
HLA genotype - Disease	-	1	7,0
HLA genotype - Chemical	-	1	1,0
Chemical - Disease	Application in prevention	1	8,0
Chemical - Disease	Application in therapy	1	102,0
Chemical - Disease	Biomarkers	1	19,0
Chemical - Disease	Investigated regarding	1	25,0
Chemical - Microorganism	Application in prevention	1	8,0
Chemical - Microorganism	Application in therapy	1	102,0

Chemical - Microorganism	Biomarkers	1	19,0
Chemical - Microorganism	Investigated regarding	1	25,0
Chemical - Microorganism	Resistance to the drug	1	2,0
Disease - Gene/Protein	Biomarkers	1	38,0
Disease - Gene/Protein	Therapy method	1	8,0
Microorganism - Gene/Protein	Biomarkers	1	2,0
Chemical - Chemical	Drugs prescribed in combination	1	42,0
Disease - Disease	Cause of the disease	1	6,0
Disease - Disease	Complex diseases	1	12,0
Disease - Disease	Main disease-symptom	1	49,0
Disease - Disease	Mimicry	1	5,0
Disease - Microorganism	Comorbidities	1	20,0
Disease - Microorganism	Complex diseases	1	4,0
Disease - Microorganism	Main disease-symptom	1	4,0
Gene/Protein - Gene/Protein	Change in clinical indicators/ level of macromolecules	1	6,0
Gene/Protein - Gene/Protein	Interactions between proteins/genes (not specified)	1	6,0
Gene/Protein - Gene/Protein	Protein/Gene families	1	1,0
Chemical - Gene/Protein	Change in clinical indicators/level of macromolecules OR unspecified effect	1	1,0
Mutation - Gene/Protein	The study of mutations associated with stability/susceptibility to therapy	1	11,0

Качество извлечения ассоциаций было проверено двумя способами. Сначала мы провели автоматическую оценку на фрагменте корпуса BioRED, включающем связи типа «химическое соединение – белок» и «белок – белок», где отношения классифицировались по бинарному принципу («есть/нет»). Дополнительно была выполнена ручная верификация выборки текстов, содержащих описания механизмов действия ингибиторов ангиотензин-превращающего фермента (АПФ). Полученные результаты представлены в таблице 11: при автоматической оценке значения составили Precision = 0,78, Recall = 0,91, F1-score = 0,74, а при ручной проверке – Precision = 0,80, Recall = 0,85, F1-score = 0,82. Следует подчеркнуть, что при ручной оценке к ложноположительным относились даже те случаи, когда ассоциация была извлечена корректно, но ошибочно классифицирован её тип, что делает данный вариант проверки более строгим.

Таблица 11 – Точность извлечения ассоциаций между объектами с применением разработанного нами и описанных в литературе подходов.

	Precision	**Recall**	**F_1-score**
Наш метод (автоматическое сравнение, корпус bioRED)	0,78	0,91	0,74
Наш метод (ручная верификация)	0,80	0,85	0,82
Сверточные + рекуррентные ИНС [170]	0,74	0,84	0,78
Графовые ИНС [171]	-	-	0,9

Сопоставление с результатами, представленными в литературе, показывает, что гибридные сверточно-рекуррентные нейронные сети достигают Precision = 0,74, Recall = 0,84 и F1-score = 0,78 [173], а графовые архитектуры достигают F_1-score до 0,90 [174]. Таким образом, разработанный нами метод обеспечивает показатели, сопоставимые с более ресурсоёмкими нейросетевыми подходами, но

при этом обладает рядом преимуществ: он интерпретируем, относительно прост в расширении за счёт добавления новых шаблонов и требует существенно меньших затрат на аннотирование обучающих корпусов.

3.2.3. Форматы представления результатов извлечения знаний о биологической активности низкомолекулярных органических соединений

Хранение всех извлечённых ассоциаций в том виде, в котором они встречаются в тексте (с указанием публикации, объектов и фразы-шаблона), приводит к формированию чрезвычайно больших массивов строк, которые оказываются затруднительными для анализа. Дополнительной проблемой является повторяемость: многие ассоциации, особенно хорошо изученные, неоднократно упоминаются в разных источниках, что приводит к дублированию записей.

В связи с этим при формировании итогового табличного формата были применены процедуры постобработки:

1. Исключение дубликатов. Фильтрация выполнялась как по самим наименованиям объектов (при отсутствии их идентификаторов в базах данных), так и по комбинациям «объект 1 – объект 2 – фраза-шаблон». При совпадении этих элементов ассоциации группировались в одну.
2. Группировка по типам биологической активности. Здесь использовалась иерархическая структура классификации фраз-шаблонов: если разные шаблоны относились к одному типу активности, ассоциации объединялись в соответствующую группу.

Не менее важным условием было сохранение информации об источниках. Поэтому при любом объединении ассоциаций сохранялись PMID статей, из которых они были извлечены. Дополнительно для каждой ассоциации рассчитывалось количество предложений, в которых она встречалась. Этот параметр служил косвенной мерой достоверности и новизны: высокая частота

упоминаний указывает на хорошо изученные связи, тогда как единичные упоминания могут свидетельствовать о недавно начатых исследованиях.

Стоит отметить, что некоторая информация о биологических и химических объектах, включенная в фактографические базы данных и онтологии, может быть важна для понимания смысла ассоциации, но может отсутствовать в тексте. К такой информации можно отнести: тип белка/гена (семейство или комплекс, конкретный белок/ген), принадлежность белка/гена к организму, принадлежность организма к высшим таксономическим категориям (эукариоты или прокариоты, вирусы или бактерии, растения или млекопитающие), и другая. Эти сведения могут отсутствовать в анализируемом тексте, но существенно повышают информативность при анализе.

С учетом вышеописанного, представление извлеченных ассоциаций в формате таблицы должно включать следующие фрагменты:

1. Объект 1 с указанием его типа (химическое соединение/заболевание/и др.), дополнительной информации из фактографических баз данных с указанием уникального идентификатора.
2. Объект 2 с указанием его типа (химическое соединение/заболевание/и др.), дополнительной информации из фактографических баз данных с указанием уникального идентификатора.
3. Тип биологической активности, который был выявлен при извлечении ассоциации.
4. Количество предложений, из которых была извлечена ассоциация.
5. Идентификаторы (PMID) статей, из которых была извлечена ассоциация.

Пример табличного оформления извлечённых ассоциаций для противовирусной активности низкомолекулярных соединений представлен в Приложении 2.

Несмотря на полноту данных, табличная форма затрудняет быструю интерпретацию большого числа ассоциаций. Для упрощения анализа была разработана система их графического представления. В этом формате объекты отображались как узлы, а ассоциации – как рёбра. Для повышения информативности применялась цветовая кодировка, учитывающая как тип объекта, так и наличие связанного с ним идентификатора в базе данных (см. рисунок 12).

Рисунок 12 – Цветовая разметка типов объектов, используемая для графической визуализации извлеченных ассоциаций. ХС – химическое соединение, БД – фактографическая база данных/онтология, SNP – однонуклеотидные полиморфизмы и аминокислотные замены, miRNA – миРНК, HLA – генотипы HLA

Если при формировании визуальной аннотации для типов объектов достаточно было ограничиться использованием цветовых оттенков ввиду их небольшого числа, то для 44 типов биологической активности и тех ассоциаций,

что были извлечены с применением совместной встречаемости, было необходимо обратить внимание еще и на форму линий, связывающих объекты. Полное соответствие формы и цвета ребер типам биологической активности представлено в таблице 12. На основании количества предложений, из которых была извлечена ассоциация, при каждом построении графического представления рассчитывалась толщина ребра. За максимум была взята толщина ребра в 10 пунктов, а за минимум – в 1. Ассоциации из визуализируемой выборки, которые чаще всего и реже всего упоминалась в предложениях, имеют фиксированную толщину ребра, а для других ассоциаций рассчитывается значение толщины ребра в пунктах, как отношение установленного для нее числа предложений на число предложений, установленное для наиболее часто встречающейся ассоциации, умноженное на 9 (разница в пунктах между самым толстым и самым тонким ребром).

Таблица 12 – Формы и цвета ребер графа и их соответствие типам биологической активности.

Типы ассоциированных объектов	**Тип биологической активности**	**Форма ребра**	**Цвет**
Chemical - Disease	Application in prevention	DASH_DOT (точка-тире)	#696BFF
Chemical - Disease	Application in therapy	DASH_DOT (точка-тире)	#2ECA00
Disease - Gene/Protein	Biomarkers	DOTS (точка)	#CA0022
Chemical - Chemical	Biotransformation	SINEWAVE (волна)	#CA9900
Disease - Gene/Protein	Cause of the disease	EQUAL_DASH (короткое тире)	#9600CA

Gene/Protein - Gene/Protein	Change in clinical indicators or levels of macromolecules	LONG_DASH (длинное тире)	#C400CA
Microorganism - Microorganism	Co-infections	DOTS (точка)	#00CAB8
Disease - Disease	Comorbidities	DOTS (точка)	#00CAB8
Chemical - Chemical	Drug-drug interactions	EQUAL_DASH (короткое тире)	#0077CA
Chemical - Chemical	Drug-drug interactions (NS)	EQUAL_DASH (короткое тире)	#C400CA
Chemical - Chemical	Drugs prescribed in combination	EQUAL_DASH (короткое тире)	#00CAB8
Chemical - Microorganism	Has an effect	DASH_DOT (точка-тире)	#CA9900
Chemical - Gene/Protein	Impact on protein/gene	SINEWAVE (волна)	#C1CA00
Gene/Protein - Gene/Protein	Interactions between proteins/genes (NS)	SINEWAVE (волна)	#00CA46
Chemical - Disease	Investigated regarding	DASH_DOT (точка-тире)	#C400CA
Disease - Disease	Main disease-symptom	DOTS (точка)	#2ECA00
Gene/Protein - Gene/Protein	Protein/Gene families	DOTS (точка)	#C1CA00
Gene/Protein - Gene/Protein	Regulation indicating impact type	SINEWAVE (волна)	#0093CA
Chemical - Disease	Side effects	SOLID (непрерывная)	#CA0022

Gene/Protein - Mutation	The study of mutations associated with stability/susceptibility to therapy	SINEWAVE (волна)	#00CAB8
Chemical - Mutation	The study of mutations associated with stability/susceptibility to therapy	SINEWAVE (волна)	#00CAB8
Disease - Gene/Protein	Therapy method	DASH_DOT (точка-тире)	#CA8100
Many	Associations on co-occurrence	SOLID (непрерывная)	#8D8D88

NS – эффект от взаимодействия объектов не уточняется

Использование визуализации в виде графов делает возможным не только наглядное разделение типов объектов и активностей, но и быструю оценку структуры взаимосвязей, выявление ключевых элементов, к которым сходятся многочисленные связи, а также поиск потенциально новых, но пока малоизученных ассоциаций. Дополнительным преимуществом является возможность применения специализированного программного обеспечения, такого как CytoScape [175], что расширяет возможности последующего анализа (кластеризация, расчёт топологических характеристик и др.).

В итоге использование табличной и графической форм представления информации позволяет совместить детальность и полноту данных с их наглядной интерпретацией. Табличный формат сохраняет максимально возможный объём сведений и обеспечивает аналитическую гибкость, тогда как графический – облегчает восприятие и делает видимой общую структуру связей. Совместное применение этих подходов обеспечивает целостное и удобное для анализа представление извлечённых знаний о биологической активности низкомолекулярных органических соединений.

3.2.4. Применение разработанного интегрального подхода для извлечения информации из биомедицинских текстов

Для демонстрации практической работоспособности и универсальности разработанного алгоритма нами были использованы коллекции текстов, описанные в предыдущем разделе. Эти корпусы различаются по тематике, объёму и уровню детализации, что позволяет проверить метод в различных условиях – от узконаправленных исследований (например, публикации о сигнальном пути Hedgehog) до масштабных массивов литературы о противовирусной активности соединений. Такой выбор обеспечивает не только валидацию алгоритма на специализированных задачах, но и оценку его применимости в условиях высокой вариативности источников и исследовательских контекстов.

В последующих разделах будут приведены конкретные примеры того, как интегральный подход позволяет извлекать и систематизировать ассоциации между биомедицинскими объектами. Каждый из примеров иллюстрирует определённый сценарий применения: выявление метаболических реакций, анализ механизмов патогенеза заболеваний, поиск новых направлений для терапевтического использования химических соединений. Это позволяет не только подтвердить корректность работы алгоритма, но и продемонстрировать его потенциал в качестве инструмента для интеграции знаний из разнородных биомедицинских источников.

3.2.4.1. Исследование роли сигнального пути Hedgehog в патогенезе неопластических заболеваний

Для тестирования разработанного нами интегрального подхода был проведён анализ сигнального пути Hedgehog (Hh), играющего ключевую роль в эмбриональном развитии и вовлечённого в формирование опухолевых процессов [125]. С использованием комбинации методов на основе словарей и алгоритма условных случайных полей были извлечены наименования белков и генов, взаимодействующих с компонентами пути Hedgehog, а также ассоциации между

ними и опухолевыми заболеваниями. Всего было выделено более 2 300 уникальных взаимодействий между белками Hh и другими белками человека, а также около 3 600 ассоциаций «белок – заболевание», что позволило построить сеть взаимодействий, включающую узлы, отвечающие за ключевые процессы канцерогенеза (например, участие белков Shh, Gli1–3, Ptch1 и Smo).

Для оценки корректности извлечённых ассоциаций были использованы данные из корпуса bioRED, что позволило получить F_1-score 0,84 (precision 0,78; recall 0,91). Кроме того, сопоставление с транскриптомными данными из базы OncoDB показало, что часть генов, выявленных как ассоциированные с белками Hh, действительно демонстрирует статистически значимые изменения экспрессии в опухолевых тканях по сравнению с нормальными образцами (p-adj < 0,05, $\log_2$FC > 0,5). Таким образом, результаты анализа не только согласуются с литературными данными, но и подтверждаются экспериментальными исследованиями на реальных биологических образцах.

3.2.4.2. Анализ механизмов большого депрессивного расстройства и возможных методов терапии

Анализ текстов, связанных с большим депрессивным расстройством (БДР), позволил выявить широкий спектр ассоциаций, отражающих как молекулярные механизмы заболевания, так и возможные подходы к терапии. Среди них – подтверждённые в литературе коморбидные связи (например, с зависимостью от алкоголя и никотина), а также механистические цепочки, указывающие на участие в патогенезе заболевания воспалительных процессов. Характерным примером является каскад «интерферон-α – IDO1 – кинуренин – хинолиновая кислота», подтверждающий роль активации иммунного ответа и метаболизма триптофана в патогенезе БДР [176-179].

Особое внимание привлекли молекулярные маркеры: в текстах отмечались ассоциации между экспрессией BDNF (нейротрофического фактора мозга)

(Рисунок 13) и эффективностью антидепрессантов, а также упоминания отдельных микроРНК, включая miR-146a-5p, рассматриваемую как биомаркер ответа на терапию дулоксетином [179-182].

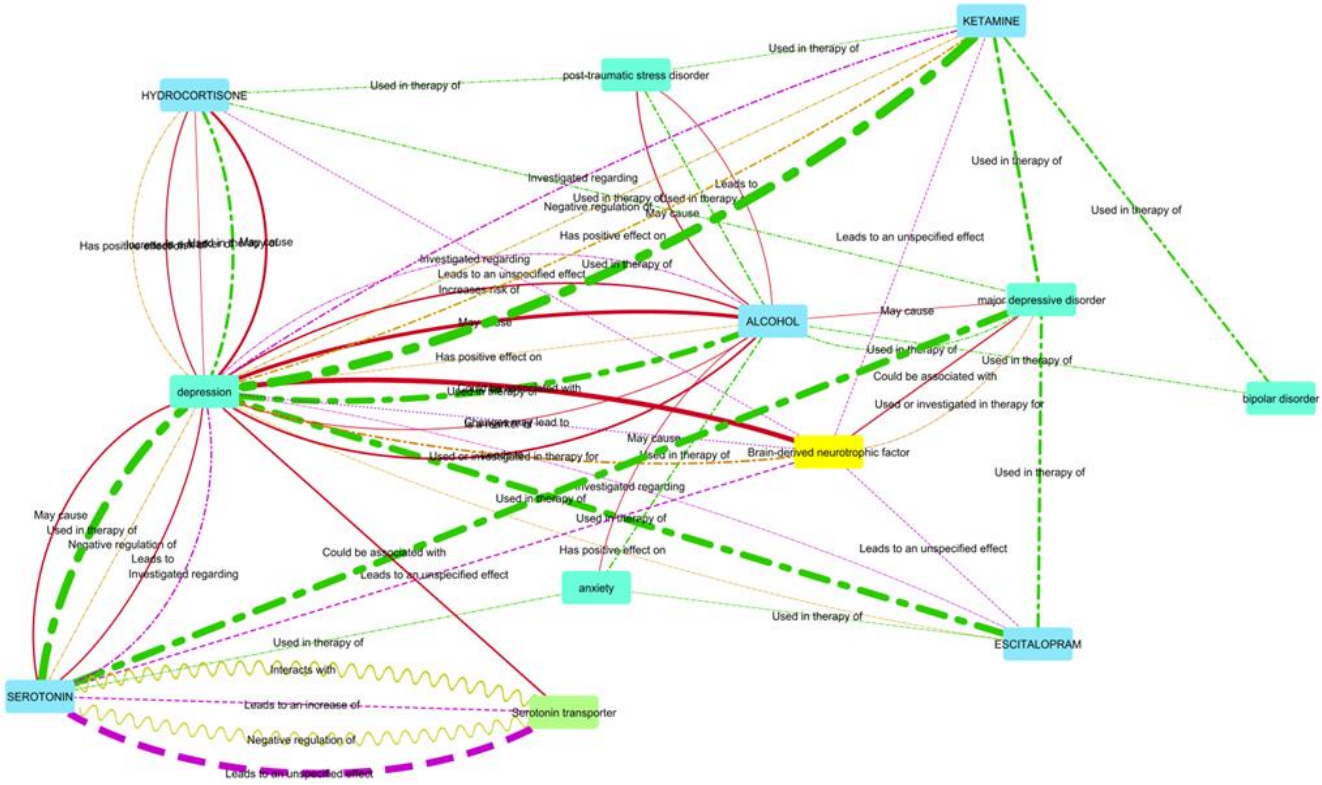

Рисунок 13 – Фрагмент представления результатов извлечения ассоциаций в формате графа для ассоциаций BDNF с механизмами большого депрессивного расстройства.

Помимо этого, в ряде публикаций упоминался куркумин как соединение с потенциальным антидепрессивным действием. Анализ ассоциаций показал его связь с регуляцией провоспалительных цитокинов и экспрессией BDNF [183], что подчёркивает универсальность разработанного нами интегрального подхода: он позволяет одновременно систематизировать молекулярные механизмы и выявлять перспективные терапевтические мишени.

При анализе механизмов большого депрессивного расстройства было важно учесть не только молекулярные механизмы патогенеза, но и предрасполагающие

внешние факторы. Мы разработали отдельный словарь таких внешних факторов, в который вошли термины вроде «childhood», «trauma» и другие. С применением разработанного алгоритма интеллектуального анализа текстов нам удалось выявить ассоциации не только между химическими соединениями, генами и заболеваниями, но и между психическими расстройствами и такими факторами, как вирусные инфекции, стрессовые события или неблагоприятный детский опыт [184]. Этот результат является демонстрацией того, что разработанный нами интегральный подход может быть масштабируем для решения различных биологических задач, и не требует значительных изменений в самом алгоритме.

3.3. База данных о наименованиях низкомолекулярных органических соединений и ассоциированных с ними видах биологической активности

3.3.1. Логическая структура базы данных

В логическую структуру базы данных были включены все четыре компонента, описанные более подробно в разделе материалы и методы: тексты, сущности, ассоциации и сведения об объектах из фактографических баз данных. Логическая структура разработанной нами базы данных представлена на рисунке 14.

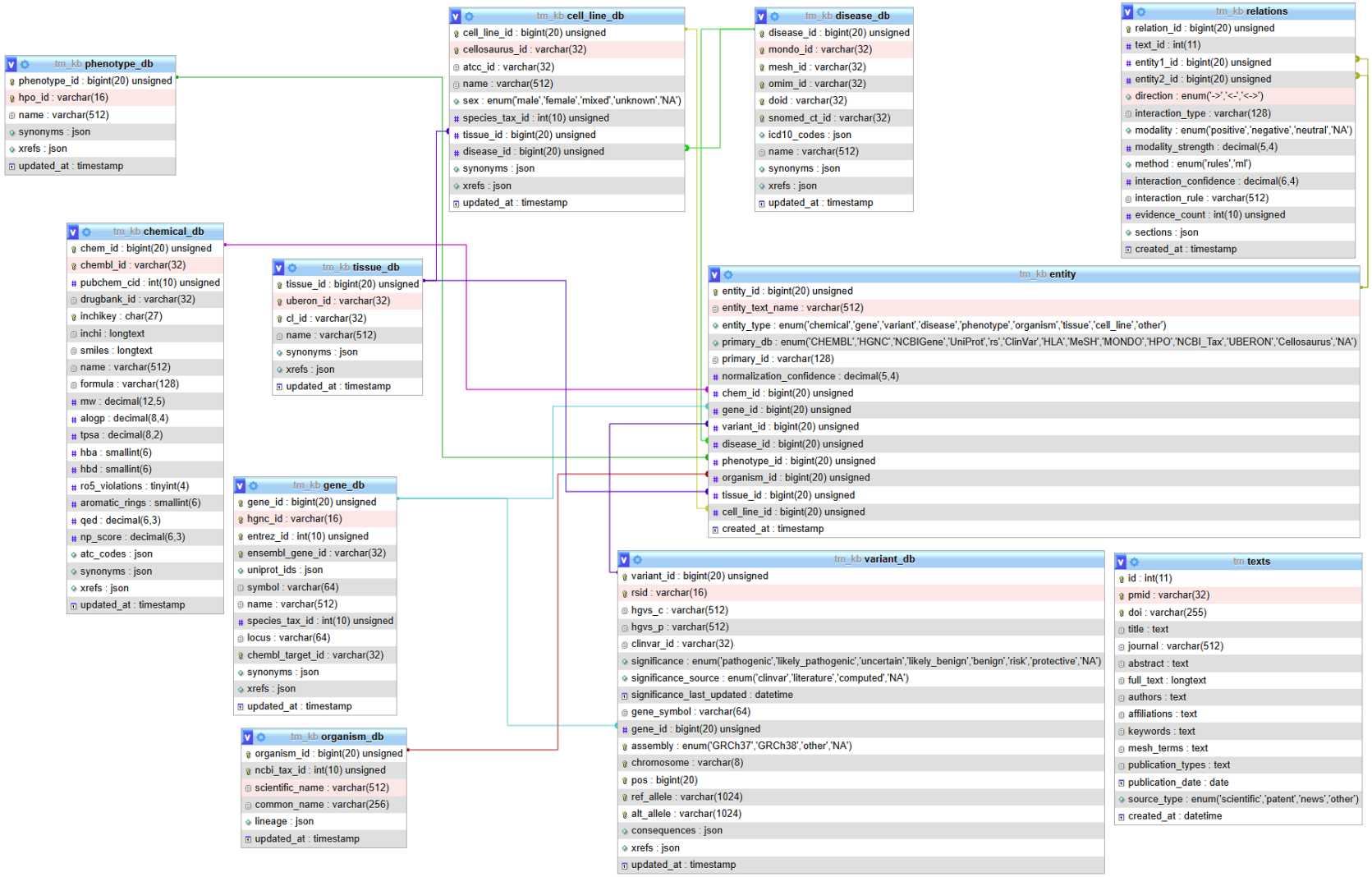

Рисунок 14 - Логическая структура базы данных о биологической активности низкомолекулярных органических соединений.

Таблица texts включает в себя следующие поля:

- Идентификатор: внутренний уникальный идентификатор текста. Позволяет хранить даже те тексты, которые не имеют PMID/DOI.
- PMID и DOI: внешние уникальные идентификаторы текстов. Эти поля могут оставаться пустыми.
- Заголовок, абстракт: текст заголовка публикации и его аннотации. Поле abstract может оставаться пустым.
- Метаданные: журнал (полное название), авторы (список), аффилиации (список), ключевые слова, MeSH-термины, типы публикации по PubMed, год публикации.
- Тип источника: научный или другой. Статус «научный» присваивается текстам научных публикаций с известными внешними идентификаторами, статус «другой» присваивается текстам без известных внешних идентификаторов, добавленных вручную. Наличие данного поля позволит расширить базу данных при дальнейшей работе за счет включения иных источников текстов (патенты, клинические записи и другие).
- Дата загрузки текста.

По внутреннему идентификатору текста связана таблица извлеченных ассоциаций (relations). В нее включены следующие поля:

- Уникальный внутренний идентификатор ассоциации внутри текста.
- Уникальный идентификатор сущности 1 и 2 в словаре сущностей (отдельная таблица, будет описана далее).
- Направление ассоциации. Указано в формате стрелок (->, <-, <->). Считается, что сущность 1 всегда слева, а сущность 2 – всегда справа. В случае, если направление взаимодействия определить не удалось, используется двунаправленная стрелка.
- Тип взаимодействия. Здесь указана группа, в которое включено сработавшее при извлечении ассоциации правило (например, «Side effect»).

- Сработавшее правило. Сюда включаются фразы-шаблоны, которые сработали непосредственно при извлечении ассоциации.
- Число уникальных ассоциаций данного типа в конкретном тексте.

К уникальным идентификаторам сущности 1 и 2 привязана таблица сущностей (entities). В нее включены следующие поля:

- Уникальный идентификатор сущностей. Является ключом для таблицы ассоциаций.
- Строка с наименованием в том виде, в котором оно упоминается в тексте.
- Тип сущности: химическое соединение, белок/ген, заболевание, биологический вид, миРНК, генотип HLA, вариант (однонуклеотидные полиморфизмы/аминокислотные замены)
- Первичная база данных: название фактографической базы данных, на которую происходит нормализация наименования.
- Уникальный идентификатор сущности во внешней базе данных. На каждый тип сущности – отдельное поле. Это позволяет избежать внедрения дополнительных таблиц и запросов для выбора таблицы объекта (описано далее) со сведениями из фактографически баз данных.

В случае, если нормализация наименования прошла успешно, то хотя бы одно поле с уникальным идентификатором сущности во внешней базе данных будет не пустым. Всего таких таблиц семь: по одной на каждый тип сущности. Каждая из них включает в себя основные сведения об объекте из фактографических баз данных. Например, для химических соединений – это InChiKey, SMILES, физико-химические признаки, формула, и другие. Для белков и генов – последовательность, принадлежность к организму, аннотация по Gene Ontology.

3.3.2. Доступ к базе данных через веб-интерфейс

В рамках работы была сформирована специализированная база данных, содержащая наименования низкомолекулярных органических соединений (НМОС)

и связанные с ними типы биологической активности, извлечённые из текстов научных публикаций. База размещена в открытом доступе [184-186] и позволяет исследователям использовать полученные сведения для анализа и поиска новых закономерностей.

При наполнении базы данных были применены процедуры постобработки: исключение дубликатов ассоциаций, нормализация наименований объектов с использованием фактографических ресурсов (ChEMBL, NCBI Taxonomy, Human Disease Ontology и др.), а также группировка ассоциаций по типам биологической активности. Для каждого соединения фиксировались его уникальный идентификатор, варианты наименований (систематические, тривиальные, сокращённые), список типов активности, а также PubMed ID публикаций, в которых упоминалась данная ассоциация. Дополнительно сохранялось количество предложений, из которых ассоциация была извлечена, что позволяет оценивать степень изученности и потенциальную новизну.

Особенностью базы является наличие отдельного поля «Status», в котором фиксируется факт ручной проверки достоверности извлечённой ассоциации. Таким образом, пользователь может отличить ассоциации, прошедшие ручную верификацию, от тех, которые ещё не проверялись.

Веб-интерфейс базы данных предусматривает возможность просмотра документов, из которых была извлечена ассоциация, с дальнейшей подсветкой наименований, которые стали основанием для формирования ассоциаций (рисунок 15).

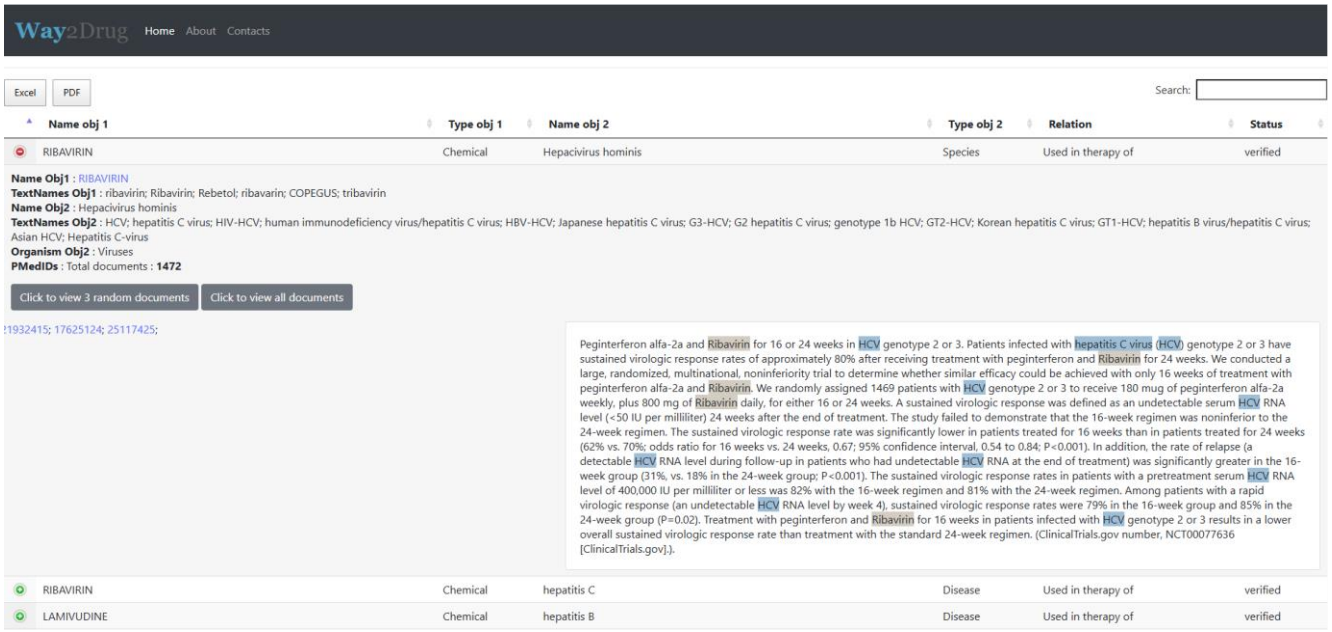

Рисунок 16 – Фрагмент базы данных, содержащей информацию о биологической активности низкомолекулярных химических соединений, извлеченных из текстов

В базу данных вошли более трех миллионов уникальных ассоциаций, которые могут быть использованы для получения всесторонних сведений о биологической активности низкомолекулярных органических соединений, описанной в текстах научных публикаций.

Важным преимуществом ресурса является наличие интерактивных ссылок на связанные объекты: пользователь может переходить от конкретного соединения к информации о соответствующих заболеваниях, белках, генах или других сущностях. Такой формат облегчает навигацию и существенно расширяет возможности анализа, позволяя рассматривать каждую ассоциацию как часть более широкой сети биомедицинских знаний.

Разработанная база данных обеспечивает систематизацию большого числа текстовых сведений, объединяет их с фактографическими ресурсами и предлагает удобные механизмы для анализа. Её использование может способствовать как подтверждению известных закономерностей, так и поиску новых знаний, например, если известна роль мишени в патогенезе заболевания и химическое

соединение, воздействующее на нее, но достоверно известна его эффективность в терапии только одного из них.

4. Проверка согласованности и полноты информации о биологической активности низкомолекулярных органических соединений.

Для оценки согласованности информации, извлекаемой методами интеллектуального анализа текстов, мы сопоставили её с данными, доступными в существующих базах данных. Основное внимание было уделено разделу Bioassays базы данных ChEMBL [187], так как именно здесь аккумулированы сведения о биологической активности соединений, полученные в доклинических исследованиях. Использование этого ресурса обосновано, во-первых, высокой степенью формализации представленной информации, а во-вторых, наличием регулярных обновлений представленных в ChEMBL сведений.

Так как собранный нами корпус текстов охватывал широкий спектр публикаций о противовирусной активности - как доклинических, так и клинических, - сравнение с ChEMBL позволяло оценить степень согласованности, в основном на доклиническом уровне.

На первом этапе мы сопоставили количество источников, использованных в обоих случаях. Как видно на рисунке 16, пересечение оказалось относительно небольшим: лишь 5 248 публикаций (1,86%) из нашей выборки одновременно присутствуют в ChEMBL. При этом 276 542 источника (98,14%) отсутствуют в ChEMBL. В обратном сопоставлении 3 835 публикаций (42,22%) из ChEMBL пересекались с нашей выборкой, тогда как 5 248 (57,78%) в неё не вошли (рисунок 16).

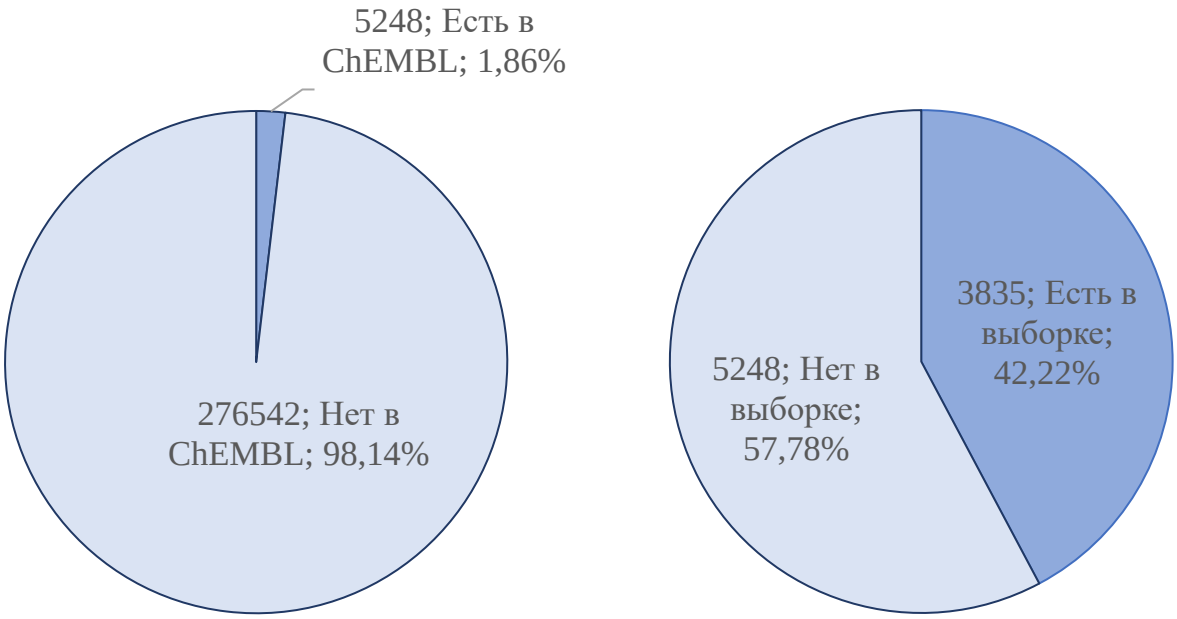

Рисунок 16 – Сопоставление публикаций в нашей выборке и в базе данных ChEMBL: А) доля публикаций из нашей выборки, представленных в ChEMBL, Б) доля публикаций из ChEMBL, вошедших в нашу выборку.

Причины таких расхождений можно разделить на несколько групп:

- в ChEMBL отсутствуют более свежие публикации, включённые в нашу выборку (самая «свежая» публикация в ChEMBL датируется июлем 2023 года, тогда как в нашей выборке – декабрём 2024 года);
- в ChEMBL представлены преимущественно доклинические исследования, тогда как наша выборка была составлена так, чтобы охватить максимально широкий спектр аспектов (репозиционирование, побочные эффекты и др.);
- некоторые публикации из ChEMBL не попадали в нашу выборку из-за отсутствия MeSH-терминов «Antiviral agents» или наименований противовирусных лекарственных препаратов, либо из-за того, что это были обзоры, исключённые на этапе запроса.

На втором этапе мы провели сопоставление ассоциаций по идентификаторам объектов (химическое соединение, белок/ген, вирус). Результаты представлены на рисунке 17. Полученные ассоциации были разделены на категории:

- Obj1Obj2 – полное совпадение пары «соединение – белок/ген/вирус» при одном и том же PMID;
- Obj1 – совпадение химического соединения и PMID при различии второго объекта;
- Obj2 – совпадение белка/гена/вируса и PMID при различии первого объекта;
- Nan – отсутствие в ChEMBL записи для данной ассоциации или идентификаторов объектов.

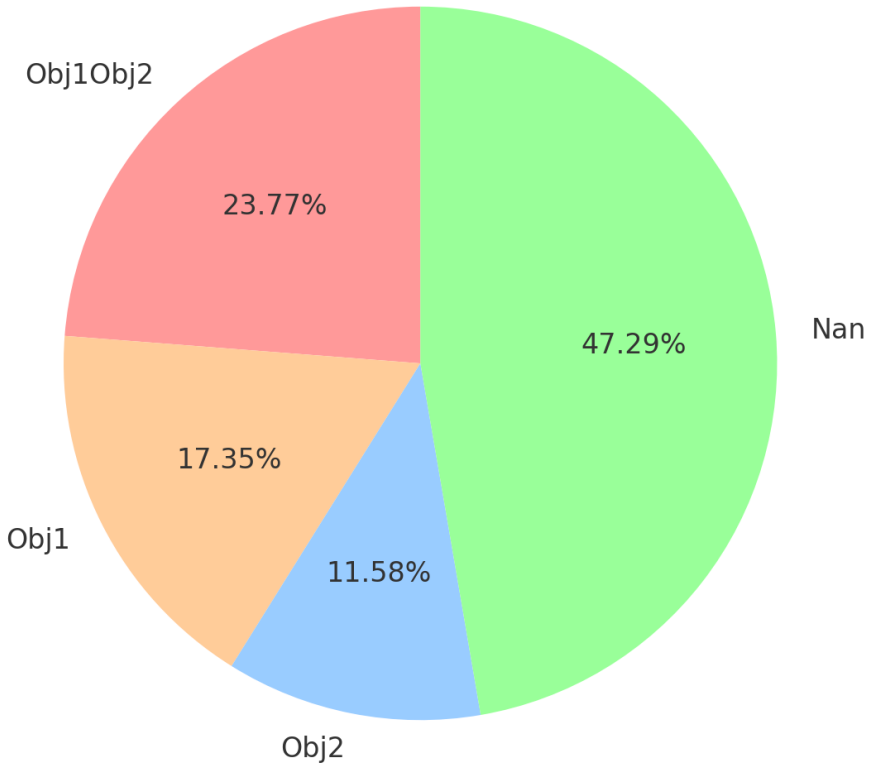

Рисунок 17 – Категории сопоставления ассоциаций, извлечённых из текстов, с базой данных ChEMBL. Obj1Obj2 - полное совпадение пары «химическое соединение – биологический объект»; Obj1 - совпадает химическое соединение,

различается биологический объект; Obj2 - совпадает биологический объект, различается химическое соединение; Nan – ассоциация отсутствует в ChEMBL.

В результате проведенного анализа было выявлено, что почти половина ассоциаций (47,29%) попала в категорию Nan, что указывает на существенно более широкий охват информации в нашей выборке. В то же время 23,77% ассоциаций полностью совпали с ChEMBL (Obj1Obj2), подтверждая корректность работы метода. Остальные случаи распределились между категориями Obj1 (17,35%) и Obj2 (11,58%), что демонстрирует различие в уровне детализации аннотаций.

Для иллюстрации результатов сопоставления были проанализированы отдельные публикации, вошедшие в нашу выборку. Например, при анализе текстов было выявлена ассоциация «Hydroxychloroquine – SARS-CoV-2» [188]. В ChEMBL же указано сразу несколько химических форм: «R-HCQ sulfate, Rac-HCQ sulfate, S-HCQ sulfate – SARS-CoV-2». Таким образом, различия связаны с нормализацией: текст фиксирует международное непатентованное наименование препарата, тогда как ChEMBL детализирует стереоизомеры и иные формы соединения.

Таким образом, проведённый анализ показал, что наш метод воспроизводит значительную часть информации, уже зафиксированной в ChEMBL, и при этом выявляет новые связи, отсутствующие в этой базе данных.

ЗАКЛЮЧЕНИЕ

Современные исследования в области биомедицины являются источником большого количества информации о методах, результатах, новых молекулах и их свойствах. Научные публикации являются основным первоисточником информации о различных характеристиках, в том числе, о биологической активности низкомолекулярных органических соединений. Автоматизация процессов агрегации и структурирования информации из текстов научных публикаций, содержащих описание различных видов биологической активности лекарственно-подобных соединений, позволяет быстро извлекать релевантные сведения о свойствах этих молекул, что необходимо при разработке компьютерных методов дизайна лекарственных препаратов.

В настоящей работе нами разработан и реализован автоматизированный метод извлечения ассоциаций между наименованиями низкомолекулярных органических соединений и видами их биологической активности. Была достигнута цель, поставленная в диссертационной работе, –создание интегрального подхода к автоматизированному извлечению знаний о биологической активности химических соединений.

Тестирование разработанного нами интегрального подхода было проведено на корпусах биомедицинских текстов для решения различных биомедицинских задач - от тщательно аннотированных коллекций (XenoMet) до сотен тысяч публикаций о противовирусной активности соединений и патогенезе заболеваний. Такой широкий охват позволил одновременно проверить метод на строго верифицированных задачах (например, извлечение знаний о биотрансформации) и на масштабных гетерогенных данных (анализ механизмов большого депрессивного расстройства), где оценивалась его устойчивость и воспроизводимость. Дополнительно была проведена валидация на примере поиска механизмов онкологических заболевания, ассоциированных с функционированием сигнального пути Hedgehog, которая подтвердила возможность масштабирования алгоритма для его применения в различных задачах биомедицины. Полученные результаты

показали, что метод сопоставим по точности с современными нейросетевыми архитектурами, при этом обладает рядом преимуществ: он более интерпретируем, проще в расширении за счёт добавления новых шаблонов, требует меньших затрат на аннотирование обучающих корпусов и сохраняет согласованность с существующими базами знаний. Это позволяет применять предложенный подход для систематизации информации о биологической активности низкомолекулярных органических соединений и связанных с ними механизмах патогенеза и терапии.

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 БД ChEMBL (v.34), а также выявлять новые взаимосвязи, отсутствующие в этой базе данных.

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа выполнена при поддержке Программы фундаментальных научных исследований в Российской Федерации на долгосрочный период (2021–2030 гг.) (№ 124050800018-9); Программы фундаментальных научных исследований в Российской Федерации на долгосрочный период (2021–2030 гг.) (№ 122030100170-5), в рамках проекта 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); гранта Российского научного фонда № 24-25-00453.

СПИСОК ЛИТЕРАТУРЫ

1. SCImago Journal & Country Rank [Электронный ресурс]. URL: https://www.scimagojr.com/ (дата обращения: 12.09.2025).
2. Peng Y., Yan S., Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets // Proceedings of the 18th BioNLP Workshop and Shared Task. Florence, Italy: Association for Computational Linguistics, 2019. С. 58–65.
3. Poroikov V., Druzhilovskiy D. Drug Repositioning: New Opportunities for Older Drugs // In Silico Drug Design. Elsevier, 2019. С. 3–17.
4. Аладышева Ж. И., Беляев В. В., Береговых В. В. и др. Промышленная фармация. Путь создания продукта : монография / под ред. А. Л. Хохлова, Н. В. Пятигорской ; Российская академия наук, Отделение медицинских наук. – Москва : Российская академия наук, 2019. – 536 с.
5. Ashburn T.T., Thor K.B. Drug repositioning: identifying and developing new uses for existing drugs // Nat Rev Drug Discov. 2004. Т. 3, № 8. С. 673–683.
6. Biziukova N., Tarasova O., Poroikov V. Revealing potential drug targets and their ligands through text-mining of massive literature data // ACS Fall 2023: National Meeting & Exposition. San Francisco, USA, Aug. 13–17, 2023.
7. Moreau E. et al. Mining impactful discoveries from the biomedical literature // BMC Bioinformatics. 2024. Т. 25, № 1. С. 303.
8. Cohen A.M. A survey of current work in biomedical text mining // Briefings in Bioinformatics. 2005. Т. 6, № 1. С. 57–71.
9. Krallinger M. et al. CHEMDNER: The drugs and chemical names extraction challenge // J Cheminform. 2015. Т. 7, № Suppl 1 Text mining for chemistry and the CHEMDNER track. С. S1.
10. Wei C.-H. et al. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task // Database (Oxford). 2016. Т. 2016. С. baw032.

11. Simmons M., Singhal A., Lu Z. Text Mining for Precision Medicine: Bringing Structure to EHRs and Biomedical Literature to Understand Genes and Health // Adv Exp Med Biol. 2016. Т. 939. С. 139–166.
12. Grissette H., Nfaoui E.H. Semisupervised neural biomedical sense disambiguation approach for aspect-based sentiment analysis on social networks // J Biomed Inform. 2022. Т. 135. С. 104229.
13. Correia R.B. et al. Mining Social Media Data for Biomedical Signals and Health-Related Behavior // Annu Rev Biomed Data Sci. 2020. Т. 3. С. 433–458.
14. Saffer J.D., Burnett V.L. Introduction to biomedical literature text mining: context and objectives // Methods Mol Biol. 2014. Т. 1159. С. 1–7.
15. Rodriguez-Esteban R., Bundschus M. Text mining patents for biomedical knowledge // Drug Discov Today. 2016. Т. 21, № 6. С. 997–1002
16. PubMed [Электронный ресурс]. URL: https://pubmed.ncbi.nlm.nih.gov/ (дата обращения: 12.09.2025).
17. PubMed Central (PMC) [Электронный ресурс]. URL: https://pmc.ncbi.nlm.nih.gov/ (дата обращения: 12.09.2025).
18. bioRxiv [Электронный ресурс]. URL: http://biorxiv.org/ (дата обращения: 12.09.2025).
19. medRxiv [Электронный ресурс]. URL: https://www.medrxiv.org/ (дата обращения: 12.09.2025).
20. World Intellectual Property Organization (WIPO) [Электронный ресурс]. URL: https://www.wipo.int/portal/en/index.html (дата обращения: 12.09.2025).
21. United States Patent and Trademark Office (USPTO) [Электронный ресурс]. URL: https://www.uspto.gov/ (дата обращения: 12.09.2025).
22. European Patent Office (EPO) [Электронный ресурс]. URL: https://www.epo.org/en (дата обращения: 12.09.2025).
23. China National Intellectual Property Administration (CNIPA) [Электронный ресурс]. URL: https://english.cnipa.gov.cn/ (дата обращения: 12.09.2025).

24. Роспатент. Патентный поиск [Электронный ресурс]. URL: https://searchplatform.rospatent.gov.ru/patents (дата обращения: 12.09.2025).
25. Google Patents [Электронный ресурс]. URL: https://patents.google.com/ (дата обращения: 12.09.2025).
26. Clarivate. Derwent Innovation [Электронный ресурс]. URL: https://clarivate.com/intellectual-property/patent-intelligence/derwent-innovation/ (дата обращения: 12.09.2025).
27. Sarker A., DeRoos A., Perrone J. Mining social media for prescription medication abuse monitoring: a review and proposal for a data-centric framework // J Am Med Inform Assoc. 2020. Т. 27, № 2. С. 315–329.
28. Singh T. et al. Social Media as a Research Tool (SMaaRT) for Risky Behavior Analytics: Methodological Review // JMIR Public Health Surveill. 2020. Т. 6, № 4. С. e21660.
29. You J., Shaik N., Chen H. Data Mining on COVID-19 Vaccines: Side Effects // Proc Assoc Inf Sci Technol. 2021. Т. 58, № 1. С. 869–871.
30. Lyu J.C., Han E.L., Luli G.K. COVID-19 Vaccine-Related Discussion on Twitter: Topic Modeling and Sentiment Analysis // J Med Internet Res. 2021. Т. 23, № 6. С. e24435.
31. Sarker A. et al. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource // J Am Med Inform Assoc. 2020. Т. 27, № 8. С. 1310–1315.
32. Lopez-Castroman J. et al. Mining social networks to improve suicide prevention: A scoping review // J of Neuroscience Research. 2020. Т. 98, № 4. С. 616–625.
33. Percha B. Modern Clinical Text Mining: A Guide and Review // Annu Rev Biomed Data Sci. 2021. Т. 4. С. 165–187.
34. Ahmad P.N., Shah A.M., Lee K. A Review on Electronic Health Record Text-Mining for Biomedical Name Entity Recognition in Healthcare Domain // Healthcare (Basel). 2023. Т. 11, № 9. С. 1268.
35. Ross M.K., Wei W., Ohno-Machado L. «Big data» and the electronic health record // Yearb Med Inform. 2014. Т. 9, № 1. С. 97–104.

36. Johnson A. et al. MIMIC-IV. PhysioNet.
37. eICU Collaborative Research Database [Электронный ресурс]. URL: https://eicu-crd.mit.edu/ (дата обращения: 12.09.2025).
38. Pollard T.J. et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research // Sci Data. 2018. Т. 5. С. 180178.
39. ClinicalTrials.gov [Электронный ресурс]. URL: https://clinicaltrials.gov/ (дата обращения: 12.09.2025).
40. EU Clinical Trials Register [Электронный ресурс]. URL: https://www.clinicaltrialsregister.eu/ (дата обращения: 12.09.2025).
41. WHO International Clinical Trials Registry Platform (ICTRP) [Электронный ресурс]. URL: https://www.who.int/clinical-trials-registry-platform (дата обращения: 12.09.2025).
42. DailyMed (FDA Drug Labels) [Электронный ресурс]. URL: https://dailymed.nlm.nih.gov/ (дата обращения: 12.09.2025).
43. European Medicines Agency (EMA) [Электронный ресурс]. URL: https://www.ema.europa.eu/ (дата обращения: 12.09.2025).
44. Biziukova N. et al. Automated Extraction of Information From Texts of Scientific Publications: Insights Into HIV Treatment Strategies // Front Genet. 2020. Т. 11. С. 618862.
45. Medical Subject Headings (MeSH) [Электронный ресурс]. URL: https://www.ncbi.nlm.nih.gov/mesh/ (дата обращения: 12.09.2025).
46. CORD-19 Research Dataset [Электронный ресурс]. URL: https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge (дата обращения: 12.09.2025).
47. Wang L.L. et al. CORD-19: The Covid-19 Open Research Dataset // ArXiv. 2020. С. arXiv:2004.10706v4.
48. OpenAlex [Электронный ресурс]. URL: https://openalex.org/ (дата обращения: 12.09.2025).

49. Priem J., Piwowar H., Orr R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts: arXiv:2205.01833. arXiv, 2022.
50. Renganathan V. Text Mining in Biomedical Domain with Emphasis on Document Clustering // Healthc Inform Res. 2017. Т. 23, № 3. С. 141–146.
51. Hofmann T. Unsupervised Learning by Probabilistic Latent Semantic Analysis // Machine Learning. 2001. Т. 42, № 1–2. С. 177–196.
52. Cheng X., Cao Q., Liao S.S. An overview of literature on COVID-19, MERS and SARS: Using text mining and latent Dirichlet allocation // J Inf Sci. 2022. Т. 48, № 3. С. 304–320.
53. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure: arXiv:2203.05794. arXiv, 2022.
54. Spärck Jones K. A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation. 2004. Т. 60, № 5. С. 493–502.
55. Joachims T. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization / T. Joachims. – In: Proceedings of the 14th International Conference on Machine Learning (ICML-97). – Nashville, US, 1997. – P. 143-151.
56. da Rocha M.A. et al. The Text Mining Technique Applied to the Analysis of Health Interventions to Combat Congenital Syphilis in Brazil: The Case of the «Syphilis No!» Project // Front Public Health. 2022. Т. 10. С. 855680.
57. Jurafsky D., Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition / Jurafsky D., Martin J. H. – 3rd ed. – Stanford, CA.; London: Prentice Hall, 2025.
58. Singhal A. Modern Information Retrieval: A Brief Overview. 2001 [Электронный ресурс]. URL: http://singhal.info/ieee2001.pdf (дата обращения: 12.09.2025).
59. Lee J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining // Bioinformatics / под ред. Wren J. 2020. Т. 36, № 4. С. 1234–1240.
60. Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding: arXiv:1810.04805. arXiv, 2019.

61. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain // Psychol Rev. 1958. Т. 65, № 6. С. 386–408.
62. Vaswani A. et al. Attention Is All You Need: arXiv:1706.03762. arXiv, 2023.
63. Tin Kam Ho. The random subspace method for constructing decision forests // IEEE Trans. Pattern Anal. Machine Intell. 1998. Т. 20, № 8. С. 832–844.
64. Vapnik V.N. The Support Vector method // Artificial Neural Networks – ICANN'97 / под ред. Gerstner W. et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997. Т. 1327. С. 261–271.
65. Cover T., Hart P. Nearest neighbor pattern classification // IEEE Trans. Inform. Theory. 1967. Т. 13, № 1. С. 21–27.
66. Rabby G., Berka P. Multi-class classification of COVID-19 documents using machine learning algorithms // J Intell Inf Syst. 2023. Т. 60, № 2. С. 571–591.
67. Névéol A. et al. Clinical Natural Language Processing in languages other than English: opportunities and challenges // J Biomed Semantics. 2018. Т. 9, № 1. С. 12.
68. Wang Y. et al. Clinical information extraction applications: A literature review // J Biomed Inform. 2018. Т. 77. С. 34–49.
69. Mogotsi I.C. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze: Introduction to information retrieval: Cambridge University Press, Cambridge, England, 2008, 482 pp, ISBN: 978-0-521-86571-5 // Inf Retrieval. 2010. Т. 13, № 2. С. 192–195.
70. Chai C.P. Comparison of text preprocessing methods // Nat. Lang. Eng. 2023. Т. 29, № 3. С. 509–553.
71. Kang N., van Mulligen E.M., Kors J.A. Comparing and combining chunkers of biomedical text // J Biomed Inform. 2011. Т. 44, № 2. С. 354–360.
72. Sennrich R., Haddow B., Birch A. Neural Machine Translation of Rare Words with Subword Units: arXiv:1508.07909. arXiv, 2016.

73. Kudo T., Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing: arXiv:1808.06226. arXiv, 2018.
74. Mikolov T. et al. Distributed Representations of Words and Phrases and their Compositionality: arXiv:1310.4546. arXiv, 2013.
75. Liu H. et al. BioLemmatizer: a lemmatization tool for morphological processing of biomedical text // J Biomed Semantics. 2012. Т. 3. С. 3.
76. Natural Language Toolkit (NLTK) [Электронный ресурс]. URL: https://www.nltk.org/ (дата обращения: 11.09.2025).
77. Wiesner M. DE-Lemma: A Maximum-Entropy Based Lemmatizer for German Medical Text // Stud Health Technol Inform. 2023. Т. 307. С. 189–195.
78. DeepPavlov [Электронный ресурс]. URL: https://deeppavlov.ai/ (дата обращения: 12.09.2025).
79. Natasha. Инструменты для обработки русского языка [Электронный ресурс]. URL: https://natasha.github.io/ (дата обращения: 12.09.2025).
80. Joachims T. A Probabilistic Analysis of the Rocchio Algorithm with TF-IDF for Text Categorization / T. Joachims. – In: Proceedings of the 14th International Conference on Machine Learning (ICML-97). – Nashville, US, 1997. – P. 143-151. – [Электронный ресурс]. URL: https://www.cs.cornell.edu/~tj/publications/joachims_97a.pdf (дата обращения: 12.09.2025).
81. Qorib M. et al. Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset // Expert Syst Appl. 2023. Т. 212. С. 118715.
82. Phang Y.C., Kassim A.M., Mangantig E. Concerns of Thalassemia Patients, Carriers, and their Caregivers in Malaysia: Text Mining Information Shared on Social Media // Healthc Inform Res. 2021. Т. 27, № 3. С. 200–213.
83. Qorib M. et al. COVID-19 Vaccine Hesitancy: A Global Public Health and Risk Modelling Framework Using an Environmental Deep Neural Network, Sentiment

Classification with Text Mining and Emotional Reactions from COVID-19 Vaccination Tweets // Int J Environ Res Public Health. 2023. Т. 20, № 10. С. 5803.

84. Bird S., Klein E., Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit / S. Bird, E. Klein, E. Loper. – Sebastopol, CA: O'Reilly Media, Inc., 2009. – ISBN 978-0-596-51649-9
85. spaCy [Электронный ресурс]. URL: https://spacy.io/ (дата обращения: 12.09.2025).
86. Burtsev M. et al. DeepPavlov: Open-Source Library for Dialogue Systems // Proceedings of ACL 2018, System Demonstrations. Melbourne, Australia: Association for Computatioal Linguistics, 2018. С. 122–127.
87. Natural Language Toolkit (NLTK) Documentation [Электронный ресурс]. URL: https://www.nltk.org/howto.html (дата обращения: 12.09.2025).
88. Соколовский Д.Е. и др. Оценка использования инструментов библиотеки spaCy и DeepPavlov для задачи извлечения именованных сущностей из описаний результатов осмотров пациентов с COVID-19 // ПромКибернетика. 2023. Т. 1, № 2. С. 46–53.
89. Honnibal M., Johnson M. An Improved Non-monotonic Transition System for Dependency Parsing // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015. С. 1373–1378.
90. Neumann M. et al. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing // Proceedings of the 18th BioNLP Workshop and Shared Task. Florence, Italy: Association for Computational Linguistics, 2019. С. 319–327.
91. Natasha. Corus Dataset [Электронный ресурс]. URL: https://github.com/natasha/corus (дата обращения: 12.09.2025).
92. Tutubalina E. et al. The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews // Bioinformatics / под ред. Wren J. 2021. Т. 37, № 2. С. 243–249.

93. Liu Y. et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach: arXiv:1907.11692. arXiv, 2019.
94. OpenAI. GPT Research [Электронный ресурс]. URL: https://openai.com/research/gpt (дата обращения: 12.09.2025).
95. Biziukova N.Yu. et al. Automatic Recognition of Chemical Entity Mentions in Texts of Scientific Publications // Autom. Doc. Math. Linguist. 2020. Т. 54, № 6. С. 306–315.
96. Krallinger M. et al. Information Retrieval and Text Mining Technologies for Chemistry // Chem Rev. 2017. Т. 117, № 12. С. 7673–7761.
97. UniProt [Электронный ресурс]. URL: https://www.uniprot.org/ (дата обращения: 12.09.2025).
98. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025 // Nucleic Acids Res. 2025. Т. 53, № D1. С. D609–D617.
99. Chemical Entities of Biological Interest (ChEBI) [Электронный ресурс]. URL: https://www.ebi.ac.uk/chebi/ (дата обращения: 12.09.2025).
100. Degtyarenko K. et al. ChEBI: a database and ontology for chemical entities of biological interest // Nucleic Acids Research. 2007. Т. 36, № Database. С. D344–D350.
101. HUGO Gene Nomenclature Committee (HGNC) [Электронный ресурс]. URL: https://www.genenames.org/ (дата обращения: 12.09.2025).
102. Seal R.L. et al. Genenames.org: the HGNC resources in 2023 // Nucleic Acids Res. 2023. Т. 51, № D1. С. D1003–D1009.
103. Manning C. D., Raghavan P., Schütze H. Evaluating search engines // Introduction to Information Retrieval. – Cambridge: Cambridge University Press, 2008. – P. 151–174.
104. Unified Medical Language System (UMLS) [Электронный ресурс]. URL: https://www.nlm.nih.gov/research/umls/index.html (дата обращения: 12.09.2025).

105. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology // Nucleic Acids Research. 2004. Т. 32, № 90001. С. 267D – 270.
106. PubChem [Электронный ресурс]. URL: https://pubchem.ncbi.nlm.nih.gov/ (дата обращения: 12.09.2025).
107. Kim S. et al. PubChem Substance and Compound databases // Nucleic Acids Res. 2016. Т. 44, № D1. С. D1202-1213.
108. Disease Ontology [Электронный ресурс]. URL: https://disease-ontology.org/ (дата обращения: 11.09.2025).
109. Schriml L.M. et al. Human Disease Ontology 2018 update: classification, content and workflow expansion // Nucleic Acids Res. 2019. Т. 47, № D1. С. D955–D962.
110. Gene Ontology [Электронный ресурс]. URL: https://geneontology.org/ (дата обращения: 12.09.2025).
111. Ashburner M. et al. Gene Ontology: tool for the unification of biology // Nat Genet. 2000. Т. 25, № 1. С. 25–29.
112. SNOMED International [Электронный ресурс]. URL: https://www.snomed.org/ (дата обращения: 12.09.2025).
113. El-Sappagh S. et al. SNOMED CT standard ontology based on the ontology for general medical science // BMC Med Inform Decis Mak. 2018. Т. 18, № 1. С. 76.
114. Song M., Yu H., Han W.-S. Developing a hybrid dictionary-based bio-entity recognition technique // BMC Med Inform Decis Mak. 2015. Т. 15 Suppl 1, № Suppl 1. С. S9.
115. Левенштейн В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии наук СССР. – 1965. – Т. 163, № 4. – С. 845–848.
116. Nastou K. et al. Improving dictionary-based named entity recognition with deep learning // Bioinformatics. 2024. Т. 40, № Suppl 2. С. ii45–ii52.
117. Aronson A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program // Proc AMIA Symp. 2001. С. 17–21.

118. Leaman R., Wei C.-H., Lu Z. tmChem: a high performance approach for chemical named entity recognition and normalization // J Cheminform. 2015. T. 7, № Suppl 1 Text mining for chemistry and the CHEMDNER track. C. S3.
119. Leaman R., Islamaj Dogan R., Lu Z. DNorm: disease name normalization with pairwise learning to rank // Bioinformatics. 2013. T. 29, № 22. C. 2909–2917.
120. Eftimov T., Koroušić Seljak B., Korošec P. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations // PLoS One. 2017. T. 12. № 6. C. e0179488.
121. Peng Y., Wei C.-H., Lu Z. Improving chemical disease relation extraction with rich features and weakly labeled data // J Cheminform. 2016. T. 8. C. 53.
122. Shardlow M. et al. A Text Mining Pipeline Using Active and Deep Learning Aimed at Curating Information in Computational Neuroscience // Neuroinformatics. 2019. T. 17. № 3. C. 391–406.
123. Tarasova O. A. et al. Extraction of Data on Parent Compounds and Their Metabolites from Texts of Scientific Abstracts // J Chem Inf Model. 2021. T. 61. № 4. C. 1683–1690.
124. Biziukova N.Yu. et al. XenoMet: A Corpus of Texts to Extract Data on Metabolites of Xenobiotics // ACS Omega. 2025. T. 10, № 3. C. 2459–2471.
125. Biziukova N.Yu. et al. Identification of Proteins and Genes Associated with Hedgehog Signaling Pathway Involved in Neoplasm Formation Using Text-Mining Approach // Big Data Min. Anal. 2024. T. 7, № 1. C. 107–130.
126. Mallick R. et al. Accelerated variant curation from scientific literature using biomedical text mining // MicroPubl Biol. 2022. T. 2022.
127. Leaman R., Khare R., Lu Z. Challenges in clinical natural language processing for automated disorder normalization // J Biomed Inform. 2015. T. 57. C. 28–37.
128. Wu H. и др. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022 // npj Digit. Med. 2022. T. 5, № 1. C. 186.

129. Krallinger M. et al. The CHEMDNER corpus of chemicals and drugs and its annotation principles // J Cheminform. 2015. T. 7, № Suppl 1 Text mining for chemistry and the CHEMDNER track. C. S2.
130. Doğan R.I., Leaman R., Lu Z. NCBI disease corpus: A resource for disease name recognition and concept normalization // Journal of Biomedical Informatics. 2014. T. 47. C. 1–10.
131. Kim J.-D. et al. Introduction to the bio-entity recognition task at JNLPBA // Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications - JNLPBA '04. Geneva, Switzerland: Association for Computational Linguistics, 2004. C. 70.
132. Wei C.-H. et al. tmVar: a text mining approach for extracting sequence variants in biomedical literature // Bioinformatics. 2013. T. 29, № 11. C. 1433–1439.
133. Тутубалина Е. В. Модели и методы автоматической обработки неструктурированных данных в биомедицинской области : дис. … канд. техн. наук. — Казань : Казанский (Приволжский) федеральный университет, 2023. — 167 с.
134. Tutubalina E., Miftahutdinov Z., Nokhiz P., Nedumov Y., Miftahutdinova L., Zolotarev A., Panchenko A., Gorbachev A., Nikolenko S. The Russian Drug Reaction Corpus: Modeling and Evaluation. // Bioinformatics. 2020. Vol. 36, No. 21. P. 5445–5453.
135. Tutubalina E., Miftahutdinov Z., Gareev B., Isachenko V., Kornilov A., Nedumov Y., Nikolenko S. NEREL-BIO: A Dataset for Nested Named Entity Recognition in the Biomedical Domain in Russian and English. // Findings of the Association for Computational Linguistics: ACL 2022. Dublin : Association for Computational Linguistics, 2022. P. 254–266
136. Kazama J. et al. Tuning support vector machines for biomedical named entity recognition // Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain -. Phildadelphia, Pennsylvania: Association for Computational Linguistics, 2002. T. 3. C. 1–8.

137. Lafferty J., McCallum A., Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 2001. P. 282–289.
138. Cohen W., Sarawagi S. Exploiting Dictionaries in Named Entity Extraction. ACL Anthology (2004) [Электронный ресурс]. URL: https://aclanthology.org/W04-1221/ (дата обращения: 12.09.2025).
139. Leaman R., Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition // Pac Symp Biocomput. 2008. С. 652–663.
140. Habibi M. et al. Deep learning with word embeddings improves biomedical named entity recognition // Bioinformatics. 2017. Т. 33, № 14. С. i37–i48.
141. Beltagy I., Lo K., Cohan A. SciBERT: A Pretrained Language Model for Scientific Text: arXiv:1903.10676. arXiv, 2019.
142. Ben Abacha A., Zweigenbaum P. Automatic extraction of semantic relations between medical entities: a rule based approach // J Biomed Semantics. 2011. Т. 2 Suppl 5, № Suppl 5. С. S4.
143. Ravikumar K.E., Rastegar-Mojarad M., Liu H. BELMiner: adapting a rule-based relation extraction system to extract biological expression language statements from bio-medical literature evidence sentences // Database. 2017. Т. 2017.
144. Kim P.-J., Price N.D. Genetic co-occurrence network across sequenced microbes // PLoS Comput Biol. 2011. Т. 7, № 12. С. e1002340.
145. Al-Aamri A. et al. Analyzing a co-occurrence gene-interaction network to identify disease-gene association // BMC Bioinformatics. 2019. Т. 20, № 1. С. 70.
146. Пономаренко Е. А. Автоматический анализ научных текстов для создания семантических сетей белков / Е. А. Пономаренко. – Москва : Институт биомедицинской химии РАН, 2009. – 150 с. – [Электронный ресурс]. URL: https://www.ibmc.msk.ru/content/dissertations/PonomarenkoEA.pdf (дата обращения: 12.09.2025).

147. Li J. et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction // Database (Oxford). 2016. Т. 2016. С. baw068.
148. NCBI BioCreative VI Challenge Dataset (BC6PPIm) [Электронный ресурс]. URL: https://ftp.ncbi.nlm.nih.gov/pub/lu/BC6PPIm/ (дата обращения: 12.09.2025).
149. Miranda-Escalada A. et al. Overview of DrugProt task at BioCreative VII: data and methods for large-scale text mining and knowledge graph generation of heterogenous chemical-protein relations // Database (Oxford). 2023. Т. 2023. С. baad080.
150. Herrero-Zazo M. et al. The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions // J Biomed Inform. 2013. Т. 46, № 5. С. 914–920.
151. MEDLINE – About MEDLINE [Электронный ресурс] / National Library of Medicine. URL: https://www.nlm.nih.gov/medline/medline_overview.html. – (дата обращения: 11.09.2025).
152. DrugBank [Электронный ресурс] / OMx Personal Health Analytics Inc.; University of Alberta et al. URL: https://go.drugbank.com/. – (дата обращения: 11.09.2025).
153. van Mulligen E.M. et al. The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships // J Biomed Inform. 2012. Т. 45, № 5. С. 879–884.
154. Luo L. et al. BioRED: a rich biomedical relation extraction dataset // Briefings in Bioinformatics. 2022. Т. 23, № 5. С. bbac282.
155. Segura-Bedmar I., Martínez P., Herrero-Zazo M. SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013) // Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013). – Atlanta : Association for Computational Linguistics, 2013. – P. 341–350. – URL: https://aclanthology.org/S13-2056.pdf (дата обращения: 11.09.2025).

156. Hata K. et al. Limited inhibitory effects of oseltamivir and zanamivir on human sialidases // Antimicrob Agents Chemother. 2008. Т. 52, № 10. С. 3484–3491.
157. Nguyen D.Q., Verspoor K. Convolutional neural networks for chemical-disease relation extraction are improved with character-based word embeddings: arXiv:1805.10586. arXiv, 2018.
158. Wang W. et al. Dependency-based long short term memory network for drug-drug interaction extraction // BMC Bioinformatics. 2017. Т. 18, № Suppl 16. С. 578.
159. Lai P.-T., Lu Z. BERT-GT: cross-sentence n-ary relation extraction with BERT and Graph Transformer // Bioinformatics. 2021. Т. 36, № 24. С. 5678–5685.
160. scikit-learn: Machine Learning in Python [Электронный ресурс]. URL: https://scikit-learn.org/stable/index.html (дата обращения: 11.09.2025).
161. Keras: Deep Learning API [Электронный ресурс]. URL: https://keras.io/ (дата обращения: 11.09.2025).
162. Biopython: Bio.Entrez module [Электронный ресурс]. URL: https://biopython.org/docs/latest/api/Bio.Entrez.html (дата обращения: 11.09.2025).
163. Sänger M. et al. HunFlair2 in a cross-corpus evaluation of biomedical named entity recognition and normalization tools // Bioinformatics. 2024. Т. 40, № 10. С. btae564.
164. NCBI Taxonomy [Электронный ресурс]. URL: https://www.ncbi.nlm.nih.gov/taxonomy (дата обращения: 11.09.2025).
165. Bui Q.-C. et al. Extracting causal relations on HIV drug resistance from literature // BMC Bioinformatics. 2010. Т. 11, № 1. С. 101.
166. PubMed Help: Publication Types [Электронный ресурс]. URL: https://pubmed.ncbi.nlm.nih.gov/help/#publication-types (дата обращения: 11.09.2025).
167. BioAssay Ontology [Электронный ресурс]. URL: http://bioassayontology.org/ (дата обращения: 11.09.2025).

168. Visser U. и др. BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results // BMC Bioinformatics. 2011. Т. 12, № 1. С. 257.
169. scispaCy: Biomedical Natural Language Processing [Электронный ресурс]. URL: https://allenai.github.io/scispacy/ (дата обращения: 11.09.2025).
170. Weber L. и др. HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition // Bioinformatics / под ред. Wren J. 2021. Т. 37, № 17. С. 2792–2794.
171. sigNER2: Tool for Biomedical Named Entity Recognition [Электронный ресурс]. URL: https://www.way2drug.com/hiv-host/sigNER2/ (дата обращения: 11.09.2025).
172. Matevosyan M. et al. Design of new chemical entities targeting both native and H275Y mutant influenza a virus by deep reinforcement learning // J Biomol Struct Dyn. 2023. Т. 41, № 20. С. 10798–10812.
173. Zhang Y. et al. A hybrid model based on neural networks for biomedical relation extraction // J Biomed Inform. 2018. Т. 81. С. 83–92.
174. Fei H. et al. A span-graph neural model for overlapping entity relation extraction in biomedical texts // Bioinformatics. 2021. Т. 37, № 11. С. 1581–1589.
175. Shannon P. et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks // Genome Res. 2003. Т. 13, № 11. С. 2498–2504.
176. Lai J.Y. et al. Interferon therapy and its association with depressive disorders – A review // Front. Immunol. 2023. Т. 14. С. 1048592.
177. Su K.-P. et al. Interferon-alpha-induced depression: Comparisons between early- and late-onset subgroups and with patients with major depressive disorder // Brain, Behavior, and Immunity. 2019. Т. 80. С. 512–518.
178. Baranyi A. et al. Quinolinic Acid Responses during Interferon-α-Induced Depressive Symptomatology in Patients with Chronic Hepatitis C Infection - A

Novel Aspect for Depression and Inflammatory Hypothesis // PLoS ONE / под ред. Guillemin G.J. 2015. Т. 10, № 9. С. e0137022.

179. Moghaddan H.S., Akhondzadeh S. The Clinician Scientist Training Program in Iran: Catalyzing Clinical Science Advancements // AJMB. 2023.
180. Zosen D. et al. Antidepressants escitalopram and venlafaxine up-regulate BDNF promoter IV but down-regulate neurite outgrowth in differentiating SH-SY5Y neurons // Neurochemistry International. 2023. Т. 169. С. 105571.
181. Talaee N. et al. Comparing the effect of fluoxetine, escitalopram, and sertraline, on the level of BDNF and depression in preclinical and clinical studies: a systematic review // Eur J Clin Pharmacol. 2024. Т. 80, № 7. С. 983–1016.
182. Nikolac Perkovic M. et al. The association of brain-derived neurotrophic factor with the diagnosis and treatment response in depression // Expert Review of Molecular Diagnostics. 2023. Т. 23, № 4. С. 283–296.
183. Fan C. et al. Prophylactic treatment of curcumin in a rat model of depression by attenuating hippocampal synaptic loss // Food Funct. 2021. Т. 12, № 22. С. 11202–11213.
184. Tarasova O.A. и др. Extracting information on virus-human interactions and on antiviral compounds based on automated analysis of large text collections // BIOMED KHIM. 2024. Т. 70, № 6. С. 469–474.
185. Way2Drug. Viruses NLP [Электронный ресурс]. URL: https://way2drug.com/viruses/nlp/ (дата обращения: 12.09.2025).
186. Way2Drug. TextOmics [Электронный ресурс]. URL: https://way2drug.com/TextOmics/ (дата обращения 21.10.2025).
187. ChEMBL. Assays [Электронный ресурс]. URL: https://www.ebi.ac.uk/chembl/explore/assays/ (дата обращения: 12.09.2025).
188. Ni Y. et al. Synthesis and evaluation of enantiomers of hydroxychloroquine against SARS-CoV-2 in vitro // Bioorg Med Chem. 2022. Т. 53. С. 116523.

Приложение 1

Иллюстрация соответствия MeSH-терминов отдельным видам исследований

In vivo (Animals[MeSH] NOT Humans[MeSH] NOT Review[PublicationType])				**In vitro** ("Cells, Cultured"[MeSH] NOT Review[PublicationType])			
Ссылка	***MeSH-термин***	***Краткое содержание***	***Соответствует типу исследования?***	***Ссылка***	***MeSH-термин***	***Краткое содержание***	***Соответствует типу исследования?***
[1]	Animals; Mice	Эксперименты на мышах по тестированию анти-SARS-CoV-2 антител.	Да	[21]	Cells, Cultured	Исследование функции NET-структур ("нейтрофильные ловушки") в патогенезе рака легкого	Да
[2]	Animals; Mice	Исследование антиоксидантов на модели диабета у мышей.	Да	[8]	Cells, Cultured	Дифференцировка дофаминергических нейронов при болезни Паркинсона при стимулах	Да
[3]	Animals; Rats	Эксперимент на крысах: влияние соединений на сердечно-сосудистую систему.	Да	[10]	Cells, Cultured	Стимуляция диссоциированных нейронов крыс	Да
[4]	Animals; Dogs	Модель сердечной недостаточности на собаках.	Да	[22]	Cell Line (Mutant Retinal)	Работа с мутантными клеточными линиями сетчатки	Да
[5]	Animals; Mice	Модель воспаления кишечника на мышах.	Да	[23]	Cells, Cultured	Воздействие вещества на культивированные нейроны мышей	Да
[6]	Animals	Биобанк образцов, без экспериментов на животных.	Нет	[24]	Cells, Cultured	Тестирование ингибитора ALPK1 в клеточных культурах	Да
[7]	Animals	Анализ данных пациентов и клеточных линий, животные	Нет	[25]	Cell Line, Tumor	Активация cGAS с применением	Да

		фактически не использовались.				синтетических олигонуклеотидов	
[8]	Animals; Rats	Эксперименты на крысах с исследованием фармакокинетики.	Да	[26]	Cell Line, Tumor	Цитотоксичность ивермектина на клеточной линии SUP-B15	Да
[9]	Animals; Mice	Модель мышей с опухолями для тестирования препарата.	Да	[27]	Cells, Cultured	AgNP–ампициллин-конъюгаты: оценка цитотоксичности в Vero-клетках	Да
[10]	Animals; Rabbits	Эксперимент на кроликах: тестирование биоматериала.	Да	[28]	Cells, Cultured	Функция CYP4B1 в кардиомиоцитах и клетках A549/H1703; влияние Ang II и NNK	Да
[11]	Animals	Использование баз данных и компьютерного моделирования, животные не задействованы.	Нет	[29]	Cell Line	Оценка опсонизации и фагоцитоза RAW 264.7 макрофагами	Да
[12]	Animals; Mice	Эксперименты на мышах (иммунология, инфекционная модель).	Да	[30]	Cells, Cultured	Микрофлюидная со-культура разных типов клеток	Да
[13]	Animals; Rats	Эксперимент на крысах (фармакология, токсикология).	Да	[31]	Cell Line, Tumor	Ниосомы: цитотоксичность в MCF-7 и MDA-MB-231	Да
[14]	Animals; Mice	Исследование воздействия вещества на мозг у мышей.	Да	[32]	Cells, Cultured	Со-культура HUVEC и фибробластов с оценкой ангиогенеза	Да
[15]	Animals	Моделирование с данными биобанка,	Нет	[33]	Cells, Cultured	Влияние вакцины на культуру лимфоцитов	Да

		животные не использовались.					
[16]	Animals; Rats	Эксперимент на крысах (метаболизм соединения).	Да	[34]	Cell Line (PC3)	Оптимизация 5'UTR-библиотек и анализ экспрессии в клеточных культурах PC3	Да
[17]	Animals	Обзор биоинформатических данных, животные не использовались.	Нет	[35]	Cells, Cultured	Гипоксически-индуцированные экзосомы кардиомиоцитов и их влияние на макрофаги RAW 264.7	Да
[18]	Animals; Mice	Эксперименты на мышах (онкологическая модель).	Да	[36]	Cell Line / Cells, Cultured	Downregulation miR-605-5p в клеточных линиях рака молочной железы, влияние на пролиферацию и прогнозирование	Да
[19]	Animals; Mice	Исследование на мышах (иммунология, инфекционные болезни).	Да	[37]	Cell Line, Tumor	Сеть взаимодействий lncRNA-miRNA/RBP-mRNA при тройном негативном раке молочной железы	Да
[20]	Animals; Swine	Эксперименты на поросятах (трансплантация почки).	Да	[38]	Cell Line, Tumor	Метод минимально инвазивного введения опухолевых клеток в лёгкое мышей для создания ортотопической модели рака лёгкого и последующего тестирования терапии	Нет

Список литературы:

1. Schwartz D.W. et al. Suspected Bartholin gland cystic-like structure and associated Corynebacterium pseudotuberculosis in a 1-year-old Nigerian dwarf doe // Can Vet J. 2025. T. 66, № 9. C. 992–996.
2. Brookhart M. et al. Tongue lesions in feedlot cattle associated with ergot alkaloid consumption // Can Vet J. 2025. T. 66, № 9. C. 997–1003.
3. Nury C., Klostermann C.A., Nichols S. Omphalophlebectomy with partial hepatectomy in a Holstein heifer calf presented for concurrent septic arthritis of the shoulder // Can Vet J. 2025. T. 66, № 9. C. 979–985.
4. Graeber M.P., Weatherton L. Suspected relay pentobarbital intoxication of a dog after ingestion of contaminated tissue // Can Vet J. 2025. T. 66, № 9. C. 974–978.
5. Pollock C.M. et al. Leptospirosis seroprevalence in Canadian beef calves at or near fall weaning // Can Vet J. 2025. T. 66, № 9. C. 955–960.
6. Devine M.E. et al. Intussusception associated with congenital lymphangioma in a dog // Can Vet J. 2025. T. 66, № 9. C. 986–991.
7. Darby S., DeNotta S., Gomez D.E. Ivermectin toxicosis in a foal: Use of intravenous lipid emulsion therapy // Can Vet J. 2025. T. 66, № 9. C. 1004–1008.
8. Peng S. et al. Paeoniflorin Combined with Neural Stem Cell Transplantation for Parkinson’s Disease: Dual Mechanism of Cell Therapy and Inflammation Regulation // Drug Des Devel Ther. 2025. T. 19. C. 7745–7761.
9. Onda R. et al. Spike-timing-dependent plasticity offers delay-gated oscillatory potentiation for autaptic weights // Front Neural Circuits. 2025. T. 19. C. 1646317.

10. Zhang Z. et al. Deviance detection and regularity sensitivity in dissociated neuronal cultures // Front Neural Circuits. 2025. T. 19. C. 1584322.
11. Martin A.M., Cordero-De La Cruz D., Swierk L. High dissolved oxygen extends dive duration and suggests physical gill use in a vertebrate // J Exp Biol. 2025. T. 228, № 17. C. jeb250627.
12. Hou R. et al. A thermochromic tissue-mimicking phantom for catheter-based radiofrequency ablation of heterogenous lesions // Med Phys. 2025. T. 52, № 9. C. e18112.
13. Prabhu D. et al. Harnessing the potential of phytochemicals to design anti-filarial molecules targeting the MurE enzyme of Brugia malayi: a hierarchical virtual screening and molecular dynamics simulation study // SAR QSAR Environ Res. 2025. T. 36, № 8. C. 753–773.
14. Kelm N. et al. Enhanced ISGylation via USP18 Isopeptidase Inactivation Fails to Mitigate the Inflammatory or Functional Course of Coxsackievirus B3-Induced Myocarditis // Cell Physiol Biochem. 2025. T. 59, № S3. C. 1–21.
15. Xie J., Wang H., Du J. High-resolution dual-ended readout PET detectors based on 0.5 mm pitch LYSO arrays with various reflectors // Med Phys. 2025. T. 52, № 9. C. e18103.
16. Wang X., Scheven U.M., Liu Z. Dynamic Magnetic Resonance Imaging of Whole-Stomach Motility in Rats // NMR Biomed. 2025. T. 38, № 10. C. e70138.
17. Viel L. et al. Avian reovirus in Italy: three episodes of abnormal losses in offspring of vaccinated broiler breeders // Vet Ital. 2025. T. 61, № 3.
18. Zacarias A.C.C. et al. Seroprevalence of small ruminant lentivirus infections (SRLV) in family farming goats from Alagoas semiarid region, Brazil // Vet Ital. 2025. T. 61, № 4.

19. Singh M., Yadav J.P. Microbiology of Otitis externa in dogs reveals wide variation in Staphylococcus species // Vet Ital. 2025. T. 61, № 4.
20. Navratil P. et al. Protection of the Endothelium and Endothelial Glycocalyx by Albumin and Sulodexide in Porcine Model of Kidney Transplant // Exp Clin Transplant. 2025. T. 23, № 8. C. 509–516.
21. Alsharif N. et al. Human lung cancer neutrophils generate NETs with preserved anti-tumor cytotoxicity but impaired anti-migratory activity // Front Immunol. 2025. T. 16. C. 1643609.
22. Villa-Vasquez S.S. et al. Rescue of ciliogenesis and hyperglutamylation mutant phenotype in AGBL5-/- cell model of retinitis pigmentosa // BMC Mol Cell Biol. 2025. T. 26, № 1. C. 27.
23. Torres J. et al. Neddylation regulates the development and function of glutamatergic neurons // Commun Biol. 2025. T. 8, № 1. C. 1338.
24. Fan J. et al. Discovery of a selective alpha-kinase 1 inhibitor for the rare genetic disease ROSAH syndrome // Nat Commun. 2025. T. 16, № 1. C. 8251.
25. Sun Y. et al. CircPSD3 aggravates tumor progression… via regulating SUCLG2 in thyroid carcinoma // Cell Death Dis. 2025. T. 16, № 1. C. 590.
26. Siregar O.R. et al. Ivermecatin induces cytotoxic effects in SUP-B15 cell line // Turk J Pediatr. 2025. T. 67, № 4. C. 569–574.
27. Ayubee M.S. et al. Synergistic antibacterial action of AgNP-ampicillin conjugates: Evading β-lactamase degradation in ampicillin-resistant clinical isolates // PLoS One. 2025. T. 20, № 9. C. e0331669.

28. Dai L. et al. 1p-Enh-regulated CYP4B1 alleviates NNK-induced heart failure and lung cancer via the STAT3 pathway // PLoS One. 2025. T. 20, № 9. C. e0331471.
29. Slarve M. et al. Evaluating antibody mediated opsonophagocytosis of bacteria via lab protocol: RAW 264.7 cell phagocytosis assay // PLoS One. 2025. T. 20, № 9. C. e0331445.
30. Dong Z. et al. Hybridoma-inspired strategy crafts tailored multifunctional exosomes for precision therapy // Proc Natl Acad Sci U S A. 2025. T. 122, № 37. C. e2424547122.
31. Eftekhari Z., Chiani M., Kazemi-Lomedasht F. NPY-functionalized niosomes for targeted delivery of margatoxin in breast cancer therapy // Med Oncol. 2025. T. 42, № 10. C. 465.
32. Mansouri A. et al. Targeting the tumor microenvironment in colorectal cancer: the effect of Rapamycin on angiogenesis, apoptosis, and STAT5A/TORC1 signaling // Mol Biol Rep. 2025. T. 52, № 1. C. 875.
33. Jiang B. et al. Mitoribosome-Targeting Antibiotics Suppress Osteoclastogenesis and Periodontitis-Induced Bone Loss by Blocking Mitochondrial Protein Synthesis // FASEB J. 2025. T. 39, № 17. C. e71037.
34. Wang L. et al. Machine learning-based analysis of the impact of 5’ untranslated region on protein expression // Nucleic Acids Res. 2025. T. 53, № 17. C. gkaf861.
35. Wang M. et al. Hypoxia-conditioned cardiomyocyte-derived exosomes attenuate myocardial injury via ANP-mediated M2 macrophage polarization // Gen Physiol Biophys. 2025. T. 44, № 5. C. 377–389.
36. Zheng T. et al. Multivalent DNA Origami Enables Single-Molecule Dissection of Integrin αvβ6–RTK Crosstalk in Cancer Biology // ACS Nano. 2025. T. 19, № 35. C. 31467–31480.

37. Yun W. et al. lncRNA-miRNA/RBP-mRNA Network Involved in Human Triple-Negative Breast Cancer: An Integrated Approach // J Environ Pathol Toxicol Oncol. 2025. Т. 44, № 3. С. 75–87.
38. Foster D.J. et al. A Minimally Invasive Method for Generating a Syngeneic Orthotopic Mouse Model of Lung Cancer // J Vis Exp. 2025. № 222.

Приложение 2

Извлеченные из текста ассоциации, представленные в формате таблицы (фрагмент).

Сведения об Объекте 1					Сведения об Объекте 2					Информация об ассоциации и источниках		
Obj1_TextNames	**Obj1_Type**	**Obj1_DBid**	**Obj1_DBPrefName**	**Obj1_Organism**	**Obj2_TextNames**	**Obj2_Type**	**Obj2_DBid**	**Obj2_DBPrefName**	**Obj2_Organism**	**BiolActivity_Concept**	**Nsenteces**	**PMIDs**
ribavirin; Ribavirin; Rebetol; ribavarin; COPEGUS; tribavirin	Chemical	CHEMBL1643	RIBAVIRIN	-	HCV; hepatitis C virus; HIV-HCV; human immunodeficiency virus/hepatitis C virus; HBV-HCV; Japanese hepatitis C virus; G3-HCV; G2 hepatitis C virus; genotype 1b HCV; GT2-HCV; Korean hepatitis C virus; GT1-HCV; hepatitis B virus/hepatitis C virus; Asian HCV; Hepatitis C-virus	Species	3052230	Hepacivirus hominis	Viruses	Used in therapy of	5414	18237873; 18238889; 18240861; 18297995; 18299740; 18299740; …
ribavirin; Ribavirin; ribavarin; COPEGUS; Rebetol; Copegus; rebetol; tribavirin	Chemical	CHEMBL1643	RIBAVIRIN	-	chronic hepatitis C; hepatitis C; hepatitis C infection; viral hepatitis C; Hepatitis C infection; Chronic hepatitis C	Disease	DOID_1883	hepatitis C	-	Used in therapy of	2954	21221800; 21397729; 21480995; 21623851; …
lamivudine	Chemical	CHEMBL141	LAMIVUDINE	-	chronic hepatitis B; hepatitis B; hepatitis B infection	Disease	DOID_2043	hepatitis B	-	Used in therapy of	2105	15065004; 15074316; …
Lamivudine; lamivudine; Epivir; Zeffix	Chemical	CHEMBL141	LAMIVUDINE	-	HBV; hepatitis B virus; HIV-HBV; HIV-1-HBV; Hepatitis B e antigen-negative/hepatitis B virus; human HBV; G-HBV	Species	10407	Hepatitis B virus	Viruses	Used in therapy of	2041	15364971; 15504286; 15506667; 15564745; …
HCV; hepatitis C virus; HIV-HCV; Egyptian hepatitis C virus; GT2-HCV; G3-HCV; Hepatitis C-virus; human immunodeficiency virus/hepatitis C virus	Species	3052230	Hepacivirus hominis	Viruses	sofosbuvir; Sofosbuvir; PSI-7977; Sovaldi; GS-7977	Chemical	CHEMBL1259059	SOFOSBUVIR	-	Used in therapy of	1530	35110949; 35110949; 35581112; 35595682; 35595682; …
CMV; TI-CMV; ganR-CMV; non-CMV; human CMV	Species	10358	Cytomegalovirus	Viruses	ganciclovir; Gancyclovir; Cymevene; Vitrasert	Chemical	CHEMBL182	GANCICLOVIR	-	Used in therapy of	1293	2851272; 10071457; …
entecavir; Entecavir; entecavir hydrate; Entecavir Hydrate	Chemical	CHEMBL5314362	ENTECAVIR	-	chronic hepatitis B; Hepatitis B; hepatitis B infection; hepatitis B	Disease	DOID_2043	hepatitis B	-	Used in therapy of	1219	27687792; 30864558; 31166004; …