

DOI: <https://doi.org/10.21323/2618-9771-2026-9-2-282-293>



Поступила 28.10.2025

Поступила после рецензирования 04.06.2026

Принята в печать 09.06.2026

© Кузнецов М. А., Никитин И. А., 2026

<https://www.fsjour.com/jour>

Научная статья

Открытый доступ

ХЕМОИНФОРМАТИЧЕСКАЯ ПРИОРИТИЗАЦИЯ БИОАКТИВНЫХ СОЕДИНЕНИЙ, АССОЦИИРОВАННЫХ С ХЛЕБНОЙ МАТРИЦЕЙ НА ОСНОВЕ ФАРМАКОЛОГИЧЕСКИХ АННОТАЦИЙ ChEMBL И МАШИННОГО ОБУЧЕНИЯ

Кузнецов М. А., Никитин И. А.*

Российский экономический университет имени Г. В. Плеханова, Москва, Россия

КЛЮЧЕВЫЕ СЛОВА: АННОТАЦИЯ

хлебная матрица;
хемоинформатика;
ChEMBL; RDKit;
фенольные кислоты;
случайный лес;
молекулярные
дескрипторы;
внутренняя проверка;
приоритизация
соединений

Хлеб и продукты на основе зернового сырья являются не только источником макронутриентов, но и сложной пищевой матрицей, содержащей низкомолекулярные соединения фенольной, ароматической и ферментационной природы. Для ряда таких соединений в экспериментальных и фармакологических базах описаны антиоксидантные, противовоспалительные, противогрибковые и рецепторно-опосредованные эффекты. При этом такие сведения относятся преимущественно к тест-системам для индивидуальных веществ и не доказывают проявление соответствующих эффектов непосредственно в составе хлеба. Целью работы была разработка воспроизводимой хемоинформатической схемы внутренней приоритизации соединений, ассоциированных с хлебной и зерновой матрицей, на основе структурных признаков и фармакологических аннотаций ChEMBL. В анализ включили 43 куративно обработанных соединения. После стандартизации структур и удаления дубликатов по InChIKey для каждой молекулы рассчитали 2058 признаков, включающих физико-химические дескрипторы и топологические фингерпринты. Функциональные метки формировали по данным ChEMBL для четырёх направлений активности: антиоксидантной, противовоспалительной, противогрибковой и AhR-модулирующей. Модели логистической регрессии, случайного леса, CatBoost и XGBoost сравнивали во вложенной повторной перекрёстной проверке в многометочной постановке; независимый внешний набор данных в настоящей работе не использовали. Наилучшее сочетание качества и устойчивости показал случайный лес: усреднённая точность ранжирования положительных примеров составила 0,8041 при 95 % доверительном интервале 0,7766–0,8320, ROC-AUC — 0,8056. Наиболее устойчивым оказался противовоспалительный класс, тогда как антиоксидантный характеризовался большей неоднородностью. Верхние позиции ранжирования заняли протокатеховая, кофейная, галловая и феруловая кислоты, кверцетин и (+)-катехин. Проверка при разделении соединений по молекулярным каркасам снизила усреднённую точность до 0,6245, что указывает на ограниченную переносимость модели. Полученные результаты следует рассматривать как инструмент предварительного отбора соединений для последующей внешней и экспериментальной проверки, а не как подтверждение их функционального действия в хлебной матрице.

ФИНАНСИРОВАНИЕ: Исследование выполнено за счет средств субсидии на выполнение государственного задания Министерства науки и высшего образования Российской Федерации в рамках научно-исследовательского проекта № FSSW-2025-0004.

Received 28.10.2025

Accepted in revised 04.06.2026

Accepted for publication 09.06.2026

© Kuznetsov M. A., Nikitin I. A., 2026

Available online at <https://www.fsjour.com/jour>

Original scientific article

Open access

CHEMOINFORMATIC PRIORITIZATION OF BIOACTIVE COMPOUNDS ASSOCIATED WITH THE BREAD MATRIX BASED ON ChEMBL PHARMACOLOGICAL ANNOTATIONS AND MACHINE LEARNING

Maksim A. Kuznetsov, Igor A. Nikitin*

Plekhanov Russian University of Economics, Moscow, Russia

KEYWORDS:

bread matrix;
chemoinformatics;
ChEMBL; RDKit;
phenolic acids; random
forest; molecular
descriptors; internal
validation; compound
prioritization

ABSTRACT

Bread and cereal-based products represent not only a source of macronutrients but also a complex food matrix containing low-molecular-weight phenolic, aromatic and fermentation-derived compounds. A number of these molecules have been associated with antioxidant, anti-inflammatory, antifungal and receptor-mediated effects in experimental studies and pharmacological databases. However, these records mainly describe test systems for individual compounds and do not demonstrate the corresponding effects directly in bread products. The objective was to develop a reproducible chemoinformatics framework for internal prioritization of compounds associated with bread and cereal matrices based on structural features and ChEMBL pharmacological annotations. The study included 43 curated compounds. After structural standardization and InChIKey deduplication, 2,058 molecular features were calculated, including physicochemical descriptors and topological fingerprints. Functional labels were generated from ChEMBL records for four activity classes: antioxidant, anti-inflammatory, antifungal, and AhR-modulating. Logistic regression, random forest, CatBoost, and XGBoost were compared using nested repeated multilabel cross-validation; no independent external dataset was used in this study. Random forest showed the best balance of performance and robustness, with a macro average precision of 0.8041 (95 % CI: 0.7766–0.8320) and a macro ROC-AUC of 0.8056. The anti-inflammatory class showed the most stable performance, whereas the antioxidant class was more heterogeneous. The highest-ranking compounds were protocatechuic, caffeic, gallic and ferulic acids, quercetin, and (+)-catechin. Evaluation with molecular scaffold-based splitting reduced macro average precision to 0.6245, indicating limited

ДЛЯ ЦИТИРОВАНИЯ: Кузнецов, М. А., Никитин, И. А. (2026). Хемоинформатическая приоритизация биоактивных соединений, ассоциированных с хлебной матрицей на основе фармакологических аннотаций ChEMBL и машинного обучения. *Пищевые системы*, 9(2), 282–293. <https://doi.org/10.21323/2618-9771-2026-9-2-282-293>

FOR CITATION: Kuznetsov, M. A., Nikitin, I. A. (2026). Chemoinformatic prioritization of bioactive compounds associated with the bread matrix based on ChEMBL pharmacological annotations and machine learning. *Food Systems*, 9(2), 282–293. <https://doi.org/10.21323/2618-9771-2026-9-2-282-293>

transferability to structurally distant compounds. The proposed approach should be regarded as a tool for preliminary candidate selection for further external and biological validation, rather than as experimental confirmation of their functional effects in the bread matrix.

FUNDING: This study was funded by a subsidy for the implementation of a state assignment from the Ministry of Science and Higher Education of the Russian Federation within the framework of research project No. FSSW-2025-0004.

1. Введение

Хлеб и продукты на основе зернового сырья традиционно рассматриваются как один из базовых компонентов рациона человека, определяющих поступление углеводов, белков, пищевых волокон, минеральных веществ и ряда микронутриентов. Вместе с тем хлебная матрица представляет собой сложную многокомпонентную систему, в которой присутствуют низкомолекулярные соединения с потенциальной биологической активностью. Показано, что содержание феруловой кислоты и других фенольных кислот изменяется на разных этапах перехода от зерна к готовому хлебу, а также зависит от вида зернового сырья, типа ферментации и условий выпечки [1–4]. Отдельные исследования указывают, что использование алейроновых и цельнозерновых фракций влияет не только на содержание фенольных кислот, но и на их доступность в пищевой матрице [5].

Помимо фенольных кислот, хлеб и связанные с ним ингредиенты содержат ароматические альдегиды и спирты, летучие продукты термических превращений, а также метаболиты дрожжевой и молочнокислой ферментации. Ванилин рассматривается как ароматическое соединение с описанным терапевтическим потенциалом [6]. Работы, посвящённые закваскам, бобово-зерновым матрицам и летучему профилю хлеба, показывают, что ферментационные процессы способны существенно изменять химический состав и сенсорные свойства готового продукта [7–10].

Наиболее изученной группой биоактивных компонентов зернового и хлебного происхождения остаются фенольные соединения. Для кверцетина, апигенина, кемпферола и некоторых производных нарингина описаны противовоспалительные, антимикробные и связанные с кишечной микробиотой эффекты [11–15]. Для фенольных кислот зернового происхождения установлены изменения профиля и антиоксидантной активности в процессе хлебопечения, а добавление фенольных соединений в хлеб может сопровождаться как антиоксидантными преимуществами, так и изменением структуры и сенсорных характеристик продукта [16–18]. Отдельно показано, что феруловая кислота способна проявлять противовоспалительные эффекты через регуляцию аутофагии и блокирование активации NLRP3-инфламасомы [19].

Значительный интерес представляет также возможная связь пищевых флавоноидов и продуктов микробной трансформации с модуляцией арилулглюкозородного рецептора (AhR), участвующего в регуляции барьерной функции кишечного эпителия, иммунного ответа и взаимодействия организма с микробными метаболитами. Для ряда флавоноидов показана способность выступать агонистами или антагонистами AhR, причём характер эффекта зависит как от структуры соединения, так и от клеточного контекста [20–22].

Хемоинформатика и методы машинного обучения позволяют перейти от разрозненных описательных сведений к формализованной оценке структурно-биологических закономерностей. Использование базы ChEMBL обосновано тем, что она объединяет вручную курируемые сведения о химических структурах, биологических активностях и типах измерений, применимых для построения прогнозных моделей [23]. Топологические молекулярные фингерпринты служат одним из распространённых способов представления структуры малых молекул в таких задачах [24]. Вместе с тем крупномасштабные сравнения алгоритмов на данных ChEMBL показывают, что качество прогнозов зависит не только от сложности модели, но и от схемы валидации, а повторное использование тех же данных при подборе параметров способно завышать оценку качества [25,26].

Следует отметить, что применение машинного обучения для прогнозной оценки пищевых и природных низкомолекулярных соединений уже имеет ряд методических прецедентов, хотя они решают иные задачи. Так, инструмент PhyteByte использовал фармакологические данные ChEMBL и сведения о пищевом составе для поиска продуктов, содержащих соединения с прогнозируемыми целевыми биологическими свойствами [27]. Reker et al. применили машинное обучение для выявления ранее неочевидных биологических эффектов пищевых и вспомогательных соединений, включая взаимодействия с белками, влияющими на фармакокинетику [28]. В работе Sanchez-Ruiz и Colmenarejo химическое пространство пищевых соединений сопоставлялось с белковыми мишенями человека на основе FoodB

и ChEMBL, что позволило выявлять пробелы в хемобиологической аннотации и прогнозировать потенциальные мишени [29]. Для природных соединений также показана возможность оценки функционального сходства с лекарственными молекулами по структурным признакам и известным белковым мишеням [30]. Ранее авторами настоящей работы были разработаны близкие по логике, но предметно и методически отличающиеся алгоритмы: модель прогнозирования активности пищевых и микробных метаболитов в отношении ER α [31] и *in silico*-пайплайн мультитаргетного скрининга флавоноидных нутрицевтических кандидатов при MASLD [32]. В отличие от перечисленных исследований, настоящая работа фокусируется не на универсальном поиске мишеней или одной рецепторной системе, а на матрично-ориентированной приоритизации соединений, ассоциированных с хлебной матрицей по нескольким функциональным классам.

В этой связи применение машинного обучения к соединениям хлебной матрицы должно строиться не как попытка получить окончательное доказательство их биологического действия, а как инструмент вычислительной приоритизации. Его задача состоит в том, чтобы на основе доступных структурных признаков и фармакологических аннотаций выделить группу соединений, наиболее перспективных для дальнейшего целевого изучения. Такая постановка особенно важна при малом числе объектов, поскольку чрезмерно сложные модели и недостаточно строгие схемы валидации могут приводить к завышенным оценкам качества и формировать необоснованно сильные выводы.

Методологически корректная структурно-фармакологическая приоритизация пищевых соединений требует соблюдения нескольких условий. Необходимо стандартизировать химические структуры и исключить скрытое дублирование молекул; сформировать прозрачную карту соответствия между функциональными классами и биологическими эффектами; учитывать неоднородность исходных метрик активности; сравнивать несколько алгоритмов, а не исходить из априорного выбора одной модели; использовать кросс-валидационные схемы, пригодные для малых мульти-лейбл наборов; анализировать не только средние метрики качества, но и ошибки, устойчивость ранжирования и переносимость модели на структурно отличающиеся соединения.

Настоящее исследование направлено на решение указанной задачи применительно к низкомолекулярным соединениям, ассоциированным с хлебной и зерновой матрицей. В работе предложена хемоинформатическая схема bread-bioact, объединяющая курацию структур, извлечение и нормализацию фармакологических данных ChEMBL, расчёт молекулярных дескрипторов, формирование многомерного пространства активности и сравнительную внутреннюю оценку моделей машинного обучения. В отличие от подходов, ориентированных только на один функциональный биологический эффект или один алгоритм, данная схема рассматривает соединения одновременно по четырём классам активности: антиоксидантной, противовоспалительной, противомикробной и AhR-модулирующей. При этом объектом моделирования являются фармакологические аннотации индивидуальных веществ, а не биологический эффект соединений в составе готового хлеба.

Научная новизна исследования заключается в разработке воспроизводимой схемы предварительного отбора соединений, ассоциированных с хлебной матрицей, по совокупности прогнозируемых структурно-фармакологических признаков в условиях малого и неоднородного набора данных. При построении модели особое значение придавалось контролю переобучения. Для этого сопоставление алгоритмов выполняли во вложенной повторной схеме перекрёстной проверки с учётом многометочной структуры данных; итоговые показатели качества и ранжирование соединений рассчитывали только на основании out-of-fold предсказаний, полученных для объектов, исключённых из обучения на соответствующем этапе. Дополнительно оценивали устойчивость результатов при изменении критериев формирования функциональных меток и при проверке обобщающей способности модели на группах соединений, различающихся по молекулярным каркасам. Независимая внешняя валидация в настоящей работе не проводилась и рассматривается как следующий этап исследования.

Целью исследования являлась разработка и внутренняя оценка хемоинформатической схемы для вероятностной приоритизации биоактивных соединений, ассоциированных с хлебной и зерновой матрицей, на основе структурных дескрипторов и нормализованных фармакологических аннотаций ChEMBL.

Для достижения поставленной цели были решены следующие задачи: сформировать и куративно обработать корпус низкомолекулярных соединений, ассоциированных с хлебной и зерновой матрицей; сопоставить указанные соединения с релевантными данными ChEMBL; нормализовать разнородные показатели биологической активности и сформировать бинарные мульти-лейбл метки для четырёх функциональных классов; рассчитать молекулярные дескрипторы и топологические фингерпринты; сравнить несколько алгоритмов машинного обучения в строгой кросс-валидационной схеме; выполнить анализ чувствительности, ошибок и устойчивости ранжирования; а также оценить роль доступных концентрационных сведений как дополнительного описательного слоя.

Рабочая гипотеза состояла в том, что структурные признаки соединений, ассоциированных с хлебной матрицей, в сочетании с куративно нормализованными фармакологическими аннотациями позволяют выявить воспроизводимый сигнал, достаточный для вероятностной приоритизации соединений по нескольким функциональным классам. При этом результаты такого анализа должны рассматриваться как основание для выбора кандидатов для последующей независимой и экспериментальной проверки, но не как замена биологической валидации и не как доказательство активности соединений непосредственно в составе хлеба.

2. Объекты и методы

Объектом исследования являлись низкомолекулярные соединения, ассоциированные с хлебной и зерновой матрицей, а также с ингредиентами и технологическими процессами, характерными для хлебопечения. В корпус включали фенольные кислоты, флавоноиды, ароматические альдегиды и спирты, летучие продукты термических превращений, а также метаболиты дрожжевой и молочнокислой ферментации. Отбор соединений проводили с учётом их описанного присутствия в зерновом сырье, отрубях, солоде, заквасках, дрожжевых культурах, пряностях и продуктах тепловой обработки. После сопоставления с химическими записями ChEMBL, стандартизации структур и удаления повторов итоговый набор включал 43 уникальных соединения; для каждого из них была сформирована единая структурная запись и рассчитан набор молекулярных признаков. Аудит ревизионного набора подтвердил отсутствие дубликатов по InChIKey и отсутствие соединений, исключённых из-за недоступности дескрипторов.

Первичная структурная информация была приведена к единому виду с использованием канонических SMILES и InChIKey. На этапе курации удаляли соли и малые фрагменты, выбирали основной органический фрагмент, нормализовали зарядовое состояние и таутомерную форму молекулы. Такая процедура была необходима не только для химической корректности описания, но и для исключения скрытого дублирования, особенно существенного при малом объёме выборки. В дальнейшем единицей анализа считалась не строка ис-

ходного списка названий, а одна уникальная химическая структура, идентифицированная по InChIKey.

Фармакологические сведения для отобранных соединений извлекали из базы ChEMBL. Для построения функциональных меток использовали только те записи, которые содержали числовые показатели активности и могли быть отнесены к одному из четырёх заранее заданных направлений: антиоксидантной, противовоспалительной, противогрибковой и AhR-модулирующей активности. Эти сведения рассматривались как фармакологические аннотации индивидуальных молекул, полученные в различных экспериментальных тест-системах, а не как данные о проявлении активности в хлебной матрице. Класс широкой антимикробной активности в основную ревизионную модель не включали, чтобы сохранить согласованность между системой меток, порогами активности и итоговыми таблицами результатов. В отличие от предварительной версии исследования, в которой число классов в разных частях текста колебалось между четырьмя и пятью, в настоящей редакции все расчёты и интерпретации относятся к четырём указанным функциональным классам.

Поскольку сведения ChEMBL получены в различных экспериментальных системах, показатели активности предварительно нормализовали и разделяли по типу биологической метрики. Концентрационно выраженные показатели, такие как IC₅₀, EC₅₀, Ki, Kd, AC₅₀ и GI₅₀, приводили к молярным единицам и преобразовывали в величину $p_{Activity}$ по формуле:

$$p_{Activity} = -\log_{10}(CM), \quad (1)$$

где CM — концентрация вещества, моль/л.

Минимальные ингибирующие концентрации, процентные показатели ингибирования или индукции, а также кратность изменения эффекта сохраняли в собственных шкалах, фиксируя направление действия и тип исходного теста. Такой порядок позволял использовать разнородные фармакологические записи, но одновременно требовал дополнительной проверки устойчивости выводов, поскольку ферментативные, клеточные, микробиологические и рецепторные тест-системы не являются биологически равнозначными.

На основе нормализованных записей для каждого соединения формировали бинарные метки активности. Соединение относили к положительным примерам по соответствующему классу, если для него имелись данные, удовлетворяющие заранее заданным порогам активности: достаточная активность в концентрационно выраженных тестах, минимальная ингибирующая концентрация не выше 32 мкг/мл, эффект не менее 50% в процентных тестах или не менее чем двукратное изменение в тестах с оценкой кратности эффекта. Для проверки влияния неоднородности исходных метрик дополнительно сравнивали три варианта разметки: основной, включающий все допустимые типы числовых наблюдений; более строгий, основанный только на концентрационно выраженных показателях активности; и промежуточный, в котором к концентрационным показателям добавлялись данные минимальной ингибирующей концентрации. Такое сопоставление позволило оценить, сохраняются ли результаты при исключении наиболее гетерогенных типов наблюдений. Подробные критерии формирования функциональных меток приведены в Таблице 1.

Таблица 1. Критерии формирования функциональных меток

Table 1. Criteria for formation of functional labels

А. Правила перевода исходных фармакологических наблюдений в бинарные метки

A. Rules for transforming initial pharmacological observations into binary labels

Тип исходного наблюдения	Принцип нормализации	Критерий положительной метки	Методическое примечание
Концентрационно выраженные показатели активности (IC ₅₀ , EC ₅₀ , Ki, Kd, AC ₅₀ , GI ₅₀ и др.)	Приведение к молярным единицам с последующим расчётом $p_{Activity}$	Значение не выше предварительно заданного порога активности для соответствующего функционального класса	Данный критерий использовали во всех вариантах разметки
Минимальная ингибирующая концентрация (MIC, MBC, MFC и родственные показатели)	Приведение к единой шкале концентраций, мкг/мл	≤ 32 мкг/мл	Включалась в основную и промежуточную схемы разметки
Процентный эффект (ингибирование, индукция, биологическая активность в процентах)	Сохранение в исходной шкале после проверки направления эффекта	≥ 50 %	Использовался только в расширенной схеме разметки
Кратность изменения эффекта (отношение значений)	Сохранение в нормализованной шкале кратности	≥ 2	Использовался только в расширенной схеме разметки
Направление эффекта	Ингибирующие эффекты учитывали напрямую; индуцирующие эффекты приводили к единой направленности относительно референсного состояния	—	Этап выполняли до бинаризации наблюдений
Отнесение к отрицательному классу	После фильтрации по релевантным конечным точкам и типам наблюдений	Соединение считали отрицательным, если не было ни одного наблюдения, удовлетворяющего критерию положительности	Оценка проводилась отдельно для каждого функционального класса

Таблица 1. Окончание / Table 1. End

Б. Функциональные классы, использованные в ревизионной модели
 B. Functional classes used in the revisionary model

Функциональный класс	Содержательная интерпретация	Число положительных соединений, n (%)	Примечание
Антиоксидантная активность	Показатели, отражающие антиоксидантный ответ и/или ингибирование процессов, связанных с окислительным повреждением	21 (48,8)	Наиболее распространённый класс в ревизионном корпусе
Противовоспалительная активность	Показатели, характеризующие ослабление воспалительного ответа	11 (25,6)	По итогам ревизионного прогона данный класс оказался наиболее устойчивым
Противогрибковая активность	Микробиологические и фармакологические конечные точки, относящиеся к подавлению роста грибов	12 (27,9)	Более редкий и менее стабильный класс по сравнению с противовоспалительным
AhR-модулирующая активность	Показатели, отражающие модуляцию сигнального пути арилуглеводородного рецептора	10 (23,3)	Включён во все основные ревизионные расчёты

Примечание. В статье использовали четыре функциональных класса; класс широкой антимикробной активности в основную оценку не включали. Для анализа чувствительности дополнительно сопоставляли три схемы разметки: расширенную, объединявшую все допустимые типы числовых наблюдений; строгую, основанную только на концентрационно выраженных показателях активности; и промежуточную, в которой к концентрационным показателям добавлялись данные минимальной ингибирующей концентрации. Переход к более строгим схемам сопровождался умеренным снижением усреднённой точности ранжирования положительных примеров, тогда как добавление MIC-сигнала не изменяло общую структуру ранжирования.

Источник: составлено авторами.

Таблица 2. Параметры сравниваемых моделей и схема перекрёстной проверки
 Table 2. Parameters of the compared models and scheme of cross-validation

А. Сравниваемые модели
 A. Compared models

Модель	Статус в сравнении	Параметры внутреннего подбора	Реализация
Логистическая регрессия	Линейная базовая модель	Параметры регуляризации и штрафа	CPU
Случайный лес	Ансамблевый метод	Число деревьев, глубина деревьев, минимальные размеры узлов, объём случайного подпространства признаков	CPU, многопоточный режим
CatBoost	Градиентный бустинг на решающих деревьях	Скорость обучения, глубина деревьев, число итераций, параметры регуляризации	GPU
XGBoost	Градиентный бустинг на решающих деревьях	Скорость обучения, глубина деревьев, число деревьев, параметры подвыборки объектов и признаков, регуляризация	GPU

Б. Схема перекрёстной проверки и отбора модели
 B. Scheme of cross-validation and model selection

Элемент методики	Реализация в ревизионном анализе	Назначение
Постановка задачи	Для каждого функционального класса строили отдельный бинарный классификатор на общем пространстве молекулярных признаков	Обеспечение прозрачной многометочной постановки на малом корпусе
Внешний контур перекрёстной проверки	5 разбиений × 10 повторов, всего 50 out-of-fold проверочных разбиений; не является независимой внешней валидацией	Внутренняя оценка качества без переиспользования тестовой информации
Внутренний контур настройки	Подбор параметров выполняли только внутри обучающей части каждого внешнего разбиения перекрёстной проверки	Исключение утечки информации при выборе гиперпараметров
Тип стратификации	Многометочная стратификация	Сохранение структуры совместной встречаемости меток при разбиении выборки
Итоговые предсказания	Использовали только out-of-fold проверочные предсказания	На их основе рассчитывали метрики, выполняли анализ ошибок и строили итоговое ранжирование
Основной критерий качества	Macro Average Precision	Наиболее информативная метрика для задач приоритизации при несбалансированных классах
Дополнительные показатели	Macro ROC-AUC и средняя ошибка Brier	Оценка разделимости классов и калибровки вероятностей
Критерий выбора итоговой модели	Максимальное среднее значение Macro Average Precision с применением правила одной стандартной ошибки	Предпочтение более устойчивой модели при близком уровне качества
Доверительные интервалы	95 % доверительные интервалы, рассчитанные бутстреп-процедурой	Количественная оценка неопределённости итоговых метрик
Обработка вырожденных разбиений	Для отдельных классов при необходимости использовали константную вероятностную модель	Предотвращение численной нестабильности в обучении
Дополнительная проверка обобщающей способности	Стресс-проверка с разделением по структурным каркасам молекул: 3 разбиения по 13 группам молекулярных каркасов	Оценка переносимости внутри текущего корпуса на химически более удалённые соединения

Примечание. В таблице приведены группы параметров, определяющие структуру и сложность сравниваемых моделей, а также схема внутренней перекрёстной проверки. Результаты сравнительного отбора модели приведены в разделе «Результаты и обсуждение».

Источник: составлено авторами.

Молекулярное описание соединений включало два взаимодополняющих блока признаков. Первый блок составляли физико-химические дескрипторы, отражающие размер молекулы, полярность, липофильность, способность к образованию водородных связей, ароматичность, насыщенность и конформационную гибкость. Второй блок представлял собой топологические фингерпринты Morgan, кодирующие локальное атомное окружение молекулы. В совокупности для каждого соединения было рассчитано 2058 признаков, что обеспечивало достаточно подробное описание химической структуры, но одновременно усиливало необходимость строгой проверки модели из-за несоответствия между числом признаков и числом объектов.

Модельный анализ был построен как задача многометочной классификации, при которой для каждого функционального класса обучался отдельный бинарный классификатор на общем пространстве молекулярных признаков. Такой подход был выбран как наиболее прозрачный для малого корпуса соединений: он позволял независимо оценивать качество прогноза по каждому направлению активности и избегать искусственного моделирования всех возможных комбинаций меток. Для исключения необоснованной привязки к одному алгоритму сравнивали несколько моделей различной сложности: логистическую регрессию как линейный базовый метод, случайный лес как устойчивый ансамблевый алгоритм [33], а также CatBoost и XGBoost как реализации градиентного бустинга. Окончательный выбор модели проводили не априорно, а по результатам сравнительной проверки, представленной в разделе «Результаты и обсуждение».

С учётом малого объёма корпуса основную оценку качества выполняли не на одном разбиении обучающей и тестовой выборки, а во вложенной повторной перекрёстной проверке. Внешний контур этой перекрёстной проверки включал пять разбиений и десять повторов, а внутренний контур использовался только для подбора гиперпараметров. Следует подчеркнуть, что под внешним контуром в данном случае понимается только проверочная часть вложенной кросс-валидации; независимый внешний набор данных в настоящей работе не использовали. Итоговые метрики рассчитывались по out-of-fold предсказаниям для тех соединений, которые в соответствующем внешнем разбиении не участвовали в обучении модели. Это принципиально отличает ревизионную схему от предварительного варианта анализа, в котором сохранялся риск завышения качества из-за использования одних и тех же объектов при обучении и оценке. Параметры сравниваемых моделей и схема перекрёстной проверки приведены в Таблице 2.

Основной метрикой качества была средняя точность ранжирования положительных примеров, поскольку она лучше отражает пригодность модели для приоритизации соединений при несбалансированных классах. Дополнительно рассчитывали площадь под ROC-кривой и вероятностную ошибку Brier, характеризующую калибровку предсказанных вероятностей. Метрики определяли как для каждого функционального класса отдельно, так и в усреднённом виде по всем четырём классам. Для ключевых сводных показателей рассчитывали доверительные интервалы, что позволило не ограничиваться точечными оценками качества.

Особое внимание уделяли анализу устойчивости выводов. Во-первых, сравнивали результаты при разных вариантах формирования функциональных меток, чтобы оценить возможное влияние неоднородности IC₅₀, EC₅₀, Ki, MIC, процентных эффектов и других типов биологических измерений. Во-вторых, анализировали ошибки классификации, выделяя ложноположительные и ложноотрицательные случаи по каждому классу активности. В-третьих, выполняли дополнительную проверку с разделением соединений по структурным каркасам молекул. Такая проверка является более строгой, чем обычная перекрёстная валидация, поскольку оценивает переносимость модели на химически менее близкие соединения. В ревизионном анализе она рассматривалась как стресс-тест обобщающей способности, а не как замена основной схемы оценки.

Для итогового ранжирования соединений использовали среднюю предсказанную вероятность активности по четырём функциональным классам, рассчитанную на основе out-of-fold проверочных предсказаний. Все классы при этом имели равный вес. Такой способ ранжирования был выбран для того, чтобы не усиливать искусственно вклад редких классов и не смещать итоговый рейтинг за счёт поправочных коэффициентов, зависящих от распространённости меток. Мультипликативный индекс, применявшийся в предварительной версии анализа, был оставлен только для сопоставления с прежними результатами и не использовался как основной критерий приоритизации.

Концентрационные данные о присутствии соединений в хлебной матрице рассматривались отдельно от модельного ранжирования. Они не включались в обучающие признаки, не использовались при расчёте предсказанных вероятностей и не изменяли итоговый рейтинг соединений. Такой подход обусловлен неполным покрытием корпуса: сопоставленные концентрационные сведения были доступны только для 4 из 43 соединений. Поэтому концентрационный слой использовали исключительно как дополнительный контекст при обсуждении физиологической интерпретируемости результатов, но не как доказательство функциональной значимости ведущих кандидатов.

Все этапы анализа были реализованы в едином воспроизводимом вычислительном контуре на языке Python: от сопоставления соединений с ChEMBL и стандартизации структур до расчёта дескрипторов, формирования меток, сравнения моделей, оценки ошибок и построения итогового ранга. Для подготовки данных использовали pandas и NumPy; для сопоставления с ChEMBL — chembl_webresource_client; для стандартизации структур, расчёта дескрипторов и Morgan-фингерпринтов — RDKit; для стратифицированной перекрёстной проверки, логистической регрессии, случайного леса и расчёта метрик — scikit-learn и iterative-stratification; для градиентного бустинга — CatBoost и XGBoost. Повторяемость расчётов обеспечивали фиксированными начальными состояниями генераторов случайных чисел. При подготовке настоящей редакции использовали только результаты ревизионного контура, основанные на повторной перекрёстной проверке и out-of-fold предсказаниях; предварительные расчёты с одиночным разбиением или оценкой на обучающих объектах в итоговые выводы не включались.

3. Результаты и обсуждение

После сопоставления исходного перечня соединений с ChEMBL, стандартизации химических структур и удаления повторов был сформирован итоговый корпус, основные параметры которого представлены в Таблице 3. Таблица фиксирует состав набора и распределение функциональных меток и служит отправной точкой для последующей оценки моделей.

Таблица 3. Характеристика сформированного набора данных и распределение функциональных меток

Table 3. Characteristics of the formed data set and distribution of functional labels

Показатель	Значение
Общая характеристика набора	
Число уникальных соединений	43
Число молекулярных признаков на одно соединение	2058
Соединений с рассчитанными дескрипторами	43
Исключено из-за недоступности дескрипторов	0
Дубликаты по InChIKey после курации	0
Распределение положительных функциональных меток	
Антиоксидантная активность	21 (48,8%)
Противовоспалительная активность	11 (25,6%)
Противогрибковая активность	12 (27,9%)
AhR-модулирующая активность	10 (23,3%)

Примечание. Одно и то же соединение могло иметь более одной функциональной метки; поэтому доли по классам не являются взаимоисключающими.

Источник: составлено авторами.

Распределение положительных меток по функциональным классам было неравномерным: наиболее представленным оказался антиоксидантный класс, тогда как противовоспалительное, противогрибковое и AhR-модулирующее направления были более редкими. Поскольку одно и то же соединение могло иметь несколько функциональных меток, доли по классам не являются взаимоисключающими. Такая структура набора подтверждает необходимость использовать метрики ранжирования и вероятностной оценки, а не ограничиваться простой долей верных классификаций.

Сравнительный анализ моделей показал преимущество случайного леса по основной метрике и его наиболее благоприятное соотношение качества и устойчивости. Логистическая регрессия показала близкий результат и использовалась как референтная модель, тогда как CatBoost и XGBoost не дали выигрыша на малом корпусе. Сводные численные показатели приведены в Таблице 4.

Таблица 4. Сравнение моделей по результатам вложенной повторной перекрёстной проверки

Table 4. Comparison of the models by the results of the nested repeated cross-validation

Модель	Средняя точность (Macro AP)	95 % ДИ для Macro AP	Macro ROC-AUC	Средняя ошибка Brier	Итог сравнительного отбора
Случайный лес	0,8041	0,7766–0,8320	0,8056	0,1750	Выбрана как итоговая модель
Логистическая регрессия	0,7886	0,7579–0,8191	0,7944	0,1760	Референтная модель
XGBoost	0,7610	0,7253–0,7971	0,7765	0,1741	Не показала преимуществ
CatBoost	0,7042	0,6730–0,7350	0,7430	0,1838	Не показала преимуществ

Примечание. Итоговая модель выбиралась по основной метрике и правилу одной стандартной ошибки.

Источник: составлено авторами.

Полученное соотношение моделей имеет принципиальное значение для интерпретации дальнейших результатов. Более сложные бустинговые алгоритмы не улучшили качество прогноза по сравнению со случайным лесом и линейной референтной моделью. Это указывает на то, что в сформированном наборе присутствует структурно-фармакологический сигнал, однако его извлечение не требует чрезмерного усложнения алгоритма. Поэтому в последующих расчётах использовалась модель случайного леса, выбранная по основной метрике и правилу одной стандартной ошибки как наиболее обоснованный вариант для малого и неоднородного корпуса. Такой результат согласуется с методическими данными о том, что при малых и структурно связанных химических наборах усложнение алгоритма не всегда повышает обобщающую способность, а выбор модели должен сопровождаться вложенной проверкой, исключающей смещение при подборе параметров [25,26].

Сопоставление с опубликованными работами показывает, что полученные показатели находятся в ожидаемом диапазоне для задач вычислительной приоритизации, однако не являются напрямую сопоставимыми из-за различий в объектах, целевых переменных и схемах валидации. В ранее опубликованных моделях ERα и мультитаргетного отбора флавоноидов использовались иные наборы данных, постановки задач и процедуры проверки [31,32]. Внешние исследования пищевых и природных соединений также подтверждают релевантность такого подхода: PhyteByte прогнозирует биоактивные пищевые молекулы по данным ChEMBL и пищевого состава [27], Reker et al. показали применимость машинного обучения для выявления биологических эффектов потребляемых соединений и вспомогательных веществ [28], а Sanchez-Ruiz и Colmenarejo продемонстрировали возможность системного анализа и прогнозирования белковых мишеней пищевых соединений [29]. На этом фоне разработанная схема bread-bioact занимает более узкую нишу: она не заменяет экспериментальную проверку, а обеспечивает матрично-ориентированную, но основанную на ChEMBL, приоритизацию соединений, ассоциированных с хлебной матрицей.

Поклассовая оценка выбранной модели выявила различия в степени предсказуемости отдельных функциональных направлений.

Наиболее устойчивым оказался противовоспалительный класс, тогда как антиоксидантная активность, несмотря на наибольшее число положительных примеров, характеризовалась менее однородным вероятностным разделением. Поклассовые значения средней точности, ROC-AUC и ошибки Brier приведены в Таблице 5.

Таблица 5. Поклассовые метрики выбранной модели

Table 5. Class-level metrics of the selected model

Функциональный класс	Положительных соединений, n (%)	Средняя точность (AP)	ROC-AUC	Ошибка Brier
Противовоспалительная активность	11 (25,6 %)	0,8680	0,9116	0,1444
AhR-модулирующая активность	10 (23,3 %)	0,8095	0,8142	0,1524
Антиоксидантная активность	21 (48,8 %)	0,7759	0,7364	0,2208
Противогрибковая активность	12 (27,9 %)	0,7630	0,7602	0,1821

Примечание. Наиболее устойчивые результаты получены для противовоспалительного класса; наименее однородным оказался антиоксидантный класс.

Источник: составлено авторами.

Таким образом, наиболее устойчивый сигнал был связан с противовоспалительным направлением, тогда как антиоксидантный класс объединял химически и, вероятно, механистически различающиеся группы соединений. Это обстоятельство важно для интерпретации, поскольку антиоксидантные свойства пищевых фенолов часто рассматриваются как наиболее очевидный эффект, тогда как модельная оценка показывает необходимость более осторожной трактовки. Поклассовые значения средней точности и площади под ROC-кривой представлены на Рисунке 1 и Рисунке 2 как визуальное сопровождение количественных метрик Таблицы 5.

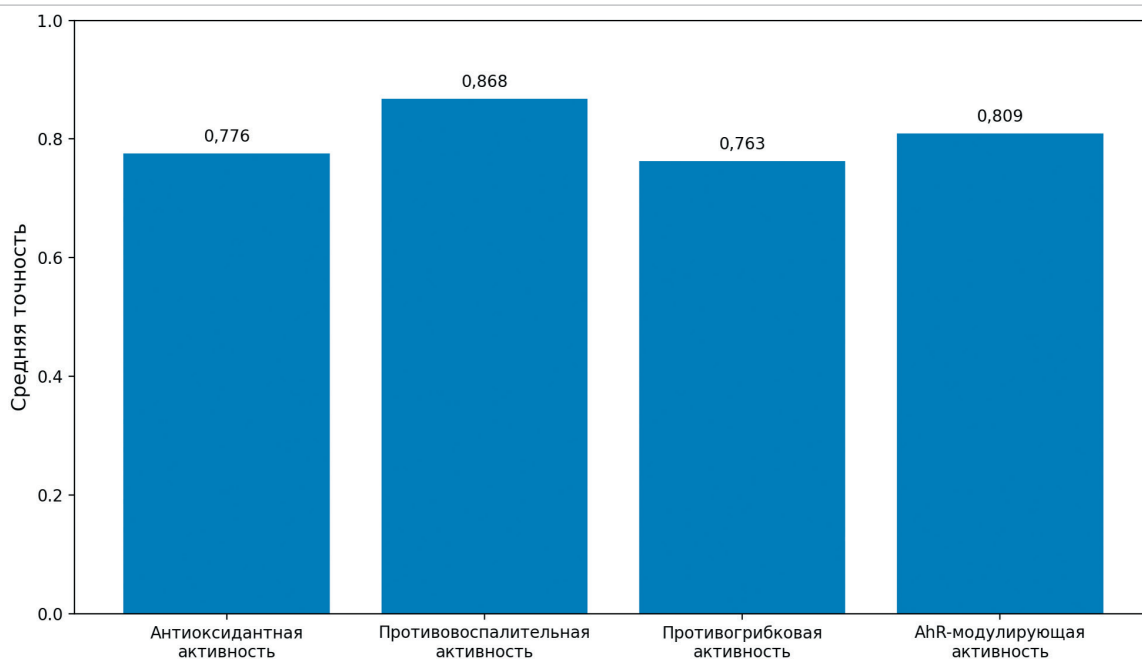


Рисунок 1. Поклассовые значения средней точности выбранной модели

Figure 1. Class-level average precision values of the selected model

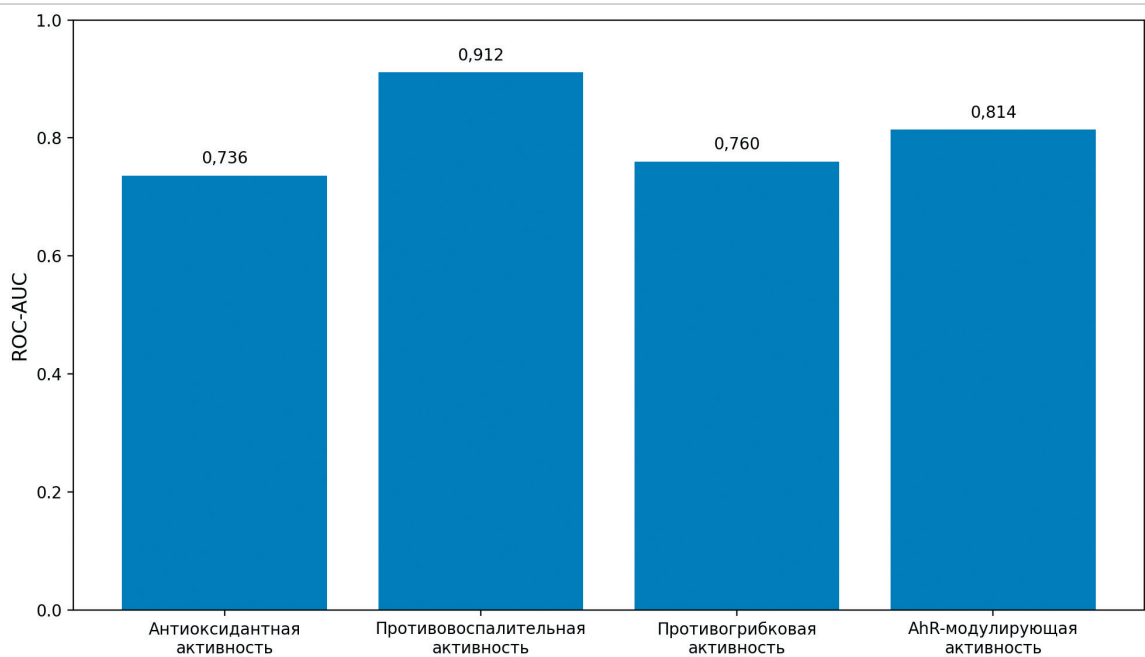


Рисунок 2. Поклассовые значения ROC-AUC выбранной модели
Figure 2. Class-level ROC-AUC values of the selected model

С учётом разнородности фармакологических данных была отдельно проверена устойчивость результатов к способу формирования функциональных меток. В основной схеме учитывались все допустимые числовые типы наблюдений, включая концентрационно выраженные показатели активности, минимальные ингибирующие концентрации, процентные эффекты и показатели кратности изменения. При переходе к более строгой схеме, основанной только на концентрационно выраженных показателях, средняя точность снизилась с 0,8041 до 0,7678, тогда как площадь под ROC-кривой увеличилась с 0,8056 до 0,8230, а ошибка Brier уменьшилась с 0,1750 до 0,1605. Добавление минимальных ингибирующих концентраций к концентрационным показателям не изменило итоговых значений. Эти результаты представлены в Таблице 6.

Такое соотношение показателей свидетельствует о том, что расширенная схема разметки несколько лучше поддерживает ранжирование положительных примеров, тогда как более строгая схема обеспечивает более благоприятную разделимость и калибровку вероятностей. Важно, однако, что изменение способа разметки не приводило к перестройке всей иерархии соединений. Между основной и более строгой схемами совпадали 8 из 10 первых позиций, а коэффициент ранговой корреляции Спирмена составлял 0,915. При сравнении двух строгих вариантов разметки совпадение первой десятки было полным, а коэффициент Спирмена равнялся 1,000. Следовательно, верхняя часть рейтинга не является артефактом одного конкретного способа бинаризации фармакологических данных.

Итоговое ранжирование соединений проводилось по средней вероятностной оценке активности, рассчитанной по четырём функциональным классам с равным вкладом каждого класса. Такой подход позволил отказаться от прежнего мультипликативного индекса, в котором редкие классы могли получать чрезмерное влияние за счёт поправочного коэффициента распространённости.

Переход к равновзвешенной вероятностной оценке сохранил верхнюю часть рейтинга, но сделал принцип приоритизации более прозрачным.

Наибольшие интегральные оценки были получены преимущественно для соединений фенольной и полифенольной природы. В первую десятку вошли протокатеховая, кофейная, галловая и феруловая кислоты, кверцетин, (+)-катехин, гидроксibenзойная кислота, ванилин, сиригновая кислота и эвгенол. Список соединений и значения интегральной оценки приведены в Таблице 7, а распределение вероятностей по отдельным функциональным классам для ведущих кандидатов представлено на Рисунке 3.

Рисунок 3 необходим для интерпретации не только общего ранга, но и профиля каждого кандидата. Интегральная оценка может формироваться разными сочетаниями классов: одни соединения занимают высокие позиции за счёт более сбалансированного профиля, тогда как другие демонстрируют локальный вклад одного-двух направлений активности. Поэтому тепловая карта позволяет различать соединения, перспективные для широкой предварительной проверки, и кандидаты, более релевантные для отдельных гипотез, например противовоспалительной или AhR-модулирующей. Одновременно такой профиль подчёркивает ограничение исследования: высокий ранг отражает структурно-фармакологическое сходство с активными аннотациями ChEMBL, но не является прямым доказательством биологической эффективности молекулы в составе хлеба.

Анализ ошибок подтвердил различия между функциональными направлениями. Наиболее благоприятный профиль был характерен для противовоспалительного класса, тогда как антиоксидантная активность сопровождалась наибольшим числом как ложноположительных, так и ложноотрицательных случаев. Сводный профиль ошибок при пороге вероятности 0,5 приведён в Таблице 8.

Таблица 6. Устойчивость результатов при различных схемах формирования функциональных меток
Table 6. Stability of the results in different schemes of the formation of functional labels

Схема формирования функциональных меток	Используемые типы фармакологических наблюдений	Средняя точность (Macro AP)	Macro ROC-AUC	Средняя ошибка Brier	Совпадение первой десятки с основной схемой	Коэффициент Спирмена
Основная схема	Концентрационно выраженные показатели, минимальные ингибирующие концентрации, процентные эффекты и показатели кратности изменения	0,8041	0,8056	0,1750	—	—
Строгая схема	Только концентрационно выраженные показатели активности	0,7678	0,8230	0,1605	8/10	0,915
Промежуточная схема	Концентрационно выраженные показатели активности и минимальные ингибирующие концентрации	0,7678	0,8230	0,1605	8/10	0,915

Примечание. При сравнении строгой и промежуточной схем совпадение первой десятки составило 10/10, коэффициент Спирмена равен 1,000.
Источник: составлено авторами.

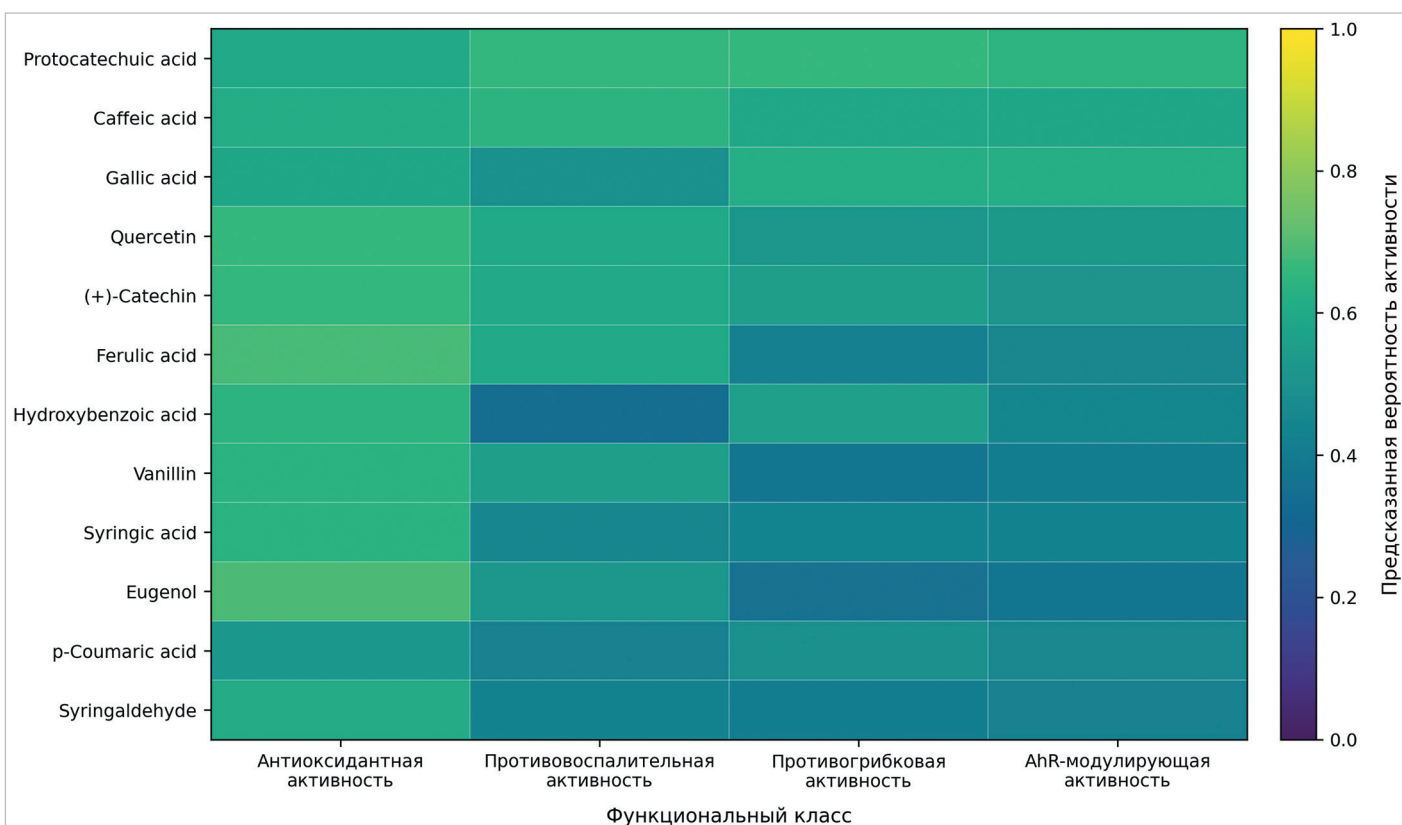


Рисунок 3. Вероятностные профили соединений с наибольшей интегральной оценкой структурно-фармакологической приоритизации. Цветовая шкала отражает предсказанную вероятность активности по каждому функциональному классу

Figure 3. Probability profiles of compounds with the highest integral score of structural-pharmacological prioritization. The color scale represents the predicted probability of activity in each functional class

Таблица 7. Десять соединений с наибольшей интегральной вероятностной оценкой структурно-фармакологической приоритизации

Table 7. Ten compounds with the highest integral probability-based score of structural-pharmacological prioritization

№	Соединение	Источник происхождения	Средняя out-of-fold вероятностная оценка
1	Протокатеховая кислота	Отруби	0,6416
2	Кофейная кислота	Отруби	0,6057
3	Галловая кислота	Добавки / отруби	0,5793
4	Кверцетин	Добавки / отруби	0,5773
5	(+)-Катехин	Добавки / солод	0,5771
6	Феруловая кислота	Отруби	0,5401
7	Гидроксibenзойная кислота	Отруби	0,5003
8	Ванилин	Термически образованные соединения	0,4934
9	Сиринговая кислота	Отруби	0,4903
10	Эвгенол	Пряности	0,4872

Примечание. Интегральная оценка рассчитывалась как средняя out-of-fold вероятностная оценка активности по четырём функциональным классам при равном вкладе каждого класса.

Источник: составлено авторами.

Наиболее выраженные ложноположительные предсказания в антиоксидантном классе относились к эвгенолу, гидроксibenзойной кислоте, ацетону, индол-3-уксусной кислоте и р-кумаровой кислоте. Среди ложноотрицательных примеров находились γ -аминомасляная кислота, фенол, 2-метилфуран, индол, триптофол, тирамин и гистамин. Эти данные позволяют уточнить характер ограничений модели: антиоксидантный класс, по-видимому, объединяет соединения с различными структурными основаниями активности, что снижает однозначность вероятностной классификации. Подробные примеры ложноположительных и ложноотрицательных случаев приведены в Таблице 9.

Таблица 8. Сводный профиль ошибок выбранной модели по функциональным классам

Table 8. Summary error profile of the selected model by functional classes

Функциональный класс	Истинно положительные	Истинно отрицательные	Ложноположительные	Ложноотрицательные
Антиоксидантная активность	14	15	7	7
Противовоспалительная активность	7	31	1	4
Противогрибковая активность	6	31	0	6
AhR-модулирующая активность	5	33	0	5

Примечание. Профиль ошибок приведён при пороговом значении вероятности 0,5.

Источник: составлено авторами.

Дополнительная проверка переносимости модели была выполнена с разделением соединений по молекулярным каркасам. В стандартной повторной перекрёстной проверке усреднённая средняя точность составляла 0,8041, а площадь под ROC-кривой 0,8056. При более строгом разделении по структурным каркасам эти значения снижались до 0,6245 и 0,6627 соответственно. Данная процедура включала три проверочных разбиения по 13 группам молекулярных каркасов и рассматривалась как дополнительная проверка обобщающей способности модели, а не как замена основной схемы оценки. Результаты сопоставления представлены на Рисунке 4.

Снижение качества при разделении по структурным каркасам показывает, что часть предсказательной способности модели связана со структурной близостью соединений внутри текущего корпуса. Следовательно, полученные результаты характеризуют прежде всего внутреннюю устойчивость модели в пределах сформированного химического пространства и не должны без дополнительной

Таблица 9. Репрезентативные ложноположительные и ложноотрицательные предсказания выбранной модели
Table 9. Representative false positive and false negative predictions of the selected model

Функциональный класс	Тип случая	Соединение	Предсказанная вероятность	Методическое значение примера
Антиоксидантная активность	Ложноположительный	Эвгенол	0,6897	Высокая вероятность при отсутствии положительной метки; пример возможного структурного сходства с активными фенольными соединениями
Антиоксидантная активность	Ложноположительный	Гидроксibenзойная кислота	0,6422	Соединение входит в верхнюю часть ранга, но является ошибочным случаем для данного класса
Антиоксидантная активность	Ложноположительный	Ацетоин	0,5798	Показывает возможность переноса сигнала на химически менее типичное соединение
Антиоксидантная активность	Ложноположительный	Индол-3-уксусная кислота	0,5549	Иллюстрирует пограничную область между ароматическими метаболитами и фенольными соединениями
Антиоксидантная активность	Ложноположительный	p-Кумаровая кислота	0,5241	Вероятность близка к пороговому уровню; пример неоднозначного фенольного профиля
Антиоксидантная активность	Ложноотрицательный	γ-Аминомасляная кислота	0,2658	Низкая вероятность при наличии положительной метки; пример химически отличающегося положительного объекта
Антиоксидантная активность	Ложноотрицательный	Фенол	0,3587	Показывает ограниченность структурного сигнала для простых ароматических соединений
Антиоксидантная активность	Ложноотрицательный	2-Метилфуран	0,3825	Пограничный случай для летучих соединений термического происхождения
Антиоксидантная активность	Ложноотрицательный	Индол	0,4129	Иллюстрирует неоднозначность индольных производных для антиоксидантного класса
Антиоксидантная активность	Ложноотрицательный	Триптофол	0,4199	Пограничный метаболит ферментационного происхождения
Антиоксидантная активность	Ложноотрицательный	Тирамин	0,4340	Пример амина, недостаточно хорошо описываемого антиоксидантным структурным профилем
Антиоксидантная активность	Ложноотрицательный	Гистамин	0,4538	Вероятность ниже порога при положительной метке; отражает неоднородность класса
Противовоспалительная активность	Ложноположительный	Карвакрол	0,5435	Единственный выраженный ложноположительный пример для наиболее устойчивого класса
Противовоспалительная активность	Ложноотрицательный	Триптофол	—	Химически правдоподобный кандидат, не достигший порогового уровня вероятности
Противовоспалительная активность	Ложноотрицательный	Гидроксibenзойная кислота	—	Пример расхождения между ранговым положением соединения и отдельным функциональным классом
Противовоспалительная активность	Ложноотрицательный	Тимол	—	Пограничный случай среди ароматических соединений пряностей
Противовоспалительная активность	Ложноотрицательный	Галловая кислота	—	Высокоранговое соединение, не всегда корректно классифицируемое по отдельному классу

Примечание. Знак «—» означает, что точная агрегированная вероятность для данного соединения не была приведена в сводном отчёте; случай включён как репрезентативный пример ошибки соответствующего класса.

Источник: составлено авторами.

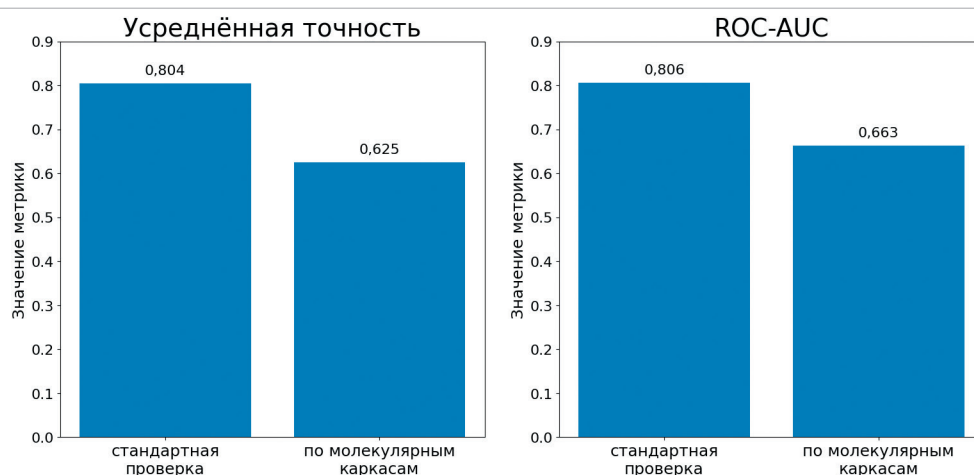


Рисунок 4. Сопоставление качества модели при стандартной повторной перекрёстной проверке и при проверке с разделением соединений по молекулярным каркасам

Figure 4. Comparison of model performance under standard repeated cross-validation and validation with compound separation by molecular scaffolds

Таблица 10. Концентрационные сведения для соединений, сопоставленных с хлебной матрицей
Table 10. Concentration data for compounds compared with the bread matrix

Ранг по вероятностной оценке	Соединение	Средняя вероятностная оценка	Концентрационные сведения, мг/кг	Интерпретация для настоящей работы
21	4-Винилфенол	0,3846	Имеется совпавшая ссылка, но без устойчивого числового резюме в объединённой таблице	Концентрационные сведения подтверждают наличие соединения в хлебной матрице, но не позволяют использовать его как количественный ранжирующий фактор
23	Ацетоин	0,3747	0,05–0,10	Соединение относится к среднему рангу; концентрационный слой имеет описательное значение
33	2,3-Бутандиол	0,2877	0,01–0,50	Низкая вероятностная оценка при наличии концентрационных сведений; пример независимости концентрационного слоя от модельного ранга
40	Этанол	0,2604	0,058–12800,0	Широкий диапазон концентраций отражает технологическую вариабельность и не используется для изменения ранга

Примечание. Концентрационные значения нормированы к мг/кг и использованы только как описательный слой. Они не входили в обучающие признаки, не изменяли предсказанные вероятности и не использовались при формировании основного рейтинга соединений.
Источник: составлено авторами.

проверки переносится на широкие внешние библиотеки соединений. Этот вывод является существенным для последующего обсуждения, поскольку задаёт границы применимости разработанной вычислительной схемы. Полученный результат согласуется с общей проблемой смещения, связанного со структурной близостью соединений в хемоинформатических наборах данных, и подчёркивает необходимость проверки моделей на химически независимых группах объектов [25,26].

Концентрационные сведения о присутствии соединений в хлебной матрице были проанализированы отдельно от модельного ранжирования. В ревизионном наборе такие сведения удалось сопоставить только для 4 из 43 соединений. Все значения были приведены к единой шкале, однако не использовались при обучении моделей, расчёте вероятностей или изменении итогового ранга. Это решение было обусловлено неполным покрытием корпуса концентрационными данными и необходимостью отделить структурно-фармакологическую приоритизацию от оценки реального содержания соединений в пищевой матрице.

Ни одно из соединений, вошедших в первую десятку рейтинга, не имело надёжного концентрационного резюме в объединённой таблице ревизионного анализа. Сопоставленные концентрационные данные относились главным образом к соединениям среднего и нижнего рангов, включая ацетоин, 2,3-бутандиол и этанол. Поэтому концентрационный блок в настоящей работе выполняет только описательную функцию и не может служить самостоятельным подтверждением физиологической значимости ведущих кандидатов. Детализированные концентрационные сведения приведены в Таблице 10.

В совокупности результаты показывают, что в сформированном корпусе соединений, ассоциированных с хлебной матрицей, выявляется воспроизводимый структурно-фармакологический сигнал, достаточный для вероятностной приоритизации кандидатов. Вместе с тем различия между функциональными классами, неодинаковый профиль ошибок, снижение качества при разделении по молекулярным каркасам, отсутствие независимого внешнего набора данных и ограниченность концентрационных сведений задают необходимые границы интерпретации. Поэтому полученные данные следует рассматривать как основание для отбора перспективных соединений для последующей внешней и экспериментальной проверки, а не как окончательное подтверждение их функциональной эффективности в составе пищевого продукта.

4. Выводы

В результате проведённого исследования разработана и апробирована хемоинформатическая схема внутренней вероятностной приоритизации низкомолекулярных соединений, ассоциированных с хлебной и зерновой матрицей. Сформированный корпус включал 43 уникальных соединения, для которых после структурной курации и исключения дубликатов был рассчитан комплекс молекулярных признаков и сформированы бинарные метки по четырём функциональным направлениям: антиоксидантному, противовоспалительному, противогрибковому и AhR-модулирующему. Тем самым была создана воспроизводимая основа для сопоставления химической структуры соединений с доступными фармакологическими аннотациями ChEMBL.

Сравнительный анализ моделей показал, что на данном наборе соединений наилучшее сочетание качества и устойчивости обеспечивает алгоритм случайного леса. Его усреднённая по четырём классам средняя точность ранжирования положительных примеров

составила 0,8041 при 95 % доверительном интервале 0,7766–0,8320; площадь под ROC-кривой равнялась 0,8056, а средняя ошибка Brier 0,1750. Полученные показатели характеризуют внутреннюю проверку во вложенной повторной перекрёстной схеме и не должны трактоваться как результат независимой внешней валидации.

Поклассовый анализ выявил неодинаковую степень предсказуемости функциональных направлений. Наиболее согласованный результат был получен для противовоспалительной активности, тогда как антиоксидантный класс, несмотря на наибольшее число положительных примеров, характеризовался более выраженной неоднородностью и менее благоприятным профилем ошибок. Это указывает на то, что антиоксидантная активность в рамках рассматриваемого корпуса, вероятно, объединяет несколько различных химико-биологических механизмов и потому требует более осторожной интерпретации, чем классы с более однородным модельным сигналом.

Итоговое ранжирование соединений позволило выделить группу фенольных и полифенольных кандидатов с наибольшей интегральной вероятностной оценкой структурно-фармакологической приоритизации. В верхние позиции вошли протокатеховая, кофейная, галловая и феруловая кислоты, кверцетин и (+)-катехин. Переход от ранее использовавшегося мультипликативного индекса к равновзвешенной вероятностной оценке не изменил принципиально состав ведущих кандидатов, но устранил риск непропорционального усиления редких классов активности. Это повышает методическую прозрачность ранжирования и делает его более пригодным для последующего отбора соединений для независимой проверки.

Проверка чувствительности показала, что изменение способа формирования функциональных меток не приводит к радикальной перестройке верхней части рейтинга. Вместе с тем разделение соединений по структурным каркасам сопровождалось снижением усреднённой точности с 0,8041 до 0,6245, что указывает на ограниченную переносимость модели на химически более удалённые соединения. Следовательно, разработанная схема наиболее надёжна как инструмент внутренней приоритизации в пределах заданного химического пространства, а её применение к расширенным библиотечкам пищевых молекул и к независимым наборам данных требует дополнительной валидации.

Концентрационные сведения о присутствии соединений в хлебной матрице в настоящей работе имели вспомогательное значение. Их покрытие оказалось недостаточным для включения в основной ранговый показатель: сопоставленные данные были доступны только для 4 из 43 соединений и не охватывали ведущие позиции рейтинга. Поэтому полученные модельные оценки не должны трактоваться как прямое свидетельство физиологической значимости соединений в составе готового пищевого продукта. Они отражают прежде всего структурно-фармакологический потенциал индивидуальных молекул по данным ChEMBL и задают направление для дальнейших экспериментальных исследований.

Таким образом, предложенный подход позволяет перейти от разрозненного описания отдельных пищевых биоактивных соединений к формализованной системе их вероятностной приоритизации. Научная значимость работы состоит в демонстрации возможности аккуратного применения методов хемоинформатики и машинного обучения к малому корпусу соединений, ассоциированных с хлебной матрицей, при условии строгой курации данных, сравнительной оценки моделей, анализа ошибок и явного указания границ

применимости результатов. Практическое значение исследования заключается в формировании обоснованного перечня соединений-кандидатов для последующей внешней, литературной и экспериментальной валидации. Дальнейшее развитие работы должно быть связано с расширением корпуса соединений, уточнением перечня

оцениваемых биологических эффектов и критериев их отнесения к функциональным классам, включением данных о концентрациях, биодоступности и метаболических превращениях, а также с проверкой ведущих кандидатов в релевантных *in vitro* моделях и на независимых наборах данных.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК / REFERENCES

- Boudaoud, S., Sicard, D., Suc, L., Conéjéro, G., Segond, D., Aouf, C. (2021). Ferulic acid content variation from wheat to bread. *Food Science and Nutrition*, 9(5), 2446–2457. <https://doi.org/10.1002/fsn3.2171>
- Boz, H. (2015). Ferulic acid in cereals — A review. *Czech Journal of Food Sciences*, 33(1), 1–7. <https://doi.org/10.17221/401/2014-CJFS>
- Skrajda-Brdak, M., Konopka, I., Tańska, M., Czaplicki, S. (2019). Changes in the content of free phenolic acids and antioxidative capacity of wholemeal bread in relation to cereal species and fermentation type. *European Food Research and Technology*, 245, 2247–2256. <https://doi.org/10.1007/s00217-019-03331-y>
- Abdel-Aal, E.-S.M., Rabalski, I. (2022). Changes in phenolic acids and antioxidant properties during baking of bread and muffin made from blends of hairless canary seed, wheat, and corn. *Antioxidants*, 11(6), Article 1059. <https://doi.org/10.3390/antiox11061059>
- Dall'Asta, M., Bresciani, L., Calani, L., Cossu, M., Martini, D., Melegari, C. et al. (2016). *In vitro* bioaccessibility of phenolic acids from a commercial aleurone-enriched bread compared to a whole grain bread. *Nutrients*, 8(1), Article 42. <https://doi.org/10.3390/nu8010042>
- Arya, S.S., Rookes, J.E., Cahill, D.M., Lenka, S.K. (2021). Vanillin: A review on the therapeutic prospects of a popular flavouring molecule. *Advances in Traditional Medicine*, 21(3), 1–17. <https://doi.org/10.1007/s13596-020-00531-w>
- Viretto, C., Tlais, A.Z.A., Arora, K., Ameer, H., Tuccillo, F., Polo, A. et al. (2025). Unraveling the functional potential of microbial resources and pulse-based matrices for sourdough breadmaking. *Future Foods*, 11, Article 100643. <https://doi.org/10.1016/j.fufo.2025.100643>
- Sanmartín, G., Sánchez-Adriá, I.E., Salvador, A., Prieto, J.A., Estruch, F., Rande-Gil, F. (2024). Quantitative assessment of volatile profile and sensory perception of artisan bread made in the city of Valencia. *Foods*, 13(23), Article 3872. <https://doi.org/10.3390/foods13233872>
- Huang, C., Huang, J., Zhang, B., Omedi, J.O., Chen, C., Zhou, L. et al. (2023). Rheo-fermentation dough properties, bread-making quality and aroma characteristics of red bean (*Vigna angularis*) sourdough induced by LAB *Weissella confusa* Q5813 strain fermentation. *Foods*, 12(3), Article 605. <https://doi.org/10.3390/foods12030605>
- Mărgăoan, R., Cornea-Cipcigan, M., Topal, E., Kösoğlu, M. (2020). Impact of fermentation processes on the bioactive profile and health-promoting properties of bee bread, mead and honey vinegar. *Processes*, 8(9), Article 1081. <https://doi.org/10.3390/pr8091081>
- Aghababaei, F., Hadidi, M. (2023). Recent advances in potential health benefits of quercetin. *Pharmaceuticals*, 16(7), Article 1020. <https://doi.org/10.3390/ph16071020>
- Suchowilska, E., Biełkowska, T., Stuper-Szablewska, K., Wiwart, M. (2020). Concentrations of phenolic acids, flavonoids and carotenoids and the antioxidant activity of the grain, flour and bran of *Triticum polonicum* as compared with three cultivated wheat species. *Agriculture*, 10(12), Article 591. <https://doi.org/10.3390/agriculture10120591>
- Wang, M., Firman, J., Liu, L., Yam, K. (2019). A review on flavonoid apigenin: Dietary intake, ADME, antimicrobial effects, and interactions with human gut microbiota. *BioMed Research International*, 2019, Article 7010467. <https://doi.org/10.1155/2019/7010467>
- Kim, S.H., Park, J.G., Lee, J., Yang, W.S., Park, G.W., Kim, H.G. et al. (2015). The dietary flavonoid kaempferol mediates anti-inflammatory responses via the Src, Syk, IRAK1, and IRAK4 molecular targets. *Mediators of Inflammation*, 2015, Article 904142. <https://doi.org/10.1155/2015/904142>
- Ruiz-Álvarez, B.E., Padilla-de la Rosa, J.D., González Avila, M.G., Sandoval, G., Desjardins, Y. (2025). Novel acylated naringin enhances propionate release and stimulates the growth of flavanone-metabolizing bacteria in an *in vitro* batch fermentation model. *Life*, 15(6), Article 967. <https://doi.org/10.3390/life15060967>
- Tian, W., Chen, G., Tilley, M., Li, Y. (2021). Changes in phenolic profiles and antioxidant activities during the whole wheat bread-making process. *Food Chemistry*, 345, Article 128851. <https://doi.org/10.1016/j.foodchem.2020.128851>
- Yu, L., Beta, T. (2015). Identification and antioxidant properties of phenolic compounds during production of bread from purple wheat grains. *Molecules*, 20(9), 15525–15549. <https://doi.org/10.3390/molecules200915525>
- Czajkowska-González, Y.A., Alvarez-Parrilla, E., Martínez-Ruiz, N.R., Vázquez-Flores, A.A., Gaytán-Martínez, M., de la Rosa, L.A. (2021). Addition of phenolic compounds to bread: Antioxidant benefits and impact on food structure and sensory characteristics. *Food Production, Processing and Nutrition*, 3, Article 25. <https://doi.org/10.1186/s43014-021-00068-8>
- Liu, Y., Shi, L., Qiu, W., Shi, Y. (2022). Ferulic acid exhibits anti-inflammatory effects by inducing autophagy and blocking NLRP3 inflammasome activation. *Molecular and Cellular Toxicology*, 18(4), 509–519. <https://doi.org/10.1007/s13273-021-00219-5>
- Goya-Jorge, E., Jorge Rodríguez, M.E., Sylla-Iyarreta Veitia, M., Giner, R.M. (2021). Plant occurring flavonoids as modulators of the aryl hydrocarbon receptor. *Molecules*, 26(8), Article 2315. <https://doi.org/10.3390/molecules26082315>
- Jin, U.-H., Park, H., Li, X., Davidson, L.A., Allred, C., Patil, B. et al. (2018). Structure-dependent modulation of aryl hydrocarbon receptor-mediated activities by flavonoids. *Toxicological Sciences*, 164(1), 205–217. <https://doi.org/10.1093/toxsci/kfy075>
- Zhang, S., Qin, C., Safe, S.H. (2003). Flavonoids as aryl hydrocarbon receptor agonists/antagonists: Effects of structure and cell context. *Environmental Health Perspectives*, 111(16), 1877–1882. <https://doi.org/10.1289/ehp.6322>
- Zdrzil, B., Felix, E., Hunter, F., Manners, E.J., Blackshaw, J., Corbett, S. et al. (2024). The ChEMBL Database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1), D1180–D1192. <https://doi.org/10.1093/nar/gkad1004>
- Rogers, D., Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5), 742–754. <https://doi.org/10.1021/ci100050t>
- Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J.K., Ceulemans, H. et al. (2018). Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical Science*, 9(24), 5441–5451. <https://doi.org/10.1039/c8sc00148k>
- Varma, S., Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, Article 91. <https://doi.org/10.1186/1471-2105-7-91>
- Westerman, K.E., Harrington, S., Ordovas, J.M., Parnell, L.D. (2020). PhyteByte: Identification of foods containing compounds with specific pharmacological properties. *BMC Bioinformatics*, 21, Article 238. <https://doi.org/10.1186/s12859-020-03582-7>
- Reker, D., Shi, Y., Kirtane, A.R., Hess, K., Zhong, G.J., Crane, E. et al. (2020). Machine learning uncovers food- and excipient-drug interactions. *Cell Reports*, 30(11), 3710–3716. <https://doi.org/10.1016/j.celrep.2020.02.094>
- Sanchez-Ruiz, A., Colmenarejo, G. (2022). Systematic Analysis and Prediction of the Target Space of Bioactive Food Compounds: Filling the Chemobiological Gaps. *Journal of Chemical Information and Modeling*, 62(16), 3734–3751. <https://doi.org/10.1021/acs.jcim.2c00888>
- Periwal, V., Bassler, S., Andrejev, S., Gabrielli, N., Patil, K.R., Typas, A. et al. (2022). Bioactivity assessment of natural compounds using machine learning models trained on target similarity between drugs. *PLoS Computational Biology*, 18(4), Article e1010029. <https://doi.org/10.1371/journal.pcbi.1010029>
- Kuznetsov, M., Chernyavskaya, O., Kutuzov, M., Vilkova, D., Novichenko, O., Stolyarova, A. et al. (2025). Development of a predictive model for the biological activity of food and microbial metabolites toward estrogen receptor alpha ($\text{ER}\alpha$) using machine learning. *Big Data and Cognitive Computing*, 9(4), Article 86. <https://doi.org/10.3390/bdcc9040086>
- Kuznetsov, M., Klein, E., Velina, D., Mutallibzoda, S., Orlovtsseva, O., Tefikova, S. et al. (2025). Using Machine learning for the discovery and development of multitarget flavonoid-based functional products in MASLD. *Molecules*, 30(21), Article 4159. <https://doi.org/10.3390/molecules30214159>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>

СВЕДЕНИЯ ОБ АВТОРАХ	AUTHOR INFORMATION
Принадлежность к организации	Affiliation
Кузнецов Максим Александрович — аспирант, кафедра пищевых технологий и биоинженерии, Российский экономический университет имени Г. В. Плеханова 115054, Москва, Стремянный пер, 36 E-mail: 5726219@gmail.com ORCID: https://orcid.org/0009-0003-2652-0303	Maksim A. Kuznetsov, PhD student, Research Laboratory «Biotechnology of Food Systems», Plekhanov Russian University of Economics 36, Stremyanny lane, 115054, Moscow, Russia E-mail: 5726219@gmail.com ORCID: https://orcid.org/0009-0003-2652-0303
Никитин Игорь Алексеевич — доктор технических наук, доцент, ведущий кафедрой пищевых технологий и биоинженерии, Российский экономический университет имени Г. В. Плеханова 115054, Москва, Стремянный пер, 36 E-mail: Nikitin.IA@rea.ru ORCID: https://orcid.org/0000-0002-8988-5911 * автор для контактов	Igor A. Nikitin, Doctor of Technical Sciences, Docent, Head of the Department of Food Technology and Bioengineering, Plekhanov Russian University of Economics 36, Stremyanny lane, 115054, Moscow, Russia E-mail: Nikitin.IA@rea.ru ORCID: https://orcid.org/0000-0002-8988-5911 * corresponding author
Критерии авторства	Contribution
Авторы в равных долях имеют отношение к написанию рукописи и одинаково несут ответственность за плагиат.	Authors equally relevant to the writing of the manuscript, and equally responsible for plagiarism.
Конфликт интересов	Conflict of interest
Авторы заявляют об отсутствии конфликта интересов.	The authors declare that there is no conflict of interest.