МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
НАУЧНОЕ УЧРЕЖДЕНИЕ
«ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР ИНСТИТУТ ЦИТОЛОГИИ и ГЕНЕТИКИ СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК»
(ИЦиГ СО РАН)

Пр-т. Академика Лаврентьева, д. 10, Новосибирск, 630090
Телефон: (383) 363-49-80
Факс (383) 333-12-78
E-mail: icg-adm@bionet.nsc.ru
https://www.icgbio.ru
ИНН 5408100138/КПП 540801001
ОКПО 03533895 ОГРН 1025403657410

от 30.03.2026 № 15345-29-38/463
на №_____ от _____

«УТВЕРЖДАЮ»
директор ИЦиГ СО РАН
академик РАН
_____Кочетов А.В.
«30» 03 2026 г

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

на диссертацию **БИЗЮКОВОЙ Надежды Юрьевны**
«Формирование знаний о биологической активности низкомолекулярных органических соединений на основе автоматизированного анализа текстов»,
представленную на соискание ученой степени кандидата биологических наук по специальности 1.5.8. Математическая биология, биоинформатика.

Исследование методов автоматизированного извлечения знаний из неструктурированных биомедицинских текстов относится к числу актуальных направлений современной математической биологии, биоинформатики, системной биологии и фармакологии. Постоянный рост объема научной литературы, индексируемой в международных библиографических базах данных, в том числе PubMed, а также в патентных источниках, существенно затрудняет использование традиционных подходов к анализу публикаций для поиска новых лекарственных мишеней, изучения свойств химических соединений и систематизации накопленных экспериментальных данных. В этой связи разработка специализированных биоинформатических методов, направленных на структурирование информации о взаимосвязях типа «химическое соединение — биологическая активность», является, безусловно, актуальной и имеет существенное значение для ускорения исследований в области поиска и разработки новых лекарственных средств.

Диссертационная работа Бизюковой Н.Ю. посвящена решению важной научной задачи, связанной с разработкой алгоритмических и программных средств для автоматизированного формирования знаний о свойствах химических веществ, прежде всего о биологической активности низкомолекулярных органических соединений, на основе гибридных методов обработки естественного языка. Целью диссертационной работы является разработка, реализация и тестирование метода извлечения взаимосвязей между наименованиями низкомолекулярных органических соединений и известными видами их биологической активности.

Актуальность темы исследования

В настоящее время одной из ключевых проблем биомедицины является фрагментарность знаний, распределенных по миллионам научных публикаций. Традиционные поисковые системы, как правило, не обеспечивают извлечения содержательных семантических связей между биомедицинскими сущностями, такими как химическое соединение, биологическая активность, фармакологический эффект и механизм действия. В связи с этим работа соискателя, направленная на адаптацию современных методов обработки естественного языка и разработку собственного интегрального метода извлечения ассоциаций между наименованиями низкомолекулярных органических соединений и видами их биологической активности, представляется актуальной и своевременной.

Следует также отметить, что важным результатом исследования является создание пополняемой базы данных, объединяющей сведения, извлекаемые из текстовых источников, с фактографической информацией. Это имеет существенное значение для решения задач поиска новых направлений применения известных соединений, репозиционирования лекарственных средств, а также систематизации накопленных данных о биологической активности низкомолекулярных органических соединений.

Дополнительную значимость работе придает возможность расширения разработанного подхода на различные типы специализированных текстовых источников, включая медицинские документы, научные публикации и патентные материалы в различных междисциплинарных областях.

Научная новизна и практическая ценность исследований

В диссертационной работе Бизюковой Н.Ю. получены новые научные результаты, имеющие существенное значение для развития методов математической биологии, биоинформатики и автоматизированного анализа биомедицинских текстов.

Прежде всего, следует отметить разработку оригинального интегрального метода автоматизированного извлечения ассоциаций между наименованиями низкомолекулярных органических соединений и видами их биологической активности. Научная новизна данного результата определяется тем, что предложенный подход объединяет методы на основе условных случайных полей (CRF), предобученную модель HunFlair (LSTM-CRF), словарные методы и систему более чем из тысячи фразовых шаблонов, сгруппированных в 44 категории типов биологической активности. Такое сочетание методов позволило обеспечить высокую точность распознавания именованных сущностей при сравнительно невысоких требованиях к вычислительным ресурсам и сохранить интерпретируемость получаемых результатов, что представляет несомненную научную и методическую ценность.

К числу новых результатов диссертации следует также отнести создание масштабируемой системы, адаптируемой к решению широкого круга биомедицинских задач. Существенно, что разработанный подход обеспечивает автоматическое пополнение базы данных и допускает перенос предложенных решений на смежные направления биомедицинских исследований, не ограничиваясь только анализом сведений о низкомолекулярных органических соединениях.

Несомненный научный интерес представляет разработанный автором метод автоматизированной категоризации научных публикаций по типам исследований — *in vitro*, *in vivo* и клинические испытания — на основе комбинации MeSH-терминов и фильтров

типов публикаций PubMed. Практически важным достоинством данного подхода является возможность формирования репрезентативных выборок текстов для последующего анализа без проведения трудоемкой ручной разметки больших текстовых корпусов.

Особого внимания заслуживает тот факт, что на основе разработанного и апробированного интегрального подхода автором извлечено более трех миллионов уникальных ассоциаций между наименованиями низкомолекулярных органических соединений, белками, генами и микроРНК. На этой основе создана база данных, включающая наименования соединений, их синонимы и ассоциированные с ними виды биологической активности. Указанный результат свидетельствует о высокой результативности предложенного подхода и открывает новые возможности для систематизации и последующего анализа биомедицинских знаний.

Важным элементом научной новизны диссертации является также проведенная автором проверка согласованности и полноты сведений о биологической активности низкомолекулярных соединений, извлекаемых из текстов научных публикаций. Показано, что разработанный метод позволяет извлекать значительную часть информации, содержащейся в базе данных ChEMBL, а также выявлять новые взаимосвязи, отсутствующие в данном ресурсе. Тем самым диссертационная работа вносит вклад не только в развитие методов интеллектуального анализа биомедицинских текстов, но и в совершенствование подходов к пополнению, верификации и актуализации специализированных фактографических баз данных и баз знаний.

Практическая ценность результатов

Практическая ценность диссертационной работы определяется тем, что разработанные автором методы интеллектуального анализа текстов позволяют существенно сократить временные и финансовые затраты на поиск, систематизацию и последующее исследование сведений о биологической активности соединений. Это имеет важное значение для фармакологии, молекулярной биологии, биоинформатики и других смежных областей, в которых полнота и оперативность учета опубликованных данных непосредственно влияют на результативность научных исследований.

Следует особо отметить прикладное значение полученных результатов для решения задач репозиционирования лекарственных средств. Показано, что предложенные алгоритмы анализа текстов позволяют выявлять ранее неизвестные виды биологической активности уже одобренных препаратов и обнаруживать новые показания к применению существующих молекул. Это создает предпосылки для сокращения сроков и снижения рисков, связанных с выводом лекарственных средств на рынок.

Практически значимым результатом является создание базы данных о низкомолекулярных органических соединениях, включающей их наименования, синонимы и ассоциированные виды биологической активности. Данный ресурс может найти применение в научных и прикладных исследованиях в области фармакологии, молекулярной биологии, биоинформатики, а также при разработке новых терапевтических стратегий и поиске перспективных молекулярных мишеней.

Разработанная система представляет интерес и как инструмент анализа больших массивов научной литературы. Она может быть использована в деятельности научных организаций, исследовательских центров и фармакологических компаний для мониторинга новых данных и автоматического пополнения баз знаний. Созданные в рамках работы аннотированные текстовые корпуса и словари также обладают самостоятельной

практической ценностью, поскольку могут служить основой для обучения и тестирования новых моделей машинного обучения в области биоинформатики.

Следует также подчеркнуть, что на основе разработанных алгоритмов создан свободно доступный веб-ресурс SigNER, интегрированный в портал Way2Drug и обеспечивающий пользователям возможность автоматического распознавания сущностей в биомедицинских текстах. Наличие такого ресурса подтверждает прикладной характер выполненного исследования и расширяет возможности практического использования полученных результатов.

К числу организаций, в деятельности которых могут быть применены результаты диссертационной работы, следует отнести Федеральный исследовательский центр «Институт цитологии и генетики Сибирского отделения Российской академии наук» (ИЦиГ СО РАН), Институт общей генетики им. Н.И. Вавилова Российской академии наук, ФГБУН «ФИЦ питания и биотехнологии», а также иные научно-исследовательские и научно-образовательные организации, имеющие подразделения, работающие в области биоинформатики, молекулярной биологии и фармакологии. Среди российских фармацевтических и биотехнологических компаний, в которых могут найти применение результаты работы, следует отметить BIOCAD, АО «ГЕНЕРИУМ», АО «ГЕРОФАРМ» и ГК «Фармасинтез».

В целом научная новизна и практическая ценность результатов диссертационной работы Бизюковой Н.Ю. не вызывают сомнений. Полученные автором результаты являются значимыми, обоснованными и перспективными как с точки зрения дальнейшего развития методов интеллектуального анализа биомедицинских текстов, так и с точки зрения их практического применения в задачах фармакологии, биоинформатики и поиска новых лекарственных средств.

Структура и содержание работы

Диссертация изложена на 151 странице, содержит 17 рисунков и 12 таблиц. Работа состоит из введения, трех глав (обзор литературы, материалы и методы, результаты и обсуждение), заключения, выводов и списка литературы, включающего 188 источников.

Во введении четко сформулированы цель и задачи, обоснована актуальность. Первая глава содержит обзор литературы по современным методам обработки естественного языка (NLP) и источникам биомедицинских данных, а также сравнительный анализ инструментов (NLTK, spaCy, Natasha, DeepPavlov) для анализа русскоязычных и англоязычных текстов. Вторая глава описывает методологию исследования: алгоритмы формирования обучающих выборки на основе MeSH-дескрипторов, методы предобработки данных и архитектуры нейронных сетей (LSTM, CRF) для распознавания именных сущностей и алгоритмов извлечения ассоциаций. Третья глава посвящена результатам работы: валидации разработанных алгоритмов, описанию созданной базы данных, а также практическому применению методов для анализа сигнальных путей (на примере Hedgehog) и классификации биологической активности. Заключение обобщает основные результаты работы и формулирует выводы. Приложения включают дополнительные таблицы и материалы. Структура работы логична и соответствует требованиям ВАК, материал изложен последовательно с четким разделением теоретической базы, методологии и экспериментальных результатов.

Обоснованность и достоверность научных положений

Достоверность результатов обеспечивается использованием репрезентативных объемов данных (десятки тысяч аннотаций), применением общепризнанных метрик оценки качества (precision, recall, F1-score) моделей и валидацией полученных биологических ассоциаций по независимым базам данных (ChEMBL, OncoDB). Результаты работы прошли апробацию на 8 международных и всероссийских конференциях и опубликованы в рецензируемых журналах, индексируемых в Scopus и Web of Science.

Соответствие содержания диссертации указанной специальности

Содержание диссертации полностью соответствует паспорту специальности 1.5.8. Математическая биология, биоинформатика.

Соответствие автореферата содержанию диссертации

Автореферат полностью и адекватно отражает основное содержание диссертации, методы исследования и полученные выводы.

Замечания и вопросы

Высоко оценивая диссертационную работу Бизюковой Н.Ю., отмечая ее актуальность, научную новизну, теоретическую и практическую значимость, ведущая организация считает целесообразным высказать следующие замечания и вопросы, носящие в основном уточняющий и дискуссионный характер.

1. В работе предложен интегральный подход, сочетающий CRF-модель, предобученную модель HunFlair, словарные методы и систему фразовых шаблонов. Вместе с тем из текста диссертации следует, что для отдельных типов связей, в частности связанных с miRNA и однонуклеотидными полиморфизмами, число устойчивых фразовых конструкций оказалось ограниченным, и в этих случаях извлечение ассоциаций осуществлялось на основе совместной встречаемости терминов в пределах предложения. В этой связи представляет интерес, насколько устойчивы результаты метода для таких классов ассоциаций и в какой степени указанное упрощение может влиять на увеличение числа ложноположительных связей.

2. Автор показывает, что разработанный подход обеспечивает высокие показатели качества распознавания именованных сущностей и приемлемые показатели точности извлечения ассоциаций. В то же время из представленных в диссертации данных следует, что по задаче relation extraction современные графовые нейросетевые архитектуры в ряде случаев могут демонстрировать более высокий F1-score. В связи с этим представляет интерес уточнение границ применимости предложенного подхода: в каких задачах, по мнению автора, приоритетными являются интерпретируемость и вычислительная экономичность, а в каких случаях может быть оправдан переход к более ресурсоемким нейросетевым моделям.

3. Существенная часть работы связана с формированием репрезентативной коллекции текстов и их категоризацией по типам исследований. При этом в диссертации отмечается, что для построения качественных обучающих выборок при использовании методов машинного обучения с учителем требуется значительный объем ручной разметки. В этой связи хотелось бы увидеть более развернутый комментарий относительно трудоемкости подготовки обучающих и тестовых выборок, а также относительно того, насколько предложенная схема масштабируется при переносе на другие предметные области.

4. Проверка полноты и согласованности извлеченных сведений проведена на материале противовирусных соединений с использованием базы данных ChEMBL. Такой выбор представляется вполне обоснованным, однако в определенной степени сужает область валидации. Кроме того, в диссертации показано, что пересечение публикаций собственной выборки с публикациями, представленными в ChEMBL, было относительно небольшим, что, по-видимому, связано как с различием временных диапазонов, так и с различием типов включенных исследований. В связи с этим представляет интерес, в какой степени полученные оценки полноты могут быть экстраполированы на другие классы соединений и иные биомедицинские направления.

5. В диссертации показано, что нормализация наименований объектов выполнялась с использованием внешних фактографических ресурсов, а в базе данных предусмотрены специальные поля для хранения внешних идентификаторов. Вместе с тем из текста работы не в полной мере ясно, каким образом обрабатываются случаи неоднозначной или неполной нормализации, особенно для редко встречающихся соединений, новых обозначений и вариантов написания. Было бы полезно уточнить, какова доля сущностей, для которых нормализация остается неполной, и как это может влиять на качество последующего анализа и агрегации данных.

6. Практически значимым результатом работы является создание открытой базы данных и веб-ресурса. Вместе с тем в диссертации лишь кратко обозначено наличие поля, фиксирующего факт ручной проверки достоверности извлеченной ассоциации. В этой связи представляет интерес, предполагается ли в дальнейшем систематическая экспертная валидация ресурса, по каким критериям будет осуществляться приоритизация записей для ручной проверки и каким образом планируется обеспечивать актуализацию базы данных по мере появления новых публикаций.

7. В качестве одного из перспективных направлений применения результатов работы рассматривается репозиционирование лекарственных средств. Данное направление представляется весьма важным и практически значимым. Вместе с тем хотелось бы видеть более детальное обсуждение того, какие именно критерии, помимо наличия текстовых ассоциаций, автор считает необходимыми для перехода от автоматически сформированной гипотезы к ее биологической и фармакологической верификации.

8. В обзоре литературы подробно обосновывается значимость подсловной токенизации для химических наименований, однако в описании собственного метода автора токенизация химических соединений представлена как сегментация «по пробелам и символам» без уточнения того, какие именно символы используются в качестве разделителей. Между тем в номенклатуре IUPAC отдельные символы и суффиксы несут самостоятельную химическую семантику: дефисы связывают позицию с заместителем, скобки обозначают стереохимию или разветвление, запятые разделяют позиции замещения, а суффиксы указывают на функциональные группы и, соответственно, на ожидаемые химические свойства соединения. Набор признаков токенов, приведенный в работе, включает главным образом общие дескрипторы (длину, наличие цифр, регистр, последние символы), но не содержит специализированных химических признаков. При этом следует отметить, что структурно близкие соединения, как правило, обладают сходной биологической активностью, поэтому и без учета тонкостей номенклатуры модель, по-видимому, способна корректно распознавать значительную часть наименований. В этой связи было бы интересно узнать, оценивалось ли влияние включения специализированных химических дескрипторов, характерных IUPAC (суффиксов, стереохимических префиксов, нумерации позиций заместителей) на точность распознавания наименований химических соединений, и если да, то к какому изменению метрик это привело.

9. Следует отметить отдельные редакционные неточности в оформлении диссертации. В частности, нарушена нумерация рисунков: на странице 115 рисунок, на который в тексте дана ссылка как на рисунок 15, имеет подпись «Рисунок 16», в результате чего на страницах 115 и 117 присутствуют два разных рисунка с одинаковым номером 16. Кроме того, в тексте встречаются отдельные опечатки и неточности оформления ссылок: в разделе «Положения, выносимые на защиту» указано «обеспечивает высокую точности» вместо «обеспечивает высокую точность»; на странице 13 - «дальшейшее» вместо «дальнейшее»; в формулировке задачи 2 на странице 13 имеется избыточная запятая перед союзом «и» в выражении «Разработка, тестирование, и программная реализация». Также ссылка [167] из списка литературы не цитируется в тексте, что, вероятно, связано с технической неточностью при цитировании на странице 80, где указано [166, 168] вместо [166–168].

Указанные замечания и вопросы не снижают общей высокой оценки диссертационной работы, которая представляет собой завершенное научно-квалификационное исследование, выполненное на актуальную тему и содержащее новые научные и практически значимые результаты.

Заключение о соответствии работы требованиям ВАК

Диссертационная работа Бизюковой Надежды Юрьевны «Формирование знаний о биологической активности низкомолекулярных органических соединений на основе автоматизированного анализа текстов» является завершенной научно-квалификационной работой, в которой решена важная научная задача, имеющая существенное значение для развития методов математической биологии, биоинформатики и автоматизированного извлечения фармакологически значимых знаний из биомедицинских текстов.

В работе представлен оригинальный подход, основанный на сочетании классических методов машинного обучения, современных предобученных моделей и методов, основанных на правилах. Такое сочетание позволило обеспечить рациональный баланс между точностью, интерпретируемостью результатов и вычислительной эффективностью, что имеет важное значение как для развития соответствующих методов, так и для их практического применения в информационно-аналитических системах биомедицинского и фармакологического профиля.

По своей актуальности, научной новизне, объему выполненных исследований, теоретической и практической значимости диссертационная работа полностью соответствует требованиям пункта 9 «Положения о присуждении ученых степеней», утвержденного Постановлением Правительства Российской Федерации от 24 сентября 2013 г. № 842 (с последующими изменениями), предъявляемым к кандидатским диссертациям, а ее автор, Бизюкова Надежда Юрьевна, заслуживает присуждения ученой степени кандидата биологических наук по специальности 1.5.8. Математическая биология, биоинформатика.

Диссертация, автореферат и настоящий отзыв были обсуждены, и отзыв был утвержден на заседании объединенного семинара подразделений Отдела системной биологии ИЦиГ СО РАН (Протокол №2 от 27 марта 2026 г).

Отзыв ведущей организации был составлен 25 марта 2026 года заведующим Лабораторией компьютерной протеомики, ведущим научным сотрудником ИЦиГ СО РАН, кандидатом биологических наук Владимиром Александровичем Иванисенко.

Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук» (ИЦиГ СО РАН)
630090, Новосибирск, пр-т Академика Лаврентьева, д.10
Телефон: +7(383) 363-49-80, Факс: +7(383) 333-12-78
E-mail: icg-adm@bionet.nsc.ru
https://www.icgbio.ru

Заведующий лабораторией компьютерной протеомики ИЦиГ СО РАН, ведущий научный сотрудник,
кандидат биологических наук
Телефон: +7(383)363-49-63*1320
E-mail: salix@bionet.nsc.ru

В.А. Иванисенко

Ученый секретарь ИЦиГ СО РАН
кандидат биологических наук

Г.В. Орлова