

На правах рукописи

Столбов Леонид Алексеевич

**РАЗРАБОТКА ПОДХОДОВ К ВИРТУАЛЬНОМУ СКРИНИНГУ АНТИВИРУСНЫХ
СОЕДИНЕНИЙ С УЧЕТОМ ГЕТЕРОГЕННОСТИ ИНФОРМАЦИИ**

1.5.8. – математическая биология, биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание учёной степени

кандидата биологических наук

Москва – 2023

Работа выполнена в Федеральном государственном бюджетном научном учреждении «Научно-исследовательский институт биомедицинской химии имени В. Н. Ореховича» (ИБМХ)

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

Научный консультант кандидат физико-математических наук
Филимонов Дмитрий Алексеевич

Официальные
оппоненты: **Васильев Павел Михайлович,**
доктор биологических наук, ФГБОУ ВО Волгоградский
государственный медицинский университет Минздрава России,
Научный центр инновационных лекарственных средств,
заведующий лабораторией

Порозов Юрий Борисович,
кандидат медицинских наук, ФГАОУ ВО Первый Московский
государственный медицинский университет им. И. М. Сеченова
Минздрава России, Центр био- и хемоинформатики,
руководитель

Ведущая организация: Федеральное государственное автономное научное учреждение
«Федеральный научный центр исследований и разработки
иммунобиологических препаратов им. М. П. Чумакова РАН»
(Институт полиомиелита)

Защита состоится «26» октября 2023 г. в 13 часов на заседании диссертационного совета
24.1.172.01 (Д 001.010.01) при Федеральном государственном бюджетном научном
учреждении «Научно-исследовательский институт биомедицинской химии имени
В. Н. Ореховича» (ИБМХ) по адресу: 119121, Москва, ул. Погодинская, д. 10, стр. 8.

С диссертацией можно ознакомиться в библиотеке и на сайте ИБМХ.

Автореферат разослан «__» _____ 2023 г.

Ученый секретарь Диссертационного совета
кандидат химических наук



Карпова Е.А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

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 (Geronikaki A. et al., 2016). Пандемия COVID-19, вызванная широким распространением в популяции вируса SARS-CoV-2 и его высокой контагиозностью, стимулировала разработку новых подходов к созданию противовирусных препаратов для оперативного ответа на новые биогенные угрозы (Adamson C.S. et al., 2021).

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 (Muratov E.N. et al., 2021). 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 (Sulimov V.B. et al., 2021; Gentile F. et al., 2022). 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 (Muratov E.N. et al., 2020).

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 (Medina-Franco J.L. et al., 2022). Накопление этой информации в свободно-доступных базах данных (PubChem, ChEMBL, и др.) обеспечило предпосылки для ее использования при создании обучающих выборок с целью построения количественных и классификационных моделей зависимостей «структура-активность» (Ishola A.A. et al., 2021). 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, Lagunin A.A. et al., 2011), 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 (Sakamuru et al., 2022), которые часто применяются для сравнения новых (Q)SAR моделей.

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 были решены 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.

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.

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 противовирусных 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.

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).

Публикации. По материалам диссертации опубликована 21 работа в российских и международных научных изданиях, включая 6 статей в рецензируемых научных журналах, входящих в перечень, рекомендованный ВАК, и 14 публикаций в трудах конференций. Получено свидетельство Федеральной службы по интеллектуальной собственности о регистрации программы для ЭВМ Sarmath №: RU2022662930 от 07.07.2022 г.

Объём и структура диссертации

Диссертационная работа состоит из введения, обзора литературы, описания материалов и методов исследования, представления результатов и их обсуждения, заключения, выводов, списка литературы, включающего 169 источников, и приложения. Работа изложена на 132 страницах, содержит 35 рисунков и 13 таблиц. Приложение содержит 3 таблицы.

МАТЕРИАЛЫ И МЕТОДЫ

Формирование обучающих и тестовых выборок

Для отладки разрабатываемых методов мы сгенерировали выборки с искусственными данными, для которых связь между зависимой и независимыми переменными заведомо известна. Их генерацию осуществляли следующим образом: регрессионные матрицы X были получены в различных случайных комбинациях регрессоров. Наблюдаемые величины y были рассчитаны по известным коэффициентам регрессии с добавлением случайных ошибок. Наблюдения были классифицированы с использованием различных порогов для получения выборок с разным уровнем сбалансированности.

Для валидации разрабатываемых методов и их сравнения с существующими подходами нами были извлечены и обработаны данные по токсичности соединений из БД Tox21. В результате подготовлены 38 выборок, каждая из которых несбалансирована по соотношению числа токсичных/нетоксичных соединений.

Экспериментальные данные об ингибировании молекулярных мишеней вируса SARS-CoV-2 были собраны из трех источников, пополняемых на протяжении пандемии COVID-19 и интегрированы в единую обучающую выборку. PostEra Moonshot, краудсорсинговый стартап по исследованиям потенциальных препаратов для терапии коронавирусной инфекции, предоставляющий экспериментальные данные об ингибиторах главной протеазы (3CL) SARS-CoV-2. COVID-19 Open Data Portal – предоставляет данные NCATS об исследованиях ингибиторов главной протеазы (3CL) и РНК-зависимой РНК-полимеразы (RdRp) SARS-CoV-2. Coronavirus Antiviral and Resistance Database – БД Стэнфордского университета с агрегацией литературных данных об ингибиторах 3CL, RdRp и папаин-подобной протеазы (PLpro) SARS-CoV-2.

Полученные из указанных источников данные, как для ВИЧ-1, так и для SARS-CoV-2, обладают высокой степенью гетерогенности, характеризуемой разнообразием химических классов и различиями в экспериментальных методах определения значений активности, а также разными механизмами действия ингибиторов. Для формирования обучающих выборок мы отбирали записи, содержащие значения полуэффективных концентраций ингибирования в молях на литр (IC_{50} (M)), значения IC_{50} переводили в $pIC_{50} = -\lg(IC_{50})$. Предобработку данных по структуре и активности соединений проводили перед этапом моделирования согласно современным требованиям (Fourches D. et al., 2016). Численные характеристики полученных выборок для антиретровирусных соединений приведены на рисунке 1.

Для построения и исследования классификационных моделей выборки ингибиторов ВИЧ дополнялись данными о константах связывания (K_i) и данными о неактивных

соединениях. Всего было подготовлено 6 выборок с объединением данных со значениями активности, полученными при разных экспериментальных условиях: IN (+Ki) - 4349 структур, IN (+Ki и неактивные) - 12000 структур, PR (+Ki) - 8545 структур, PR (+Ki и неактивные) - 12000 структур, RT (+Ki) - 7508 структур, RT (+Ki и неактивные) - 12000 структур.

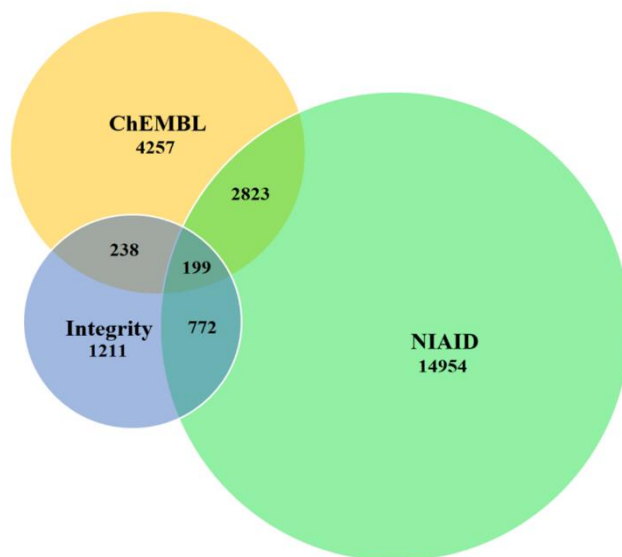


Рисунок 1 – Пересечение и общее число уникальных ингибиторов ВИЧ-1 после процедуры предобработки

Молекулярные дескрипторы

Для описания химической структуры при апробации методов мы использовали дескрипторы количественных атомных окрестностей (Quantitative Neighborhoods of Atoms – QNA). Дескрипторы QNA характеризуют каждый атом молекулы с учетом влияния на него всех остальных атомов молекулы (Filimonov D.A. et al. 2009). Получение одинакового числа дескрипторов для соединений выборки достигается проведением преобразований с использованием полиномов Чебышёва в зависимости от объема обучающей выборки. Сопоставление химического разнообразия выборок проводили с использованием дескрипторов многоуровневых атомных окрестностей (Multilevel Neighborhoods of Atoms – MNA) (Filimonov D.A. et al., 2009).

Самосогласованная регрессия

Для построения количественных и классификационных моделей были применены методы построения зависимостей «структура-активность», разработанные на основе самосогласованной регрессии (Self-Consistent Regression – SCR). Основным преимуществом SCR является строгая математическая обоснованность используемых оценок. Выбор значимых переменных происходит совместно с построением зависимостей. В задачах

построения 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 $p(\mathbf{a}|\mathbf{V})$ коэффициентов регрессии $\mathbf{a}$, где $\mathbf{V}$ – его параметры. Для нахождения апостериорного распределения вероятностей $p(\mathbf{a}|\mathbf{X}, \mathbf{y}, \mathbf{V})$ используется формула Байеса:

$$p(\mathbf{a}|\mathbf{X}, \mathbf{y}, \mathbf{V}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{a})p(\mathbf{a}|\mathbf{V})}{p(\mathbf{y}|\mathbf{X}, \mathbf{V})}$$

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

Коэффициенты регрессии $\mathbf{a}$ и параметры «регуляризации» $\mathbf{V}$ в SCR определяются согласованно:

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

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

Самосогласованная классификация

В нашем исследовании метод SCR был применен для построения количественных моделей. SCR может быть использована и для решения задач классификации, если для класса положительных примеров $y = 1$, а отрицательных $y = 0$, что позволяет провести сравнение с другими методами классификации.

Для решения задач классификации на основе SCR нами разработаны два новых метода – логистический SCLC и экстремальный SCEC классификаторы.

Большинство задач классификации соответствует схеме испытаний Бернулли. Для случая двух классов функция правдоподобия выборки имеет следующий вид:

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

где $P_k = P(\mathbf{x}_k, \mathbf{a})$ – вероятность положительного исхода (проявление активности) y -ого объекта выборки – химического соединения, которое представлено вектором дескрипторов $\mathbf{x}_k$. Логистическая регрессия следует из этого выражения, если:

$$P(\mathbf{x}, \mathbf{a}) = (1 + \text{Exp}(-\mathbf{x}'\mathbf{a}))^{-1}$$

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

$$-\text{Ln}(p(\mathbf{y}|\mathbf{X}, \mathbf{a})) = \sum_{k=1}^n \text{Ln}(1 + \text{Exp}(-u_k \mathbf{e}'_k \mathbf{X} \mathbf{a}))$$

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

Для неверного прогноза, характеризуемого $u_k \mathbf{e}'_k \mathbf{X} \mathbf{a} < 0$, функция «штрафа» $\text{Ln}(1 + \text{Exp}(-u_k \mathbf{e}'_k \mathbf{X} \mathbf{a}))$ возрастает почти линейно и стремится к нулю, для верных прогнозов $u_k \mathbf{e}'_k \mathbf{X} \mathbf{a} > 0$.

SCEC может быть получен заменой логистической функции на распределение экстремальных значений Гумбеля, чему соответствует обратный логарифм функции правдоподобия:

$$-\text{Ln}(p(\mathbf{y}|\mathbf{X}, \mathbf{a})) = \sum_{k=1}^n \text{Exp}(-u_k \mathbf{e}'_k \mathbf{X} \mathbf{a})$$

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

Статистическая регуляризация этих выражений при нормальном априорном распределении коэффициентов регрессии $p(\mathbf{a}|\mathbf{V})$ выполняется добавлением $\frac{1}{2}\mathbf{a}'\mathbf{V}\mathbf{a}$ к обратным логарифмам функции правдоподобия:

$$\text{SCLC: } -\text{Ln}(p(\mathbf{a}|\mathbf{X}, \mathbf{y}, \mathbf{V})) = \text{const} + \sum_{k=1}^n \text{Ln}(1 + \text{Exp}(-u_k \mathbf{e}'_k \mathbf{X} \mathbf{a})) + \frac{1}{2}\mathbf{a}'\mathbf{V}\mathbf{a}$$

$$\text{SCEC: } -\text{Ln}(p(\mathbf{a}|\mathbf{X}, \mathbf{y}, \mathbf{V})) = \text{const} + \sum_{k=1}^n \text{Exp}(-u_k \mathbf{e}'_k \mathbf{X} \mathbf{a}) + \frac{1}{2}\mathbf{a}'\mathbf{V}\mathbf{a}$$

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

Оценка коэффициентов регрессии осуществляется итеративно. На каждой итерации оценки коэффициентов регрессии $\mathbf{a}$ и параметров регуляризации $\mathbf{V}$ пересчитываются. При этом изменяются и веса наблюдений:

$$\text{SCLC: } W_{kk} = P(\mathbf{x}_k, \mathbf{a})(1 - P(\mathbf{x}_k, \mathbf{a})), \text{ SCEC: } W_{kk} = \text{Exp}(-u_k \mathbf{e}'_k \mathbf{X} \mathbf{a}).$$

Между итерациями происходит ранжирование используемых независимых переменных с учетом параметров регуляризации $\mathbf{V}$ и весов $\mathbf{W}$ и отбраковка незначимых переменных. Процесс завершается, когда коэффициенты регрессии $\mathbf{a}$ и параметры регуляризации $\mathbf{V}$

сходятся к своим итоговым значениям, устанавливается итоговый набор независимых переменных и эффективная размерность построенной зависимости.

Используемое программное обеспечение

Разработку 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.

Программа GUSAR (Generalized Unrestricted Structure-Activity Relationships (Generalized Unrestricted Structure-Activity Relationships) (Filimonov D.A. et al., 2009, Zakharov A.V. et al., 2014) 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.

Прогнозы спектра ВИЧ-ассоциированных заболеваний были нами получены с использованием программы PASS (Prediction of Activity Spectra for substance) (Филимонов и др., 2018).

Критерии оценки качества разрабатываемых 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 критериев: R^2 – коэффициент детерминации; Q^2 – коэффициент детерминации при перекрестном контроле; $R_{\text{тест}}^2$ – коэффициент детерминации на тестовой выборке.

Оценку точности и прогностической способности классификационных моделей выполняли на основе следующих критериев: BA – сбалансированная точность при перекрестном контроле; AUC – площадь под ROC кривой при перекрестном контроле; V – эффективная размерность модели.

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

$$EF_{\%} = \frac{N_{x\%}/N_{\%}}{N_a/N},$$

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.

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:

$$A_{\%} = \frac{1}{N_{\%}} \sum_{i=1}^{N_{\%}} A_i,$$

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 IC_{50} и нами подсчитывались их средние значения.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Валидация 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. Поскольку процесс оценки параметров моделей и параметров регуляризации итеративен, 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 2а), и эти примеры вносят наименьший вклад в получаемую модель. Для таких примеров быстро достигаются высокие абсолютные значения (рисунок 2б) 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.

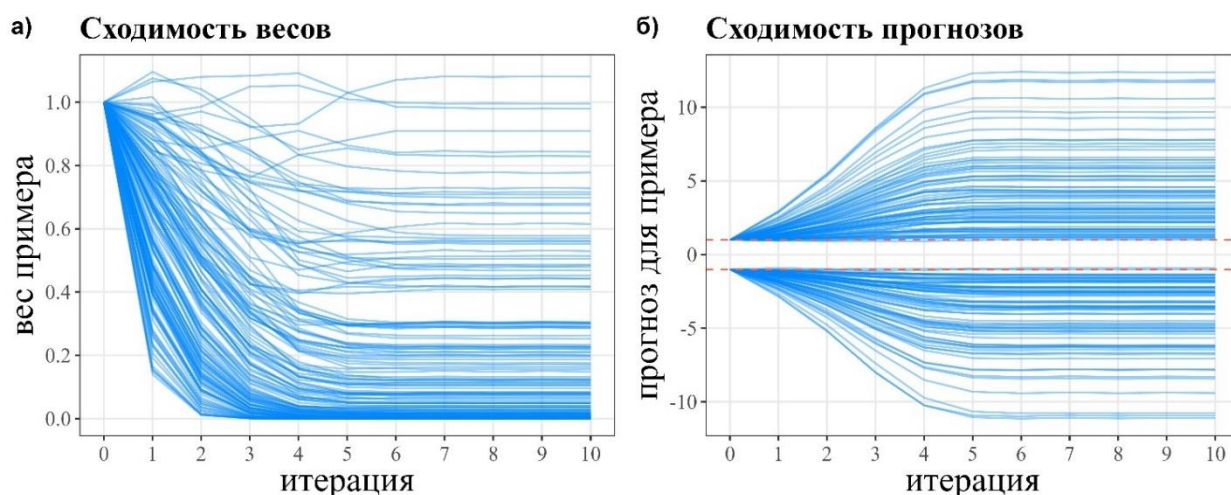


Рисунок 2 – Сходимость в процессе итераций SCEC: а) весов б) прогнозов

Было установлено, что разработанные нами методы позволяют восстанавливать значения сгенерированных коэффициентов регрессии. Проверка результатов осуществлялась на сгенерированных выборках с известными параметрами. Во всех произведенных симуляциях оценки параметров, получаемые SCLC и SCEC, отклонялись от действительных значений не более чем на 5%. Это позволило верифицировать работу алгоритмов. В SCLC и SCEC применяются разные оценки вероятности положительного исхода в схеме испытаний Бернулли, поэтому мы сравнивали значения весов соединений, в зависимости от разницы их наблюдаемых значений и порога классификации (рисунок 3).

Для оценки качества разрабатываемых методов мы построили классификационные модели SCLC, SCEC и SCR для 38 выборок из Tox21. Далее проводили перекрестный контроль и сравнивали сбалансированную точность. В случае использования SCEC были получены сравнимые (0.711 в среднем) или превосходящие SCR (0.707 в среднем) значения сбалансированной точности и сравнимые в случае с SCLC (0.704 в среднем). Как для SCLC, так и для SCEC сопоставимые результаты были получены при существенно меньшем числе независимых переменных по сравнению с SCR (81, 79 и 220 в среднем, соответственно).

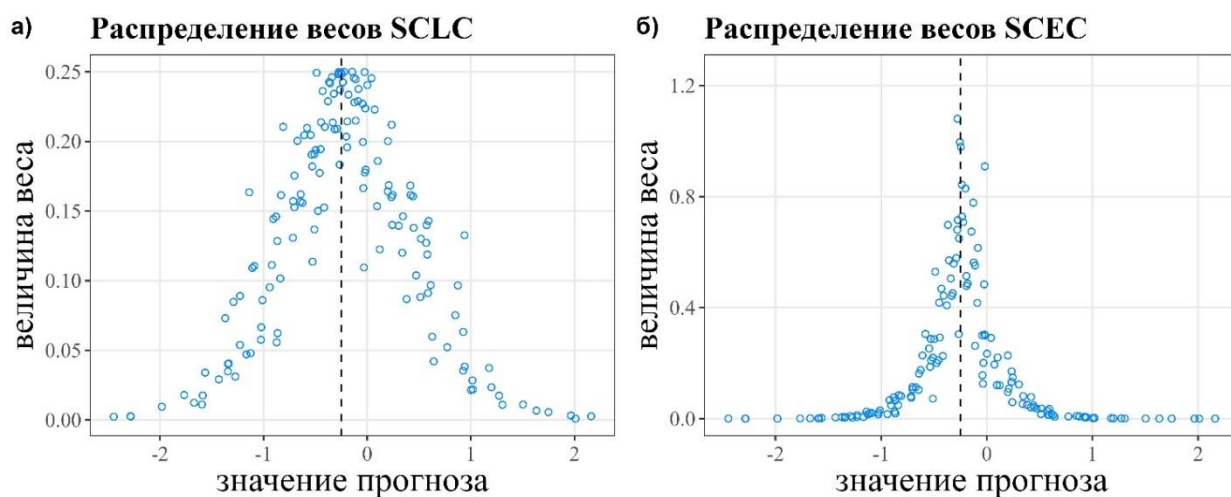


Рисунок 3 – Веса примеров в зависимости от прогноза для а) SCLC б) SCEC

QSAR модели ингибирования мишеней ВИЧ-1

Подготовленные нами выборки с данными об ингибиторах интегразы, протеазы и обратной транскриптазы ВИЧ-1 были использованы для построения QSAR моделей. Характеристики моделей, полученных с использованием выборок, содержащих данные NIAID и ChEMBL, NIAID и Integrity, представлены в таблице 1. Для оценки прогностической способности построенных количественных зависимостей «структура-активность» мы подготовили независимые тестовые выборки на основе источников данных, не входящих в обучающие выборки. Так, для выборок NIAID&ChEMBL тестовые выборки составлялись из Integrity источника с удалением дубликатов, входящих в обучающую выборку. Для выборок NIAID&Integrity использовали данные ChEMBL в качестве тестовой выборки. Сопоставление прогноза с наблюдаемыми величинами при валидации представлено на рисунке 4.

Из представленных на рисунке 4 данных видно, что результат валидации зависит от конкретной мишени. Менее высокие значения точности прогноза в случае ингибиторов интегразы и обратной транскриптазы ВИЧ-1 могут быть объяснены особенностями химических пространств в различных БД. Поскольку выборки химических соединений всегда покрывают лишь часть химического пространства, мы оценили новизну соединений, которая приводит к таким результатам, на примере тестовых выборок ChEMBL. С целью оценки новизны на уровне атомов и фрагментов использованы MNA дескрипторы. Для каждой структуры из ChEMBL оценивали процент новых дескрипторов MNA по сравнению с аналогичными данными из NIAID и Integrity (рисунок 5).

Таблица 1 – 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

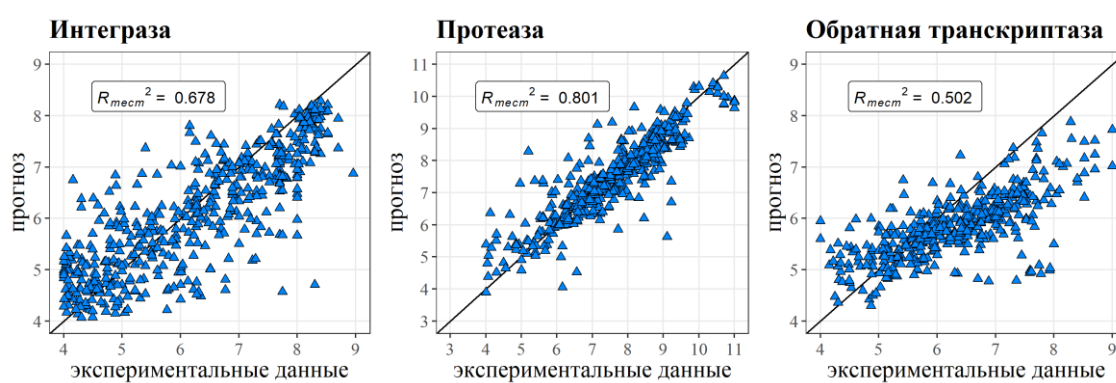


Рисунок 4 – Прогноз тестовых выборок ChEMBL моделями NIAID&Integrity

Для каждой мишени имеются новые дескрипторы относительно обучающих выборок. Для некоторых ингибиторов интегразы и обратной протеазы их доля превышает 15%. Для ингибиторов протеазы число новых дескрипторов составило 94 по сравнению с 220 и 199 для ингибиторов интегразы и обратной транскриптазы соответственно. Поэтому для покрытия всего химического пространства были использованы объединенные выборки, включающие информацию из трех БД. Характеристики моделей, полученных на этих выборках, представлены в таблице 2. Сопоставление результатов прогноза с экспериментальными данными представлено на рисунке 6. Так как для создания этих моделей использовались все данные, то оценка их качества проводилась на основе перекрестного контроля с разбиением выборок на 5 частей.

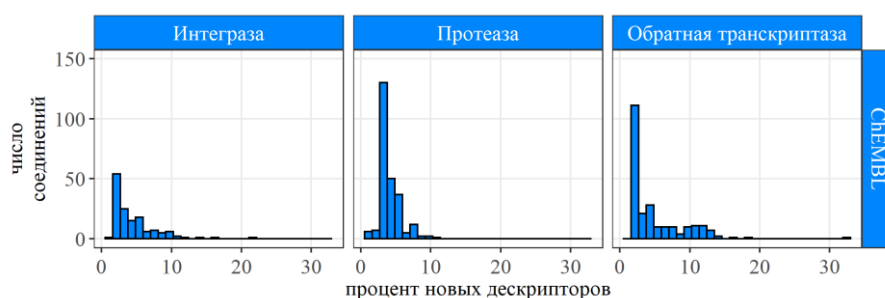


Рисунок 5 – Распределения новых дескрипторов в выборке ChEMBL относительно выборок NIAID и Integrity

Таблица 2 – Характеристики количественных моделей SCR

Выборка	R^2	Q^2	RMSD	V
IN (NIAID, ChEMBL и Integrity)	0.960	0.818	0.595	392
PR (NIAID, ChEMBL и Integrity)	0.954	0.824	0.709	470
RT (NIAID, ChEMBL и Integrity)	0.941	0.714	0.767	452

R^2 – коэффициент детерминации; Q^2 – коэффициент детерминации при перекрестном контроле; RMSD – среднеквадратичное отклонение прогноза; V – эффективная размерность.

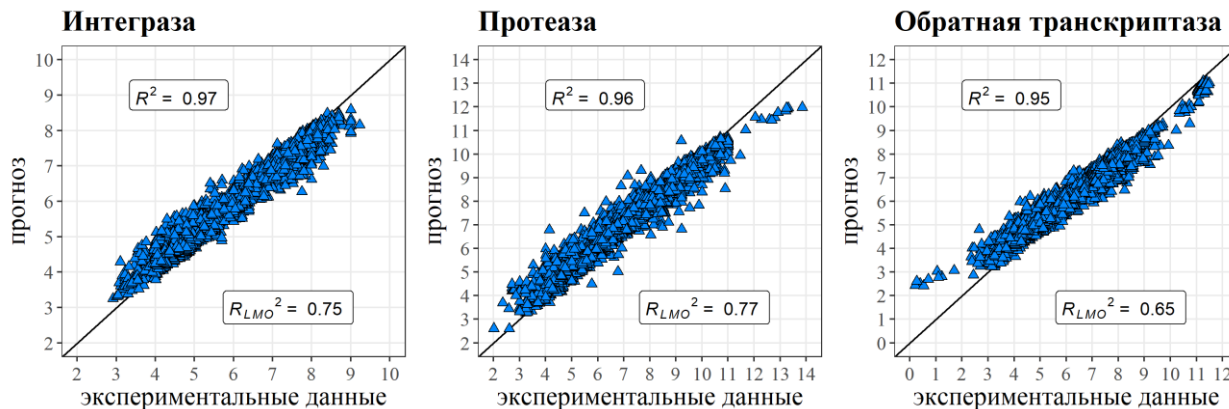


Рисунок 6 – Прогнозируемые и наблюдаемые величины pIC_{50} для количественных моделей SCR по выборкам: (A) IN, (B) PR, (C) RT

Построенные QSAR модели в случае, когда использованы все доступные экспериментальные данные, содержащие показатели IC_{50} , имеют удовлетворительные характеристики. Однако, в литературе имеется большое число данных, характеризующих активность другими показателями. Использование этих данных позволяет расширить

охватываемое химическое пространство и, соответственно, область применимости построенных классификационных моделей.

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

Используя разработанные нами методы SCLC и SCEC и выборки ингибиторов ВИЧ-1, SARS-CoV-2, мы построили классификационные модели. Качество полученных моделей сравнивали с показателями для SCR и SVM. В случае с SVM, отбор переменных осуществлялся по значениям t-критерия. Построение и валидацию моделей SCR и SVM проводили аналогичным образом.

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

Прогностическую способность моделей сравнивали по характеристикам сбалансированной точности (BA) и площади под ROC кривой при перекрестном контроле (AUC). Используемое при построении моделей число независимых переменных оценивали по величине эффективной размерности сравниваемых моделей. Из приведенных в таблицах данных видно, что модели SCLC и SCEC обладают сравнимыми с SCR значениями сбалансированной точности, однако, при этом имеют значительно меньшую размерность. BA в моделях SCEC превзошла таковую у SCR, в моделях SCLC оказалась несколько ниже во всех случаях.

Сравнение различных подходов для решения задач виртуального скрининга

Для демонстрации преимуществ использования классификационных методов мы обогатили обучающие выборки данными констант связывания (выборки (+K_i)). Другая группа выборок была составлена с использованием данных о константах связывания и добавлением полностью классификационных данных (выборки (+K_i и неактивные)). Соответствующее сравнение представлено в таблице 4. Из представленных в таблице результатах видно, что увеличение объема используемых данных и, соответственно, более широкий охват анализируемого химического пространства позволил достичь более высокой прогностической способности.

Таблица 3 – Сравнение моделей для ингибиторов ВИЧ-1 и SARS-CoV-2

Метод	N	V	BA	AUC	Метод	N	V	BA	AUC
IN					3CL				
SCR	4091	298	0.845	0.923	SCR	6230	370	0.851	0.919
SVM	4091	285	0.716	0.857	SVM	6230	373	0.834	0.901
PASS	4091	5331	0.829	0.920	PASS	6230	13037	0.849	0.920
ANN	4091	285	0.702	0.852	ANN	6230	373	0.830	
SCLC	4091	86	0.849	0.916	SCLC	6230	108	0.850	0.917
SCEC	4091	99	0.849	0.929	SCEC	6230	117	0.861	0.925
Метод	N	V	BA	AUC	Метод	N	V	BA	AUC
PR					PLpro				
SCR	6552	417	0.861	0.933	SCR	106	11	0.668	0.803
SVM	6552	373	0.802	0.893	SVM	106	12	0.660	0.789
PASS	6552	5867	0.861	0.933	PASS	106	1154	0.689	0.753
ANN	6552	373	0.798	0.828	ANN	106	12	0.662	0.799
SCLC	6552	123	0.858	0.910	SCLC	106	11	0.684	0.802
SCEC	6552	130	0.866	0.941	SCEC	106	11	0.682	0.812
RT					RdRp				
SCR	6309	360	0.788	0.867	SCR	2632	213	0.720	0.852
SVM	6309	368	0.745	0.855	SVM	2632	207	0.703	0.805
PASS	6309	7002	0.797	0.874	PASS	2632	7853	0.725	0.759
ANN	6309	368	0.760	0.863	ANN	2632	207	0.720	0.831
SCLC	6309	112	0.803	0.866	SCLC	2632	69	0.719	0.849
SCEC	6309	124	0.793	0.876	SCEC	2632	71	0.731	0.859

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

Так как предложенные методы позволяют ранжировать соединения, а, следовательно, обеспечивают возможность применять к анализируемым выборкам фильтры при проведении виртуального скрининга, были рассчитаны профили фактора обогащения (рисунок 7). В качестве анализируемых выборок использовались исключительно выборки с данными об IC₅₀.

Таблица 4 – Сравнение характеристик на перекрестном контроле для классификации QSAR моделью и классификационными моделями, построенными на обогащенных выборках, для ингибиторов ВИЧ-1

Метод	Выборка	N	V	BA	AUC
Интеграза ВИЧ-1					
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
Протеаза ВИЧ-1					
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
Обратная транскриптаза ВИЧ-1					
SCR	RT	6309	452	0.790	0.870
SCEC	IN 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

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 по pIC_{50} проводили с помощью модифицированной процедуры перекрестного контроля с разбиением выборки на 5 частей: 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 рисунке 8. Их можно интерпретировать таким образом, что наилучшие результаты получены для

классификационных моделей SCEC, обученных на гетерогенных данных без примеси неактивных соединений.

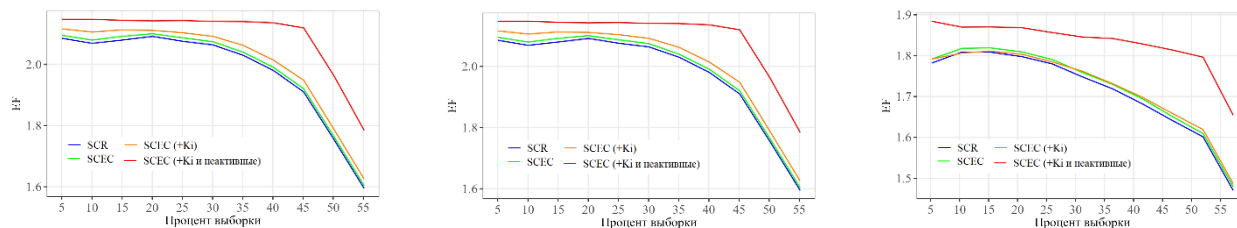


Рисунок 7 – Профили фактора обогащения для выборок: (A) IN, (B) PR, (C) RT

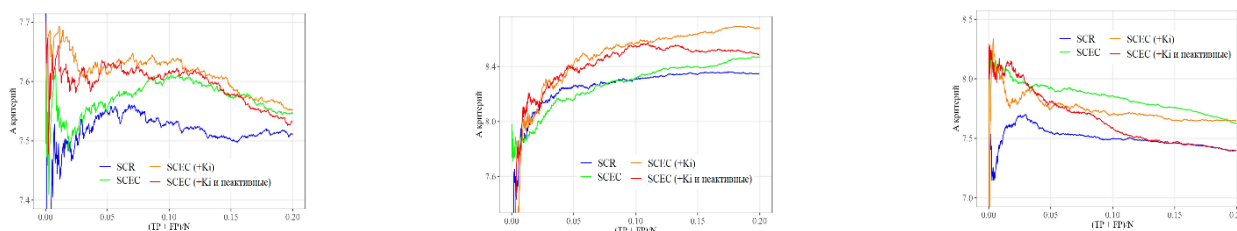


Рисунок 8 – Профили $A_{\%}$ по pIC_{50} : (A) IN, (B) PR, (C) RT

Веб-сервис AntiHIV Pred

На основе полученных (Q)SAR моделей был реализован свободно доступный веб-сервис AntiHIV Pred (<http://way2drug.com/HIV/>) (рисунок 9).

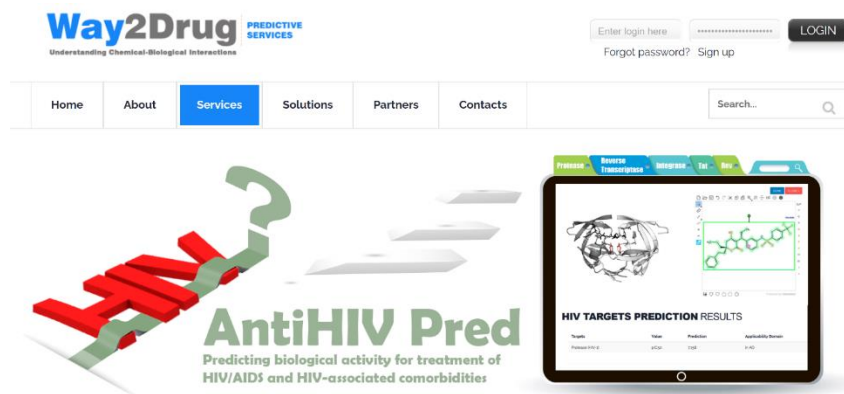


Рисунок 9 – Веб сервис AntiHIV Pred

Этот веб-сервис позволяет получить прогноз активности на основе структурной формулы соединения для каждой из мишеней ВИЧ-1 с использованием выбранной модели. В дополнение к анализу «структура-активность» для мишеней ВИЧ-1, веб-сервис позволяет прогнозировать профили активности соединений, связанные с терапией 83 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 предложенными ранее подходами, 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 также 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

1. Stolbov L.A., Filimonov, D.A., Poroikov, V.V. SAR based on Self Consistent Classifier // SAR and QSAS in Environmental Research. – 2022. – V. 33. – P. 793-804.
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.
7. Столбов Л.А., Филимонов Д.А., Поройков В.В. Программа Sarmath для ЭВМ. – 2022. – Свидетельство на программу. – Рег. Номер: RU2022662930.
8. 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.
9. 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.
10. Stolbov L.A., Filimonov D.A., Poroikov V.V. // XXXVIII Symposium of Bioinformatics and Computer-Aided Drug Discovery. Proceedings book. – Moscow, 2022. – P. 56.
11. 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.
12. 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.
13. 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.
 14. 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.
 15. 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.
 16. Poroikov V.V., Druzhilovskiy D.S., Filimonov D.A., Veselovsky A.A., 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.
 17. Stolbov L.A., Pogodin P.V., Druzhilovskiy D.S., Filimonov D.A., Poroikov V.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.
 18. 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.
 19. 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.
 20. 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.
 21. 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.