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

На правах рукописи

Столбов Леонид Алексеевич

РАЗРАБОТКА ПОДХОДОВ К ВИРТУАЛЬНОМУ СКРИНИНГУ АНТИВИРУСНЫХ СОЕДИНЕНИЙ С УЧЕТОМ ГЕТЕРОГЕННОСТИ ИНФОРМАЦИИ

Специальность 1.5.8. Математическая биология, биоинформатика

Диссертация на соискание ученой степени
кандидата биологических наук

Научный руководитель:
доктор биологических наук, профессор,
член-корреспондент РАН
Поройков Владимир Васильевич

Научный консультант:
кандидат физико-математических наук
Филимонов Дмитрий Алексеевич

Москва 2023

СОДЕРЖАНИЕ

СПИСОК СОКРАЩЕНИЙ

БД – база данных

ВИЧ-1 – вирус иммунодефицита человека типа 1

ИНС - искусственные нейронные сети

ЯМР – ядерный магнитный резонанс

ADMET – абсорбция, распределение, метаболизм, выведение, токсичность (Absorption, Distribution, Metabolism, Excretion, Toxicity)

CADD – компьютерное моделирование лекарств (Computer-Aided Drug Design)

CoMFA – сравнительный анализ молекулярных полей (Comparative Molecular Field Analysis)

CoMSIA – сравнительный анализ молекулярных индексов сходства (Comparative Molecular Similarity Indices Analysis)

IC_{50} – полуэффективная ингибирующая концентрация

IN – интеграза ВИЧ-1

Ki – константа ингибирования

LBDD – конструирование лекарственных препаратов на основе лиганда (Ligand-Based Drug Design)

LR – логистическая регрессия

MNA – (дескрипторы) многоуровневых атомных окрестностей (Multilevel Neighborhoods of Atoms)

Mpro (3CLpro) – главная (химотрипсин-подобная) протеаза SARS-CoV-2

PLpro – папаин-подобная протеаза SARS-CoV-2

PR – протеаза ВИЧ-1

QSAR – количественные зависимости «структура-активность»

RdRp – РНК-зависимая РНК-полимераза SARS-CoV-2

RT – обратная транскриптаза ВИЧ-1

SALI – индекс ландшафта «структура-активность» (Structure-Activity Landscape Index)

SAR – зависимости «структура-активность» (Structure-Activity Relationships)

SBDD – конструирование лекарственных препаратов на основе структуры макромолекулы-мишени (Structure-Based Drug Design)

SCEC – самосогласованный экстремальный классификатор (Self-Consistent Extreme Classifier)

SCLC – самосогласованный логистический классификатор (Self-Consistent Logistic Classifier)

SDF – файл, содержащий структурные данные (Structure Data File)

SMILES – строковый формат молекулы (Simplified Molecular Input Line Entry System)

SVM – метод опорных векторов (Support Vector Machine)

QNA – (дескрипторы) количественных атомных окрестностей (Quantitative Neighborhoods of Atoms)

ВВЕДЕНИЕ

Актуальность работы. Поиск новых антиретровирусных соединений является актуальной задачей, поскольку разрешенные к применению препараты для лечения ВИЧ инфекции не обеспечивают полную элиминацию вируса из организма и обладают серьезными побочными эффектами, а их клиническое применение приводит к возникновению резистентных штаммов [1]. Пандемия COVID-19, вызванная широким распространением в популяции вируса SARS-CoV-2 и его высокой контагиозностью, стимулировала разработку новых подходов к созданию антивирусных препаратов для оперативного ответа на новые биогенные угрозы [2].

Применение компьютерных методов для виртуального скрининга соединений, потенциально обладающих противовирусной активностью, позволяет существенно снизить финансовые затраты и временные издержки на проведение исследований, а также риск получения отрицательных результатов на ключевых этапах поиска и оптимизации соединений-лидеров [3]. Методы компьютерного дизайна лекарств, основанные на структуре макромолекулы-мишени (Structure-Based Drug Design), подразумевают наличие информации о пространственной структуре мишени в комплексе с лигандом, расшифрованной с достаточно высоким разрешением, и требуют использования высокопроизводительных вычислительных ресурсов [4, 5]. Подходы к компьютерному дизайну лекарств, основанные на структуре лигандов (Ligand-Based Drug Design), базируются на анализе зависимостей «структура-активность» для соединений обучающей выборки с экспериментально установленными характеристиками биологической активности. В настоящее время эти *in silico* подходы широко применяются для поиска и оптимизации фармакологических веществ, проявляющих целевую биологическую активность [6].

Изначально построение моделей количественных зависимостей «структура-активность» было основано на анализе выборок лекарственно-подобных

соединений, принадлежащих к одному химическому классу, биологическая активность которых была исследована в одних и тех же условиях эксперимента [7].

Вследствие развития медицинской химии и разработки высокопроизводительных методов исследования активности синтезированных соединений *in vitro* в биохимических или клеточных тест-системах доступная информация о структуре и биологическом действии исследованных веществ стала существенно более гетерогенной [8]. Накопление этой информации в свободно-доступных базах данных (PubChem, ChEMBL, и др.) обеспечило предпосылки для ее использования для создания обучающих выборок при построении количественных и классификационных моделей зависимостей «структура-активность» [9]. Хотя ранее разработанные методы компьютерного прогноза в ряде случаев дают возможность получать на основе анализа гетерогенных данных зависимости «структура-активность», обладающие удовлетворительной предсказательной способностью (см., например, [10]), развитие новых подходов позволит повысить качество получаемых результатов.

Валидацию разрабатываемых методов целесообразно провести на примере анализа зависимостей «структура-активность» как с использованием искусственно сгенерированных данных, для которых такие зависимости будут заранее известны, так и для данных Tox21 [11], которые часто применяются для сравнения новых (Q)SAR моделей [12].

Прошедшие валидацию методы могут быть применены для построения зависимостей «структура-активность» для ингибиторов основных мишеней ВИЧ-1 и ингибиторов репликации вируса SARS-CoV-2. Для ингибиторов ВИЧ-1 накоплен обширный экспериментальный материал, обеспечивающий формирование объемных гетерогенных выборок, содержащих информацию о структуре антиретровирусных соединений, принадлежащих к различным химическим классам и проявляющих активность в широком диапазоне значений.

Для ингибиторов репликации вируса SARS-CoV-2 число доступных данных сопоставимо с информацией об ингибиторах ВИЧ-1, однако они получены в

нестандартизированных тест-системах и в отсутствие общепринятых препаратов сравнения, что повышает гетерогенность доступной информации. Поэтому разработка (Q)SAR моделей, обладающих высокой точностью и предсказательной способностью, особенно с целью поиска новых препаратов для терапии COVID-19, является актуальной задачей.

Целью диссертационной работы является разработка и валидация методов виртуального скрининга противовирусных соединений на основе анализа зависимостей «структура-активность» в гетерогенных массивах данных.

Для достижения этой цели решались следующие задачи:

1. Разработать алгоритмы построения классификационных моделей зависимостей «структура-активность», выполнить их программную реализацию.
2. Провести валидацию новых методов и сравнить их с существующими подходами на основе сгенерированных искусственных данных и выборок Tox21, широко используемых при сравнении методов классификации.
3. Сформировать обучающие выборки для анализа зависимостей «структура-активность» ингибиторов белков ВИЧ-1 и SARS-CoV-2 из доступной в открытых источниках и прошедшей предварительную обработку информации.
4. Построить количественные и классификационные модели зависимостей «структура-активность» для ингибиторов белков ВИЧ-1 и провести сопоставление их точности, предсказательной способности и области применимости.
5. Построить классификационные модели зависимостей «структура-активность» для ингибиторов ферментов SARS-CoV-2.
6. Создать свободно доступный веб-ресурс, позволяющий оценивать фармакологические характеристики, связанные с анти-ВИЧ активностью и сопутствующими заболеваниями, на основе прогноза.

Научная новизна

Разработаны оригинальные методы самосогласованной логистической (SCLC – Self-Consistent Logistic Classifier) и экстремальной (SCEC – Self-Consistent Extreme Classifier) классификации, которые повышают эффективность виртуального скрининга противовирусных соединений на основе анализа зависимостей «структура-активность» в гетерогенных массивах данных.

С использованием SCLC и SCEC впервые построены модели оценки ингибирующей активности к мишеням ВИЧ-1 и SARS-CoV-2, обладающие хорошей точностью и предсказательной способностью.

Научно-практическая значимость

Разработанные в диссертационной работе методы анализа взаимосвязей «структура-активность» могут быть применены для поиска и конструирования новых антивирусных соединений – прототипов лекарств для терапии ВИЧ-инфекции и COVID-19.

Построенные с применением SCLC и SCEC классификационные модели обладают характеристиками, сравнимыми или превосходящими таковые у ранее созданных методов. SCLC и SCEC позволяют строить классификационные модели на основе слабо сбалансированных выборок и автоматически проводить эффективный отбор минимального числа значимых независимых переменных, что повышает объективность расчетных оценок.

Свободно доступный веб-ресурс AntiHIV Pred предоставляет широкому кругу исследователей возможность отбирать наиболее перспективные соединения для синтеза и определять приоритетные направления тестирования активности ранее синтезированных веществ.

Разработанный нами подход может быть также применен для построения зависимостей «структура-активность» в других фармакотерапевтических областях.

Личный вклад автора. Автором отобрана и проанализирована релевантная области исследования литература. Автор провел сбор и обработку данных для составления обучающих выборок и создания классификационных и регрессионных

моделей «структура-активность». Автор непосредственно участвовал в разработке и валидации методов классификации, реализующих их программ и создании веб-сервиса.

Положения, выносимые на защиту:

- Разработанная процедура извлечения и обработки информации о биологически активных веществах позволила сформировать выборки гетерогенных данных об антиретровирусных и антикоронавирусных соединениях.
- Модели, разработанные на основе самосогласованной логистической и экстремальной классификации, обладают более высокой прогностической способностью по сравнению с количественными методами.
- На основе сформированных нами выборок получены количественные зависимости «структура-активность» для ингибиторов белков ВИЧ-1 и классификационные зависимости «структура-активность» для ингибиторов белков ВИЧ-1, SARS-CoV-2, обладающие хорошей точностью и предсказательной способностью.
- Разработанный нами свободно доступный веб-сервис AntiHIV Pred позволяет прогнозировать антиретровирусное действие и виды биологической активности, связанные с терапией ВИЧ-ассоциированных заболеваний.

Апробация работы. Основные положения диссертации были представлены на российских и международных конференциях и симпозиумах: XXXVIII Symposium of Bioinformatics and Computer-Aided Drug Discovery (Virtual, 2022), XXVII Российский национальный конгресс «Человек и лекарство». (Москва, 2020), Third international School-Seminar «From Empirical to Predictive Chemistry» (Казань, 2018), XXV Российский национальный конгресс «Человек и лекарство» (Москва 2018), IX Международный конгресс «Биотехнология: состояние и перспективы развития. Науки о жизни» (Москва, 2019).

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ

Поиск и разработка новых фармакологических веществ является сложным процессом, требующим значительных временных и финансовых издержек. Традиционный подход к поиску прототипов новых лекарственных препаратов включает в себя многостадийный химический синтез и экспериментальное исследование большого числа соединений в биохимических и клеточных тест-системах [13].

В настоящее время все большую роль приобретают вычислительные методы, которые применяются для анализа молекулярных мишеней, структурных и физико-химических характеристик лигандов и взаимодействия лиганд-мишень с целью обнаружения потенциальных соединений-лидеров [14]. Эти методы носят название CADD – компьютерное конструирование лекарств. Они позволяют осуществлять вычислительные эксперименты с целью отбора наиболее перспективных соединений для синтеза и определения приоритетных направлений тестирования их биологической активности.

Использование методов компьютерного конструирования лекарств снижает временные издержки и финансовые затраты на поиск и разработку новых лекарственных препаратов. Методы CADD применяются для поиска и конструирования новых фармакологических веществ, расчета фармакокинетических характеристик (абсорбция, распределение в крови и тканях, метаболизм и выведение из организма), оценки профилей токсичности, расчета физико-химических свойств и др. В настоящее время компьютерные методы широко применяются на всех стадиях разработки новых лекарственных препаратов [15].

Подходы к конструированию лекарственных препаратов подразделяют на две категории: методы, основанные на структуре макромолекулы-мишени (SBDD), и методы, основанные на структуре лигандов (LBDD). Рассмотрим эти подходы более подробно.

1.1 Компьютерное конструирование лекарств

1.1.1 Методы, основанные на структуре макромолекулы-мишени

В этих подходах используется информация о пространственной структуре молекулы-мишени (как правило, белка) для поиска и конструирования потенциальных лигандов, взаимодействующих с мишенью. Такого рода информацию получают экспериментальными методами (рентгеноструктурный анализ, ЯМР высокого разрешения, криоэлектронная микроскопия) или моделируют с учетом гомологии аминокислотных последовательностей.

Получение пространственной структуры мишени.

В случае отсутствия экспериментальных данных о пространственной структуре мишени иногда применяют вычислительные методы. Надежные 3D модели можно получить методами моделирования по гомологии, основанными на сходстве аминокислотной последовательности изучаемой макромолекулы-мишени и другого белка, трехмерная структура которого определена экспериментально. Необходимым условием надежности модели изучаемого белка является превышающая ~30% степень идентичности первичных структур макромолекул [16, 17]. В процессе моделирования белок с известной пространственной структурой служит шаблоном, по которому происходит определение координат атомов искомого белка в пространстве [18, 19]. В настоящее время для решения данной задачи также используются системы искусственного интеллекта, такие как AlphaFold, после обучения на белках с известной пространственной структурой [20].

Информация об известных 3D структурах макромолекул и их комплексах с лигандами, которая может быть использована как в качестве исходных экспериментальных данных, так и для моделирования, доступна в различных источниках, например в банке данных белковых структур (Protein Data Bank, PDB) [21]. В нем на 25 июня 2023 года содержатся сведения о более чем 190 тысячах структур. Для целей SBDD в основном применяются структуры, расшифрованные

с высоким разрешением <2 Å. Количество таких данных о структуре белка составляет около 45 тысяч записей [22].

Важными являются особенности экспериментального определения структуры белка, например, необходимо учитывать возможные изменения структуры белка в зависимости от условий эксперимента, таких как температура, pH и наличие лигандов. Все эти факторы могут существенно влиять на структуру белка и его функциональные свойства, поэтому необходимо учитывать их при интерпретации результатов и проведении дальнейших исследований [23, 24].

В задачах поиска новых лигандов желательно наличие информации о пространственной структуре комплекса «белок-лиганд», что позволяет установить механизм взаимодействия между ними. Наличие такой информации дает возможность дополнить недостающие аминокислотные остатки и уточнить пространственное расположение активного центра белка. Далее проводится подготовка белка в соответствии с общими требованиями к анализу, включающая дополнение структурных элементов, таких как связи и атомы водорода, восстановление недостающих фрагментов, а также добавление или удаление молекул воды и ионов металлов [25].

Имея в наличии подготовленную соответствующим образом 3D структуру молекулярной мишени, проводят поиск лигандов методами виртуального скрининга с применением докинга или молекулярной динамики в библиотеках известных соединений, либо осуществляют конструирование лигандов *de novo*.

SBDD методы виртуального скрининга.

Для оценки конформации лиганда, обеспечивающей наилучшее связывание с активным центром белка, применяется молекулярный докинг. Подход заключается в генерации различных возможных конформаций и ориентаций лиганда в области связывания. При этом вводятся дополнительные правила и ограничения, такие как запрет на докирование определенных областей, допустимость или недопустимость вращений аминокислотных остатков мишени относительно определенных связей [26]. Наиболее важной задачей при этом

является оценка энергии связывания. Для этого исследуемая область характеризуется различными потенциалами взаимодействия в точках, соответствующих наложенной на нее 3-х мерной сетки. Потенциалы обычно описывают эффекты силовых полей, порождаемых взаимодействиями Ван-дер-Ваальса, электростатическими взаимодействиями, водородными связями, присутствием растворителя, и др. Используя значения потенциалов при различных конформациях лиганда, составляется оценочная функция, характеризующая энергию образования комплекса «белок-лиганд» [27, 28]. Получаемые значения оценочной функции не соответствуют экспериментально проверяемым значениям, однако могут иметь с ними корреляции, что позволяет производить отбор лигандов для экспериментального тестирования. Пример расположения лиганда в определенной конформации в активном центре мишени по результатам докинга приведен на рисунке 1а.

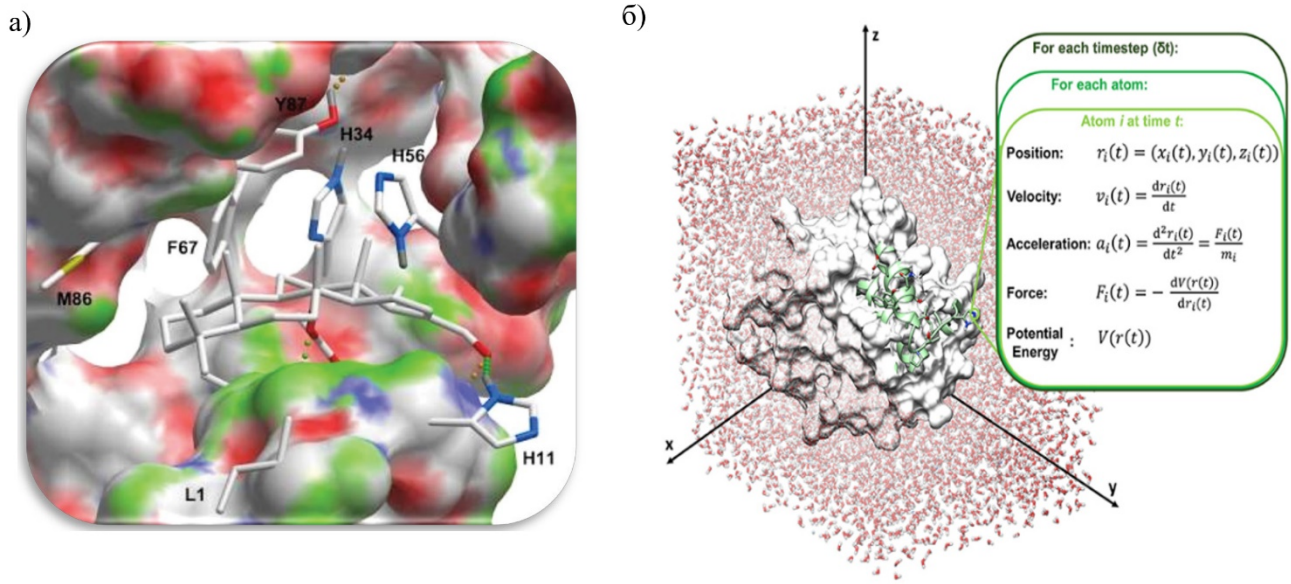


Рисунок 1 – Иллюстрация: а) докинга [28], б) молекулярной динамики [29]

В отличие от докинга, методы молекулярной динамики дают оценки свободной энергии не в статичном расположении лиганда в активном центре, а в процессе связывания на микросекундном временном диапазоне. Преимуществами

этого подхода являются учет динамических изменений структур лиганда и мишени, стабильности конформаций и возможных реакций. Таким образом дается оценка не конечным конформациям, а траекториям. Тем не менее, возможности для экспериментальной валидации каждого из результатов таких симуляций ограничены в еще большей степени, чем в случае докинга, при гораздо более высоких вычислительных затратах. Интерпретация результатов молекулярного моделирования подразумевает дополнительный анализ и кластеризацию траекторий и требует высокой подготовки исследователя в каждом конкретном случае. Иллюстрация молекулярного моделирования с использованием молекулярной динамики приведена на рисунке 1б.

Оба указанных подхода используются для проведения виртуального скрининга библиотек лигандов [30]. Выбор соответствующей библиотеки лигандов, а также проводимая предобработка актуальны и для методов LBDD и описаны в последующих разделах.

De novo дизайн.

Подходом, противоположным виртуальному скринингу, является *de novo* дизайн. На начальном этапе данного подхода производится оценка областей молекулы-мишени, которая основана на электростатическом описании особенностей активного центра. Как правило, такие электростатические особенности обусловлены наличием конкретных функциональных групп. В дальнейшем именно расположение этих групп является определяющим. На следующем этапе производится оценка связывания определенных фрагментов, из которых может состоять потенциальный лиганд. Поэтому производят оценку взаимодействий имеющихся доступных фрагментов с ранее найденными группами активного центра. На завершающем этапе производят поиск фрагментов-линкеров для их соединения в конечную структуру и оценивают свободную энергию уже для всего лиганда [31]. Этапы *de novo* дизайна проиллюстрированы на рисунке 2.

SBDD методы были успешно применены при разработке таких препаратов как Ралтитрексид (Raltitrexed) [32], Ампренавир (Amprenavir) [33], [34], Изониазид

(Isoniazid) [35], Эпалрестат (Epalrestat) [36], Флурбипрофен (Flurbiprofen) [37], [38], Дорзоламид (Dorzolamide) [39] и др. [40, 41].

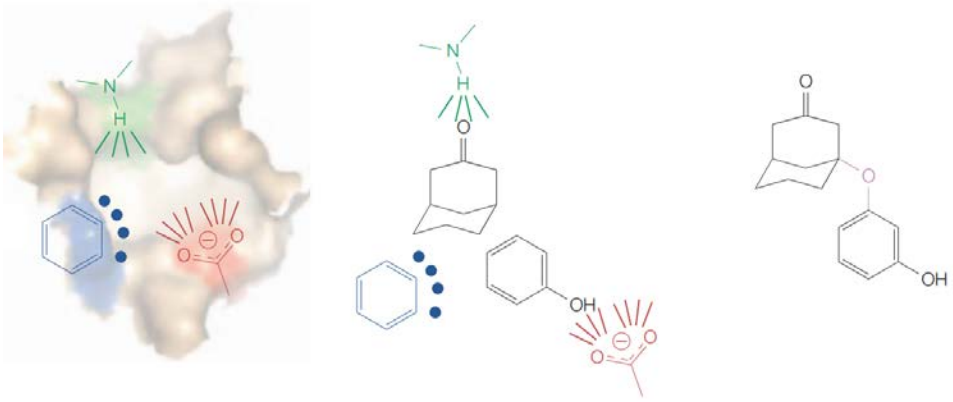


Рисунок 2 – Этапы *de novo* дизайна лекарств: определение участков активного центра, пригодных для связывания, подбор наилучших фрагментов, формирование целой структуры из фрагментов [31]

1.1.2 Методы, основанные на структуре лигандов

В этих подходах используется информация о структурах лигандов и их экспериментальной активности, например, по отношению к мишени либо фармакологическому действию. Структуры лигандов обычно известны из схем их синтеза и идентифицируются методами ЯМР-спектроскопии [42] и другими методами аналитической химии. Данные об активности могут быть получены с использованием различных экспериментальных протоколов как при биохимическом скрининге *in vitro* в биохимических или клеточных тест-системах, так и при скрининге *in vivo*. LBDD методы основаны на том предположении, что похожие структуры обладают похожими свойствами, и включают в себя анализ структурного сходства, фармакофорный анализ и анализ зависимостей структура-свойство ((Q)SAR).

Анализ структурного сходства.

Анализ структурного сходства основан на подобии молекулярных графов, фрагментов и скаффолдов, или опирается на векторизованное представление химических структур на основе структурных дескрипторов. Поэтому ключевым

вопросом является, какую именно схожесть необходимо оценить, и от ответа на него зависит, какие именно представления использовать [43].

При описании структуры важным является понятие скаффолда и боковых заместителей (R-групп). Скаффолд, также представляющий собой структуру Маркуша, определяется как структурный фрагмент, характеризующий группу молекул, принадлежащую к определенному химическому классу [44]. Синтез структур с одинаковом скаффолдом проводится по одинаковым или похожим схемам и, соответственно, такой фрагмент может служить шаблоном для синтеза соединений целого химического класса, что часто используется в комбинаторной химии [45]. Одна из гипотез анализа сходства заключается в предположении, что при сохранении боковых групп и замене скаффолда на похожий скаффолд полученная молекула будет обладать свойствами сравнимыми с исходной молекулой (рис. 3). Аналогичная гипотеза выдвигается и для боковых заместителей.

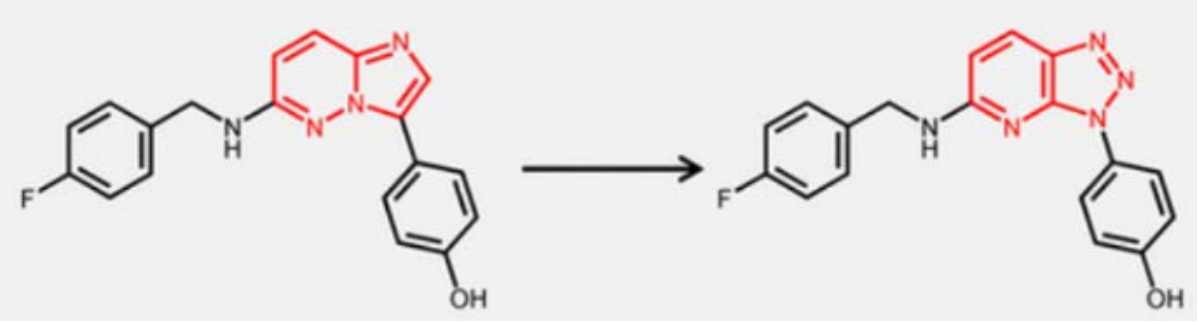


Рисунок 3 – Замена скаффолда [45]

При использовании иерархической кластеризации с использованием скаффолдов [46] или поиске наибольшей общей подструктуры [47] оценка сходства для каждой пары соединений заключается в классификации на схожие или несхожие.

Напротив, для сравнения векторов, описывающих молекулу, могут быть применены различные метрики с характерной оценкой сходства пары молекул в

интервале [0, 1]. Так, например, для битовых векторов часто используется сравнение по Танимото [48] для каждой пары молекул:

$$T_{AB} = \frac{|A \wedge B|}{|A| + |B| - |A \wedge B|}$$

где A и B векторы как упорядоченные множества, описывающие соответствующие молекулы.

Применение понятия сходства сопряжено с изучением и первичным анализом данных о структурах для кластеризации и укладки соединений на двумерную плоскость таким образом, что схожие структуры оказываются близко расположенными, используя такие методы, как t-SNE [49], UMAP [50], [51], k-ближайших соседей [52] (рис. 4а).

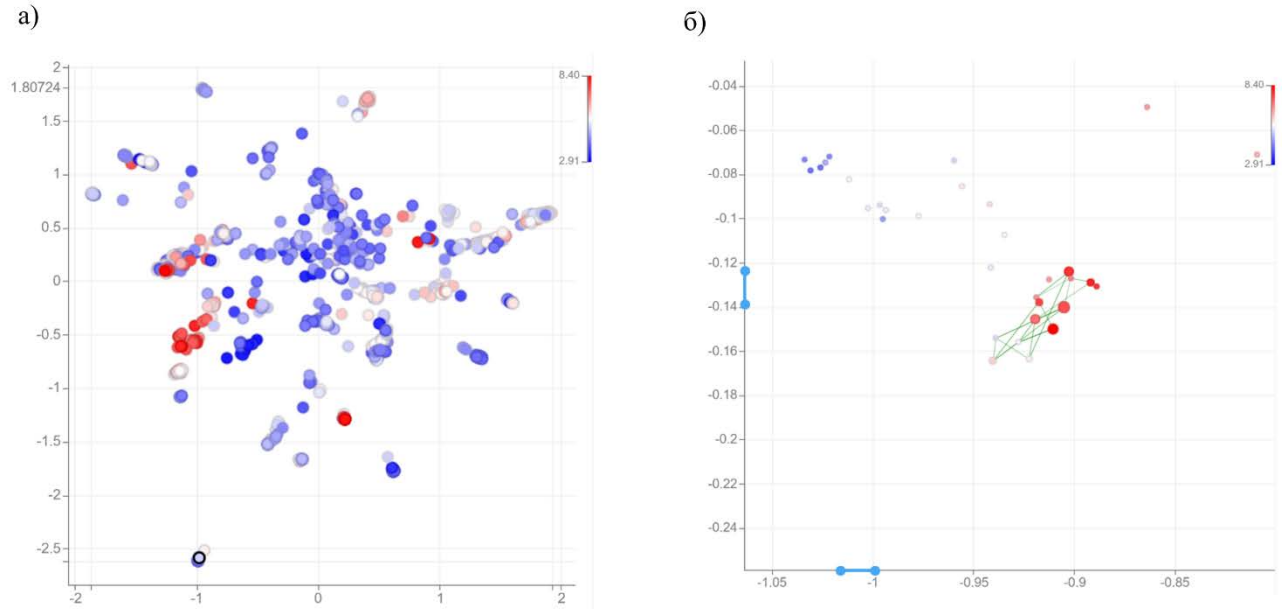


Рисунок 4 – Применение сходства для а) визуализации химического пространства (красные - активные, синие - неактивные), б) пики активности (чем больше разница в размере точек в одном графе, тем выше значение SALI)

Подобный подход используется не только для визуализации исследуемого химического пространства, но и для оценки ассоциированной со структурами биологической активности. Явным примером такого использования является

выявление «пиков активности» (activity cliffs) [53] – ситуаций, когда небольшие структурные отличия сопровождаются значительной разницей в величине активности. Для детекции самих пиков между двумя структурами (рис. 4б) используются различные индексы, например, SALI [54]:

$$SALI_{AB} = \frac{|Y_A - Y_B|}{1 - T_{AB}}$$

Несмотря на то, что похожие соединения не всегда обладают схожими свойствами, для проведения виртуального скрининга может быть использован поиск по сходству. Этот подход обеспечивает обогащение результирующей выборки соединениями со свойствами, близкими к запросу [55]. Например, при отборе сходных соединений по коэффициенту Танимото, превышающем 0,85, около 30% отобранных молекул будут иметь схожие свойства [56]. При этом поиск по сходству не требует наличия данных об экспериментальной активности большого числа соединений, на основе которых могут быть построены зависимости «структура-активность».

Фармакофорное моделирование.

Понятие фармокофора неоднозначно и менялось с течением времени. В современном контексте это понятие означает совокупность молекулярных особенностей, достаточных для описания взаимодействия лиганда с потенциальной мишенью [57, 58]. Фармакофор не является конкретной функциональной группой или сочетанием таких групп, но является абстрактным представлением различных электростатических и других свойств, которые эти группы могут проявлять. При этом не имеет значения, что именно может обуславливать конкретный фармакофор – специфическую гидрофобную или гидрофильную группу, всю молекулу или целый химический класс, молекулу в определенной конформации или набор ее конформаций. На практике моделируют фармакофоры нескольких типов, описывающие гидрофобные, положительные и отрицательные полярные области, катионы, анионы, проявления ароматичности структур и возможности образования водородных связей [59]. В LBDD методах фармакофоры рассчитываются для уже известных лигандов, прогноз свойств

новых соединений осуществляется на основе этой информации [60]. Фармакофоры также находят применение и в SBDD, где фармакофоры для потенциального лиганда рассчитываются на основе особенностей активного центра [61, 62].

(Q)SAR.

Анализ (количественных) зависимостей «структура-активность» ((Q)SAR) использует статистические оценки и методы машинного обучения для описания наблюдаемых величин активности y_i конкретной структуры по m-мерному набору $\boldsymbol{x}_i$ ее особенностей. Отображение $f: X \to Y$ в терминах (Q)SAR называется моделью, что соответствует и статистической интерпретации [6]. Соответственно, двумя основополагающими задачами при построении зависимостей «структура-активность» являются принципы, по которым будут выбраны характеристики молекул X и подходы к созданию модели.

Существует ряд способов описания молекулы через перечисление её характеристик. Как и в случае анализа по сходству, распространенным способом является представление молекулы в виде вектора двоичных, целых или действительных чисел – дескрипторов. Зачастую дескрипторы характеризуют некоторые физико-химические свойства молекулы или присутствие в ней определенных структурных фрагментов. Более подробно примеры таких дескрипторов представлены в разделе 1.2. Тем не менее, существует и ряд подходов с иным описанием молекулы.

Одним из примеров таких подходов является 3D-QSAR, использующий 3D дескрипторы, рассчитываемые по геометрическим представлениям пространственной структуры молекулы. Методы 3D-QSAR отличаются по тому, какие геометрические представления они используют, например, извлекаются численные характеристики различных полей, получаемых из модели фармакофоров. Примерами таких методов являются CoMFA и CoMSIA [63]. Существенной особенностью 3D-QSAR является использование информации о

геометрии расположения функциональных групп и экспериментальных данных для конкретных стереоизомеров, однако, такая информация зачастую отсутствует.

Другим примером использования иных представлений молекул являются методы QSAR с применением графовых нейросетей. В этом подходе дескрипторы для молекул не рассчитываются, а используется весь молекулярный граф в качестве входной информации [64].

Для построения (Q)SAR модели используются различные математические подходы. Эти подходы включают в себя метод опорных векторов (SVM) [65], логистическую регрессию, метод частичных наименьших квадратов [66], случайный лес, искусственные нейронные сети (ИНС) и их многочисленные модификации. Модели, получаемые с помощью различных подходов, будут отличаться по своим характеристикам, количеству используемых переменных, интерпретируемости и прогностической способности. Подробнее эти детали описаны в разделе 1.2.4.

(Q)SAR модели широко используются в фармацевтических исследованиях при разработке новых лекарственных препаратов с целью:

- виртуального скрининга молекул (обогащение исследуемых выборок соединениями с наибольшей вероятностью наличия активности);
- оптимизации соединений-лидеров;
- оценки ADMET свойств.

1.1.3 Источники данных о лигандах

Любую молекулу можно рассматривать как элемент так называемого химического пространства. Концепция химического пространства основана на описании всего набора возможных химических структур, а также связанных с этими структурами всевозможных свойств. Это понятие имеет различные определения, как например, данное Варнеком и Баскиным [67] – набор графов или векторов дескрипторов, для которых необходимо установить взаимосвязи. Оценка размеров химического пространства является нетривиальной задачей и обычно

сводится к пространству малых молекул или даже к пространству лекарственно-подобных соединений, однако и в этом случае потенциальное число входящих в состав такого пространства молекул достигает от 10^{20} до 10^{60} [68, 69]. Существующие наборы данных, конечно, содержат гораздо меньшее число структур.

В задачах компьютерного конструирования лекарств используются источники данных, принципиально отличающиеся по своему назначению. Данные о химической структуре, ассоциированные с определенной биологической активностью, используются при создании, валидации и тестировании (Q)SAR моделей. Такими данными являются, например, результаты *in vitro* или *in vivo* скрининга, которые содержатся в литературных источниках и патентах. Для удобства использования они агрегируются в крупные базы данных и аннотируются. Наиболее широко используемыми свободно доступными базами данных такого типа являются ChEMBL [70] и PubChem [71].

С целью проведения виртуального (*in silico*) скрининга определяется та часть химического пространства, среди которой могут быть найдены новые фармакологически активные соединения. Это может быть сравнительно небольшой набор уже синтезированных соединений, либо структуры, которые планируются к синтезу. С другой стороны, это может быть сравнительно большая библиотека структур коммерчески доступных образцов лекарственно-подобных соединений, содержащая информацию о миллионах молекул, сгенерированных *in silico* структур. Для экономии времени, необходимого для проведения соответствующих расчетов, как правило, используют различные методы фильтрации по физико-химическим или потенциальным ADMET свойствам. Помимо коммерчески доступных образцов и синтезированных соединений, виртуальный скрининг проводят по различным базам данных, которые могут содержать потенциальные фармакологически активные соединения, примеры которых приведены ниже.

DrugBank [72] – база данных уже существующих лекарственных препаратов и их фармакологически активных субстанций. На 01.06.2023 она содержит более 12 000 структур, относящихся к уже существующим препаратам, и может быть использована в исследованиях по репозиционированию лекарств.

ZINC [73] – база данных, содержащая информацию о трехмерной структуре более 200 миллионов коммерчески доступных веществ. Данная информация может быть использована для докинга. Кроме того, в базе данных имеется информация о 750 миллионов структур, подготовленных для поиска соединений по сходству и тестирования их биологической активности.

GDB [74] – база данных, содержащая сгенерированные молекулы, состоящие из 11, 13 и 17 тяжелых атомов. Она содержит более 166 миллиардов структур, сгенерированных по основным принципам образования химической связи.

SAVI [75] – база данных потенциально синтезируемых соединений, содержащая более 1,5 миллиардов структур и пути их синтеза. Генерация данных произведена на основе правил о химических реакциях.

1.2 Применение (Q)SAR анализа для виртуального скрининга

1.2.1 Использование данных для обучения

Важной предпосылкой для построения (Q)SAR моделей является выбор надежных данных для обучения, поскольку это напрямую влияет на результаты построения зависимостей «структура-активность».

Традиционно считалось, что (Q)SAR модель, обладающую хорошей прогностической способностью, можно получить при выполнении двух условий. Во-первых, все соединения выборки принадлежат к одному химическому классу или схожи между собой; во-вторых, значения активности получены в одних и тех же экспериментальных условиях. Однако, при таком подходе область применимости моделей будет включать лишь малую часть всего химического пространства. В настоящее время при проведении (Q)SAR анализа часто используют гетерогенные данные из многочисленных доступных источников,

которые включают структуры из разных химических классов и значения активности, полученные при использовании различных экспериментальных протоколов [6].

Модели, построенные с использованием таких данных, являются более общими и могут быть применены к существенно большей части химического пространства. При использовании такого подхода на исследователя возлагается решение дополнительных задач, включая: возможные существенные различия величин активности при разных условиях эксперимента для конкретного химического соединения; отсутствие возможности перевода классификационных значений в количественные; асимметрия распределения значения активности в ту или иную сторону, и т.д. Объединение таких данных в одну обучающую выборку приводит к проблемам их сопоставимости, зависящей от подробного описания методики экспериментальных исследований, что зачастую недоступно. Все указанные факторы приводят к снижению точности и предсказательной способности получаемых моделей по сравнению с подходом, когда используются соединения конкретного химического класса.

Описанная ситуация вынуждает исследователя к выбору между более низкой прогностической способностью обобщенной модели и ограничениями, наложенными на модель, построенную на однородной выборке [76].

Подходом, позволяющим преодолеть трудности, связанных с качеством и гетерогенностью данных, является применение классификационных моделей для определения активных и неактивных структур. В отличие от преобразования данных в количественные значения, преобразование в классификационные величины возможно практически всегда. Так, количественные значения могут быть классифицированы, невзирая на различные экспериментальные условия и наличие ошибок измерения. Сама классификация при этом происходит по заранее принятому порогу активности [77, 78]. Также, для задач классификации существуют приемы, позволяющие увеличить количество неактивных соединений,

например, путем добавления структур, для которых не предполагается проявление выбранной активности [79].

Другим аспектом построения моделей, который непосредственно связан с данными, является предобработка структур и значений активности. Этот этап необходим для устранения неточностей, обработки дублированной информации и приведения всех данных к стандартизованному виду. Согласно современным требованиям [80, 81] предобработка структур подразумевает следующие процедуры:

- удаление записей с ошибками в структуре (недопустимые значения валентности и т.п.), неорганическими и металлоорганическими соединениями;
- удаление структур с редкими элементами;
- выделение фармакологического компонента в смесях и солях;
- стандартизация структурных формул;
- приведение химических структур к единой ароматической форме;
- приведение химических структур к незаряженной форме;
- приведение химических структур к единой форме таутомеров;
- удаление стереохимической информации.

Для предобработки значений активностей выполняются следующие действия:

- удаление записей, содержащих информацию о нерелевантных для исследования штаммах и неверно аннотированных мишенях;
- расчет медианных значений активности для структурных дубликатов, удаление записей с противоречивой информацией;
- нормализация значений активности, полученных в различных тест-системах.

1.2.2 Дескрипторы для построения (Q)SAR моделей

Химическую структуру можно описать набором экспериментальных или теоретических дескрипторов в зависимости от выбранного молекулярного представления. Экспериментальными дескрипторами являются различные физико-химические величины (как, например, коэффициент распределения октанол-вода), значения которых получены в определенных экспериментальных условиях. Теоретические дескрипторы получают с использованием конкретного выбранного хемоинформатического алгоритма с использованием однозначного молекулярного представления. К настоящему времени существует большое число различных типов дескрипторов, применимых к широкому спектру задач. Большое количество различных типов дескрипторов приводит к проблеме выбора конкретного типа/ов для проведения анализа зависимостей «структура-активность» [82].

Желательно, чтобы алгоритмы для расчета дескрипторов удовлетворяли следующим свойствам:

- однозначно определяли набор дескрипторов для конкретной структуры;
- имели структурную или физико-химическую интерпретацию;
- были применимы для любого типа/класса соединений;
- получали набор дескрипторов одной природы.

Такими дескрипторами, в частности, являются дескрипторы многоуровневых атомных окрестностей (MNA) и дескрипторы количественных атомных окрестностей (QNA), разработанные Д. А. Филимоновым [83-85].

Дескрипторы MNA описывают структуру молекулы множеством символьных строк, являющихся линейной нотацией атомно центрированных фрагментов молекулярного графа. Построение MNA дескрипторов для каждого атома выполняется рекурсивно:

- дескриптор MNA 0-го уровня это метка самого атома A;

- дескриптор MNA любого последующего уровня – условное обозначение структурного фрагмента $A(D_1 D_2 \dots D_i \dots)$, где D_i – дескриптор MNA предыдущего уровня для i-го непосредственного соседа данного атома A.

Дескрипторы соседей $D_1 D_2 \dots D_i \dots$ перечисляются в каком-нибудь однозначном, например, лексикографическом порядке. Подобная итерационная процедура может быть продолжена до любого необходимого уровня [83].

Метки атомов могут включать не только обозначение атома согласно периодической таблице Д. И. Менделеева, но и любую однозначную дополнительную информацию, например, обозначение принадлежности атома к линейной или циклической части молекулы.

Дескрипторы QNA [85] рассчитываются на основе матрицы связности молекулы **C** и стандартных значений потенциала ионизации *IP* и сродства к электрону *EA* каждого атома молекулы. Дескрипторы QNA описывают каждый из атомов в молекуле двумя действительными значениями *P* и *Q*. Для i-ого атома они вычисляются следующим образом:

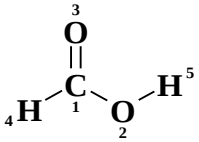
$$P_i = B_i \sum_k \left(\mathrm{Exp}\left(-\frac{1}{2}\mathbf{C}\right)\right)_{ik} B_k \qquad (1),$$

$$Q_i = B_i \sum_k \left(\mathrm{Exp}\left(-\frac{1}{2}\mathbf{C}\right)\right)_{ik} B_k A_k \qquad (2),$$

где $A_k = \frac{1}{2}(IP_k + EA_k)$, $B_k = (IP_k - EA_k)^{-\frac{1}{2}}$ для атома k. Использованные в диссертационной работе значения EA и IP приведены в Приложении 1.

Дескрипторы QNA описывают каждый атом в молекуле, при этом значения P и Q зависят от атомного состава и структуры всей молекулы (рис. 5).

Вычисление дескрипторов QNA представляет собой достаточно простую процедуру. Поскольку матрица связности $\mathbf{C}$ состоит только из нулей и единиц и необходим только результат умножения матрицы $\mathrm{Exp}\left(-\frac{1}{2}\mathbf{C}\right)$ на векторы, то и отдельный расчет самой матричной экспоненты не нужен.



а)

$$\mathbf{C} = \begin{vmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{vmatrix}$$

б)

$$\mathrm{Exp}\left(-\tfrac{1}{2}\mathbf{C}\right) = \begin{vmatrix} 1.40 & -0.59 & -0.57 & -0.57 & 0.14 \\ -0.59 & 1.27 & 0.14 & 0.14 & -0.54 \\ -0.57 & 0.14 & 1.13 & 0.13 & -0.02 \\ -0.57 & 0.14 & 0.13 & 1.13 & -0.02 \\ 0.14 & -0.54 & -0.02 & -0.02 & 1.13 \end{vmatrix}$$

в)

	EA	*IP*	*A*	*B*	*P*	*Q*
C	*1.263*	*11.26*	*6.262*	*0.316*	*-0.00218*	*-0.1820*
O	*1.461*	*13.62*	*7.541*	*0.287*	*0.02944*	*0.3019*
O	*1.461*	*13.62*	*7.541*	*0.287*	*0.06199*	*0.5297*
H	*0.754*	*13.60*	*7.177*	*0.279*	*0.05812*	*0.4706*
H	*0.754*	*13.60*	*7.177*	*0.279*	*0.05304*	*0.3533*

г)

Рисунок 5 – Пример расчета дескрипторов QNA для муравьиной кислоты: (а) молекулярный граф, (б) матрица связности, (в) экспонента от матрицы связанности, (г) значения сродства к электрону, потенциалов ионизации, параметров A и B, и величин P и Q [85]

Значения P и Q коррелируют и имеют разный масштаб, поэтому необходима их нормализация:

$$P' = \frac{P - \mu_P}{\sigma_P}\ Q' = \frac{Q - \mu_Q}{\sigma_Q} \qquad (3),$$

$$U = \frac{P' + Q'}{\sqrt{2\left(1 + R_{PQ}\right)}}, V = \frac{P' - Q'}{\sqrt{2\left(1 - R_{PQ}\right)}} \qquad (4),$$

где μ_P и μ_Q – средние значения P и Q, σ_P и σ_Q – стандартные отклонения, R_{PQ} – коэффициент корреляции величин P и Q, соответственно.

Средние и стандартные отклонения рассчитываются при этом для конкретной выборки. Величины U и V являются ортонормальными (а, следовательно, не коррелируют), с нулевым средним и единичным стандартным отклонением.

Дескрипторы MNA и QNA были использованы нами в диссертационной работе ввиду соответствия указанным выше требованиям.

1.2.3 Способы построения зависимостей «структура-активность»

К настоящему времени разработаны разнообразные методы машинного обучения, которые широко используются в задачах био- и хемоинформатики. В задачах построения зависимостей «структура-активность» рассчитанные структурные дескрипторы являются независимыми переменными, а значения активности – зависимыми переменными. Построение таких зависимостей используется для решения задач кластеризации, ранжирования, снижения размерности и т.д. Для решения задач прогноза и проведения виртуального скрининга строят количественные (зависимая переменная является действительным числом) или классификационные модели (зависимая переменная принимает дискретное значение), что с точки зрения машинного обучения относится к задачам «обучения с учителем» [86].

Одним из исторически первых способов построения модели является применение множественной линейной регрессии:

$$y_k = a_0 + \sum_{i=1}^{m} a_i x_{ik} + \varepsilon.$$

Или, что то же самое, в векторном виде:

$$y_k = \boldsymbol{x}_k^T \boldsymbol{a} + \varepsilon\,,$$

где y_k – зависимая переменная, $\boldsymbol{x_k}$ – вектор-столбец m независимых переменных соединения k, $\boldsymbol{a}$ – вектор-столбец коэффициентов регрессии, ε – ошибка.

Для всех переменных выражение принимает вид:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a} + \varepsilon \qquad (5)\,,$$

где $\boldsymbol{y}$ – вектор-столбец n зависимых переменных, $\boldsymbol{X}$ – регрессионная матрица m столбцов (по независимым переменным) и n строк (по зависимым переменным).

Множественная линейная регрессия предполагает, что ошибка, обусловленная ненаблюдаемыми остатками, является случайной величиной с нормальным распределением и нулевым средним. Коэффициенты регрессии являются определяемыми величинами, их нахождение возможно, например, из выражения, получаемого методом максимального правдоподобия:

$$\boldsymbol{a} = \mathrm{ArgMin}\left[\sum_{k=1}^{n}\left(y_k - \sum_{i=0}^{m}(a_i x_{ik})\right)^2\right] \qquad (6)$$

Ограничения линейной регрессии хорошо изучены. К ним относятся: необходимость использования некоррелированных или слабо коррелированных независимых переменных; существенно большее число зависимых переменных по сравнению с определяемыми коэффициентами регрессии; невозможность заранее произвести выбор независимых переменных, которые вносят наибольший вклад в модель [87].

Последнее ограничение можно преодолеть, проводя пошаговую оценку значимости переменных с повторением процедуры определения коэффициентов регрессии на каждой итерации. Оценка значимости каждого из коэффициентов регрессии производится, например, на основе расчёта t-критерия Стьюдента:

$$t_i = \frac{a_i\sqrt{n}\sigma_i}{S_{\text{ост}}} \qquad (7),$$

где σ_i – стандартное отклонение соответствующей независимой переменной, $S_{\text{ост}}$ – стандартное отклонение остатков.

Полученные значения сравнивают со значениями квантилей распределения Стьюдента, и принимается решение о том, будет ли эта переменная использоваться в последующих итерациях. Следует отметить, что этот способ применим при отсутствии корреляции независимых переменных.

Для построения (Q)SAR моделей применяют также другие подходы (например, случайный лес и ИНС). В основе алгоритма случайного леса лежит использование ансамблей деревьев принятия решений [88]. На первом этапе производится построение большого числа деревьев. Из обучающей выборки посредством процедуры бутстрепа формируются случайные подвыборки, из всех независимых переменных извлекаются их случайные подмножества. Для каждой подвыборки и соответствующего подмножества случайных величин строится дерево принятия решений [89]. Для получения предсказаний результаты прогноза каждого дерева усредняются, при этом число деревьев выбирается таким образом,

чтобы минимизировать ошибку прогноза. Случайный лес является широко применяемым методом машинного обучения, в частности, из-за того, что в процессе обучения не происходит оценка параметров [90]. Этот факт, в свою очередь, влечет ограничения метода: невозможность определить область применимости модели, постановки четких критериев предсказательной способности и механистической интерпретации.

Искусственные нейронные сети в настоящее время являются одним из наиболее исследуемых и развиваемых методов. Существенным преимуществом ИНС является возможность построения сложных нелинейных моделей, с высокой точностью отражающих используемые для обучения данные. В основе работы ИНС лежит преобразование сигналов (кодирующих обрабатываемые данные) внутри нейронов и их передача между соседними нейронами. Нейроны располагаются в соответствующих им слоях, число которых может варьироваться. При создании ИНС важной задачей является создание архитектуры – топологии соединений между нейронами и слоями. Суть обучения заключается в минимизации выбираемой целевой функции (функции ошибки) и определении гиперпараметров, определяющих значения весов связей между нейронами. Для определения гиперпараметров используются стохастические методы, как, например, стохастический градиентный спуск, при котором значения целевой функции попадает в локальный минимум. Из-за случайного выбора параметров, по которым происходит минимизация целевой функции, следует неоднозначность локального минимума, в котором окажется целевая функция. Кроме того, ИНС не всегда могут быть применены релевантно для прогнозирования биологических данных из-за их особенностей (неявное представление химической структуры и экспериментальные ошибки значений активности) по сравнению с данными, для которых они хорошо себя зарекомендовали – изображения, звуки, тексты и т.д. [91].

Поскольку существует большое количество методов для построения количественных моделей, возникает необходимость выбора конкретного подхода.

В нашей работе для построения количественных моделей использовалась самосогласованная регрессия (SCR), благодаря её устойчивости к шуму и выбросам, однозначности решения, учету всех входящих переменных и интерпретируемости прогнозируемых значений.

SCR предполагает линейную зависимость, аналогичную формуле (5). Преимуществом SCR является строгая математическая обоснованность используемых оценок [92, 93]. Выбор значимых переменных происходит одновременно с процессом построения зависимостей. В задачах анализа зависимостей «структура-активность» по эмпирическим данным любая независимая переменная может иметь только ограниченное влияние на отклик. Это дает основание предполагать, что вероятность появления больших величин коэффициентов регрессии невелика, что можно выразить в виде априорного распределения вероятностей $p(\boldsymbol{a}|\mathbf{V})$ коэффициентов регрессии $\boldsymbol{a}$, где $\boldsymbol{V}$ – его параметры. Для нахождения апостериорного распределения вероятностей $p(\boldsymbol{a}|\boldsymbol{X},\boldsymbol{y},\mathbf{V})$ используется формула Байеса:

$$p(\boldsymbol{a}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{V}) = \frac{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})p(\boldsymbol{a}|\boldsymbol{V})}{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{V})} \qquad (8),$$

где $p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})$ – функция правдоподобия для экспериментальных значений $\boldsymbol{y}$ с матрицей независимых переменных $\mathbf{X}$ и коэффициентами регрессии $\boldsymbol{a}$, условная вероятность $p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{V})$ в знаменателе является суммой значений числителя $p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})p(\boldsymbol{a}|\boldsymbol{V})$ по всем значениям $\boldsymbol{a}$.

Коэффициенты регрессии $\boldsymbol{a}$ и параметры регуляризации $\boldsymbol{V}$ в SCR определяются самосогласованно:

$$\boldsymbol{a} = \underset{\boldsymbol{a}}{\mathrm{ArgMax}}\, p(\boldsymbol{a}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{V}) \text{ при условии } \boldsymbol{V} = \underset{\boldsymbol{V}}{\mathrm{ArgMax}}\, p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{V})$$

Если распределения $p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})$ и $p(\boldsymbol{a}|\boldsymbol{V})$ нормальные, то и апостериорные распределения $p(\boldsymbol{a}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{V})$ и $p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{V})$ тоже нормальные. Таким образом, задача поиска наилучших коэффициентов регрессии может быть решена с применением метода наименьших квадратов с регуляризацией, что позволяет разработать эффективные алгоритмы построения моделей (Q)SAR методом SCR.

При нормально распределенных ошибках по методу максимального правдоподобия получаем:

$$P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a}) = Exp\left(\frac{n}{2}\sum_{i}^{n}\mathrm{Ln}\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{a})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{a})\right)$$

При нормальном априорном распределении коэффициентов регрессии:

$$P(\boldsymbol{a}|\boldsymbol{V}) = Exp\left(\frac{1}{2}tr\left(\mathrm{Ln}\left(\frac{1}{2\pi}\boldsymbol{V}\right)\right) - \frac{1}{2}\boldsymbol{a}^T\boldsymbol{V}\boldsymbol{a}\right)$$

Функция правдоподобия выборки относительно параметров регуляризации:

$$P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{V}) = Exp\left(\frac{1}{2}tr\left(\mathrm{Ln}\left(\frac{1}{2\pi}(\boldsymbol{E}-\boldsymbol{X}^T\boldsymbol{T}\boldsymbol{X})\right)\right) - \frac{1}{2}\boldsymbol{y}^T(\boldsymbol{E}-\boldsymbol{X}^T\boldsymbol{T}\boldsymbol{X})\boldsymbol{y}\right)$$

Тогда итоговая апостериорная плотность вероятности:

$$p(\boldsymbol{a}|\mathbf{X},\boldsymbol{y},\boldsymbol{V}) = Exp\left(\frac{1}{2}tr\left(\mathrm{Ln}\left(\frac{1}{2\pi\sigma^2}\boldsymbol{T}^{-1}\right)\right) - \frac{1}{2\sigma^2}(\boldsymbol{a}-\boldsymbol{T}\boldsymbol{X}^T\boldsymbol{y})^T\boldsymbol{T}^{-1}(\boldsymbol{a}-\boldsymbol{T}\boldsymbol{X}^T\boldsymbol{y})\right) \quad (9),$$

$\boldsymbol{T} = (\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{V})^{-1}$ является апостериорной матрицей ковариации коэффициентов регрессии.

Таким образом, нахождение коэффициентов регрессии является задачей минимизации, аналогично формуле (6) для линейной регрессии:

$$\boldsymbol{a} = \mathrm{ArgMin}\left[\sum_{k=1}^{n}\left(y_k - \sum_{i=0}^{m}(a_i x_{ik})\right)^2 + \sigma^2\sum_{i=1}^{m}(v_i {a_i}^2)\right],$$

где n – число объектов в обучающей выборке, y_k – наблюдаемое значение k-ой зависимой переменной, m – число независимых переменных, x_{ik} – значение i-ой независимой переменной k-ого объекта обучающей выборки, v_i – значение i-го параметра регуляризации.

Согласно формуле (9) максимум апостериорной плотности вероятности $p(\boldsymbol{a}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{V})$ достигается при:

$$\boldsymbol{a} = \boldsymbol{T}\boldsymbol{X}^T\boldsymbol{y}$$

При диагональной матрице параметров их взаимосвязь с коэффициентами регрессии выглядит следующим образом:

$$V_{ii}({a_i}^2 + \sigma^2 T_{ii}) = 1, i = 1, \dots, m$$

Из-за нелинейной связи между $\boldsymbol{a}$ и $\boldsymbol{V}$ в SCR, соответствующие уравнения решаются итеративными методами. При этом, несмотря на то, что в процессе формирования SCR модели изначально рассматриваются все регрессоры $\boldsymbol{X}$, лишь часть из них входит в финальную модель.

Величина σ^2 также может быть выведена из принципа максимального правдоподобия и имеет следующую самосогласованную оценку:

$$\sigma^2 = \frac{(\boldsymbol{y} - \mathbf{X}\boldsymbol{a})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{a})}{n - tr(\mathbf{X}\mathbf{T}\mathbf{X}^T)}.$$

Для построения классификационных моделей также может быть использован ряд методов, включая упомянутый выше случайный лес и ИНС. Большинство задач бинарной классификации соответствует схеме испытаний Бернулли. Для случая двух классов функция правдоподобия выборки имеет следующий вид:

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{a}) = Exp\left(\sum_{k=1}^{n}\left(y_k LnP_k + (1 - y_k)Ln(1 - P_k)\right)\right) \qquad (10),$$

где y_k – соответствующие дискретные значения зависимой переменной 0 или 1, $P_k = P(\boldsymbol{x}_k, \boldsymbol{a})$ – вероятность положительного исхода (вероятность того, что значение зависимой переменной равно 1) k-ого объекта выборки, который представлен вектором дескрипторов $\boldsymbol{x}_k$. Это выражение непосредственно соответствует логистической регрессии. Тем не менее, логистическая регрессия очень чувствительна к сбалансированности выборок, большому числу независимых переменных и аналогично линейной регрессии не имеет встроенных математических инструментов для выбора переменных. Подход, использующий логистическую регрессию, может быть улучшен при добавлении различных ограничений на значения коэффициентов регрессии [94]. Проблему с несбалансированностью выборок обычно обходят, сокращая количество примеров с одним из значений зависимой переменной [95].

Другим подходом к классификации является применение метода опорных векторов (SVM). SVM представляет собой универсальный классификатор, основанный на теории статистического обучения Вапника-Червоненкиса. Суть

метода заключается в нахождении оптимальной гиперплоскости, позволяющей разделить два класса точек в пространстве. Итогом применения SVM является линейное выражение вида:

$$\boldsymbol{w}^T\boldsymbol{x} + b = 0,$$

где $\boldsymbol{w}$ – вектор-столбец весов модели, b – свободный коэффициент.

Оптимальная гиперплоскость может быть найдена при минимизации нормы вектора $\boldsymbol{w}$ [96] при условии:

$$\boldsymbol{w}^T\boldsymbol{x} + b \geq +1, y_i = +1$$
$$\boldsymbol{w}^T\boldsymbol{x} + b \leq -1, y_i = -1$$

К настоящему времени создано много различных методов бинарной классификации, однако, в применяемых подходах остается актуальной проблема несбалансированности выборок [97] и задача отбора наиболее значимых переменных [98]. Другим недостатком разработанных ранее подходов является низкая обобщающая способность прогноза: при получении хороших характеристик моделей для обучающей выборки (точность модели), хорошие характеристики для внешней тестовой выборки (предсказательная способность модели), не гарантированы. В задачах хемоинформатики по прогнозированию биологической активности или токсичности применяются подходы, основанные на модифицированных под определенные условия логистической регрессии (LR) и SVM [94]. Будучи эффективными в случае небольшого числа переменных, эти подходы требуют дополнительных манипуляций для выбора значимых независимых переменных в многомерных случаях. Так, в SVM нет встроенного аппарата отбора таких переменных и возможен пошаговый вариант с отбором по формуле (7) или иные подходы [99].

В нашей работе представлены разработка и применение классификационных подходов, что позволило преодолеть описанные выше недостатки существующих классификационных методов.

1.2.4 Оценка качества (Q)SAR моделей

Для оценки качества получаемых регрессионных и классификационных моделей применяются различные критерии, включая внутренние критерии согласия и внешние, такие как предсказательная способность.

Основным критерием для оценки согласия регрессионных моделей является коэффициент детерминации:

$$R^2 = 1 - \frac{\sum_{k=1}^{n}(y_k - \hat{y}_k)^2}{\sum_{k=1}^{n}(y_k - \bar{y})^2} \qquad (11),$$

где $\hat{y}_k$ – прогнозируемые значения зависимой переменной, $\bar{y}$ – среднее значение по обучающей выборке зависимой переменной. Значение R^2 показывает долю описанной дисперсии наблюдаемых значений по отношению к наблюдаемой дисперсии. Это значение также является показателем стабильности расчетных значений коэффициентов регрессии.

Оценка прогностической способности регрессионной модели может быть осуществлена с применением скользящего контроля или с оценкой значения коэффициента детерминации на тестовой внешней выборке.

Скользящий контроль подразумевает разбиение обучающей выборки на несколько частей, одна из которой используется в качестве тестовой, а все остальные используются для обучения. Аналогом коэффициента детерминации при таком подходе является коэффициент детерминации при кросс-валидации Q^2, рассчитываемый аналогично R^2. Помимо различных коэффициентов детерминации для оценки прогностической способности оценивается среднеквадратичное отклонение прогноза:

$$RMSD = \sqrt{\frac{\sum_{k=1}^{n}(y_k - \hat{y}_k)^2}{n}} \qquad (12)$$

При наличии независимой тестовой выборки рассчитывается соответствующий коэффициент детерминации $R_{\text{тест}}^2$ для тестовых данных.

Для классификационных моделей существует несколько основных критериев согласия. Расчет данных критериев основан на четырех основополагающих

величинах: число истинно положительных примеров (TP), число ложно положительных примеров (FP), число истинно отрицательных примеров (TN) и число ложно отрицательных примеров (FN). Наиболее часто используют следующие оценки:

Чувствительность:

$$Sensitivity = TPR = \frac{TP}{TP + FN} \qquad (13)$$

Специфичность:

$$Specificity = 1 - FPR = \frac{TN}{TN + FP} \qquad (14)$$

Сбалансированная точность:

$$BA = \frac{Sensitivity + Specificity}{2} \qquad (15)$$

Для более наглядной интерпретации и сравнения различных моделей строится ROC кривая в координатах TPR и FPR. Численным же выражением качества, которое следует из построения такой кривой, является площадь под ней (AUC).

Оценка прогностической способности классификационной модели может быть произведена с применением этих же величин, как при скользящем контроле, так и на основе прогноза для внешней выборки.

С целью сравнения подходов по возможности снижения размерности (отбор наиболее значимых независимых переменных) не всегда стоит учитывать лишь количество используемых в модели переменных. Вместо этого, можно применить понятие эффективной размерности V – величины, полученной при некотором преобразовании исходного пространства без потери свойств. V всегда меньше количества независимых переменных, использованных для построения модели. V может быть рассчитано различными способами, в зависимости от выбранного подхода.

1.3 Актуальность применения анализа «структура-активность» к поиску ингибиторов мишеней ВИЧ-1 и SARS-CoV-2

Разработка новых противовирусных препаратов для терапии инфекционных заболеваний является актуальной потому, что во многих случаях существующие препараты:

- не приводят к полному излечению;
- обладают серьезными побочными эффектами;
- их применение сопряжено с положительным отбором резистентных штаммов.

Проблема имеет особое значение по отношению к ВИЧ инфекции и в случае пандемии COVID-19.

1.3.1 Обзор мишеней и ингибиторов ВИЧ

К вирусам иммунодефицита человека (ВИЧ) относят два вида вирусов, относящихся к роду Lentivirus: ВИЧ-1 и ВИЧ-2. Эти вирусы вызывают тяжелое поражение иммунной системы, которое при отсутствии или неэффективности терапии приводит к синдрому приобретенного иммунодефицита человека (СПИД). По оценкам специалистов на 2021 год, во всем мире с диагнозом «ВИЧ-инфекция» живут от 33,9 до 43,8 миллионов человек; до 2 миллионов пациентов с впервые установленным диагнозом «ВИЧ-инфекция» и до 860 тысяч погибших от осложнений СПИД [100]. Ситуация с ВИЧ-инфекцией в Российской Федерации существенно хуже, чем в странах Европы, Азии и Америки. Данные для сравнения приведены в таблице 1 [101].

При этом исследователи отмечают серьезный характер поражения трудоспособного населения и растущую смертность от ВИЧ-инфекции в РФ, проявляющийся как в увеличении числа заболевших, так и в уменьшении активных лет жизни [102].

Таблица 1 – Эпидемиологическая и демографическая статистика по ВИЧ-инфекции в некоторых странах мира на 2021 год [103]

Страна	***Число случаев заражения (на 100 000 населения)***	***Годы жизни, скорректированные по нетрудоспособности (на 100 000 населения)***	***Общее число случаев***
Германия	*94*	*27*	*80 000*
Индия	*131*	*187*	*1 826 000*
Великобритания	*196*	*30*	*132 000*
США	*531*	*127*	*1 743 000*
РФ	***776***	***730***	***1 138 000***
Танзания	*2650*	*2715*	*1 503 000*

Человек является источником и резервуаром ВИЧ-инфекции на всех стадиях заболевания. Максимальная концентрация вирионов наблюдается в крови и других биологических жидкостях, а заражение происходит через прямой контакт с ними [103]. ВИЧ-1 распространен повсеместно в отличие от ВИЧ-2. ВИЧ-2 распространен в странах Западной Африки, Франции и Испании и имеет меньшую вирусную нагрузку и скорость развития клинических признаков [104].

Вирус иммунодефицита человека заражает любые клетки, на поверхности которых есть белок CD4: Т-хелперы, макрофаги, дендритные клетки - с помощью своего комплекса гликопротеинов gp-120-gp41 и корецепторов CCR5 и/или CXCR4 на мембране клетки-хозяина. В первые дни болезни вирус заражает почти все CD4+ клетки хозяина, в результате чего большинство из них погибает; в выживших клетках вирус переходит в латентное состояние, накапливает мутации, приобретает способность заражать широкий спектр клеток-хозяина и, в итоге, количество иммунных клеток становится критически малым, и организм погибает от оппортунистических инфекций и их осложнений [105].

Жизненный цикл ВИЧ условно разделяют на следующие этапы: прикрепление, слияние, обратная транскрипция, транспорт в ядро, интеграция,

транскрипция, транспорт из ядра, трансляция, сборка, высвобождение, созревание [107]. Однако не каждый этап жизненного цикла следует ингибировать, так как на некоторых этапах вирус использует ферменты хозяина, ингибирование которых может повлечь за собой нежелательные эффекты. Существуют пять групп мишеней, на которые воздействуют препараты, одобренные Управлением по санитарному надзору за качеством пищевых продуктов и медикаментов США (FDA) для применения в клинической практике:

- ингибиторы прикрепления/антагонисты ко-рецепторов;
- ингибиторы слияния;
- ингибиторы обратной транскриптазы (нуклеозидные, ненуклеозидные);
- ингибиторы протеазы;
- ингибиторы интегразы.

Для терапии рекомендовано использовать как минимум комбинацию из трех препаратов, воздействующих на две мишени [107]. Подобный подход называется высокоактивной антиретровирусной терапией (ВААРТ) и позволяет значительно повысить качество жизни пациентов. Эффективные схемы лечения содержатся в клинических рекомендациях, и лечащие врачи принимают решение о тактике лечения, исходя из клинической ситуации в соответствии с этими рекомендациями. Однако, в нынешних условиях невозможно полностью вывести вирус из организма, и инфицированные пациенты вынуждены принимать препараты пожизненно. На рисунке 6 схематично представлен жизненный цикл ВИЧ и мишени, на которые действуют существующие препараты.

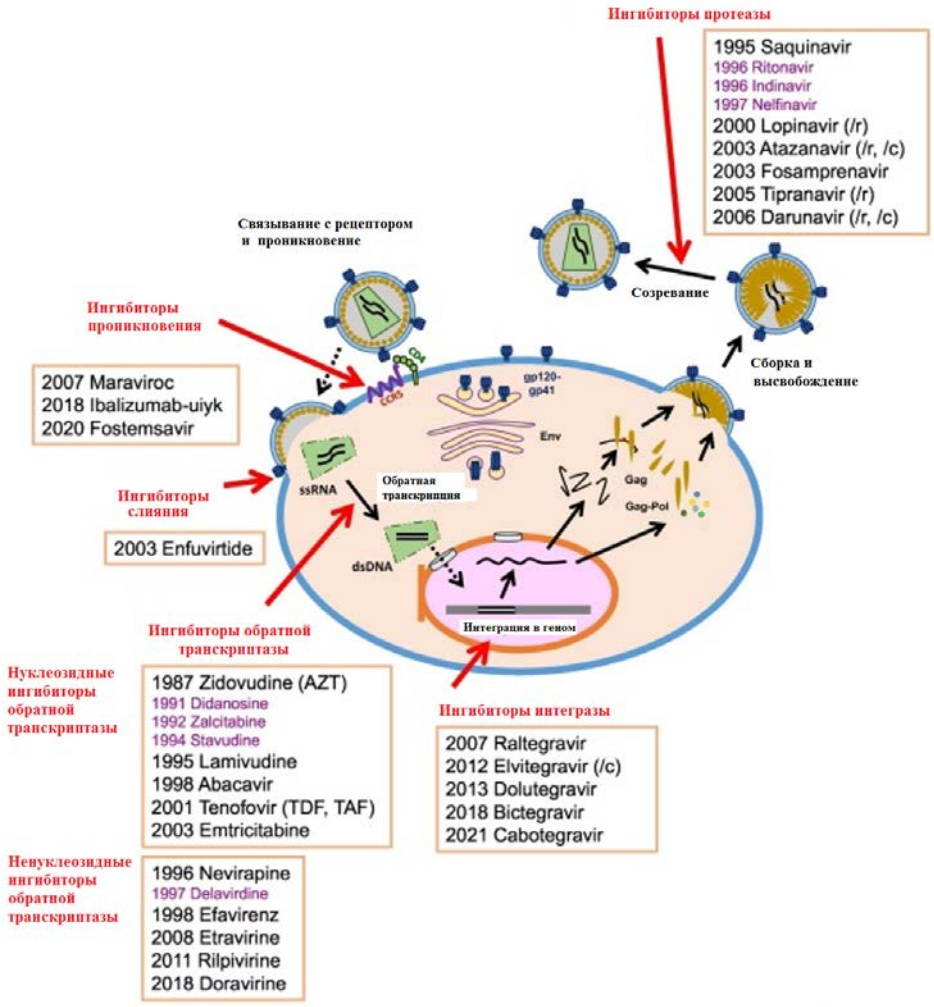


Рисунок 6 – Препараты против ВИЧ и мишени для их ингибирования в жизненном цикле вируса (адаптировано из [108])

Первыми были созданы ингибиторы обратной транскриптазы ВИЧ-1. Структура, функция, механизмы ингибирования и появление резистентности к ингибиторам обратной транскриптазы приведены в работах [109, 110]. Обратная транскриптаза (ОТ) – гетеродимер, состоящий из малой субъединицы (p51), несущей структурную функцию, и большой субъединицы (p66), обладающей каталитической активностью. ОТ обладает несколькими типами каталитической активности: РНК-зависимый ДНК синтез, ДНК-зависимый ДНК синтез, РНКаза H, трансфер цепи, синтез с замещением цепи. Для полимеразной активности ключевыми являются три остатка аспарагиновой кислоты и два катиона магния. Реакция начинается со связывания праймера и матрицы, при этом конформация

каталитического центра изменяется с закрытой на открытую. Праймер 3’-концом связывается ковалентно с остатком аспартата. Другой остаток аспартата связывает дезокситрифосфатнуклеотид. Далее, активный центр переходит в закрытое состояние, выравнивая 3’-ОН конец праймера и α-фосфат дезокситрифосфатнуклеотида. После этого происходит связывание нуклеотида с высвобождением пирофосфата, открыванием каталитического центра и транслокацией цепи.

За активность РНКазы Н (способность расщеплять РНК в комплексе с ДНК) отвечают остатки аспартата, остаток глутамата и 2 катиона магния. Реакция протекает по механизму нуклеофильного замещения с атакой на фосфатную группу, связывающей нуклеотиды в цепь. В настоящее время, нет одобренных для применения в клинике ингибиторов РНКазы Н, но ведутся активные доклинические испытания перспективных молекул [111, 112]. Механизм ингибирования «кандидатов» заключается в хелатировании ионов магния в активном центре и прочном нековалентном связывании с активными остатками аспартата и глутамата за счет образования водородных связей [111].

Ингибиторы ОТ по химическому строению делятся на две группы: аналоги нуклеозидов и ненуклеозидные ингибиторы. Нуклеозидные ингибиторы ОТ – первые препараты от ВИЧ-инфекции. К ним относят зидовудин, абакавир, тенофовир, эмтрицитабин. В организме они фосфорилируются до 5’-трифосфонуклеотидов, которые выступают в качестве альтернативного субстрата для полимеразы, но в их составе нет 3’-ОН группы, поэтому, когда такой нуклеотид встраивается в цепь, репликация цепи прерывается. Ненуклеозидные ингибиторы, к которым относятся невирапин, эфавиренц, этравирин, рилпивирин и доравирин, действуют иначе. Во-первых, они нарушают мобильность «большого пальца» и «пальцев», мотивов, отвечающих за открытие и закрытие полимеразного центра. Во-вторых, они изменяют конформацию каталитического центра. В-третьих, такие ингибиторы меняют сеть межмолекулярных взаимодействий в гетеродуплексе РНК-ДНК путем репозиционирования участников [113].

Позже были разработаны ингибиторы протеазы ВИЧ-1. Протеаза – фермент, ответственный за расщепление исходного первичного транскрипта на отдельные функциональные белки. Зрелая протеаза – гомодимер, каждая субъединица которого состоит из 99 аминокислот и имеет один каталитический центр, состоящий из триплета аспартат-серин/треонин-глицин [114]. Ключевыми в каталитическом акте являются остатки аспартата: один выступает в качестве донора протона, другой в качестве акцептора протона. Реакция протекает в 4 стадии: асимметричное связывание субстрата и последующее гидратация карбонила разрезаемой связи молекулой воды в активном центре (стадия 1); конформационная переориентация азота, приводящая к гош-конформации гидратированной пептидной связи (стадия 2); одновременный обмен протонами, при котором одна из каталитических аспарагиновых кислот отдает протон к неподелённой паре азота, а другой заряженный аспартат принимает один протон от гидроксила диола (стадия 3); разрыв пептидной связи и восстановление исходного протонированного состояния активного сайта (стадия 4) [115]. К ингибиторам протеазы ВИЧ-1 относят саквинавир, лопинавир, атазанавир, дарунавир. Их механизм действия заключается в имитации переходного состояния субстрата: гидроксильная группа ингибитора взаимодействует с карбоксильной группой остатков активного центра протеазы, D25 и D25', посредством водородных связей. Остатки протеазы ВИЧ-1, контактирующие с ингибитором, относительно консервативны, включая G27, D29, D30 и G48, но накопление мутаций изменяет структуру протеазы ВИЧ-1 и приводит к возникновению резистентности и снижению эффективности лечения [116].

Интеграза – фермент, отвечающий за встраивание генома вируса в геном клетки-хозяина. Он состоит из 288 аминокислот, поделенных на три домена: N-концевой, каталитическое ядро и С-концевой [117]. N-концевой домен представляет из себя компактный трехспиральный пучок, стабилизированный координацией иона Zn2+ инвариантными остатками гистидина и цистеина мотива HHCC. Каталитическое ядро содержит активный центр и имеет общие

характеристики α/β-кратности с семейством различных нуклеаз, включая РНКазу H, резольвазы соединения Холлидея и ДНК-транспозазы. С-концевой домен характеризуется SH3-подобной β-бочкообразной складкой [118]. Интеграция протекает в два этапа. На первом этапе, процессинге 3'-конца, два нуклеотида отщепляются от каждого 3'-конца ДНК, которая изначально имеет тупые концы. Это обнажает 3'-концевые консервативные динуклеотиды, которые образуют соединение между вирусной и ДНК-мишенью при интеграции. На следующем этапе, переносе цепи ДНК, гидроксильные группы на 3'-концах вирусной ДНК атакуют пару фосфодиэфирных связей в ДНК-мишени, которые, в случае ВИЧ, разделены пятью нуклеотидами на каждой цепи ДНК. В результате образуется интермедиат интеграции, в котором 3'-концы вирусной ДНК ковалентно присоединены к ДНК-мишени. 3'-Концы ДНК-мишени и 5'-концы вирусной ДНК, два нуклеотида которой неспарены, не соединяются [119]. Далее с помощью клеточных ферментов идет репарация нарушений: удаление неспаренных нуклеотидов и лигирование цепей. Эти реакции протекают в специальном нуклеопротеиновом комплексе – интасоме. Механизм действия ингибиторов интегразы, к которым относят ралтегравир, элвитогравир, долутегравир, каботегравир, биктегравир, заключается в связывании с интасомой и препятствовании захвата цепей ДНК клетки-хозяина с ней.

Вирионы ВИЧ прикрепляются к мембране клетки хозяина, связываясь своим белковым комплексом gp41-gp120 с CD4, а взаимодействие с CCR5 и CXCR4 запускает процесс слияния мембран и проникновения капсида внутрь цитоплазмы [120]. Существуют две группы препаратов, направленных на предотвращение заражения клеток вирионами. Первая группа – ингибиторы слияния. В этой группе, пока один одобренный для применения в клинике препарат – энфувиртид. Препарат представляет собой синтетический пептид, который соответствует последовательности из 36 аминокислот второго гептадного повтора gp41 ВИЧ-1. Энфувиртид конкурентно связывается с первым гептадным повтором gp41, предотвращая образование шести-спирального пучка, который имеет решающее

значение для слияния мембран [121]. Вторая группа – антагонисты хемокиновых рецепторов CCR5 и CXCR4. Вирионы отличаются тропизмом к рецепторам: у каких-то больше сродство к CCR5, у каких-то к CXCR4, у каких-то к их двойному комплексу, причем пока обнаружено небольшое количество молекул, которые блокируют оба рецептора, несмотря на их структурное сходство [122]. FDA одобрены 3 препарата: маравирок, ибализумаб, фостемсавир. Маравирок – малая молекула, первая в своем классе, является антагонистом CCR5 [124]. Ибализумаб – моноклональное антитело ко второму домену CD4, которое не оказывает иммуносупрессивного эффекта, так как не влияет на взаимодействие CD4 с MHC второго класса [124]. Фостемсавир связывается с gp120, поэтому активен против CCR5-тропных и CXCR4-тропных вирусов [125].

Поиск новых ингибиторов ферментов ВИЧ-1 остается актуальной задачей и для ее решения разрабатываются и применяются SBDD и LBDD подходы, в частности QSAR модели, описанные в современных работах [126, 127].

1.3.2 Обзор мишеней и ингибиторов COVID-19

Пандемия болезни, называемой COVID-19, началась в декабре 2019 года со случаев пневмонии неизвестного происхождения в городе Ухань, Китай. Было выяснено, что возбудителем инфекции является вирус SARS-CoV-2, относящийся к роду *Betacoronavirus* [128]. На 31 января 2023 года выявлено более 670 миллионов случаев заражения и более 6,5 миллиона смертельных исходов по всему миру [129]. На территории РФ на 30 января 2023 года выявлено почти 22 миллиона случаев заражения, и 395 022 случая оказались со смертельным исходом [130]. На февраль 2023 года интенсивность заражения не такая высокая, как в начале пандемии. Динамика заболеваемости волнообразная, и пик волны приходится на преобладание какого-либо варианта в популяции, который получил эволюционное преимущество [131, 132]. Как правило, это связано с появлением мутаций, направленных на увеличение трансмиссивности и на уход от иммунного ответа [133]. Это приводит к снижению эффективности вакцин и различной симптоматической картине, вызываемой разными штаммами [133–135].

Вирусные частицы SARS-CoV-2 обычно имеют сферическую форму и заключены в липидную мембрану эндоплазматического ретикулума клетки-хозяина [136]. Внутренняя оболочка вируса (капсид) состоит из оболочечных и мембранных белков, образующих структурные элементы, с вкраплениями булавовидных шиповидных белков, выступающих из поверхности, что придает ему характерный коронообразный вид и позволяет отнести данный вирус к семейству коронавирусов [136, 137]. Эта оболочка достаточно стабильна, чтобы обеспечить вирусным частицам структуру и защиту геномного материала вне клеток-хозяев. В то же время, она все еще способна разрушаться и высвобождать геномный материал внутри клеток-хозяев. В центре каждой вирусной частицы находится материал генома, который эффективно сворачивается в компактные структуры, связывая белок нуклеокапсида, который в конечном итоге образует белковое покрытие вокруг РНК вируса [138].

Жизненный цикл SARS-CoV-2 подразделяется на следующие этапы: прикрепление к мембране клетки, слияние, репликация и трансляция, сборка вирионов, выход из клетки [139]. Несмотря на то, что срок разработки лекарства от идеи до готового препарата составляет 10-12 лет, фармацевтические компании получают временные разрешения для вывода на рынок препаратов для лечения COVID-19 из-за экстренной ситуации. Их регистрация во всех странах мирах идет по нестандартным, ускоренным, «чрезвычайным» протоколам.

Поиск мишеней для противовирусной терапии проводят с учетом имеющейся информации о жизненном цикле вируса. Противовирусные препараты можно разделить на классы по механизму их действия. Широкий спектр мишеней для лекарств позволяет использовать различные препараты, которые подразделяются на несколько классов. Для ингибиторов белков SARS-CoV-2 можно выделить три основные группы:

- Ингибиторы слияния;
- Ингибиторы главной и папаин-подобной протеазы;
- Ингибиторы полимеразы.

Инфекционный процесс начинается с процесса слияния, обусловленного взаимодействием между шиповидным белком (Spike) вируса и ангиотензин-превращающим ферментом (АПФ) клеток-хозяев. Роль ингибиторов проникновения SARS-CoV-2 в качестве терапевтических средств для лечения COVID-19 заключается в блокировании/предотвращении проникновения вируса в клетку-хозяина. Эта группа препаратов может включать, помимо прочего, реконвалесцентную плазму, моноклональные антитела и небольшие молекулы или пептиды, нацеленные на факторы хозяина, участвующие в проникновении [140, 141]. Кроме того, к ингибиторам проникновения относятся вещества, блокирующие проникновение вирусной РНК в цитоплазму. К ним относятся препараты, которые действуют путем предотвращения распознавания, прикрепления, ингибировании слияния, в частности, на основе применения растворимых АПФ [140–143]. К таким препаратам можно отнести, например, ингибиторы взаимодействия Spike-АПФ (умифеновир), ингибиторы трансмембранной сериновой протеазы TMPRSS2 (нафамостат), препараты, изменяющие эндосомальный pH (хлорохин и гидроксихлорохин) [144].

У SARS-CoV-2 есть две протеазы, которые являются неотъемлемой частью жизненного цикла вируса: главная (Mpro или 3CLpro) и папаин-подобная протеаза (PLpro) [142, 145]. Главная протеаза расщепляет 75% сайтов расщепления двух полипротеинов (pp1a и pp1ab) на функциональные белки, в то время как PLpro расщепляет 25% сайтов расщепления и несколько белков-хозяев [145]. Главная протеаза в зрелой форме – димер, хотя мономеры, состоящие из трёх доменов, также обладают, пусть и меньшей, каталитической активностью [146, 147]. Активный сайт, расположенный на пересечении первого и второго доменов, образован пятью гидрофобными карманами, в которых главную роль играют остатки H41, C145 [148]. Папаин-подобная протеаза – фермент, состоящий из 315 аминокислот, богат цистеином и имеет структурный мотив «большой палец – ладонь – пальцы» [149]. Активный сайт PLpro содержит в себе каноничную каталитическую триаду цистеиновой протеазы – C111, H272, D286 [150]. Хотя обе

протеазы играют решающую роль в высвобождении и созревании функциональных белков для репликации вируса, разрабатываются преимущественно ингибиторы Mpro. Такое предпочтение связано с тем, что PLpro также обладает деубиквитинирующей активностью и распознает С-концевую последовательность убиквитина. Это затрудняет поиск препаратов против PLpro, поскольку ингибиторы PLpro, полученные из субстрата, будут не только ингибировать вирусную PLpro, но также ингибировать деубиквитиназы клеток-хозяев [145, 149].

За некоторыми незначительными исключениями, большинству РНК-вирусов требуется фермент РНК-зависимая РНК-полимераза для репликации и транскрипции вирусного геномного материала, что делает его важным компонентом для их выживания [151]. Помимо решающей роли в жизненном цикле вируса, структурные и функциональные особенности полимеразы высококонсервативны среди различных групп РНК-содержащих вирусов, при этом активный сайт фермента является наиболее консервативной и доступной областью, что позволяет рассматривать ее в качестве эффективной терапевтической мишени [142, 151]. SARS-CoV-2 использует полимеразу, состоящую из трех субъединиц nsp (nsp7, 8 и 12), для осуществления транскрипции своего РНК-генома, который относительно сложнее, чем у других вирусов, из-за экзорибонуклеазной активности и большого размера генома [152]. Благодаря высокой консервативности структурных и функциональных характеристик, возникает возможность применения существующих ингибиторов полимеразы (нуклеозидных аналогов), которые могут быть перепрофилированы, например, рибавирин, который использовался во время распространения инфекции SARS-CoV, и ремдесивир, одобренный для экстренного применения при терапии COVID-19 [143].

После начала пандемии COVID-19 (Q)SAR модели разрабатывались и применялись для виртуального скрининга, направленного на поиск ингибиторов мишеней вируса [153–155].

Несмотря на достигнутые результаты антиретровирусной терапии и достижения в области исследований коронавирусов, имеется необходимость в

создании как новых препаратов для ВААРТ из-за возникновения резистентных штаммов, и в разработке подходов для оперативного ответа на новые биогенные угрозы, такие как пандемия COVID-19. Темпы и массовость исследования активности новых соединений и, как следствие, накопление гетерогенной информации выдвигает новые требования к методам, которые применяются для ее анализа. Развитие новых подходов позволит повысить качество получаемых с использованием гетерогенных данных зависимостей «структура-активность» на основе методов машинного обучения.

ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ

2.1 Формирование обучающих и тестовых выборок

2.1.1 Искусственно сгенерированные данные

Для отладки работы разрабатываемых методов нами были сгенерированы выборки с искусственными данными, которые имитируют определенные взаимосвязи «структура-активность», как линейные комбинации:

$$y_i = a_0 + \sum_j^m a_j x_{ij} + \varepsilon_i \qquad (16),$$

где y_i – наблюдаемые величины активности (зависимая переменная), x_{ij} – «дескрипторы», описывающие химическую структуру (m независимых переменных), a_j – коэффициенты регрессии, ε_i – ошибки, независимые случайные величины с нормальным распределением $N(0, \sigma)$.

Генерация этих данных осуществлялась следующим образом:

- все коэффициенты регрессии a_j принимались равными 1, a_0 равным 0;
- регрессионная матрица $\boldsymbol{X}$ для каждого набора данных составлялась из независимых переменных, сгенерированных согласно $N(0, 1)$ с использованием генератора равномерно распределенных случайных чисел на основе простых чисел Мерсенна [156] и их преобразование к нормальному распределению по методу Бокса-Мюллера [157];
- величины $\boldsymbol{y}$ рассчитывались по приведенной выше формуле (16), ошибки сгенерированы аналогично генерации независимых переменных с учетом соответствующего стандартного отклонения;
- зависимые переменные относились к двум классам (активные и неактивные) в зависимости от значений y_i и выбираемых пороговых величин p, соответствующих каждому набору данных. При $y_i \geq p$ зависимая переменная относилась к классу активных и наоборот, к классу неактивных при $y_i < p$. Путем варьирования пороговых значений были получены

выборки (пример в Приложении 2) с разным уровнем сбалансированности (варьировалось соотношение между количеством активных и неактивных «соединений»).

Подобные выборки генерировались для имитации случаев, которые могут возникнуть при расчете дескрипторов для экспериментальных данных различного происхождения.

2.1.2 Выборки данных Tox21

Для валидации различных методов анализа связи «структура-активность» и их сравнения друг с другом широко используются массивы данных Tox21 [158], которые являются несбалансированными, представляя различные соотношения «активных» и «неактивных» (токсичных и нетоксичных) соединений (рис. 7). Из БД Tox21 были извлечены 10 тысяч записей, содержащих структурные формулы и данные о 38 различных видах токсичности химических соединений. После соответствующей обработки (перевод кодов SMILES в SDF формат, удаление недостоверных сведений и т.п.) было подготовлено 38 выборок для построения классификационных моделей (Приложение В).

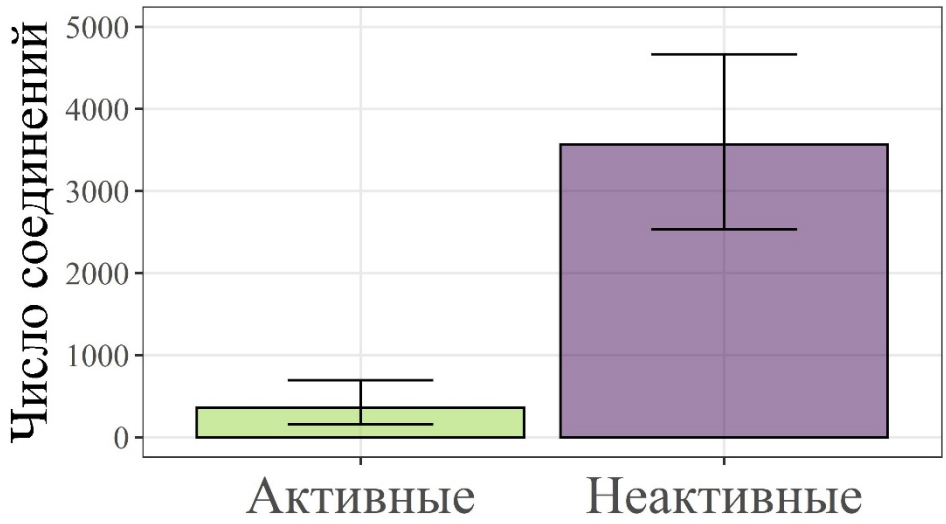


Рисунок 7 – Соотношение активных/неактивных веществ в 38 выборках из Tox21

2.1.3 Выборки ингибиторов ВИЧ-1

Сведения о структуре и ингибиторной к молекулярным мишеням ВИЧ-1 активности были нами извлечены из трех различных баз данных, обработаны и проанализированы. Для каждой из трех мишеней (протеаза, обратная транскриптаза и интеграза ВИЧ-1) и каждого информационного источника (NIAID, ChEMBL, Integrity) были сформированы отдельные выборки. Построение регрессионных и классификационных моделей осуществлялось с использованием попарных комбинаций выборок из различных информационных источников и объединённых выборок по каждой из мишеней.

NIAID ChemDB HIV [159] (Opportunistic Infection and Tuberculosis Therapeutics Database – NIAID DB) является свободно доступной БД, которая содержит информацию о структуре химических соединений и их фармакологическом действии на ферменты ВИЧ-1, ВИЧ-ассоциированные заболевания и сопутствующие инфекции. Для последующего анализа были извлечены сведения о 10 377, 7 604 и 8 936 ингибиторах интегразы, протеазы и обратной транскриптазы ВИЧ-1, соответственно.

ChEMBL [70] является свободно доступной БД, которая содержит информацию о структуре и биологической активности лекарственно-подобных соединений. Из ChEMBL нами были извлечены сведения о 2 283, 2 387 и 2 149 ингибиторах интегразы, протеазы и обратной транскриптазы ВИЧ-1, соответственно.

Integrity [160] – это коммерчески доступная БД компании Clarivate Analytics (в настоящее время распространяемая под названием Cortellis Drug Discovery Intelligence). Из этой базы данных были извлечены сведения о 563, 316 и 731 ингибиторах интегразы, протеазы и обратной транскриптазы ВИЧ-1, соответственно.

2.1.4 Выборки ингибиторов SARS-CoV-2

Сведения о структуре и ингибировании ферментов вируса SARS-CoV-2 были нами подготовлены на основе трех источников, пополняемых на протяжении пандемии COVID-19, и интегрированы в единую обучающую выборку.

PostEra Moonshot [161] – стартап, осуществляющий исследования потенциальных антикоронавирусных соединений и предоставляющий в открытом доступе экспериментальные данные об ингибиторах главной протеазы (3CLpro) SARS-CoV-2.

COVID-19 Open Data Portal [162] – предоставляет данные NCATS об исследованиях ингибиторов 3CLpro и РНК-зависимой РНК-полимеразы (RdRp) SARS-CoV-2.

Coronavirus Antiviral and Resistance Database – БД Стэнфордского университета [163], агрегирующая литературные данные об ингибиторах 3CLpro, RdRp и папаин-подобной протеазы (PLpro) SARS-CoV-2.

2.1.5 Предобработка выборок ингибиторов ВИЧ-1 и SARS-CoV-2

Собранные нами данные, как для ВИЧ-1, так и для SARS-CoV-2, обладают высокой степенью гетерогенности, характеризуемой разнообразием химических классов и механизмов действия ингибиторов, и различиями в экспериментальных методах определения значений активности.

Для формирования обучающих выборок были отобраны записи, содержащие значения полуэффективных концентраций ингибирования в молях на литр (IC_{50} (M)), значения IC_{50} переводили в $pIC_{50} = -\lg(IC_{50})$. Предобработка данных о структурах соединений и их активности проводилась нами перед этапом моделирования согласно современным требованиям (см. раздел 1.2.1).

Число структур в отдельных выборках, а также их пересечения друг с другом, представлены на рисунке 8.

Для оценки прогностической способности построенных количественных зависимостей «структура-активность» были подготовлены независимые

тестовые выборки с данными, не использованными при обучении. Так, для объединения выборок NIAID и ChEMBL использовались тестовые выборки из Integrity с исключением дубликатов, входящих в обучающую выборку; для объединения выборок NIAID и Integrity использовались тестовые выборки из ChEMBL. Соединения были классифицированы при пороге в 1 мкМ, как активные при IC_{50} < 1 мкМ и неактивные при IC_{50} ≥1 мкМ.

Были подготовлены обучающие выборки ингибиторов белков ВИЧ-1 путем объединения данных NIAID и ChEMBL, NIAID и Integrity, и всех доступных данных:

- выборка IN (NIAID и ChEMBL) – 3 987 структур ингибиторов интегразы;
- выборка IN (NIAID и Integrity) – 3 597 структур ингибиторов интегразы;
- выборка IN, 4 091структура ингибиторов интегразы, диапазон pIC_{50} от 1,75 до 10,15;
- тестовая выборка IN (ChEMBL) - 490 структур ингибиторов интегразы;
- тестовая выборка IN (Integrity) - 104 структуры ингибиторов интегразы;
- выборка PR (NIAID и ChEMBL) – 6 462 структуры ингибиторов протеазы;
- выборка PR (NIAID и Integrity) – 6 068 структур ингибиторов протеазы;
- выборка PR, 6 552 структуры ингибиторов протеазы, диапазон pIC_{50} от 2,00 до 13,85;
- тестовая выборка PR (ChEMBL) - 486 структур ингибиторов протеазы;
- тестовая выборка PR (Integrity) - 92 структуры ингибиторов протеазы;
- выборка RT (NIAID и ChEMBL) – 6 093 структуры ингибиторов обратной транскриптазы;
- выборка RT (NIAID и Integrity) – 5 894 структуры ингибиторов обратной транскриптазы;

- выборка RT, 6 309 структур ингибиторов обратной транскриптазы, диапазон pIC_{50} от 0,20 до 11,49;
- тестовая выборка RT (ChEMBL) - 415 структур ингибиторов обратной транскриптазы;
- тестовая выборка RT (Integrity) - 216 структур ингибиторов обратной транскриптазы;

Аналогичным образом были подготовлены выборки с качественными данными по активности для ингибиторов белков SARS-CoV-2:

- выборка 3CLpro: 6 230 структур, 447 активных;
- выборка PLpro: 106 структур, 68 активных;
- выборка RdRp: 2 632 структуры, 91 активная.

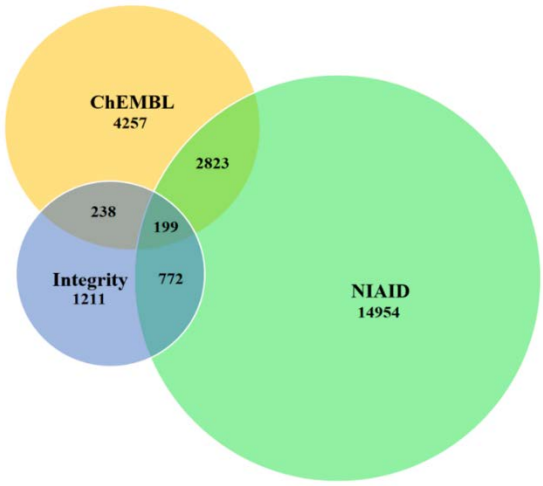


Рисунок 8 – Пересечение и общее число уникальных ингибиторов для трех мишеней ВИЧ-1 после процедуры предобработки

2.1.6 Формирование расширенных выборок ВИЧ-1, содержащих качественные данные об активности

Наряду с данными по количественным характеристикам IC_{50} ингибиторов ферментов ВИЧ-1 (>15 000 структур), имеется большой массив данных о значениях

Ki (~8 000 структур), и качественной характеристики активности (активные/неактивные) экспериментально изученных соединений (~1 000 структур). Использование всей совокупности доступных данных позволяет существенно расширить охват анализируемого химического пространства за счет химических классов, не вошедших в исходную выборку, и может повысить предсказательную способность классификационных моделей. Поэтому выборки, содержащие информацию о значениях IC_{50} ингибиторов ВИЧ, были дополнены данными о константах связывания (Ki) и данными о неактивных соединениях. С целью построения классификационных моделей значения IC_{50} и Ki были переведены в классификационные величины по соответствующим порогам. Всего было подготовлено 6 комбинированных выборок: IN (+Ki) – 4 349 структур, IN (+Ki и неактивные) – 12 000 структур, PR (+Ki) – 8 545 структур, PR (+Ki и неактивные) – 12 000 структур, RT (+Ki) – 7 508 структур, RT (+Ki и неактивные) – 12 000 структур.

2.2 Молекулярные дескрипторы

Для описания структуры химических соединений в диссертационной работе использованы дескрипторы атомных окрестностей, которые основаны на таком представлении структуры молекул, которое включает только ковалентные связи и все атомы водорода в соответствии с валентностями и зарядами тяжелых атомов, и не включает типы связей. При построении классификационных и регрессионных моделей использовались дескрипторы количественных атомных окрестностей (Quantitative Neighborhoods of Atoms – QNA) [85], для сравнения химических пространств выборок – дескрипторы многоуровневых атомных окрестностей (Multilevel Neighborhoods of Atoms – MNA) [83].

2.2.1 Расчет QNA дескрипторов

Для расчета дескрипторов QNA применялись формулы (1-3). В формуле (3) не использовались статистические величины, относящиеся к самой обучающей

выборке. Средние значения и стандартные отклонения P и Q, использованные в диссертационной работе, были ранее рассчитаны для выборки из пятидесяти тысяч лекарственно подобных соединений, входящих в основную обучающую выборку программы PASS [84].

2.2.2 Использование QNA для разработки моделей

С целью получения независимых величин проводилась ортогонализация величин P' и Q' отличным от формулы (4) образом:

$$U = P', V = \alpha P' - \beta Q',$$

где величины α и β подбираются таким образом, чтобы $E[V^2] = 1$, $E[UV] = 0$: $\alpha = E[P'Q']/\sqrt{1 - E^2[P'Q']}$, $\beta = 1/\sqrt{1 - E^2[P'Q']}$.

Таким образом, ортонормированные значения U и V обладают следующими свойствами – среднее равно нулю, стандартное отклонение равно единице, сами величины не коррелируют.

Очевидно, что различные молекулы будут иметь различные по размеру наборы величин U и V. В задачах (Q)SAR обычным представлением выборки молекул является набор векторов одинакового размера. Векторы дескрипторов необходимой длины получали с использованием многочленов Чебышёва с учетом объема обучающей выборки.

Расчет соответствующего вектора представим в виде:

$$T_{uv}(P, Q) = \mathrm{Cos}\left(u * \mathrm{ArcCos}\big(\mathrm{TanH}(U)\big)\right) * \mathrm{Cos}\left(v * \mathrm{ArcCos}\big(\mathrm{TanH}(V)\big)\right),$$

где $u, v = 0,1,2 \ldots$ являются целыми числами и определяют порядок многочлена. Многочлены Чебышёва упорядочены по возрастанию сумм $u + v$. Для $u + v = 1$ это $T_{1,0}$, $T_{0,1}$; для $u + v = 2$ это $T_{2,0}$, $T_{1,1}$, $T_{0,2}$ и так далее.

2.3 Методы самосогласованной регрессии и классификации

2.3.1 Самосогласованная регрессия

Для разработки регрессионных моделей и прогнозирования количественных

значений активности использовалась самосогласованная регрессия (SCR), описанная в разделе 1.2.3.

2.3.2 Самосогласованная классификация

Хотя метод SCR был разработан для построения количественных моделей, он может быть использован также для решения задач классификации, если в классе положительных примеров положить $y = 1$, а в отрицательных примерах $y = 0$. Такой простой подход, реализованный ранее А. В. Захаровым в программе GUSAR [164], не учитывает, однако, отличие классификации от регрессии, поэтому для решения задач классификации на основе SCR нами разработаны два новых метода: логистический SCLC и экстремальный SCEC классификаторы.

SCLC следует из выражения (10), если:

$$P(\boldsymbol{x}, \boldsymbol{a}) = \left(1 + Exp(-\boldsymbol{x}^T \boldsymbol{a})\right)^{-1}$$

Обратный логарифм функции правдоподобия $-Ln\left(p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{a})\right)$ можно представить так:

$$-Ln\left(p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{a})\right) = \sum_{k=1}^{n} Ln\left(1 + Exp(-u_k \boldsymbol{e}_k^T \boldsymbol{X}\boldsymbol{a})\right),$$

$$u_k = 2y_k - 1, u_k = \pm 1, u_k^2 \equiv 1,$$

где $\boldsymbol{e}_k^T$ – единичный транспонированный вектор.

Однако, согласно формуле (8) в разделе 1.2.3, нам нужно выражение для обратного логарифма числителя формулы Байеса:

$$-Ln\left(p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{a})p(\boldsymbol{a}|\boldsymbol{V})\right) \qquad (17)$$

Для неверного прогноза, характеризуемого $u_k \boldsymbol{e}_k^T \boldsymbol{X}\boldsymbol{a} < 0$, функция «штрафа» $Ln\left(1 + Exp(-u_k \boldsymbol{e}_k^T \boldsymbol{X}\boldsymbol{a})\right)$ возрастает почти линейно, а для верных прогнозов при $u_k \boldsymbol{e}_k^T \boldsymbol{X}\boldsymbol{a} > 0$ – стремится к нулю.

Из-за нелинейности $\boldsymbol{p}(\boldsymbol{X}, \boldsymbol{a})$ задачу необходимо сформулировать относительно малых приращений $d\boldsymbol{a}$ коэффициентов регрессии с подстановкой $\boldsymbol{a} \to \boldsymbol{a} + d\boldsymbol{a}$. Тогда обратный логарифм (17) будет выражен как:

$$-Ln\big(p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})p(\boldsymbol{a}|\boldsymbol{V})\big)$$

$$= -\sum_{k=1}^{n}\big(y_k LnP_k + (1-y_k)Ln(1-P_k)\big) + \frac{1}{2}\boldsymbol{a}^T\boldsymbol{V}\boldsymbol{a} - \frac{1}{2}tr\big(Ln(\boldsymbol{V})\big)$$

$$+ d\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{p}(\boldsymbol{X},\boldsymbol{a}) + \boldsymbol{V}\boldsymbol{a} - \boldsymbol{X}^T\boldsymbol{y}) + \frac{1}{2}d\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \boldsymbol{V})d\boldsymbol{a} \qquad (18)$$

Для нахождения параметров приравняем производную выражения (18) к нулю и получим систему уравнений:

$$(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \boldsymbol{V})(\boldsymbol{a} + d\boldsymbol{a}) - \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{p}(\boldsymbol{X},\boldsymbol{a}) + \boldsymbol{W}\boldsymbol{X}\boldsymbol{a}) = \boldsymbol{0} \qquad (19),$$

где $\boldsymbol{p}(\boldsymbol{X},\boldsymbol{a})$ – вектор столбец вероятностей такой, что $\boldsymbol{e}_k^T\boldsymbol{p}(\boldsymbol{X},\boldsymbol{a}) = P(\boldsymbol{x}_k,\boldsymbol{a})$,

$\boldsymbol{W}$ – диагональная матрица весов входящих в обучающую выборку примеров

$$W_{kk} = P(\boldsymbol{x}_k,\boldsymbol{a})\big(1 - P(\boldsymbol{x}_k,\boldsymbol{a})\big) \qquad (20)$$

Как и в SCR коэффициенты регрессии связаны с параметрами регуляризации:

$$V_{ii}({a_i}^2 + T_{ii}) = 1, i = 1, \dots, m$$

SCEC может быть получен заменой логистической функции в SCLC на распределении экстремальных значений Гумбеля [165], чему соответствует обратный логарифм функции правдоподобия:

$$-Ln\big(p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})\big) = \sum_{k=1}^{n} Exp\big(-u_k\boldsymbol{e}_k^{\mathrm{T}}\boldsymbol{X}\boldsymbol{a}\big),$$

$$u_k = 2y_k - 1, u_k = \pm 1, u_k^2 \equiv 1.$$

В этом случае штраф для ошибочных предсказаний будет экспоненциально возрастать. Аналогично получаем выражение с приращениями:

$$-Ln\big(p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})p(\boldsymbol{a},\boldsymbol{V})\big)$$

$$= -\sum_{k=1}^{n}\big(y_k LnP_k + (1-y_k)Ln(1-P_k)\big) + \frac{1}{2}\boldsymbol{a}^T\boldsymbol{V}\boldsymbol{a} - \frac{1}{2}tr\big(Ln(\boldsymbol{V})\big)$$

$$+ d\boldsymbol{a}^T(\boldsymbol{V}\boldsymbol{a} - \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{u}) + \frac{1}{2}d\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \boldsymbol{V})d\boldsymbol{a} \qquad (21)$$

со значением весов:

$$W_{kk} = Exp(-u_k\boldsymbol{e}_k^T\boldsymbol{X}\boldsymbol{a}) \qquad (22)$$

Дифференцирование выражения (21) дает искомые уравнения для нахождения коэффициентов регрессии:

$$(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \boldsymbol{V})(\boldsymbol{a} + d\boldsymbol{a}) - \boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{X}\boldsymbol{a} + \boldsymbol{u}) = \boldsymbol{0} \qquad (23)$$

Поскольку SCLC и SCEC предоставляют количественное значение результата прогноза, была введена оценка порога классификации модели для разделения активных и неактивных соединений. В качестве порога классификации модели в данной работе использовалось значение, при котором достигается максимальная сбалансированная точность модели.

2.3.3 Ортогонализация и итерирование

Итерационные выражения для SCLC и SCEC определяют общий ход решения систем нелинейных уравнений и определения таким образом значений коэффициентов регрессии на каждом шаге итерации. Ниже приведены критерии для оценки значимости переменных, а также пути их использования при формировании финальной модели.

Независимые переменные могут быть скоррелированы и, соответственно, вносят в описание зависимой переменной частично перекрывающиеся вклады. Поэтому необходимо провести их ортогонализацию относительно друг друга. В SCLC и SCEC требуется решение похожих систем уравнений (19) и (23) для чего используется модифицированный алгоритм ортогонализации Грама-Шмидта [166].

Начиная с первого столбца матрицы $\boldsymbol{X}$, на каждом шаге выполняется ортогонализация к данному столбцу $\boldsymbol{x}_i$ всех последующих столбцов матрицы $\boldsymbol{X}$. То есть, к уже построенной модели добавляется одна новая переменная, как если бы выполнялась инкрементная пошаговая регрессия.

На каждом шаге ортогонализации столбец $\boldsymbol{x}_j$ замещается своим остатком:

$$\delta\boldsymbol{x}_j = \boldsymbol{x}_j - \boldsymbol{X}\boldsymbol{T}\boldsymbol{X}^T\boldsymbol{x}_j$$

Столбец зависимой переменной после шага ортогонализации в SCLC преобразуется следующим образом:

$$\delta\boldsymbol{y} = (\boldsymbol{E} - \boldsymbol{X}\boldsymbol{T}\boldsymbol{X}^T)(\boldsymbol{y} - \boldsymbol{p}(\boldsymbol{X}, \boldsymbol{a}) + \boldsymbol{W}\boldsymbol{X}\boldsymbol{a})$$

Столбец зависимой переменной после шага ортогонализации в SCEC преобразуется следующим образом:

$$\delta \boldsymbol{y} = (\boldsymbol{E} - \boldsymbol{XTX}^T \boldsymbol{W})(\boldsymbol{u} + \boldsymbol{Xa})$$

При этом выражения для параметров регуляризации и параметров $\boldsymbol{a}$:

$$V_{jj} = \frac{\left(\boldsymbol{x_j}^T \boldsymbol{W} \delta \boldsymbol{x}_j\right)^2}{\left(\boldsymbol{x_j}^T \boldsymbol{W} \delta \boldsymbol{y}\right)^2 - \boldsymbol{x_j}^T \boldsymbol{W} \delta \boldsymbol{x}_j}$$

$$a_j + da_j = \frac{\boldsymbol{x_j}^T \boldsymbol{W} \delta \boldsymbol{y}}{\boldsymbol{x_j}^T \boldsymbol{W} \delta \boldsymbol{x}_j} \left(1 - \frac{\boldsymbol{x_j}^T \boldsymbol{W} \delta \boldsymbol{x}_j}{\left(\boldsymbol{x_j}^T \boldsymbol{W} \delta \boldsymbol{y}\right)^2}\right)$$

Выражение в скобках при расчете $a_j + da_j$ является эффективной размерностью переменной:

$$\dim \boldsymbol{x}_j = 1 - \frac{\boldsymbol{x_j}^T \boldsymbol{W} \delta \boldsymbol{x}_j}{\left(\boldsymbol{x_j}^T \boldsymbol{W} \delta \boldsymbol{y}\right)^2} \qquad (24)$$

Эффективная размерность может быть использована для оценки вклада переменной в модель: использование переменной имеет смысл, если $\dim \boldsymbol{x}_j > 0$, при получении значения $\dim \boldsymbol{x}_j$, равного нулю либо меньше нуля, переменная не добавляется в модель. После описанной выше процедуры ортогонализации происходит ранжирование переменных по величинам их текущей эффективной размерности. Следующий шаг ортогонализации происходит по той переменной, эффективная размерность которой наибольшая и, если ортогонализация по этой переменной еще не проводилась.

Эффективная размерность всей модели оценивается как сумма всех эффективных размерностей переменных после завершения ортогонализации:

$$Dim = \sum_{j=1}^{m} \dim \boldsymbol{x}_j.$$

В процессе разработки подходов SCLC и SCEC был эмпирически установлен наиболее вычислительно эффективный способ итерирования, суть которого заключена в последовательном выполнении двух процедур. После ортогонализации производится ранжирование используемых независимых

переменных с учетом параметров регуляризации $\boldsymbol{V}$ и отбраковка незначимых переменных. До начала повторной ортогонализации переменные также ранжируются. Ранжирование переменных в этом случае происходит по аналогичным выражениям для размерности по нормированным значениям векторов:

$$\dim \boldsymbol{x}_j = 1 - \frac{\boldsymbol{x}_j^T \boldsymbol{W} \boldsymbol{x}_j}{\left(\boldsymbol{x}_j^T \boldsymbol{W} \boldsymbol{y}\right)^2}$$

Процесс завершается, когда параметры $\boldsymbol{a}$ и параметры регуляризации $\boldsymbol{V}$ сходятся к своим итоговым значениям, устанавливается итоговый набор независимых переменных и эффективная размерность построенной зависимости. Критерием завершения является условие, которое должно выполняться для каждого коэффициента:

$$da_j < 0.001.$$

2.4 Используемое программное обеспечение

Разработка классификационных моделей и прогнозирование активностей проводили с использованием разработанных нами новых методов SCLC и SCEC, алгоритмы которых были реализованы на языке C++ для непосредственного применения через команды языка R посредством Rcpp в качестве промежуточной среды (пример кода для основных методов дан в Приложении 4).

Программа GUSAR (Generalized Unrestricted Structure-Activity Relationships) [165] использована для построения QSAR моделей и прогноза количественных значений активности.

Сравнение точности и предсказательной способности новых методов проводили по отношению к SCR из программы GUSAR, методу опорных векторов (SVM) пакета R ‘caret’, байесовскому классификатору из программы PASS и ИНС пакета R ‘neuralnet’ для построения классификационных моделей.

Прогноз спектров биологической активности для ВИЧ-ассоциированных заболеваний был получен с использованием программы PASS (Prediction of Activity Spectra for Substance) [84].

Для предобработки данных были разработаны процедуры на языке Python с применением хемоинформатической библиотеки RDKit [167].

2.5 Критерии оценки качества и сравнения моделей

Оценка характеристик моделей проводилась с применением перекрестного контроля с разбиением на 5 частей. В процессе валидации выборки разбивались на 5 равных частей таким образом, чтобы 4 части составляли обучающую выборку и 1 часть – тестовую. Для полученной обучающей выборки строили модель, которая использовалась для прогноза активности соединений в тестовой выборке. Обучение и прогноз осуществлялись для каждого разбиения.

Оценку точности и прогностической способности количественных моделей выполняли на основе следующих внутренних и внешних критериев (раздел 1.2.4):

- R^2 – коэффициент детерминации;
- Q^2 – коэффициент детерминации при кросс-валидации;
- $R_{тест}^2$ – коэффициент детерминации на тестовой выборке;
- RMSD – среднеквадратичное отклонение прогноза.

С этой целью использовались формулы (11), (12).

Оценку точности классификационных моделей выполняли на основе следующих критериев:

- BA – сбалансированная точность при кросс-валидации;
- AUC – площадь под ROC кривой при кросс-валидации;
- V – эффективная размерность модели.

При этом указанные критерии использовались и для оценки прогностической способности на перекрестном контроле.

С этой целью использовались формулы (13)-(15) (см. раздел 1.2.4).

2.6 Оценка методов виртуального скрининга

Модель также возможно оценить с точки зрения того, насколько она будет эффективна при проведении виртуального скрининга. Указанный в разделе 1.2.4 критерий ROC-AUC позволяет оценить способность модели отделять активные соединения от неактивных, однако не характеризует качество процедуры виртуального скрининга, при котором производится отбор соединений, ранжированных согласно расчетным оценкам вероятности наличия (или количественной величины) активности. При применении (Q)SAR методов необходимо провести такое ранжирование соединений на основе численной оценки, получаемой из модели. Для количественных моделей такой оценкой является непосредственная оценка, получаемая при прогнозе. Точный количественный прогноз должен приводить и к правильному упорядочиванию по убыванию величины активности, но и достижение такого упорядочивания классификационными моделями также возможно, при условии, что в процессе классификации получают некоторую количественную оценку (например, оценку вероятностей отнесения соединения к классам «активных» и «неактивных»).

В SCLC и SCEC в качестве такого рода параметра, скоринга, выступают значения $\boldsymbol{x}_k^T\boldsymbol{a}$.

Для оценки ранжирования в работе использован фактор обогащения (Enrichment Factor – EF) как явно интерпретируемая величина:

$$EF_{\%} = \frac{N_{x\%}/N_{\%}}{N_a/N} \qquad (25),$$

где $N_{x\%}$ – число активных соединений в верхних $x\%$ упорядоченной по скорингу выборки, $N_{\%}$ – общее число соединений в верхних $x\%$ упорядоченной по скорингу выборки, N_a – число активных соединений в выборке, N – число соединений в выборке.

EF является параметрической метрикой, зависящей от $x\%$ отобранных из упорядоченной по скорингу выборки соединений [168].

Для оценки методов виртуального скрининга по тестовой выборке с известными количественными данными нами предложена дополнительная характеристика $A_{\%}$ – среднее значение активности A в верхних $x\%$ упорядоченной по скорингу выборки:

$$A_{\%} = \frac{1}{N_{\%}} \sum_{i=1}^{N_{\%}} A_i \qquad (26),$$

где A_i – значение активности A для i-ого наблюдения, суммирование выполняется по $N_{\%}$ первых по скорингу соединений. Для основных выборок ингибиторов белков ВИЧ-1 известны pIC_{50} и нами подсчитывались их средние значения.

ГЛАВА 3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Учитывая актуальность усовершенствования и развития подходов к созданию (Q)SAR моделей, нами разработаны методы классификации SCEC и SCLC. С целью оценки качества наших методов были сформированы выборки, на основе которых проведены верификация и валидация разработанных алгоритмов.

Апробацию методов проводили с использованием выборок об активности ингибиторов ферментов ВИЧ-1 и SARS-CoV-2. Выборки были собраны нами из различных источников и была проведена их предварительная обработка. Для создания количественных моделей использованы только данные о полуингибирующей концентрации IC_{50} как наиболее представительные. При создании обучающих выборок для классификационных моделей использованы как данные по IC_{50}, так и другие количественные характеристики активности, а также качественные данные (активно/неактивно).

Как с использованием новых методов, так и с применением ранее предложенных подходов были построены количественные и классификационные модели зависимостей «структура-активность». Полученные модели использованы для сравнения различных подходов и, таким образом, были выявлены преимущества и недостатки разработанных методов, а также проведено сопоставление точности и предсказательной способности количественного и классификационного подходов. На последнем этапе оценены потенциальные преимущества классификационного подхода для проведения виртуального скрининга с учетом имеющихся сведений об экспериментальных данных по активности.

В дальнейшем полученные модели были использованы для создания свободно-доступного веб-сервиса, который предоставляет возможность прогнозирования соответствующих видов активности на основе структурной формулы нового соединения широкому кругу пользователей.

3.1 Методы SCEC и SCLC

Перед разработкой новых методов анализа «структура-активность» нами были сформулированы критерии, которым должны удовлетворять современные подходы к классификации для учета гетерогенности используемых при обучении данных:

- проведение автобалансировки для несбалансированных выборок;
- отбор и использование наиболее значимых переменных;
- возможность снижения эффективной размерности получаемых моделей;
- точность и предсказательная способность моделей не должна уступать аналогичным характеристикам ранее реализованных подходов.

Разработанные нами классификационные методы SCEC и SCLC были протестированы на искусственно сгенерированных данных (раздел 2.1.1), а также была проведена их валидация с использованием выборок Tox21. Методы были апробированы путем анализа зависимостей «структура-активность» для ингибиторов ферментов ВИЧ-1 и SARS-CoV-2. Описание данных приведено в разделах 2.1.1-2.1.4; разработанные нами математические алгоритмы SCEC и SCLC представлены в разделе 2.3.2. Проведено сопоставление качества классификации, полученного на этих выборках с применением методов SCEC и SCLC, с подходами, основанными на SCR, SVM, байесовским классификатором и ИНС. Продемонстрировано, что классификаторы, полученные на обширных обучающих выборках, обеспечивают более высокую точность и предсказательную способность по сравнению с использованием выборок меньшего объема. Показаны преимущества разработанных нами классификаторов по сравнению с количественными моделями, благодаря расширению химического разнообразия вовлекаемых в анализ данных, для которых активность соединений обучающей выборки охарактеризована различными количественными параметрами, либо экспериментальные результаты представлены в качественном виде (активно/неактивно).

3.1.1 Тестирование методов SCEC и SCLC

Для анализа точности и предсказательной способности предложенных нами методов использовались сгенерированные выборки, имитирующие реальные количественные данные, классифицированные на положительные и отрицательные примеры (соответствующие активным/неактивным соединениям) по выбранному порогу (раздел 2.1.1). Поскольку процесс оценки параметров моделей и параметров регуляризации итеративен (раздел 2.3.3), было необходимо убедиться, что веса соединений, коэффициенты регрессии и расчетные величины зависимой переменной сходятся к конечным значениям. Изменения всех упомянутых выше параметров отслеживались на каждой итерации, что позволило установить критерии сходимости процесса вычислений к стабильным значениям. Так, для примеров, которые легко классифицируются, значение весов быстро убывает (рисунок 9а), и эти примеры вносят наименьший вклад в получаемую модель. Рисунок 9б иллюстрирует, что для таких примеров быстро достигаются высокие абсолютные значения оценок зависимой переменной. И наоборот, примеры, расположенные близко к порогу классификации и, соответственно, имеющие низкие абсолютные величины оценок, сохраняют высокие значения весов и вносят наибольший вклад в модель.

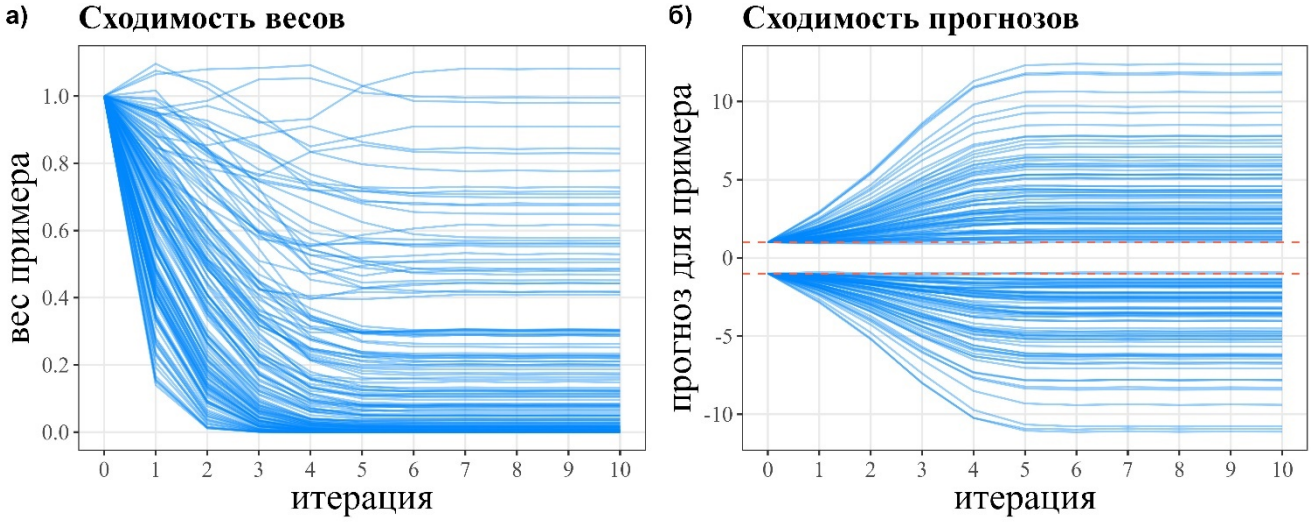


Рисунок 9 – Сходимость в процессе итераций SCEC: а) весов б) прогнозов (на примере выборки 6 приложения 2)

На рисунке 9 представлен результат расчета для идеализированного случая, когда возможно точное разделение между классами положительных и отрицательных примеров. На практике такое разделение производится с помощью определения порога классификации модели (раздел 2.3.2).

Принципиальным различием SCEC и SCLC является способ расчета весов конкретных примеров и оценки соответствующего вклада этих примеров в модель. Веса SCLC ограничены значением 0.25, тогда как в SCEC веса могут иметь любые положительные значения. В случае сгенерированных данных мы обладаем всей полнотой информации о значениях зависимой переменной до проведения классификации, поэтому оценка адекватности разработанных методов проводилась путем анализа значений весов примеров в зависимости от их удаленности от порога разделения. Для точек, близких к порогу классификации модели, значения весов наибольшие, и наоборот (рисунок 10). Следует отметить, что веса распределяются согласно наложенным штрафным функциям.

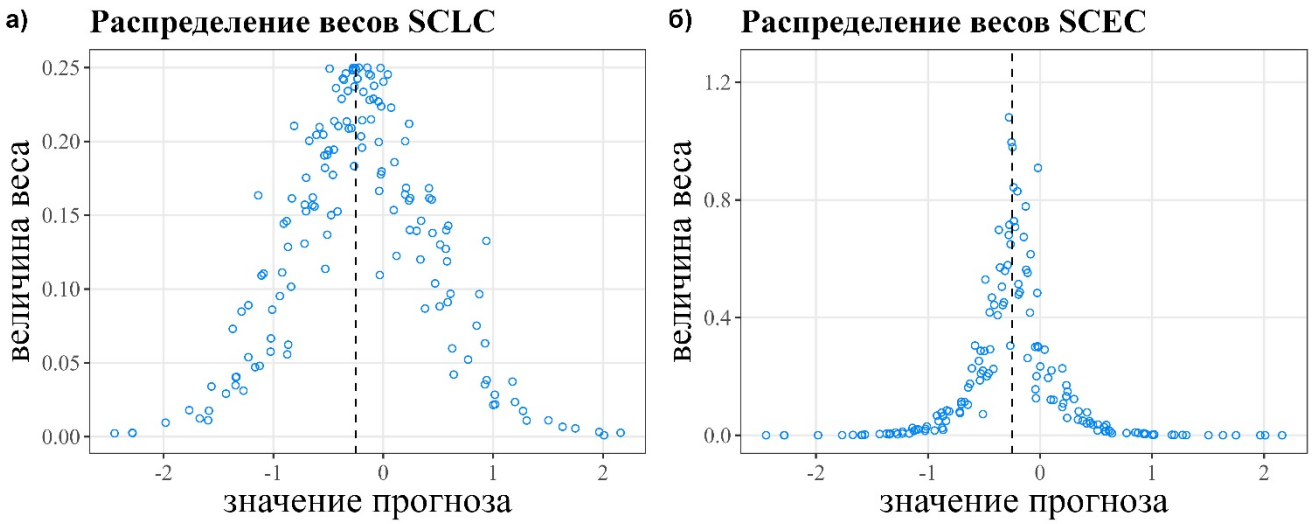


Рисунок 10 – Веса примеров в зависимости от значений прогноза для а) SCLC б) SCEC; штриховой линией показан порог классификации модели

В SCLC и SCEC применяются разные оценки вероятности положительного исхода в схеме испытаний Бернулли, поэтому мы сравнивали значения весов соединений, в зависимости от удаленности наблюдаемых значений от порога

классификации. В SCLC такая зависимость куполообразна, на основном участке веса растут почти линейно до 1/4 (рисунок 10а). В SCEC такая зависимость экспоненциальная и вес может достигать любых значений (рисунок 10б).

Проведенные нами расчеты показали, что для обоих методов получаемые оценки параметров близки к таковым у сгенерированных данных (отклонялись от действительных значений не более, чем на 5%). Таким образом, проведенные нами вычислительные эксперименты подтвердили адекватность предложенных нами алгоритмов. Результаты исследования SCLC и SCEC на сгенерированных данных (Приложение Б) приведены в таблице 2.

Таблица 2 – Результаты проверки SCLC и SCEC на сгенерированных данных

Метод	***Выборка***	$\boldsymbol{\sigma_a}$	$\boldsymbol{N_{it}}$
SCLC	*Выборка 1*	*0.0381*	*7*
SCLC	*Выборка 2*	*0.298*	*8*
SCLC	*Выборка 3*	*0.0305*	*7*
SCLC	*Выборка 4*	*0.264*	*7*
SCLC	*Выборка 5*	*0.0291*	*7*
SCLC	*Выборка 6*	*0.235*	*8*
SCLC	*Выборка 7*	*0.0319*	*7*
SCLC	*Выборка 8*	*0.343*	*7*
SCLC	*Выборка 9*	*0.0189*	*7*
SCLC	*Выборка 10*	*0.339*	*7*
SCLC	*Выборка 11*	*0.0334*	*16*
SCLC	*Выборка 12*	*0.353*	*17*
SCLC	*Выборка 13*	*0.0330*	*16*
SCLC	*Выборка 14*	*0.362*	*18*
SCLC	*Выборка 15*	*0.0107*	*16*
SCLC	*Выборка 16*	*0.210*	*18*

Метод	***Выборка***	$\boldsymbol{\sigma_a}$	$\boldsymbol{N_{it}}$
SCEC	*Выборка 1*	*0.0426*	*8*
SCEC	*Выборка 2*	*0.436*	*8*
SCEC	*Выборка 3*	*0.0331*	*8*
SCEC	*Выборка 4*	*0.255*	*8*
SCEC	*Выборка 5*	*0.0341*	*8*
SCEC	*Выборка 6*	*0.347*	*8*
SCEC	*Выборка 7*	*0.0176*	*8*
SCEC	*Выборка 8*	*0.351*	*8*
SCEC	*Выборка 9*	*0.0294*	*21*
SCEC	*Выборка 10*	*0.267*	*21*
SCEC	*Выборка 11*	*0.0334*	*21*
SCEC	*Выборка 12*	*0.233*	*22*
SCEC	*Выборка 13*	*0.0240*	*21*
SCEC	*Выборка 14*	*0.089*	*21*
SCEC	*Выборка 15*	*0.0260*	*21*
SCEC	*Выборка 16*	*0.224*	*21*

$\boldsymbol{\sigma_a}$ – стандартная ошибка оценок значений параметров, $\boldsymbol{N_{it}}$ – количество итераций до удовлетворения условий сходимости.

3.1.2 Валидация методов SCEC и SCLC

Для валидации разрабатываемых методов нами были построены классификационные модели SCLC, SCEC и SCR для 38 выборок из Tox21. Далее проводилась кросс-валидация и сравнение сбалансированной точности моделей, построенных с применением разных методов.

С использованием SCEC были получены сравнимые (0.711 – среднее для всех моделей) или превосходящие SCR (0.707 – среднее для всех моделей) значения сбалансированной точности и меньшие в случае при использовании SCLC (0.704 – среднее для всех моделей). Как в случае SCLC, так и в случае SCEC соответствующая точность оценок была достигнута при существенно меньшем количестве независимых переменных (81 и 79 в среднем, соответственно) по сравнению с SCR (220 в среднем).

Сравнение результатов SCR и SCEC приведено на рисунке 11. Характеристики точности и прогностической способности моделей перечислены в приложении 3. Полученные результаты свидетельствуют о возможности применения разработанных нами методов для построения моделей с использованием несбалансированных выборок. При этом не происходит значительного снижения точности и прогностической способности получаемых моделей. Эта особенность может быть использована для добавления в обучающую выборку данных, относящихся к конкретному классу (положительных или отрицательных примеров) для увеличения описываемого химического пространства. Также построение моделей с использованием данных Tox21 указывает на возможность применения новых подходов в ситуациях, когда несбалансированность соотношения активных и неактивных структур очень велика (вплоть до 1:73).

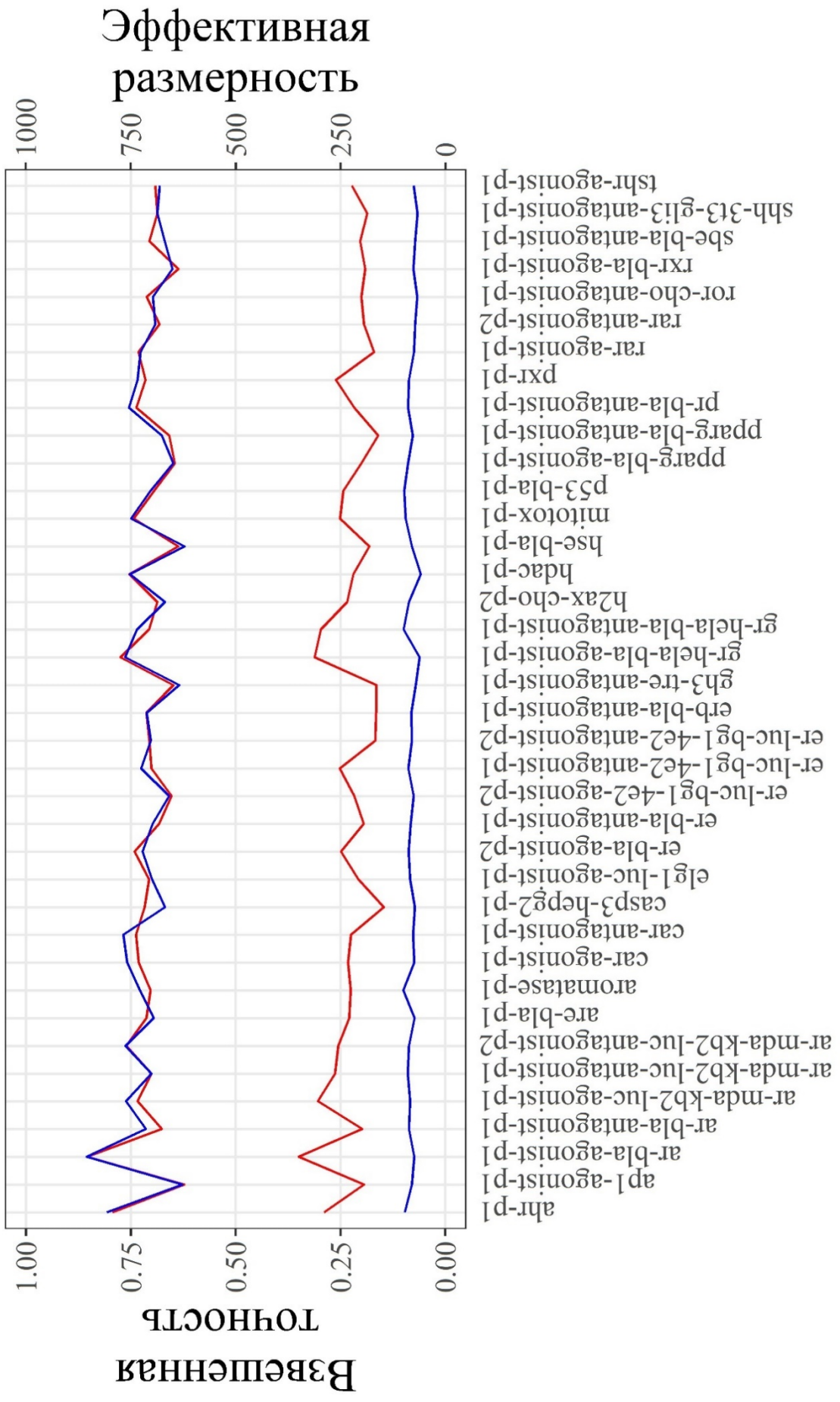


Рисунок 11 – Результаты валидации на данных Tox21. SCEC – синяя линия, SCR – красная линия

3.2 Предобработка данных

При формировании обучающих выборок нами были разработаны полуавтоматические процедуры обработки структурной информации в выборках и ассоциированных значений активности.

Подробное перечисление проведенных процедур описано в разделе 2.1.5. Примеры подобных преобразований приведены на рисунке 12.

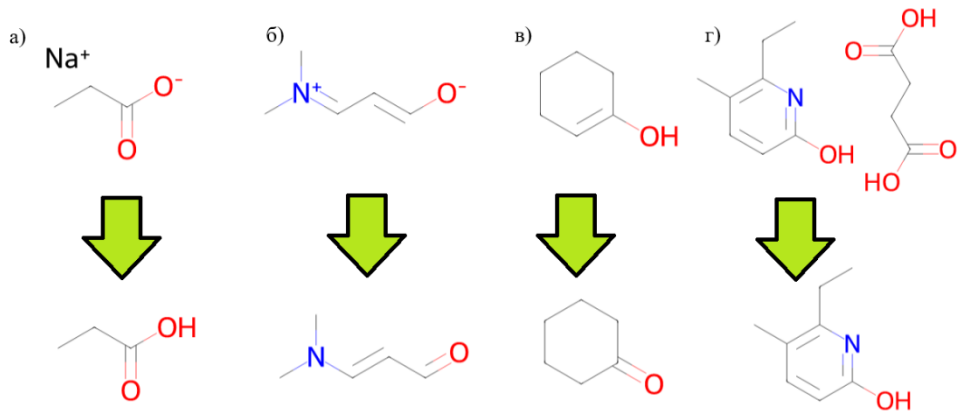


Рисунок 12 – Процесс автоматизированной обработки структур: а) обработка солей; б) нормализация; в) таутомеризация; г) выделение фармакологического компонента

Автоматизация обработки данных позволила провести подготовку обучающих выборок ингибиторов ферментов ВИЧ-1 и SARS-CoV-2 единообразным способом. Так, например, удалось избежать ситуации, когда два различных со структурной и идентичных с фармакологической точки зрения соединения входят в обучающую выборку, что может произойти при необработанных таутомерах, либо использовании различных солей одного и того же фармакологического компонента. За счет обработки экспериментальных значений активности снижены отрицательные эффекты от дуплицированных векторов дескрипторов при построении моделей. Такая процедура позволяет избежать возникновения линейной зависимости строк в регрессионной матрице и избежать смещения получаемых результатов в пользу того химического класса, который представлен дубликатами.

3.3 QSAR модели ингибирования мишеней ВИЧ-1

Подготовленные выборки с данными об ингибиторах интегразы, протеазы и обратной транскриптазы ВИЧ-1 были использованы для построения QSAR моделей.

3.3.1 Создание моделей из попарных комбинаций выборок

Характеристики моделей, полученных с использованием выборок, содержащих данные NIAID и ChEMBL, NIAID и Integrity об ингибиторах интегразы, протеазы и обратной транскриптазы ВИЧ-1, представлены в таблице 3.

Таблица 3 – Характеристики количественных моделей, полученных с использованием выборок NIAID и ChEMBL, NIAID и Integrity

Выборка	R^2	Q^2	***RMSD***	***V***
IN (NIAID и ChEMBL)	*0.960*	*0.821*	*0.587*	*384*
PR (NIAID и ChEMBL)	*0.956*	*0.829*	*0.696*	*494*
RT (NIAID и ChEMBL)	*0.943*	*0.723*	*0.760*	*455*
IN (NIAID и Integrity)	*0.960*	*0.819*	*0.592*	*371*
PR (NIAID и Integrity)	*0.957*	*0.827*	*0.710*	*485*
RT (NIAID и Integrity)	*0.942*	*0.715*	*0.776*	*441*

R^2 – коэффициент детерминации; Q^2 – коэффициент детерминации предсказания; ***RMSD*** – среднеквадратичное отклонение прогноза; V – эффективная размерность.

В качестве независимых тестовых выборок для оценки прогностической способности построенных количественных зависимостей «структура-активность» были подготовлены выборки с данными, не входящими в обучение (раздел 2.1.5). Сопоставление результатов прогноза с наблюдаемыми величинами при валидации представлено на рисунках 13-15.

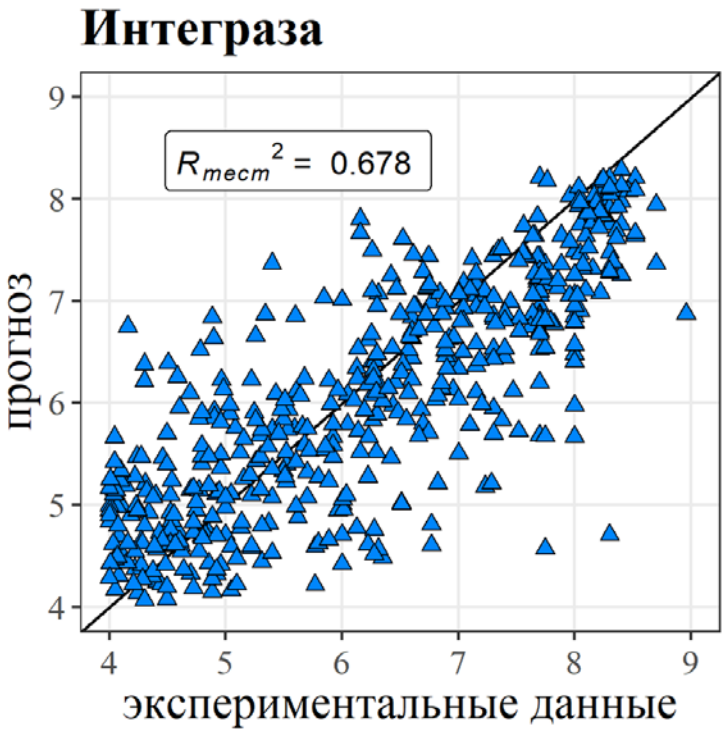


Рисунок 13 – Прогноз pIC_{50} М ингибиторов интегразы для тестовой выборки ChEMBL на основе моделей, построенных на основе NIAID и Integrity

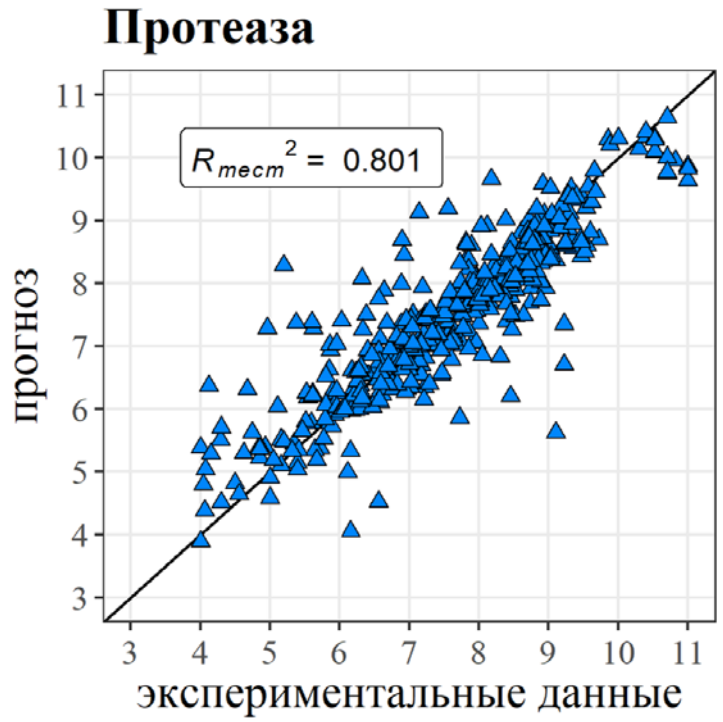


Рисунок 14 – Прогноз pIC50 М ингибиторов протеазы для тестовой выборки ChEMBL на основе моделей, построенных на основе NIAID и Integrity

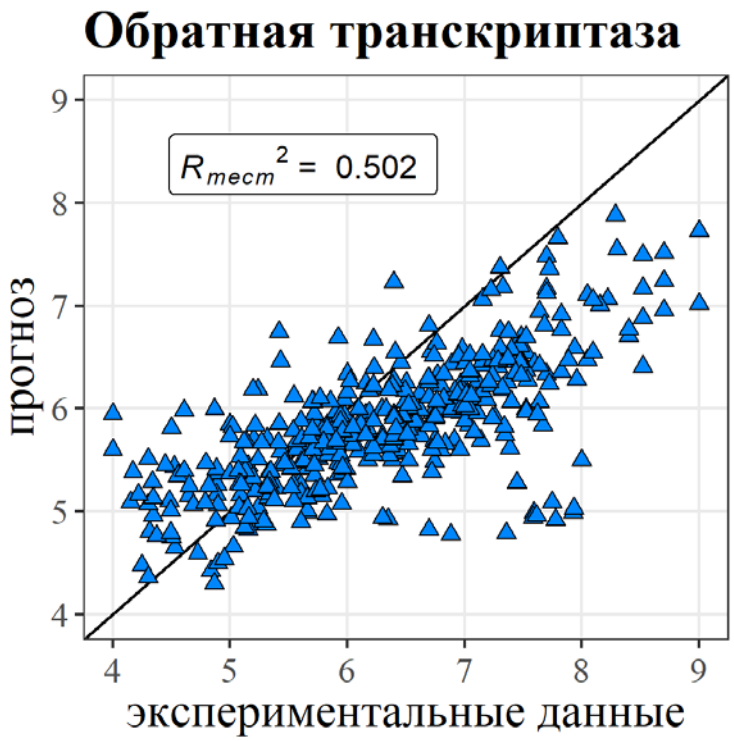


Рисунок 15 – Прогноз pIC50 M ингибиторов обратной транскриптазы для тестовых выборок ChEMBL на основе моделей, построенных на основе NIAID и Integrity

Из представленных на рисунках 13-15 данных видно, что результат валидации зависит от конкретной мишени. Менее высокая точность прогноза в случае ингибиторов интегразы и обратной транскриптазы может быть объяснена особенностями химических пространств по отношению к информации, связанной с конкретными ферментами (см. ниже).

Для выяснения причин снижения качества прогноза была проведена оценка числа новых MNA дескрипторов в тестовых выборках ChEMBL по сравнению с обучающими выборками, поскольку любые выборки химических соединений охватывают лишь часть теоретически возможного химического пространства. Анализ проводили таким образом, что для каждой структуры из ChEMBL оценивали процент новых MNA дескрипторов по сравнению с представленными в БД NIAID и Integrity (рисунки 16-18).

Для всех трех мишеней имеются новые дескрипторы, не содержащиеся в соединениях обучающих выборок. У некоторых ингибиторов интегразы и обратной транскриптазы их доля превышает 15%. По отношению к ингибиторам протеазы число новых дескрипторов составило 94 по сравнению с 220 и 199 для ингибиторов интегразы и обратной транскриптазы, соответственно. Поэтому с целью расширения рассматриваемого химического пространства были использованы выборки, включающие в себя все доступные данные.

Полученные нами результаты согласуются с результатами Погодина с соавт. [169].

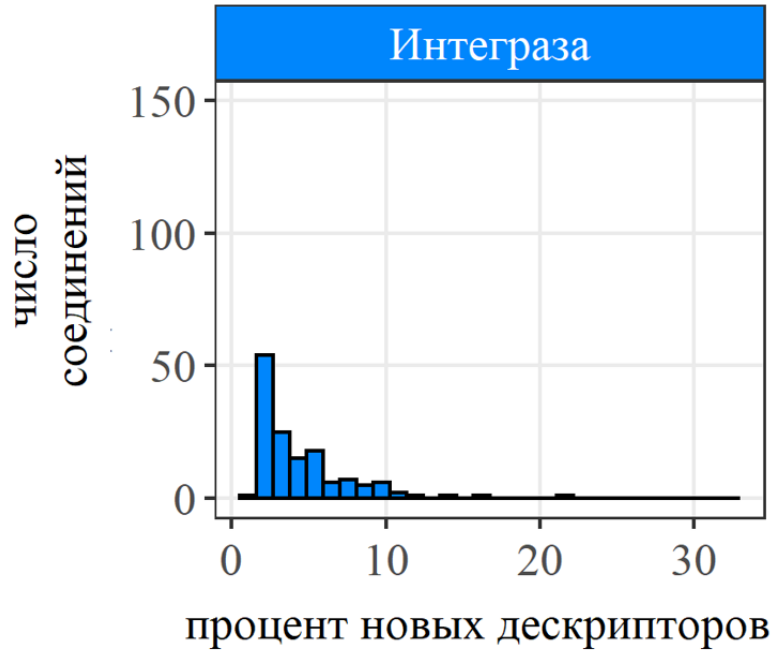


Рисунок 16 – Гистограмма распределения новых дескрипторов в выборке ингибиторов интегразы ChEMBL по сравнению с выборками NIAID и Integrity

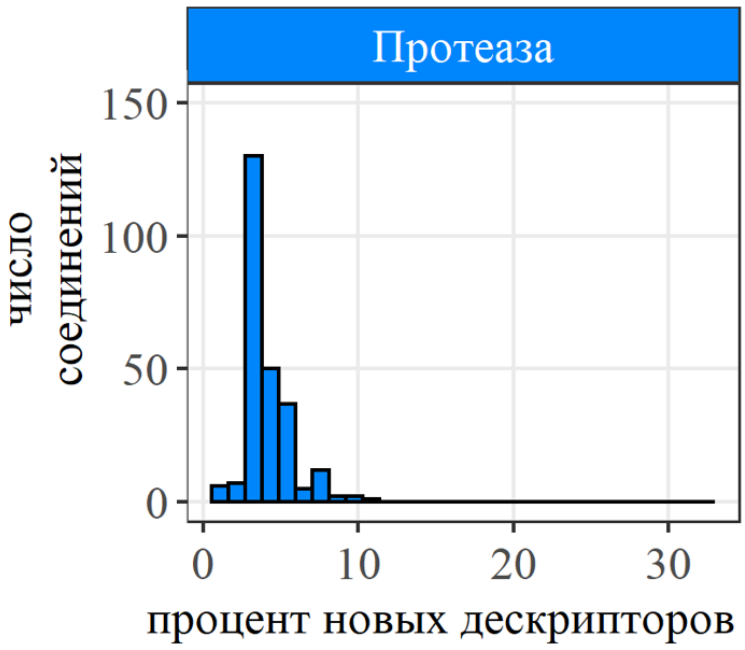


Рисунок 17 – Гистограмма распределения новых дескрипторов в выборке ингибиторов протеазы ChEMBL по сравнению с выборками NIAID и Integrity

Рисунок 18 – Гистограмма распределения новых дескрипторов в выборке ингибиторов обратной транскриптазы ChEMBL по сравнению с выборками NIAID и Integrity

3.3.2 Создание моделей с использованием всех доступных количественных данных о полуингибирующей концентрации

В ходе выполнения диссертационной работы нами были построены количественные зависимости «структура-активность» (QSAR модели) с использованием всех доступных данных о полуингибирующей концентрации (IC_{50}) ингибиторов ферментов ВИЧ-1. Характеристики полученных моделей представлены в таблице 4. Сопоставление результатов прогноза с экспериментальными данными приведено на рисунках 19-21. Поскольку для создания этих моделей использовались объединенные выборки, то их валидацию проводили не с помощью прогноза для тестовой выборки, а путем пятикратной кросс-валидации.

Таблица 4 – Характеристики количественных моделей SCR

Выборка	$\boldsymbol{R^2}$	$\boldsymbol{Q^2}$	***RMSD***	***V***	$\boldsymbol{R^2_{LMO}}$
IN (NIAID, ChEMBL и Integrity)	*0.966*	*0.818*	*0.595*	*392*	*0.75*
PR (NIAID, ChEMBL и Integrity)	*0.957*	*0.824*	*0.709*	*470*	*0.77*
RT (NIAID, ChEMBL и Integrity)	*0.945*	*0.714*	*0.767*	*452*	*0.65*

$\boldsymbol{R^2}$ – коэффициент детерминации; $\boldsymbol{Q^2}$ – коэффициент детерминации кросс-валидации; ***RMSD*** – среднеквадратичное отклонение прогноза; ***V*** – эффективная размерность; $\boldsymbol{R^2_{LMO}}$ – коэффициент детерминации на тестовой выборке.

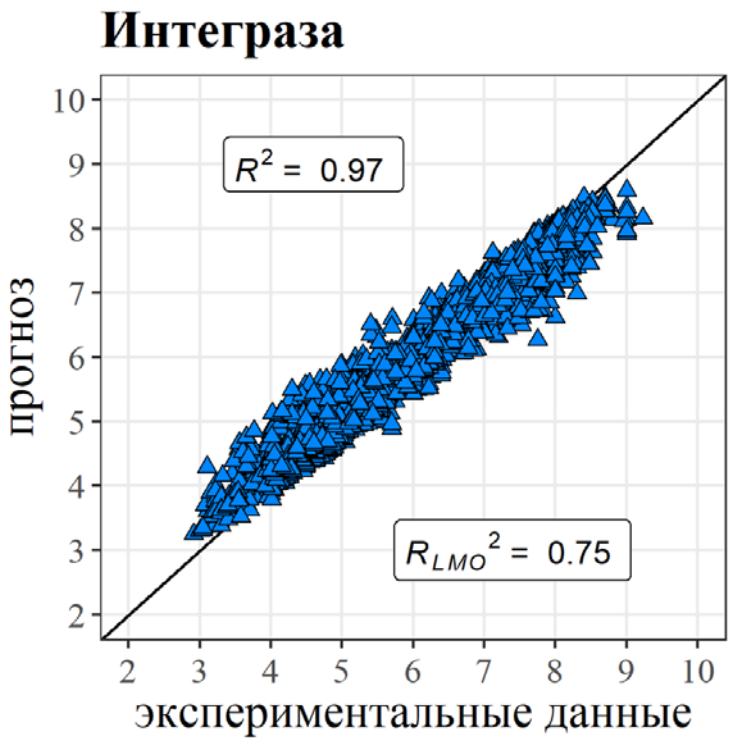


Рисунок 19 – Расчетные и наблюдаемые величины pIC50 для количественных моделей SCR для выборки IN

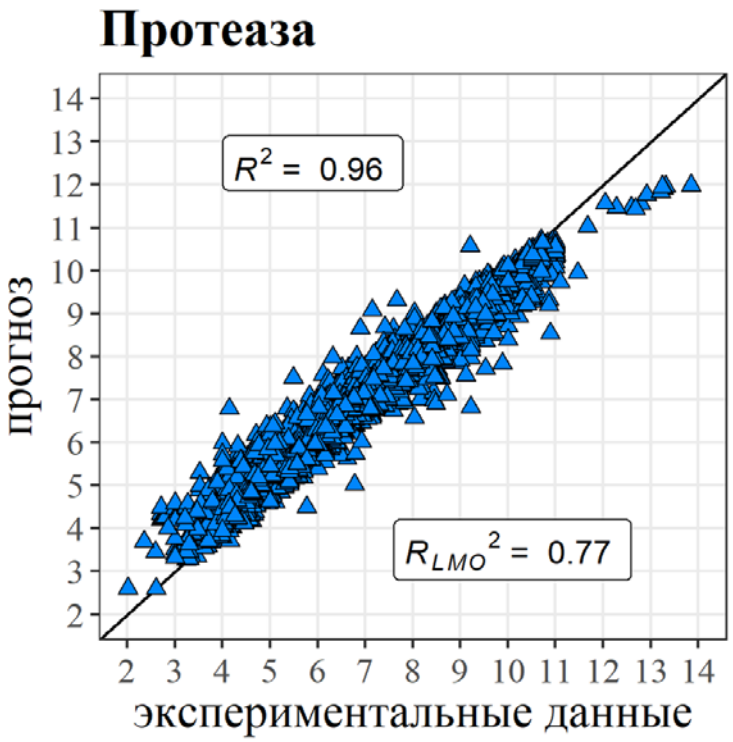


Рисунок 20 – Расчетные и наблюдаемые величины pIC50 для количественных моделей SCR для выборки PR

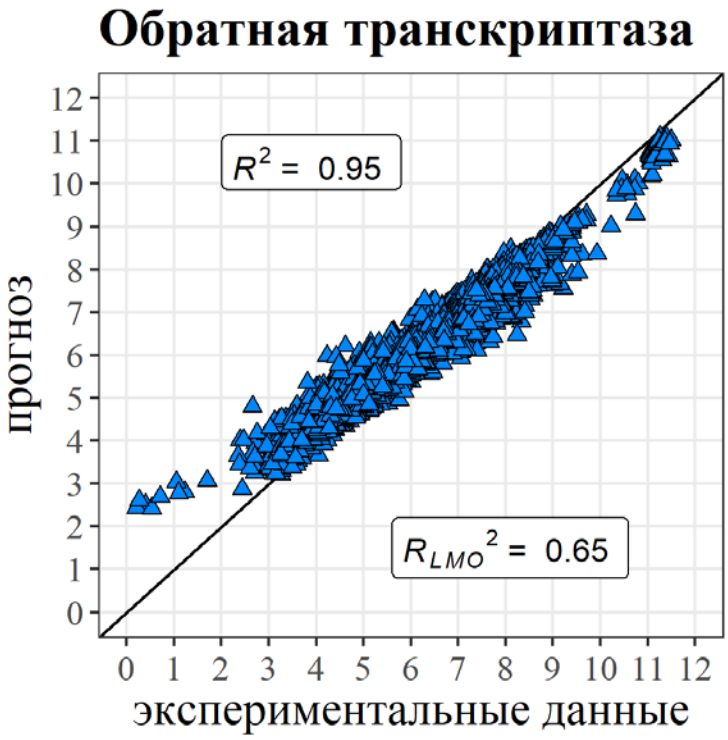


Рисунок 21 – Расчетные и наблюдаемые величины pIC50 для количественных моделей SCR для выборки RT

Как видно из приведенных в таблице 4 данных, построенные QSAR модели имеют удовлетворительные характеристики. Однако, необходимо подчеркнуть, что при построении этих моделей была использована лишь информация по полуингибирующей концентрации IC_{50}, которая составляет менее 60% доступных данных. Можно полагать, что добавление в обучающие выборки данных, охарактеризованных иными количественными показателями активности (Ki и др.), либо качественными оценками активности (активно/неактивно) позволит расширить охватываемое химическое пространство и, соответственно, область применимости моделей. Очевидно, что в этом случае можно использовать только классификационные модели.

3.4 Классификационные модели зависимостей «структура-активность» для ингибиторов ферментов ВИЧ-1 и SARS-CoV-2

Построение классификационных моделей выполнено на основе выборок по ингибиторам белков ВИЧ-1 и SARS-CoV-2 (разделы 2.1.5 и 2.1.6). Используя программную реализацию методов SCLC и SCEC, были построены классификационные модели, качество которых сравнили с моделями SCR, SVM, PASS и ИНС. Алгоритмы SCR, SCEC и SCLC обеспечивают автоматический отбор значимых переменных при построении моделей (раздел 2.3.3). В случае SVM и ИНС автоматического отбора значимых переменных не происходит, поэтому они были отобраны с использованием SCEC. Построение всех моделей проводили на одних и тех же обучающих выборках при одинаковых условиях; качество моделей оценивали на основе пятикратной кросс-валидации. Для оценки числа независимых переменных, использованных при построении полученных моделей, были рассчитаны показатели эффективной размерности.

3.4.1 Классификационные модели зависимостей «структура-активность» для ингибиторов ферментов ВИЧ-1

Моделирование проводили для выборок, содержащих данные по ингибиторам интегразы, протеазы и обратной транскриптазы ВИЧ-1. Сопоставление характеристик моделей представлено в таблицах 5-7.

Из представленных в таблицах данных можно заключить, что модели SCEC и SCLC имеют сопоставимые с SCR, SVM, PASS и ИНС характеристики предсказательной способности, при этом используют гораздо меньшую эффективную размерность независимых переменных.

Таблица 5 – Сравнение характеристик моделей для ингибиторов интегразы ВИЧ-1

Метод	***N***	***V***	***BA***	***AUC***
SCR	*4091*	*298*	*0.845*	*0.923*
SVM	*4091*	*285*	*0.716*	*0.857*
PASS	*4091*	*5331*	*0.829*	*0.920*
ANN	*4091*	*285*	*0.702*	*0.852*
SCLC	*4091*	*86*	*0.849*	*0.916*
SCEC	*4091*	*99*	*0.849*	*0.929*

Таблица 6 – Сравнение характеристик моделей для ингибиторов протеазы ВИЧ-1

Метод	***N***	***V***	***BA***	***AUC***
SCR	*6552*	*417*	*0.861*	*0.933*
SVM	*6552*	*373*	*0.802*	*0.893*
PASS	*6552*	*5867*	*0.861*	*0.933*
ANN	*6552*	*373*	*0.798*	*0.828*
SCLC	*6552*	*123*	*0.858*	*0.910*
SCEC	*6552*	*130*	*0.866*	*0.941*

Таблица 7 – Сравнение характеристик моделей для ингибиторов обратной транскриптазы ВИЧ-1

Метод	***N***	***V***	***BA***	***AUC***
SCR	*6309*	*360*	*0.788*	*0.867*
SVM	*6309*	*368*	*0.745*	*0.855*
PASS	*6309*	*7002*	*0.797*	*0.874*
ANN	*6309*	*368*	*0.760*	*0.863*
SCLC	*6309*	*112*	*0.803*	*0.866*
SCEC	*6309*	*124*	*0.793*	*0.876*

N – количество соединений в обучающей выборке; ***V*** – размерность модели; ***BA*** – сбалансированная точность при кросс-валидации; ***AUC*** – площадь под ROC кривой при кросс-валидации.

Большинство ошибочных прогнозов происходит из-за наличия схожих соединений, величина активности которых находится близко к порогу классификации. Подобное несоответствие наблюдаемых и прогнозируемых значений проиллюстрировано на рисунке 22 для одного из ингибиторов интегразы ВИЧ-1 в обучающей выборке. В обучающей выборке данный ингибитор входит в класс активных со значением IC_{50} = 0.41 мкМ, что близко к пороговому значению 1 мкМ. Было отобрано девять наиболее схожих с указанной структурой соединений. При этом экспериментальные значения полуэффективной концентрации для всех девяти структур превышает 1 мкМ, поэтому в обучающей выборке они относятся к классу неактивных. В определенных случаях разработанные модели не могли различить конкретные структурные особенности, которые могут влиять на активность, так как вклад в модель схожих неактивных соединений оказывает решающие значение для целого класса соединений.

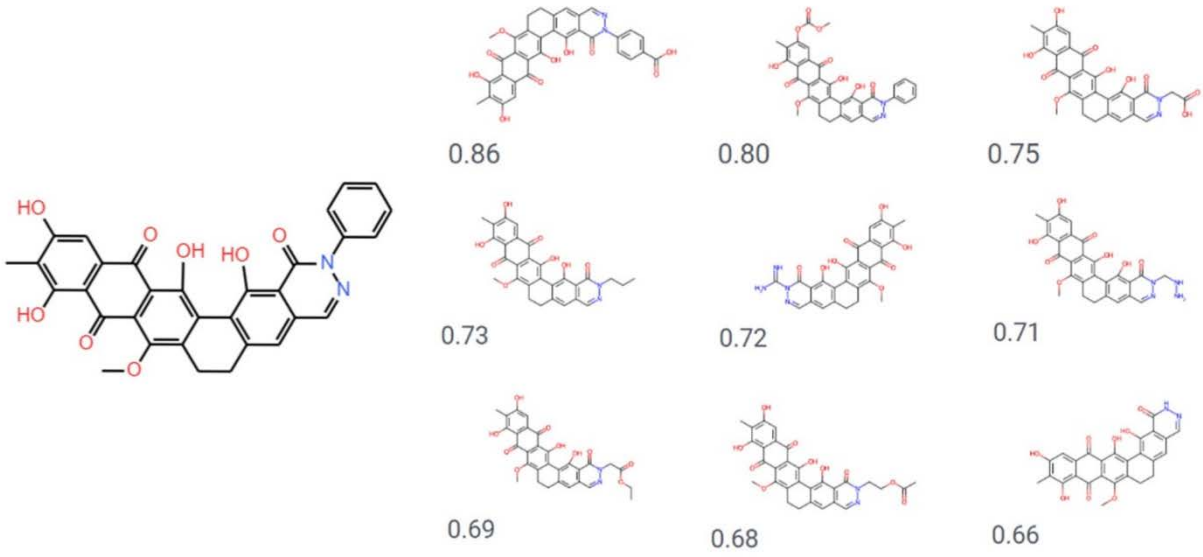


Рисунок 22 – Пример активного соединения, для которого все похожие структуры являются неактивными; приведены значения схожести по Танимото.

Более детальная характеристика для сравнения прогностической способности классификационных моделей SCR, SCLC и SCEC для ферментов ВИЧ-1 представлена на рисунках 23-25 в виде ROC кривых, построенных в координатах TPR и FPR (раздел 1.2.4, формулы (13), (14)).

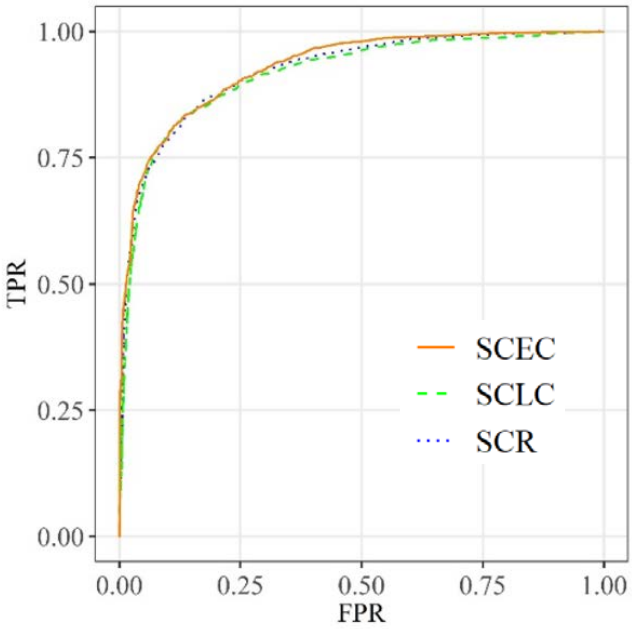


Рисунок 23 – Валидационные ROC кривые для моделей, построенных с использованием выборки IN; SCEC – оранжевая непрерывная линия, SCLC – штрих зеленая линия, SCR – пунктирная синяя линия

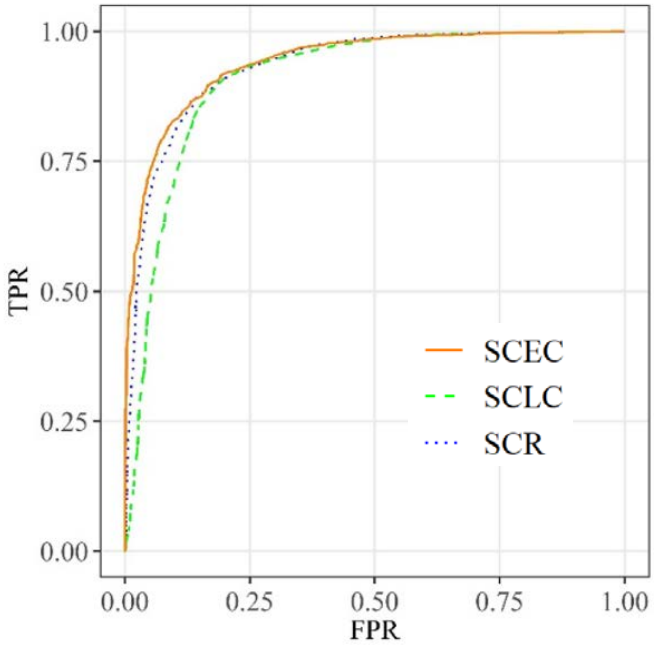


Рисунок 24 – Валидационные ROC кривые для моделей, построенных с использованием выборки PR; SCEC – оранжевая непрерывная линия, SCLC – штрих зеленая линия, SCR – пунктирная синяя линия

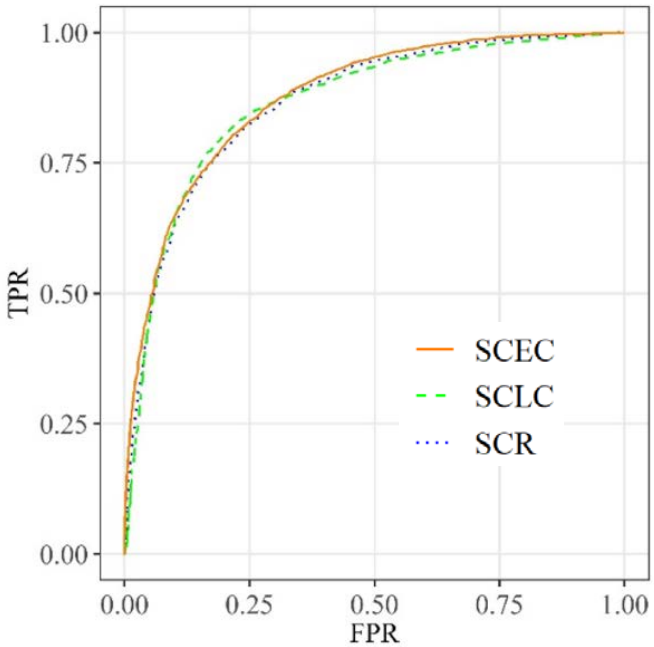


Рисунок 25 – Валидационные ROC кривые для моделей, построенных с использованием выборки RT; SCEC – оранжевая непрерывная линия, SCLC – штрих зеленая линия, SCR – пунктирная синяя линия

3.4.2 Классификационные модели зависимостей «структура-активность» для ингибиторов ферментов SARS-CoV-2

Моделирование проводилось для выборок с ингибиторами интегразы, протеазы и обратной транскриптазы ВИЧ-1 и для выборок с ингибиторами главной протеазы, папаин-подобной протеазы (PLpro) и РНК-зависимой РНК-полимеразы (RdRp) SARS-CoV-2. Результаты сравнения характеристик моделей для ингибиторов ВИЧ-1 и SARS-CoV-2 представлены в таблицах 3 и 4, соответственно.

ROC кривые для сравнения прогностической способности классификационных моделей для ферментов SARS-CoV-2 представлены на рисунках 26-28.

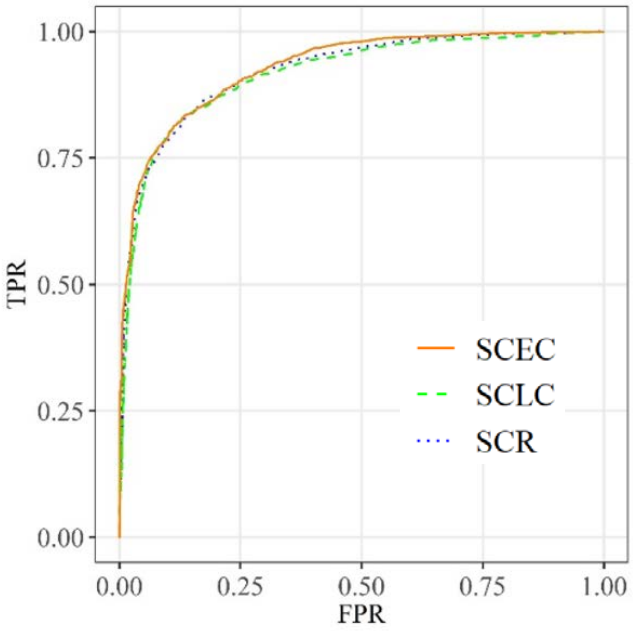


Рисунок 26 – Валидационные ROC кривые для моделей, построенных с использованием выборки 3CLpro; SCEC – оранжевая непрерывная линия, SCLC – штрих зеленая линия, SCR – пунктирная синяя линия

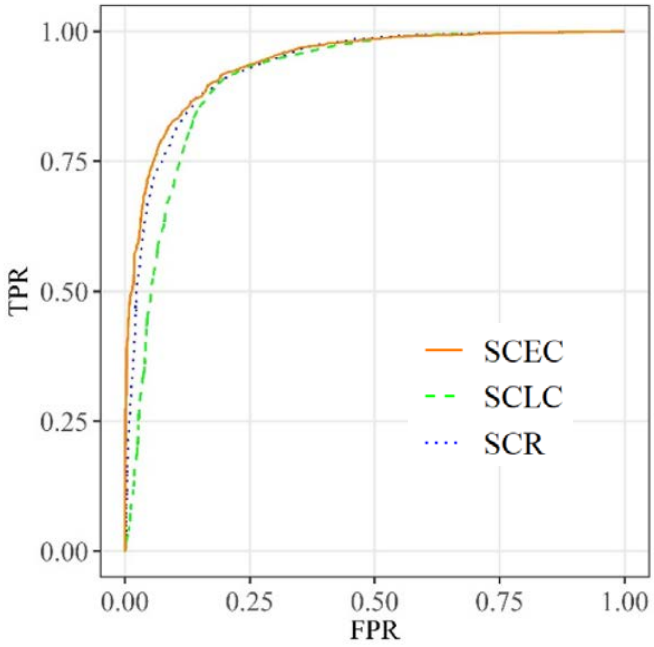


Рисунок 27 – Валидационные ROC кривые для моделей, построенных с использованием выборки PLpro; SCEC – оранжевая непрерывная линия, SCLC – штрих зеленая линия, SCR – пунктирная синяя линия

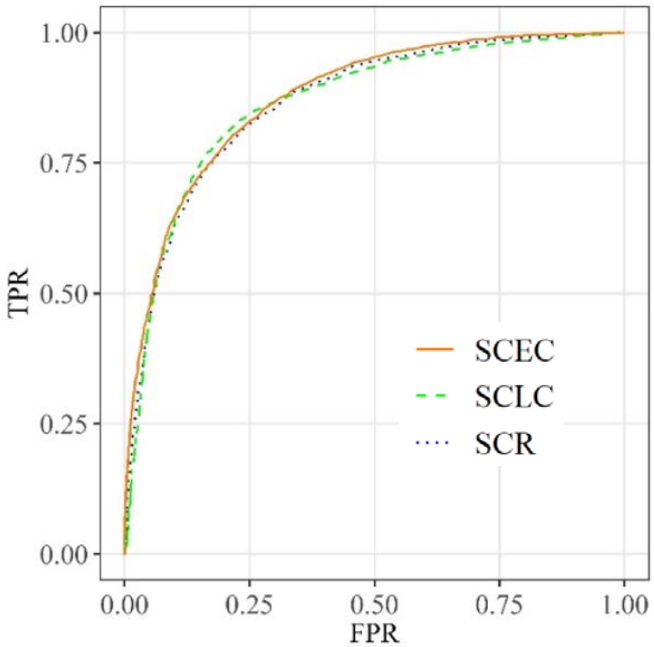


Рисунок 27 – Валидационные ROC кривые для моделей, построенных с использованием выборки RdRp; SCEC – оранжевая непрерывная линия, SCLC – штрих зеленая линия, SCR – пунктирная синяя линия

Таблица 8 – Сравнение характеристик моделей для ингибиторов главной протеазы SARS-CoV-2

Метод	***N***	***V***	***BA***	***AUC***
SCR	*6230*	*370*	*0.851*	*0.919*
SVM	*6230*	*373*	*0.834*	*0.901*
PASS	*6230*	*13037*	*0.849*	*0.920*
ANN	*6230*	*373*	*0.830*	
SCLC	*6230*	*108*	*0.850*	*0.917*
SCEC	*6230*	*117*	*0.861*	*0.925*

Таблица 9 – Сравнение характеристик моделей для ингибиторов папаин-подобной протеазы SARSs-CoV-2

Метод	***N***	***V***	***BA***	***AUC***
SCR	*106*	*11*	*0.668*	*0.803*
SVM	*106*	*12*	*0.660*	*0.789*
PASS	*106*	*1154*	*0.689*	*0.753*
ANN	*106*	*12*	*0.662*	*0.799*
SCLC	*106*	*11*	*0.684*	*0.802*
SCEC	*106*	*11*	*0.682*	*0.812*

Таблица 10 – Сравнение характеристик моделей для ингибиторов РНК-зависимой РНК-полимеразы SARS-CoV-2

Метод	***N***	***V***	***BA***	***AUC***
SCR	*2632*	*213*	*0.720*	*0.852*
SVM	*2632*	*207*	*0.703*	*0.805*
PASS	*2632*	*7853*	*0.725*	*0.759*
ANN	*2632*	*207*	*0.720*	*0.831*
SCLC	*2632*	*69*	*0.719*	*0.849*
SCEC	*2632*	*71*	*0.731*	*0.859*

N – количество соединений в обучающей выборке; ***V*** – размерность модели; ***BA*** – сбалансированная точность при кросс-валидации; ***AUC*** – площадь под ROC кривой при кросс-валидации.

Из представленных в таблицах 8-10 данных видно, что модели SCLC и SCEC обладают сопоставимыми с SCR значениями сбалансированной точности, однако при этом имеют значительно меньшую размерность. ВА в моделях SCEC превосходит таковую для SCR; в моделях SCLC оказалась ниже во всех случаях, что видно из данных в таблицах 8-10.

3.4.3 Сравнение различных подходов для решения задач виртуального скрининга

Для оценки, обеспечивают ли классификационные модели какие-либо преимущества по сравнению с количественными, мы сравнили характеристики моделей, построенных как исключительно на данных IC_{50}, так и на обогащенных данных. С целью обогащения мы добавляли данные о константах связывания (выборки (+Ki)). Другая группа выборок была составлена с использованием данных о константах связывания и добавлением полностью классификационных данных (выборки (+Ki и неактивные)) (раздел 2.1.6).

В качестве сравнения были созданы модели с применением SCR, обученные исключительно на данных IC_{50}. С использованием SCEC были построены классификационные модели как исключительно на данных по IC_{50}, так и с добавлением данных о константах связывания и с использованием всех имеющихся данных. Результаты сравнения представлены в таблицах 11-13.

Из представленных в таблицах данных видно, что увеличение объема используемых данных и, соответственно, более широкий охват анализируемого химического пространства обеспечивает более высокие показатели прогностической способности.

Так как SCEC позволяет ранжировать соединения, а, следовательно, дает возможность применять к анализируемым выборкам фильтры при проведении виртуального скрининга, согласно формуле (25), были рассчитаны профили фактора обогащения (рисунки 28-30). В качестве анализируемых выборок использовались исключительно выборки с данными об IC_{50}.

При исследовании данной характеристики сравнение проводилось для расчетных значений, получаемых с использованием моделей для ранжированных по экспериментальным значениям полуэффективной концентрации. Сравнение проводилось для ранжирования, произведенного SCEC (см. раздел 2.6), с использованием исходной выборки с данными IC_{50}, так и расширенных выборок. Для SCR использовались только исходные выборки с данными IC_{50}.

Проведенный анализ показывает, что добавление данных при построении классификационных моделей позволяет улучшить ранжирование в выборках, по которым проводится виртуальный скрининг. Эти результаты и охват более широкой части химического пространства такими моделями повышает их эффективность при проведении виртуального скрининга.

Сравнение методов виртуального скрининга на основе $A_{\%}$ согласно формуле (26) по pIC_{50} проводили с помощью модифицированной процедуры скользящего контроля: для обучения отбирались 4/5 части выборки данных с IC_{50}, которые дополнялись всеми остальными данными (Ki и Ki + неактивные), оставшаяся 1/5 часть использовалась в качестве тестовой выборки. Процедура повторялась 5 раз для каждой исследуемой модели. Результаты расчетов $A_{\%}$ представлены на рисунках 31-33, их можно интерпретировать таким образом, что наилучшие результаты получены для классификационных моделей SCEC, обученных на гетерогенных данных без примеси неактивных соединений.

Таблица 11 – Сравнение характеристик при пятикратной кросс-валидации для классификации QSAR моделью и классификационными моделями, построенными на обогащенных выборках, для ингибиторов интегразы ВИЧ-1

Метод	***Выборка***	***N***	***V***	***BA***	***AUC***
SCR	*IN*	*4091*	*392*	*0.846*	*0.925*
SCEC	*IN*	*4091*	*99*	*0.849*	*0.929*
SCEC	*IN (+Ki)*	*4349*	*103*	*0.849*	*0.930*
SCEC	*IN (+Ki и неактивные)*	*12000*	*186*	*0.869*	*0.952*

Таблица 12 – Сравнение характеристик при пятикратной кросс-валидации для классификации QSAR моделью и классификационными моделями, построенными на обогащенных выборках, для ингибиторов протеазы ВИЧ-1

Метод	***Выборка***	***N***	***V***	***BA***	***AUC***
SCR	*PR*	*6552*	*470*	*0.860*	*0.932*
SCEC	*PR*	*6552*	*130*	*0.866*	*0.941*
SCEC	*PR (+Ki)*	*8545*	*163*	*0.889*	*0.953*
SCEC	*PR (+Ki и неактивные)*	*12000*	*190*	*0.891*	*0.954*

Таблица 13 – Сравнение характеристик при пятикратной кросс-валидации для классификации QSAR моделью и классификационными моделями, построенными на обогащенных выборках, для ингибиторов обратной транскриптазы ВИЧ-1

Метод	***Выборка***	***N***	***V***	***BA***	***AUC***
SCR	*RT*	*6309*	*452*	*0.790*	*0.870*
SCEC	*RT*	*6309*	*124*	*0.793*	*0.876*
SCEC	*RT (+Ki)*	*7508*	*155*	*0.801*	*0.899*
SCEC	*RT (+Ki и неактивные)*	*12000*	*189*	*0.800*	*0.905*

N – количество соединений в обучающей выборке; ***V*** – размерность модели; ***BA*** – сбалансированная точность при кросс-валидации; ***AUC*** – площадь под ROC кривой при кросс-валидации.

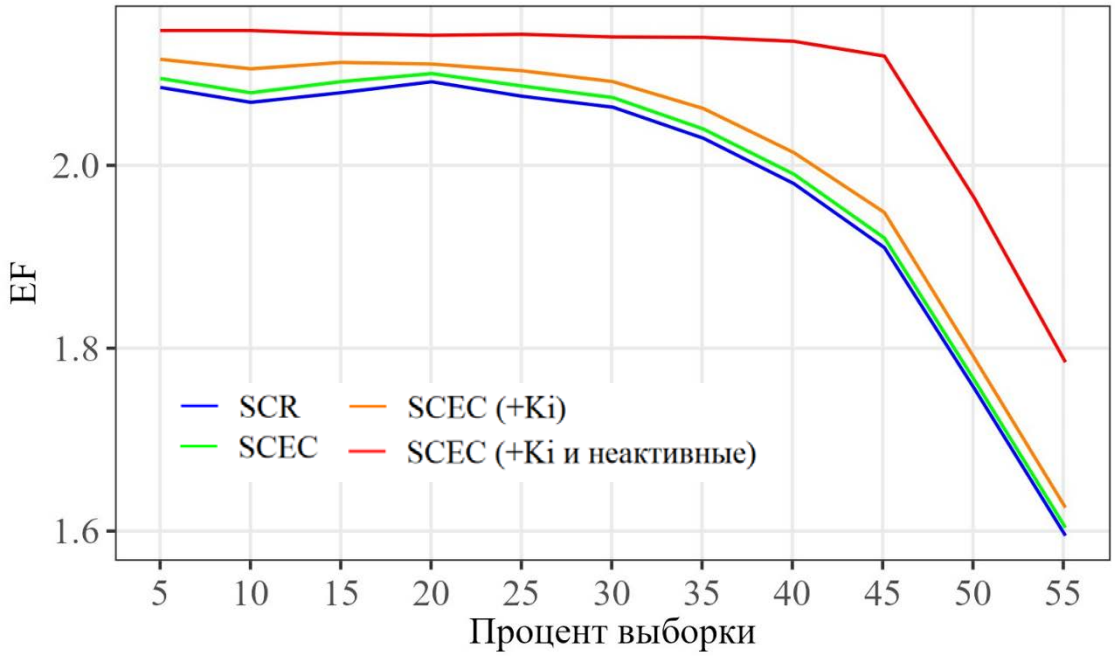


Рисунок 28 – Профили фактора обогащения для выборки IN: SCEC (выборка IN), SCEC (выборка IN (+Ki)), SCEC (выборка IN (+Ki и неактивные)), SCR (классификационная модель), SCR (QSAR модель)

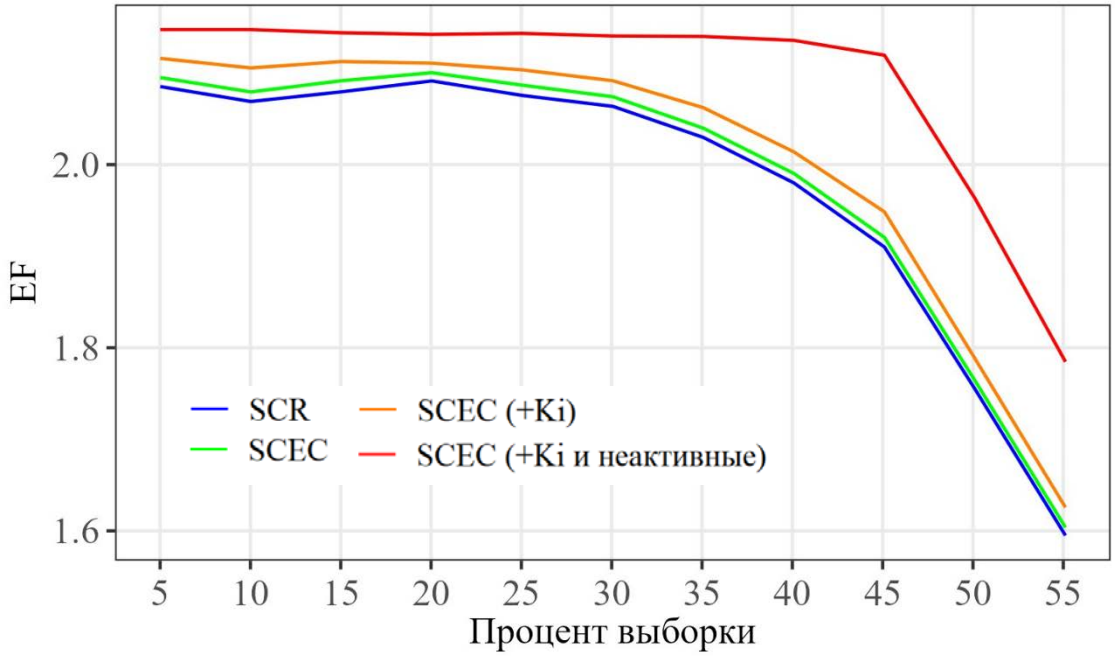


Рисунок 29 – Профили фактора обогащения для выборки PR: SCEC (выборка PR), SCEC (выборка PR (+Ki)), SCEC (выборка PR (+Ki и неактивные)), SCR (классификационная модель), SCR (QSAR модель)

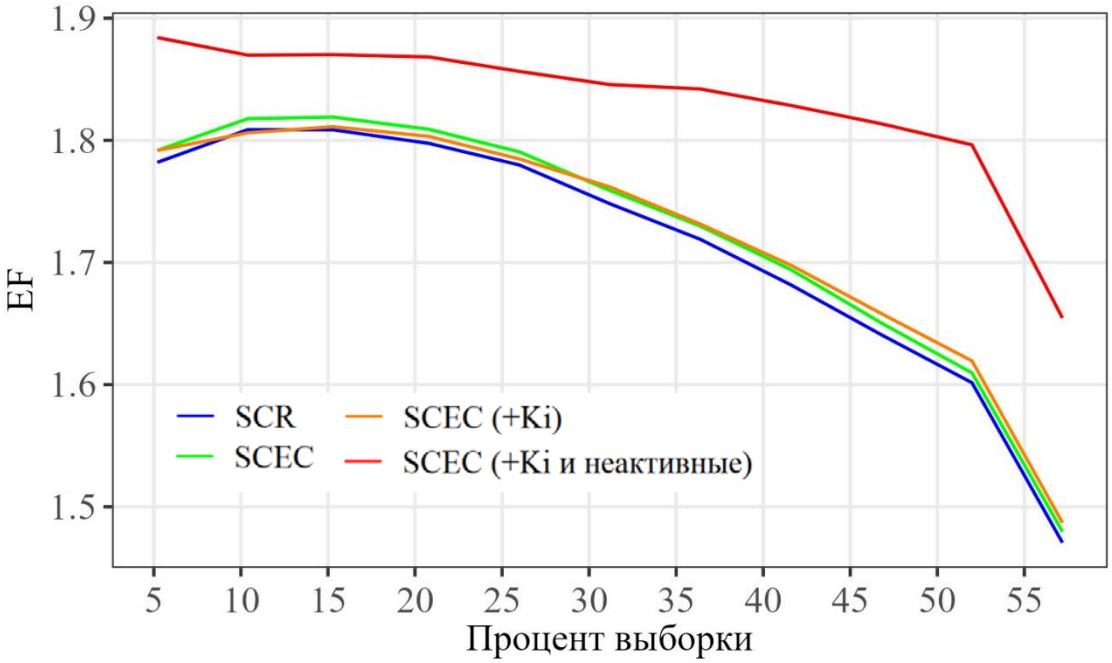


Рисунок 30 – Профили фактора обогащения для выборки RT: SCEC (выборка RT), SCEC (выборка RT (+Ki)), SCEC (выборка RT (+Ki и неактивные)), SCR (классификационная модель), SCR (QSAR модель)

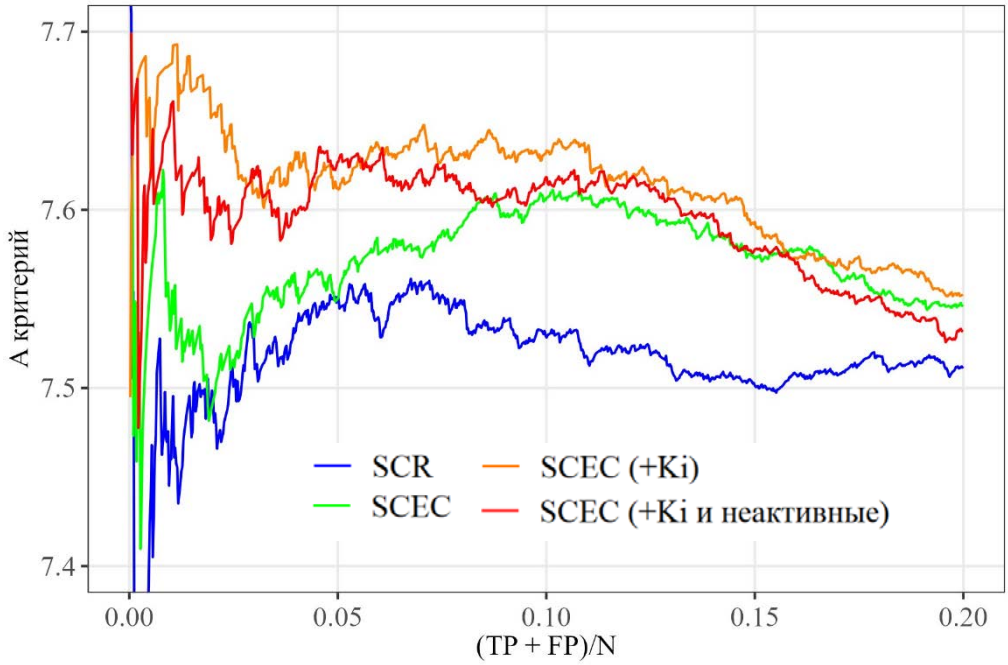


Рисунок 31 – Профили $A_{\%}$ по pIC_{50} для выборки IN: SCEC (выборка IN), SCEC (выборка IN (+Ki)), SCEC (выборка IN (+Ki и неактивные)), SCR (выборка IN)

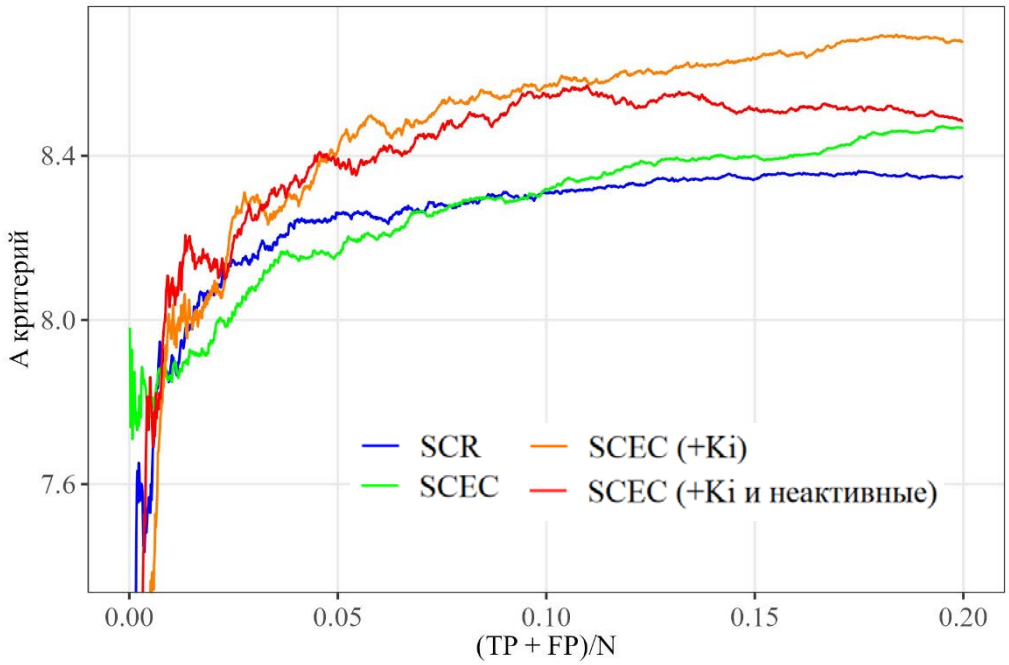


Рисунок 32 – Профили $A_{\%}$ по pIC_{50} для выборки PR: SCEC (выборка PR), SCEC (выборка PR (+Ki)), SCEC (выборка PR (+Ki и неактивные)), SCR (выборка PR)

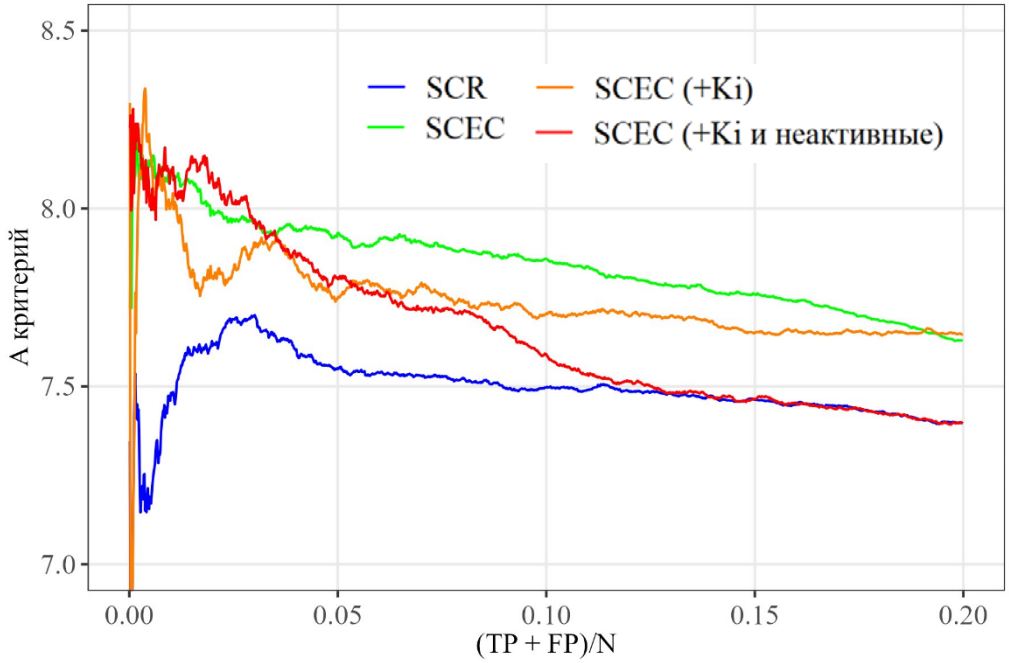


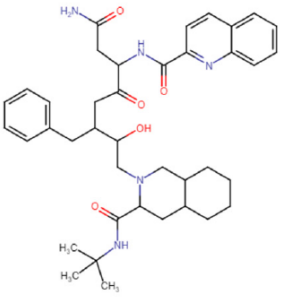
Рисунок 33 – Профили $A_{\%}$ по pIC_{50} для выборки RT: SCEC (выборка RT), SCEC (выборка RT (+Ki)), SCEC (выборка RT (+Ki и неактивные)), SCR (выборка RT)

3.4 Веб сервис AntiHIV Pred

На основе полученных (Q)SAR моделей нами был реализован свободно доступный веб-сервис AntiHIV Pred (http://way2drug.com/HIV/) (рисунок 34).

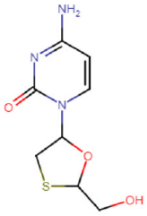
Рисунок 34 – Веб сервис AntiHIV Pred

Этот веб-сервис позволяет получить прогноз активности на основе структурной формулы соединения для каждой из мишеней ВИЧ-1 с использованием выбранной модели, соответствующей конкретной мишени ВИЧ-1 и источнику данных, на которых была построена модель (рисунок 35). Прогноз осуществляется для однокомпонентных незаряженных молекул, с молекулярной массой менее 1250 а.е.м. и содержащих не менее 3-х атомов углерода. В дополнение к анализу «структура-активность» для мишеней ВИЧ-1 веб-сервис дает возможность прогнозирования различных видов активности соединений с целью терапии 83 ВИЧ-ассоциированных заболеваний, что может быть использовано для создания ингибиторов двойного действия. Инструкция по использованию разработанного нами веб-сервиса представлена в приложении 5.



HIV TARGETS PREDICTION RESULTS

Targets	Value	Prediction	Applicability Domain
Protease (HIV-1):	p(IC50)	8.977	in AD
Reverse Transcriptase (HIV-1):	p(IC50)	7.404	not in AD
Integrase (HIV-1):	p(IC50)	4.604	not in AD



HIV COMORBIDITIES ACTIVITY SPECTRA

Comorbidity	Pa	Pi
Antineoplastic (myeloid leukemia)	0.777	0.007
Myelodysplastic syndrome treatment	0.768	0.005
Antiviral (Hepatitis B)	0.71	0.002
Antileukemic	0.607	0.009

Рисунок 35 – Пример результата прогноза. Верхняя таблица – значение рассчитанной активности для мишеней ВИЧ-1. Нижняя таблица – вероятность активности и неактивности для ВИЧ-ассоциированных заболеваний

ЗАКЛЮЧЕНИЕ

В настоящее время остается актуальным ряд задач, связанных с поиском новых фармакологически активных веществ. Современные подходы, которые используются для решения этих задач, должны учитывать возрастающую сложность, проявляющуюся в гетерогенности этих данных.

В диссертационной работе была исследована возможность построения обобщенных моделей для классификационных прогнозов на примере данных Tox21, ингибиторов молекулярных мишеней ВИЧ-1 и SARS-CoV-2. Продемонстрировано, что разработанные нами методы SCLC и SCEC позволяют создавать классификационные модели с высокой прогностической способностью за счет внедрения в алгоритм построения модели балансировки выборок и выбора ключевых переменных, обуславливающих активность. Сравнение точности методов с созданными ранее методами SAR, в частности, с SCR, показало, что сопоставимые характеристики моделей могут быть получены при существенно меньшей размерности используемых переменных.

Анализ доступных данных показал преимущества классификационных методов, связанные с возможностью описания более широкого химического пространства, для проведения виртуального скрининга при использовании гетерогенных данных. Проведенный количественный и классификационный анализ взаимосвязей «структура-активность» в отношении ингибиторов ВИЧ-1 является опорной точкой для поиска новых фармакологически активных соединений в рамках виртуального скрининга. Реализованный в виде веб-сервиса AntiHIV Pred этот инструментарий доступен широкому кругу исследователей. Наряду с прогнозом противовирусной активности, сервис обеспечивает и прогноз спектра активности соединения по отношению к сопутствующим ВИЧ заболеваниям, что обеспечивает дополнительную поддержку принятия решения для отбора соответствующей структуры для синтеза и тестирования и дает возможность многокритериальной оценки соединения как потенциального лекарственного препарата.

ВЫВОДЫ

1. Для решения задач классификации предложены и реализованы алгоритмы самосогласованной логистической и экстремальной классификации (SCLC и SCEC).
2. Валидация разработанных нами методов на сгенерированных искусственных данных и выборках Tox21 показала, что SCEC и SCLC не уступают по точности альтернативным подходам, но позволяют создавать модели зависимости «структура-активность» с более высокой прогностической способностью при меньшем числе независимых переменных.
3. Разработана процедура извлечения и обработки сведений о биологически активных веществах, сформированы выборки гетерогенных данных об ингибиторах протеазы, обратной транскриптазы и интегразы ВИЧ-1, а также об ингибиторах главной протеазы 3CLpro, папаин-подобной протеазы PLpro, РНК-полимеразы RdRp SARS-CoV-2.
4. Построены количественные и классификационные модели зависимостей «структура-активность» для ингибиторов белков ВИЧ-1, оценка которых показала преимущества применения классификационных моделей при использовании гетерогенных данных.
5. Показано, что разработанные классификационные модели для ингибиторов ферментов SARS-CoV-2 обладают удовлетворительной точностью и предсказательной способностью.
6. Создан свободно доступный веб-ресурс AntiHIV Pred, который предоставляет широкому кругу исследователей возможность осуществлять выбор наиболее перспективных молекул с требуемой активностью для синтеза на основе виртуального скрининга.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в рецензируемых журналах

1. Stolbov L.A., Filimonov, D.A., Poroikov, V.V. SAR based on Self Consistent Classifier // SAR and QSAR in Environmental Research. – 2022.
2. Druzhilovskiy D.S., Stolbov L.A., Savosina P.I., Pogodin P.V., Filimonov D.A., Veselovsky A.V., Stefanisko K., Tarasova N.I., Nicklaus M., Poroikov V.V. Computational approaches to identify a hidden pharmacological potential in large chemical libraries // Supercomputing frontiers and innovations. – 2020. – V. 7. – P. 57-76.
3. Stolbov L.A., Druzhilovskiy D.S., Filimonov D.A., Nicklaus M., Poroikov V.V. (Q)SAR models of HIV-1 proteins inhibition by drug-like compounds // Molecules. – 2020. – V. 25(1). – P. 87.
4. Stolbov L.A., Rudik A.V., Filimonov D.A., Poroikov V.V., Nicklaus M. AntiHIV-Pred: Web-resource for in silico prediction of anti-HIV/AIDS activity // Bioinformatics. – 2020. – V. 36(3). – P. 978–979.
5. Поройков В.В., Филимонов Д.А., Глориозова Т.А., Лагунин А.А., Дружиловский Д.С., Рудик А.В., Столбов Л.А., Дмитриев А.В., Тарасова О.А., Иванов С.М., Погодин П.В. Компьютерный прогноз спектров биологической активности органических соединений: возможности и ограничения // Вестник РАН серия химическая. – 2019. – № 12. – С. 2143–2154.
6. Савосина П.И., Столбов Л.А., Дружиловский Д.С., Филимонов Д.А., Никлаус Марк, Поройков В.В. Поиск новых антиретровирусных соединений в химическом пространстве “Больших Данных” библиотеки SAVI // Биомедицинская химия. – 2019. – № 65(2). – С. 73–79.

Свидетельства на регистрацию программ для ЭВМ

1. Столбов Л.А., Филимонов Д.А., Поройков В.В., программа Sarmath, свидетельство на программу для ЭВМ: Рег. Номер: RU2022662930, выдано 07.07.2022 Федеральной службой по интеллектуальной собственности.

Работы, опубликованные в сборниках материалов научных конференций

1. Stolbov L.A., Filimonov D.A., Poroikov,V.V. An Application of Self Consistent Classifier to predict the inhibitors of Cytochromes P450 // IX Международная

конференция молодых ученых: вирусологов, биотехнологов, биофизиков, молекулярных биологов и биоинформатиков, 2022. – Новосибирск, 2022. – С. 48.

2. Druzhilovskiy D.S., Savosina P.I., Biziukova N.Yu., Filimonov D.A., Ivanov S.M., Karasev D.A., Pogodin P.V., Stolbov L.A., Tarasova O.A., Veselovsky A.V., Poroikov V.V. Analysis of big multidisciplinary data for optimization of the development and utilization of pharmaceutical agents against coronavirus infection // XXXVIII Symposium of Bioinformatics and Computer-Aided Drug Discovery. Proceedings book. – Moscow, 2022. – P. 106.
3. Stolbov L.A., Filimonov D.A., Poroikov V.V. Self consistent classifier sar approach // XXXVIII Symposium of Bioinformatics and Computer-Aided Drug Discovery. Proceedings book. – Moscow, 2022. – P. 56.
4. Поройков В.В., Дружиловский Д.С., Погодин П.В., Столбов Л.А., Иванов С.М., Савосина П.И., Ионов Н.С., Бизюкова Н.Ю., Веселовский А.В., Филимонов Д.А. Компьютерный поиск фармакологических веществ для терапии инфекции sars-cov-2: иллюзии и реальность // 5-я Российская конференция МедХим-Россия. – Волгоград, 2021. – С. 164.
5. Поройков В.В., Дружиловский Д.С., Савосина П.И., Гомазков О.А., Соболев Б.Н., Веселовский А.В., Тарасова О.А., Бизюкова Н.Ю., Ионов Н.С., Иванов С.М., Рудик А.В., Столбов Л.А., Погодин П.В., Карасев Д.А., Дмитриев Д.А., Лагунин А.А., Филимонов Д.А. Репозиционирование лекарств для терапии covid-19 // Биотехнология: состояние и перспективы развития. материалы международного конгресса. – Москва, 2021. – С. 209–211.
6. Савосина П.И., Столбов Л.А., Погодин П.В., Дружиловский Д.С., Поройков В.В. Covid-19: анализ практики репозиционирования лекарственных препаратов // Сборник тезисов докладов Шестой Междисциплинарной конференции «Молекулярные и биологические аспекты химии, фармацевтики и фармакологии». – Нижний Новгород, 2020. – С. 224.
7. Поройков В.В., Дружиловский Д.С., Филимонов Д.А., Веселовский А.А., Столбов Л.А., Погодин П.В., Карасев Д.А., Савосина П.И., Соболев Б.Н. Поиск

соединений с антивирусной активностью среди миллиарда лекарственноподобных соединений // Сборник тезисов докладов Шестой Междисциплинарной конференции «Молекулярные и биологические аспекты химии, фармацевтики и фармакологии». – Нижний Новгород, 2020. – С. 87.

8. Столбов Л.А., Дружиловский Д.С., Рудик А.В., Филимонов Д.А., Поройков В.В. Построение (Q)SAR моделей для ингибиторов белков ВИЧ-1 // Сборник трудов XXVII Российского национального конгресса «Человек и лекарство». – 2020. – С. 71–72.
9. Poroikov V.V., Druzhilovskiy D.S., Filimonov D.A., Veselovsky A.V., Pogodin P.V., Stolbov, L.A., Savosina, P.I., Pevzner Yu., Patel H., Nicklaus M. Synthetically available virtual inventory: can new anti-HIV agents be found among the 283 mln molecules? // Third international School-Seminar «From Empirical to Predictive Chemistry». – Kazan, 2018. – P. 13.
10. Stolbov L.A., Pogodin P.V., Druzhilovskiy D.S., Filimonov D.A., PoroikovV.V. QSAR models for predicting HIV-1 protease inhibiting activity among the molecules from SAVI // Third international School-Seminar «From Empirical to Predictive Chemistry». – Kazan, 2018. – P. 40.
11. Столбов Л.А., Погодин П.В., Дружиловский Д.С., Филимонов Д.А., Поройков В.В. Поиск ингибиторов протеазы ВИЧ-1 среди потенциально синтезируемых соединений (SAVI) с использованием количественных моделей взаимосвязи // Сборник трудов XXV Российского национального конгресса «Человек и лекарство». – 2018. – С. 86–87.
12. Druzhilovskiy D.S., Filimonov D.A., Veselovsky A.A., Bezhentsev V., Stolbov L.A., Poroikov V.V., Nicklaus M. Analysis of large chemical-biological space aimed at discovery of novel anti-HIV agents // 256th ACS National Meeting. – Boston, 2018, C 151.
13. Stolbov L.A., Druzhilovskiy D.S., Filimonov D.A., Rudik A.V., Poroikov V.V. Web-service for antiretroviral activity prediction based on models of quantitative

structure-activity relationships // IX Международный конгресс «Биотехнология: состояние и перспективы развития. Науки о жизни».– Москва, 2019.

14. Druzhilovskiy D.S., Filimonov D.A., Pogodin P.V., Stolbov L.A., Savosina P.I., Tarasova N.I., Nicklaus M., Poroikov V.V. AI-based Big Data Application in Drug Discovery // 259th ACS National Meeting. – Philadelphia, 2020, C 169.

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа выполнена при поддержке гранта РФФИ-NIH № 17-54-30015-NIH_a и гранта РНФ № 19-15-00396.

БЛАГОДАРНОСТИ

▪Автор выражает благодарность научному руководителю доктору биологических наук, профессору, члену-корреспонденту РАН Поройкову Владимиру Васильевичу за руководство на всех этапах исследования.

▪Автор выражает благодарность научному консультанту кандидату физико-математических наук Дмитрию Алексеевичу Филимонову за существенный вклад в математический аппарат и поддержку при разработке методов SCLC и SCEC.

▪Автор выражает благодарность кандидату биологических наук Дмитрию Сергеевичу Дружиловскому, кандидату биологических наук Анастасии Владимировне Рудик за помощь в разработке веб-сервиса.

СПИСОК ЛИТЕРАТУРЫ

[1] Geronikaki A., Eleftheriou P., Poroikov V. Anti-HIV Agents: Current Status and Recent Trends // *Communicable Diseases of the Developing World*. 2016. P. 37–95. doi: 10.1007/7355_2015_5001.

[2] Adamson C.S., Chibale K., Goss R. J. M., Jaspars M., Newman D. J., Dorrington R. A. Antiviral drug discovery: preparing for the next pandemic // *Chem Soc Rev*. 2021. V. 50, Iss. 6. P. 3647–3655. doi: 10.1039/D0CS01118E.

[3] Muratov E.N., Amaro R., Andrade C.H., Brown N., Ekins S., Fouches D., Isayev O., Kozakov D., Medina-Franco J.L., Mertz K.M., Oprea T.I., Poroikov V., Scneider G., Todd M.H., Varnek A., Winkler D.A., Zakharov A. V., Cherkasov A., Tropsha A. A critical overview of computational approaches employed for COVID-19 drug discovery // *Chem Soc Rev*. 2021. V. 50, Iss. 16. P. 9121-9151. doi: 10.1039/D0CS01065K.

[4] Sulimov V. B., Kutov D. C., Taschilova A. S., Ilin I. S., Tyrtyshnikov E. E., Sulimov A. V. Docking Paradigm in Drug Design // *CTMC*. 2021. V.21, Iss. 6. doi: 10.2174/1568026620666201207095626.

[5] Gentile F., Yaacoub J.C., Gleave J., Fernandez M., Ton A.T., Ban F., Stern A., Cherkasov. Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking» // *Nat Protoc*. 2022. V. 17, Iss. 3. P. 672–697. doi: 10.1038/s41596-021-00659-2.

[6] Muratov E. N., Bajorath J., Sheridan R.P., Tetko I.V., Filimonov D., Poroikov V., Oprea T.I., Baskin I.I., Varnek A., Roitberg A., Isayev O., Curtalolo S., Fouchers D., Cjhen Y., Aspuru-Guzik A., Winkler D.A., Agrafiotis D., Cherkasov A., Tropsha A. QSAR without borders // *Chem Soc Rev*. 2020. V. 49, Iss. 11. P. 3525–3564. doi: 10.1039/D0CS00098A.

[7] Franke R., Gruska A. General Introduction to QSAR // In book: *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens*. 2003. ch1. doi: 10.1201/9780203010822.

[8] Medina-Franco J. L., Chávez-Hernández A. L., López-López E., Saldívar-González F. I. Chemical Multiverse: An Expanded View of Chemical Space // *Mol Informatics*. 2022. V. 41. Iss. 11. P. 2200116. doi: 10.1002/minf.202200116.

[9] Ishola A. A., Adedirin O., Joshi T., Chandra S. QSAR modeling and pharmacoinformatics of SARS coronavirus 3C-like protease inhibitors // *Computers in Biology and Medicine*. 2021. V. 134. P. 104483. doi: 10.1016/j.compbiomed.2021.104483.

[10] Lagunin A., Zakharov A., Filimonov D., Poroikov V. QSAR Modelling of Rat Acute Toxicity on the Basis of PASS Prediction // *Mol Inf*. 2011. V. 30. Iss. 2–3. P. 241–250. doi: 10.1002/minf.201000151.

[11] Sakamuru S., Huang R., Xia M. Use of Tox21 Screening Data to Evaluate the COVID-19 Drug Candidates for Their Potential Toxic Effects and Related Pathways // *Front Pharmacol*. 2022. V. 13. P. 935399. doi: 10.3389/fphar.2022.935399.

[12] Huang R., Xia M. Editorial: Tox21 Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways As Mediated by Exposure to Environmental Toxicants and Drugs // *Fron. Environ. Sci*. 2017. V. 5. doi: 10.3389/fenvs.2017.00003.
[13] Kapetanovic I.M. Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach // *Chem Biol Interact*. 2008. V. 171. Iss. 2. P. 165–176. doi: 10.1016/j.cbi.2006.12.006.
[14] Shaikh S., Jain T., Sandhu G., Latha N., Jayaram B. From Drug Target to Leads-Sketching A Physicochemical Pathway for Lead Molecule Design In Silico // *CPD*. 2007. V. 13. Iss. 34. P. 3454–3470. doi: 10.2174/138161207782794220.
[15] Jorgensen W.L. The Many Roles of Computation in Drug Discovery // *Science*. 2004. V. 303. Iss. 5665. P. 1813–1818. doi: 10.1126/science.1096361.
[16] Kroemer R.T. Structure-Based Drug Design: Docking and Scoring // *CPPS*. 2007. V. 8. Iss. 4. P. 312–328. doi: 10.2174/138920307781369382.
[17] Carvalho A.L., Santos-Silva T., Romao M.J., Cabrita E.J., Marcelo F. Structural Elucidation of Macromolecules // In book: *Essential Techniques for Medical and Life Scientists: A Guide to Contemporary Methods and Current Applications with the Protocols*. 2018. P. 30–91. doi: 10.2174/9781681087092118010005.
[18] Krieger E., Nabuurs S.B., Vriend G. Homology Modeling // In book: *Methods of Biochemical Analysis*. 2005. P. 509–523. doi: 10.1002/0471721204.ch25.
[19] Cavasotto C.N., Phatak S.S. Homology modeling in drug discovery: current trends and applications // *Drug Discovery Today*. 2009. V. 14. Iss. 13–14. P. 676–683. doi: 10.1016/j.drudis.2009.04.006.
[20] Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., Tunyasuvunakool K., Bates R., Zidek A., Potapenko A., Bridgland A., Meyer C., Kohl A.A.N., Ballard A.J., Cowie A., Romera-Paredes B., Nikolov S., Jain R., Adler J., Back T., Petersen S., Reiman D., Clancy E., Zielinski M., Steinegger M., Pacholska M., Berghammer T., Bodenstein S., Silver D., Vinyals O., Senior A.W., Kavukcuoglu K., Lohli P., Hassabis D. Highly accurate protein structure prediction with AlphaFold // *Nature*. 2021. V. 596. Iss. 7873. P. 583–589. doi: 10.1038/s41586-021-03819-2.
[21] «Protein Data Bank». https://www.rcsb.org/
[22] Burley S.K., Berman H.M., Duarte J.M., Feng Z., Flatt J.W., Hudson B.P., Lowe R., Peisach E., Piehl D.W., Rose Y., Sali A., Sekharan M., Shao C., Vallat B., Voint M., Westbrook J.D., Young J.Y., Zardecki C. Protein Data Bank: A Comprehensive Review of 3D Structure Holdings and Worldwide Utilization by Researchers, Educators, and Students // *Biomolecules*. 2022. V. 12. Iss. 10. P. 1425. doi: 10.3390/biom12101425.
[23] Timofeev V., Samygina V. Protein Crystallography: Achievements and Challenges // *Crystals*. 2023. V. 13 № 1. P. 71. doi: 10.3390/cryst13010071.
[24] Laskowski R.A., Swaminathan G.J. Problems of Protein Three-Dimensional Structures // In book: *Reference Module in Chemistry, Molecular Sciences and Chemical Engineering*. 2013. P. B978012409547202535X. doi: 10.1016/B978-0-12-409547-2.02535-X.

[25] Warren G.L., Do T.D., Kelley B.P., Nicholls A., Warren S.D. Essential considerations for using protein–ligand structures in drug discovery // *Drug Discovery Today*. 2012 V. 17. Iss. 23–24. P. 1270–1281. doi: 10.1016/j.drudis.2012.06.011.
[26] Meng X.-Y., Zhang H.-X., Mezei M., Cui M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery // *CAD*. 2011. V. 7. Iss. 2. P. 146–157. doi: 10.2174/157340911795677602.
[27] Li J., Fu A., Zhang L. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking // *Interdiscip Sci Comput Life Sci*. 2019. V. 11. Iss. 2. P. 320–328. doi: 10.1007/s12539-019-00327-w.
[28] Dar A.M., Mir S. Molecular Docking: Approaches, Types, Applications and Basic Challenges // *J Anal Bioanal Tech*. 2017. V. 08. Iss. 02. doi: 10.4172/2155-9872.1000356.
[29] Salmaso V., Moro S. Bridging Molecular Docking to Molecular Dynamics in Exploring Ligand-Protein Recognition Process: An Overview // *Front Pharmacol.* 2018. V. 9. P. 923. doi: 10.3389/fphar.2018.00923.
[30] Shoichet B. K. Virtual screening of chemical libraries // *Nature*. 2004. V. 432. Iss. 7019. P. 862–865. doi: 10.1038/nature03197.
[31] Schneider G., Fechner U. Computer-based de novo design of drug-like molecules // *Nat Rev Drug Discov*. 2005. V. 4. Iss. 8. P. 649–663. doi: 10.1038/nrd1799.
[32] Anderson A.C. The Process of Structure-Based Drug Design // *Chem Biol*. 2003. V. 10. Iss. 9. P. 787–797. doi: 10.1016/j.chembiol.2003.09.002.
[33] Wlodawer A., Vondrasek J. Inhibitors of HIV-1 protease A Major Success of Structure-Assisted Drug Design // *Annu Rev Biophys Biomol Struct.* 1998. V. 27. Iss. 1. P. 249–284. doi: 10.1146/annurev.biophys.27.1.249.
[34] Clark D. E. What has computer-aided molecular design ever done for drug discovery? // *Expert Opin Drug Discov*. 2006. V. 1. Iss. 2. P. 103–110. doi: 10.1517/17460441.1.2.103.
[35] Marrakchi H., Lanéelle G., Quémard A. InhA, a target of the antituberculous drug isoniazid, is involved in a mycobacterial fatty acid elongation system, FAS-II // *Microbiology*. 2000. V. 146. Iss. 2. P. 289–296. doi: 10.1099/00221287-146-2-289.
[36] Wang L., Gu Q., Zheng X., Ye J., Liu Z., Li J., Hu X., Hagler A., Xu J. Discovery of New Selective Human Aldose Reductase Inhibitors through Virtual Screening Multiple Binding Pocket Conformations // *J Chem Inf Model.* 2013. V. 53. Iss. 9. P. 2409–2422. doi: 10.1021/ci400322j.
[37] Dadashpour S., Tuylu Kucukkilinc T., Tan O. U., Ozadali K., Irannejad H., Emami S. Design, Synthesis and *In Vitro* Study of 5,6-Diaryl-1,2,4-triazine-3-ylthioacetate Derivatives as COX-2 and β-Amyloid Aggregation Inhibitors: Triazine-3-thioacetates as COX-2 and Aβ Formation Inhibitors // *Arch Pharm Chem Life Sci.* 2015. V. 348. Iss. 3. P. 179–187.doi: 10.1002/ardp.201400400.
[38] Miller Z., Kim K.-S., Lee D.-M., Kasam V., Baek S.E., Lee K.H., Zhang Y.-Y., Ao L., Carmony K., Lee N.-R., Zhou S., Zhao Q., Jang Y., Jeong H.-Y., Zhan C.-G., Lee W., Kim D.-E., Kim K.B. Proteasome Inhibitors with Pyrazole Scaffolds from

Structure-Based Virtual Screening // *J Med Chem.* 2015. V. 58. Iss. 4. P. 2036–2041. doi: 10.1021/jm501344n.

[39] Grover S., Apushkin M.A., Fishman G. A. Topical Dorzolamide for the Treatment of Cystoid Macular Edema in Patients With Retinitis Pigmentosa // *Am J Ophthalmol*. 2006. V. 141. Iss. 5. P. 850–858. doi: 10.1016/j.ajo.2005.12.030.

[40] Ren J.-X., Li L.-L., Zheng R.-L., Xie H.-Z., Cao Z.-Q., Feng S., Pan Y.-L., Chen X., Wei Y.-Q., Yang S.-Y. Discovery of Novel Pim-1 Kinase Inhibitors by a Hierarchical Multistage Virtual Screening Approach Based on SVM Model, Pharmacophore, and Molecular Docking // *J Chem Inf Model.* 2011. V. 51. Iss. 6. P. 1364–1375. doi: 10.1021/ci100464b.

[41] Matsuno K., Masuda Y., Uehara Y., Sato H., Muroya A., Takanashi O., Yokotagawa T., Furuya T., Okawara T., Otsuka M., Ogo N., Ashizawa Y., Akiyama Y., Asai A. Identification of a New Series of STAT3 Inhibitors by Virtual Screening // *ACS Med Chem Lett.* 2010. V. 1. Iss. 8. P. 371–375. doi: 10.1021/ml1000273.

[42] Elyashberg M. «dentification and structure elucidation by NMR spectroscopy // *TrAC Trends in Analytical Chemistry*. 2015. V. 69. P. 88–97. doi: 10.1016/j.trac.2015.02.014.

[43] Bender A. How similar are those molecules after all? Use two descriptors and you will have three different answers // *Expert Opin Drug Discov*. 2010. V. 5. Iss. 12. P. 1141–1151. doi: 10.1517/17460441.2010.517832.

[44] Macarron R. How dark is HTS dark matter? // *Nat Chem Biol*. 2015. V. 11. Iss. 12. P. 904–905. doi: 10.1038/nchembio.1937.

[45] Manelfi C., Gemei M., Talarici C., Cerchia C., Fava A., Lunghini F., Beccari A.R.*u ∂p.* "Molecular Anatomy": a new multi-dimensional hierarchical scaffold analysis tool // *J Cheminform*. 2021. V. 13. Iss. 1. P. 54. doi: 10.1186/s13321-021-00526-y.

[46] Hu Y., Stumpfe D., Bajorath J. Lessons Learned from Molecular Scaffold Analysis // *J Chem Inf Model.* 2011. V. 51. Iss. 8. P.. 1742–1753. doi: 10.1021/ci200179y.

[47] Englert P., Kovács P. Efficient Heuristics for Maximum Common Substructure Search // *J Chem Inf Model.* 2015. V. 55. Iss. 5. P. 941–955. doi: 10.1021/acs.jcim.5b00036.

[48] Bajusz D., Rácz A., Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? // *J Cheminform*. 2015. V. 7. Iss. 1. P. 20. doi: 10.1186/s13321-015-0069-3.

[49] Chakravarti S. K. Distributed Representation of Chemical Fragments // *ACS Omega*. 2018. V. 3. Iss. 3, P. 2825–2836. doi: 10.1021/acsomega.7b02045.

[50] Rugard M., Jaylet T., Taboureau O., Tromelin A., Audouze K. Smell compounds classification using UMAP to increase knowledge of odors and molecular structures linkages // *PLoS ONE*. 2021. V. 16. Iss. 5. P. e0252486. doi: 10.1371/journal.pone.0252486.

[51] Probst D., Reymond J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees // *J Cheminform*. 2020. V. 12. Iss. 1. P. 12. doi: 10.1186/s13321-020-0416-x.

[52] Ortega F., Algar M.J., Diego I.M., Moguerza J.M. Unconventional application of k-means for distributed approximate similarity search // *IR*. 2022. doi: 10.48550/ARXIV.2208.02734.
[53] Stumpfe D., Hu H., Bajorath J. Advances in exploring activity cliffs // *J Comput Aided Mol Des*. 2020. V. 34. Iss. 9. P. 929–942. doi: 10.1007/s10822-020-00315-z.
[54] Guha R., Van Drie J.H. Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs // *J Chem Inf Model.* 2008. V. 48. Iss. 3. P. 646–658. doi: 10.1021/ci7004093.
[55] Baldi P., Nasr R. When is Chemical Similarity Significant? The Statistical Distribution of Chemical Similarity Scores and Its Extreme Values // *J Chem Inf Model.* 2010. V. 50. Iss. 7. P. 1205–1222. doi: 10.1021/ci100010v.
[56] Martin Y.C., Kofron J.L., Traphagen L.M. Do Structurally Similar Molecules Have Similar Biological Activity? // *J Med Chem.* 2002. V. 45. Iss. 19. P. 4350–4358. doi: 10.1021/jm020155c.
[57] Khedkar S., Malde A., Coutinho E., Srivastava S. Pharmacophore Modeling in Drug Discovery and Development: An Overview // *MC*. 2007. V. 3. Iss. 2. P. 187–197. doi: 10.2174/157340607780059521.
[58] Muhammed M.T., Aki-Yalcin E. Pharmacophore Modeling in Drug Discovery: Methodology and Current Status // *J Turk Chem Soc Sect A Chem.* 2021. P. 759–772. doi: 10.18596/jotcsa.927426.
[59] Duarte C., Barreiro E., Fraga C. Privileged Structures: A Useful Concept for the Rational Design of New Lead Drug Candidates // *MRMC*. 2007. V. 7. Iss. 11. P. 1108–1119. doi: 10.2174/138955707782331722.
[60] Leelananda S.P., Lindert S. Computational methods in drug discovery // *J. Org. Chem.* 2016. V. 12. P. 2694–2718. doi: 10.3762/bjoc.12.267.
[61] Peach M.L., Nicklaus M.C. Combining docking with pharmacophore filtering for improved virtual screening // *J Cheminform*. 2009. V. 1. Iss. 1. P. 6. doi: 10.1186/1758-2946-1-6.
[62] Hindle S.A., Rarey M., Buning C., Lengauer T. Flexble dockind under pharmacophore type constraints // *J Comput Aided Mol Des.* 2002. V. 16. Iss. 2. P. 129–149. doi: 10.1023/A:1016399411208.
[63] Verma J., Khedkar V., Coutinho E. 3D-QSAR in Drug Design - A Review // *CTMC*. 2010. V. 10. Iss. 1.P. 95–115. doi: 10.2174/156802610790232260.
[64] Hung C., Gini G. QSAR modeling without descriptors using graph convolutional neural networks: the case of mutagenicity prediction // *Mol Divers*. 2021. V. 25. Iss. 3. P. 1283–1299. doi: 10.1007/s11030-021-10250-2.
[65] Geppert H., Horváth T., Gärtner T., Wrobel S., Bajorath J. Support-Vector-Machine-Based Ranking Significantly Improves the Effectiveness of Similarity Searching Using 2D Fingerprints and Multiple Reference Compounds // *J Chem Inf Model.* 2008. V. 48. Iss. 4. P. 742–746. doi: 10.1021/ci700461s.
[66] Wold S., Sjöström M., Eriksson L. PLS-regression: a basic tool of chemometrics // *Chemom Intell Lab Syst.* 2001. V. 58. Iss. 2. P. 109–130. doi: 10.1016/S0169-7439(01)00155-1.

[67] Varnek A., Baskin I. I. Chemoinformatics as a Theoretical Chemistry Discipline // *Mo. Inf.* 2011. V. 30. Iss. 1. P. 20–32. doi: 10.1002/minf.201000100.
[68] Ertl P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups // *J Chem Inf Comput Sci.* 2003. V. 43. Iss. 2. P. 374–380. doi: 10.1021/ci0255782.
[69] Kirkpatrick P., Ellis C. Chemical space // *Nature*. 2004. V. 432. Iss. 7019. P. 823–823. doi: 10.1038/432823a.
[70] «ChEMBL database URL». https://www.ebi.ac.uk/chembl/
[71] «PubChem database URL ». https://pubchem.ncbi.nlm.nih.gov/
[72] «DrugBank database URL ». https://go.drugbank.com/
[73] «ZINC database URL ». http://zinc.docking.org/
[74] «GDB database URL ». https://gdb.unibe.ch/downloads/
[75] «SAVI database URL ». https://cactus.nci.nih.gov/download/savi_download/
[76] Tarasova O.A., Urusova A.F., Filimonov D.A., Nicklaus M.C., Zakharov A.V., Poroikov V.V. QSAR Modeling Using Large-Scale Databases: Case Study for HIV-1 Reverse Transcriptase Inhibitors // *J Chem Inf Model.* 2015. V. 55. Iss 7. P. 1388–1399. doi: 10.1021/acs.jcim.5b00019.
[77] An A., Wang Y. Comparisons of classification methods for screening potential compounds // *IEEE Comput Soc*. 2001. P. 11–18. doi: 10.1109/ICDM.2001.989495.
[78] Chen J.J., Tsai C.-A., Moon H., Ahn H., Young J.J., Chen C.-H. Decision threshold adjustment in class prediction // *SAR QSAR Environ Res*. 2006. V. 17. Iss. 3. P. 337–352. doi: 10.1080/10659360600787700.
[79] Zakharov A.V., Peach M.L., Sitzmann M., Nicklaus, M. C. QSAR Modeling of Imbalanced High-Throughput Screening Data in PubChem // *J Chem Inf Model.* 2014. V. 54. Iss. 3. P. 705–712. doi: 10.1021/ci400737s.
[80] Fourches, D. Muratov E., Tropsha A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research // *J Chem Inf Model.* 2010. V. 50. Iss. 7. P. 1189–1204. doi: 10.1021/ci100176x.
[81] Fourches D., Muratov E., Tropsha A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation // *J Chem Inf Model.* 2016. V. 56. Iss. 7. P. 1243–1252. doi: 10.1021/acs.jcim.6b00129.
[82] Mauri A., Consonni V., Todeschini R. Molecular Descriptors // In book: *Handbook of Computational Chemistry*. 2017. P. 2065–2093. doi: 10.1007/978-3-319-27282-5_51.
[83] Filimonov D., Poroikov V., Borodina Y., Gloriozova T. Chemical Similarity Assessment through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors // *J Chem Inf Comput Sci.* 1999. V. 39. Iss. 4. P. 666–670. doi: 10.1021/ci980335o.
[84] Filimonov D.A., Druzhilovsky D.S., Lagunin A.A., Gloriziva T.A., Rudik A.V., Dmitriev A.V., Pogodin P.V., Poroikov V.V. Computer-aided prediction of biological activity spectra for chemical compounds: opportunities and limitation // *BMCRM*. 2018. V. 1. Iss. 1. P. e00004. doi: 10.18097/BMCRM00004.

[85] Filimonov D.A., Zakharov, A.V., Lagunin A.A., Poroikov V.V. QNA-based ‘Star Track’ QSAR approach // *SAR QSAR Environ Res* 2009. V. 20. Iss. 7–8. P. 679–709. doi: 10.1080/10629360903438370.

[86] Баскин И.И., Маджидов Т.И., Варнек А.А. // *Введение в хемоинформатику*. Казань. 2015. т. 4.

[87] Wold S., Dunn W. J. Multivariate quantitative structure-activity relationships (QSAR): conditions for their applicability // *J Chem Inf Comput Sci.* 1983. V. 23. Iss. 1. P. 6–13. doi: 10.1021/ci00037a002.

[88] Breiman L. Random Forest // *Machine Learning*. 2001. V. 45. Iss. 1. P. 5–32. doi: 10.1023/A:1010933404324.

[89] Breiman L., Friedman J. H., Olshen R. A., Stone C. J. // In book: *Classification And Regression Trees*. 2017. doi: 10.1201/9781315139470.

[90] Svetnik V., Liaw A., Tong C., Culberson J.C., Sheridan R.P., Feuston B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling // *J Chem Inf Comput. Sci.* 2003. V. 43. Iss. 6. P. 1947–1958. doi: 10.1021/ci034160g.

[91] Bender A., Cortes-Ciriano I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data // *Drug Discov Today*. 2021. V. 26. Iss. 4. P. 1040–1052. doi: 10.1016/j.drudis.2020.11.037.

[92] Filimonov D.A., Akimov D.V., Poroikov V.V. The Method of Self-Consistent Regression for the Quantitative Analysis of Relationships Between Structure and Properties of Chemicals // *J Pharm Chem* . 2004. V. 38. Iss. 1. P. 21–24. doi: 10.1023/B:PHAC.0000027639.17115.5d.

[93] Lagunin A.A., Zakharov A.V., Filimonov D.A., Poroikov V.V. A new approach to QSAR modelling of acute toxicity // *SAR QSAR Environ Res*. 2007. V. 18. Iss. 3–4. P. 285–298. doi: 10.1080/10629360701304253.

[94] Alharthi A.M., Lee M.H., Algamal Z.Y., Al-Fakih A. M. Quantitative structure-activity relationship model for classifying the diverse series of antifungal agents using ratio weighted penalized logistic regression // *SAR QSAR Environ Res*. 2020. V. 31. Iss. 8. P. 571–583. doi: 10.1080/1062936X.2020.1782467.

[95] Chen J.J., Tsai C.A., Young J.F., Kodell R. L. Classification ensembles for unbalanced class sizes in predictive toxicology // *SAR QSAR Environ Res*. 2005. V. 16. Iss. 6. P. 517–529. doi: 10.1080/10659360500468468.

[96] Cortes C., Vapnik V. Support-vector networks // *Mach Learn*. 1995. V. 20. Iss. 3. P. 273–297. doi: 10.1007/BF00994018.

[97] Shahlaei M. Descriptor Selection Methods in Quantitative Structure–Activity Relationship Studies: A Review Study // *Chem Rev.* 2013. V. 113. Iss. 10. P. 8093–8103. doi: 10.1021/cr3004339.

[98] Czarnecki W. M., Rataj K. Compounds Activity Prediction in Large Imbalanced Datasets with Substructural Relations Fingerprint and EEM // *IEEE Trustcom/BigDataSE/ISPA*. 2015. P. 192–192. doi: 10.1109/Trustcom.2015.581.

[99] Li S., Fedorowicz A., Andrew M. E. A new descriptor selection scheme for SVM in unbalanced class problem: a case study using skin sensitisation dataset // *SAR*

QSAR Environ Res. 2007. V. 18. Iss. 5–6. P. 423–441. doi: 10.1080/10629360701428474.

[100] Global HIV & AIDS statistics — Fact sheet | UNAIDS // https://www.unaids.org/en/resources/fact-sheet.

[101] Govender R.D., Hashim M.J., Khan M.A., Mustafa H., Khan G. Global Epidemiology of HIV/AIDS: A Resurgence in North America and Europe //*JEGH*. 2021. V. 11. Iss. 3. P. 296. doi: 10.2991/jegh.k.210621.001.

[102] Покровский В.В., Ладная Н.Н., Покровская А. В. ВИЧ/СПИД сокращает число россиян и продолжительность их жизни // *ДО*. 2017. т. 4, вып. 1, с. 65–82. doi: 10.17323/demreview.v4i1.6988.

[103] Becken B., Multani A., Padival S., Cunningham C. K. Human Immunodeficiency Virus I: History, Epidemiology, Transmission, and Pathogenesis // In book: *Introduction to Clinical Infectious Diseases*. 2019. P. 417–423. doi: 10.1007/978-3-319-91080-2_40.

[104] Visseaux B., Damond F., Matheron S., Descamps D., Charpentier C. Hiv-2 molecular epidemiology //*Infect Genet Evol* 2016. V. 46. P 233–240. doi: 10.1016/j.meegid.2016.08.010.

[105] А. В. Пиневич, А. К. Сироткин, О. В. Гаврилова, и А. А. Потехин, *Вирусология*, 2-е изд. Санкт-Петербург: Издательство Санкт-Петербургского университета, 2020.

[106] Kirchhoff F. HIV Life Cycle: Overview // In book; *Encyclopedia of AIDS*. 2013. P. 1–9. doi: 10.1007/978-1-4614-9610-6_60-1.

[107] Eggleton J.S., Nagalli S. Highly Active Antiretroviral Therapy (HAART) // In Book *StatPearls*. 2023. http://www.ncbi.nlm.nih.gov/books/NBK554533/

[108] Menéndez-Arias L., Delgado R. Update and latest advances in antiretroviral therapy // *Trends Pharmacol Sci*. 2022. V. 43. Iss. 1. P. 16–29. doi: 10.1016/j.tips.2021.10.004.

[109] Esposito F., Corona A., Tramontano E. HIV-1 Reverse Transcriptase Still Remains a New Drug Target: Structure, Function, Classical Inhibitors, and New Inhibitors with Innovative Mechanisms of Actions // *Mol Biol Int*. 2012. V. 2012. P. 1–23. doi: 10.1155/2012/586401.

[110] Sarafianos S.G., Marchand B., Das K., Himmel D.M., Parniak M., Hughes S.H., Arnold E. Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition // *J Mol Biol*. 2009. V. 385. Iss. 3. P. 693–713. doi: 10.1016/j.jmb.2008.10.071.

[111] Boyer P.L., Smith S.J., Zhao X.Z., Das K., Gruber K., Arnold E., Burke T.R., Hughes S.H. Developing and Evaluating Inhibitors against the RNase H Active Site of HIV-1 Reverse Transcriptase // 2018. *J Virol*. V. 92. Iss. 13. P. e02203-17. doi: 10.1128/JVI.02203-17.

[112] Adamson C.S., Freed E. O. Anti-HIV-1 Therapeutics: From FDA-approved Drugs to Hypothetical Future Targets // *Mol Interv*. 2009. V. 9. Iss. 2. P. 70–74. doi: 10.1124/mi.9.2.5.

[113] Clercq E. The Nucleoside Reverse Transcriptase Inhibitors, Nonnucleoside Reverse Transcriptase Inhibitors, and Protease Inhibitors in the Treatment of HIV Infections

(AIDS) // *Adv Pharmacol*. 2013. P. 317–358. doi: 10.1016/B978-0-12-405880-4.00009-3.
[114] Weber I.T., Wang Y.-F., Harrison R.W. HIV Protease: Historical Perspective and Current Research // *Viruses*. 2021. V. 13. Iss. 5. P. 839. doi: 10.3390/v13050839.
[115] Gulnik S., Erickson J.W., Xie D. HIV protease: Enzyme function and drug resistance // *Vitam Horm*. 2000. P. 213–256. doi: 10.1016/S0083-6729(00)58026-1.
[116] Wang Y., Lv Z., Chu Y. HIV protease inhibitors: a review of molecular selectivity and toxicity // *HIV*. 2015. P. 95. doi: 10.2147/HIV.S79956.
[117] Esposito D., Craigie R. HIV Integrase Structure and Function // *Adv Virus Res*. 1999. P. 319–333. doi: 10.1016/S0065-3527(08)60304-8.
[118] Maertens G.N., Engelman A.N., Cherepanov P. Structure and function of retroviral integrase // *Nat Rev Microbiol*. 2022. V. 20. Iss. 1. P. 20–34. doi: 10.1038/s41579-021-00586-9.
[119] Craigie R. The molecular biology of HIV integrase // *Future Virol*. 2012. V. 7. Iss. 7. P. 679–686. doi: 10.2217/fvl.12.56.
[120] Klasse P.J. The molecular basis of HIV entry // *Cellular Microbiology*. 2012. V. 14. Iss. 8. P. 1183–1192. doi: 10.1111/j.1462-5822.2012.01812.x.
[121] Dando T.M., Perry C.M. Enfuvirtide // *Drugs*. 2003. V. 63. Iss. 24. P. 2755–2766. doi: 10.2165/00003495-200363240-00005.
[122] Grande F., Occhiuzzi M.A., Rizzuti B., Ioele G., Luca M., Tucci P., Svicher V., Aquaro S., Garofalo A. CCR5/CXCR4 Dual Antagonism for the Improvement of HIV Infection Therapy // *Molecules*. 2019. V. 24. Iss. 3. P. 550. doi: 10.3390/molecules24030550.
[123] Ryst E. V. Maraviroc - a CCR5 Antagonist for the Treatment of HIV-1 Infection // *Front Immunol*. 2015. V. 6. doi: 10.3389/fimmu.2015.00277.
[124] Bettiker R.L., Koren D.E., Jacobson J. M. Ibalizumab // *Curr Opin HIV AIDS*. 2018. V. 13. Iss. 4. P. 354–358. doi: 10.1097/COH.0000000000000473.
[125] Chahine E.B. Fostemsavir: The first oral attachment inhibitor for treatment of HIV-1 infection // *Am J Health Syst Pharm*. 2021. V. 78. Iss. 5. P. 376–388. doi: 10.1093/ajhp/zxaa416.
[126] Darnag R., Minaoui B., Fakir M. QSAR models for prediction study of HIV protease inhibitors using support vector machines, neural networks and multiple linear regression // *Arab J Chem*. 2017. V. 10. P. S600–S608. doi: 10.1016/j.arabjc.2012.10.021.
[127] Baassi M., Moussaoui M., Soufi H., Rajkhowa S., Sharma A., Sinha S., Belaaouad S. Towards designing of a potential new HIV-1 protease inhibitor using QSAR study in combination with Molecular docking and Molecular dynamics simulations // *PLoS ONE*. 2023. V. 18. Iss. 4. P. e0284539. doi: 10.1371/journal.pone.0284539.
[128] Gorbalenya A.E., Baker S.C., Baric R.S., Groot R.J., Drosten C., Gulyaeva A.A., Haagmans B.L., Lauber C., Leontovich A.M., Neuman B.W., Penzar D., Perlman S., Poon L.L., Samborskiy D.V., Sidorov I.A., Sola I., Ziebuhr J. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and

naming it SARS-CoV-2 // *Nat Microbiol*. 2020. V. 5. Iss. 4. P. 536–544. doi: 10.1038/s41564-020-0695-z.

[129] Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) // *Johns Hopkins Coronavirus Resource Center*. https://coronavirus.jhu.edu/map.html

[130] Оперативные данные. https://xn--80aesfpebagmfblc0a.xn--p1ai/information.

[131] Karako K., Song P., Chen Y., Tang W., Kokudo N. Overview of the characteristics of and responses to the three waves of COVID-19 in Japan during 2020-2021 // *BST*. 2021. V. 15. Iss. 1. P. 1–8. doi: 10.5582/bst.2021.01019.

[132] Zeiser F.A., Donida B., Costa C.A., Ramos G.O., Scherer J.N., Barcellos N.T., Alegretti A.P., Ikeda M.L.R., Muller A.P.W., Bohn H.C., Santos I., Boni L., Antunes R.S., Righi R.R., Rigo S.J. First and second COVID-19 waves in Brazil: A cross-sectional study of patients' characteristics related to hospitalization and in-hospital mortality // *Lancet Reg Health Am*.2022. V. 6. P. 100107. doi: 10.1016/j.lana.2021.100107.

[133] Tao K., Tzou P.L., Nouhin J., Gupta R.K., Oliveira T., Pond S.L.K., Fera D., Shafer R.W. The biological and clinical significance of emerging SARS-CoV-2 variants // *Nat Rev Genet*. 2021. V. 22. Iss. 12. P. 757–773. doi: 10.1038/s41576-021-00408-x.

[134] Karim S.S.A., Karim Q.A. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic // *Lancet*. 2021. V. 398. Iss. 10317. P. 2126–2128. doi: 10.1016/S0140-6736(21)02758-6.

[135] Mendiola-Pastrana I.R., López-Ortiz E., Río de la Loza-Zamora J.G., González J., Gómez-García A., López-Ortiz G. SARS-CoV-2 Variants and Clinical Outcomes: A Systematic Review // *Life*. 2022. V. 12. Iss. 2. P. 170. doi: 10.3390/life12020170.

[136] Neuman B.W., Buchmeier M.J. Supramolecular Architecture of the Coronavirus Particle // *Adv Virus Res*. 2016. P. 1–27. doi: 10.1016/bs.aivir.2016.08.005.

[137] Zhao X., Ding Y., Du J., Fan Y. 2020 update on human coronaviruses: One health, one world // *Med Nov Technol Devices*. 2020. V. 8. P. 100043. doi: 10.1016/j.medntd.2020.100043.

[138] McBride R., Zyl M., Fielding B. The Coronavirus Nucleocapsid Is a Multifunctional Protein // *Viruses*. 2014. V. 6. Iss. 8. P. 2991–3018. doi: 10.3390/v6082991.

[139] V'kovski P., Kratzel A., Steiner S., Stalder H., Thiel V. Coronavirus biology and replication: implications for SARS-CoV-2 // *Nat Rev Microbiol*. 2021. V. 19. Iss. 3. P. 155–170. doi: 10.1038/s41579-020-00468-6.

[140] Chitsike L., Duerksen-Hughes P. Keep out! SARS-CoV-2 entry inhibitors: their role and utility as COVID-19 therapeutics // *Virol J*. 2021. V. 18. Iss. 1. P. 154. doi: 10.1186/s12985-021-01624-x.

[141] Cannalire R., Stefanelli I., Cerchia C., Beccari A.R., Pelliccia S., Summa V. SARS-CoV-2 Entry Inhibitors: Small Molecules and Peptides Targeting Virus or Host Cells // *IJMS*. 2020. V. 21. Iss. 16. P. 5707. doi: 10.3390/ijms21165707.

[142] Tao K., Tzou P. L., Nouhin J., Bonilla H., Jagannathan P., Shafer R.W. SARS-CoV-2 Antiviral Therapy // *Clin Microbiol Rev*. 2021. V. 34. Iss. 4. P. e00109-21. doi: 10.1128/CMR.00109-21.

[143] Jackson C.B., Farzan M., Chen B., Choe H. Mechanisms of SARS-CoV-2 entry into cells // *Nat Rev Mol Cell Biol*. 2022. V. 23. Iss. 1. P. 3–20. doi: 10.1038/s41580-021-00418-x.

[144] Menéndez J.C. Approaches to the Potential Therapy of COVID-19: A General Overview from the Medicinal Chemistry Perspective // *Molecules*. 2022. V. 27. Iss. 3. P. 658. doi: 10.3390/molecules27030658.

[145] Ullrich S., Nitsche C. The SARS-CoV-2 main protease as drug target // *Bioorg & Med Chem Lett*. 2020. V. 30. Iss. 17. P. 127377. doi: 10.1016/j.bmcl.2020.127377.

[146] Denesyuk A.I., Permyakov E.A., Johnson M.S., Permyakov S.E., Denessiouk K., Uversky V. N. Structural and functional significance of the amino acid differences Val35Thr, Ser46Ala, Asn65Ser, and Ala94Ser in 3C-like proteinases from SARS-CoV-2 and SARS-CoV // *Int J Biol Macromol*. 2021. V. 193. P. 2113–2120. doi: 10.1016/j.ijbiomac.2021.11.043.

[147] Abe K., Kabe Y., Uchiyama S., Iwasaki Y.W., Ishizu H., Uwamilo Y., Takenouchi T., Uno S., Ishii M., Maruno T., Noda M., Murata M., Hasegawa Y., Fukunaga K., Amagai M., Siomi H., Suematsu M., Kosaki K., Project K.D. Pro108Ser mutation of SARS-CoV-2 3CLpro reduces the enzyme activity and ameliorates the clinical severity of COVID-19 // *Sci Rep*. 2022. V. 12. Iss. 1. P. 1299. doi: 10.1038/s41598-022-05424-3.

[148] Hu Q., Xiong Y., Zhu G.-H., Zhang Y.-N., Zhang Y.-W., Huang P., Ge G.-B. The SARS-CoV-2 main protease (M^{pro}): Structure, function, and emerging therapies for COVID-19 // *Med Comm*. 2022. V. 3. Iss. 3.P. 151. doi: 10.1002/mco2.151.

[149] Osipiuk J., Azizi S.-A., Dvorkin S., Endres M., Jedrzejczak R., Jones K.A., Kang S., Kathayat R.S., Kim Y., Lisnyak V.G., Maki S.L., Nicolaescu V., Taylor C.A., Tesar C., Zhang Y.-A., Zhou Z., Randall G., Michalska K., Snyder S.A., Dickinson B.C., Joachimiak A. Structure of papain-like protease from SARS-CoV-2 and its complexes with non-covalent inhibitors // *Nat Commun*. 2021. V. 12. Iss. 1. c. 743. doi: 10.1038/s41467-021-21060-3.

[150] Báez-Santos Y.M., John St.S.E., Mesecar, A.D. The SARS-coronavirus papain-like protease: Structure, function and inhibition by designed antiviral compounds // *Antiviral Res*. 2015. V. 115. P. 21–38. doi: 10.1016/j.antiviral.2014.12.015.

[151] Aftab S.O., Ghouri M.Z., Masood M.U., Haider Z., Khan Z., Ahmad A., Manawar N. Analysis of SARS-CoV-2 RNA-dependent RNA polymerase as a potential therapeutic drug target using a computational approach // *J Transl Med*. 2020. V. 18. Iss. 1. P. 275. doi: 10.1186/s12967-020-02439-0.

[152] Kokic G., Hillen H.S., Tegunov D., Dienemann C., Seitz F., Schmitzova J., Farnung L., Siewert A., Hobather C., Cramer P. Mechanism of SARS-CoV-2 polymerase stalling by remdesivir // *Nat Commun*. 2021. V. 12. Iss. 1. P. 279. doi: 10.1038/s41467-020-20542-0.

[153] Tinkov O.V., Grigorev V.Yu., Grigoreva L. D. Virtual Screening and Molecular Design of Potential SARS-COV-2 Inhibitors // *Moscow Univ Chem Bull*. 2021. V. 76. Iss. 2. P. 95–113, мар. 2021, doi: 10.3103/S0027131421020127.

[154] Jawarkar R.D., Bakai R., Zaki M.E.A., Al-Hussain S., Grosh A., Gandhni A., Mukerjee N., Samad A., Masand V.H., Lewaa I. QSAR based virtual screening

derived identification of a novel hit as a SARS CoV-229E 3CLpro Inhibitor: GA-MLR QSAR modeling supported by molecular Docking, molecular dynamics simulation and MMGBSA calculation approaches // *Arab J Chem*. 2022. V. 15. Iss. 1. P. 103499. doi: 10.1016/j.arabjc.2021.103499.

[155] Edache E.I., Uzairu A., Mamza P.A., Shallangwa G.A. QSAR, homology modeling, and docking simulation on SARS-CoV-2 and pseudomonas aeruginosa inhibitors, ADMET, and molecular dynamic simulations to find a possible oral lead candidate // *J Genet Eng Biotechnol*. 2022. V. 20. Iss. 1. P. 88. doi: 10.1186/s43141-022-00362-z.

[156] Matsumoto M., Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator // *ACM Trans Model Comput Simul.* 1998. V. 8. Iss. 1. P. 3–30. doi: 10.1145/272991.272995.

[157] Box G.E.P., Muller M.E. A Note on the Generation of Random Normal Deviates // *Ann Math Statist.* 1958. V. 29. Iss. 2. P. 610–611. doi: 10.1214/aoms/1177706645.

[158] «Tox21 public data URL». https://tripod.nih.gov//tox21/pubdata/

[159] «NIAID HIV/OI/TB database URL». https://chemdb.niaid.nih.gov/

[160] «Clarivate Analytics Integrity database URL». https://integrity.clarivate.com/

[161] «PostEra Moonshot data URL». https://covid.postera.ai/covid/activity_data

[162] «NCATS COVID-19 screening collection URL». https://www.ebi.ac.uk/chembl/

[163] «Stanford University Coronavirus Antiviral & Resistance Database URL». https://covdb.stanford.edu/search/?virus=SARS-CoV-2

[164] Zakharov A.V., Peach M.L., Sitzmann M., Nicklaus M. C. A New Approach to Radial Basis Function Approximation and Its Application to QSAR // *J Chem Inf Model.* 2014. V. 54. Iss. 3. P. 713–719. мар. doi: 10.1021/ci400704f.

[165] Pinheiro E.C., Ferrari S.L.P. A comparative review of generalizations of the Gumbel extreme value distribution with an application to wind speed data // *J Statist Comput Simul*. 2016. V. 86. Iss. 11. P. 2241–2261. doi: 10.1080/00949655.2015.1107909.

[166] Seber G.A.F. // In book: *Linear Regression Analysis*. 1977.

[167] «RDKit». https://www.rdkit.org/

[168] Баскин И.И., Маджидов Т.И., Варнек А.А. Введение в Хемоинформатику // Учебное пособие. Казань. 2015. т. 6.

[169] Pogodin P.V., Lagunin A.A., Rudik A.V., Filimonov D.A., Druzhilovskiy D.S., Nicklaus M.C., Poroikov V.V. How to Achieve Better Results Using PASS-Based Virtual Screening: Case Study for Kinase Inhibitors // *Front Chem.* 2018. V. 6. P. 133. doi: 10.3389/fchem.2018.00133.

Приложение А

Сродство к электрону *EA* и потенциал ионизации *IP* элементов

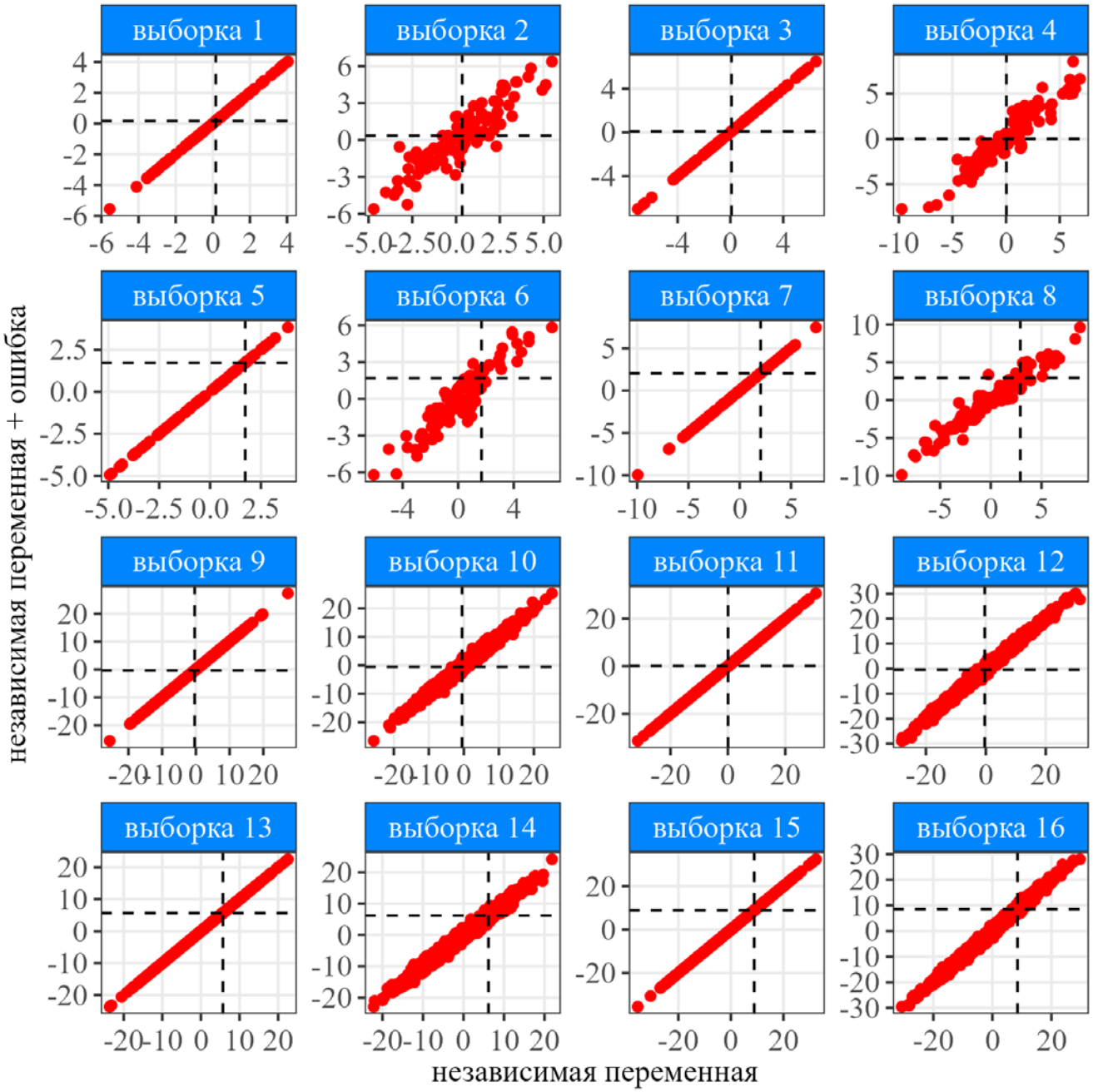
Atom	*EA*	*IP*	*Atom*	*EA*	*IP*	*Atom*	*EA*	*IP*
H	*0.75*	*13.60*	*Kr*	*-0.42*	*14.00*	*Lu*	*0.20*	*6.15*
He	*0.08*	*24.59*	*Rb*	*0.49*	*4.18*	*Hf*	*0.33*	*7.50*
Li	*0.62*	*5.39*	*Sr*	*-0.15*	*5.69*	*Ta*	*0.40*	*7.89*
Be	*-0.20*	*9.32*	*Y*	*0.31*	*6.22*	*W*	*0.67*	*7.98*
B	*0.28*	*8.30*	*Zr*	*0.33*	*6.84*	*Re*	*0.23*	*7.88*
C	*1.26*	*11.26*	*Nb*	*0.51*	*6.88*	*Os*	*1.44*	*8.73*
N	*0.44*	*14.53*	*Mo*	*0.68*	*7.09*	*Ir*	*1.57*	*9.10*
O	*1.46*	*13.62*	*Tc*	*0.54*	*7.23*	*Pt*	*1.10*	*8.96*
F	*3.45*	*17.42*	*Ru*	*1.10*	*7.37*	*Au*	*1.25*	*9.23*
Ne	*0.00*	*21.57*	*Rh*	*1.14*	*7.46*	*Hg*	*-0.19*	*10.44*
Na	*0.55*	*5.14*	*Pd*	*1.11*	*8.34*	*Tl*	*0.31*	*6.11*
Mg	*-0.31*	*7.64*	*Ag*	*1.22*	*7.58*	*Pb*	*1.39*	*7.42*
Al	*0.30*	*5.99*	*Cd*	*-0.43*	*8.99*	*Bi*	*0.97*	*7.29*
Si	*1.39*	*8.15*	*In*	*0.31*	*5.79*	*Po*	*1.97*	*8.42*
P	*0.75*	*10.49*	*Sn*	*1.39*	*7.34*	*At*	*2.90*	*9.20*
S	*2.00*	*10.36*	*Sb*	*0.90*	*8.64*	*Rn*	*-0.15*	*10.75*
Cl	*3.61*	*12.97*	*Te*	*1.97*	*9.01*	*Fr*	*0.48*	*3.98*
Ar	*-0.37*	*15.76*	*I*	*3.23*	*10.45*	*Ra*	*-0.15*	*5.28*
K	*0.50*	*4.34*	*Xe*	*-0.25*	*12.13*	*Ac*	*0.80*	*5.20*
Ca	*-0.19*	*6.11*	*Cs*	*0.47*	*3.89*	*Th*	*0.80*	*6.10*
Sc	*0.19*	*6.56*	*Ba*	*-0.15*	*5.21*	*Pa*	*0.84*	*6.00*
Ti	*0.33*	*6.82*	*La*	*0.30*	*5.59*	*U*	*0.82*	*6.19*
V	*0.53*	*6.74*	*Ce*	*0.25*	*5.54*	*Np*	*0.82*	*6.20*
Cr	*0.67*	*6.77*	*Pr*	*0.20*	*5.47*	*Pu*	*0.84*	*6.06*
Mn	*-0.17*	*7.43*	*Nd*	*0.20*	*5.53*	*Am*	*0.85*	*6.00*
Fe	*0.50*	*7.90*	*Pm*	*0.20*	*5.58*	*Cm*	*0.85*	*6.09*
Co	*0.66*	*7.86*	*Sm*	*0.20*	*5.64*	*Bk*	*0.82*	*6.23*
Ni	*1.16*	*7.64*	*Eu*	*0.20*	*5.67*	*Cf*	*0.84*	*6.27*
Cu	*1.23*	*7.72*	*Gd*	*0.20*	*6.15*	*Es*	*0.86*	*6.47*
Zn	*-0.44*	*9.39*	*Tb*	*0.20*	*5.86*	*Fm*	*0.86*	*6.60*
Ga	*0.30*	*6.00*	*Dy*	*0.20*	*5.94*	*Md*	*0.83*	*6.68*
Ge	*1.39*	*7.90*	*Ho*	*0.20*	*6.02*	*No*	*0.79*	*6.58*
As	*0.80*	*9.79*	*Er*	*0.20*	*6.11*	*Lr*	*0.85*	*6.69*
Se	*2.02*	*9.75*	*Tm*	*0.20*	*6.18*	*Db*	*0.46*	*6.43*
Br	*3.45*	*11.81*	*Yb*	*0.20*	*6.25*	*Jl*	*0.50*	*6.78*

Приложение Б

Примеры сгенерированных выборок

Выборка	N	N^+	n	m_a	σ
Выборка 1	*100*	*50*	*25*	*5*	*0.01*
Выборка 2	*100*	*50*	*25*	*5*	*1*
Выборка 3	*100*	*50*	*25*	*10*	*0.01*
Выборка 4	*100*	*50*	*25*	*10*	*1*
Выборка 5	*100*	*20*	*25*	*5*	*0.01*
Выборка 6	*100*	*20*	*25*	*5*	*1*
Выборка 7	*100*	*20*	*25*	*10*	*0.01*
Выборка 8	*100*	*20*	*25*	*10*	*1*
Выборка 9	*1000*	*500*	*250*	*50*	*0.01*
Выборка 10	*1000*	*500*	*250*	*50*	*1*
Выборка 11	*1000*	*500*	*250*	*100*	*0.01*
Выборка 12	*1000*	*500*	*250*	*100*	*1*
Выборка 13	*1000*	*200*	*250*	*50*	*0.01*
Выборка 14	*1000*	*200*	*250*	*50*	*1*
Выборка 15	*1000*	*200*	*250*	*100*	*0.01*
Выборка 16	*1000*	*200*	*250*	*100*	*1*

N – число примеров; N^+– число положительных примеров; n – число независимых переменных; m_a – число используемых параметров; σ – стандартное отклонение сгенерированной ошибки.



Независимые переменные без и с добавлением ошибки; штрихованная линия – порог разделения «активных» и «неактивных»

Приложение В

Выборки и модели Tox21

Выборка\Модель	*Активные*	*Неактивные*	*V(SCR)*	*V(SCEC)*	*BA(SCR)*	*BA(SCEC)*
ahr-p1	*579*	*4413*	*289*	*97*	*0.794*	*0.807*
ap1-agonist-p1	*241*	*3523*	*194*	*80*	*0.623*	*0.627*
ar-bla-agonist-p1	*178*	*4661*	*350*	*75*	*0.854*	*0.856*
ar-bla-antagonist-p1	*343*	*4065*	*198*	*87*	*0.676*	*0.715*
ar-mda-kb2-luc-agonist-p1	*193*	*4394*	*304*	*84*	*0.734*	*0.762*
ar-mda-kb2-luc-antagonist-p1	*439*	*4144*	*263*	*90*	*0.701*	*0.701*
ar-mda-kb2-luc-antagonist-p2	*590*	*2939*	*255*	*87*	*0.761*	*0.763*
are-bla-p1	*429*	*2417*	*229*	*74*	*0.713*	*0.695*
aromatase-p1	*466*	*3940*	*226*	*101*	*0.703*	*0.728*
car-agonist-p1	*689*	*3184*	*232*	*75*	*0.731*	*0.759*
car-antagonist-p1	*661*	*2830*	*225*	*77*	*0.738*	*0.768*
casp3-hepg2-p1	*136*	*3580*	*147*	*73*	*0.717*	*0.669*
elg1-luc-agonist-p1	*212*	*4714*	*207*	*84*	*0.707*	*0.699*
er-bla-agonist-p2	*282*	*4539*	*249*	*88*	*0.741*	*0.722*
er-bla-antagonist-p1	*289*	*3948*	*195*	*83*	*0.683*	*0.698*
er-luc-bg1-4e2-agonist-p2	*725*	*3527*	*217*	*76*	*0.653*	*0.658*
er-luc-bg1-4e2-antagonist-p1	*446*	*4303*	*252*	*89*	*0.701*	*0.725*
er-luc-bg1-4e2-antagonist-p2	*298*	*3408*	*167*	*80*	*0.706*	*0.702*
erb-bla-antagonist-p1	*244*	*3060*	*165*	*82*	*0.713*	*0.712*
gh3-tre-antagonist-p1	*202*	*3790*	*164*	*71*	*0.649*	*0.635*
gr-hela-bla-agonist-p1	*128*	*4745*	*312*	*62*	*0.775*	*0.764*
gr-hela-bla-antagonist-p1	*356*	*3999*	*297*	*100*	*0.707*	*0.736*
h2ax-cho-p2	*199*	*3722*	*234*	*87*	*0.687*	*0.668*
hdac-p1	*193*	*3526*	*219*	*59*	*0.754*	*0.753*
hse-bla-p1	*170*	*2937*	*181*	*80*	*0.637*	*0.622*

mitotox-p1	*680*	*3423*	*251*	*95*	*0.743*	*0.75*
p53-bla-p1	*288*	*4404*	*243*	*99*	*0.695*	*0.703*
pparg-bla-agonist-p1	*186*	*4554*	*200*	*90*	*0.645*	*0.649*
pparg-bla-antagonist-p1	*185*	*2766*	*160*	*78*	*0.659*	*0.676*
pr-bla-antagonist-p1	*450*	*2878*	*217*	*90*	*0.737*	*0.754*
pxr-p1	*796*	*2536*	*261*	*87*	*0.715*	*0.734*
rar-agonist-p1	*350*	*2815*	*170*	*75*	*0.732*	*0.727*
rar-antagonist-p2	*636*	*2655*	*194*	*73*	*0.682*	*0.692*
ror-cho-antagonist-p1	*378*	*2691*	*200*	*67*	*0.712*	*0.697*
rxr-bla-agonist-p1	*325*	*3114*	*191*	*77*	*0.637*	*0.651*
sbe-bla-antagonist-p1	*197*	*3371*	*203*	*72*	*0.706*	*0.669*
shh-3t3-gli3-antagonist-p1	*383*	*2536*	*186*	*67*	*0.687*	*0.687*
tshr-agonist-p1	*165*	*3686*	*223*	*76*	*0.692*	*0.682*

Активные – количество активных соединений в обучающей выборке;

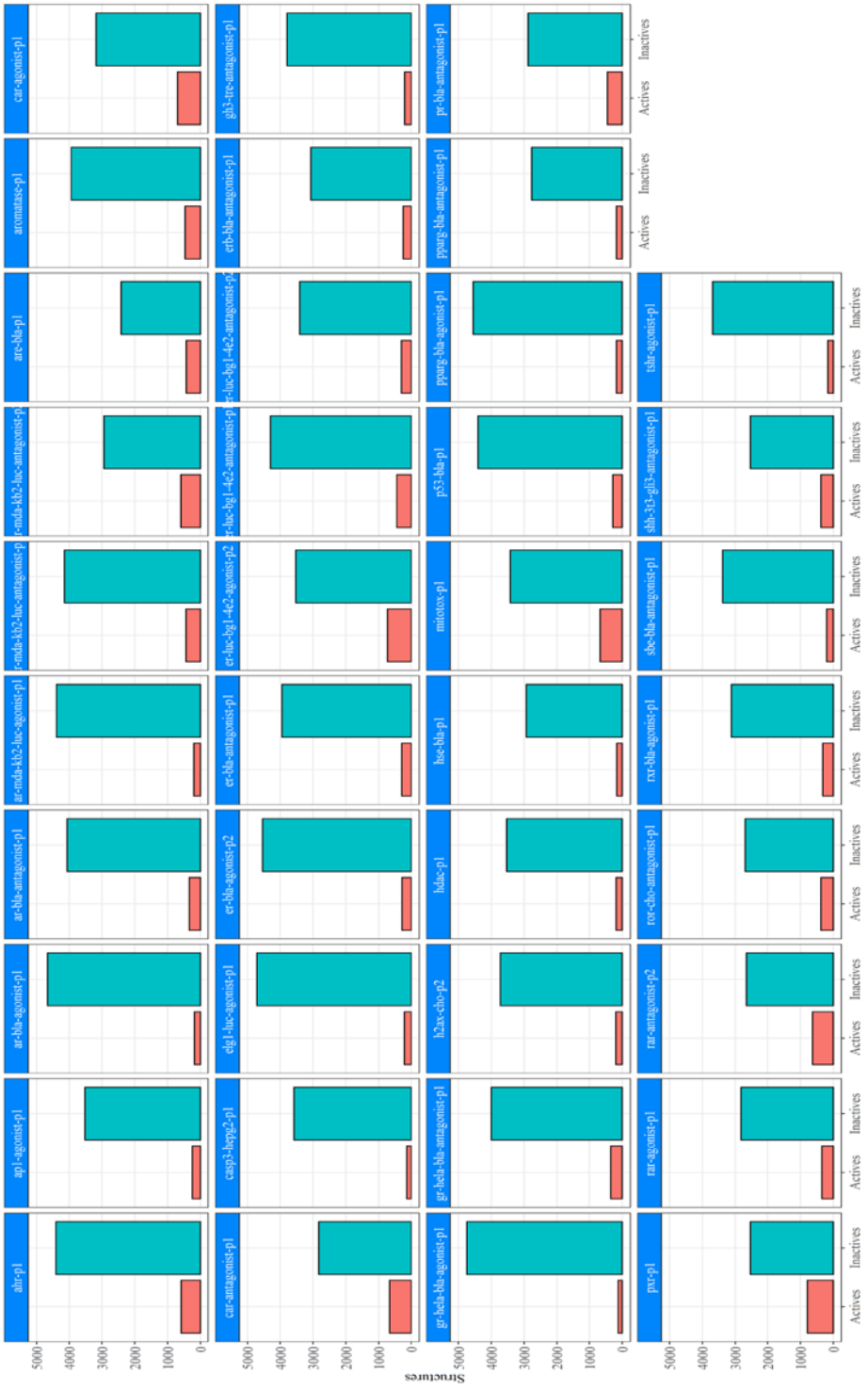
Неактивные – количество неактивных соединений в обучающей выборке;

V(SCR) – эффективная размерность при использовании SCR;

V(SCEC) – эффективная размерность при использовании SCEC;

BA(SCR) – сбалансированная точность при кросс-валидации для моделей SCR model;

BA(SCEC) – сбалансированная точность при кросс-валидации для моделей SCEC.


ahr-p1
ap1-agonist-p1
ar-bla-agonist-p1
ar-bla-antagonist-p1
ar-msda-kb2-luc-agonist-p1
ar-msda-kb2-luc-antagonist-p1
ar-msda-kb2-luc-antagonist-p2
arc-bla-p1
aromatase-p1
car-agonist-p1
car-antagonist-p1
casp3-hepg2-p1
cellq1-luc-agonist-p1
er-bla-agonist-p2
er-bla-antagonist-p1
er-luc-bg1-4e2-agonist-p2
er-luc-bg1-4e2-antagonist-p1
er-luc-bg1-4e2-antagonist-p2
erb-bla-antagonist-p1
gh3-nre-antagonist-p1
gr-hela-bla-agonist-p1
gr-hela-bla-antagonist-p1
h2ax-cho-p2
hdac-p1
hse-bla-p1
mitotox-p1
p53-bla-p1
pparg-bla-agonist-p1
pparg-bla-antagonist-p1
pr-bla-antagonist-p1
pxr-p1
rar-agonist-p1
rar-antagonist-p2
ror-cho-antagonist-p1
rxr-bla-agonist-p1
sbe-bla-antagonist-p1
shh-3t3-gli3-antagonist-p1
tshr-agonist-p1
Structures
0
1000
2000
3000
4000
5000
Actives
Inactives

Приложение Г

Реализация SCLC и SCEC в C++

```
void SetSCLEC(vector<vector<double>>* FInDF, bool SCLC, bool restrict)
{
  ordered FRows = FInDF->size();
  ordered FCols = FInDF->at(0).size();

  ordered FDim = FCols - 1;
  ordered FDimIt = FDim;
  ordered FDimItSurvival = 0;
  ordered FDimItSurvivalHistory = 0;
  double Epsilon = std::pow(10, -24);
  ordered counter = 0;
  bool afterburn = false;
  ordered afterBurnIteration = 5;


  vector<vector<double>>* FIn = new vector<vector<double>>(0);
  for (ordered i = 0; i < FRows; i++)
    FIn->push_back(vector<double>(FCols));

  for (ordered j = 0; j < FCols; j++) {
    for (ordered i = 0; i < FRows; i++)
      (*FIn)[i][j] = (*FInDF)[i][j];
  }

  for (ordered i = 0; i < FRows; i++)
    (*FIn)[i][0] = (*FIn)[i][0] == 1 ? 1 : -1;
  std::vector<double> a(FCols, 0.0);
  std::vector<double> ah(FCols, 0.0);

  std::vector<double> W(FRows + FCols, 1.0);
  std::vector<double> Preds(FRows, 0);

  vector<vector<double>>* FF = new vector<vector<double>>(0);
  for (ordered i = 0; i < FRows + FCols; i++)
    FF->push_back(vector<double>(FCols));

  vector<vector<double>>* FFnorm = new vector<vector<double>>(0);
  for (ordered i = 0; i < FRows + FCols; i++)
    FFnorm->push_back(vector<double>(FCols));

  double convergencyIndicatorFull = 10000;
  double convergencyIndicator = 10000;
  double convergencyLimit = std::pow(10, -2);
  double delta = 0;
  bool stop = false;


  std::vector<ordered> FN(FCols, 0.0);
  std::vector<double> XWX(FCols, 0);
  std::vector<double> XWY(FCols, 0);
  std::vector<double> FEDim(FCols, 0.0);
  std::vector<double> FxWv(FCols, 0.0);

  ordered reductionCounter = 0;
  ordered reductionCounterLimit = 20;

  while (!stop && counter < 20) {
    counter++;
```

```
for (ordered i = 0; i < FRows; i++)
  for (ordered j = 0; j < FCols; j++)
    (*FF)[i + 1][j] = (*FIn)[i][j];
for (ordered i = 1; i < FCols; i++)
  for (ordered j = 1; j < FCols; j++)
    (*FF)[FRows + i][j] = i == j ? 1 : 0;
for (ordered j = 0; j < FCols; j++)
  (*FF)[0][j] = 0;
for (ordered j = 1; j < FCols; j++)
  (*FF)[FRows + j][0] = 0;

if (SCLC) {
  for (ordered i = 1; i <= FRows; i++) {

    double xa = 0;
    for (ordered j = 1; j < FCols; j++)
      xa += (*FF)[i][j] * a[j];

    xa += a[0];
    double p = 1 / (1 + exp(-xa));

    W[i] = p * (1 - p);
    (*FF)[i][0] = (*FF)[i][0] - p + W[i] * xa;
    Preds[i - 1] = (*FF)[i][0];
  }
}
else {
  for (ordered i = 1; i <= FRows; i++) {
    double orderederrim = (*FF)[i][0];

    for (ordered j = 1; j < FCols; j++)
      (*FF)[i][0] += (*FF)[i][j] * a[j];

    (*FF)[i][0] += a[0];


    W[i] = std::exp(-orderederrim * ((*FF)[i][0] - orderederrim));

    Preds[i - 1] = (*FF)[i][0];
  }
}

W[0] = 0;
for (ordered j = 1; j < FCols; j++)
  W[FRows + j] = 0;

double totalWeight = 0;
for (ordered i = 1; i <= FRows; i++)
  totalWeight += W[i];

(*FF)[0][0] = 0;
for (ordered i = 1; i <= FRows; i++)
  (*FF)[0][0] += (*FF)[i][0] * W[i];

for (ordered j = 1; j < FCols; j++) {
  (*FF)[0][j] = 0;
  double add = 0;
  for (ordered i = 1; i <= FRows; i++) {
    add = (*FF)[i][j] * W[i];
    (*FF)[0][j] += add;
  }
}

for (ordered j = 0; j < FCols; j++)
```

```
      (*FF)[0][j] *= -1 / totalWeight;

    for (ordered i = 1; i <= FRows; i++)
      (*FF)[i][0] += (*FF)[0][0];

    for (ordered j = 1; j < FCols; j++)
      for (ordered i = 1; i <= FRows; i++)
        (*FF)[i][j] += (*FF)[0][j];

    for (ordered i = 0; i < (*FF).size(); i++)
      for (ordered j = 0; j < (*FF)[0].size(); j++)
        (*FFnorm)[i][j] = (*FF)[i][j];

    for (ordered i = 0; i < FCols; i++) {
      FN[i] = counter == 1 ? i : FN[i];
      XWX[i] = 0;
      XWY[i] = 0;
      FEDim[i] = FEDim[i] == -1 ? -1 : 1;
      FxWv[i] = 0;
    }

    ordered n = 0;
    while (n < FDim) {

      ordered k = FN[n + 1];

      if (FEDim[k] <= 0) {
        n++;
        W[FRows + k] = 0;
        continue;
      }

      for (ordered j = n + 1; j < FCols; j++) {

        double add = 0;
        double s = 0;
        ordered kk = FN[j];
        XWX[kk] = 0;
        XWY[kk] = 0;

        if (FEDim[kk] >= 0) {
          for (ordered i = 1; i <= FRows; i++) {
            add = (*FFnorm)[i][kk] * W[i];
            XWX[kk] += add * (*FF)[i][kk];
            XWY[kk] += add * (*FF)[i][0];
          }

          s = XWY[kk] * XWY[kk] / XWX[kk];
          FEDim[kk] = s < 1 ? 0 : 1 - 1 / s;
        }

        W[FRows + kk] = FEDim[kk] > Epsilon ? XWX[kk] * XWX[kk] / (XWY[kk] * XWY[kk] -
XWX[kk]) : 0;
      }

      if (!afterburn) {
        ordered i = FDim;
        while (i > n + 2) {
          k = n + 1;

          while (k < i) {
            k++;

            if (FEDim[FN[k - 1]] < FEDim[FN[k]]) {
```

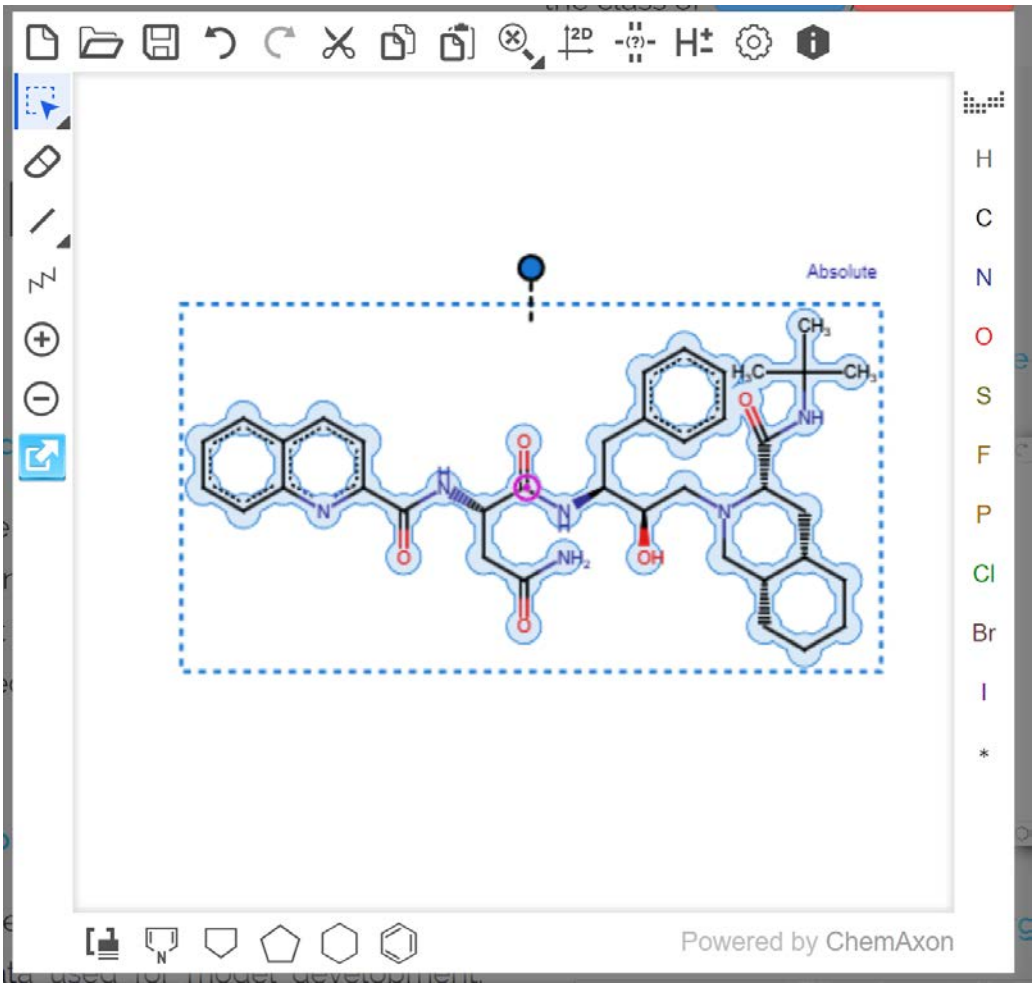
```
          ordered j = FN[k - 1];
          FN[k - 1] = FN[k];
          FN[k] = j;
        }
      }
      i--;
    }
  }

  k = FN[n + 1];
  double s = 0;
  for (int i = 1; i <= FRows; i++)
    s += (*FF)[i][k] * W[i] * (*FF)[i][0];
  for (ordered i = 0; i <= n; i++) {
    ordered l = FN[i + 1];
    s += (*FF)[l + FRows][k] * W[l + FRows] * (*FF)[l + FRows][0];
  }
  FxWv[0] = s;


  for (ordered j = n; j < FDim; j++) {
    ordered d = FN[j + 1];
    s = 0;
    for (int i = 0; i <= FRows; i++)
      s += (*FF)[i][k] * W[i] * (*FF)[i][d];
    for (ordered i = 0; i <= n; i++) {
      ordered l = FN[i + 1];
      s += (*FF)[l + FRows][k] * W[l + FRows] * (*FF)[l + FRows][d];
    }
    FxWv[d] = s;
  }

  for (ordered i = 0; i <= FRows; i++)
    (*FF)[i][0] -= (*FF)[i][k] * FxWv[0] / FxWv[k];
  for (ordered i = 0; i <= n; i++) {
    ordered l = FN[i + 1];
    (*FF)[l + FRows][0] -= (*FF)[l + FRows][k] * FxWv[0] / FxWv[k];
  }

  if (n + 1 < FDim) {
    for (ordered j = n + 1; j < FDim; j++) {
      ordered d = FN[j + 1];

      for (ordered i = 0; i <= FRows; i++)
        (*FF)[i][d] -= (*FF)[i][k] * FxWv[d] / FxWv[k];
      for (ordered i = 0; i <= n; i++) {
        ordered l = FN[i + 1];
        (*FF)[l + FRows][d] -= (*FF)[l + FRows][k] * FxWv[d] / FxWv[k];
      }
    }
  }

  n++;
}
if (afterburn)
  for (ordered i = 0; i < FEDim.size(); i++)
    FEDim[i] = FEDim[i] <= 0 ? -1 : FEDim[i];

for (ordered j = 1; j < FCols; j++)
  a[j] = -(*FF)[FRows + j][0];
a[0] = -(*FF)[0][0];

FDimItSurvivalHistory = FDimItSurvival;
FDimItSurvival = 0;
```

```
    for (ordered j = 1; j < FCols; j++)
      a[j] != 0 ? FDimItSurvival++ : FDimItSurvival;

    if (!afterburn && FDimItSurvival <= FDimItSurvivalHistory) {
      afterburn = true;
    }

    convergencyIndicator = 0;
    delta = 0;
    convergencyIndicatorFull = 0;
    for (ordered i = 0; i < a.size(); i++) {
      delta = std::abs(a[i] - ah[i]);
      convergencyIndicatorFull += delta;
      convergencyIndicator = delta > convergencyIndicator ? delta : convergencyIndicator;
      ah[i] = a[i];
    }

    if (convergencyIndicator <= convergencyLimit)
      stop = true;
  }

  delete FIn;
  delete FF;
  delete FFnorm;
}
```

Приложение Д

Использование веб-сервиса AntiHIV pred

Для ввода структуры необходимо воспользоваться встроенным редактором молекул. Редактор позволяет подстановку структуры, если в буфер обмена скопирована строка SMILES или mol файл.



Далее необходимо выбрать мишень и модель, специфицируя данные, на которых она была построена.