Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»

На правах рукописи

Веселов Марк Сергеевич

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РАЗРАБОТКИ НОВЫХ МОЛЕКУЛ С АНТИБАКТЕРИАЛЬНОЙ АКТИВНОСТЬЮ

Специальность: 03.01.02 – Биофизика

АВТОРЕФЕРАТ
диссертация на соискание ученой степени
кандидата биологических наук

Москва – 2019

Работа прошла апробацию на кафедре инновационной фармацевтики, медицинской техники и биотехнологии федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)»

Научный руководитель: Иваненков Ян Андреевич, кандидат биологических наук, заведующий лабораторией медицинской химии и биоинформатики федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)».

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования «Казанский (Приволжский) федеральный университет»

Защита состоится «24» декабря 2019 года в 10.00 на заседании диссертационного совета ФБМФ03.01.02.005, по адресу: 141701, г. Долгопрудный Московской обл., Институтский переулок, д. 9.

С диссертацией можно ознакомиться в библиотеке и на сайте Московского физико-технического института (национальный исследовательский университет) https://mipt.ru/education/post-graduate/soiskateli-biologicheskie-nauki.php.

Работа представлена «3» октября 2019 г. в Аттестационную комиссию федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)» для рассмотрения советом по защите диссертаций на соискание ученой степени кандидата наук в соответствии с п .3.1 ст . 4 Федерального закона «О науке и государственной научно - технической политике»

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. В последние годы у большого числа патогенов возрастает устойчивость к антибактериальным препаратам, что представляет угрозу для здравоохранения, увеличивается количество инфекций с множественной лекарственной устойчивостью. За последние два десятилетия значительно увеличилась устойчивость патогенных микроорганизмов к терапии первой линии[1]. Кроме того, наблюдается рост числа штаммов, устойчивых к препаратам второй и третьей линии терапии. Все это приводит к серьезным клиническим последствиям, неудачному лечению заболеваний, высокой смертности и длительным госпитализациям, а также увеличению расходов на здравоохранение.

Многие крупные фармацевтические компании не инвестируют ресурсы в разработку новых антибактериальных молекул по ряду причин, одна из которых – слишком малая вероятность получить положительный результат в ходе клинических испытаний. Таким образом, ключевую роль в исследованиях по поиску новых антибактериальных препаратов начинают играть небольшие фармацевтические компании и академические учреждения. Повышение эффективности программ по поиску первоначальных активных молекул для дальнейшей оптимизации является важной задачей, и ее решение возможно с привлечением современных методов компьютерного моделирования. Тем не менее, доступные на данный момент компьютерные модели имеют множество недостатков и не применимы для прогнозирования антибактериальной активности молекул с высоким структурным разнообразием. С учетом этого главная цель настоящего исследования состояла в том, чтобы разработать эффективную компьютерную модель, которая бы учла недостатки уже опубликованных моделей и обладала бы достаточной прогностической способностью в отношении антибактериальной активности.

Степень разработанности темы. За последнее время было опубликовано большое количество работ, посвященных применению методов компьютерного моделирования для прогнозирования антибактериальной активности малых молекул. Однако подавляющая часть из них сфокусирована на каком-либо одном химическом классе (хемотипе) молекул. Как правило, такие модели не применимы к библиотекам с высоким структурным разнообразием, поскольку диапазон значений признаков, описывающих обучающую выборку, очень узок и характеризует только конкретно заданный хемотип. Тем не менее, был опубликован и ряд работ с моделями, построенными с использованием выборок, включающих несколько разных хемотипов (химических классов). Однако, эти выборки были

[1] Fowler T., Walker D., Davies S.C. The risk/benefit of predicting a post-antibiotic era: Is the alarm working? // Annals of the New York Academy of Sciences. 2014. Т. 1323. № 1. С. 1–10.

подготовлены на основе ограниченных и недостаточно представительных источников данных[2]. В недавних работах были описаны модели, обученные на более качественных выборках[3,4]. К недостаткам опубликованных моделей можно также отнести то, что в большинстве случаев их прогностическая способность не была оценена с помощью кросс-валидации или независимой тестовой выборки с высоким структурным разнообразием[5]. И только небольшая часть этих моделей была протестирована в экспериментальных условиях (биологическое тестирование), в результате чего были обнаружены новые молекулы, обладающие антибактериальной активностью. В настоящей работе большое внимание было уделено вышеупомянутым недостаткам: подготовке разнообразной обучающей выборки, хорошо покрывающей химическое пространство, изучению различных архитектур моделей и их экспериментальной валидации.

Целью работы является разработка компьютерной модели для прогнозирования антибактериальной активности малых молекул с высоким структурным разнообразием с применением методов машинного обучения. Для достижения этой цели ставились следующие **задачи**:

1. Создание базы данных известных активных молекул и соединений, которые в ходе биологических тестов не продемонстрировали антибактериальную активность, для дальнейшего учета при отборе молекул.
2. Отбор молекул, доступных в коммерческих коллекциях, на первый этап высокопроизводительного скрининга (ВПС) с использованием метода, позволяющего отбирать молекулы с высоким структурным разнообразием.
3. Анализ полученных в ходе биологического скрининга результатов, описание и характеристика химического пространства и подготовка представительной обучающей выборки. Расчет и отбор молекулярных дескрипторов.
4. Тестирование различных алгоритмов машинного обучения и разработка оптимальной прогностической модели.

---

[2] Yang X.-G. et al. Prediction of antibacterial compounds by machine learning approaches // J. Comput. Chem. 2009. Vol. 30, № 8. P. 1202–1211.

[3] Masalha M. et al. Capturing antibacterial natural products with in silico techniques // Mol. Med. Rep. 2018. Vol. 18, № 1. P. 763–770.

[4] Wang L. и др. Discovering new agents active against methicillin-resistant Staphylococcus aureus with ligand-based approaches // J. Chem. Inf. Model. 2014. Т. 54. № 11. С. 3186–3197.

[5] Durrant J.D., Amaro R.E. Machine-learning techniques applied to antibacterial drug discovery // Chem. Biol. Drug Des. 2015. Vol. 85, № 1. P. 14–21.

5. Валидация построенной модели с использованием независимой тестовой выборки молекул, анализ полученных в ходе биологического тестирования результатов, оценка прогностической способности модели.

Научная новизна. В ходе исследования был предложен и реализован метод рационального отбора малых органических молекул с высоким структурным разнообразием, позволяющий сохранить покрытие химического пространства. С помощью этого метода был отобран и протестирован большой набор молекул на предмет их антибактериальной активности. На основе результатов скрининга была сформирована обучающая выборка, состоящая из более чем 74 тыс. малых молекул с экспериментально определенной антибактериальной активностью по отношению к штамму *E. coli* (ΔtolC) в единых условиях. Важно отметить, что аналогов такого представительного обучающего набора молекул в научной литературе не описано. Впервые был проведен анализ привилегированных подструктур, встречающихся в активных и неактивных молекулах, и обнаружены значимые закономерности. Впервые для решения задачи прогнозирования антибактериальной активности был применен метод генеративного топографического картирования. Многие найденные с помощью модели молекулы показали высокую антибактериальную активность и могут рассматриваться в качестве перспективных соединений для дальнейшей оптимизации. Некоторые обнаруженные молекулы ингибируют трансляцию и обладают низкой цитотоксичностью (CC_{50}) по отношению к панели эукариотических клеточных линий, обеспечивая тем самым высокий индекс селективности (ИС). На основе предварительного патентного исследования, ряд молекул можно отнести к классу патентоспособных.

Научно-практическая значимость. Разработанный метод рационального отбора потенциально активных и обладающих высоким структурным разнообразием молекул, позволяет снизить общее количество молекул для биологического тестирования, при этом сохраняя хорошее покрытие химического пространства. Разработанная компьютерная модель позволяет прогнозировать антибактериальную активность соединений (хит-рейт 24% на независимой тестовой выборке с высоким структурным разнообразием). В ходе экспериментов были обнаружены ранее не описанные молекулы с высокой антибактериальной активностью, в том числе соединение, активное в отношении клинически значимых штаммов (5'-[(4-бромбензоил)амино]-2,3'-битиофен-4'-карбоновая кислота).

Результаты диссертационного исследования внесли значительный вклад в работу по гранту РНФ №17-74-30012 «Новый рациональный подход к разработке антибактериальных и противоопухолевых лекарственных молекул с применением технологии высокопроизводительного скрининга»

Положения, выносимые на защиту.

1. Наибольший вклад в разделение активных и неактивных молекул вносят следующие химические группы: карбоксильная группа, α,β-ненасыщенные карбонилы и аллилы, имидазол, хинолин и бензимидазол (характерны для активных молекул); фуран, пиперазин, пропаноильная группа и бензодиоксольный фрагмент (характерны для неактивных молекул).
2. Наибольший вклад в разделение активных и неактивных молекул вносят следующие молекулярные дескрипторы: HBD (количество потенциальных доноров водородной связи), Hy (индекс гидрофильности), RB (число свободно вращающихся связей), logS (логарифм растворимости в воде), и др.
3. Наилучшую точность для прогнозирования антибактериальной активности на обучающей выборке в проведенном эксперименте показал алгоритм градиентного бустинга.
4. При валидации прогностической способности на независимой тестовой выборке с высоким структурным разнообразием хит-рейт эксперимента составил 24%, что гораздо выше по сравнению с показателем в 2% при рандомном скрининге.
5. Обнаруженные с помощью модели хемотипы соединений обладают активностью, сравнимой с известными лекарствами (левофлоксацин и эритромицин). Одно из соединений (соединение **1** – 5'-[(4-бромбензоил)амино]-2,3'-битиофен-4'-карбоновая кислота) показало активность в отношении клинически значимого штамма *S. aureus*.

Личный вклад автора и апробация работы.

Автором были проведены работы по сбору данных и референсных баз молекул, их подготовке и обработке. Были реализованы оригинальные программные модули для отбора молекул и модели для прогнозирования активности. Был проведен анализ данных результатов биологического тестирования.

Результаты, полученные в ходе проделанной работы, были представлены на всероссийских и международных конференциях: XXVIII Зимняя молодежная научная школа "Перспективные направления физико-химической биологии и биотехнологии", Институт Биоорганической Химии РАН, Россия, 8-11 февраля 2016; Международная конференция Chemical Biology 2016, EMBL Heidelberg, Германия, 31 августа - 3 сентября 2016; V Съезд физиологов СНГ, V Съезд Биохимиков России, Сочи, Россия, 4-8 октября 2016; Международная конференция FEBS 2018, Прага, Чехия, 7-12 июля 2018 (итого – 4 научных конференции).

Публикации. По теме диссертации опубликовано 7 печатных работ, среди них: 6 в журналах, рекомендованных ВАК, 1 в тезисах международных научных конференций, 6 в международных журналах, индексируемых в базах данных Scopus и WoS.

Структура и объем диссертации. Работа состоит из введения, трех глав, заключения, выводов, списка сокращений, и списка литературы. Общее количество страниц: 113. Работа содержит 17 иллюстраций и 18 таблиц; список литературы включает 129 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении дается краткая характеристика работы, ее актуальность, научная новизна и практическая значимость. Рассмотрены цели и задачи работы, а также положения, выносимые на защиту.

В первой главе приведен краткий обзор основных классов антибактериальных веществ, их механизмов действия и видов лекарственной устойчивости микроорганизмов. Также рассматриваются актуальные проблемы в сфере медицинской химии антибактериальных молекул. Приведен обзор опубликованных компьютерных моделей для прогнозирования антибактериальной активности, обозначены их основные недостатки.

Во второй главе описана работа с базами данных (сбор, обработка, очистка и подготовка данных), расчет молекулярных дескрипторов, отбор соединений на стадию ВПС, методы компьютерного моделирования и методы биологического тестирования.

Для того, чтобы не проводить тестирование молекул, для которых уже была описана антибактериальная активность, был проведен анализ баз данных ChEMBL и Thompson Integrity Database. На основе этого анализа была собрана выборка активных (>30 тыс. молекул) и протестированных, но неактивных молекул (>200 тыс.), которые были впоследствии исключены из рассмотрения. Отбор молекул на стадию ВПС осуществлялся с использованием доступных коллекций органических соединений, в частности компаний ХимРар и InterBioScreen (IBS). Для этого была подготовлена общая база данных, содержащая 2.2 млн. структур. В то время как коллекция ХимРар в основном содержит органические молекулы синтетического происхождения, коллекция IBS содержит большое количество природных молекул и их близких аналогов. Для отбора молекул на стадию ВПС из этих коллекций был предложен метод, позволяющий сохранить покрытие химического пространства

молекул с учетом их 3D формы (для расчета дескрипторов формы молекулы использовался метод USR – *ultrafast shape similarity*).

Биологическое тестирование на стадии ВПС проводилось с использованием уникальной платформы, описанной ранее в ряде научных публикаций[6]. Эта платформа позволяет не только измерить антибактериальную активность молекул, но и выяснить их механизм действия (ингибирование трансляции, повреждение ДНК или какой-либо другой). Антибактериальная активность была предварительно оценена тщательным визуальным анализом и измерением области ингибирования роста бактерий: 0-4 мм («-»), 4-7 мм («+/-»), 7-11 мм («+»), 11-16 мм («++»), 16-20 мм («+++»), 20-25 мм («++++»). При дальнейшем формировании обучающей выборки, соединения с незначительной зоной ингибирования роста («-», «+/-» и «+») были отнесены к классу неактивных, поскольку на этой стадии использовались относительно высокие концентрации соединений. Молекулы, которые вызывали сильное ингибирование роста бактерий («++», «+++», «++++»), были классифицированы как активные. Для наиболее активных молекул проводили изучение ингибирования трансляционной активности *in vivo* (с использованием штамма *E.coli* ΔtolC) и *in vitro* (в бесклеточной системе на основе изолированных рибосом (S30 из *E. coli*) и мРНК люциферазы светлячка).

Обучающая выборка была сформирована на основе базы из 140 тыс. соединений, состоящей из результатов биологического тестирования, и дополненной активными молекулами из базы Thomson Reuters Integrity (12 тыс. структур). Структуры, которые не соответствуют наиболее общим критериям лекарственного подобия, были удалены. Затем база данных была кластеризована с использованием программы ChemoSoft: порог 2D подобия Танимото ≥0.5, количество структур в кластере ≥5. С целью увеличения общего разнообразия выборки и уменьшения количества похожих хемотипов, из каждого кластера отобрали по 30 структур с максимальным коэффициентом разнообразия, а также оставили уникальные соединения, не вошедшие ни в один кластер. В результате финальная база данных содержала 74567 структур (8724 активных и 65843 неактивных молекул). Для того, чтобы иллюстрировать покрытие химического пространства сформированным набором соединений, из каждой выборки было отобрано случайным образом по 1000 молекул, и для этих соединений было сделано отображение с помощью метода нелинейного снижения размерности и визуализации многомерных переменных (t-*distributed stochastic neighbor embedding*, t-SNE). Из рисунка 1 видно, что выборка равномерно покрывает коммерчески доступное химическое пространство.

[6] Osterman I.A. и др. Sorting Out Antibiotics' Mechanisms of Action: a Double Fluorescent Protein Reporter for High-Throughput Screening of Ribosome and DNA Biosynthesis Inhibitors // Antimicrob. Agents Chemother. 2016. Т. 60. № 12. С. 7481–7489.

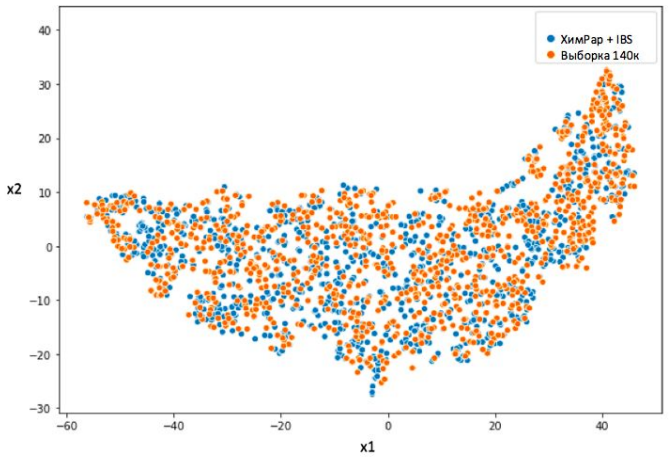


Рисунок 1. Визуализация результатов, полученных с применением алгоритма t-SNE: из каждого подмножества было выбрано 1000 молекул случайным образом.

Молекулярные дескрипторы (всего: 1749) были рассчитаны для всего набора обучающих данных с использованием программных пакетов Dragon, ChemoSoft, MOE и SmartMining. Количество дескрипторов было уменьшено до 1243 за счет удаления постоянных, почти постоянных и высокоскоррелированных (R^2>0.9) дескрипторов. Для всех дескрипторов был рассчитан t-критерий Стьюдента, и на основе него был отобран финальный список из 40 дескрипторов.

Обучающий набор данных был разделен на выборку для кросс-валидации (80% от общего набора) и отложенную тестовую выборку (20% от общего набора). Кросс-валидация (с разделением на 4 части) использовалась, чтобы избежать переобучения модели, а отложенная выборка использовалась для предварительной оценки прогностической способности разработанных моделей.

В ходе экспериментов были исследованы следующие методы машинного обучения: градиентный бустинг, метод k-ближайших соседей, случайный лес, нейронная сеть прямого распространения, логистическая регрессия и генеративное топографическое картирование. Для построения моделей применялись алгоритмы, реализованные в библиотеках на языке *Python*: *xgboost*, *sklearn*, *ugtm*.

В третьей главе описаны основные результаты, которые были получены в ходе компьютерного моделирования и экспериментальной валидации.

В ходе ВПС было протестировано 139152 молекулы, из которых 106143 – из коллекции компании ХимРар и 33509 – из коллекции компании IBS. С помощью ранее описанной репортерной системы для этих соединений, помимо активности в отношении *E. coli*, был также определен один из трех механизмов действия – ингибирование трансляции, синтеза ДНК, либо другой механизм действия, не связанный с двумя предыдущими. По результатам первого раунда скрининга был обнаружен ряд

соединений, активных в отношении *E. coli* (ΔTolC). Из 139152 молекул, 2024 показали заметный эффект (больше "+").

Химическое пространство протестированных молекул было исследовано на предмет наличия привилегированных подструктур, различающих активные и неактивные соединения. Среди категории негетероциклических фрагментов метокси- (30.5% и 35% для активных и неактивных соединений, соответственно) и карбонильная группа (39% и 25%) являются наиболее представленными. Неактивные соединения содержат в 1.56 раза больше карбонильных фрагментов в отличие от активных, в то время как метоксигруппа не обеспечивает значимого разделения между двумя рассматриваемыми классами. Частота встречаемости пропаноильного фрагмента среди неактивных молекул в 3 раза выше, чем у активных. Карбоксильные, α,β-ненасыщенные карбонилы и аллилы являются наиболее характерными подструктурами для антибактериальных соединений: их частота в 3.75, 6 и 9 раз выше по сравнению с неактивным классом. Среди гетероциклических фрагментов индол является наиболее представленным (12%) в антибактериальных соединениях. Доля фрагментов имидазола, хинолина и бензимидазола значительно смещена в сторону антибактериальных соединений, в то время как фуран и пиперазин (~7%) в 2.3 раза чаще встречается в неактивных молекулах. Кроме того, 1,3-бензодиоксольный фрагмент является предпочтительным для неактивных молекул, в то время как изоксазол одинаково представлен в обоих классах.

На основе данных биологического тестирования была подготовлена обучающая выборка из 74567 тыс. структур, для которых были рассчитаны и отобраны молекулярные дескрипторы (на основе t-критерия Стьюдента). Следует отметить, что молекулярные дескрипторы, отобранные для процедуры обучения, отражают статистические наблюдения, приведенные выше, и тесно связаны с важными физико-химическими свойствами молекул. Например, общая полярность, представленная как S(-OH), S(-O-), S(=N-), S(>N-), HB2, a_acc, O-057/061, PEOE_VSA_FPOS и TPSA, соответствует метокси-, карбонил-, пропаноил-, карбоксил-, α,β-ненасыщенным карбонильным группам и гетероциклам. Дескрипторы Hy и SlogP_VSA0 отражают липофильность молекулы, особенно в случае линейных и разветвленных алкильных фрагментов, а также ароматических фрагментов. Топология молекулярной структуры описывается, например, дескрипторами M1, SPI, EEig07x, Q', VEA2 и GATS1m.

Для первоначальной оценки прогностической способности моделей были построены базовые классификаторы со стандартными параметрами: модель k-ближайших соседей, градиентного бустинга на решающих деревьях, случайного леса, логистической регрессии, нейронной сети и генеративного топографического картирования. Обучение и оценка классифицирующей способности моделей производилась на обучающей выборке с помощью кросс-валидации (с разделением на 4 части), а также

на отложенной выборке. В ходе экспериментов для всех шести моделей были рассчитаны значения *ROC AUC* (площадь под кривой *ROC*), *F1* (гармоническое среднее между точностью и полнотой), *Precision* (точность) и *Recall* (полнота) для активных молекул. Наилучшие показатели метрики *Precision* получились у градиентного бустинга (*Precision* = 0.846) и случайного леса (*Precision* = 0.866). Этот показатель важен в том случае, когда необходимо максимизировать долю активных молекул при биологическом тестировании. При этом, стоит также учитывать показатель *Recall*, значение которого отражает долю найденных активных соединений из тех, которые в принципе можно было найти. Из базовых моделей, самый высокий показатель этой метрики получился у генеративного топографического картирования (*Recall* = 0.678). Далее были проведены вычислительные эксперименты по поиску оптимальных гиперпараметров для каждой из моделей. Поиск производился с помощью модуля *RandomizedSearchCV* (из библиотеки *sklearn*) на заранее заданном множестве гиперпараметров, с максимизацией метрики *F1*. Результаты представлены в таблице 1.

Таблица 1. Результаты обучения моделей после подбора гиперпараметров

Градиентный бустинг				
Выборка	*ROC AUC*	*F1*	*Precision*	*Recall*
Кросс-валидация	0.934	0.928	0.970	0.891
Отложенная выборка	0.928	0.789	0.945	0.834
Метод ближайших соседей				
Выборка	*ROC AUC*	*F1*	*Precision*	*Recall*
Кросс-валидация	0.774	0.549	0.596	0.510
Отложенная выборка	0.761	0.6	0.641	0.564
Случайный лес				
Выборка	*ROC AUC*	*F1*	*Precision*	*Recall*
Кросс-валидация	0.929	0.646	0.920	0.498
Отложенная выборка	0.754	0.657	0.91	0.514
Логистическая регрессия				
Выборка	*ROC AUC*	*F1*	*Precision*	*Recall*
Кросс-валидация	0.841	0.391	0.703	0.271
Отложенная	0.628	0.391	0.707	0.27

выборка				
Нейронная сеть				
Выборка	*ROC AUC*	*F1*	*Precision*	*Recall*
Кросс-валидация	0.854	0.597	0.621	0.574
Отложенная выборка	0.773	0.625	0.675	0.583
Генеративное топографическое картирование*				
Выборка	*ROC AUC*	*F1*	*Precision*	*Recall*
Отложенная выборка	0.763	0.774	0.781	0.753
Независимая тестовая выборка	0.751	0.753	0.764	0.742

*обучение проводилось на выборке со сбалансированными классами

Оценка прогностической способности построенных моделей была проведена с использованием независимой тестовой выборки из 5000 малых органических молекул, обладающих сравнительно низким структурным подобием – менее 0.5 (коэффициент Танимото) по отношению к обучающим примерам. Эти молекулы были также предоставлены компаниями ХимРар и IBS. Антибактериальная активность соединений была спрогнозирована с использованием разработанных моделей, а затем оценена в соответствии с биологическими протоколами, описанными ранее. На основе полученных результатов были рассчитаны метрики *ROC AUC*, *F1*, *Precision* и *Recall* для независимого тестового набора молекул (таблица 2). Это позволило оценить способность модели прогнозировать активность для молекул, отличающихся от тех, которые представлены в обучающей выборке.

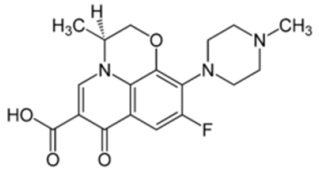
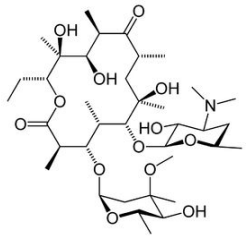
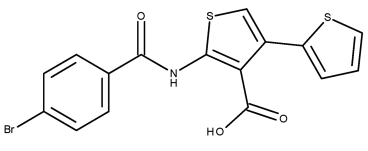
Таблица 2. Результаты валидации модели с применением независимого тестового набора молекул

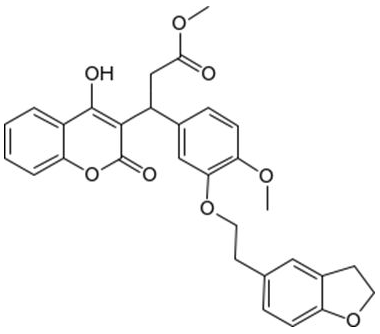
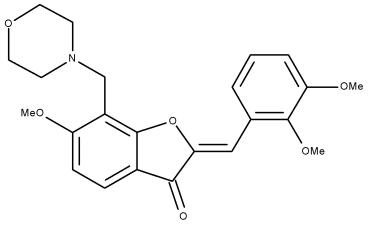
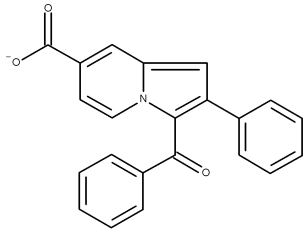
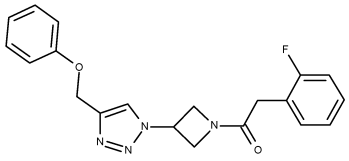
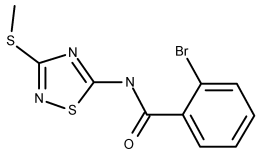
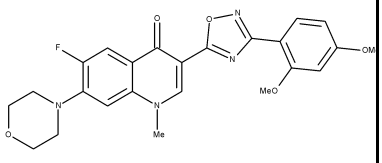
Градиентный бустинг				
Выборка	*ROC AUC*	*F1*	*Precision*	*Recall*
Отложенная выборка	0.934	0.928	0.970	0.891
Независимая тестовая выборка	0.651	0.362	0.243	0.791

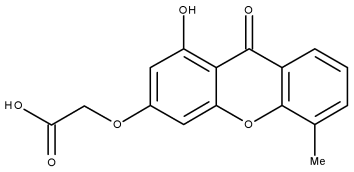
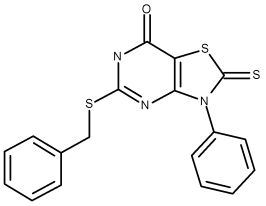
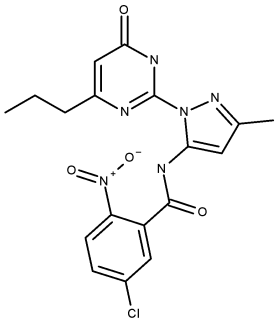
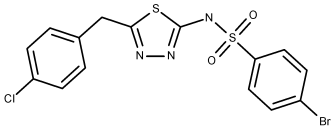
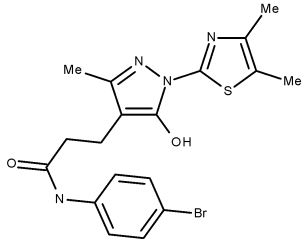
*для построения модели использовались следующие параметры: *booster* = 'gbtree', *colsample_bytree* = 0.817, *gamma* = 0.463, *learning_rate* = 0.233, *max_delta_step* = 0, *max_depth* = 8, *min_child_weight* = 2, *n_estimators* = 950, *objective* = 'binary:logistic', *subsample* = 0.731

В ходе биологического тестирования было обнаружено 105 активных соединений (доля найденных активных молекул в общем эксперименте = 2.1%). Из них верно спрогнозированными алгоритмом активными молекулами оказались 24%. Следует особо отметить, что среди всех активных соединений, обнаруженных в ходе первого раунда ВПС (молекулы, включенные в обучающую выборку) и из независимой тестовой выборки, только несколько соединений продемонстрировали значительную ингибирующую активность в отношении *E. coli* (дикий тип). Несколько молекул вызывали устойчивый SOS-ответ или ингибирование трансляции. Некоторые соединения индуцировали оба сигнала, но с относительно низкой интенсивностью. Наиболее активные молекулы, найденные в результате биологического тестирования представлены в таблице 3.

Таблица 3. Примеры активных молекул, для которых была верно спрогнозирована антибактериальная активность

ID	**Структура**	**Активность**	**МИК (мкг/мл, ΔtolC)**	**Механизм действия**	***In vitro* трансляция**
LVX		++++	0.016±0.009	SOS	-
ERY		++++	2.5±0.5	T	+
1		+++	1.8±0.8	T	+

2		+++	2±0.4	T	+
3		++++	3.9±1.4	S+T	+
4		+++	6.25±1.3	T	+
5		++	12.5±1.9	T	±
6		+	42±5	T	+
7		+++	0.8	SOS	-

8		+++	20.8	SOS	-
9		+++	<0.2	O	-
10		+++	<0.2	O	-
11		+++	0.8	O	-
12		+++	0.8	O	-

* LVX – левофлоксацин; ERY – эритромицин; МИК – минимальная ингибирующая концентрация; ИС = индекс селективности = CC_{50}(мкг/мл или %)/МИК(мкг/мл): H – высокий, ИС>100; M – средний, 20<ИС<100; L – низкий, ИС<20; T – ингибирование трансляции, SOS – SOS-ответ, O – другой механизм действия.

Как показано в таблице 3, среди представленных молекул наибольшая антибактериальная активность была выявлена для аналога фторхинолона **7** (МИК=0.8 мкг/мл), 6*H*-тиазоло[4,5-d]пиримидинона **9** (МИК<0.2 мкг/мл), (6-оксо-1*H*-пиримидин-2-ил)пиразола **10**

(МИК<0.2 мкг/мл), замещенного тиадиазола **11** (МИК=0.8 мкг/мл) и гидроксипиразола **12** (МИК=0.8 мкг/мл). Соединения **1** и **2** сильно ингибировали трансляцию при 16 мкг/мл и показали хороший индекс селективности. Кроме того, соединение **2** продемонстрировало антибактериальную активность в отношении нескольких мутантных штаммов. Два соединения **7** и **8** вызывали значительный SOS-ответ (МИК=0.8 и 20.8 мкг/мл соответственно), однако соединение **8** показало более низкий индекс селективности. Среди молекул, действующих по иным механизмам, соединение **11** можно отнести к широкому классу ингибиторов дигидроптероат-синтетазы на основе сульфаниламидов. Патентоспособность молекул оценивали с использованием баз данных SciFinder и Integrity Database.

Соединение **1** было изучено на предмет активности по отношению к клинически значимым штаммам: *E. coli* (ATCC 25922), *K. pneumoniae* (181210171-2), *P. aeruginosa* (ATCC 27853), *S. aureus* (ATCC USA 206) и *C. albicans* (181210169-1) (таблица 4). Молекула продемонстрировала умеренную активность в отношении грамотрицательных бактерий *K. pneumoniae* и незначительно ингибировала рост бактерий кишечной палочки. Аналогичный эффект наблюдался в случае мультирезистентного штамма *C. albicans*. Активности по отношению к *K. pneumoniae* обнаружено не было. Высокая антибактериальная активность соединения **1** была выявлена в тестах на грамположительных штаммах золотистого стафилококка. Зона ингибирования роста бактерий превышала 20 мм.

Таблица 4. Антибактериальная активность соединения **1** в отношении выбранных клинически значимых бактериальных штаммов

Вид	ID штамма	Коллекция	Активность
Escherichia coli	ATCC 25922	ATCC*	±
Klebsiella pneumoniae	181210171-2	Клиника БГМУ	+
Pseudomonas aeruginosa	ATCC 27853	ATCC	-
Staphylococcus aureus	ATCC USA 206	Клиника БГМУ	++++
Candida albicans	181210169-1	ATCC	±

*ATCC – Американская коллекция типовых культур; **Башкирский Государственный Медицинский Университет

ВЫВОДЫ

1. При анализе химического пространства активных и неактивных антибактериальных молекул были обнаружены характерные структурные особенности для обоих классов молекул: среди активных чаще всего встречаются карбоксильная группа, α,β-ненасыщенные карбонилы и аллилы, имидазол, хинолин и бензимидазол. Среди неактивных молекул чаще всего встречаются фуран, пиперазин, пропаноильная группа и бензодиоксольный фрагмент.
2. Обнаружены молекулярные дескрипторы, которые вносят наибольший вклад в разделение активных и неактивных молекул, среди них: HBD (количество потенциальных доноров водородной связи), Hy (индекс гидрофильности), RB (число свободно вращающихся связей), logS (логарифм растворимости в воде), и др.
3. Наилучшую точность для прогнозирования антибактериальной активности на обучающей выборке в проведенном эксперименте показал алгоритм градиентного бустинга (*Precision* = 97%). Также довольно высокую точность показали случайный лес (*Precision* = 92%) и логистическая регрессия (*Precision* = 70%).
4. На независимой тестовой выборке была показана хорошая способность построенной модели прогнозировать активность для новых соединений, структурно отличающихся от молекул в обучающей выборке (общий хит-рейт составил 24% по сравнению с 2% при рандомном скрининге).
5. В результате экспериментальной валидации модели были найдены ранее не описанные в научной литературе хемотипы соединений (6*H*-тиазоло[4,5-d]пиримидинон, 6-оксо-1*H*-пиримидин-2-ил)пиразол и др.), обладающие активностью, сравнимой с известными лекарствами (левофлоксацин и эритромицин, средний MIC~1-2 мкг/мл). Одно из соединений (соединение **1** – 5'-[(4-бромбензоил)амино]-2,3'-битиофен-4'-карбоновая кислота) показало активность в отношении клинически значимого штамма *S. aureus*.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в рецензируемых научных журналах:

1. **Veselov MS**, Sergiev PV, Osterman IA, Skvortsov DA, Golovina AY, Andreyanova ES, Laptev IG, Pletnev PI, Evfratov SA, Marusich EI, Leonov SV, Ivanenkov YA, Bogdanov AA, Dontsova OA. Common features of antibacterial compounds: an analysis of 10^4 compounds library. // Biomed Khim., 2015; 61(6):785-90. doi: 10.18097/PBMC20156106785.
2. Sergiev PV, Osterman IA, Golovina AY, Andreyanova ES, Laptev IG, Pletnev FI, Evfratov SA, Marusich EI, **Veselov MS**, Leonov SV, Ivanenkov YA, Bogdanov AA, Dontsova OA. High throughput screening platform for new inhibitors of protein biosynthesis. // Moscow University Chemistry Bulletin, 2016; 71:65-67. doi: 10.3103/S0027131416010144.
3. Ivanenkov YA, Zhavoronkov A, Yamidanov RS, Osterman IA, Sergiev PV, Aladinskiy VA, Aladinskaya AV, Terentiev VA, **Veselov MS**, Ayginin AA, Kartsev VG, Skvortsov DA, Chemeris AV, Baimiev AK, Sofronova AA, Malyshev AS, Filkov GI, Bezrukov DS, Zagribelnyy BA, Putin EO, Puchinina MM, Dontsova OA. Identification of Novel Antibacterials Using Machine Learning Techniques. // Front Pharmacol., 2019; 10:913. doi: 10.3389/fphar.2019.00913.
4. **Veselov MS**, Ivanenkov YA, Yamidanov RS, Osterman IA, Sergiev PV, Aladinskiy VA, Aladinskaya AV, Terentiev VA, Ayginin AA, Skvortsov DA, Komarova KS, Chemeris AV, Baimiev AK, Sofronova AA, Machulkin AE, Petrov RA, Maklakova SY, Bezrukov DS, Filkov GI, Zainullina LF, Maximova MA, Zileeva ZR, Kartsev VG, Vakhitova YV, Dontsova OA. Identification of pyrrolo-pyridine derivatives as novel class of antibacterials. // Mol Divers., 2019. doi: 10.1007/s11030-019-09946-3.
5. Ivanenkov YA, Yamidanov RS, Osterman IA, Sergiev PV, Aladinskiy VA, Aladinskaya AV, Terentiev VA, **Veselov MS**, Ayginin AA, Skvortsov DA, Komarova KS, Zagribelnyy BA, Baimiev AK, Shvetc KY, Baimiev AK, Sofronova AA, Machulkin AE, Petrov RA, Zainullina LF, Maximova MA, Zileeva ZR, Vakhitova YV, Bezrukov DS, Puchinina MM, Dontsova OA. Large-scale high-throughput screening revealed 5'-(carbonylamino)-2,3'-bithiophene-4'-carboxylate as novel template for antibacterial agents. // Curr Drug Discov Technol., 2019. doi: 10.2174/1570163816666190603095521.
6. Ivanenkov YA, Yamidanov RS, Osterman IA, Sergiev PV, Aladinskiy VA, Aladinskaya AV, Terentiev VA, **Veselov MS**, Ayginin AA, Skvortsov DA, Komarova KS, Chemeris AV, Baimiev AK, Sofronova AA, Malyshev AS, Machulkin AE, Petrov RA, Bezrukov DS, Filkov GI, Puchinina MM, Zainullina LF, Maximova MA, Zileeva ZR, Vakhitova YV, Dontsova OA. Identification of N-Substituted Triazolo-azetidines as Novel Antibacterials using pDualrep2 HTS Platform. // Comb Chem High Throughput Screen., 2019. doi: 10.2174/1386207322666190412165316.

Тезисы конференций:

7. Sergiev PV, Komarova ES, Osterman IA, Pletnev PhI, Golovina AY , Laptev IG, Evfratov SA, Marusich EI, **Veselov MS**, Leonov SV, Ivanenkov YA, Bogdanov AA, Dontsova OA. Overview of 17,856 Compound Screening for Translation Inhibition and DNA Damage in Bacteria. // Proceedings of the Scientific-Practical Conference "Research and Development-2016", 2018. p. 601-608.